

# Stochastics, Statistics, Skedastics

Notes in Probability and its Applications

Robert Chin

June 5, 2021

# Contents

|   |            |
|---|------------|
| <b>Preface</b>  | <b>xix</b> |
| <b>I Fundamentals</b>   | <b>1</b>   |
| <b>1 Introductory Probability</b>                                   | <b>2</b>   |
| 1.1 Probability Laws . . . . .                                      | 2          |
| 1.1.1 Events . . . . .  | 2          |
| 1.1.2 Classical Definition of Probability . . . . .                 | 3          |
| 1.1.3 Addition Rule of Probability . . . . .                        | 3          |
| 1.1.4 Complementary Probabilities . . . . .                         | 3          |
| 1.1.5 Mutual Exclusivity . . . . .                                  | 3          |
| 1.1.6 Conditional Probability . . . . .                             | 3          |
| 1.1.7 Chain Rule of Probability . . . . .                           | 3          |
| 1.1.8 Law of Total Probability . . . . .                            | 4          |
| 1.1.9 Bayes' Theorem . . . . .                                      | 4          |
| 1.1.10 DeMorgan's Laws . . . . .                                    | 5          |
| 1.2 Counting . . . . .  | 5          |
| 1.2.1 Product Sets [83] . . . . .                                   | 5          |
| 1.2.2 Permutations . . . . .  | 6          |
| 1.2.3 Combinations . . . . .  | 7          |
| 1.2.4 Arrangements [41] . . . . .                                   | 8          |
| 1.3 Probability Distributions . . . . .                             | 9          |
| 1.3.1 Random Variables . . . . .                                    | 9          |
| 1.3.2 Distribution Functions . . . . .                              | 9          |
| 1.3.3 Joint Distributions . . . . .                                 | 13         |
| 1.3.4 Conditional Distributions . . . . .                           | 14         |
| 1.4 Expectation . . . . .   | 17         |
| 1.4.1 Linearity of Expectation . . . . .                            | 18         |
| 1.4.2 Conditional Expectation . . . . .                             | 18         |
| 1.4.3 Law of Iterated Expectations . . . . .                        | 19         |
| 1.4.4 Expectation of Indicator Random Variables . . . . .           | 20         |
| 1.4.5 Expectations Using Cumulative Distribution Function . . . . . | 20         |
| 1.4.6 Expectations Using Quantile Function . . . . .                | 21         |
| 1.5 Variance . . . . .  | 21         |
| 1.5.1 Standard Deviation . . . . .                                  | 22         |
| 1.5.2 Precision . . . . .   | 22         |
| 1.5.3 Covariance . . . . .  | 22         |
| 1.5.4 Correlation . . . . .   | 24         |
| 1.5.5 Conditional Variance . . . . .                                | 26         |
| 1.5.6 Law of Total Variance . . . . .                               | 27         |

|        |   |    |
|--------|---|----|
| 1.5.7  | Conditional Covariance . . . . .                                      | 27 |
| 1.5.8  | Law of Total Covariance . . . . .                                     | 28 |
| 1.6    | Independence . . . . .  | 28 |
| 1.6.1  | Independent Events . . . . .  | 28 |
| 1.6.2  | Independent Random Variables . . . . .                                | 29 |
| 1.6.3  | Uncorrelatedness . . . . .  | 31 |
| 1.6.4  | Mean Independence . . . . .   | 32 |
| 1.6.5  | Conditional Independence . . . . .                                    | 34 |
| 1.6.6  | Orthogonality [219] . . . . .   | 34 |
| 1.6.7  | Exchangeability . . . . .   | 35 |
| 1.6.8  | Variance Using Independent Copies . . . . .                           | 35 |
| 1.7    | Transformations of Random Variables . . . . .                         | 36 |
| 1.7.1  | Linear Transformations of Random Variables . . . . .                  | 36 |
| 1.7.2  | Parametric Distributions . . . . .                                    | 36 |
| 1.7.3  | Sums of Random Variables . . . . .                                    | 37 |
| 1.7.4  | Strictly Monotonic Transformations of Random Variables . . . . .      | 39 |
| 1.7.5  | Probability Integral Transform . . . . .                              | 40 |
| 1.7.6  | Inverse Transform Sampling . . . . .                                  | 40 |
| 1.7.7  | Gauss' Approximation Theorem [26] . . . . .                           | 41 |
| 1.7.8  | Variance-Stabilising Transforms . . . . .                             | 42 |
| 1.7.9  | Independence of Transformed Random Variables . . . . .                | 42 |
| 1.7.10 | Products of Random Variables . . . . .                                | 43 |
| 1.7.11 | Ratios of Random Variables . . . . .                                  | 44 |
| 1.7.12 | Decompositions of Random Variables . . . . .                          | 45 |
| 1.7.13 | Mixture Distributions . . . . .                                       | 46 |
| 1.7.14 | Compound Distributions . . . . .                                      | 46 |
| 1.7.15 | Truncated Distributions . . . . .                                     | 47 |
| 1.8    | Families of Continuous Univariate Probability Distributions . . . . . | 48 |
| 1.8.1  | Dirac Delta Distribution . . . . .                                    | 48 |
| 1.8.2  | Continuous Uniform Distribution . . . . .                             | 49 |
| 1.8.3  | Exponential Distribution . . . . .                                    | 49 |
| 1.8.4  | Gaussian Distribution . . . . .                                       | 51 |
| 1.8.5  | Laplace Distribution . . . . .  | 60 |
| 1.8.6  | Cauchy Distribution . . . . .   | 60 |
| 1.8.7  | Gamma Distribution . . . . .  | 61 |
| 1.8.8  | Beta Distribution . . . . .   | 63 |
| 1.8.9  | Chi-Squared Distribution . . . . .                                    | 64 |
| 1.8.10 | Student's $t$ Distribution . . . . .                                  | 67 |
| 1.8.11 | $F$ -Distribution . . . . .   | 70 |
| 1.8.12 | Pareto Distribution . . . . .   | 72 |
| 1.8.13 | Gumbel Distribution . . . . .   | 72 |
| 1.8.14 | Fréchet Distribution . . . . .  | 72 |
| 1.8.15 | Weibull Distribution . . . . .  | 73 |
| 1.8.16 | Logistic Distribution . . . . .                                       | 73 |
| 1.8.17 | Cantor Distribution . . . . .   | 73 |
| 1.9    | Families of Discrete Univariate Probability Distributions . . . . .   | 73 |
| 1.9.1  | Kronecker Delta Distribution . . . . .                                | 73 |
| 1.9.2  | Discrete Uniform Distribution . . . . .                               | 74 |
| 1.9.3  | Bernoulli Distribution . . . . .                                      | 74 |
| 1.9.4  | Rademacher Distribution . . . . .                                     | 74 |
| 1.9.5  | Binomial Distribution . . . . .                                       | 74 |

---

|          |   |           |
|----------|---|-----------|
| 1.9.6    | Categorical Distribution . . . . .                    | 76        |
| 1.9.7    | Poisson Distribution . . . . .                        | 77        |
| 1.9.8    | Skellam Distribution . . . . .                        | 78        |
| 1.9.9    | Geometric Distribution . . . . .                      | 78        |
| 1.9.10   | Hypergeometric Distribution . . . . .                 | 79        |
| 1.9.11   | Negative Hypergeometric Distribution . . . . .        | 81        |
| 1.9.12   | Negative Binomial Distribution . . . . .              | 81        |
| 1.9.13   | Beta-Binomial Distribution . . . . .                  | 84        |
| 1.9.14   | Benford Distribution . . . . .                        | 84        |
| 1.10     | Distribution Relationships [123] . . . . .            | 84        |
| 1.10.1   | Cauchy and Gaussian Distribution . . . . .            | 84        |
| 1.10.2   | Box-Muller Transform . . . . .                        | 85        |
| 1.10.3   | Exponential and Geometric Distribution . . . . .      | 86        |
| 1.10.4   | Beta and Gamma Distribution . . . . .                 | 87        |
| <b>2</b> | <b>Introductory Statistics</b>                        | <b>89</b> |
| 2.1      | Data Generating Processes . . . . .                   | 89        |
| 2.1.1    | Populations . . . . .                                 | 89        |
| 2.1.2    | Samples . . . . .                                     | 89        |
| 2.2      | Descriptive Statistics . . . . .                      | 90        |
| 2.2.1    | Measures of Central Tendency . . . . .                | 90        |
| 2.2.2    | Measures of Dispersion . . . . .                      | 93        |
| 2.2.3    | Measures of Dependence . . . . .                      | 98        |
| 2.2.4    | Measures of Shape . . . . .                           | 106       |
| 2.3      | Normal Statistics . . . . .                           | 109       |
| 2.3.1    | $z$ -Scores . . . . .                                 | 109       |
| 2.3.2    | Normal Sample Mean . . . . .                          | 109       |
| 2.3.3    | Normal Sample Variance . . . . .                      | 109       |
| 2.3.4    | Excess Kurtosis . . . . .                             | 111       |
| 2.3.5    | Qualitative Central Limit Theorem . . . . .           | 112       |
| 2.4      | Inferential Statistics . . . . .                      | 112       |
| 2.4.1    | Confidence Intervals . . . . .                        | 112       |
| 2.4.2    | Confidence Intervals on Population Mean . . . . .     | 113       |
| 2.4.3    | Confidence Intervals on Population Variance . . . . . | 114       |
| 2.4.4    | Prediction Intervals [137] . . . . .                  | 114       |
| 2.4.5    | Tolerance Intervals [137] . . . . .                   | 114       |
| 2.4.6    | Null Hypothesis Statistical Testing . . . . .         | 115       |
| 2.4.7    | Hypothesis Tests for Population Mean . . . . .        | 115       |
| 2.4.8    | Hypothesis Tests for Population Variance . . . . .    | 115       |
| 2.4.9    | Chi-Squared Goodness-of-Fit Testing . . . . .         | 115       |
| 2.5      | Two-Sample Inference . . . . .                        | 117       |
| 2.5.1    | Pooled Variance . . . . .                             | 117       |
| 2.5.2    | Matched Pairs $t$ -test . . . . .                     | 118       |
| 2.5.3    | Independent Samples Tests . . . . .                   | 118       |
| 2.5.4    | Fisher's Exact Test . . . . .                         | 121       |
| 2.6      | Simple Linear Regression . . . . .                    | 121       |
| 2.6.1    | Simple Least Squares Estimator . . . . .              | 122       |
| 2.6.2    | Coefficient of Determination . . . . .                | 122       |
| 2.6.3    | Inference for Linear Regressions . . . . .            | 124       |
| 2.7      | Design of Experiments . . . . .                       | 124       |
| 2.7.1    | Survey Methods . . . . .                              | 124       |

---

|          |  |            |
|----------|--|------------|
| 2.7.2    | Factorial Experiments . . . . .                                | 124        |
| 2.8      | Statistical Graphics . . . . .                                 | 124        |
| 2.8.1    | Scatter Plots . . . . .  | 124        |
| 2.8.2    | Histograms . . . . .   | 124        |
| 2.8.3    | Q-Q Plots . . . . .  | 124        |
| 2.9      | Method of Moments [72] . . . . .                               | 124        |
| 2.9.1    | Method of Moments for Normal Distribution . . . . .            | 125        |
| 2.9.2    | Method of Moments for Negative Binomial Distribution . . . . . | 126        |
| 2.9.3    | Method of Percentiles [60] . . . . .                           | 127        |
| <b>3</b> | <b>Intermediate Probability</b>                                | <b>128</b> |
| 3.1      | Random Vectors . . . . .                                       | 128        |
| 3.1.1    | Multivariate Probability Distributions . . . . .               | 128        |
| 3.1.2    | Mean Vectors . . . . .   | 131        |
| 3.1.3    | Covariance Matrices . . . . .                                  | 131        |
| 3.1.4    | Independence of Random Vectors . . . . .                       | 134        |
| 3.1.5    | Transformations of Random Vectors . . . . .                    | 135        |
| 3.2      | Families of Multivariate Probability Distributions . . . . .   | 140        |
| 3.2.1    | Multivariate Gaussian Distribution . . . . .                   | 140        |
| 3.2.2    | Multinomial Distribution . . . . .                             | 141        |
| 3.2.3    | Multivariate Hypergeometric Distribution . . . . .             | 143        |
| 3.2.4    | Dirichlet Distribution . . . . .                               | 143        |
| 3.2.5    | Dirichlet-Multinomial Distribution . . . . .                   | 145        |
| 3.2.6    | Multivariate Cauchy Distribution . . . . .                     | 145        |
| 3.2.7    | Elliptical Distributions . . . . .                             | 145        |
| 3.3      | Inequalities in Probability . . . . .                          | 145        |
| 3.3.1    | Boole's Inequality . . . . .                                   | 145        |
| 3.3.2    | Comparison of Random Variables . . . . .                       | 146        |
| 3.3.3    | Jensen's Inequality . . . . .                                  | 146        |
| 3.3.4    | Markov's Inequality . . . . .                                  | 147        |
| 3.3.5    | Chebychev's Inequality . . . . .                               | 149        |
| 3.3.6    | Gauss' Inequality [158] . . . . .                              | 151        |
| 3.3.7    | Vysochanskij-Petunin Inequality [158] . . . . .                | 151        |
| 3.3.8    | Cantelli's Inequality . . . . .                                | 151        |
| 3.3.9    | Cauchy-Schwarz Inequality . . . . .                            | 152        |
| 3.3.10   | Paley-Zygmund Inequality . . . . .                             | 153        |
| 3.3.11   | Hölder's Inequality . . . . .                                  | 153        |
| 3.3.12   | Minkowski's Inequality . . . . .                               | 154        |
| 3.3.13   | Lyapunov's Inequality . . . . .                                | 155        |
| 3.3.14   | Popoviciu's Inequality . . . . .                               | 155        |
| 3.3.15   | Bhatia-Davis Inequality . . . . .                              | 156        |
| 3.4      | Notions of Probabilistic Convergence . . . . .                 | 157        |
| 3.4.1    | Convergence in Distribution . . . . .                          | 157        |
| 3.4.2    | Convergence in Mean . . . . .                                  | 158        |
| 3.4.3    | Convergence in Mean Square . . . . .                           | 158        |
| 3.4.4    | Convergence in $p$ -Mean . . . . .                             | 158        |
| 3.4.5    | Convergence in Probability . . . . .                           | 158        |
| 3.4.6    | Almost Sure Convergence . . . . .                              | 160        |
| 3.4.7    | Complete Convergence . . . . .                                 | 160        |
| 3.4.8    | With High Probability . . . . .                                | 161        |
| 3.4.9    | Continuous Mapping Theorem . . . . .                           | 161        |

---

---

|  |            |
|--|------------|
| 3.4.10 Slutsky's Theorem . . . . .   | 161        |
| 3.4.11 Cramér-Wold Theorem [24] . . . . .                                      | 162        |
| 3.5 Moments . . . . .  | 162        |
| 3.5.1 Central Moments . . . . .  | 162        |
| 3.5.2 Standardised Moments . . . . .   | 162        |
| 3.5.3 Moment Generating Functions . . . . .                                    | 163        |
| 3.5.4 Sums of Random Variables with Moment Generating Functions . . . . .      | 165        |
| 3.5.5 Chernoff Bound . . . . .   | 167        |
| 3.5.6 Hoeffding's Lemma . . . . .  | 169        |
| 3.5.7 Hoeffding's Inequality . . . . .   | 170        |
| 3.5.8 Joint Moment Generating Functions . . . . .                              | 172        |
| 3.6 Probability Generating Functions . . . . .                                 | 173        |
| 3.6.1 Sums of Random Variables with Probability Generating Functions . . . . . | 174        |
| 3.6.2 Probability Generating Function of Poisson Distribution . . . . .        | 174        |
| 3.7 Characteristic Functions . . . . .   | 174        |
| 3.7.1 Sums of Random Variables with Characteristic Functions . . . . .         | 175        |
| 3.7.2 Subindependence . . . . .  | 175        |
| 3.7.3 Characteristic Functions of Gaussians . . . . .                          | 176        |
| 3.8 Cumulants . . . . .  | 176        |
| 3.8.1 Cumulant Generating Functions [186] . . . . .                            | 176        |
| 3.8.2 Law of Total Cumulance . . . . .   | 179        |
| 3.9 Exponential Families . . . . .   | 179        |
| 3.9.1 Single-Parameter Exponential Families . . . . .                          | 179        |
| 3.9.2 Multiple-Parameter Exponential Families . . . . .                        | 180        |
| 3.9.3 Multiple-Parameter Multivariate Exponential Families . . . . .           | 181        |
| <b>4 Intermediate Statistics</b> . . . . .                                     | <b>182</b> |
| 4.1 Multivariate Statistics . . . . .  | 182        |
| 4.1.1 Multivariate Sample Mean . . . . .                                       | 182        |
| 4.1.2 Multivariate Medians . . . . .   | 182        |
| 4.1.3 Sample Variance as Quadratic Forms . . . . .                             | 183        |
| 4.1.4 Sample Covariance Matrix . . . . .                                       | 184        |
| 4.1.5 Partial Correlation . . . . .  | 188        |
| 4.1.6 Mahalanobis Distance . . . . .   | 193        |
| 4.1.7 Higher Co-Moments [139] . . . . .  | 193        |
| 4.1.8 Confidence Regions . . . . .   | 194        |
| 4.2 Statistical Decision Theory . . . . .                                      | 194        |
| 4.2.1 Optimal Prediction . . . . .   | 194        |
| 4.2.2 Binary Hypothesis Testing . . . . .                                      | 196        |
| 4.2.3 Admissible Decision Rules [17] . . . . .                                 | 200        |
| 4.2.4 Unbiased Tests [126] . . . . .   | 201        |
| 4.2.5 Consistent Tests [126] . . . . .   | 201        |
| 4.2.6 Uniformly Most Powerful Tests . . . . .                                  | 201        |
| 4.2.7 Uncertainty Quantification . . . . .                                     | 201        |
| 4.3 Least Squares . . . . .  | 203        |
| 4.3.1 Ordinary Least Squares . . . . .   | 203        |
| 4.3.2 Weighted Least Squares . . . . .   | 213        |
| 4.3.3 Generalised Least Squares [205] . . . . .                                | 213        |
| 4.3.4 Total Least Squares . . . . .  | 214        |
| 4.3.5 Regularised Least Squares . . . . .                                      | 214        |
| 4.3.6 Constrained Least Squares . . . . .                                      | 218        |

---

|        |   |     |
|--------|---|-----|
| 4.3.7  | Recursive Least Squares . . . . .                                       | 219 |
| 4.3.8  | Partial Least Squares [175] . . . . .                                   | 221 |
| 4.3.9  | Stochastic Least Squares [206] . . . . .                                | 221 |
| 4.3.10 | Multiple Output Least Squares [80] . . . . .                            | 221 |
| 4.3.11 | Nonlinear Least Squares . . . . .                                       | 222 |
| 4.4    | Estimation Theory . . . . .   | 224 |
| 4.4.1  | Asymptotic Consistency . . . . .  | 224 |
| 4.4.2  | Weak Law of Large Numbers . . . . .                                     | 225 |
| 4.4.3  | Strong Law of Large Numbers . . . . .                                   | 226 |
| 4.4.4  | Sufficient Statistics . . . . .   | 229 |
| 4.4.5  | Ancillary Statistics [41] . . . . .                                     | 232 |
| 4.4.6  | Complete Statistics [41] . . . . .                                      | 232 |
| 4.4.7  | Basu's Theorem [41, 176] . . . . .                                      | 233 |
| 4.4.8  | U-Statistics [122] . . . . .  | 234 |
| 4.4.9  | Minimum Variance Unbiased Estimators . . . . .                          | 235 |
| 4.4.10 | Best Linear Unbiased Estimators . . . . .                               | 235 |
| 4.4.11 | Gauss-Markov Theorem . . . . .  | 236 |
| 4.4.12 | Rao-Blackwell Theorem [41] . . . . .                                    | 237 |
| 4.4.13 | Lehmann-Scheffé Theorem . . . . .                                       | 238 |
| 4.4.14 | Minimum Mean Square Error Estimators . . . . .                          | 239 |
| 4.4.15 | James-Stein Estimation . . . . .  | 240 |
| 4.5    | Maximum Likelihood Estimation . . . . .                                 | 240 |
| 4.5.1  | Likelihood Function . . . . .   | 240 |
| 4.5.2  | Maximum Likelihood Estimator . . . . .                                  | 241 |
| 4.5.3  | Asymptotic Consistency of Maximum Likelihood Estimator . . . . .        | 246 |
| 4.5.4  | Maximum Likelihood Justification of Least Squares . . . . .             | 247 |
| 4.5.5  | Maximum Likelihood Justification of Least Absolute Deviations . . . . . | 249 |
| 4.5.6  | Log Concavity of Likelihoods . . . . .                                  | 249 |
| 4.5.7  | Maximum Likelihood of Exponential Families [143] . . . . .              | 251 |
| 4.5.8  | Expectation Maximisation . . . . .                                      | 253 |
| 4.5.9  | M-Estimators [84] . . . . .   | 255 |
| 4.5.10 | Extremum Estimators [84] . . . . .                                      | 255 |
| 4.6    | Maximum Likelihood Inference . . . . .                                  | 255 |
| 4.6.1  | Score Function . . . . .  | 255 |
| 4.6.2  | Fisher Information . . . . .  | 256 |
| 4.6.3  | Observed Fisher Information . . . . .                                   | 257 |
| 4.6.4  | Fisher Information in Linear Regression . . . . .                       | 257 |
| 4.6.5  | Cramer-Rao Bound . . . . .  | 258 |
| 4.6.6  | Efficient Estimators . . . . .  | 259 |
| 4.6.7  | Asymptotic Normality of Maximum Likelihood Estimator . . . . .          | 259 |
| 4.6.8  | Standard Errors for Maximum Likelihood Estimator . . . . .              | 260 |
| 4.6.9  | Hypothesis Testing for Maximum Likelihood Estimation [136] . . . . .    | 262 |
| 4.7    | Multiple Hypothesis Testing . . . . .                                   | 262 |
| 4.7.1  | Multiple Competing Hypothesis Testing . . . . .                         | 262 |
| 4.7.2  | Multiple Simultaneous Hypothesis Testing . . . . .                      | 262 |
| 4.7.3  | Multiple Comparisons [189] . . . . .                                    | 264 |
| 4.7.4  | Analysis of Variance [169] . . . . .                                    | 264 |
| 4.7.5  | False Discovery Rate Control . . . . .                                  | 264 |
| 4.8    | Generalised Linear Models . . . . .                                     | 264 |
| 4.8.1  | Poisson Regression . . . . .  | 264 |
| 4.8.2  | Logistic Regression . . . . .   | 264 |

---

|          |   |            |
|----------|---|------------|
| 4.8.3    | Probit Regression . . . . .   | 265        |
| 4.8.4    | Multinomial Logistic Regression . . . . .                             | 265        |
| 4.8.5    | Ordinal Regression . . . . .  | 265        |
| 4.9      | Quantile Regression . . . . .   | 265        |
| <b>5</b> | <b>Advanced Probability</b>   | <b>266</b> |
| 5.1      | Multivariate Gaussian Properties . . . . .                            | 266        |
| 5.1.1    | Joint Moment Generating Function of Multivariate Gaussian . . . . .   | 266        |
| 5.1.2    | Uncorrelatedness and Independence of Multivariate Gaussians . . . . . | 267        |
| 5.1.3    | Marginal Gaussians and Joint Gaussians . . . . .                      | 267        |
| 5.1.4    | Conditional Gaussian Densities . . . . .                              | 268        |
| 5.1.5    | Product of Gaussian Densities . . . . .                               | 272        |
| 5.1.6    | Quotient of Gaussian Densities . . . . .                              | 275        |
| 5.1.7    | Marginalisation of Gaussians . . . . .                                | 275        |
| 5.1.8    | Exchangeability of Gaussian Sequences . . . . .                       | 278        |
| 5.1.9    | Bivariate Gaussian Properties . . . . .                               | 278        |
| 5.2      | Stochastic Processes . . . . .  | 281        |
| 5.2.1    | Properties of Stochastic Processes . . . . .                          | 282        |
| 5.2.2    | Stationarity . . . . .  | 284        |
| 5.2.3    | Ergodicity . . . . .  | 286        |
| 5.2.4    | Karhunen-Loëve Theorem . . . . .                                      | 288        |
| 5.3      | Families of Stochastic Processes . . . . .                            | 288        |
| 5.3.1    | Bernoulli Processes . . . . .   | 288        |
| 5.3.2    | Binomial Processes . . . . .  | 289        |
| 5.3.3    | Poisson Processes . . . . .   | 289        |
| 5.3.4    | Random Walks . . . . .  | 290        |
| 5.3.5    | Gaussian Processes . . . . .  | 291        |
| 5.3.6    | Wiener Process . . . . .  | 291        |
| 5.3.7    | Dirichlet Processes . . . . .   | 296        |
| 5.4      | Branching Processes . . . . .   | 296        |
| 5.4.1    | Galton-Watson Processes . . . . .                                     | 296        |
| 5.5      | Renewal Theory . . . . .  | 296        |
| 5.6      | Central Limit Theorems . . . . .                                      | 296        |
| 5.6.1    | Lindberg-Levy Central Limit Theorem [106] . . . . .                   | 296        |
| 5.6.2    | De Moivre-Laplace Theorem [171] . . . . .                             | 297        |
| 5.6.3    | Lindberg-Feller Central Limit Theorem . . . . .                       | 298        |
| 5.6.4    | Multivariate Central Limit Theorem [125] . . . . .                    | 298        |
| 5.6.5    | Finite Population Central Limit Theorem . . . . .                     | 298        |
| 5.6.6    | Functional Central Limit Theorem . . . . .                            | 300        |
| 5.6.7    | Stationary Process Central Limit Theorem [2] . . . . .                | 301        |
| 5.7      | Concentration Inequalities . . . . .                                  | 301        |
| 5.7.1    | Sub-Gaussian Random Variables . . . . .                               | 301        |
| 5.7.2    | Sub-Exponential Random Variables . . . . .                            | 307        |
| 5.7.3    | Sub-Gamma Random Variables [30] . . . . .                             | 307        |
| 5.7.4    | Azuma-Hoeffding Inequality [55] . . . . .                             | 310        |
| 5.7.5    | McDiarmid's Inequality [140] . . . . .                                | 312        |
| 5.7.6    | Bernstein's Inequality [30] . . . . .                                 | 314        |
| 5.8      | Large Deviations Theory . . . . .                                     | 316        |
| 5.9      | Random Matrices . . . . .   | 316        |
| 5.9.1    | Matrix Gaussian Distribution . . . . .                                | 316        |
| 5.9.2    | Gaussian Ensembles . . . . .  | 317        |

---

---

|          |   |            |
|----------|---|------------|
| 5.9.3    | Wishart Distribution . . . . .                          | 317        |
| <b>6</b> | <b>Bayesian Probability &amp; Statistics</b>            | <b>318</b> |
| 6.1      | Bayes' Theorem Extensions . . . . .                     | 318        |
| 6.1.1    | Bayes' Theorem for Multiple Events . . . . .            | 318        |
| 6.1.2    | Bayes' Theorem for Distributions . . . . .              | 318        |
| 6.1.3    | Bayes' Theorem for Mixed Distributions . . . . .        | 321        |
| 6.1.4    | Bayes' Theorem for Multivariate Distributions . . . . . | 321        |
| 6.2      | Bayesian Priors . . . . .                               | 322        |
| 6.2.1    | Principle of Indifference . . . . .                     | 322        |
| 6.2.2    | Cromwell's Rule . . . . .                               | 322        |
| 6.2.3    | Improper Priors . . . . .                               | 322        |
| 6.2.4    | Uninformative Priors . . . . .                          | 322        |
| 6.2.5    | Jeffrey's Priors . . . . .                              | 322        |
| 6.2.6    | Conjugate Priors . . . . .                              | 324        |
| 6.2.7    | Empirical Bayes . . . . .                               | 325        |
| 6.3      | Bayesian Updating . . . . .                             | 325        |
| 6.3.1    | Rule of Succession . . . . .                            | 325        |
| 6.3.2    | Odds Ratio Updating . . . . .                           | 326        |
| 6.3.3    | Log Odds Updating . . . . .                             | 327        |
| 6.3.4    | Bayes Filters [35, 198] . . . . .                       | 327        |
| 6.4      | Bayesian Inference . . . . .                            | 331        |
| 6.4.1    | Maximum a Posteriori Estimation . . . . .               | 331        |
| 6.4.2    | Bayes Estimators . . . . .                              | 331        |
| 6.4.3    | Credible Intervals . . . . .                            | 331        |
| 6.4.4    | Posterior Predictive Distributions . . . . .            | 331        |
| 6.4.5    | Bayes Factors [19] . . . . .                            | 331        |
| 6.4.6    | Bayesian Regularisation . . . . .                       | 332        |
| 6.4.7    | Bayesian Classifiers . . . . .                          | 334        |
| 6.4.8    | Bayesian Linear Regression [160] . . . . .              | 336        |
| 6.4.9    | Type II Maximum Likelihood [143] . . . . .              | 340        |
| 6.4.10   | Hierarchical Bayes Modelling [160] . . . . .            | 340        |
| 6.5      | Posterior Approximations . . . . .                      | 341        |
| 6.5.1    | Laplace's Approximation . . . . .                       | 341        |
| 6.5.2    | Expectation Propagation [65, 196] . . . . .             | 342        |
| 6.5.3    | Variational Inference [65] . . . . .                    | 345        |
| 6.5.4    | Posterior Sampling . . . . .                            | 348        |
| 6.6      | Bayesian Networks . . . . .                             | 349        |
| 6.6.1    | Factor Graphs . . . . .                                 | 349        |
| 6.6.2    | Probabilistic Graphical Models . . . . .                | 349        |
| 6.6.3    | Structure Learning in Graphical Models . . . . .        | 349        |
| 6.6.4    | Causality in Graphical Models . . . . .                 | 349        |
| 6.7      | Bernstein-von Mises Theorem [199] . . . . .             | 349        |
| 6.8      | Cox's Theorem [13, 100] . . . . .                       | 349        |
| 6.9      | Subjective Probability [87, 176] . . . . .              | 349        |
| 6.9.1    | Dempster-Shafer Theory . . . . .                        | 349        |
| <b>7</b> | <b>Markov Processes</b>                                 | <b>350</b> |
| 7.1      | Finite-State Discrete-Time Markov Chains . . . . .      | 350        |
| 7.1.1    | Markov Models . . . . .                                 | 350        |
| 7.1.2    | Stochastic Matrices . . . . .                           | 350        |
| 7.1.3    | Markov Chain Probabilities . . . . .                    | 352        |

---

---

|          |   |            |
|----------|---|------------|
| 7.1.4    | Markov Chain Properties . . . . .                                   | 354        |
| 7.1.5    | Absorbing Markov Chains . . . . .                                   | 356        |
| 7.1.6    | Stationary Distributions . . . . .                                  | 358        |
| 7.1.7    | Reversible Markov Chains [170] . . . . .                            | 367        |
| 7.1.8    | Strong Markov Property [149] . . . . .                              | 368        |
| 7.1.9    | Maximum Likelihood Estimation of Markov Chains . . . . .            | 368        |
| 7.2      | Infinite-State Discrete-Time Markov Chains . . . . .                | 368        |
| 7.2.1    | Countable-State Discrete-Time Markov Chains . . . . .               | 368        |
| 7.2.2    | Uncountable-State Discrete-Time Markov Chains [114] . . . . .       | 369        |
| 7.3      | Continuous-Time Markov Processes [3] . . . . .                      | 370        |
| 7.3.1    | Countable-State Continuous-Time Markov Processes . . . . .          | 370        |
| 7.3.2    | Uncountable-State Continuous-Time Markov Processes . . . . .        | 370        |
| 7.4      | Time-Inhomogeneous Markov Chains . . . . .                          | 370        |
| 7.5      | Hidden Markov Models . . . . .                                      | 370        |
| 7.5.1    | Discrete-Time Hidden Markov Models . . . . .                        | 370        |
| 7.5.2    | Forward Algorithm . . . . .   | 371        |
| 7.5.3    | Forward-Backward Algorithm [15] . . . . .                           | 373        |
| 7.5.4    | Hidden Markov Model Prediction . . . . .                            | 374        |
| 7.5.5    | Viterbi Algorithm . . . . .   | 375        |
| 7.5.6    | Baum-Welch Algorithm [76] . . . . .                                 | 377        |
| 7.5.7    | Hidden Markov Model Estimation by Method of Moments [114] . . . . . | 380        |
| 7.6      | Markov Decision Processes . . . . .                                 | 382        |
| 7.6.1    | Discrete-Time Markov Decision Processes . . . . .                   | 382        |
| 7.6.2    | Partially Observable Markov Decision Processes . . . . .            | 383        |
| 7.6.3    | Continuous-Time Markov Decision Processes . . . . .                 | 383        |
| 7.7      | Markov Networks . . . . .   | 383        |
| 7.7.1    | Belief Propagation . . . . .  | 383        |
| 7.8      | Semi-Markov Chains [153, 203] . . . . .                             | 383        |
| 7.9      | Quasistationary Distributions [45] . . . . .                        | 383        |
| <b>8</b> | <b>Measure Theoretic Probability</b> . . . . .                      | <b>384</b> |
| 8.1      | Probability Spaces . . . . .  | 384        |
| 8.1.1    | Concepts in Probability Spaces . . . . .                            | 384        |
| 8.1.2    | Probability Triple . . . . .  | 386        |
| 8.1.3    | Measurability . . . . .   | 387        |
| 8.1.4    | Borel-Cantelli Lemma . . . . .                                      | 389        |
| 8.2      | Lebesgue Integration . . . . .                                      | 391        |
| 8.2.1    | Riemann Integrability . . . . .                                     | 391        |
| 8.2.2    | Lebesgue Integral over Probability Spaces . . . . .                 | 393        |
| 8.2.3    | Monte-Carlo Characterisation of Lebesgue Integral [112] . . . . .   | 394        |
| 8.2.4    | Measure-Theoretic Random Variables . . . . .                        | 394        |
| 8.2.5    | Measure-Theoretic Expectation . . . . .                             | 396        |
| 8.2.6    | Fatou's Lemma . . . . .   | 396        |
| 8.2.7    | Dominated Convergence Theorem . . . . .                             | 396        |
| 8.3      | Radon-Nikodym Derivatives . . . . .                                 | 397        |
| 8.3.1    | Radon-Nikodym Theorem . . . . .                                     | 397        |
| 8.3.2    | Smoothing Law . . . . .   | 397        |
| 8.4      | Product Measures . . . . .  | 397        |
| 8.4.1    | Fubini's Theorem . . . . .  | 397        |
| 8.5      | Convergence of Probability Measures . . . . .                       | 397        |
| 8.6      | Measure Theoretic Stochastic Processes . . . . .                    | 397        |

---

|          |   |            |
|----------|---|------------|
| 8.6.1    | Concept of a Stochastic Process [9] . . . . .           | 397        |
| 8.6.2    | Filtrations . . . . .                                   | 398        |
| 8.6.3    | Martingales . . . . .                                   | 398        |
| 8.6.4    | Stopping Times . . . . .                                | 398        |
| 8.6.5    | Law of the Iterated Logarithm . . . . .                 | 398        |
| 8.6.6    | Doob Decomposition Theorem . . . . .                    | 398        |
| 8.7      | Ergodic Theory . . . . .                                | 398        |
| 8.7.1    | Birkhoff's Ergodic Theorem [112] . . . . .              | 398        |
| <b>9</b> | <b>Advanced Statistics</b>                              | <b>399</b> |
| 9.1      | Asymptotic Statistics . . . . .                         | 399        |
| 9.1.1    | Law of Large Numbers for Correlated Sequences . . . . . | 399        |
| 9.1.2    | Pointwise Convergence in Probability . . . . .          | 400        |
| 9.1.3    | Uniform Convergence in Probability . . . . .            | 401        |
| 9.1.4    | Delta Method [204] . . . . .                            | 401        |
| 9.1.5    | Weierstrass Approximation Theorem [161] . . . . .       | 402        |
| 9.1.6    | Edgeworth Series Expansions . . . . .                   | 402        |
| 9.2      | Empirical Measures . . . . .                            | 402        |
| 9.2.1    | Empirical Distribution Function . . . . .               | 402        |
| 9.2.2    | Glivenko-Cantelli Theorem . . . . .                     | 403        |
| 9.2.3    | Dvoretzky-Kiefer-Wolfowitz Inequality [113] . . . . .   | 404        |
| 9.2.4    | Kolmogorov-Smirnov Distance . . . . .                   | 405        |
| 9.3      | Order Statistics . . . . .                              | 405        |
| 9.3.1    | Distribution of a Single Order Statistic . . . . .      | 405        |
| 9.3.2    | Joint Distribution of Order Statistics . . . . .        | 407        |
| 9.3.3    | Conditional Distribution of Order Statistics . . . . .  | 409        |
| 9.3.4    | Spacings of Order Statistics . . . . .                  | 410        |
| 9.3.5    | Fisher-Tippett-Gnedenko Theorem [50] . . . . .          | 411        |
| 9.3.6    | Central Limit Theorem for Order Statistics . . . . .    | 412        |
| 9.3.7    | L-Statistics . . . . .                                  | 415        |
| 9.4      | Computational Statistics . . . . .                      | 416        |
| 9.4.1    | Monte-Carlo Estimation . . . . .                        | 416        |
| 9.4.2    | Acceptance-Rejection Sampling [165, 166] . . . . .      | 416        |
| 9.4.3    | Importance Sampling [115] . . . . .                     | 418        |
| 9.4.4    | Markov Chain Monte-Carlo . . . . .                      | 420        |
| 9.4.5    | Latin Hypercube Sampling . . . . .                      | 425        |
| 9.4.6    | Quasi Monte-Carlo Estimation [148] . . . . .            | 425        |
| 9.4.7    | Monte-Carlo Confidence Intervals [162] . . . . .        | 425        |
| 9.5      | Resampling Methods . . . . .                            | 425        |
| 9.5.1    | Jackknife . . . . .                                     | 425        |
| 9.5.2    | Bootstrapping . . . . .                                 | 426        |
| 9.5.3    | Bootstrap Confidence Intervals . . . . .                | 427        |
| 9.5.4    | Bootstrap Hypothesis Tests . . . . .                    | 428        |
| 9.5.5    | Permutation Tests . . . . .                             | 429        |
| 9.6      | Survival Analysis . . . . .                             | 432        |
| 9.6.1    | Survival Function . . . . .                             | 432        |
| 9.6.2    | Censored Data . . . . .                                 | 432        |
| 9.6.3    | Kaplan-Meier Estimator [89] . . . . .                   | 432        |
| 9.6.4    | Greenwood's Formula . . . . .                           | 432        |
| 9.7      | Nonparametric Statistics . . . . .                      | 432        |
| 9.7.1    | Nonparametric Mass Estimation . . . . .                 | 432        |

---

---

|           |   |            |
|-----------|---|------------|
| 9.7.2     | Nonparametric Density Estimation [91]                         | 432        |
| 9.7.3     | Nonparametric Regression . . . . .                            | 435        |
| 9.7.4     | Splines . . . . .   | 438        |
| 9.7.5     | Nonparametric Hypothesis Testing . . . . .                    | 438        |
| 9.7.6     | Nonparametric Confidence Intervals . . . . .                  | 441        |
| 9.8       | Robust Statistics . . . . .                                   | 443        |
| 9.8.1     | Robust Point Estimation . . . . .                             | 443        |
| 9.8.2     | Robust Regression . . . . .                                   | 443        |
| 9.8.3     | Sandwich Estimators . . . . .                                 | 443        |
| 9.8.4     | Robust Design . . . . .                                       | 443        |
| <b>10</b> | <b>Stochastic Calculus</b>                                    | <b>444</b> |
| 10.1      | Continuity of Stochastic Processes . . . . .                  | 444        |
| 10.1.1    | Continuity in Mean-Square . . . . .                           | 444        |
| 10.1.2    | Continuity in Probability . . . . .                           | 445        |
| 10.1.3    | Continuity in Distribution . . . . .                          | 445        |
| 10.1.4    | Continuity with Probability One . . . . .                     | 445        |
| 10.1.5    | Sample Continuity . . . . .                                   | 445        |
| 10.1.6    | Càdlàg Stochastic Processes . . . . .                         | 445        |
| 10.1.7    | Kolmogorov-Chentsov Continuity Theorem [12, 16] . . . . .     | 445        |
| 10.2      | Mean-Square Stochastic Calculus . . . . .                     | 445        |
| 10.2.1    | Differentiability in Mean-Square . . . . .                    | 445        |
| 10.2.2    | Integrability in Mean-Square . . . . .                        | 447        |
| 10.2.3    | Mean Square Stochastic Differential Equations [190] . . . . . | 448        |
| 10.3      | Continuous-Time Martingales . . . . .                         | 448        |
| 10.3.1    | Semimartingales . . . . .                                     | 448        |
| 10.3.2    | Doob-Meyer Decomposition Theorem . . . . .                    | 448        |
| 10.4      | Itô Calculus . . . . .  | 448        |
| 10.4.1    | Itô Integral [38] . . . . .                                   | 448        |
| 10.4.2    | Itô Processes . . . . .                                       | 448        |
| 10.4.3    | Itô's Lemma . . . . .   | 448        |
| 10.5      | Stratonovich Integral . . . . .                               | 448        |
| 10.6      | Stochastic Differential Equations . . . . .                   | 448        |
| 10.6.1    | Diffusions . . . . .  | 448        |
| 10.6.2    | Stochastic Partial Differential Equations . . . . .           | 448        |
| 10.6.3    | Backward Stochastic Differential Equations . . . . .          | 448        |
| 10.7      | Malliavin Calculus . . . . .                                  | 448        |
| 10.8      | Numerical Stochastic Differential Equations . . . . .         | 448        |
| 10.8.1    | Euler–Maruyama Method . . . . .                               | 448        |
| 10.8.2    | Milstein Method . . . . .                                     | 448        |
| 10.8.3    | Runge-Kutta Method . . . . .                                  | 448        |
| <b>11</b> | <b>Probabilistic Combinatorics</b>                            | <b>449</b> |
| 11.1      | Stirling's Approximation . . . . .                            | 449        |
| 11.2      | Inclusion-Exclusion Principle . . . . .                       | 449        |
| 11.2.1    | Probabilistic Inclusion-Exclusion Principle . . . . .         | 450        |
| 11.2.2    | Complementary Inclusion-Exclusion Principle . . . . .         | 450        |
| 11.2.3    | Bonferroni Inequalities [119] . . . . .                       | 451        |
| 11.3      | Pigeonhole Principle . . . . .                                | 451        |
| 11.4      | Partitions . . . . .  | 451        |
| 11.4.1    | Integer Compositions . . . . .                                | 451        |
| 11.4.2    | Stirling Numbers of the Second Kind . . . . .                 | 452        |

---

---

|           |   |            |
|-----------|---|------------|
| 11.4.3    | Bell Numbers . . . . .  | 454        |
| 11.5      | Catalan Numbers [119] . . . . .                                   | 454        |
| 11.6      | Derangements . . . . .  | 454        |
| 11.7      | Twelffold Way . . . . .   | 455        |
| 11.8      | Probabilisitic Method . . . . .                                   | 455        |
| 11.9      | Random Graphs . . . . .   | 455        |
| 11.9.1    | Erdős-Rényi Graphs . . . . .                                      | 455        |
| <b>II</b> | <b>Applications</b>   | <b>456</b> |
| <b>12</b> | <b>Information Theory</b>   | <b>457</b> |
| 12.1      | Entropy . . . . .   | 457        |
| 12.1.1    | Shannon Entropy . . . . .   | 457        |
| 12.1.2    | Differential Entropy . . . . .                                    | 458        |
| 12.1.3    | Joint Entropy . . . . .   | 458        |
| 12.1.4    | Conditional Entropy . . . . .                                     | 459        |
| 12.1.5    | Chain Rule of Entropy . . . . .                                   | 459        |
| 12.1.6    | Entropy of Functions . . . . .                                    | 460        |
| 12.1.7    | Cross Entropy . . . . .   | 463        |
| 12.1.8    | Entropy Rate . . . . .  | 463        |
| 12.1.9    | Asymptotic Equipartition Property . . . . .                       | 463        |
| 12.1.10   | Typicality . . . . .  | 464        |
| 12.1.11   | Rényi Entropy . . . . .   | 464        |
| 12.2      | Kullback-Leibler Divergence . . . . .                             | 465        |
| 12.2.1    | Gibbs' Inequality . . . . .                                       | 465        |
| 12.2.2    | Chain Rule of KL Divergence . . . . .                             | 467        |
| 12.2.3    | Mutual Information . . . . .                                      | 468        |
| 12.2.4    | Information Processing Inequality . . . . .                       | 470        |
| 12.2.5    | Asymptotic Equipartition Property for the KL Divergence . . . . . | 471        |
| 12.2.6    | Equivalence Between Minimum KL Divergence and MLE . . . . .       | 471        |
| 12.2.7    | Symmetrised KL Divergence . . . . .                               | 472        |
| 12.2.8    | Method of Types . . . . .   | 472        |
| 12.3      | Maximum Entropy Distributions [75] . . . . .                      | 477        |
| 12.3.1    | Principle of Maximum Entropy . . . . .                            | 477        |
| 12.3.2    | Maximum Entropy Distributions on Finite Support . . . . .         | 477        |
| 12.3.3    | Maximum Entropy Distributions on Bounded Support . . . . .        | 478        |
| 12.3.4    | Maximum Entropy Distributions on Unbounded Support . . . . .      | 478        |
| 12.3.5    | Maximum Entropy of Exponential Families [75, 143] . . . . .       | 478        |
| 12.4      | Coding Theory [47] . . . . .                                      | 478        |
| 12.4.1    | Source Coding . . . . .   | 478        |
| 12.4.2    | Channel Coding . . . . .  | 483        |
| 12.4.3    | Differential Privacy . . . . .                                    | 485        |
| 12.4.4    | Perplexity . . . . .  | 485        |
| 12.5      | Information Criteria . . . . .                                    | 486        |
| 12.5.1    | Akaike Information Criterion [130] . . . . .                      | 486        |
| 12.5.2    | Bayesian Information Criterion . . . . .                          | 488        |
| 12.5.3    | Deviance Information Criterion . . . . .                          | 489        |
| 12.6      | Optimal Experimental Design . . . . .                             | 489        |
| 12.6.1    | Optimal Experimental Design for Least Squares . . . . .           | 489        |
| 12.7      | Statistical Distances . . . . .                                   | 491        |
| 12.7.1    | Total Variation Distance . . . . .                                | 491        |

---

---

|           |  |            |
|-----------|--|------------|
| 12.7.2    | Hellinger Distance . . . . .                             | 495        |
| 12.7.3    | <i>f</i> -Divergence [49] . . . . .                      | 495        |
| 12.7.4    | Wasserstein Distance . . . . .                           | 496        |
| 12.8      | Algorithmic Information Theory . . . . .                 | 507        |
| 12.8.1    | Kolmogorov Complexity [47] . . . . .                     | 507        |
| 12.8.2    | Universal Probability [47] . . . . .                     | 510        |
| 12.8.3    | Minimum Description Length [47] . . . . .                | 513        |
| 12.9      | Information Geometry . . . . .                           | 514        |
| 12.9.1    | Statistical Manifolds [40] . . . . .                     | 514        |
| 12.9.2    | Fisher Information Metric . . . . .                      | 515        |
| 12.9.3    | Natural Gradients [97] . . . . .                         | 520        |
| 12.9.4    | Bregman Divergence [97] . . . . .                        | 522        |
| 12.9.5    | Information Projection . . . . .                         | 522        |
| <b>13</b> | <b>Econometrics</b> . . . . .                            | <b>523</b> |
| 13.1      | Economic Data . . . . .                                  | 523        |
| 13.1.1    | Generation of Economic Data . . . . .                    | 523        |
| 13.1.2    | Types of Economic Data . . . . .                         | 523        |
| 13.2      | Model Specification . . . . .                            | 523        |
| 13.2.1    | Causal Interpretation of Models . . . . .                | 523        |
| 13.2.2    | Statistical Interpretation of Models . . . . .           | 524        |
| 13.2.3    | Log-Level Models . . . . .                               | 525        |
| 13.2.4    | Level-Log Models . . . . .                               | 527        |
| 13.2.5    | Log-Log Models . . . . .                                 | 527        |
| 13.2.6    | Quadratic Models . . . . .                               | 528        |
| 13.2.7    | Interaction Terms [192] . . . . .                        | 528        |
| 13.2.8    | Specification Tests . . . . .                            | 528        |
| 13.2.9    | Quasi-Maximum Likelihood Estimates . . . . .             | 529        |
| 13.3      | Regression Analysis . . . . .                            | 530        |
| 13.3.1    | Sample Regression Function Coefficients . . . . .        | 530        |
| 13.3.2    | Unbiasedness of Ordinary Least Squares . . . . .         | 531        |
| 13.3.3    | Gauss-Markov Theorem in Econometrics . . . . .           | 531        |
| 13.3.4    | Asymptotic Normality of Ordinary Least Squares . . . . . | 532        |
| 13.3.5    | Consistency for Regression Variance . . . . .            | 533        |
| 13.3.6    | Homoskedasticity-Only Standard Errors . . . . .          | 534        |
| 13.3.7    | Heteroskedasticity . . . . .                             | 535        |
| 13.3.8    | Tests for Heteroskedasticity . . . . .                   | 535        |
| 13.3.9    | White Standard Errors . . . . .                          | 535        |
| 13.3.10   | Multicollinearity [72] . . . . .                         | 535        |
| 13.3.11   | Frisch-Waugh-Lovell Theorem [72] . . . . .               | 538        |
| 13.3.12   | <i>F</i> -Tests for Linear Restrictions . . . . .        | 540        |
| 13.3.13   | Wald Tests . . . . .                                     | 543        |
| 13.4      | Instrumental Variables Regression . . . . .              | 544        |
| 13.4.1    | Endogeneity . . . . .                                    | 544        |
| 13.4.2    | Omitted Variable Bias . . . . .                          | 544        |
| 13.4.3    | Measurement Errors [193] . . . . .                       | 545        |
| 13.4.4    | Simultaneous Causal Equations . . . . .                  | 548        |
| 13.4.5    | Systems of Simultaneous Causal Equations [72] . . . . .  | 549        |
| 13.4.6    | Two-Stage Least Squares . . . . .                        | 551        |
| 13.4.7    | Two-Stage Least Squares Inference . . . . .              | 555        |
| 13.5      | Panel Data Regression . . . . .                          | 558        |

---

|           |  |            |
|-----------|--|------------|
| 13.5.1    | Pooled Ordinary Least Squares . . . . .                                      | 558        |
| 13.5.2    | Fixed Effects Models . . . . .   | 558        |
| 13.5.3    | Differences-in-Differences Estimation . . . . .                              | 562        |
| 13.5.4    | Seemingly Unrelated Regressions . . . . .                                    | 564        |
| 13.5.5    | Random Effects Models [14] . . . . .   | 564        |
| 13.5.6    | Mixed Effects Models [14] . . . . .  | 565        |
| 13.6      | Time-Series Models . . . . .   | 565        |
| 13.6.1    | Autoregressive (AR) Models . . . . .   | 565        |
| 13.6.2    | Autoregressive Distributed Lag (ARDL) Models . . . . .                       | 570        |
| 13.6.3    | Moving Average (MA) Models . . . . .   | 570        |
| 13.6.4    | Autoregressive Moving Average (ARMA) Models . . . . .                        | 571        |
| 13.6.5    | Autoregressive Integrated Moving Average (ARIMA) Models . . . . .            | 576        |
| 13.6.6    | Autoregressive Moving Average with Exogenous Inputs (ARMAX) Models           | 578        |
| 13.6.7    | Vector Autoregressive (VAR) Models . . . . .                                 | 578        |
| 13.6.8    | Vector Autoregressive Exogenous (VARX) Models . . . . .                      | 578        |
| 13.6.9    | Nonlinear Autoregressive Exogeneous (NARX) Models . . . . .                  | 579        |
| 13.6.10   | Trend Models [95] . . . . .  | 579        |
| 13.6.11   | Seasonal Models . . . . .  | 579        |
| 13.7      | Time-Series Regression . . . . .   | 579        |
| 13.7.1    | AR Estimation . . . . .  | 579        |
| 13.7.2    | ARMA Estimation . . . . .  | 580        |
| 13.7.3    | ARIMA Estimation . . . . .   | 584        |
| 13.7.4    | VAR Estimation [132] . . . . .   | 584        |
| 13.7.5    | VARX Estimation [132] . . . . .  | 587        |
| 13.8      | Time-Series Analysis . . . . .   | 589        |
| 13.8.1    | Residual Autocorrelation . . . . .   | 589        |
| 13.8.2    | Structural Breaks . . . . .  | 589        |
| 13.8.3    | Unit Root . . . . .  | 590        |
| 13.8.4    | Cointegration . . . . .  | 597        |
| 13.8.5    | Spurious Regression . . . . .  | 597        |
| 13.8.6    | Seasonality [33, 42, 200, 222] . . . . .                                     | 598        |
| 13.8.7    | Box-Jenkins Method . . . . .   | 598        |
| 13.8.8    | Cholesky Impulse . . . . .   | 598        |
| 13.8.9    | Slutsky-Yule Effect . . . . .  | 598        |
| 13.9      | Time-Series Forecasting . . . . .  | 599        |
| 13.9.1    | Granger Causality . . . . .  | 599        |
| 13.9.2    | Innovations Algorithm [33] . . . . .   | 599        |
| 13.10     | Generalised Method of Moments . . . . .                                      | 601        |
| 13.10.1   | Ordinary Least Squares as Generalised Method of Moments . . . . .            | 602        |
| 13.10.2   | Maximum Likelihood as Generalised Method of Moments . . . . .                | 602        |
| 13.10.3   | Instrumental Variables Regression as Generalised Method of Moments . . . . . | 603        |
| 13.10.4   | Consistency of Generalised Method of Moments [72] . . . . .                  | 605        |
| 13.10.5   | Asymptotic Normality of Generalised Method of Moments [72] . . . . .         | 605        |
| 13.10.6   | Asymptotic Efficiency of Generalised Method of Moments . . . . .             | 606        |
| 13.10.7   | Sargan-Hansen Overidentifying Restrictions <i>J</i> -Test . . . . .          | 607        |
| <b>14</b> | <b>Machine Learning</b> . . . . .  | <b>609</b> |
| 14.1      | Concepts in Machine Learning . . . . .                                       | 609        |
| 14.1.1    | Machine Learning Datasets . . . . .  | 609        |
| 14.1.2    | Cross-Validation . . . . .   | 612        |
| 14.1.3    | Machine Learning Models [101] . . . . .                                      | 612        |

---

|   |     |
|---|-----|
| 14.2 Statistical Classification . . . . .                               | 612 |
| 14.2.1 Performance Metrics in Classification . . . . .                  | 612 |
| 14.2.2 Confusion Matrices . . . . .                                     | 618 |
| 14.2.3 Receiver Operating Characteristic . . . . .                      | 618 |
| 14.2.4 $k$ -Nearest Neighbours [80] . . . . .                           | 620 |
| 14.2.5 Linear Discriminant Analysis . . . . .                           | 620 |
| 14.2.6 Quadratic Discriminant Analysis . . . . .                        | 622 |
| 14.2.7 Support Vector Machines . . . . .                                | 622 |
| 14.2.8 Multiclass Classification . . . . .                              | 627 |
| 14.3 Unsupervised Learning . . . . .                                    | 628 |
| 14.3.1 $k$ -means Clustering . . . . .                                  | 628 |
| 14.3.2 Mode-Seeking Algorithms . . . . .                                | 630 |
| 14.3.3 Gaussian Mixture Models [25] . . . . .                           | 631 |
| 14.3.4 Dirichlet Process Mixtures . . . . .                             | 632 |
| 14.4 Artificial Neural Networks . . . . .                               | 632 |
| 14.4.1 Multi-Layer Perceptrons . . . . .                                | 632 |
| 14.4.2 Convolutional Neural Networks . . . . .                          | 638 |
| 14.4.3 Recurrent Neural Networks [70] . . . . .                         | 638 |
| 14.4.4 Mixture Density Networks . . . . .                               | 642 |
| 14.4.5 Generative Adversarial Networks [39] . . . . .                   | 642 |
| 14.5 Gaussian Process Regression . . . . .                              | 642 |
| 14.5.1 Gaussian Process Classification . . . . .                        | 642 |
| 14.6 Ensemble Methods . . . . .   | 642 |
| 14.6.1 Bagging . . . . .  | 642 |
| 14.6.2 Boosting . . . . .   | 642 |
| 14.6.3 Stacking [80, 221] . . . . .                                     | 643 |
| 14.6.4 Condorcet's Jury Theorem . . . . .                               | 643 |
| 14.7 Decision Trees [99] . . . . .                                      | 644 |
| 14.7.1 Regression Trees . . . . .                                       | 644 |
| 14.7.2 Classification Trees . . . . .                                   | 645 |
| 14.7.3 Random Forests . . . . .   | 646 |
| 14.8 Dimensionality Reduction . . . . .                                 | 646 |
| 14.8.1 Principal Component Analysis . . . . .                           | 646 |
| 14.8.2 Factor Analysis [103] . . . . .                                  | 648 |
| 14.8.3 Canonical Correlation Analysis . . . . .                         | 648 |
| 14.8.4 Random Projections . . . . .                                     | 650 |
| 14.8.5 Multidimensional Scaling [48] . . . . .                          | 655 |
| 14.8.6 $t$ -Distributed Stochastic Neighbour Embedding . . . . .        | 655 |
| 14.8.7 Autoencoders . . . . .   | 655 |
| 14.9 Statistical Learning Theory . . . . .                              | 656 |
| 14.9.1 Agnostic Probably Approximately Correct Learning [179] . . . . . | 656 |
| 14.9.2 Vapnik-Chervonenkis Dimension . . . . .                          | 661 |
| 14.9.3 Rademacher Complexity [140, 179] . . . . .                       | 662 |
| 14.9.4 Growth Function [140, 179] . . . . .                             | 667 |
| 14.9.5 Fundamental Theorem of Statistical Learning [179] . . . . .      | 668 |
| 14.9.6 Regression Generalisation Bounds [81] . . . . .                  | 668 |

---

---

|   |            |
|---|------------|
| <b>15 Statistical Signal Processing</b>                               | <b>669</b> |
| 15.1 Random Signals and Systems . . . . .                             | 669        |
| 15.1.1 Random Linear Time Invariant Systems . . . . .                 | 669        |
| 15.1.2 Discrete-Time Stochastic State-Space Models [114] . . . . .    | 673        |
| 15.1.3 Continuous-Time Stochastic State-Space Models . . . . .        | 676        |
| 15.2 Power Spectral Density . . . . .                                 | 676        |
| 15.2.1 Wiener-Khintchine Theorem [219] . . . . .                      | 677        |
| 15.2.2 Discrete-Time Wiener-Khintchine Theorem . . . . .              | 678        |
| 15.2.3 Cross Spectral Density . . . . .                               | 680        |
| 15.2.4 Spectral Density Characterisation of Filters . . . . .         | 680        |
| 15.2.5 Coloured Noise . . . . .                                       | 682        |
| 15.2.6 Spectral Factorisation Theorem [9] . . . . .                   | 685        |
| 15.2.7 Wold Decomposition Theorem [197] . . . . .                     | 688        |
| 15.2.8 Discretisation of Continuous-Time Stochastic Systems . . . . . | 688        |
| 15.3 Spectral Density Estimation [194] . . . . .                      | 688        |
| 15.3.1 Periodograms . . . . .   | 688        |
| 15.3.2 Correlograms . . . . .   | 688        |
| 15.3.3 Whittle Likelihood . . . . .                                   | 688        |
| 15.3.4 Maximum Entropy Spectral Density Estimation [47] . . . . .     | 688        |
| 15.4 Linear Filtering . . . . .                                       | 688        |
| 15.4.1 Linear Prediction Filter [219] . . . . .                       | 688        |
| 15.4.2 Matched Filtering [57, 197] . . . . .                          | 689        |
| 15.4.3 Wiener-Kolmogorov Filtering . . . . .                          | 691        |
| 15.4.4 Recursive Least Squares Filter [85, 189] . . . . .             | 695        |
| 15.4.5 Least Mean Squares Filter [85] . . . . .                       | 696        |
| 15.4.6 Deconvolution . . . . .  | 697        |
| 15.5 Kalman Filtering . . . . .                                       | 697        |
| 15.5.1 Linear Kalman Filter . . . . .                                 | 697        |
| 15.5.2 Linearised Kalman Filter [105] . . . . .                       | 703        |
| 15.5.3 Extended Kalman Filter . . . . .                               | 705        |
| 15.5.4 Unscented Kalman Filter . . . . .                              | 707        |
| 15.5.5 Information Filter . . . . .                                   | 710        |
| 15.5.6 Kalman-Bucy Filter . . . . .                                   | 710        |
| 15.5.7 Kalman Smoother [114, 184] . . . . .                           | 710        |
| 15.6 Particle Filtering . . . . .                                     | 710        |
| 15.6.1 Bootstrap Filter [53] . . . . .                                | 710        |
| 15.6.2 Sequential Importance Sampling [53, 196] . . . . .             | 711        |
| 15.6.3 Rao-Blackwellised Particle Filter [53, 114] . . . . .          | 714        |
| 15.6.4 Particle Smoothing [54] . . . . .                              | 715        |
| 15.7 Independent Component Analysis [96, 196] . . . . .               | 715        |
| 15.8 Wavelets [134, 210, 211] . . . . .                               | 715        |
| 15.8.1 Wavelet Denoising . . . . .                                    | 715        |
| 15.9 Compressed Sensing [62, 81] . . . . .                            | 715        |
| <b>16 Stochastic Control</b>  | <b>716</b> |
| 16.1 System Identification . . . . .                                  | 716        |
| 16.1.1 Quasistationary Signals [130] . . . . .                        | 716        |
| 16.1.2 Persistency of Excitation [206] . . . . .                      | 721        |
| 16.1.3 Frequency Domain Identification [98] . . . . .                 | 722        |
| 16.1.4 Identifiability of Dynamic Systems . . . . .                   | 722        |
| 16.1.5 Closed Loop Identification [116] . . . . .                     | 723        |

---

---

|   |            |
|---|------------|
| 16.1.6 Subspace Identification [130] . . . . .                            | 723        |
| 16.2 Queueing Theory [127] . . . . .                                      | 731        |
| 16.3 Stochastic Stability . . . . .                                       | 731        |
| 16.3.1 Moment Stability [5] . . . . .                                     | 731        |
| 16.3.2 Stability in Probability . . . . .                                 | 733        |
| 16.3.3 Stochastic Input-Output Stability . . . . .                        | 734        |
| 16.3.4 Almost Sure Stability [110] . . . . .                              | 734        |
| 16.3.5 Markov Chain Stability [138] . . . . .                             | 734        |
| 16.4 Stochastic Games . . . . .   | 734        |
| 16.4.1 Non-Sequential Decision under Uncertainty [20] . . . . .           | 734        |
| 16.4.2 Multi-Player Stochastic Games [146] . . . . .                      | 736        |
| 16.4.3 Stochastic Dynamic Games . . . . .                                 | 736        |
| 16.5 Stochastic Dynamic Programming . . . . .                             | 736        |
| 16.5.1 Stochastic Programming [180] . . . . .                             | 736        |
| 16.5.2 Stochastic Dynamic Programming over Finite Horizons [20] . . . . . | 738        |
| 16.5.3 Stochastic Dynamic Programming over Infinite Horizons . . . . .    | 741        |
| 16.6 Stochastic Optimal Control . . . . .                                 | 744        |
| 16.6.1 Hamilton-Jacobi-Bellman Equation [191] . . . . .                   | 744        |
| 16.6.2 Linear Quadratic Gaussian Control . . . . .                        | 744        |
| 16.6.3 Stochastic Model Predictive Control . . . . .                      | 744        |
| 16.7 Stochastic Approximation [7, 189] . . . . .                          | 744        |
| 16.7.1 Robbins-Monro Algorithm . . . . .                                  | 744        |
| 16.7.2 Stochastic Gradient Descent . . . . .                              | 746        |
| 16.7.3 Stochastic Average Approximation . . . . .                         | 747        |
| 16.7.4 Asymptotic Normality of Stochastic Approximation . . . . .         | 748        |
| 16.8 Multi-Armed Bandits . . . . .  | 748        |
| 16.8.1 Stochastic Bandits . . . . .                                       | 748        |
| 16.8.2 Regret . . . . .   | 748        |
| 16.8.3 Bandit Algorithms [121] . . . . .                                  | 750        |
| 16.8.4 Contextual Bandits . . . . .                                       | 758        |
| 16.8.5 Adversarial Bandits . . . . .                                      | 759        |
| 16.8.6 Non-Stationary Bandits . . . . .                                   | 759        |
| 16.8.7 Markovian Bandits . . . . .  | 760        |
| 16.8.8 Bayesian Bandits [185] . . . . .                                   | 761        |
| 16.9 Reinforcement Learning . . . . .                                     | 762        |
| 16.9.1 Markov Decision Problems . . . . .                                 | 762        |
| 16.9.2 Monte-Carlo Approximate Dynamic Programming . . . . .              | 764        |
| 16.9.3 Temporal Differences . . . . .                                     | 764        |
| 16.9.4 Value Function Approximation . . . . .                             | 766        |
| 16.9.5 Policy Gradients . . . . .   | 767        |
| 16.9.6 Actor-Critic Methods . . . . .                                     | 769        |
| <b>17 Quantitative Finance</b> . . . . .                                  | <b>770</b> |
| 17.1 Copulae . . . . .  | 770        |
| 17.1.1 Sklar's Theorem [102] . . . . .                                    | 771        |
| 17.1.2 Fréchet–Hoeffding Bounds [102] . . . . .                           | 771        |
| 17.1.3 Copula Density Functions . . . . .                                 | 773        |
| 17.1.4 Maximum Likelihood Copulae Fitting . . . . .                       | 774        |
| 17.1.5 Gaussian Copula . . . . .  | 775        |
| 17.1.6 Archimedean Copulae . . . . .                                      | 776        |
| 17.2 Heavy-Tailed Distributions [61] . . . . .                            | 778        |

---

---

|                     |  |            |
|---------------------|--|------------|
| 17.2.1              | Long-Tailed Distributions . . . . .  | 778        |
| 17.2.2              | Subexponential Distributions . . . . .   | 778        |
| 17.2.3              | States of Randomness . . . . .   | 778        |
| 17.3                | Stochastic Orders . . . . .  | 778        |
| 17.3.1              | First-Order Stochastic Dominance . . . . .   | 778        |
| 17.3.2              | Second-Order Stochastic Dominance . . . . .  | 781        |
| 17.3.3              | Higher-Order Stochastic Dominance [180] . . . . .                                  | 786        |
| 17.3.4              | Multivariate Stochastic Dominance [178] . . . . .                                  | 787        |
| 17.4                | Portfolio Optimisation . . . . .   | 788        |
| 17.4.1              | Kelly Criterion . . . . .  | 788        |
| 17.4.2              | Modern Portfolio Theory . . . . .  | 789        |
| 17.4.3              | Capital Asset Pricing Model [118] . . . . .  | 795        |
| 17.5                | Discrete-Time Derivatives Pricing [27, 182, 202] . . . . .                         | 795        |
| 17.5.1              | Binomial Trees . . . . .   | 795        |
| 17.6                | Continuous-Time Derivatives Pricing [93, 144, 217] . . . . .                       | 795        |
| 17.6.1              | Black-Scholes Model . . . . .  | 795        |
| 17.7                | Optimal Stopping . . . . .   | 795        |
| 17.7.1              | Odds Algorithm [36] . . . . .  | 795        |
| 17.7.2              | Changepoint Detection [114] . . . . .  | 795        |
| 17.7.3              | Optional Stopping Theorem . . . . .  | 795        |
| 17.8                | Ruin Theory . . . . .  | 795        |
| 17.9                | Financial Econometrics . . . . .   | 795        |
| 17.9.1              | Beta Regression . . . . .  | 795        |
| 17.9.2              | Autoregressive Conditional Heteroskedasticity (ARCH) Models [172] . . . . .        | 795        |
| 17.9.3              | Generalised Autoregressive Conditional Heteroskedasticity (GARCH) Models . . . . . | 800        |
| <b>18 Physics</b>   |  | <b>805</b> |
| 18.1                | Hamiltonian Monte-Carlo [34] . . . . .   | 806        |
| 18.2                | Statistical Mechanics . . . . .  | 806        |
| 18.2.1              | Maxwell-Boltzmann Distribution . . . . .   | 806        |
| 18.2.2              | Gibbs Distribution . . . . .   | 806        |
| 18.2.3              | Brownian Motion . . . . .  | 806        |
| 18.2.4              | Fokker-Planck Equations . . . . .  | 806        |
| 18.2.5              | Statistical Thermodynamics [112] . . . . .   | 806        |
| 18.3                | Statistical Physics . . . . .  | 806        |
| 18.3.1              | Mean Sojourn Time . . . . .  | 806        |
| 18.4                | Langevin Dynamics . . . . .  | 806        |
| 18.4.1              | Langevin Monte-Carlo . . . . .   | 806        |
| 18.5                | Mean Field Theory . . . . .  | 806        |
| 18.6                | Quantum Mechanics [78, 214, 216] . . . . .   | 806        |
| 18.6.1              | Quantum Probability . . . . .  | 806        |
| 18.6.2              | Quantum Computing . . . . .  | 806        |
| 18.6.3              | Quantum Stochastic Calculus . . . . .  | 806        |
| 18.7                | Econophysics [135] . . . . .   | 806        |
| 18.7.1              | Statistical Finance [208] . . . . .  | 806        |
| 18.7.2              | Quantum Finance . . . . .  | 806        |
| <b>Bibliography</b> |  | <b>807</b> |

---

# Preface

These are a collection of continuously-evolving notes in probability and statistics, that I started writing as a graduate student. They begin at a high-school or early-undergraduate introduction, and then go on to cover topics mainly from the undergraduate to early-graduate level.

They may be particularly useful for someone who:

- is studying probability and statistics from an electrical engineering, computer science or econometrics curriculum.
- wishes to use the notes as a handy reference, that complements other resources (much like an encyclopaedia).
- is a practitioner who wishes to gain deeper theoretical understanding in the methods and techniques they use.
- is interested in the overlaps and connections between different fields that apply probability/statistics.
- wants to use these notes as a model for writing their own collection of notes.

While all concepts in probability and statistics aim to be self-contained, some sections inevitably assume background knowledge and familiarity with outside topics such as (multivariate) calculus, real analysis, linear algebra, mathematical optimisation, and systems theory. Whenever one of these topics is invoked however, we usually refer to them by their commonly-known names, so that they can be easily looked up elsewhere. We will also occasionally make forward-references to sections later in the document, so the notes certainly do not need to be read strictly in the presented order.

In writing these notes, the goal was to strike a balance between rigour and pedagogy. To this end, we aim to motivate, develop intuition, and highlight connections behind the methods and formulae. At the same time, we do not shy away from deriving things from first principles or providing proofs. Occasionally however, when a proof is omitted, it may be because the result feels intuitive on its own, or perhaps because the proof is advanced, lengthy or not particularly instructive. We will at the very least try to explain why a result might hold intuitively or heuristically. Also as these notes are intended to be viewed digitally, space constraints are not a main issue, so proof/derivation steps are sometimes outlined in more detail than what printed textbooks may allow.

Etymology:

- *stochastic* from Greek, meaning to ‘aim at’ or to ‘guess’.
- *statistic* from Latin, meaning ‘of the state/council’.
- *skedastic* from Greek, meaning ‘scattered’.

Recently edited sections:

- GARCH models
- Second-order stochastic dominance
- Multivariate stochastic dominance
- Bias-complexity tradeoff

Upcoming sections to be edited:

- Bias-variance tradeoff
- Gaussian process regression
- Channel coding

Editing actions TODO:

- Add more links connecting sections
- Convert  $\exp[\cdot]$  and  $\log[\cdot]$  into  $\exp(\cdot)$  and  $\log(\cdot)$
- Convert  $dx$  into  $\mathrm{d}x$
- Minor typos

# Part I

## Fundamentals

# Chapter 1

## Introductory Probability

### 1.1 Probability Laws

#### 1.1.1 Events

When expressing probabilities, we express probabilities of occurrences of events. Notionally, we can think of some *random experiment* which leads to outcomes. Certain outcomes may be associated with a particular event, so we can represent an event as a set which contains as its elements the outcomes associated with it.

#### Unions

Given events  $A$  and  $B$ , the event of  $A$  or  $B$  occurring is given by the set union  $A \cup B$ .

#### Intersections

Given events  $A$  and  $B$ , the event of  $A$  and  $B$  both occurring is given by the set intersection  $A \cap B$ .

#### Complements

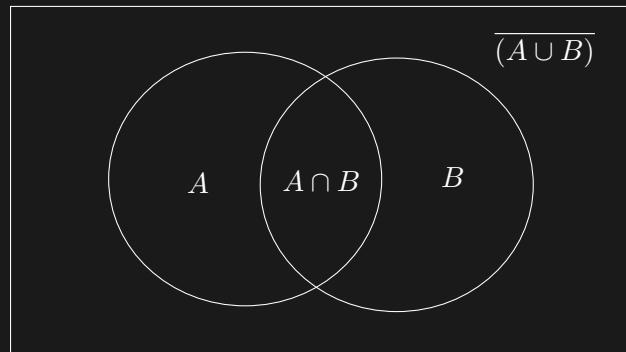
The complementary event of  $A$  is denoted  $\overline{A}$ . This is the event that  $A$  does not happen.

The relative complement of event  $A$  with respect to event  $B$  is denoted  $B \setminus A$ . This is the event of  $B$  occurring, but not  $A$  occurring. We have the relation

$$B \setminus A = B \cap \overline{A} \quad (1.1.1)$$

#### Venn Diagrams

Venn diagrams provide a way to visualise sets and events.



### 1.1.2 Classical Definition of Probability

In an experiment with outcomes given by the set  $\Omega$ , the outcomes associated with event  $A$  is a subset of  $\Omega$  (i.e.  $A \subseteq \Omega$ ). Suppose we have a way to count these outcomes. Let  $|\Omega|$  be the number of outcomes in  $\Omega$  (called the *cardinality* of  $\Omega$ ), and similarly let  $|A|$  be the number of outcomes in  $A$ . Then by the classical definition of probability, the probability of event  $A$ , denoted  $\Pr(A)$ , is

$$\Pr(A) = \frac{|A|}{|\Omega|} \quad (1.1.2)$$

This means that probabilities are always between 0 and 1 (inclusive). We can interpret this probability by saying if we were able to conduct an infinite number of experiments, then the proportion of experiments which resulted in the occurrence of  $A$  would be  $\Pr(A)$ .

### 1.1.3 Addition Rule of Probability

The probability of the intersection of  $A$  and  $B$  is written  $\Pr(A \cap B)$ , or alternatively may be denoted  $\Pr(A, B)$ . To obtain the probability of the union between  $A$  and  $B$ , we have the formula (called the addition rule of probability):

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \quad (1.1.3)$$

### 1.1.4 Complementary Probabilities

The complementary probability of  $\Pr(A)$  is denoted  $\Pr(\bar{A})$ . We have that

$$\Pr(\bar{A}) = 1 - \Pr(A) \quad (1.1.4)$$

### 1.1.5 Mutual Exclusivity

Events  $A$  and  $B$  are mutually exclusive if they cannot both happen together. That means

$$\Pr(A \cap B) = 0 \quad (1.1.5)$$

Hence the addition rule of probability in the case of mutual exclusivity simplifies to

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \quad (1.1.6)$$

### 1.1.6 Conditional Probability

The probability of event  $A$  occurring given event  $B$  has occurred is denoted by  $\Pr(A|B)$  and is known as a conditional probability. The conditional probability can be calculated by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (1.1.7)$$

The term  $\Pr(B)$  can be thought of as a ‘normalising constant’ for  $\Pr(A \cap B)$ , as we only want to consider the space of outcomes where  $B$  has occurred.

### 1.1.7 Chain Rule of Probability

Given the conditional probability  $\Pr(A|B)$  and the marginal probability  $\Pr(B)$ , the chain rule of probability says that the *joint* probability  $\Pr(A \cap B)$  is equal to

$$\Pr(A \cap B) = \Pr(A|B) \Pr(B) \quad (1.1.8)$$

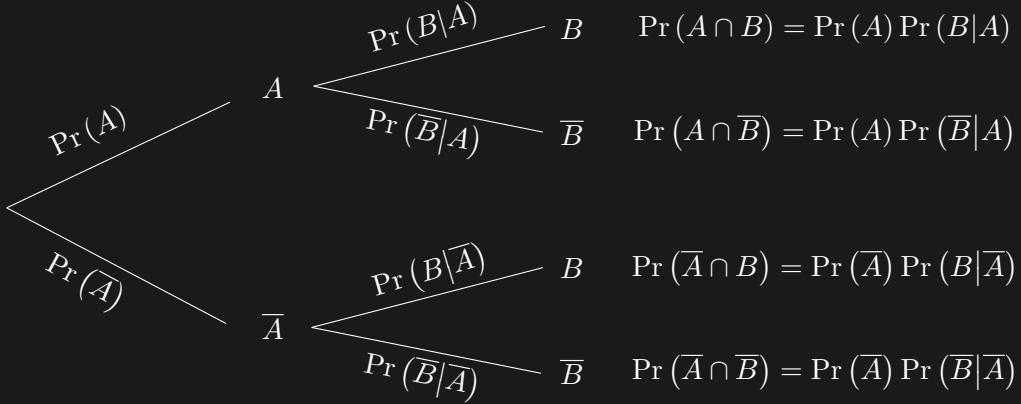
### 1.1.8 Law of Total Probability

The law of total probability gives a way to write the probability of an event  $B$  in terms of its conditional probabilities, using the chain rule of probability. For events  $A$  and  $B$ , the law of total probability says that

$$\Pr(B) = \Pr(A)\Pr(B|A) + \Pr(\bar{A})\Pr(B|\bar{A}) \quad (1.1.9)$$

### Probability Tree Diagrams

A probability tree diagram for two events  $A, B$  is given below.



Probability tree diagrams allow for the law of total probability to be visualised. One can think of obtaining  $\Pr(B)$  across all branches to the tree to find the joint probabilities, then summing over the probabilities which are favourable to event  $B$ .

### 1.1.9 Bayes' Theorem

By definitions of conditional probability, we have for two events  $A, B$ :

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (1.1.10)$$

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \quad (1.1.11)$$

Rearranging both gives

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) \quad (1.1.12)$$

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A) \quad (1.1.13)$$

Then equating them gives Bayes' theorem

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A) \quad (1.1.14)$$

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \quad (1.1.15)$$

where  $\Pr(A|B)$  is known as the posterior probability,  $\Pr(A)$  is known as the prior probability,  $\Pr(B|A)$  is known as the likelihood and  $\Pr(B)$  is known as the marginal likelihood.

### 1.1.10 DeMorgan's Laws

DeMorgan's laws can be arrived at via some logical postulates.

- The event of  $A$  or  $B$  not happening is the same as neither  $A$  happening nor  $B$  happening.
- The event that  $A$  and  $B$  both do not happen is the same as either  $A$  not happening or  $B$  not happening.

This can be written down using set notation:

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \quad (1.1.16)$$

$$\overline{A \cap B} = \overline{A} \cup \overline{B} \quad (1.1.17)$$

And in terms of probability:

$$\Pr(\overline{A \cup B}) = \Pr(\overline{A} \cap \overline{B}) \quad (1.1.18)$$

$$\Pr(\overline{A \cap B}) = \Pr(\overline{A} \cup \overline{B}) \quad (1.1.19)$$

Or by using complements:

$$1 - \Pr(A \cup B) = \Pr(\overline{A} \cap \overline{B}) \quad (1.1.20)$$

$$1 - \Pr(A \cap B) = \Pr(\overline{A} \cup \overline{B}) \quad (1.1.21)$$

## 1.2 Counting

In the context of the classical definition of probability, combinatorics allows us to count the number of outcomes, relating to both the number of outcomes in an experiment, and the number of outcomes favourable to an event.

### 1.2.1 Product Sets [83]

For sets  $A$  and  $B$ , the product  $A \times B$  is the set of all ordered (i.e. order matters) pairs  $(a, b)$ , where  $a$  is from  $A$  and  $b$  is from  $B$ .

### Fundamental Theorem of Counting [41]

Consider a sequence of symbols where the  $j^{\text{th}}$  symbol in the sequence comes from an ‘alphabet’ of size  $n_j$ . Then for a sequence of length  $m$ , there are  $n_1 \times n_2 \times \dots \times n_m$  different ways to form such sequences without restriction. This can be used to count the number of elements in a product set, and calculate the total number of outcomes at the end of a probability tree. For instance, if sets  $A_1, \dots, A_m$  have  $n_1, \dots, n_m$  elements respectively, then

$$|A_1 \times \dots \times A_m| = |A_1| \times \dots \times |A_m| \quad (1.2.1)$$

$$= n_1 \times \dots \times n_m \quad (1.2.2)$$

As a special case, if each of  $A_1, \dots, A_m$  have the same number of elements as the set  $A$  with  $n$  elements, then

$$|A_1 \times \dots \times A_m| = |A|^m \quad (1.2.3)$$

$$= n^m \quad (1.2.4)$$

### 1.2.2 Permutations

Consider the number of different ways to order  $n$  different symbols. Each way is called a permutation. There are  $n$  possibilities for the first symbol,  $n - 1$  possibilities for the second symbol, etc. Hence there are

$$n \times (n - 1) \times \cdots \times 2 \times 1 = n! \quad (1.2.5)$$

different permutations.

### Permutations in a Circle

Consider the number of different ways to arrange  $n$  distinct objects in a circle (where it only matters which object is next to which). The first object can be placed in any arbitrary position. There are  $n - 1$  possibilities for the second object placed next to the first object, and  $n - 2$  possibilities for the third object placed next to the second object, etc. Hence there are

$$(n - 1) \times \cdots \times 2 \times 1 = (n - 1)! \quad (1.2.6)$$

different ways. Note that if it did matter which position in the circle each object was placed, the number of arrangements becomes the same as the number of arrangements in a line.

### $k$ -Permutations

Consider the number of different ways to pick  $k$  objects from a pool of  $n$  distinct objects, where the order in which they are picked matters. There are  $n$  possibilities for the first pick,  $n - 1$  possibilities for the second pick, up to  $n - k + 1$  possibilities for the  $k^{\text{th}}$  pick. Hence there are

$$n \times (n - 1) \times \cdots \times (n - k + 1) = \frac{n!}{(n - k)!} \quad (1.2.7)$$

different ways. This is called the number of  $k$ -permutations (sometimes referred to as just the number of permutations), and may be denoted as  ${}^n P_k$  (which can be informally read as ‘ $n$  pick  $k$ ’). This also gives the number of different ways to arrange  $k$  distinct objects among  $(n - k)$  other homogeneous objects in a line (i.e. there are  $n$  objects total). This is because there are  $n$  possible locations for the first object, up to  $n - k + 1$  possible locations for the  $k^{\text{th}}$  object.

### $k$ -Permutations in a Circle

With  $n$  total objects, consider the number of different ways to arrange  $k$  distinct objects among  $(n - k)$  other homogeneous objects in a circle (where it only matters which object is next to which). The first object can be placed in any arbitrary position. There are  $n - 1$  possible positions for where the second object can be placed relative to first object,  $n - 2$  possible positions for where the third object can be placed relative to first object, up to  $n - k + 1$  possible positions for where the  $k^{\text{th}}$  object can be placed relative to first object. Hence there are

$$(n - 1) \times \cdots \times (n - k + 1) = \frac{(n - 1)!}{(n - k)!} \quad (1.2.8)$$

different ways. Note that we can arrive at this number by counting  ${}^n P_k$  permutations in a circle where absolute position does matter (equivalent to permutations in a line), but then dividing by the possible number of rotations in a circle without affecting relative positions, which is  $n$ .

### 1.2.3 Combinations

#### Binomial Coefficient

Consider the number of different ways to pick  $k$  objects from a pool of  $n$  distinct objects, where the order in which they are picked does not matter (i.e. it only matters which group of  $k$  are picked). We know that there are  ${}^n P_k$  different ways to pick  $k$  objects when order does matter, but this number counts all the  $k!$  different ways for every possible grouping of  $k$  objects from the pool. So dividing the number of permutations by  $k!$  gives

$$\frac{{}^n P_k}{k!} = \frac{n!}{k! (n-k)!} \quad (1.2.9)$$

which is the number of different groups of  $k$  can be picked from  $n$  distinct objects. This is known as the binomial coefficient, and denoted  ${}^n C_k$ , which may also be read as “ $n$  choose  $k$ ”. It is also sometimes denoted

$${}^n C_k = \binom{n}{k} \quad (1.2.10)$$

This number can also be thought of as the number of different ways to arrange  $k$  homogeneous objects among  $(n-k)$  other homogeneous objects in a line. Note that the following properties hold for the binomial coefficient:

$${}^n C_k = {}^n C_{n-k} \quad (1.2.11)$$

which says that asking for the number of ways to choose  $k$  from  $n$  is no different to asking for the number of ways to choose  $n-k$  from  $n$ . Also

$${}^n C_0 + {}^n C_1 + \cdots + {}^n C_n = 2^n \quad (1.2.12)$$

since there are  $2^n$  different binary sequences of length  $n$ , and the above sum exhaustively enumerates through all combinations of binary sequences from when there are zero ‘1’s up to  $n$  ‘1’s.

#### Multinomial Coefficient

As a generalisation to the binomial coefficient, suppose there are  $n$  objects and  $m$  different groups, with the numbers of homogeneous objects in each group given by  $k_1, k_2, \dots, k_m$  and so we have

$$k_1 + k_2 + \cdots + k_m = n \quad (1.2.13)$$

To derive the number of different ways these objects can be arranged, first consider the number of ways to arrange only the first group (which has  $k_1$  members), without regard for the other groups (call it the ‘remainder’ group). By the binomial coefficient, there are

$${}^n C_{k_1} = \frac{n!}{k_1! (n-k_1)!} \quad (1.2.14)$$

ways. Then for every one of these arrangements, there are  ${}^{n-k_1} C_{k_2}$  ways to arrange members of group 2 among the remainder. Hence there are

$${}^n C_{k_1} \times {}^{n-k_1} C_{k_2} = \frac{n!}{k_1! (n-k_1)!} \times \frac{(n-k_1)!}{k_2! (n-k_1-k_2)!} \quad (1.2.15)$$

different ways to arrange objects in groups 1 and 2, without regard for the others. By continuing this process of induction, we find that

$${}^n C_{k_1} \times {}^{n-k_1} C_{k_2} \times \cdots \times {}^{n-k_1-\cdots-k_{m-1}} C_{k_m} = \frac{n!}{k_1! (n-k_1)!} \times \frac{(n-k_1)!}{k_2! (n-k_1-k_2)!} \times \cdots$$

$$\times \frac{(n - k_1 - \dots - k_{m-2})!}{k_{m-1}! (n - k_1 - \dots - k_{m-1})!} \times \frac{(n - k_1 - \dots - k_{m-1})!}{k_m! 0!} \quad (1.2.16)$$

gives the number of different ways to arrange all objects in all groups. By cancellation of terms, the multinomial coefficient can be denoted as

$$\binom{n}{k_1, \dots, k_m} = \frac{n!}{k_1! k_2! \dots k_m!} \quad (1.2.17)$$

This number can also be thought of as the number of ways of depositing  $n$  distinct objects into  $m$  bins, where the  $j^{\text{th}}$  bin requires  $k_j$  objects. The number of permutations can also be written as special case of the multinomial coefficient. We have

$${}^n P_k = \binom{n}{n-k, 1, 1, \dots, 1} \quad (1.2.18)$$

$$= \frac{n!}{(n-k)!} \quad (1.2.19)$$

### Combinations in a Circle

With  $n$  total objects, consider the number of different ways to arrange  $k$  homogeneous objects among  $(n - k)$  other homogeneous objects in a circle (where it only matters which object is next to which). Like with arrangements and permutations in a circle, we can find this number by dividing the total number of combinations in a line by the number of possible rotations without affecting relative position, which is  $n$ . This gives

$$\frac{{}^n C_k}{n} = \frac{(n-1)!}{k! (n-k)!} \quad (1.2.20)$$

#### 1.2.4 Arrangements [41]

Using the Fundamental Theorem of Counting as well as the concepts of permutations/combinations, we can summarise the number of possible arrangements of size  $k$  from  $n$  objects. We qualify each arrangement as being:

- Ordered or unordered (i.e. whether order matters when counting the arrangements).
- With or without replacement (i.e. whether the same object is allowed to be used twice in the arrangement).

|           | Without replacement | With replacement   |
|-----------|---------------------|--------------------|
| Ordered   | $\frac{n!}{(n-k)!}$ | $n^k$              |
| Unordered | $\binom{n}{k}$      | $\binom{n+k-1}{k}$ |

The formula  $\frac{n!}{(n-k)!}$  is the number of  $k$  permutations of  $n$ , the formula  $n^k$  is by the Fundamental

Theorem of Counting, and  $\binom{n}{k}$  is the binomial coefficient. To obtain the formula  $\binom{n+k-1}{k}$  for an unordered arrangement with replacement, we use the following explanation. If  $k = 1$ , the formula is the same as without replacement. If  $k = 2$ , we need to replace the first chosen object, so by the time we choose the second object, this is effectively the same as choosing from  $n + 1$  objects without replacement. Generalising for arbitrary  $k$ , by the time we choose the  $k^{\text{th}}$  object, this is effectively the same as having chosen from  $n + k - 1$  objects without replacement, since we need to have replaced  $k - 1$  objects up to that point. And since we are considering unordered arrangements, it does not matter what order we choose the ‘artificial’ replacement objects, so it is valid to choose from  $n + k - 1$  objects from the beginning.

## 1.3 Probability Distributions

### 1.3.1 Random Variables

Random variables are variables which are determined as a result of a random experiment, i.e. they are not fixed and will depend on the outcome of the random experiment. This is in contrast to *deterministic* variables, which are always fixed regardless of the outcome. Deterministic variables can even be thought to be a special case of random variables, which are determined at the end of the random experiment but always take on the same value.

The value of a random variable after the experiment has occurred is said to have been *realised*.

#### Degenerate Random Variables

A degenerate random variable takes on a single value with probability one, and thus can be regarded as a deterministic variable.

#### Binary Random Variables

Binary random variables are random variables which can take on either of two possible values, typically chosen to be 0 and 1.

#### Indicator Random Variables

An indicator random variable for an event  $A$  is a binary random variable that takes on the value of 1 if  $A$  occurred, and 0 if  $A$  did not occur. An indicator for  $A$  can be denoted by  $\mathbb{I}_A$ .

#### Simple Random Variables

A simple random variable is a weighted sum of indicator random variables for mutually exclusive events. Suppose  $A_1, \dots, A_n$  are  $n$  mutually exclusive events. Then a simple random variable  $X$  is

$$X = \sum_{i=1}^n a_i \mathbb{I}_{A_i} \quad (1.3.1)$$

where  $a_1, \dots, a_n$  are the weights.

#### Discrete Random Variables

A discrete random variable takes on discrete values (e.g. within  $\{0, 1, 2, \dots\}$ ).

#### Continuous Random Variables

In some random experiments, a random variable can take on a continuum of values (e.g.  $[0, 1]$  or  $(-\infty, \infty)$ ). These are called continuous random variables.

#### Mixed Random Variables

In some random experiments, a random variable can take on a mix of discrete or continuous values (e.g.  $[0, 1]$  or  $(-\infty, \infty)$ ). These are called mixed random variables.

### 1.3.2 Distribution Functions

The possible values which a random variable can take, as well as probabilities of random variables taking on certain values, is represented using a distribution function.

## Probability Mass Functions

For a discrete random variable  $X$ , the probability mass function (PMF) describes the probability of  $X$  being equal to a particular value, denoted

$$p_X(x) = \Pr(X = x) \quad (1.3.2)$$

Note that for random experiments with discrete outcomes which do not explicitly result in the realisation of a random variable, a discrete random variable can still be defined by assigning each outcome to the set of integers (and if an integer  $x$  is not assigned to any outcome, this simply means  $\Pr(X = x) = 0$ ). In this way, a probability mass function can still be defined for any random experiment with discrete outcomes. A probability mass function satisfies the following properties:

$$\sum_{x=-\infty}^{\infty} \Pr(X = x) = 1 \quad (1.3.3)$$

and

$$0 \leq \Pr(X = x) \leq 1 \quad (1.3.4)$$

for all integers  $x$ .

## Probability Density Functions

A continuous random variable  $X$  can have a probability density function (PDF), denoted  $f_X(x)$ . Integrating this function over regions gives the probability that  $X$  will lie in that region. For example, the probability that  $a \leq X \leq b$  is given by

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x) dx \quad (1.3.5)$$

Note that it does not matter if this is integrated with open intervals or closed intervals. Hence for a continuous random variable,  $\Pr(a \leq X \leq b) = \Pr(a < X < b)$ . Also, the probability that  $X$  takes on any one particular value is zero, i.e.  $\Pr(X = x) = 0$ . That means it only makes sense to talk about  $X$  lying within regions, rather than for specific points. A probability density function satisfies the following properties:

$$\int_{-\infty}^{\infty} f_X(x) dx = 1 \quad (1.3.6)$$

and

$$f_X(x) \geq 0 \quad (1.3.7)$$

for all  $x \in (-\infty, \infty)$ . Note that unlike probability masses, probability densities are allowed to be greater than 1.

## Cumulative Distribution Functions

A cumulative distribution function (CDF)  $F_X(x)$  for a random variable  $X$  is defined as a function for the ‘cumulative’ probability up to  $x$ :

$$F_X(x) = \Pr(X \leq x) \quad (1.3.8)$$

The CDF is the ‘purest’ description of a random variable, in that any random variable, whether it is continuous, discrete, or mixed, has a CDF. Thus, the “distribution function” of a random variable is usually taken by default to mean the cumulative distribution function. If  $X$  is a discrete random variable, then the CDF will be discontinuous (i.e. it has ‘jumps’), and the

relation between the cumulative distribution function and the probability mass function is given by

$$F_X(a) = \sum_{x=-\infty}^a \Pr(X = x) \quad (1.3.9)$$

Whereas if  $X$  is a continuous random variable, then the CDF will be a continuous function, and the relation between the cumulative distribution function and the probability density function is given by

$$F_X(a) = \int_{-\infty}^a f_X(x) dx \quad (1.3.10)$$

The cumulative distribution satisfies the following properties:

$$0 \leq F_X(x) \leq 1 \quad (1.3.11)$$

for all  $x$ , and

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad (1.3.12)$$

$$\lim_{x \rightarrow \infty} F_X(x) = 1 \quad (1.3.13)$$

Furthermore, the probability density function can be obtained from the cumulative density function by

$$f_X(x) = \frac{dF_X(x)}{dx} \quad (1.3.14)$$

(wherever the cumulative density function is differentiable) and if given the probability density function, then

$$F_X(x) = \int f_X(x) dx \quad (1.3.15)$$

with a constant of integration such that  $\lim_{x \rightarrow \infty} F_X(x) = 1$ . The cumulative distribution function is always non-decreasing, since probabilities and densities are non-negative. If  $X$  is discrete, then  $F_X(x)$  will have ‘jumps’, but is by convention treated as right-continuous, to keep up with the definition  $F_X(x) = \Pr(X \leq x)$ .

### Support of a Distribution

The support of a distribution for a random variable is the set of values for which the probability density function (or probability mass function, in the case of discrete random variables) is non-zero. For a discrete random variable  $X$ , we can denote the support of  $X$  as

$$\mathcal{X} := \{x \in \mathbb{R} : \Pr(X = x) > 0\} \quad (1.3.16)$$

whereas if  $X$  were continuous with density  $f_X(x)$ , the support can be denoted as

$$\mathcal{X} := \{x \in \mathbb{R} : f_X(x) > 0\} \quad (1.3.17)$$

We may also say that  $X$  (or the distribution of  $X$ ) is supported on  $\mathcal{X}$ .

### Decomposition of Cumulative Distribution Functions

The cumulative distribution function for any random variable  $X$  can be decomposed into

$$F_X(x) = p_c F_c(x) + p_d F_d(x) \quad (1.3.18)$$

where  $p_c \geq 0$ ,  $p_d \geq 0$  and  $p_c + p_d = 1$ . Also,  $F_c(x)$  is the CDF of a continuous random variable, while  $F_d(x)$  is the CDF of a discrete random variable. Thus:

- If  $X$  is continuous, then  $p_c = 1$  and  $p_d = 0$ .
- If  $X$  is discrete, then  $p_c = 0$  and  $p_d = 1$ .
- If  $X$  is mixed, then we will have some combination of  $p_c > 0$  and  $p_d > 0$ .

## Elemental Probabilities

Consider the probability of a continuous random variable  $X$  belonging to a small strip from  $x$  to  $x + dx$ , where  $dx$  is the differential of  $x$ . We can write this probability in terms of the probability density function  $f_X(x)$  by

$$\Pr(x \leq X \leq x + dx) = f_X(x) dx \quad (1.3.19)$$

Alternatively, we can also write this using the differential of the cumulative distribution function,  $dF_X(x)$ , since  $dF_X(x) = f_X(x) dx$ .

## Quantile Functions

The quantile function  $Q(p) : (0, 1) \rightarrow \mathbb{R}$  can roughly be thought of as the inverse of the cumulative distribution function. However, the CDF  $F(x) = \Pr(X \leq x)$  may not always be invertible. There are different levels of definitions for the quantile function based on generality.

- If  $X$  is a continuous random variable supported on  $\mathbb{R}$ , then its CDF  $F(x)$  is strictly increasing everywhere, then its inverse  $F^{-1}(p)$  is well defined as the unique value  $x$  that makes  $F(x) = p$ , and we can take the quantile function as

$$Q(p) = F^{-1}(p) \quad (1.3.20)$$

- If  $X$  is a continuous random variable but not necessarily supported on  $\mathbb{R}$ , then  $F(x)$  is a non-decreasing continuous function, so for every  $p \in (0, 1)$  we can find a value  $x$  that makes  $F(x) = p$ . However, this value of  $x$  may not be unique (i.e. the CDF can have flat sections). So the appropriate way to define the quantile function is by taking the infimum:

$$Q(p) = \inf \{x \in \mathbb{R} : F(x) = p\} \quad (1.3.21)$$

which is loosely thought of as the smallest value  $x$  which makes  $F(x) = p$ .

- If  $X$  is allowed to be a discrete random variable, then its CDF  $F(x)$  could have jumps, so there does not necessarily exist a value  $x$  such that  $F(x) = p$  for a given  $p \in (0, 1)$ . In that case, we can define a *generalised inverse* for the quantile function, which is the smallest  $x$  such that  $F(x) \geq p$ . More formally, we can write this in terms of the infimum:

$$Q(p) = \inf \{x \in \mathbb{R} : F(x) \geq p\} \quad (1.3.22)$$

which is loosely thought of as the smallest value  $x$  which makes  $F(x) \geq p$ .

The quantile function can also be denoted  $F^{-1}(p)$ , and this is taken to mean the generalised inverse, whenever  $F(x)$  cannot be inverted conventionally. Choosing this definition for the quantile function means that the following conditions:

$$p \leq F(c) \quad (1.3.23)$$

$$F^{-1}(p) \leq c \quad (1.3.24)$$

are equivalent.

*Proof.* To show this in the forward direction starting with  $p \leq F(c)$ , let

$$x^* = \inf \{x : F(x) \geq p\} \quad (1.3.25)$$

$$= F^{-1}(p) \quad (1.3.26)$$

Then since the cumulative distribution is non-decreasing,  $F(x^*) = \inf\{F(x) : F(x) \geq p\}$  and from the condition  $p \leq F(c)$  this must mean  $F(x^*) \leq F(c)$ . Apply the non-decreasing property again to obtain  $x^* \leq c$ , and therefore by definition of  $x^* = F^{-1}(p)$ :

$$F^{-1}(p) \leq c \quad (1.3.27)$$

To show the reverse beginning with  $F^{-1}(p) \leq c$ , define yet again  $x^* = \inf\{x : F(x) \geq p\}$  so that  $F(x^*) \geq p$  and  $x^* = F^{-1}(p) \leq c$ . Using the non-decreasing property of  $F(\cdot)$ , this implies  $F(x^*) \leq F(c)$  hence:

$$p \leq F(c) \quad (1.3.28)$$

□

Since by convention the cumulative distribution function is right-continuous, then this definition also ensures the quantile function will be left-continuous.

### 1.3.3 Joint Distributions

Consider a random experiment which produces two random variables,  $X$  and  $Y$ . Then we say that  $X$  and  $Y$  has a joint distribution, which can be expressed in several ways.

#### Joint Probability Mass Functions

If the random variables  $X$  and  $Y$  are discrete (taking on integer values), we can express a joint probability mass function over all combinations of  $X$  and  $Y$ , given by

$$p_{XY}(x, y) = \Pr(X = x, Y = y) \quad (1.3.29)$$

This joint distribution has the property that

$$\sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} \Pr(X = x, Y = y) = 1 \quad (1.3.30)$$

We call  $\Pr(X = x)$  and  $\Pr(Y = y)$  the marginal distributions, which can be calculated from the joint distribution by a summation:

$$\Pr(X = x) = \sum_{y=-\infty}^{\infty} \Pr(X = x, Y = y) \quad (1.3.31)$$

$$\Pr(Y = y) = \sum_{x=-\infty}^{\infty} \Pr(X = x, Y = y) \quad (1.3.32)$$

#### Joint Probability Density Functions

If  $X$  and  $Y$  are continuous random variables, then it can have a joint density function denoted  $f_{XY}(x, y)$  with the relation

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1 \quad (1.3.33)$$

and the marginal densities  $f_X(x)$ ,  $f_Y(y)$  given by the integrals

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy \quad (1.3.34)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx \quad (1.3.35)$$

Note that while it is possible to obtain marginal distributions from the joint distribution (via the process known as ‘marginalisation’), it is generally not possible to obtain the joint distribution from the marginal distributions (without further information), as there may be more than one joint distribution which would yield those marginal distributions.

### Joint Cumulative Distribution Functions

For random variables  $X$  and  $Y$ , its joint cumulative distribution function is given by

$$F_{XY}(x, y) = \Pr(X \leq x, Y \leq y) \quad (1.3.36)$$

If  $X$  and  $Y$  are continuous with joint density  $f_{XY}(x, y)$ , then the relation between the joint CDF and joint density is

$$F_{XY}(x, y) = \int_{-\infty}^y \int_{-\infty}^x f_{XY}(x', y') dx' dy' \quad (1.3.37)$$

Conversely, this operation can be reversed by taking the second-order derivative:

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \quad (1.3.38)$$

If  $X$  and  $Y$  are integer-valued discrete random variables, the joint CDF can be obtained with a double sum over the joint probability mass function:

$$F_{XY}(x, y) = \sum_{j=-\infty}^y \sum_{i=-\infty}^x \Pr(X = i, Y = i) \quad (1.3.39)$$

Alternatively, deriving the joint probability mass function from the joint PDF can be done using the addition rule. First define the events

$$A = \{X \leq x, Y < y\} \quad (1.3.40)$$

$$B = \{X < x, Y \leq y\} \quad (1.3.41)$$

Then logically

$$A \cup B = \{X \leq x, Y \leq y\} \setminus \{X = x, Y = y\} \quad (1.3.42)$$

$$A \cap B = \{X < x, Y < y\} \quad (1.3.43)$$

Thus the joint probability mass function can be expressed as

$$p_{XY}(x, y) = \Pr(X = x, Y = y) \quad (1.3.44)$$

$$= \Pr(X \leq x, Y \leq y) - \Pr(A \cup B) \quad (1.3.45)$$

$$= \Pr(X \leq x, Y \leq y) - \Pr(A) - \Pr(B) + \Pr(A \cap B) \quad (1.3.46)$$

$$\begin{aligned} &= \Pr(X \leq x, Y \leq y) - \Pr(X \leq x, Y < y) - \Pr(X < x, Y \leq y) \\ &\quad + \Pr(X < x, Y < y) \end{aligned} \quad (1.3.47)$$

$$\begin{aligned} &= \Pr(X \leq x, Y \leq y) - \Pr(X \leq x, Y \leq y-1) - \Pr(X \leq x-1, Y \leq y) \\ &\quad + \Pr(X \leq x-1, Y \leq y-1) \end{aligned} \quad (1.3.48)$$

$$= F_{XY}(x, y) - F_{XY}(x, y-1) - F_{XY}(x-1, y) + F_{XY}(x-1, y-1) \quad (1.3.49)$$

### 1.3.4 Conditional Distributions

#### Conditional Probability Mass Functions

By the definition of the conditional probability, we know (for discrete  $X$  and  $Y$ ):

$$\Pr(X = x|Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} \quad (1.3.50)$$

In terms of the joint mass function and marginal mass function of  $Y$ , we write

$$p_{XY}(x|y) = \Pr(X = x|Y = y) \quad (1.3.51)$$

$$= \frac{p_{XY}(x,y)}{p_Y(y)} \quad (1.3.52)$$

which is known as the conditional probability mass function of  $X$  given  $Y = y$ , and changes depending on the value of  $y$ . Notice that it is the quotient between the joint distribution of  $X, Y$  and the marginal distribution of  $Y$ , and is defined only when  $\Pr(Y = y) > 0$ , otherwise we can presume  $\Pr(X = x|Y = y) = 0$ . Likewise, the conditional distribution of  $Y$  given  $X$  is

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} \quad (1.3.53)$$

Using the law of total probability, we can relate the marginal distribution of  $X$  to the conditional distribution of  $X$  given  $Y$  by

$$\Pr(X = x) = \sum_{y=-\infty}^{\infty} \Pr(X = x, Y = y) \quad (1.3.54)$$

$$= \sum_{\{y: \Pr(Y=y)>0\}} \Pr(X = x|Y = y) \Pr(Y = y) \quad (1.3.55)$$

We can also show that the conditional distribution  $\Pr(X = x|Y = y)$  is a proper distribution (summing over  $x$  to 1) for any given  $y$ :

$$\sum_{x=-\infty}^{\infty} \Pr(X = x|Y = y) = \sum_{x=-\infty}^{\infty} \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} \quad (1.3.56)$$

$$= \frac{1}{\Pr(Y = y)} \sum_{x=-\infty}^{\infty} \Pr(X = x, Y = y) \quad (1.3.57)$$

$$= \frac{\Pr(Y = y)}{\Pr(Y = y)} \quad (1.3.58)$$

$$= 1 \quad (1.3.59)$$

### Conditional Cumulative Distribution Functions

For integer-valued discrete random variables  $X$  and  $Y$ , the conditional cumulative distribution function of  $X$  given  $Y = y$  is given by the analogous relation between the CDF and PMF:

$$F_{X|Y}(x|y) = \Pr(X \leq x|Y = y) \quad (1.3.60)$$

$$= \sum_{i=-\infty}^x \Pr(X = i|Y = y) \quad (1.3.61)$$

$$= \sum_{i=-\infty}^x p_{XY}(i|y) \quad (1.3.62)$$

If  $X$  and  $Y$  are continuous, then defining a conditional CDF  $F_{X|Y}(x|y)$  is slightly more problematic because  $\Pr(Y = y) = 0$ , so  $\Pr(X \leq x|Y = y)$  is not well defined. However, we can instead take the limit of the conditional probability given that  $Y$  is in a small strip next to  $y$  [102].

$$F_{X|Y}(x|y) = \lim_{\varepsilon \rightarrow 0} \frac{\Pr(X \leq x, y \leq Y < y + \varepsilon)}{\Pr(y \leq Y < y + \varepsilon)} \quad (1.3.63)$$

$$= \lim_{\varepsilon \rightarrow 0} \frac{\Pr(X \leq x, Y < y + \varepsilon) - \Pr(X \leq x, Y < y)}{\Pr(Y < y + \varepsilon) - \Pr(Y < y)} \quad (1.3.64)$$

$$= \lim_{\varepsilon \rightarrow 0} \frac{\Pr(X \leq x, Y < y + \varepsilon) - \Pr(X \leq x, Y < y)}{\varepsilon} \quad (1.3.65)$$

$$\div \lim_{\varepsilon \rightarrow 0} \frac{\Pr(Y < y + \varepsilon) - \Pr(Y < y)}{\varepsilon}$$

Recognising the definition of the derivative, then

$$F_{X|Y}(x|y) = \frac{\partial}{\partial y} \Pr(X \leq x, Y < y) \div \frac{d}{dy} \Pr(Y < y) \quad (1.3.66)$$

$$= \frac{\frac{\partial}{\partial y} F_{X,Y}(x,y)}{f_Y(y)} \quad (1.3.67)$$

This can be shown to be a valid CDF, by taking the limit of  $x \rightarrow \infty$  and  $x \rightarrow -\infty$ . First note that

$$\lim_{x \rightarrow \infty} \frac{\partial}{\partial y} F_{X,Y}(x,y) = \lim_{x \rightarrow \infty} \frac{\partial}{\partial y} \Pr(X \leq x, Y \leq y) \quad (1.3.68)$$

$$= \frac{\partial}{\partial y} \Pr(X \leq \infty, Y \leq y) \quad (1.3.69)$$

$$= \frac{\partial}{\partial y} \Pr(Y \leq y) \quad (1.3.70)$$

$$= f_Y(y) \quad (1.3.71)$$

Thus

$$\lim_{x \rightarrow \infty} F_{X|Y}(x|y) = \frac{f_Y(y)}{f_Y(y)} \quad (1.3.72)$$

$$= 1 \quad (1.3.73)$$

Similarly, with  $\lim_{x \rightarrow -\infty} \frac{\partial}{\partial y} F_{X,Y}(x,y) = 0$ , we have

$$\lim_{x \rightarrow -\infty} F_{X|Y}(x|y) = 0 \quad (1.3.74)$$

### Conditional Probability Density Functions

If  $X$  and  $Y$  are continuous, then it can have conditional densities  $f_{X|Y}(x|y)$  and  $f_{Y|X}(y|x)$ . These can be obtained by differentiating the conditional CDF, which becomes in the case for  $f_{X|Y}(x|y)$ :

$$f_{X|Y}(x|y) = \frac{\partial}{\partial x} F_{X|Y}(x|y) \quad (1.3.75)$$

$$= \frac{\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)}{f_Y(y)} \quad (1.3.76)$$

$$= \frac{f_{XY}(x,y)}{f_Y(y)} \quad (1.3.77)$$

This is defined wherever  $f_Y(y) > 0$ , and can be interpreted as the conditional density of  $X$  given that  $Y$  is in the infinitesimal interval  $(y, y + dx)$ . Likewise, the conditional density of  $Y$  given  $X$  is

$$f_Y(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \quad (1.3.78)$$

which is defined wherever  $f_X(x) > 0$ . Rearranging these expressions, we have

$$f_{XY}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad (1.3.79)$$

$$f_{XY}(x, y) = f_{Y|X}(y|x) f_X(x) \quad (1.3.80)$$

These forms are analogous to the chain rule of probability. We can show similar marginalisation properties (as with discrete distributions) hold:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy \quad (1.3.81)$$

$$= \int_{\{y: f_Y(y) > 0\}} f_{X|Y}(x|y) f_Y(y) dy \quad (1.3.82)$$

and the density integrates to one:

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = \int_{-\infty}^{\infty} \frac{f_{XY}(x, y)}{f_Y(y)} dx \quad (1.3.83)$$

$$= \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{XY}(x, y) dx \quad (1.3.84)$$

$$= \frac{f_Y(y)}{f_Y(y)} \quad (1.3.85)$$

$$= 1 \quad (1.3.86)$$

### Conditional Random Variables

A ‘conditional random variable’ denoted  $[X|Y = y]$  can be informally defined as the random variable which has its own distribution function being the conditional distribution of  $X$  given  $Y = y$ .

## 1.4 Expectation

The expectation (or expected value, or mean) of a random variable can be thought about in several ways:

- The average realisation of the random variable over an infinite number of trials.
- The ‘center of mass’ of the probability distribution. If there were a rigid thin rod with density (or mass) along the rod equal to the probability density (or mass), the expectation would be the point at which the rod could be perfectly balanced.
- In the case of continuous distributions, the expectation is the net area between the function  $x f_X(x)$  and the horizontal axis.

The expected value of a random variable  $X$  is a constant denoted  $\mathbb{E}[X]$  and for a discrete random variable may be calculated as

$$\mathbb{E}[X] = \sum_{x=-\infty}^{\infty} x \Pr(X = x) \quad (1.4.1)$$

If  $X$  is a continuous random variable, then the expectation is calculated by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \quad (1.4.2)$$

Expectations of functions of random variables (which are also random variables) are readily calculable, for example some with  $g(X)$ , the so-called *Law of the Unconscious Statistician* is given by:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad (1.4.3)$$

If an experiment results in two (continuous) random variables  $X$  and  $Y$ , we can also take the expectation of some function of the two random variables  $g(X, Y)$  by

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy \quad (1.4.4)$$

and analogously for discrete random variables. Sometimes, particularly when there are multiple random variables present, an expectation may be taken with respect to a random variable, which expresses which marginal distribution to sum/integrate over. For example, with a continuous random variable  $X$  and a function  $g(X)$ , the expectation of  $g(X)$  with respect to  $X$  is denoted with  $\mathbb{E}_X[g(X)]$ , and defined like above as

$$\mathbb{E}_X[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad (1.4.5)$$

This is more for emphasis and clarity rather than any technical difference, since  $g(X)$  may be written using another symbol which does not make the connection with  $X$  explicit.

### 1.4.1 Linearity of Expectation

Consider two random variables  $X, Y$  and two constants  $a, b$ . The expectation of  $aX + bY$  (supposing  $X$  and  $Y$  are discrete) is written as

$$\mathbb{E}[aX + bY] = \sum_x \sum_y (ax + by) \Pr(X = x, Y = y) \quad (1.4.6)$$

$$= a \sum_x \sum_y x \Pr(X = x, Y = y) + b \sum_x \sum_y y \Pr(X = x, Y = y) \quad (1.4.7)$$

$$= a \sum_x x \sum_y \Pr(X = x, Y = y) + b \sum_y y \sum_x \Pr(X = x, Y = y) \quad (1.4.8)$$

$$= a \sum_x x \Pr(X = x) + b \sum_y y \Pr(Y = y) \quad (1.4.9)$$

$$= a\mathbb{E}[X] + b\mathbb{E}[Y] \quad (1.4.10)$$

This analogously holds for continuous  $X$  and  $Y$ . Since sums and integrals are linear operators, and the expectation is defined either in terms of sums or integrals, this gives the property known as the linearity of expectation.

### 1.4.2 Conditional Expectation

The conditional expectation of  $X$  given  $Y$  is denoted  $\mathbb{E}[X|Y]$ , and may be interpreted in two ways:

- Given a particular value of  $Y = y$ , the conditional expectation  $\mathbb{E}[X|Y = y]$  is the average of  $X$  over all events where  $Y = y$ . It may be calculated by taking the expectation of the conditional distribution of  $X$  given  $Y$ , i.e. for discrete random variables

$$\mathbb{E}[X|Y = y] = \sum_x x \Pr(X = x|Y = y) \quad (1.4.11)$$

For continuous random variables,

$$\mathbb{E}[X|Y=y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|Y=y) dx \quad (1.4.12)$$

This form of the conditional expectation may be treated as a deterministic function in the realisation  $y$ .

- The conditional expectation  $\mathbb{E}[X|Y]$  without being given a particular value of  $Y$  is a random variable, because  $Y$  is random. For discrete  $X$  and  $Y$ , the random variable  $\mathbb{E}[X|Y]$  is defined to take on value  $\mathbb{E}[X|Y=y]$  with probability  $\Pr(Y=y)$ . For continuous  $X$  and  $Y$ , the random variable  $\mathbb{E}[X|Y]$  has density  $f_{X|Y}(x|Y=y)$  at value  $\mathbb{E}[X|Y=y]$ .

The conditional expectation of  $X$  conditional on its own realisation  $X=x$  becomes  $x$ , i.e.  $\mathbb{E}[X|X=x]=x$ , because given  $X=x$ , we expect that  $X$  naturally can only be  $x$ . We can show this more formally (for the case of a discrete random variable  $X$ ) by using the facts that  $\Pr(X=x|X=x)=1$  and  $\Pr(X=x'|X=x)=0$  when  $x' \neq x$ :

$$\mathbb{E}[X|X=x] = \sum_{x'} x' \Pr(X=x'|X=x) \quad (1.4.13)$$

$$= x \quad (1.4.14)$$

It can then be reasoned that  $\mathbb{E}[X|X]=X$ , because by the definition above  $X$  is the random variable which takes on value  $\mathbb{E}[X|X=x]=x$  with probability  $\Pr(X=x)$ . Generalising this, we can say that for a function  $f(\cdot)$ , we have

$$\mathbb{E}[f(X)|X=x] = f(x) \quad (1.4.15)$$

and

$$\mathbb{E}[f(X)|X] = f(X) \quad (1.4.16)$$

### 1.4.3 Law of Iterated Expectations

The law of iterated expectations (also sometimes referred to the law of total expectation) states that the expectation of the random variable  $\mathbb{E}[X|Y]$  (over  $Y$ ) is  $\mathbb{E}[X]$ . That is,

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \quad (1.4.17)$$

Sometimes the law is written  $\mathbb{E}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}[X]$  to explicitly show which variables the expectations are taken with respect to. To prove the law for discrete random variables,

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_y \mathbb{E}[X|Y=y] \Pr(Y=y) \quad (1.4.18)$$

$$= \sum_y \sum_x x \Pr(X=x|Y=y) \Pr(Y=y) \quad (1.4.19)$$

$$= \sum_y \sum_x x \Pr(X=x, Y=y) \quad (1.4.20)$$

$$= \mathbb{E}[X] \quad (1.4.21)$$

This is similarly shown for continuous random variables. The law of iterated expectations can also be extended to when there are more than two random variables. Some variations are:

$$\mathbb{E}[\mathbb{E}[X|Y, Z]] = \mathbb{E}[X] \quad (1.4.22)$$

$$\mathbb{E}[\mathbb{E}[X|Y, Z]|Z] = \mathbb{E}[X|Z] \quad (1.4.23)$$

Note that it is necessary to condition the inner expectation  $\mathbb{E}[X|Y, Z]$  in the second equation to also be conditional on  $Z$ . This is because if we instead wrote  $\mathbb{E}[\mathbb{E}[X|Y]|Z]$ , the outer expectation means  $Z$  is given, while the inner expectation means  $Z$  is not given, so this will not be very clearly defined.

#### 1.4.4 Expectation of Indicator Random Variables

Let  $\mathbb{I}_A$  be an indicator random variable for event  $A$ . The expectation of  $\mathbb{I}_A$  is the probability of  $A$ , as shown below.

$$\mathbb{E}[\mathbb{I}_A] = 0 \times \Pr(\bar{A}) + 1 \times \Pr(A) \quad (1.4.24)$$

$$= \Pr(A) \quad (1.4.25)$$

#### 1.4.5 Expectations Using Cumulative Distribution Function

Let  $X$  be a non-negative random variable with cumulative distribution function  $F_X(x)$ . Then the expectation of  $X$  can be alternatively characterised using  $F_X(x)$  by noting

$$X = \int_0^X 1 \cdot dx \quad (1.4.26)$$

$$= \int_0^\infty \mathbb{I}_{\{X>x\}} dx \quad (1.4.27)$$

where  $\mathbb{I}_{\{X>x\}}$  is the indicator random variable for the event  $\{X > x\}$ . Hence

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbb{I}_{\{X>x\}} dx\right] \quad (1.4.28)$$

$$= \int_0^\infty \mathbb{E}[\mathbb{I}_{\{X>x\}}] dx \quad (1.4.29)$$

$$= \int_0^\infty \Pr(X > x) dx \quad (1.4.30)$$

$$= \int_0^\infty (1 - \Pr(X \leq x)) dx \quad (1.4.31)$$

$$= \int_0^\infty (1 - F_X(x)) dx \quad (1.4.32)$$

This formula can be generalised further when  $X$  takes on negative values. Suppose now  $X$  is real-valued (not just non-negative). By combining both cases when  $X$  may be positive or negative, we can write

$$X = \int_0^\infty \mathbb{I}_{\{x < X\}} dx - \int_{-\infty}^0 \mathbb{I}_{\{x \geq X\}} dx \quad (1.4.33)$$

Taking expectations as before,

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbb{I}_{\{x < X\}} dx - \int_{-\infty}^0 \mathbb{I}_{\{x \geq X\}} dx\right] \quad (1.4.34)$$

$$= \int_0^\infty \mathbb{E}[\mathbb{I}_{\{x < X\}}] dx - \int_{-\infty}^0 \mathbb{E}[\mathbb{I}_{\{x \geq X\}}] dx \quad (1.4.35)$$

$$= \int_0^\infty \Pr(X > x) dx - \int_{-\infty}^0 \Pr(X \leq x) dx \quad (1.4.36)$$

$$= \int_0^\infty (1 - F_X(x)) dx - \int_{-\infty}^0 F_X(x) dx \quad (1.4.37)$$

A special case occurs when  $X$  is a discrete random variable, on support  $\{0, 1, 2, \dots\}$ . From above where we had  $\mathbb{E}[X] = \int_0^\infty \Pr(X > x) dx$ , we can replace the integral exactly using a sum of area of rectangles with width one, since  $\Pr(X > x)$  is constant over the interval  $[n, n+1]$ . Therefore

$$\mathbb{E}[X] = \int_0^\infty \Pr(X > x) dx \quad (1.4.38)$$

$$= \sum_{n=0}^{\infty} \Pr(X > n) \quad (1.4.39)$$

Another proof of this fact follows by observing

$$\sum_{n=0}^{\infty} \Pr(X > n) = \sum_{n=0}^{\infty} \sum_{i=n+1}^{\infty} \Pr(X = i) \quad (1.4.40)$$

$$= (\Pr(X = 1) + \Pr(X = 2) + \Pr(X = 3) + \dots) \quad (1.4.41)$$

$$+ (\Pr(X = 2) + \Pr(X = 3) + \Pr(X = 4) + \dots) + \dots$$

$$= \Pr(X = 1) + \Pr(X = 2) + \Pr(X = 2) \quad (1.4.42)$$

$$+ \Pr(X = 3) + \Pr(X = 3) + \Pr(X = 3)$$

$$+ \Pr(X = 4) + \Pr(X = 4) + \Pr(X = 4) + \Pr(X = 4) + \dots$$

$$= \sum_{i=0}^{\infty} i \Pr(X = i) \quad (1.4.43)$$

$$= \mathbb{E}[X] \quad (1.4.44)$$

#### 1.4.6 Expectations Using Quantile Function

The expectation of a continuous random variable  $X$  with probability density function  $f(x)$ , cumulative distribution function  $F(x)$  and quantile function  $Q(u)$  by applying the change of variables  $u = F(x)$  in the integral

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) dx \quad (1.4.45)$$

By noting that  $du = f(x)dx$ , and by definition of the quantile function  $x = F^{-1}(u) = Q(u)$ , and by the properties  $\lim_{x \rightarrow \infty} F(x) = 1$  and  $\lim_{x \rightarrow 0} F(x) = 0$ , this change of variables yields

$$\mathbb{E}[X] = \int_0^1 Q(u) du \quad (1.4.46)$$

## 1.5 Variance

The variance of a random variable  $X$  measures the spread of the distribution of  $X$  about its mean. Realisations of a random variable with high variance can be thought to deviate far from its mean on average, while realisations of a random variable with low variance are expected to be concentrated close to the mean. The variance of  $X$  is defined as

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad (1.5.1)$$

We can interpret  $(X - \mathbb{E}[X])^2$  as the squared deviation of  $X$  from its mean, and so  $\text{Var}(X)$  is the expected squared deviation of  $X$  from its mean. Another way to interpret the variance of a random variable is that it gives an idea of the ‘uncertainty’. That is, the more uncertain a random variable is, the more spread out its distribution becomes and so the harder it is to predict. An alternative formula for the variance is

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (1.5.2)$$

This can be shown as follows. Starting by expanding the original definition given for the variance,

$$\text{Var}(X) = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \quad (1.5.3)$$

Then using the linearity of expectation:

$$\text{Var}(X) = \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]^2] \quad (1.5.4)$$

Note that  $\mathbb{E}[X\mathbb{E}[X]] = \mathbb{E}[X]^2$  and  $\mathbb{E}[\mathbb{E}[X]^2] = \mathbb{E}[X]^2$  as  $\mathbb{E}[X]$  is treated as a ‘constant’ with respect to the distribution of  $X$ , and so can be taken outside of the expectation. Hence

$$\text{Var}(X) = \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \quad (1.5.5)$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (1.5.6)$$

The variance is always non-negative, because  $(X - \mathbb{E}[X])^2$  is always non-negative. If  $X$  is scaled by a constant  $a$ , then

$$\text{Var}(aX) = \mathbb{E}[(aX - \mathbb{E}[aX])^2] \quad (1.5.7)$$

$$= a^2 \mathbb{E}[(X - \mathbb{E}[X])^2] \quad (1.5.8)$$

$$= a^2 \text{Var}(X) \quad (1.5.9)$$

Variance is also not affected by a constant translation in the random variable, because

$$\text{Var}(X + b) = \mathbb{E}[(X + b - \mathbb{E}[X + b])^2] \quad (1.5.10)$$

$$= \mathbb{E}[(X - \mathbb{E}[X] + b - b)^2] \quad (1.5.11)$$

$$= \text{Var}(X) \quad (1.5.12)$$

### 1.5.1 Standard Deviation

The units of the variance  $\text{Var}(X)$  is in the squared units of  $X$ . A quantity which measures a random variable’s dispersion about its mean but in the same units as the random variable is the standard deviation. The standard deviation is defined as the square root of the variance, i.e.

$$\text{sd}(X) = \sqrt{\text{Var}(X)} \quad (1.5.13)$$

Like the variance, the standard deviation is also always non-negative. The standard deviation is useful because since it is in the same units as the random variable, it is more easily interpretable.

### 1.5.2 Precision

The precision of a random variable  $X$  is defined as the inverse of the variance, i.e.  $\text{Var}(X)^{-1} = \frac{1}{\text{Var}(X)}$ . Naturally, the more precise a random variable is, the more concentrated it is about its expected value.

### 1.5.3 Covariance

The covariance of a pair of random variables  $X$  and  $Y$  gives a measure of how closely the two random variables are related (in terms of their realisations). The covariance is defined by

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \quad (1.5.14)$$

This expectation is taken over the joint distribution of  $X$  and  $Y$ . If a realisation of  $X$  being ‘high’ (i.e. above the mean) means that a realisation of  $Y$  is likely to also be high, and vice-versa ( $Y$  being high means  $X$  is likely to be high), and also that  $X$  being low means  $Y$  is likely to be

low and vice-versa, then the sign of  $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$  is more likely to be positive and hence the covariance will be positive. We may say that the random variables ‘move together’. In the opposing case, i.e. the sign of  $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$  is more likely to be negative, then the covariance will be negative and we can interpret the random variables as ‘moving against’ each other. An alternative expression for the covariance is

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.5.15)$$

This can be shown by expanding the original definition given:

$$\text{Cov}(X, Y) = \mathbb{E}[XY - \mathbb{E}[X]Y - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]] \quad (1.5.16)$$

$$= \mathbb{E}[XY] - \mathbb{E}[\mathbb{E}[X]Y] - \mathbb{E}[X\mathbb{E}[Y]] + \mathbb{E}[\mathbb{E}[X]\mathbb{E}[Y]] \quad (1.5.17)$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \quad (1.5.18)$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.5.19)$$

where we have used the linearity of expectation and the facts that  $\mathbb{E}[X]$  and  $\mathbb{E}[Y]$  are treated as constant with respect to the joint distribution over  $X$  and  $Y$ , so hence may be taken outside of the respective expectations. Also note that if we take the covariance between  $X$  and itself:

$$\text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad (1.5.20)$$

$$= \text{Var}(X) \quad (1.5.21)$$

which becomes the variance of  $X$ . The covariance between a random variable and a constant is zero, because

$$\text{Cov}(X, b) = \mathbb{E}[Xb] - \mathbb{E}[X]b \quad (1.5.22)$$

$$= 0 \quad (1.5.23)$$

Also, if we were to scale random variables  $X$  and  $Y$  by constants  $a$  and  $b$  respectively, we would have

$$\text{Cov}(aX, bY) = \mathbb{E}[aXbY] - \mathbb{E}[aX]\mathbb{E}[bY] \quad (1.5.24)$$

$$= ab(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \quad (1.5.25)$$

$$= ab\text{Cov}(X, Y) \quad (1.5.26)$$

## Variance of Sums

For the sum of random variables  $Z = X + Y$ , the variance of  $Z$  is given by

$$\text{Var}(Z) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y) \quad (1.5.27)$$

This is shown using the properties of the expectation, and definitions of the variance and covariance.

$$\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \quad (1.5.28)$$

$$= \mathbb{E}[(X + Y)^2] - \mathbb{E}[X + Y]^2 \quad (1.5.29)$$

$$= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \quad (1.5.30)$$

$$= (\mathbb{E}[X^2] - \mathbb{E}[X]^2) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) + (\mathbb{E}[Y^2] - \mathbb{E}[Y]^2) \quad (1.5.31)$$

$$= \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y) \quad (1.5.32)$$

## Variance of Linear Combinations

If we now consider the variance of the linear combination  $Z = aX + bY$  where  $a, b$  are constants, then

$$\text{Var}(Z) = \text{Var}(aX + bY) \quad (1.5.33)$$

$$= \text{Var}(aX) + 2\text{Cov}(aX, bY) + \text{Var}(bY) \quad (1.5.34)$$

$$= a^2 \text{Var}(X) + 2ab \text{Cov}(X, Y) + b^2 \text{Var}(Y) \quad (1.5.35)$$

## Covariance of Linear Combinations

With random variables  $X, Y, W, Z$  and constants  $a, b, c, d$ , we can show that the covariance between linear combinations of random variables is given by

$$\text{Cov}(aX + bY, cW + dZ) = ac \text{Cov}(X, W) + ad \text{Cov}(X, Z) + bc \text{Cov}(Y, W) + bd \text{Cov}(Y, Z) \quad (1.5.36)$$

To show this, first use the definition of covariance

$$\text{Cov}(aX + bY, cW + dZ) = \mathbb{E}[(aX + bY)(cW + dZ)] - \mathbb{E}[aX + bY]\mathbb{E}[cW + dZ] \quad (1.5.37)$$

Then expanding gives

$$\begin{aligned} \text{Cov}(aX + bY, cW + dZ) &= ab\mathbb{E}[XW] + ad\mathbb{E}[XZ] + bc\mathbb{E}[YW] + bd\mathbb{E}[YZ] \\ &\quad - ac\mathbb{E}[X]\mathbb{E}[W] - ad\mathbb{E}[X]\mathbb{E}[Z] - bc\mathbb{E}[Y]\mathbb{E}[W] - bd\mathbb{E}[Y]\mathbb{E}[Z] \end{aligned} \quad (1.5.38)$$

Grouping terms, we finally see that

$$\text{Cov}(aX + bY, cW + dZ) = ac \text{Cov}(X, W) + ad \text{Cov}(X, Z) + bc \text{Cov}(Y, W) + bd \text{Cov}(Y, Z) \quad (1.5.39)$$

With this, by treating  $bY = b$  and  $dZ = d$  as constants, we can also show that the covariance is unaffected by constant translations in its arguments:

$$\text{Cov}(X + b, W + d) = \text{Cov}(X, W) + d \text{Cov}(X, 1) + b \text{Cov}(1, W) + bd \text{Cov}(1, 1) \quad (1.5.40)$$

$$= \text{Cov}(X, W) \quad (1.5.41)$$

### 1.5.4 Correlation

The correlation coefficient is a ‘standardised’ covariance. Given random variables  $X, Y$  and their standard deviations  $\text{sd}(X)$  and  $\text{sd}(Y)$ , the correlation coefficient between  $X$  and  $Y$  is defined as

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)} \quad (1.5.42)$$

As the correlation coefficient is standardised, it always lies between  $-1$  and  $1$ . We can show this follows for two random variables  $X$  and  $Y$ . For ease of notation, denote  $\mu_X := \mathbb{E}[X]$ ,  $\mu_Y := \mathbb{E}[Y]$ ,  $\sigma_X := \text{sd}(X)$  and  $\sigma_Y := \text{sd}(Y)$ . Define the ‘standardised’ random variables

$$\tilde{X} := \frac{X - \mu_X}{\sigma_X} \quad (1.5.43)$$

$$\tilde{Y} := \frac{Y - \mu_Y}{\sigma_Y} \quad (1.5.44)$$

Then start from the inequality

$$\mathbb{E}\left[\left(\tilde{X} + \tilde{Y}\right)^2\right] \geq 0 \quad (1.5.45)$$

Expanding out gives

$$\mathbb{E} [\tilde{X}^2 + 2\tilde{X}\tilde{Y} + \tilde{Y}^2] \geq 0 \quad (1.5.46)$$

By the linearity of expectation,

$$\mathbb{E} [\tilde{X}^2] + 2\mathbb{E} [\tilde{X}\tilde{Y}] + \mathbb{E} [\tilde{Y}^2] \geq 0 \quad (1.5.47)$$

Note that by standardisation,  $\mathbb{E} [\tilde{X}^2] = \text{Var} (\tilde{X}) = 1$  and  $\mathbb{E} [\tilde{Y}^2] = \text{Var} (\tilde{Y}) = 1$  so

$$2 + 2\mathbb{E} [\tilde{X}\tilde{Y}] \geq 0 \quad (1.5.48)$$

Using the definitions of the standardised random variables, we can show:

$$2 + 2\frac{\mathbb{E} [XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y]}{\sigma_X\sigma_Y} \geq 0 \quad (1.5.49)$$

$$2 + 2\frac{\mathbb{E} [XY] - \mathbb{E} [X]\mu_Y - \mathbb{E} [Y]\mu_X + \mu_X\mu_Y}{\sigma_X\sigma_Y} \geq 0 \quad (1.5.50)$$

$$2 + 2\frac{\mathbb{E} [XY] - \mu_X\mu_Y}{\sigma_X\sigma_Y} \geq 0 \quad (1.5.51)$$

Then using the definition of the covariance:

$$2 + 2\frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} \geq 0 \quad (1.5.52)$$

Hence using the definition of the correlation coefficient

$$\text{Corr}(X, Y) \geq -1 \quad (1.5.53)$$

Starting from the inequality

$$\mathbb{E} [\tilde{X}^2] \geq 0 \quad (1.5.54)$$

and applying the same steps, we get

$$\text{Corr}(X, Y) \leq 1 \quad (1.5.55)$$

Therefore combined,

$$-1 \leq \text{Corr}(X, Y) \leq 1 \quad (1.5.56)$$

The correlation coefficient allows for pairs of random variables with different units/scales to be compared.

### Correlation of Linear Transformations

If the random variables inside the correlation are linearly transformed, this does not affect the correlation, which can be shown as follows:

$$\text{Corr}(aX + b, cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{\sqrt{\text{Var}(aX + b)\text{Var}(cY + d)}} \quad (1.5.57)$$

$$= \frac{ac \text{Cov}(X, Y)}{ac\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad (1.5.58)$$

$$= \text{Corr}(X, Y) \quad (1.5.59)$$

### Subadditivity of Standard Deviations

Suppose  $X$  and  $Y$  are random variables, and we form  $Z = X + Y$ . Then we have

$$\text{sd}(Z) \leq \text{sd}(X) + \text{sd}(Y) \quad (1.5.60)$$

This is known as the subadditivity property of standard deviations. We can show subadditivity as follows.

$$\sqrt{\text{Var}(Z)} \leq \sqrt{\text{Var}(X)} + \sqrt{\text{Var}(Y)} \quad (1.5.61)$$

$$\sqrt{\text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y)} \leq \sqrt{\text{Var}(X)} + \sqrt{\text{Var}(Y)} \quad (1.5.62)$$

$$\text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y) \leq (\sqrt{\text{Var}(X)} + \sqrt{\text{Var}(Y)})^2 \quad (1.5.63)$$

$$\text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y) \leq \text{Var}(X) + 2\sqrt{\text{Var}(X)\text{Var}(Y)} + \text{Var}(Y) \quad (1.5.64)$$

$$\text{Cov}(X, Y) \leq \sqrt{\text{Var}(X)\text{Var}(Y)} \quad (1.5.65)$$

We know the last inequality holds, because correlations are upper bounded by 1, i.e.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \leq 1 \quad (1.5.66)$$

#### 1.5.5 Conditional Variance

The conditional variance of  $X$  given  $Y$ , denoted  $\text{Var}(X|Y)$  may informally be thought of as the variance of the conditional random variable  $X|Y$ . Formally, the definition of the conditional variance is analogous to the definition of the variance except the expectations are conditioned on  $Y$ :

$$\text{Var}(X|Y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2 | Y] \quad (1.5.67)$$

We can show that an analogous alternative formula holds for the conditional variance:

$$\text{Var}(X|Y) = \mathbb{E}[X^2|Y] - \mathbb{E}[X|Y]^2 \quad (1.5.68)$$

This is done by first expanding and using the linearity of expectation:

$$\text{Var}(X|Y) = \mathbb{E}[X^2 - 2\mathbb{E}[X|Y]X + \mathbb{E}[X|Y]^2 | Y] \quad (1.5.69)$$

$$= \mathbb{E}[X^2|Y] - 2\mathbb{E}[\mathbb{E}[X|Y]X|Y] + \mathbb{E}[\mathbb{E}[X|Y]^2|Y] \quad (1.5.70)$$

Now since  $\mathbb{E}[X|Y]$  is a random variable on the same sample space as  $Y$ , we can take the term out of the expectation conditioned on  $Y$ , giving

$$\text{Var}(X|Y) = \mathbb{E}[X^2|Y] - 2\mathbb{E}[X|Y]^2 + \mathbb{E}[X|Y]^2 \quad (1.5.71)$$

$$= \mathbb{E}[X^2|Y] - \mathbb{E}[X|Y]^2 \quad (1.5.72)$$

As with conditional expectation, the conditional variance with supplied realisation of  $y$ , denoted  $\text{Var}(X|Y = y)$ , may be thought of as a deterministic function in  $y$ .

The conditional variance of  $X$  given  $X$  can be shown to be

$$\text{Var}(X|X) = \mathbb{E}[X^2|X] - \mathbb{E}[X|X]^2 \quad (1.5.73)$$

$$= X^2 - X^2 \quad (1.5.74)$$

$$= 0 \quad (1.5.75)$$

Intuitively, if given a value of  $X$ , there should be no more uncertainty surrounding  $X$ , hence why the conditional variance is zero.

### 1.5.6 Law of Total Variance

The Law of Total Variance says that for random variables  $X$  and  $Y$

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X]) \quad (1.5.76)$$

To show this, first start with the variance of  $Y$  in terms of expectations involving  $Y$ .

$$\text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \quad (1.5.77)$$

By applying the Law of Iterated Expectations:

$$\text{Var}(Y) = \mathbb{E}_X[\mathbb{E}[Y^2|X]] - \mathbb{E}_X[\mathbb{E}[Y|X]]^2 \quad (1.5.78)$$

Now use the definition of conditional variance  $\text{Var}(Y|X) = \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2$  to get

$$\text{Var}(Y) = \mathbb{E}_X[\text{Var}(Y|X) + \mathbb{E}[Y|X]^2] - \mathbb{E}_X[\mathbb{E}[Y|X]]^2 \quad (1.5.79)$$

Using the linearity of expectations:

$$\text{Var}(Y) = \mathbb{E}_X[\text{Var}(Y|X)] + \mathbb{E}_X[\mathbb{E}[Y|X]^2] - \mathbb{E}_X[\mathbb{E}[Y|X]]^2 \quad (1.5.80)$$

Then finally recognising  $\text{Var}(\mathbb{E}[Y|X]) = \mathbb{E}_X[\mathbb{E}[Y|X]^2] - \mathbb{E}_X[\mathbb{E}[Y|X]]^2$  yields

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X]) \quad (1.5.81)$$

### 1.5.7 Conditional Covariance

The conditional covariance between random variables  $X$  and  $Y$  on  $Z$  is denoted  $\text{Cov}(X, Y|Z)$ . It is taken (informally) as the covariance between the conditional random variables  $X|Z$  and  $Y|Z$ . It is defined as

$$\text{Cov}(X, Y|Z) = \mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])|Z] \quad (1.5.82)$$

Expanding this out and using the Law of Iterated Expectations, we can write

$$\text{Cov}(X, Y|Z) = \mathbb{E}[XY - X\mathbb{E}[Y|Z] - Y\mathbb{E}[X|Z] + \mathbb{E}[X|Z]\mathbb{E}[Y|Z]|Z] \quad (1.5.83)$$

$$= \mathbb{E}[XY|Z] - \mathbb{E}[X\mathbb{E}[Y|Z]|Z] - \mathbb{E}[Y\mathbb{E}[X|Z]|Z] + \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]|Z] \quad (1.5.84)$$

The same way  $\mathbb{E}[XY|X] = X\mathbb{E}[Y|X]$ , we can take out terms  $\mathbb{E}[Y|Z]$  and  $\mathbb{E}[X|Z]$  from their outer expectations.

$$\text{Cov}(X, Y|Z) = \mathbb{E}[XY|Z] - \mathbb{E}[X|Z]\mathbb{E}[Y|Z] - \mathbb{E}[Y|Z]\mathbb{E}[X|Z] + \mathbb{E}[1|Z]\mathbb{E}[X|Z]\mathbb{E}[Y|Z] \quad (1.5.85)$$

$$= \mathbb{E}[XY|Z] - \mathbb{E}[X|Z]\mathbb{E}[Y|Z] \quad (1.5.86)$$

As with conditional expectations, the term  $\text{Cov}(X, Y|Z)$  may be treated as a random variable on the same sample space as  $Z$ , whereas  $\text{Cov}(X, Y|Z = z)$  is a function of the realisation  $z$ .

### 1.5.8 Law of Total Covariance

The Law of Total Covariance says that

$$\text{Cov}(X, Y) = \mathbb{E}[\text{Cov}(X, Y|Z)] + \text{Cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]) \quad (1.5.87)$$

To derive this, first begin with the definition of covariance.

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.5.88)$$

Using the Law of Iterated Expectations, rewrite the expectations as

$$\text{Cov}(X, Y) = \mathbb{E}[\mathbb{E}[XY|Z]] - \mathbb{E}[\mathbb{E}[X|Z]]\mathbb{E}[\mathbb{E}[Y|Z]] \quad (1.5.89)$$

Using the definition of conditional covariance, an alternative expression for the inner term of the first expectation gives

$$\text{Cov}(X, Y) = \mathbb{E}[\text{Cov}(X, Y|Z) + \mathbb{E}[X|Z]\mathbb{E}[Y|Z]] - \mathbb{E}[\mathbb{E}[X|Z]]\mathbb{E}[\mathbb{E}[Y|Z]] \quad (1.5.90)$$

Then using the linearity of expectation

$$\text{Cov}(X, Y) = \mathbb{E}[\text{Cov}(X, Y|Z)] + \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]] - \mathbb{E}[\mathbb{E}[X|Z]]\mathbb{E}[\mathbb{E}[Y|Z]] \quad (1.5.91)$$

Finally, recognise that the last two terms give the covariance between the random variables  $\mathbb{E}[X|Z]$  and  $\mathbb{E}[Y|Z]$ .

$$\text{Cov}(X, Y) = \mathbb{E}[\text{Cov}(X, Y|Z)] + \text{Cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]) \quad (1.5.92)$$

The Law of Total Variance is a special case of the Law of Total Covariance where  $X = Y$ .

## 1.6 Independence

### 1.6.1 Independent Events

Intuitively, two events  $A$  and  $B$  are independent if any information about the occurrence of  $B$  does not affect the probability of occurrence of  $A$ . That is,

$$\Pr(A|B) = \Pr(A) \quad (1.6.1)$$

The same will then be true about  $B$  given  $A$ , because by using Bayes' rule

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)} \quad (1.6.2)$$

$$= \frac{\Pr(A)\Pr(B)}{\Pr(A)} \quad (1.6.3)$$

$$= \Pr(B) \quad (1.6.4)$$

Then by using the definition of conditional probabilities,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (1.6.5)$$

$$\Pr(A) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (1.6.6)$$

$$\Pr(A \cap B) = \Pr(A)\Pr(B) \quad (1.6.7)$$

This says the joint probabilities of  $A$  and  $B$  can be found by multiplying their marginal probabilities.

### Independence of Complementary Events

If events  $A$  and  $B$  are independent, then  $A$  and  $\bar{B}$  are also independent. We can show this as follows:

$$\Pr(A) = \Pr(A \cap (B \cup \bar{B})) \quad (1.6.8)$$

$$= \Pr((A \cap B) \cup (A \cap \bar{B})) \quad (1.6.9)$$

$$= \Pr(A \cap B) + \Pr(A \cap \bar{B}) - \Pr((A \cap B) \cap (A \cap \bar{B})) \quad (1.6.10)$$

$$= \Pr(A \cap B) + \Pr(A \cap \bar{B}) \quad (1.6.11)$$

since  $B$  and  $\bar{B}$  are mutually exclusive. Then from independence of  $A$  and  $B$ ,

$$\Pr(A) = \Pr(A) \Pr(B) + \Pr(A \cap \bar{B}) \quad (1.6.12)$$

Then rearranging, we obtain

$$\Pr(A \cap \bar{B}) = \Pr(A)(1 - \Pr(B)) \quad (1.6.13)$$

$$= \Pr(A)\Pr(\bar{B}) \quad (1.6.14)$$

which reveals that  $A$  and  $\bar{B}$  are independent. It follows simply that the events  $\bar{A}$  and  $B$  are independent, while also the events  $\bar{A}$  and  $\bar{B}$  are independent.

### Pairwise Independent Events

For a collection of events  $A_1, \dots, A_n$ , we say that they are pairwise independent if every pair of events is independent.

### Mutually Independent Events

For a collection of events  $A_1, \dots, A_n$ , we say that they are mutually independent if each event is independent with the intersection of any other events. This implies that each event is also independent with the union of any other events, because we can obtain a union from an intersection using DeMorgan's laws, and independence still holds if we take complements. Note that mutual independence also implies pairwise independence, but the converse is not necessarily true. For mutually independent events,

$$\Pr(A_1 \cap \dots \cap A_n) = \Pr(A_1) \times \dots \times \Pr(A_n) \quad (1.6.15)$$

However, it is not enough for the above condition to hold for mutual independence. We would also require that  $\Pr(A_1 \cap \dots \cap A_n)$  can be formed by multiplication of probabilities of any intersections of events, for instance:

$$\Pr(A_1 \cap \dots \cap A_n) = \Pr(A_1 \cap A_2) \times \dots \times \Pr(A_3 \cap \dots \cap A_n) \quad (1.6.16)$$

Mutually independent events will be pairwise independent, but the converse is not necessarily true.

### 1.6.2 Independent Random Variables

Extending the definition of independent events, two (discrete) random variables  $X$  and  $Y$  are independent if the events  $X = x$  and  $Y = y$  are independent for all possible combinations of realisations  $x$  and  $y$ . Hence the joint probability is the product of the marginal probabilities:

$$\Pr(X = x \cap Y = y) = \Pr(X = x)\Pr(Y = y) \quad (1.6.17)$$

for all  $x$  and  $y$ . That is, if we can find at least one pair of  $x$  and  $y$  such that

$$\Pr(X = x \cap Y = y) \neq \Pr(X = x)\Pr(Y = y) \quad (1.6.18)$$

then  $X$  and  $Y$  are dependent. For the case of continuous random variables  $X$  and  $Y$ , the definition of independence naturally extends to being that the joint density is the product of the marginal probabilities, i.e.

$$f_{XY}(x, y) = f_X(x)f_Y(y) \quad (1.6.19)$$

for all  $x$  and  $y$ . Equivalently (this also applies to discrete random variables), we may say that  $X$  and  $Y$  are independent if their cumulative distribution functions satisfy

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x)\Pr(Y \leq y) \quad (1.6.20)$$

for all  $x$  and  $y$ . Alternatively, in terms of marginal distributions, if  $X$  and  $Y$  are independent, then for all  $x$  and  $y$

$$f_{X|Y}(x|y) = f_X(x) \quad (1.6.21)$$

$$f_{Y|X}(y|x) = f_Y(y) \quad (1.6.22)$$

because  $f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$  and  $f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$ .

### Pairwise Independent Random Variables

If every pair in a collection of  $n$  random variables  $X_1, X_2, \dots, X_n$  is independent, then they are said to be pairwise independent.

### Mutually Independent Random Variables

For a collection of  $n$  random variables  $X_1, X_2, \dots, X_n$ , if for any  $x_1, \dots, x_n$  the collection of events  $\{X_1 \leq x_1\}, \dots, \{X_n \leq x_n\}$  are mutually independent, then the collection of random variables are said to be mutually independent random variables. This means that their cumulative distributions satisfy

$$\Pr(X_1 \leq x_1, \dots, X_n \leq x_n) = \Pr(X_1 \leq x_1) \times \Pr(X_n \leq x_n) \quad (1.6.23)$$

for any  $x_1, \dots, x_n$ . It is common to drop the ‘mutual’ qualifier, so whenever random variables  $X_1, X_2, \dots, X_n$  are said to be independent, this is normally understood (unless otherwise specified) to mean that they are mutually independent.

As with events, mutually independent random variables are also pairwise independent, but the reverse is not necessarily true. Thus mutual independence is stronger than pairwise independence, as the following counterexample demonstrates. Let  $X$  and  $Y$  be indicator random variables for independent fair coin flips. Let  $Z$  be another indicator for whether exactly one of those coin flips is heads. The joint distribution between  $X, Y, Z$  is then

$$\Pr(X = x, Y = y, Z = z) = \begin{cases} 1/4, & (x, y, z) = (0, 0, 0) \\ 1/4, & (x, y, z) = (0, 1, 1) \\ 1/4, & (x, y, z) = (1, 0, 1) \\ 1/4, & (x, y, z) = (1, 1, 0) \\ 0, & \text{otherwise} \end{cases} \quad (1.6.24)$$

Denote the marginal distributions by  $p_X(\cdot)$ ,  $p_Y(\cdot)$  and  $p_Z(\cdot)$ . Observe that  $p_X(0) = p_X(1) = \frac{1}{2}$  and likewise with  $p_Y(\cdot)$  and  $p_Z(\cdot)$ . Denoting  $p_{XY}(\cdot, \cdot)$ ,  $p_{XZ}(\cdot, \cdot)$ ,  $p_{YZ}(\cdot, \cdot)$  as the bivariate distributions, observe that  $p_{XY}(x, y) = \frac{1}{4}$  everywhere over the support and likewise with  $p_{XZ}(\cdot, \cdot)$  and  $p_{YZ}(\cdot, \cdot)$ . Hence

- $X$  and  $Y$  are independent.
- $X$  and  $Z$  are independent.
- $Y$  and  $Z$  are independent.

Thus  $X$ ,  $Y$  and  $Z$  are pairwise independent. However they are not mutually independent because  $Z$  depends on both  $X$  and  $Y$ . We can also check that

$$p_X(0)p_Y(0)p_Z(0) = \frac{1}{8} \quad (1.6.25)$$

$$\neq \Pr(X = 0, Y = 0, Z = 0) \quad (1.6.26)$$

The intuition is that although  $Z$  depends on both  $X$  and  $Y$ , when considering the pair  $(X, Z)$ , they behave like independent random variables because  $Y$  is still ‘free’ to vary.

### Independent and Identically Distributed Random Variables

If a collection of  $n$  random variables  $X_1, X_2, \dots, X_n$  all have the same distribution and are mutually independent (i.e. the joint distribution of  $X_1, X_2, \dots, X_n$  is the product of the marginal distributions of  $X_1, X_2, \dots, X_n$ ), then the collection of random variables is said to be independent and identically distributed (i.i.d.).

#### 1.6.3 Uncorrelatedness

If two random variables  $X$  and  $Y$  have a covariance (hence also correlation) of zero, then they are said to be uncorrelated. We can show that if  $X$  and  $Y$  are independent, then they will always be uncorrelated (provided their covariance exists). From the definition of covariance:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.27)$$

$$= \sum_x \sum_y xy \Pr(X = x \cap Y = y) - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.28)$$

Using independence, then

$$\text{Cov}(X, Y) = \sum_x \sum_y xy \Pr(X = x) \Pr(Y = y) - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.29)$$

$$= \sum_x x \Pr(X = x) \sum_y y \Pr(Y = y) - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.30)$$

$$= \left( \sum_x x \Pr(X = x) \right) \left( \sum_y y \Pr(Y = y) \right) - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.31)$$

$$= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.32)$$

$$= 0 \quad (1.6.33)$$

However, the reverse is not true. If  $X$  and  $Y$  are uncorrelated, then this does not necessarily mean that they are independent. A simple counterexample to show this is a continuous random variable  $X$  with distribution

$$f_X(x) = \begin{cases} \frac{1}{2}, & x \in [-1, 1] \\ 0, & \text{elsewhere} \end{cases} \quad (1.6.34)$$

and  $Y = X^2$ . Then  $X$  and  $Y$  are clearly dependent. However it can be seen that  $\mathbb{E}[X] = 0$  and  $\mathbb{E}[XY] = 0$ . The latter is reasoned by considering

$$\mathbb{E}[XY] = \mathbb{E}[X^3] \quad (1.6.35)$$

$$= \int_{-1}^1 \frac{x^3}{2} dx \quad (1.6.36)$$

$$= 0 \quad (1.6.37)$$

since  $\frac{x^3}{2}$  is an odd function. Therefore

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.38)$$

$$= 0 \quad (1.6.39)$$

Subadditivity of the standard deviation is also more straightforward to show in the case where  $X$  and  $Y$  are uncorrelated. We firstly derive what is known as the subadditivity property of the square root function. For any two scalars  $a, b \geq 0$ :

$$(\sqrt{a} + \sqrt{b})^2 = a + 2\sqrt{ab} + b \quad (1.6.40)$$

$$\sqrt{a} + \sqrt{b} = \sqrt{a + 2\sqrt{ab} + b} \quad (1.6.41)$$

$$\geq \sqrt{a+b} \quad (1.6.42)$$

since  $\sqrt{ab} \geq 0$ . Then

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \quad (1.6.43)$$

In fact, we can tell this only holds with equality if and only if  $a = 0$  or  $b = 0$ . The variance of  $Z$  is given by  $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y)$  so

$$\sqrt{\text{Var}(Z)} \leq \sqrt{\text{Var}(X)} + \sqrt{\text{Var}(Y)} \quad (1.6.44)$$

which yields the subadditivity property for the standard deviation.

#### 1.6.4 Mean Independence

For random variables  $X$  and  $Y$ , if  $\mathbb{E}[X|Y] = \mathbb{E}[X]$  (that is, the conditional expectation is equal to the unconditional expectation), then  $X$  is said to be mean independent of  $Y$ . Likewise if  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ , then  $X$  is said to be mean independent of  $Y$ . This condition is something in between independence and uncorrelatedness; it is weaker than independence but stronger than uncorrelatedness.

If  $X$  and  $Y$  are independent, then  $\mathbb{E}[X|Y] = \mathbb{E}[X]$  and  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ .

*Proof.* To show this, consider that if  $X$  and  $Y$  are continuous random variables and independent, then for all  $y$ :

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} f_{X|Y}(x|y) dx \quad (1.6.45)$$

$$= \int_{-\infty}^{\infty} f_X(x) dx \quad (1.6.46)$$

$$= \mathbb{E}[X] \quad (1.6.47)$$

This holds analogously for  $\mathbb{E}[Y|X = x]$  and if  $X$  and  $Y$  were discrete.  $\square$

However, the converse is not generally true. If  $\mathbb{E}[X|Y] = \mathbb{E}[X]$  or  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ , this does not necessarily imply  $X$  and  $Y$  are independent.

*Proof.* To show with a counterexample, let  $X$  and  $Y$  be two independent zero-mean random variables. Define  $Z = XY$ . Then

$$\mathbb{E}[Z|X] = \mathbb{E}[XY|X] \quad (1.6.48)$$

$$= X\mathbb{E}[Y|X] \quad (1.6.49)$$

$$= X\mathbb{E}[Y] \quad (1.6.50)$$

$$= 0 \quad (1.6.51)$$

Also because of independence,

$$\mathbb{E}[Z] = \mathbb{E}[XY] \quad (1.6.52)$$

$$= \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.53)$$

$$= 0 \quad (1.6.54)$$

Therefore  $\mathbb{E}[Z|X] = \mathbb{E}[Z]$ . However,  $Z$  and  $X$  are not independent because  $Z$  is formed by  $Z = XY$ .  $\square$

Also note that  $\mathbb{E}[X|Y] = \mathbb{E}[X]$  does not necessarily imply  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ .

*Proof.* A counterexample of this is as follows. Let  $Y$  be a random variable defined by

$$\Pr(Y = y) = \begin{cases} 1/3, & y = -1, 0, 1 \\ 0, & \text{elsewhere} \end{cases} \quad (1.6.55)$$

so that  $\mathbb{E}[Y] = 0$ . Let  $X$  be an indicator random variable for  $Y = 0$ . Hence

$$\Pr(X = x) = \begin{cases} 2/3, & x = 0 \\ 1/3, & x = 1 \\ 0, & \text{elsewhere} \end{cases} \quad (1.6.56)$$

From this characterisation, we see that

$$\mathbb{E}[Y|X = 0] = 0 \quad (1.6.57)$$

$$\mathbb{E}[Y|X = 1] = 0 \quad (1.6.58)$$

Hence  $\mathbb{E}[Y] = \mathbb{E}[Y|X]$ . However,  $\mathbb{E}[X] \neq \mathbb{E}[X|Y]$  because

$$\mathbb{E}[X|Y = 0] = 1 \quad (1.6.59)$$

$$\mathbb{E}[X|Y = 1] = 0 \quad (1.6.60)$$

$$\mathbb{E}[X|Y = -1] = 0 \quad (1.6.61)$$

$\square$

Now if  $\mathbb{E}[X|Y] = \mathbb{E}[X]$  or  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ , then this implies  $X$  and  $Y$  are uncorrelated.

*Proof.* Suppose  $\mathbb{E}[X|Y] = \mathbb{E}[X]$ . Then using the Law of Iterated Expectations,

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|Y]] \quad (1.6.62)$$

$$= \mathbb{E}[Y\mathbb{E}[X|Y]] \quad (1.6.63)$$

$$= \mathbb{E}[Y\mathbb{E}[X]] \quad (1.6.64)$$

$$= \mathbb{E}[X]\mathbb{E}[Y] \quad (1.6.65)$$

which is one of the definitions for uncorrelatedness. The same can be analogously shown if  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ .  $\square$

However, the converse need not necessarily be true. If  $X$  and  $Y$  are uncorrelated, this does not necessarily imply that  $\mathbb{E}[X|Y] = \mathbb{E}[X]$  or  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ .

*Proof.* We demonstrate again with the counterexample where

$$\Pr(Y = y) = \begin{cases} 1/3, & y = -1, 0, 1 \\ 0, & \text{elsewhere} \end{cases} \quad (1.6.66)$$

and  $X$  is an indicator random variable for  $Y = 0$  with

$$\Pr(X = x) = \begin{cases} 2/3, & x = 0 \\ 1/3, & x = 1 \\ 0, & \text{elsewhere} \end{cases} \quad (1.6.67)$$

By this construction,  $XY = 0$  always so  $\mathbb{E}[XY] = 0$ . Also since  $\mathbb{E}[Y] = 0$ , then  $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = 0$  so  $X$  and  $Y$  are uncorrelated. However as already established,  $\mathbb{E}[X] \neq \mathbb{E}[X|Y]$ . And since the naming of  $X$  and  $Y$  is arbitrary, we can construct an analogous counterexample where  $\mathbb{E}[Y] \neq \mathbb{E}[Y|X]$ .  $\square$

### 1.6.5 Conditional Independence

#### Conditionally Independent Events

Two events  $A$  and  $B$  are conditionally independent on a third event  $C$  if the knowledge of  $C$  makes events  $A$  and  $B$  independent (hence their joint probability conditional on  $C$  can be obtained by multiplying the marginal conditional probabilities):

$$\Pr(A \cap B|C) = \Pr(A|C)\Pr(B|C) \quad (1.6.68)$$

By rearranging this and applying the definition of conditional probability, an alternative expression can be obtained.

$$\frac{\Pr(A \cap B|C)}{\Pr(B|C)} = \Pr(A|C) \quad (1.6.69)$$

$$\Pr(A|B \cap C) = \Pr(A|C) \quad (1.6.70)$$

Note that  $A$  and  $B$  may or may not be independent without  $C$ . Also, if  $A$  and  $B$  are not conditionally independent on  $C$ , then they are conditionally dependent on  $C$ .

#### Conditionally Independent Random Variables

Similarly, two random variables  $X$  and  $Y$  are conditionally independent on a third random variable  $Z$  if the conditional random variables  $X|Z$  and  $Y|Z$  are independent. Or more formally (for the case of continuous random variables),

$$f_{XY|Z}(x, y) = f_{X|Z}(x)f_{Y|Z}(y) \quad (1.6.71)$$

### 1.6.6 Orthogonality [219]

Two random variables  $X$  and  $Y$  are said to be orthogonal if  $\mathbb{E}[XY] = 0$ . If  $X$  and  $Y$  are uncorrelated and at least one of them is zero-mean, then they are also orthogonal.

### 1.6.7 Exchangeability

A sequence of random variables  $X_1, \dots, X_n$  is said to be exchangeable (or interchangeable) if for any permutation  $i_1, \dots, i_n$  of  $1, \dots, n$ , the distribution of  $X_{i_1}, \dots, X_{i_n}$  is the same as that of  $X_1, \dots, X_n$ . It is related to independence because a sequence of i.i.d. random variables automatically qualifies as exchangeable. An example of a sequence of dependent but exchangeable random variables is a (uniform) sample from a finite population without replacement.

### 1.6.8 Variance Using Independent Copies

The variance of a random variable  $X$  can alternatively be represented as

$$\text{Var}(X) = \frac{1}{2}\mathbb{E}[(X - Y)^2] \quad (1.6.72)$$

where  $Y$  is an independent copy of  $X$ . This can be shown as follows. Using the **variance of linear combinations**,

$$\text{Var}(X - Y) = \text{Var}(X) - 2\text{Cov}(X, Y) + \text{Var}(Y) \quad (1.6.73)$$

$$= 2\text{Var}(X) \quad (1.6.74)$$

since  $\text{Var}(X) = \text{Var}(Y)$  and  $\text{Cov}(X, Y) = 0$  by independence. Also due to the definition of independence:

$$\text{Var}(X - Y) = \mathbb{E}[(X - Y - \mathbb{E}[X - Y])^2] \quad (1.6.75)$$

$$= \mathbb{E}[(X - Y)^2] \quad (1.6.76)$$

because  $\mathbb{E}[X - Y] = \mathbb{E}[X] - \mathbb{E}[Y] = 0$ . Thus combining the expressions for  $\text{Var}(X - Y)$  above, it can be seen that

$$\text{Var}(X) = \frac{1}{2}\mathbb{E}[(X - Y)^2] \quad (1.6.77)$$

### Covariance Using Independent Copies

Similar to the variance, an alternative characterisation for the covariance is

$$\text{Cov}(X, Y) = \frac{1}{2}\mathbb{E}[(X - X')(Y - Y')] \quad (1.6.78)$$

where  $(X', Y')$  is an independent copy of  $(X, Y)$ . We can use essentially the same technique to show this. Using the **covariance of linear combinations**,

$$\text{Cov}(X - X', Y - Y') = \text{Cov}(X, Y) - \text{Cov}(X, Y') - \text{Cov}(X', Y) + \text{Cov}(X', Y') \quad (1.6.79)$$

$$= 2\text{Cov}(X, Y) \quad (1.6.80)$$

because  $\text{Cov}(X, Y) = \text{Cov}(X', Y')$  due to identicality, and  $\text{Cov}(X, Y') = 0$ ,  $\text{Cov}(X', Y) = 0$  due to independence. Also,

$$\text{Cov}(X - X', Y - Y') = \mathbb{E}[(X - X' - (\mathbb{E}[X] - \mathbb{E}[X']))(Y - Y' - (\mathbb{E}[Y] - \mathbb{E}[Y']))] \quad (1.6.81)$$

$$= \mathbb{E}[(X - X')(Y - Y')] \quad (1.6.82)$$

because  $\mathbb{E}[X] = \mathbb{E}[X']$  and  $\mathbb{E}[Y] = \mathbb{E}[Y']$ . Equating these two expressions for  $\text{Cov}(X - X', Y - Y')$ , we have

$$\text{Cov}(X, Y) = \frac{1}{2}\mathbb{E}[(X - X')(Y - Y')] \quad (1.6.83)$$

## 1.7 Transformations of Random Variables

### 1.7.1 Linear Transformations of Random Variables

Suppose random variable  $X$  has probability density function  $f_X(x)$  and cumulative distribution function  $F_X(x)$ . Define the linearly transformed random variable  $Y = aX + b$ , with  $a > 0$ . We derive expressions for the density function  $f_Y(y)$ , cumulative distribution  $F_Y(y)$  and quantile function  $F_Y^{-1}(p)$ . Firstly by definition of the cumulative distribution,

$$F_Y(y) = \Pr(Y \leq y) \quad (1.7.1)$$

$$= \Pr(aX + b \leq y) \quad (1.7.2)$$

$$= \Pr\left(X \leq \frac{y-b}{a}\right) \quad (1.7.3)$$

$$= F_X\left(\frac{y-b}{a}\right) \quad (1.7.4)$$

Differentiating using the chain rule, the probability density function becomes

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right) \quad (1.7.5)$$

The relationship between the quantile functions takes the same form as the linear transformation:

$$F_Y^{-1}(p) = aF_X^{-1}(p) + b \quad (1.7.6)$$

This can be shown using the definition of the quantile function:

$$F_Y^{-1}(p) = \inf\{y : F_Y(y) \geq p\} \quad (1.7.7)$$

$$= \inf\{y : \Pr(Y \leq y) \geq p\} \quad (1.7.8)$$

$$= \inf\{y : \Pr(aX + b \leq y) \geq p\} \quad (1.7.9)$$

$$= \inf\left\{y : \Pr\left(X \leq \frac{y-b}{a}\right) \geq p\right\} \quad (1.7.10)$$

$$= a \inf\left\{\frac{y-b}{a} : \Pr\left(X \leq \frac{y-b}{a}\right) \geq p\right\} + b \quad (1.7.11)$$

$$= aF_X^{-1}(p) + b \quad (1.7.12)$$

### 1.7.2 Parametric Distributions

A probability distribution may belong to a ‘family’ of distributions, characterised by a common functional form that is parametrised. A probability distribution parametrised in  $\theta$  may be denoted  $f(x; \theta)$ .

#### Scale Parameters

A scale parameter ‘stretches’ the distribution. If a particular distribution with scale parameter  $\theta$  has density  $f(x)$ , then the distribution with scale parameter  $a\theta$  has density  $f(x/a)/a$ .

#### Rate Parameters

The reciprocal of a scale parameter may be referred to as a rate parameter (i.e. it ‘concentrates’ the distribution). Hence if a particular distribution with rate parameter  $\theta$  has density  $f(x)$ , then the distribution with rate parameter  $b\theta$  has density  $bf(bx)$ .

## Location Parameters

A location parameter ‘shifts’ the distribution. This means that if a particular distribution with location parameter  $\theta$  has density  $f(x)$ , then the distribution with location parameter  $\theta + c$  has density  $f(x - c)$ . Families of distributions may typically be parametrised in terms of the mean, median or mode as a location parameter.

## Shape Parameters

Any parameter of a distribution which is neither a scale/rate nor location parameter may be referred to as a shape parameter.

### 1.7.3 Sums of Random Variables

#### Sums of Continuous Random Variables

Consider the sum of two continuous random variables  $W = X + Y$ . The PDF of  $W$  can be defined as

$$f_W(w) = \int \int_{\{x,y:x+y=w\}} f_{XY}(x, y) dx dy \quad (1.7.13)$$

We can then just integrate along the straight line  $x + y = w$  and make a substitution of either  $x = w - y$  or  $y = w - x$  (which influences the integrating variable). This gives rise to two alternative formulas.

$$f_W(w) = \int_{-\infty}^{\infty} f_{XY}(x, w - x) dx \quad (1.7.14)$$

$$f_W(w) = \int_{-\infty}^{\infty} f_{XY}(w - y, y) dy \quad (1.7.15)$$

If  $X$  and  $Y$  are independent, then

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w - x) dx \quad (1.7.16)$$

$$f_W(w) = \int_{-\infty}^{\infty} f_X(w - y) f_Y(y) dy \quad (1.7.17)$$

In this case,  $f_W(\cdot)$  resembles a convolution between  $f_X(\cdot)$  and  $f_Y(\cdot)$ . Still treating  $X$  and  $Y$  as independent, the CDF  $F_W(w)$  can be derived as:

$$F_W(w) = \Pr(W \leq w) \quad (1.7.18)$$

$$= \int_{-\infty}^{\infty} \Pr(X \leq w - y) f_Y(y) dy \quad (1.7.19)$$

since  $X \leq w - Y$  implies  $X + Y = W \leq w$ . Then

$$F_W(w) = \int_{-\infty}^{\infty} F_X(w - y) f_Y(y) dy \quad (1.7.20)$$

This can be verified by differentiating the CDF:

$$f_W(w) = \frac{\partial}{\partial w} \int_{-\infty}^{\infty} F_X(w - y) f_Y(y) dy \quad (1.7.21)$$

$$= \int_{-\infty}^{\infty} \frac{\partial}{\partial w} F_X(w - y) f_Y(y) dy \quad (1.7.22)$$

$$= \int_{-\infty}^{\infty} f_X(w - y) f_Y(y) dy \quad (1.7.23)$$

### Sums of Discrete Random Variables

For the case of the sum of two discrete random variables  $W = X + Y$  taking on integer values, the analogous formulae for the probability mass distribution are

$$\Pr(W = w) = \sum_{x=-\infty}^{\infty} \Pr(X = x, Y = w - x) \quad (1.7.24)$$

$$\Pr(W = w) = \sum_{y=-\infty}^{\infty} \Pr(X = w - y, Y = y) \quad (1.7.25)$$

and in the independent case

$$\Pr(W = w) = \sum_{x=-\infty}^{\infty} \Pr(X = x) \Pr(Y = w - x) \quad (1.7.26)$$

$$\Pr(W = w) = \sum_{y=-\infty}^{\infty} \Pr(X = w - y) \Pr(Y = y) \quad (1.7.27)$$

and the CDF in the independent case is

$$\Pr(W \leq w) = \sum_{y=-\infty}^{\infty} \Pr(X \leq w - y) \Pr(Y = y) \quad (1.7.28)$$

### Difference of Continuous Random Variables

Consider the difference of two continuous random variables  $W = X - Y$ . Analogously (by making substitutions  $x = w + y$  or  $y = -w + x$ ) we can show two different versions for the density of the difference:

$$f_W(w) = \int_{-\infty}^{\infty} f_{XY}(x, -w + x) dx \quad (1.7.29)$$

$$f_W(w) = \int_{-\infty}^{\infty} f_{XY}(w + y, y) dy \quad (1.7.30)$$

and in the independent case

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(-w + x) dx \quad (1.7.31)$$

$$f_W(w) = \int_{-\infty}^{\infty} f_X(w + y) f_Y(y) dy \quad (1.7.32)$$

For the CDF in the independent case, we can use

$$F_W(w) = \int_{-\infty}^{\infty} F_X(w + y) f_Y(y) dy \quad (1.7.33)$$

**Theorem 1.1.** *If  $X$  and  $Y$  are independent and identically distributed continuous random variables, than the distribution of their difference  $W = X - Y$  is symmetric.*

*Proof.* Note that  $W$  will necessarily have a mean of zero (assuming  $X$  and  $Y$  have finite mean). If the mean of  $W$  is not defined, its probability distribution will still have a ‘symmetry point’ of zero. It suffices to show  $f_W(w) = f_W(-w)$ . Using the formulas above in the independent case,

$$f_W(-w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w + x) dx \quad (1.7.34)$$

Since  $X$  and  $Y$  are identical,  $f_X(x) = f_Y(x)$  so

$$f_W(-w) = \int_{-\infty}^{\infty} f_Y(x) f_X(w+x) dx \quad (1.7.35)$$

'Renaming' the integrating variable to  $y$  yields

$$f_W(-w) = \int_{-\infty}^{\infty} f_Y(y) f_X(w+y) dy = f_W(w) \quad (1.7.36)$$

□

## Stable Distributions

A family of distributions is said to be stable if a linear combination of two independent random variables from that family also has a distribution in that family. Examples of stable continuous distributions are the Gaussian and Cauchy distributions. Examples of stable discrete distributions are the Poisson and Binomial distributions.

### 1.7.4 Strictly Monotonic Transformations of Random Variables

Let  $X$  be a random variable with probability density  $f_X(x)$ , and let  $\mathcal{X}$  be the support of  $X$ , which is the set

$$\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\} \quad (1.7.37)$$

Suppose that  $Y$  is another random variable which is a transformation of  $X$ , given by  $Y = g(X)$  such that  $g$  is strictly *monotonic* (i.e. either strictly increasing or decreasing) over  $\mathcal{X}$ . We derive the probability density of  $Y$ , denoted  $f_Y(y)$ , in terms of  $f_X(x)$  and  $g$ .

First assume  $g(\cdot)$  is monotonically increasing, so  $g^{-1}(\cdot)$  can be well-defined over  $\mathcal{X}$  and will also be monotonically increasing. The cumulative distribution function of  $Y$  is given by

$$F_Y(y) = \Pr(Y \leq y) \quad (1.7.38)$$

$$= \Pr(g(X) \leq y) \quad (1.7.39)$$

$$= \Pr(X \leq g^{-1}(y)) \quad (1.7.40)$$

$$= F_X(g^{-1}(y)) \quad (1.7.41)$$

Differentiating  $F_Y(y)$  with respect to  $y$  gives the density of  $Y$ :

$$f_Y(y) = \frac{d}{dy} F_Y(y) \quad (1.7.42)$$

$$= \frac{d}{dy} F_X(g^{-1}(y)) \quad (1.7.43)$$

Using the chain rule,

$$f_Y(y) = \frac{d}{dx} F_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) \quad (1.7.44)$$

$$= f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) \quad (1.7.45)$$

Now assume  $g$  is monotonically decreasing, so  $g^{-1}(\cdot)$  is also monotonically decreasing and has the ability to flip an inequality sign. Similar to before, we can write

$$F_Y(y) = \Pr(Y \leq y) \quad (1.7.46)$$

$$= \Pr(g(X) \leq y) \quad (1.7.47)$$

$$= \Pr(X \geq g^{-1}(y)) \quad (1.7.48)$$

$$= 1 - F_X(g^{-1}(y)) \quad (1.7.49)$$

Now differentiating  $F_Y$  gives

$$f_Y(y) = \frac{d}{dy} F_Y(y) \quad (1.7.50)$$

$$= -\frac{d}{dy} F_X(g^{-1}(y)) \quad (1.7.51)$$

$$= -\frac{d}{dx} F_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) \quad (1.7.52)$$

$$= -f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) \quad (1.7.53)$$

Since by monotonically decreasing  $g^{-1}(\cdot)$ , we have  $\frac{d}{dy} g^{-1}(y) < 0$  hence this density is positive.

Combining the two cases, we get

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| \quad (1.7.54)$$

### 1.7.5 Probability Integral Transform

Suppose  $X$  is a continuous random variable with CDF  $F(x)$  and PDF  $f(x)$ . Then the transformed random variable  $U = F(X)$  is uniformly distributed on  $(0, 1)$ , i.e.

$$F(X) \sim \text{Uniform}(0, 1) \quad (1.7.55)$$

To show this, first note that because  $X$  is continuous, then  $F(x)$  is strictly increasing over its support. Then applying the formula for strictly monotonic transformations of random variables with the transformation  $F(\cdot)$ , we have for the PDF of  $U$ :

$$f_U(u) = f(F^{-1}(u)) \frac{d}{du} F^{-1}(u) \quad (1.7.56)$$

Using the inverse function theorem gives

$$\frac{d}{du} F^{-1}(u) = \frac{1}{f(F^{-1}(u))} \quad (1.7.57)$$

which cancels out the first factor. Therefore  $f_U(u) = 1$  over the support of  $U$ . Since the range/image of  $F(\cdot)$  is  $[0, 1]$ , then the support of  $U$  is  $[0, 1]$ . But since  $U$  will be a continuous random variable as well, then we can equally say that the support of  $U$  is  $(0, 1)$ . Therefore  $U$  has a Uniform  $(0, 1)$  distribution.

Note that this transform will not work for discrete random variables, because if  $X$  is a discrete random variable, then  $F(X)$  will also be a discrete random variable, so it cannot be a continuous uniform distribution.

### 1.7.6 Inverse Transform Sampling

If  $U$  is uniformly distributed on  $(0, 1)$ , we can find a transformation  $T(\cdot)$  such that  $T(U)$  has any arbitrary distribution as desired. This technique is also known as the *inverse probability integral transform*, which as the name suggests, can be thought of as the reverse of the . However

unlike the probability integral transform (which only works for continuous random variables), this transformation works to generate both continuous and discrete random variables, and can be found as follows. Suppose we desire  $T(U)$  to be identically distributed to  $X$ , which has cumulative distribution function  $F_X(x)$ . Then we can choose  $T(\cdot)$  as  $F_X^{-1}(\cdot)$ , the quantile function of  $X$ . Then because  $F_X^{-1}(U) \leq x$  and  $U \leq F_X(x)$  are the same event through the quantile function, we have

$$\Pr(T(U) \leq x) = \Pr(F_X^{-1}(U) \leq x) \quad (1.7.58)$$

$$= \Pr(U \leq F_X(x)) \quad (1.7.59)$$

$$= F_X(x) \quad (1.7.60)$$

since the cumulative distribution of  $U$  is  $F_U(u) = \Pr(U \leq u) = u$ . Therefore  $T(U)$  has the same distribution as  $X$ . Note that if  $X$  is discrete, then the choice of  $T(\cdot)$  is not unique - any function which appropriately assigns values to partitions of  $[0, 1]$  can generate  $X$ .

This method is useful for sampling from arbitrary distributions, because we can simply sample  $U$  from the uniform distribution and transform it by the quantile function of the target distribution. How this works intuitively is that the steeper sections of the CDF (corresponding to more likely values of the random variable) become flatter in the quantile function, so are more likely to be ‘intersected’ by a uniform random variable. Conversely, flatter sections of the CDF (corresponding to rarer values) become steeper in the quantile function so are harder to be intersected by a uniform random variable.

### Inverse Transform Sampling from Truncated Distributions

Suppose we would like to sample from a truncated distribution, that is, given a distribution for  $X$  with CDF  $F_X(x)$  and an interval  $[a, b]$ , we would like to sample from the conditional distribution  $F_{X|a \leq X \leq b}(x) = \Pr(X \leq x | a \leq X \leq b)$ . This can be done similar as above except by generating a uniform random variable between  $F(a)$  and  $F(b)$ .

#### 1.7.7 Gauss’ Approximation Theorem [26]

For a differentiable function  $g(X)$  of a single random variable  $X$ , we can make the approximations

$$\mathbb{E}[g(X)] \approx g(\mathbb{E}[X]) \quad (1.7.61)$$

$$\text{Var}(g(X)) \approx \text{Var}(X) g'(\mathbb{E}[X])^2 \quad (1.7.62)$$

This is achieved using a first-order Taylor series expansion of  $g(X)$  about  $m := \mathbb{E}[X]$  as follows

$$g(X) \approx g(m) + (X - m) g'(m) \quad (1.7.63)$$

Taking the expectations of both sides gives

$$\mathbb{E}[g(X)] \approx \mathbb{E}[g(m)] + \mathbb{E}[(X - m) g'(m)] \quad (1.7.64)$$

$$= \mathbb{E}[g(\mathbb{E}[X])] + \mathbb{E}[(X - \mathbb{E}[X])] g'(\mathbb{E}[X]) \quad (1.7.65)$$

$$= g(\mathbb{E}[X]) + (\mathbb{E}[X] - \mathbb{E}[X]) g'(\mathbb{E}[X]) \quad (1.7.66)$$

$$= g(\mathbb{E}[X]) \quad (1.7.67)$$

Taking the variance of both sides of the expansion gives

$$\text{Var}(g(X)) \approx \text{Var}(g(m) + (X - m) g'(m)) \quad (1.7.68)$$

$$= \text{Var}((X - m) g'(m)) \quad (1.7.69)$$

$$= \text{Var}(X - m) g'(m)^2 \quad (1.7.70)$$

$$= \text{Var}(X) g'(m)^2 \quad (1.7.71)$$

$$= \text{Var}(X) g'(\mathbb{E}[X])^2 \quad (1.7.72)$$

### 1.7.8 Variance-Stabilising Transforms

Suppose there is a random variable  $X$  with mean  $\mathbb{E}[X] = \mu$  and variance as some function of the mean, given by  $\text{Var}(X) = v(\mu)$ . This applies to many single-parameter families such as the Poisson and exponential distributions. Having this dependence between the mean and variance can be undesirable. For example, if we specify the regression model  $\mathbb{E}[Y|Z] = \beta_0 + \beta_1 Z$ , then the conditional variance will be some  $\text{Var}(Y|Z) = v(\beta_0 + \beta_1 Z)$ , which would be classified as heteroskedasticity. Therefore, we seek a transformation  $g(X)$  such that  $\text{Var}(g(X))$  is (roughly) constant. This is said to be a variance-stabilising transformation. By Gauss' approximation theorem, we approximate  $\text{Var}(g(X))$  as

$$\text{Var}(g(X)) \approx \text{Var}(X) g'(\mathbb{E}[X])^2 \quad (1.7.73)$$

$$= v(\mu) g'(\mu)^2 \quad (1.7.74)$$

Imposing the relation

$$v(\mu) g'(\mu)^2 = c \quad (1.7.75)$$

for some  $c > 0$ , we rewrite this as

$$\left( \frac{dg}{d\mu} \right)^2 = \frac{c}{v(\mu)} \quad (1.7.76)$$

and obtain the differential equation

$$\frac{dg}{d\mu} = \frac{\sqrt{c}}{\sqrt{v(\mu)}} \quad (1.7.77)$$

After a separation of variables:

$$\int dg = \sqrt{c} \int \sqrt{v(\mu)} d\mu \quad (1.7.78)$$

and so we obtain the transformation

$$g(x) = \sqrt{c} \int \sqrt{v(x)} dx + d \quad (1.7.79)$$

with arbitrary constants  $c > 0$  and  $d$ .

### 1.7.9 Independence of Transformed Random Variables

If  $X$  and  $Y$  are independent random variables, then the transformed random variables  $f(X)$  and  $g(Y)$  are also independent. This is shown by first using the definition of independence that for sets  $A$  and  $B$ ,

$$\Pr(X \in A \cap Y \in B) = \Pr(X \in A) \Pr(Y \in B) \quad (1.7.80)$$

Then note the equivalence of events

$$\{f(X) \in A\} \equiv \{X \in f^{-1}(A)\} \quad (1.7.81)$$

$$\{g(Y) \in B\} \equiv \{Y \in g^{-1}(B)\} \quad (1.7.82)$$

So

$$\Pr(f(X) \in A \cap g(Y) \in B) = \Pr(X \in f^{-1}(A) \cap Y \in g^{-1}(B)) \quad (1.7.83)$$

$$= \Pr(X \in f^{-1}(A)) \Pr(Y \in g^{-1}(B)) \quad (1.7.84)$$

$$= \Pr(f(X) \in A) \Pr(g(Y) \in B) \quad (1.7.85)$$

Note that if  $X$  and  $Y$  were dependent, then  $f(X)$  and  $g(Y)$  need not necessarily be dependent. A trivial counter example is a case where  $f(X)$  is a constant.

### 1.7.10 Products of Random Variables

The product distribution of random variables  $X$  and  $Y$  is the distribution of  $Z = XY$ . If  $X$  and  $Y$  are independent and have probability density functions  $f_X(x)$  and  $f_Y(y)$  respectively, then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx \quad (1.7.86)$$

*Proof.* We can express the probability density of  $Z$  as the integration of the joint density of  $X$  and  $Y$  over the region where  $XY = Z$ .

$$f_Z(z) = \int \int_{\{x,y:xy=z\}} f_X(x) f_Y(y) dy dx \quad (1.7.87)$$

Make the substitution  $y = z/x$  and rewrite the integral using the Dirac delta function:

$$f_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x) f_Y(z/x) \delta(xy - z) dy dx \quad (1.7.88)$$

$$= \int_{-\infty}^{\infty} f_X(x) f_Y(z/x) \int_{-\infty}^{\infty} \delta(xy - z) dy dx \quad (1.7.89)$$

Note the scaling property of the Dirac delta function, that  $\delta(xy) = \frac{\delta(y)}{|x|}$  (this is to keep the property that  $\int_{-\infty}^{\infty} |x| \delta(xy) dy = \int_{-\infty}^{\infty} \delta(y) dy = 1$ ). Hence

$$\int_{-\infty}^{\infty} \delta(xy - z) dy = \frac{1}{|x|} \quad (1.7.90)$$

and

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx \quad (1.7.91)$$

□

Furthermore, for independent  $X$  and  $Y$ , the expectation of  $Z = XY$  is

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \quad (1.7.92)$$

*Proof.* Using the definition of covariance,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (1.7.93)$$

Then apply  $\text{Cov}(X, Y) = 0$  due to independence. □

If  $X$  and  $Y$  are not independent but have the joint density function  $f_{XY}(x, y)$ , then

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, z/x) \frac{1}{|x|} dx \quad (1.7.94)$$

### 1.7.11 Ratios of Random Variables

The ratio (or quotient) distribution of random variables  $X$  and  $Y$  is the distribution of  $W = \frac{X}{Y}$ . If  $X$  and  $Y$  are independent and have probability density functions  $f_X(x)$  and  $f_Y(y)$  respectively, then

$$f_W(w) = \int_{-\infty}^{\infty} f_X(wy) f_Y(y) |y| dy \quad (1.7.95)$$

The proof is similar to the derivation of the product distribution.

*Proof.* We can express the probability density of  $W$  as the integration of the joint density of  $X$  and  $Y$  over the region where  $X/Y = W$ .

$$f_W(w) = \int \int_{\{x,y:x/y=w\}} f_X(x) f_Y(y) dx dy \quad (1.7.96)$$

Make the substitution  $x = wy$  and rewrite the integral using the Dirac delta function:

$$f_W(w) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(wy) f_Y(y) \delta(x/y - w) dy dx \quad (1.7.97)$$

$$= \int_{-\infty}^{\infty} f_X(wy) f_Y(y) \int_{-\infty}^{\infty} \delta(x/y - w) dx dy \quad (1.7.98)$$

By the scaling property of the Dirac delta function,

$$\int_{-\infty}^{\infty} \delta(x/y - z) dx = |y| \quad (1.7.99)$$

Hence

$$f_W(w) = \int_{-\infty}^{\infty} f_X(wy) f_Y(y) |y| dy \quad (1.7.100)$$

□

If  $X$  and  $Y$  are not independent but have the joint density function  $f_{XY}(x, y)$ , then

$$f_W(w) = \int_{-\infty}^{\infty} f_{XY}(wy, y) \frac{1}{|y|} dy \quad (1.7.101)$$

The mean of the ratio distribution (even in the independent case) is not as straightforward compared to the product distribution, however can be approximated using Taylor expansions. To illustrate, we first develop an extension of Gauss' approximation theorem to a nonlinear function in two variables,  $f(x, y)$ . Suppose random variables  $X$  and  $Y$  have means  $\mu_X$  and  $\mu_Y$  respectively with  $\mu_Y \neq 0$ . Then a second order Taylor series approximation of  $f(X, Y)$  about the mean  $(\mu_X, \mu_Y)$  gives

$$\begin{aligned} f(X, Y) &\approx f(\mu_X, \mu_Y) + \nabla f(x, y)^\top \Big|_{(x, y)=(\mu_X, \mu_Y)} \begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix} \\ &\quad + \frac{1}{2} [X - \mu_X \quad Y - \mu_Y] \nabla^2 f(x, y) \Big|_{(x, y)=(\mu_X, \mu_Y)} \begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix} \end{aligned} \quad (1.7.102)$$

If we take the expectation of this, the first order term vanishes because

$$\mathbb{E} \left[ \begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix} \right] = \begin{bmatrix} \mu_X - \mu_X \\ \mu_Y - \mu_Y \end{bmatrix} \quad (1.7.103)$$

We are then left with

$$\begin{aligned} \mathbb{E}[f(X, Y)] &\approx f(\mu_X, \mu_Y) + \frac{1}{2}\mathbb{E}[(X - \mu_X)^2] \left( \frac{\partial^2 f(x, y)}{\partial x^2} \right) \Big|_{(x,y)=(\mu_X,\mu_Y)} \\ &+ \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \left( \frac{\partial^2 f(x, y)}{\partial x \partial y} \right) \Big|_{(x,y)=(\mu_X,\mu_Y)} + \mathbb{E}[(Y - \mu_Y)^2] \left( \frac{\partial^2 f(x, y)}{\partial y^2} \right) \Big|_{(x,y)=(\mu_X,\mu_Y)} \end{aligned} \quad (1.7.104)$$

For the case  $f(x, y) = \frac{x}{y}$ , the second partial derivatives evaluate to

$$\frac{\partial^2 f(x, y)}{\partial x^2} = 0 \quad (1.7.105)$$

$$\frac{\partial^2 f(x, y)}{\partial x \partial y} = -\frac{1}{y^2} \quad (1.7.106)$$

$$\frac{\partial^2 f(x, y)}{\partial y^2} = \frac{2x}{y^3} \quad (1.7.107)$$

Hence

$$\mathbb{E}\left[\frac{X}{Y}\right] \approx \frac{\mu_X}{\mu_Y} - \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \frac{1}{\mu_Y^2} + \mathbb{E}[(Y - \mu_Y)^2] \frac{\mu_X}{\mu_Y^3} \quad (1.7.108)$$

$$= \frac{\mathbb{E}[X]}{\mathbb{E}[Y]} - \frac{\text{Cov}(X, Y)}{\mathbb{E}[Y]^2} + \text{Var}(Y) \frac{\mathbb{E}[X]}{\mathbb{E}[Y]^3} \quad (1.7.109)$$

To approximate the variance, taking a first order Taylor approximation (for simplicity) gives

$$\text{Var}(f(X, Y)) \approx \text{Var}\left(f(\mu_X, \mu_Y) + \nabla f(x, y)^\top \Big|_{(x,y)=(\mu_X,\mu_Y)} \begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix}\right) \quad (1.7.110)$$

$$= \text{Var}\left(\nabla f(x, y)^\top \Big|_{(x,y)=(\mu_X,\mu_Y)} \begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix}\right) \quad (1.7.111)$$

$$= \text{Var}\left(X \frac{\partial f(x, y)}{\partial x} \Big|_{(x,y)=(\mu_X,\mu_Y)} + Y \frac{\partial f(x, y)}{\partial y} \Big|_{(x,y)=(\mu_X,\mu_Y)}\right) \quad (1.7.112)$$

$$= \left(\frac{\partial f(x, y)}{\partial x}\right)^2 \Big|_{(x,y)=(\mu_X,\mu_Y)} \cdot \text{Var}(X) \quad (1.7.113)$$

$$+ 2 \left(\frac{\partial f(x, y)}{\partial x} \cdot \frac{\partial f(x, y)}{\partial y}\right) \Big|_{(x,y)=(\mu_X,\mu_Y)} \cdot \text{Cov}(X, Y)$$

$$+ \left(\frac{\partial f(x, y)}{\partial y}\right)^2 \Big|_{(x,y)=(\mu_X,\mu_Y)} \cdot \text{Var}(Y)$$

Again by treating  $f(x, y) = \frac{x}{y}$ ,

$$\text{Var}\left(\frac{X}{Y}\right) \approx \frac{1}{\mu_Y^2} \text{Var}(X) - 2 \frac{\mu_X}{\mu_Y^3} \text{Cov}(X, Y) + \frac{\mu_X^2}{\mu_Y^4} \text{Var}(Y) \quad (1.7.114)$$

$$= \frac{\text{Var}(X)}{\mathbb{E}[Y]^2} - 2 \frac{\mathbb{E}[X]}{\mathbb{E}[Y]^3} \text{Cov}(X, Y) + \frac{\mathbb{E}[X]^2}{\mathbb{E}[Y]^4} \text{Var}(Y) \quad (1.7.115)$$

### 1.7.12 Decompositions of Random Variables

The distribution of random variable  $Z$  is said to be decomposable if  $Z$  can be expressed as the sum of non-constant independent random variables  $X$  and  $Y$ , i.e.  $Z = X + Y$ . If it is not possible, the distribution of  $Z$  is said to be indecomposable.

## Divisibility of Random Variables

The distribution of random variable  $Z$  is said to be divisible if  $Z$  can be expressed as the sum of independent and identically distributed random variables  $X_1$  and  $X_2$ . If  $Z$  can be expressed as a sum of arbitrarily many i.i.d. random variables, then the distribution of  $Z$  is said to be infinitely divisible. This is linked to the notion of **stable distributions**. If  $Z$  belongs to a stable family of distributions, then it is infinitely divisible (but not all infinitely divisible families of random variables need to belong to the class of stable distributions).

### Cramér's Decomposition Theorem [131]

The sum of independent Gaussian random variables is also Gaussian distributed. Cramér's decomposition theorem states the converse: that if  $X$  and  $Y$  are independent real-valued random variables whose sum  $X+Y$  is a Gaussian random variable, then  $X$  and  $Y$  must also be Gaussian. By induction, if the sum of finitely many independent random variables is Gaussian, then the summands must be Gaussian.

#### 1.7.13 Mixture Distributions

If  $f_1(x)$  and  $f_2(x)$  are probability density functions, then we see that

$$f(x) = w_1 f_1(x) + w_2 f_2(x) \quad (1.7.116)$$

with  $w_1 + w_2 = 1$  satisfies the properties of a probability density function and we say that  $f(x)$  is a mixture distribution of  $f_1(x)$  and  $f_2(x)$  with weights  $w_1$  and  $w_2$  respectively. More generally, given  $n$  density functions  $f_1(x), \dots, f_n(x)$ , then

$$f(x) = \sum_{i=1}^n w_i f_i(x) \quad (1.7.117)$$

with  $\sum_i^n w_i = 1$  is said to be a mixture distribution of  $f_1(x), \dots, f_n(x)$  with weights  $w_1, \dots, w_n$  respectively. To generate a realisation from the mixture distribution, we can imagine first selecting a distribution randomly such that each distribution  $f_i(x)$  has probability  $w_i$  of being selected, and then subsequently generating a realisation from that distribution.

Note that a mixture distribution is not the same as taking a convex combination of random variables (i.e. the random variable  $Z = \frac{1}{2}X + \frac{1}{2}Y$  will generally not have the density function  $\frac{1}{2}f_X(z) + \frac{1}{2}f_Y(z)$ ).

#### 1.7.14 Compound Distributions

A compound probability distribution is a parametrised distribution, where the parameters themselves are randomly distributed. Let  $f_{X|\theta}(x|\theta)$  denote a family of distributions parametrised in  $\theta$ . If  $\theta$  is now a random quantity which has its own distribution  $f_\theta(\theta)$  on support  $\Theta$ , then this induces an unconditional distribution for  $X$ , which can be obtained by marginalising over the parameter:

$$f_X(x) = \int_{\Theta} f_{X|\theta}(x|\theta) f_\theta(\theta) d\theta \quad (1.7.118)$$

A mixture distribution is an example of a compound distribution.

### 1.7.15 Truncated Distributions

Let  $X$  be a continuous distribution with PDF  $f_X(x)$  and CDF  $F_X(x)$ . The truncation of  $X$  on an interval  $[a, b]$  with  $-\infty < a < b < \infty$  is defined to be a random variable  $Y$  with the probability density:

$$f_Y(y) \propto \begin{cases} f_X(y), & y \in [a, b], \\ 0, & y \notin [a, b] \end{cases} \quad (1.7.119)$$

To determine the normalising constant, we can perform integration:

$$\int_a^b f_X(x) dx = F_X(b) - F_X(a) \quad (1.7.120)$$

Thus

$$f_Y(y) = \begin{cases} \frac{f_X(y)}{F_X(b) - F_X(a)}, & y \in [a, b], \\ 0, & y \notin [a, b] \end{cases} \quad (1.7.121)$$

The CDF can consequently be found as

$$F_Y(y) = \int_{-\infty}^y f_Y(t) dt \quad (1.7.122)$$

$$= \frac{1}{F_X(b) - F_X(a)} \int_b^y f_X(t) dt \quad (1.7.123)$$

$$= \frac{F_X(y) - F_X(a)}{F_X(b) - F_X(a)} \quad (1.7.124)$$

valid when  $y \in [a, b]$ , and appropriately saturated between  $[0, 1]$  otherwise. The random  $Y$  can also be interpreted as the conditional distribution of  $X$ , given  $a \leq X \leq b$ . This can be shown as follows, using  $y \in [a, b]$ :

$$\Pr(X \leq y | a \leq X \leq b) = \frac{\Pr(X \leq y, a \leq X \leq b)}{\Pr(a \leq X \leq b)} \quad (1.7.125)$$

$$= \frac{\Pr(a \leq X \leq y)}{\Pr(a \leq X \leq b)} \quad (1.7.126)$$

$$= \frac{1 - \Pr(X < a) - \Pr(X > y)}{1 - \Pr(X < a) - \Pr(X > b)} \quad (1.7.127)$$

$$= \frac{1 - F_X(a) - (1 - F_X(y))}{1 - F_X(a) - (1 - F_X(b))} \quad (1.7.128)$$

$$= \frac{F_X(y) - F_X(a)}{F_X(b) - F_X(a)} \quad (1.7.129)$$

$$= F_Y(y) \quad (1.7.130)$$

### Left-Truncated Distributions

If a distribution is truncated on  $[a, \infty)$ , we say that it is left-truncated, or alternatively *truncated from below*. Formally, we can consider truncation on  $[a, b]$  and take the limit as  $b \rightarrow \infty$ , for which  $F_X(b) \rightarrow 1$ . Hence the distribution is specified by

$$f_Y(y) = \frac{f_X(y)}{1 - F_X(a)} \mathbb{I}_{\{y \geq a\}} \quad (1.7.131)$$

$$F_Y(y) = \max \left\{ 0, \min \left\{ \frac{F_X(y) - F_X(a)}{1 - F_X(a)}, 1 \right\} \right\} \quad (1.7.132)$$

## Right-Truncated Distributions

If a distribution is truncated on  $(-\infty, b]$ , we say that it is right-truncated, or alternatively *truncated from above*. Formally, we can consider truncation on  $[a, b]$  and take the limit as  $a \rightarrow -\infty$ , for which  $F_X(a) \rightarrow 0$ . Hence the distribution is specified by

$$f_Y(y) = \frac{f_X(y)}{F_X(b)} \mathbb{I}_{\{y \leq b\}} \quad (1.7.133)$$

$$F_Y(y) = \max \left\{ 0, \min \left\{ \frac{F_X(y)}{F_X(b)}, 1 \right\} \right\} \quad (1.7.134)$$

## 1.8 Families of Continuous Univariate Probability Distributions

### 1.8.1 Dirac Delta Distribution

The Dirac delta function  $\delta(x)$  (also known as the *unit impulse function*) can be heuristically characterised by

$$\delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & x \neq 0 \end{cases} \quad (1.8.1)$$

where

$$\int_{-\infty}^{\infty} \delta(x) = 1 \quad (1.8.2)$$

A more formal way to derive this function is by starting from the rectangle function

$$f(x) = \begin{cases} 1/T, & x \in [-T/2, T/2] \\ 0, & x \notin [-T/2, T/2] \end{cases} \quad (1.8.3)$$

which integrates to one. Then we may consider  $\delta(x)$  as the limit of  $f(x)$  as  $T \rightarrow 0$ . That is, the Dirac delta function represents an infinite ‘spike’ (or alternatively, an impulse or point mass) at  $x = 0$ , which that the area underneath the spike integrates to one. Thus, we can treat the Dirac delta function as a valid probability density function. Its cumulative distribution function is the step function, given by

$$F(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (1.8.4)$$

A random variable  $X$  with a Dirac delta distribution will be a **degenerate random variable** at zero, i.e.  $\Pr(X = 0) = 1$ . By shifting the distribution, e.g.  $\delta(c - x)$ , we can represent the degenerate random variable such that  $\Pr(X = c) = 1$ .

### Density Functions of Discrete Random Variables

Every real-valued random variable (regardless of being continuous or discrete) has a cumulative distribution function  $F(x)$ . If we consider some **discrete random variable**  $X$  with masses  $\{p_1, p_2, \dots\}$  with  $\sum_i p_i = 1$  at respective support points  $\{x_1, x_2, \dots\}$ , it is evident that  $F(x)$  will look like a non-decreasing staircase function. So if we wanted to represent  $X$  with a density function satisfying  $\lim_{x \rightarrow -\infty} F(x) = 0$ ,  $\lim_{x \rightarrow \infty} F(x) = 1$  and  $F(x) = \int_{-\infty}^x f(t) dt$ , we can construct  $f(x)$  for a discrete random variable by building it out of Dirac delta functions weighted by masses  $\{p_1, p_2, \dots\}$  placed at the support points  $\{x_1, x_2, \dots\}$ . For example if there were  $n$  support points, we would define  $f(x)$  as

$$f(x) = \sum_{i=1}^n p_i \delta(x - x_i) \quad (1.8.5)$$

which agrees with what we already know about computing the cumulative distribution function for a discrete random variable:

$$F(x) = \Pr(X \leq x) \quad (1.8.6)$$

$$= \int_{-\infty}^x f(t) dt \quad (1.8.7)$$

$$= \int_{-\infty}^x \sum_{i=1}^n p_i \delta(t - x_i) dt \quad (1.8.8)$$

$$= \sum_{i:x_i \leq x} p_i \quad (1.8.9)$$

To summarise, we can imagine the density function of a discrete random variable to be a mixture of shifted Dirac delta functions.

### 1.8.2 Continuous Uniform Distribution

#### Irwin-Hall Distribution

#### Bates Distribution

### 1.8.3 Exponential Distribution

The exponential distribution is given by the PDF

$$f(x) = \lambda e^{-\lambda x} \quad (1.8.10)$$

over support  $x \geq 0$ , with rate parameter  $\lambda > 0$ . To verify that this is a valid distribution,

$$\int_0^\infty f(x) dx = \int_0^\infty \lambda e^{-\lambda x} dx \quad (1.8.11)$$

$$= \left[ -e^{-\lambda x} \right]_0^\infty \quad (1.8.12)$$

$$= 0 - (-1) \quad (1.8.13)$$

$$= 1 \quad (1.8.14)$$

The form of the CDF can also be obtained by integrating:

$$F(x) = \int f(x) dx \quad (1.8.15)$$

$$= -e^{-\lambda x} + c \quad (1.8.16)$$

To find the constant of integration  $c$ , use the property  $F(0) = 0$ , giving  $c = 1$ , and thus

$$F(x) = 1 - e^{-\lambda x} \quad (1.8.17)$$

The exponential distribution can be used to model continuous-valued waiting/arrival times, under the characterisation that at any instant of time, there is a small chance of an arrival, independent of the past. To derive the exponential distribution from this characterisation, we can use the Poisson distribution, which counts the total number of arrivals when there is a small chance of arrival at every instant of time. If the Poisson distribution has average rate  $\lambda$  per unit of time, then over  $t$  units of time the total number of arrivals will be  $Y \sim \text{Poisson}(\lambda t)$ . Thus

$$\Pr(X \leq t) = 1 - \Pr(X > t) \quad (1.8.18)$$

$$= 1 - \Pr(Y = 0) \quad (1.8.19)$$

since if  $X$  is supposed to be a waiting time, then the probability that we have to wait longer than  $t$  units of time is same as the probability that there are exactly zero arrivals in  $t$  units of time. So applying the probability mass function of the Poisson distribution at zero:

$$\Pr(X \leq t) = 1 - \frac{(x\lambda)^0 e^{-\lambda x}}{0!} \quad (1.8.20)$$

$$= 1 - e^{-\lambda x} \quad (1.8.21)$$

which is the CDF of an exponential distribution with rate  $\lambda$ .

### Memorylessness of Exponential Distribution

Due to the characterisation of the waiting time (chance of arrival is independent of the past), we can formally show that given the waiting is past  $t_1$ , the probability that we will wait further past  $t_2$  (with  $t_2 > t_1$ ) only depends on the time difference  $t_2 - t_1$ , i.e.

$$\Pr(X > t_2 | X > t_1) = \Pr(X > t_2 - t_1) \quad (1.8.22)$$

*Proof.* From the conditional probability:

$$\Pr(X > t_2 | X > t_1) = \frac{\Pr(X > t_2, X > t_1)}{\Pr(X > t_1)} \quad (1.8.23)$$

$$= \frac{\Pr(X > t_2)}{\Pr(X > t_1)} \quad (1.8.24)$$

$$= \frac{1 - \Pr(X \leq t_2)}{1 - \Pr(X \leq t_1)} \quad (1.8.25)$$

$$= \frac{e^{-\lambda t_2}}{e^{-\lambda t_1}} \quad (1.8.26)$$

$$= e^{-\lambda(t_2 - t_1)} \quad (1.8.27)$$

$$= 1 - \Pr(X \leq t_2 - t_1) \quad (1.8.28)$$

$$= \Pr(X > t_2 - t_1) \quad (1.8.29)$$

□

We call this the memorylessness property of the exponential distribution.

### Mean of Exponential Distribution

To compute the expectation of an exponentially distributed random variable  $X \sim \text{Exp}(\lambda)$ , we can use integration by parts, with  $x$  as the differentiable function and  $f(x)$  as the integrable function:

$$\mathbb{E}[X] = \int_0^\infty x f(x) dx \quad (1.8.30)$$

$$= [xF(x)]_0^\infty - \int_0^\infty F(x) dx \quad (1.8.31)$$

$$= \left[ x(1 - e^{-\lambda x}) \right]_0^\infty - \int_0^\infty (1 - e^{-\lambda x}) dx \quad (1.8.32)$$

$$= \left[ x - xe^{-\lambda x} \right]_0^\infty - \left[ x + \frac{e^{-\lambda x}}{\lambda} \right]_0^\infty \quad (1.8.33)$$

$$= - \left[ xe^{-\lambda x} \right]_0^\infty - \left[ \frac{e^{-\lambda x}}{\lambda} \right]_0^\infty \quad (1.8.34)$$

$$= - \lim_{x \rightarrow \infty} xe^{-\lambda x} + 0 - \lim_{x \rightarrow \infty} e^{-\lambda x} + 0 \frac{1}{\lambda} \quad (1.8.35)$$

Using L'Hôpital's rule, we may show

$$\lim_{x \rightarrow \infty} xe^{-\lambda x} = \lim_{x \rightarrow \infty} \frac{x}{e^{\lambda x}} \quad (1.8.36)$$

$$= \lim_{x \rightarrow \infty} \frac{1}{\lambda e^{\lambda x}} \quad (1.8.37)$$

$$= 0 \quad (1.8.38)$$

Therefore we are left with

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad (1.8.39)$$

## Hyperexponential Distribution

### 1.8.4 Gaussian Distribution

The Gaussian distribution, also known as the *normal distribution*, is a ubiquitous distribution which has the following functional form for its PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right] \quad (1.8.40)$$

where  $\mu$  is the mean (location parameter) and  $\sigma$  is the standard deviation (scale parameter). It is bell-curve shaped, symmetric, and also has a median and mode of  $\mu$ . To denote the probability density of a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , we may write  $\mathcal{N}(\mu, \sigma^2)$ . To denote that a random variable  $X$  is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , we may write

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad (1.8.41)$$

The *standard Gaussian distribution* (or standard normal distribution) is a special case of the Gaussian distribution which has a mean of  $\mu = 0$  and a standard deviation of  $\sigma = 1$  (hence also the variance  $\sigma^2 = 1$ ).

A ‘first principles’ derivation of the Gaussian distribution is as follows [78]. Consider random coordinates  $(X, Y)$  in the Cartesian plane. We place the following restrictions on  $(X, Y)$ :

- $X$  and  $Y$  are independent and identically distributed.
- Larger  $|X|$  and  $|Y|$  are less likely (i.e. have smaller probability density).
- The likelihood of landing on a point  $(x, y)$  depends only on the distance from the origin  $r = \sqrt{x^2 + y^2}$ , and not the angle  $\theta = \tan^{-1}\left(\frac{y}{x}\right)$ .

An analogy for this set of assumptions could be dropping a dart on a horizontal plane, aiming for the origin  $(0, 0)$ . We can then show that the distribution of  $X$  (or identically,  $Y$ ) is Gaussian. First, consider an arbitrary point  $(x, y)$ . The joint probability density at the point is  $f(x)f(y)$ , due to independence. We also have, under a change of coordinates,

$$f(x)f(y) = g(r) \quad (1.8.42)$$

since the likelihood only depends on  $r$  and not  $\theta$ . Taking partial derivatives of both sides with respect to  $\theta$  using the product rule:

$$f(x) \frac{\partial f(y)}{\partial \theta} + f(y) \frac{\partial f(x)}{\partial \theta} = 0 \quad (1.8.43)$$

Then using the chain rule:

$$f(x)f'(y)\frac{\partial y}{\partial \theta} + f(y)f'(x)\frac{x}{\partial \theta} = 0 \quad (1.8.44)$$

We use the relation  $x = r \cos \theta$  and  $y = r \sin \theta$  to give

$$\frac{\partial x}{\partial \theta} = -r \sin \theta \quad (1.8.45)$$

$$= -y \quad (1.8.46)$$

and

$$\frac{\partial y}{\partial \theta} = r \cos \theta \quad (1.8.47)$$

$$= x \quad (1.8.48)$$

Hence

$$f(x)f'(y)x - f(y)f'(x)y = 0 \quad (1.8.49)$$

$$\frac{f'(x)}{xf(x)} = \frac{f'(y)}{yf(y)} \quad (1.8.50)$$

This means that  $\frac{f'(x)}{xf(x)}$  must be equal to some constant, since the relation holds for any arbitrary  $(x, y)$ . That is, if we sample some realisation  $x_1$  and an independent copy  $x_2$ , it will always be true that  $\frac{f'(x_1)}{x_1 f(x_1)} = \frac{f'(x_2)}{x_2 f(x_2)}$ . Therefore we solve the differential equation

$$\frac{f'(x)}{xf(x)} = K \quad (1.8.51)$$

$$\frac{f'(x)}{f(x)} = Kx \quad (1.8.52)$$

Taking the anti-derivative of both sides,

$$\log f(x) = \frac{1}{2}Kx^2 + C \quad (1.8.53)$$

$$f(x) = e^{Kx^2/2+C} \\ = Ae^{Kx^2/2} \quad (1.8.54)$$

where  $A = e^C$ . Lastly, since the likelihood is decreasing the further away from the origin, it makes sense if  $K$  is negative. Letting  $K = -1/\sigma^2$ , we have the form of a zero-mean Gaussian:

$$f(x) = A \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (1.8.55)$$

where the normalising constant  $A$  can be computed from the Gaussian integral. Suppose  $\sigma = 1$ , so that

$$|f'(x)| = |xf(x)| \quad (1.8.56)$$

This characterises the shape of the standard Gaussian distribution as the slope of the density function being equal to the area of the box with corners being  $(0, 0)$  and  $(x, f(x))$ .

## Error Function

The error function (also called the Gauss error function) is defined as

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt \quad (1.8.57)$$

which has an image of  $(-1, 1)$ . The cumulative distribution function of the standard Gaussian distribution  $\Phi(x)$  may be defined in terms of the error function with the relationship

$$\Phi(x) = \frac{1}{2} \left( 1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right) \quad (1.8.58)$$

which we can show by

$$\Phi(x) = \frac{1}{2} \left( 1 + \frac{1}{\sqrt{\pi}} \int_{-x/\sqrt{2}}^{x/\sqrt{2}} e^{-t^2} dt \right) \quad (1.8.59)$$

$$= \frac{1}{2} \left( 1 + \frac{2}{\sqrt{\pi}} \int_0^{x/\sqrt{2}} e^{-t^2} dt \right) \quad (1.8.60)$$

With the change of variables  $u = \sqrt{2}t$  hence  $t = u/\sqrt{2}$  and  $du = \sqrt{2}dt$ , this becomes

$$\Phi(x) = \frac{1}{2} \left( 1 + \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2/2} \frac{du}{\sqrt{2}} \right) \quad (1.8.61)$$

$$= \frac{1}{2} + \int_0^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \quad (1.8.62)$$

$$= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du + \int_0^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \quad (1.8.63)$$

$$= \int_{-\infty}^x \phi(u) du \quad (1.8.64)$$

## *Q*-Function

The *Q*-function is the complementary cumulative distribution function of the standard Gaussian distribution. It is denoted

$$Q(x) = 1 - \Phi(x) \quad (1.8.65)$$

$$= \int_x^\infty \phi(u) du \quad (1.8.66)$$

$$= \int_x^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}} du \quad (1.8.67)$$

Also due symmetry about  $x = 0$ , the *Q*-function can also be written in terms of the Gaussian CDF as

$$Q(x) = \Phi(-x) \quad (1.8.68)$$

## Complementary Error Function

The complementary error function is defined as

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \quad (1.8.69)$$

which has an image of  $(0, 2)$ . An expression for  $\operatorname{erfc}(x)$  in terms of just a single integral is

$$\operatorname{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (1.8.70)$$

$$= \lim_{x \rightarrow \infty} \operatorname{erf}(x) - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (1.8.71)$$

$$= \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-t^2} dt - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (1.8.72)$$

$$= \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \quad (1.8.73)$$

By applying a change of variables  $t = u/\sqrt{2}$ , we can see that this can be written in terms of the  $Q$ -function by

$$\operatorname{erfc}(x) = 2 \int_{\sqrt{2}x}^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \quad (1.8.74)$$

$$= 2Q(\sqrt{2}x) \quad (1.8.75)$$

or the other way around,

$$Q(x) = \frac{1}{2} \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) \quad (1.8.76)$$

Also, since  $\Phi(x) = Q(-x)$ , then we also have

$$\Phi(x) = \frac{1}{2} \operatorname{erfc}\left(-\frac{x}{\sqrt{2}}\right) \quad (1.8.77)$$

### Gaussian Integrals

The Gaussian integral is the integral  $\int_{-\infty}^\infty e^{-x^2} dx$ . Note that this integral is closely related to the error function as the limits tend to infinity. However unlike the error function with finite argument, the Gaussian integral has a closed-form solution. To compute this integral, first realise

$$\int_{-\infty}^\infty e^{-x^2} dx = 2 \int_0^\infty e^{-x^2} dx \quad (1.8.78)$$

Then squaring both sides,

$$\left(\int_{-\infty}^\infty e^{-x^2} dx\right)^2 = 4 \left(\int_0^\infty e^{-x^2} dx\right)^2 \quad (1.8.79)$$

$$= 4 \int_0^\infty \int_0^\infty e^{-x^2} e^{-y^2} dy dx \quad (1.8.80)$$

$$= 4 \int_0^\infty \int_0^\infty e^{-(x^2+y^2)} dy dx \quad (1.8.81)$$

In the inner integral, make a change of variables  $y = xs$ ,  $dy = xds$  so

$$\left(\int_{-\infty}^\infty e^{-x^2} dx\right)^2 = 4 \int_0^\infty \int_0^\infty e^{-x^2(1+s^2)} x ds dx \quad (1.8.82)$$

$$= 4 \int_0^\infty \int_0^\infty e^{-x^2(1+s^2)} x dx ds \quad (1.8.83)$$

For this inner integral, make a change of variables  $u = x^2$ ,  $du = 2dx$  so that

$$\int_0^\infty e^{-x^2(1+s^2)} x dx = \int_0^\infty e^{-u(1+s^2)} \frac{du}{2} \quad (1.8.84)$$

$$= \left[ -\frac{1}{2(1+s^2)} e^{-u(1+s^2)} \right]_{u=0}^{u=\infty} \quad (1.8.85)$$

$$= - \left( -\frac{1}{2(1+s^2)} \right) \quad (1.8.86)$$

Hence

$$\left( \int_{-\infty}^{\infty} e^{-x^2} dx \right)^2 = 4 \int_0^{\infty} \frac{1}{2(1+s^2)} ds \quad (1.8.87)$$

$$= 2 [\arctan s]_{s=0}^{s=\infty} \quad (1.8.88)$$

$$= \pi \quad (1.8.89)$$

Therefore the evaluation of the Gaussian integral yields

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \quad (1.8.90)$$

More generally, integrals of the form  $\int_{-\infty}^{\infty} a \exp \left[ -(x-b)^2 / (2c^2) \right] dx$  can be evaluated. By a change of variables  $z = \frac{x-b}{c\sqrt{2}}$ , we have

$$\int_{-\infty}^{\infty} a \exp \left[ -(x-b)^2 / (2c^2) \right] dx = a \int_{-\infty}^{\infty} a \exp(-z^2) c\sqrt{2} dz \quad (1.8.91)$$

$$= ac\sqrt{2\pi} \quad (1.8.92)$$

This can be used to compute the normalising constant of a Gaussian density.

### Mill's Ratio

Mill's ratio of a distribution is defined as the ratio of the complementary cumulative distribution function to the probability density function. For the standard Gaussian distribution, Mill's ratio can be written in terms of the Q function as  $\frac{Q(x)}{\phi(x)}$ . This ratio satisfies some well-known inequalities. Firstly, we can show

$$Q(x) = \int_x^{\infty} \phi(u) du \quad (1.8.93)$$

$$< \int_x^{\infty} \frac{u}{x} \phi(u) du \quad (1.8.94)$$

for  $x > 0$ , since the ratio  $u/x > 1$  for  $u > x$  and  $\phi(u) > 0$  everywhere. Then with a change of variables  $v = \frac{u^2}{2}$  (noting that  $\frac{dv}{du} = u$ ):

$$\int_x^{\infty} \frac{u}{x} \phi(u) du = \int_x^{\infty} \frac{u}{x} \cdot \frac{e^{-u^2/2}}{\sqrt{2\pi}} du \quad (1.8.95)$$

$$= \int_{x^2/2}^{\infty} \frac{e^{-v}}{x\sqrt{2\pi}} dv \quad (1.8.96)$$

$$= - \left[ \frac{1}{x\sqrt{2\pi}} e^{-v} \right]_{v=x^2/2}^{v=\infty} \quad (1.8.97)$$

$$= \frac{e^{-x^2/2}}{x\sqrt{2\pi}} \quad (1.8.98)$$

$$= \frac{\phi(x)}{x} \quad (1.8.99)$$

Hence we have the upper bound on Mill's ratio

$$\frac{Q(x)}{\phi(x)} < \frac{1}{x} \quad (1.8.100)$$

for  $x > 0$ . A lower bound can be derived in a similar fashion. Start with

$$\left(1 + \frac{1}{x^2}\right) Q(x) = \int_x^\infty \left(1 + \frac{1}{u^2}\right) \phi(u) du \quad (1.8.101)$$

$$> \int_x^\infty \left(1 + \frac{1}{u^2}\right) \phi(u) du \quad (1.8.102)$$

since  $\frac{1}{x^2} > \frac{1}{u^2}$  for  $u > x$  and  $x$  while  $\phi(u) > 0$ . Now note that the derivative of  $\phi(u)$  is

$$\frac{d\phi(u)}{du} = -u\phi(u) \quad (1.8.103)$$

Then using the quotient rule, the derivative of  $-\phi(u)/u$  is

$$\frac{d}{du} \left(-\frac{\phi(u)}{u}\right) = -\frac{-u^2\phi(u) - \phi(u)}{u^2} \quad (1.8.104)$$

$$= \left(1 + \frac{1}{u^2}\right) \phi(u) \quad (1.8.105)$$

Hence

$$\int_x^\infty \left(1 + \frac{1}{u^2}\right) \phi(u) du = \left[-\frac{\phi(u)}{u}\right]_{u=x}^{u=\infty} \quad (1.8.106)$$

$$= \frac{\phi(x)}{x} \quad (1.8.107)$$

This gives the lower bound

$$\left(1 + \frac{1}{x^2}\right) Q(x) > \frac{\phi(x)}{x} \quad (1.8.108)$$

$$\frac{x}{1+x^2} < \frac{Q(x)}{\phi(x)} \quad (1.8.109)$$

for  $x > 0$ .

### Truncated Normal Distribution

In the truncated normal distribution, a Gaussian distribution is truncated on support  $[a, b]$  with  $-\infty < a < b < \infty$ . If we consider a truncated standard Gaussian, then the PDF and CDF over  $[a, b]$  are given by

$$f(x) = \frac{\phi(x)}{\Phi(b) - \Phi(a)} \quad (1.8.110)$$

$$F(x) = \frac{\Phi(x) - \Phi(a)}{\Phi(b) - \Phi(a)} \quad (1.8.111)$$

respectively. If we consider the truncation of a  $\mathcal{N}(\mu, \sigma^2)$ , then given this is the linear transformation of a standard Gaussian by  $\sigma X + \mu$ , the truncated PDF and CDF over  $[a, b]$  are given in terms of the standard Gaussian PDF and CDF by

$$f(x) = \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\sigma \left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right)} \quad (1.8.112)$$

$$F(x) = \frac{\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \quad (1.8.113)$$

For a left-truncated normal distribution on  $[a, \infty)$ , the PDFs and CDFs over  $[a, b]$  can be written as

$$f(x) = \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\sigma(1 - \Phi\left(\frac{a-\mu}{\sigma}\right))} \quad (1.8.114)$$

$$F(x) = \frac{\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)} \quad (1.8.115)$$

respectively, while for a right-truncated normal distribution on  $(\infty, b]$ , the PDFs and CDFs over  $[a, b]$  can be written as

$$f(x) = \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\sigma\Phi\left(\frac{b-\mu}{\sigma}\right)} \quad (1.8.116)$$

$$F(x) = \frac{\Phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)} \quad (1.8.117)$$

### Folded Normal Distribution

If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $Y = |X|$  has a folded normal distribution. Thus, the distribution can be thought of as having been ‘folded’ about zero. Let  $Z$  be a standard Gaussian random variable. Then we can derive the CDF of  $Y$  by

$$F_Y(y) = \Pr(Y \leq y) \quad (1.8.118)$$

$$= \Pr(-y \leq X \leq y) \quad (1.8.119)$$

$$= \Pr(-y \leq \sigma Z + \mu \leq y) \quad (1.8.120)$$

$$= \Pr\left(\frac{-y - \mu}{\sigma} \leq Z \leq \frac{y - \mu}{\sigma}\right) \quad (1.8.121)$$

$$= 1 - \Pr\left(Z > \frac{y - \mu}{\sigma}\right) - \Pr\left(Z < -\frac{y + \mu}{\sigma}\right) \quad (1.8.122)$$

$$= 1 - \left(1 - \Phi\left(\frac{y - \mu}{\sigma}\right)\right) - \Phi\left(-\frac{y + \mu}{\sigma}\right) \quad (1.8.123)$$

$$= \Phi\left(\frac{y - \mu}{\sigma}\right) - \Phi\left(-\frac{y + \mu}{\sigma}\right) \quad (1.8.124)$$

valid for  $y \geq 0$ . Taking the derivative, we obtain the PDF, valid for  $y \geq 0$ :

$$f_Y(y) = \frac{1}{\sigma}\phi\left(\frac{y - \mu}{\sigma}\right) + \frac{1}{\sigma}\phi\left(-\frac{y + \mu}{\sigma}\right) \quad (1.8.125)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y - \mu)^2}{2\sigma^2}\right] + \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y + \mu)^2}{2\sigma^2}\right] \quad (1.8.126)$$

Using the definition of the hyperbolic cosine  $\cosh x = \frac{e^x + e^{-x}}{2}$ , the PDF can be expressed alternatively as

$$f_Y(y) = \frac{1}{2} \left[ \frac{1}{\sigma\sqrt{\pi/2}} \exp\left(-\frac{y^2 - 2\mu y + \mu^2}{2\sigma^2}\right) + \frac{1}{\sigma\sqrt{\pi/2}} \exp\left(-\frac{y^2 + 2\mu y + \mu^2}{2\sigma^2}\right) \right] \quad (1.8.127)$$

$$= \frac{1}{\sigma\sqrt{\pi/2}} \exp\left(-\frac{y^2 + \mu^2}{2\sigma^2}\right) \cdot \frac{e^{\mu y/\sigma^2} + e^{-\mu y/\sigma^2}}{2} \quad (1.8.128)$$

$$= \frac{1}{\sigma\sqrt{\pi/2}} \exp\left(-\frac{y^2 + \mu^2}{2\sigma^2}\right) \cosh\left(\frac{\mu y}{\sigma^2}\right) \quad (1.8.129)$$

### Half-Normal Distribution

The half-normal distribution is the folded normal distribution  $Y = |X|$  when  $X$  has a mean of  $\mu = 0$ . Thus, the CDF reduces to

$$F_Y(y) = \Phi\left(\frac{y}{\sigma}\right) - \Phi\left(-\frac{y}{\sigma}\right) \quad (1.8.130)$$

$$= \Phi\left(\frac{y}{\sigma}\right) - \left(1 - \Phi\left(\frac{y}{\sigma}\right)\right) \quad (1.8.131)$$

$$= 2\Phi\left(\frac{y}{\sigma}\right) - 1 \quad (1.8.132)$$

and the PDF becomes

$$f_Y(y) = \frac{2}{\sigma} \phi\left(\frac{y}{\sigma}\right) \quad (1.8.133)$$

$$= \frac{1}{\sigma\sqrt{\pi/2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \quad (1.8.134)$$

Thus the density of the half-normal distribution becomes double the density of the zero-mean Gaussian distribution, except valid over  $y \geq 0$ .

### Lognormal Distribution

The random variable  $X$  is lognormal distributed with parameters  $\mu$  and  $\sigma^2$ , if  $\log X \sim \mathcal{N}(\mu, \sigma^2)$ . An alternatively characterisation is if  $Y \sim \mathcal{N}(\mu, \sigma^2)$ , then  $X = e^Y$  is lognormal distributed. To compute the CDF of the lognormal distribution in terms of the standard Gaussian CDF, we have

$$F(x) = \Pr(X \leq x) \quad (1.8.135)$$

$$= \Pr(\log X \leq \log x) \quad (1.8.136)$$

$$= \Pr(Y \leq \log x) \quad (1.8.137)$$

$$= \Phi\left(\frac{\log x - \mu}{\sigma}\right) \quad (1.8.138)$$

To compute the PDF, we can let  $Z \sim \mathcal{N}(0, 1)$ , and consider the transformation

$$X = g(Z) \quad (1.8.139)$$

where  $g(z) = e^{\sigma z + \mu}$  and  $g^{-1}(x) = \frac{\log x - \mu}{\sigma}$ . Then using the expression for PDFs of strictly monotonic transformations, we have

$$f(x) = \phi(g^{-1}(x)) \cdot \frac{d}{dx}g^{-1}(x) \quad (1.8.140)$$

$$= \phi\left(g^{-1}\left(\frac{\log x - \mu}{\sigma}\right)\right) \cdot \frac{1}{x\sigma} \quad (1.8.141)$$

$$= \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right) \quad (1.8.142)$$

## Inverse Gaussian Distribution

### Skew Normal Distribution [11]

The skew normal distribution has the probability density function

$$f(x; \lambda) = 2\Phi(\lambda x) \phi(x) \quad (1.8.143)$$

where  $\lambda \in \mathbb{R}$  is a shape parameter,  $\phi(\cdot)$  is the standard Gaussian PDF and  $\Phi(\cdot)$  is the standard Gaussian CDF. We observe the following.

- If  $\lambda = 0$ , then  $\Phi(\lambda x) = 1/2$  hence  $f(x; 0) = \phi(x)$  which is the standard Gaussian density.
- As  $\lambda \rightarrow \infty$ , then  $\Phi(\lambda x)$  approaches the step function, so the shape of  $f(x; \lambda)$  approaches the shape of the half-normal distribution.
- If  $\lambda > 0$ , the distribution will be positive skewed, because  $\Phi(\lambda x)$  is increasing in  $x$ , so more weight will be assigned to positive values, compared to negative values.
- Likewise, if  $\lambda < 0$ , then the distribution will be negative skewed, because  $\Phi(\lambda x)$  will be decreasing in  $x$ .

The skew normal distribution ‘skews’ the standard Gaussian distribution, because it can be characterised as ‘shifting’ some mass from one side to another. We can also show that the density is valid, by noting that  $\left(\Phi(\lambda x) - \frac{1}{2}\right)\phi(x)$  is an odd function, meaning

$$\int_{-\infty}^{\infty} \left(\Phi(\lambda x) - \frac{1}{2}\right)\phi(x) dx = 0 \quad (1.8.144)$$

Therefore

$$\int_{-\infty}^{\infty} f(x; \lambda) dx = \int_{-\infty}^{\infty} 2\Phi(\lambda x) \phi(x) dx \quad (1.8.145)$$

$$= \int_{-\infty}^{\infty} 2 \left( \Phi(\lambda x) - \frac{1}{2} \right) \phi(x) dx + \int_{-\infty}^{\infty} \phi(x) dx \quad (1.8.146)$$

$$= 1 \quad (1.8.147)$$

### Generalised Skew Normal Distribution [178]

The generalised skew normal distribution (also called the Balakrishnan skew normal distribution [181]) generalises the skew normal distribution (which in turn generalises the Gaussian distribution). A new natural-numbered parameter  $n \in \{1, 2, \dots\}$  is introduced, and its probability density function is given by

$$f(x; n, \lambda) = \frac{\Phi(\lambda x)^n \phi(x)}{C(n, \lambda)} \quad (1.8.148)$$

where

$$C(n, \lambda) = \int_{-\infty}^{\infty} \Phi(\lambda x)^n \phi(x) dx \quad (1.8.149)$$

is the normalising constant. We can see that if  $n = 1$ , it collapses to the skew normal distribution, and if furthermore  $\lambda = 0$ , then it collapses to the standard Gaussian distribution. In the same way that multiplying the density by  $\Phi(\lambda x)$  can be thought of as skewing the distribution, then multiplying the density by  $\Phi(\lambda x)$  several times again skews the distribution even further.

### 1.8.5 Laplace Distribution

### 1.8.6 Cauchy Distribution

Also known as the Lorentz distribution, the Cauchy distribution with location parameter  $x_0$  and scale parameter  $\gamma$  has probability density function

$$f(x) = \frac{1}{\pi\gamma} \left[ \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \right] \quad (1.8.150)$$

The cumulative distribution function is given by

$$F(x) = \frac{1}{\pi\gamma} \int_{-\infty}^x \frac{1}{[(t - x_0)/\gamma]^2 + 1} dt \quad (1.8.151)$$

$$= \frac{1}{\pi} \int_{-\infty}^{\frac{x-x_0}{\gamma}} \frac{1}{u^2 + 1} du \quad (1.8.152)$$

where the change of variables  $u = \frac{t - x_0}{\gamma}$  was made. From the fact that the indefinite integral  $\int \frac{1}{u^2 + 1} du = \arctan u$  plus some constant, then

$$F(x) = \left[ \frac{\arctan u}{\pi} \right]_{-\infty}^{\frac{x-x_0}{\gamma}} \quad (1.8.153)$$

$$= \frac{1}{\pi} \arctan \left( \frac{x - x_0}{\gamma} \right) + \frac{1}{2} \quad (1.8.154)$$

because  $\lim_{u \rightarrow -\infty} \arctan u = -\frac{\pi}{2}$ .

### Mean of Cauchy Distribution

The Cauchy distribution is an example of a distribution which has undefined mean. This is because if  $X$  is Cauchy-distributed with  $x_0 = 0$  and  $\gamma = 1$  (for simplicity), then

$$\mathbb{E}[X] = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x}{x^2 + 1} dx \quad (1.8.155)$$

$$= \frac{1}{\pi} \int_{-\infty}^0 \frac{x}{x^2 + 1} dx + \frac{1}{\pi} \int_0^{\infty} \frac{x}{x^2 + 1} dx \quad (1.8.156)$$

$$= \frac{1}{2\pi} \int_{-\infty}^0 \frac{1}{u+1} du + \frac{1}{\pi} \int_0^{\infty} \frac{1}{u+1} du \quad (1.8.157)$$

$$= \left[ \frac{\ln(u+1)}{2\pi} \right]_0^\infty + \left[ \frac{\ln(u+1)}{2\pi} \right]_0^\infty \quad (1.8.158)$$

with the substitution  $u = x^2$ . Since the first term is  $-\infty$  and the second term is  $\infty$ , it follows that their sum is undefined. In other words, the function  $xf(x)$  is not absolutely integrable for the Cauchy distribution.

### Variance of Cauchy Distribution

As the mean is undefined, it follows that the variance will also be undefined since  $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ . The location parameter  $x_0$  determines the median and mode of the distribution, rather than the mean. The reason for the mean and variance being undefined can be thought of as being caused by the characterisation that the Cauchy distribution has very ‘fat’ or heavy tails.

### 1.8.7 Gamma Distribution

#### Gamma Function

The (complete) Gamma function is an interpolation of the factorials defined by

$$\Gamma(z) = \int_0^\infty y^{z-1} e^{-y} dy \quad (1.8.159)$$

which is continuous over the positive reals, and satisfies the recurrence relation

$$\Gamma(1) = 1 \quad (1.8.160)$$

$$\Gamma(z+1) = z\Gamma(z) \quad (1.8.161)$$

Thus, the relation between the Gamma function and the factorials (for non-negative integer values of  $z$ ) is

$$z! = \Gamma(z+1) \quad (1.8.162)$$

or alternatively,

$$\Gamma(z) = (z-1)! \quad (1.8.163)$$

#### Incomplete Gamma Functions

The incomplete Gamma functions are defined as integrals with the same integrand as the Gamma function, but with incomplete limits. The lower incomplete Gamma function (sometimes known as the incomplete Gamma function of the first kind) is defined as

$$\gamma(z, x) = \int_0^x y^{z-1} e^{-y} dy \quad (1.8.164)$$

while the upper incomplete Gamma function (sometimes known as the incomplete Gamma function of the second kind) is defined as

$$\Gamma(z, x) = \int_x^\infty y^{z-1} e^{-y} dy \quad (1.8.165)$$

where we can see that

$$\gamma(z, x) + \Gamma(z, x) = \Gamma(z) \quad (1.8.166)$$

#### Bayesian's Gamma Distribution

From the definitions of the incomplete and complete Gamma functions we can see that  $\frac{\gamma(z, x)}{\Gamma(z)}$  (sometimes known as the regularised Gamma function) is a valid cumulative distribution function in  $x$  defined over the positive reals. More generally, by introducing a parameter  $\beta > 0$  and another parameter  $\alpha = z > 0$ , we can define a cumulative distribution function

$$F(x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} \quad (1.8.167)$$

since  $\gamma(\alpha, \beta x) \leq \Gamma(\alpha)$  for all  $\beta > 0$  and  $x > 0$ . This is known as the (Bayesian's) Gamma distribution with shape parameter  $\alpha$  and rate parameter  $\beta$ . We can differentiate the cumulative distribution function using the Fundamental Theorem of Calculus to obtain the density function  $f(x)$ :

$$f(x) = \frac{dF(x)}{dx} \quad (1.8.168)$$

$$= \frac{1}{\Gamma(\alpha)} \frac{d\gamma(\alpha, \beta x)}{dx} \quad (1.8.169)$$

$$= \frac{1}{\Gamma(\alpha)} \frac{d}{dx} \int_0^{\beta x} y^{\alpha-1} e^{-y} dy \quad (1.8.170)$$

$$= \frac{1}{\Gamma(\alpha)} \cdot \frac{1}{\beta} (\beta x)^{\alpha-1} e^{-\beta x} \quad (1.8.171)$$

$$= \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad (1.8.172)$$

### Econometrician's Gamma Distribution

The ‘econometrician’s’ Gamma distribution is an alternative parametrisation with  $k = \alpha$  as the shape parameter and  $\theta = \frac{1}{\beta}$  as the scale parameter. The PDF may be written as

$$f(x) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)} \quad (1.8.173)$$

### Erlang Distribution

The Erlang distribution with shape parameter  $k$  and rate parameter  $\lambda$  has the probability density function

$$f(x; k, \lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \quad (1.8.174)$$

with  $\lambda > 0$  and  $k \in \{1, 2, \dots\}$ . The Erlang distribution can be characterised as the distribution of the sum of  $k$  i.i.d. exponential random variables with rate parameter  $\lambda$ . To show this, we can begin with the sum of 2 exponential random variables with rate parameter  $\lambda$ ,  $S_2 = X_1 + X_2$ . Note that we must take care in applying the convolution formula for computing the density of the sum of random variables, since exponential random variables have non-negative support. By applying the convolution formula, the density of  $S_2$  is computed via the integral:

$$f_{S_2}(x) = \int_0^x \lambda e^{-\lambda t} \lambda e^{-\lambda(x-t)} dt \quad (1.8.175)$$

with bounded terminals since neither  $t$  nor  $x - t$  can be negative. Computing this integral yields:

$$f_{S_2}(x) = \lambda^2 e^{-\lambda x} \int_0^x dt \quad (1.8.176)$$

$$= \lambda^2 x e^{-\lambda x} \quad (1.8.177)$$

which is the density of an Erlang random variable with shape  $k = 2$  and rate  $\lambda$ . We can generalise this to compute the density for arbitrary  $k$  through induction, by considering the sum between an Erlang random variable with shape  $k-1$  and an exponential random variable with the same rate parameter  $\lambda$ :

$$f(x; k, \lambda) = \int_0^x \frac{\lambda^{k-1} t^{k-2} e^{-\lambda t}}{(k-2)!} \lambda e^{-\lambda(x-t)} dt \quad (1.8.178)$$

$$= \frac{\lambda^k e^{-\lambda x}}{(k-2)!} \int_0^x t^{k-2} dt \quad (1.8.179)$$

$$= \frac{\lambda^k e^{-\lambda x}}{(k-2)!} \frac{x^{k-1}}{k-1} \quad (1.8.180)$$

$$= \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \quad (1.8.181)$$

Notice that the Erlang density can alternatively be written as

$$f(x; k, \lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)} \quad (1.8.182)$$

which is the Gamma density with shape parameter  $k$  and rate parameter  $\lambda$ . Hence the Erlang distribution can be treated as a special case of the Gamma distribution (where the shape parameter takes on natural numbers), and furthermore the exponential distribution can be treated as a special case of both the Erlang and Gamma distributions.

### 1.8.8 Beta Distribution

#### Beta Function

Also known as the Euler integral of the first kind, the Beta function is defined as

$$B(z, y) = \int_0^1 t^{z-1} (1-t)^{y-1} dt \quad (1.8.183)$$

for  $\text{Re}(y) > 0$  (the Beta function is defined on the complex numbers). The Beta function has a special relation with the Gamma function as follows:

$$B(z, y) = \frac{\Gamma(z)\Gamma(y)}{\Gamma(z+y)} \quad (1.8.184)$$

This can be shown by first writing out the definitions of the product of two Gamma functions:

$$\Gamma(z)\Gamma(y) = \left( \int_0^\infty u^{z-1} e^{-u} du \right) \left( \int_0^\infty v^{y-1} e^{-v} dv \right) \quad (1.8.185)$$

$$= \int_0^\infty \int_0^\infty u^{z-1} v^{y-1} e^{-u-v} du dv \quad (1.8.186)$$

Introduce the change in variables  $u = st$ ,  $v = s(1-t)$ , which when rearranged becomes  $t = \frac{u}{v+u}$  and  $s = v+u$ . Hence for all  $u \in (0, \infty)$ ,  $v \in (0, \infty)$ , we have  $t \in (0, 1)$ ,  $s \in (0, \infty)$ . Hence rewriting the integral gives

$$\Gamma(z)\Gamma(y) = \int_0^\infty \int_0^1 (st)^{z-1} [s(1-t)]^{y-1} e^{-s} \det(\mathbf{J}) dt ds \quad (1.8.187)$$

where the Jacobian determinant

$$\det(\mathbf{J}) = \det \begin{pmatrix} \frac{\partial}{\partial t} st & \frac{\partial}{\partial s} st \\ \frac{\partial}{\partial t} s(1-t) & \frac{\partial}{\partial s} s(1-t) \end{pmatrix} = s(1-t) + st = s \quad (1.8.188)$$

So

$$\Gamma(z)\Gamma(y) = \int_0^\infty \int_0^1 (st)^{z-1} [s(1-t)]^{y-1} e^{-s} s dt ds \quad (1.8.189)$$

$$= \int_0^\infty \int_0^1 (s^{z+y-1} e^{-s}) t^{z-1} (1-t)^{y-1} dt ds \quad (1.8.190)$$

$$= \left( \int_0^\infty (s^{z+y-1} e^{-s}) ds \right) \left( \int_0^1 t^{z-1} (1-t)^{y-1} dt \right) \quad (1.8.191)$$

$$= \Gamma(z + y) B(z, y) \quad (1.8.192)$$

as required. From this property, we can also deduce the Beta function is symmetric, i.e.  $B(z, y) = B(y, z)$ . The incomplete beta function is the Beta function with incomplete limits, i.e.

$$B(x; z, y) = \int_0^x t^{z-1} (1-t)^{y-1} dt \quad (1.8.193)$$

The regularised incomplete Beta function is defined as  $\frac{B(x; z, y)}{B(z, y)}$ .

### Beta Distribution of the First Kind

From the regularised incomplete Beta function we can see that it is a valid cumulative distribution function in  $x$ , defined on  $[0, 1]$ . Hence the Beta distribution (also known as the Beta distribution of the first kind) has cumulative distribution function

$$F(x) = \frac{B(x; \alpha, \beta)}{B(\alpha, \beta)} \quad (1.8.194)$$

where  $\alpha > 0$  and  $\beta > 0$  are shape parameters. Differentiating the cumulative distribution using the Fundamental Theorem of Calculus yields the probability density function  $f(x)$ :

$$f(x) = \frac{dF(x)}{dx} \quad (1.8.195)$$

$$= \frac{1}{B(\alpha, \beta)} \frac{d}{dx} B(x; \alpha, \beta) \quad (1.8.196)$$

$$= \frac{1}{B(\alpha, \beta)} \frac{d}{dx} \int_0^x t^{\alpha-1} (1-t)^{\beta-1} dt \quad (1.8.197)$$

$$= \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (1.8.198)$$

with an alternative form in terms of Gamma functions as

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (1.8.199)$$

If  $\alpha$  and  $\beta$  are positive integers, then the Beta distribution in terms of factorials is

$$f(x) = \frac{(\alpha + \beta - 1)!}{(\alpha - 1)! (\beta - 1)!} x^{\alpha-1} (1-x)^{\beta-1} \quad (1.8.200)$$

By letting  $n = \alpha + \beta - 2$  and  $r = \alpha - 1$ , an alternative parametrisation is

$$f(x) = \frac{(n+1)!}{r! (n-r)!} x^r (1-x)^{n-r} \quad (1.8.201)$$

### 1.8.9 Chi-Squared Distribution

The chi-squared distribution with  $k$  degrees of freedom is the distribution of the sum of squares of  $k$  standard normal random variables. Thus if  $Z_1, \dots, Z_k$  are standard normal random variables, then

$$X = Z_1^2 + \dots + Z_k^2 \quad (1.8.202)$$

is chi-squared distributed with  $k$  degrees of freedom. This is commonly denoted  $X \sim \chi_k^2$ . Let  $f(x)$  be the probability density function of  $X$  which has support  $[0, \infty)$ . We can see that the following two integrals should be the same (integrating to 1):

$$\int_0^\infty f(x) dx = \int_{x=0}^{x=\infty} \int_{z_1^2 + \dots + z_n^2=x} \mathcal{N}(z_1) \dots \mathcal{N}(z_n) dz_1 \dots dz_n \quad (1.8.203)$$

where  $\mathcal{N}(\cdot)$  is the standard normal PDF. Hence we equate

$$f(x) dx = \int_{z_1^2 + \dots + z_n^2 = x} \mathcal{N}(z_1) \dots \mathcal{N}(z_n) dz_1 \dots dz_n \quad (1.8.204)$$

Using the expressions for the PDFs:

$$f(x) dx = \int_{z_1^2 + \dots + z_n^2 = x} \frac{\exp[-(z_1^2 + \dots + z_n^2)/2]}{(2\pi)^{k/2}} dz_1 \dots dz_n \quad (1.8.205)$$

Since  $x = z_1^2 + \dots + z_n^2$  is constant with respect to the integral,

$$f(x) dx = \frac{e^{-x/2}}{(2\pi)^{k/2}} \int_{z_1^2 + \dots + z_n^2 = x} dz_1 \dots dz_n \quad (1.8.206)$$

Letting  $x = r^2$ , the integral  $\int_{z_1^2 + \dots + z_n^2 = r^2} dz_1 \dots dz_n$  is the volume of an infinitesimally thin shell of radius  $r$  with thickness  $dr$ . Since  $r = \sqrt{x}$ , then

$$\frac{dr}{dx} = \frac{1}{2\sqrt{x}} \quad (1.8.207)$$

$$dr = \frac{1}{2\sqrt{x}} dx \quad (1.8.208)$$

The surface area of a  $k$ -dimensional hypersphere with radius  $r$  is  $\frac{2r^{k-1}\pi^{k/2}}{\Gamma(k/2)}$  hence

$$f(x) dx = \frac{e^{-x/2}}{(2\pi)^{k/2}} \frac{2r^{k-1}\pi^{k/2}}{\Gamma(k/2)} \frac{1}{2\sqrt{x}} dx \quad (1.8.209)$$

Substituting  $r = \sqrt{x}$  and cancelling terms, this gives the expression for the PDF of the chi-squared distribution

$$f(x) = \frac{e^{-x/2} x^{k/2-1}}{2^{k/2} \Gamma(k/2)} \quad (1.8.210)$$

The CDF can be obtained by

$$F(x) = \frac{1}{\Gamma(k/2)} \int_0^x \frac{e^{-y/2} y^{k/2-1}}{2^{k/2}} dy \quad (1.8.211)$$

Making a change of variables  $w = y/2$  so  $dw = dy/2$ , we get

$$F(x) = \frac{1}{\Gamma(k/2)} \int_0^{x/2} \frac{e^{-w} (w/2)^{k/2-1}}{2^{k/2}} 2dw \quad (1.8.212)$$

$$= \frac{1}{\Gamma(k/2)} \int_0^{x/2} e^{-w} w^{k/2-1} dw \quad (1.8.213)$$

which is a regularised incomplete Gamma function with Gamma argument  $k/2$  and upper terminal  $x/2$ . Hence the chi-squared distribution with  $k$  degrees of freedom is a special case of the gamma distribution with shape parameter  $\alpha = k/2$  and rate parameter  $\beta = 1/2$  (or scale parameter  $\theta = 2$ ).

### Mean of Chi-Squared Distribution

The variance of a standard normal random variable implies from

$$\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \quad (1.8.214)$$

that  $\mathbb{E}[Z^2] = 1$ . Hence by the linearity of expectation, the expectation of  $X \sim \chi_k^2$  is

$$\mathbb{E}[X] = \mathbb{E}[Z_1^2 + \dots + Z_k^2] \quad (1.8.215)$$

$$= k \quad (1.8.216)$$

### Variance of Chi-Squared Distribution

We first compute the variance of a  $\chi_1^2$  random variable, which is identically distributed with  $Z^2$ , where  $Z \sim \mathcal{N}(0, 1)$ .

$$\text{Var}(Z^2) = \mathbb{E}[Z^4] - \mathbb{E}[Z^2]^2 \quad (1.8.217)$$

It is shown above that  $\mathbb{E}[Z^2] = 1$  and it can be shown from the excess kurtosis that  $\mathbb{E}[Z^4] = 3$ . Thus  $\text{Var}(Z^2) = 2$  and since the random variable  $X \sim \chi_k^2$  is the sum of  $k$  independent terms, then

$$\text{Var}(X) = \text{Var}(Z_1^2 + \dots + Z_k^2) \quad (1.8.218)$$

$$= \text{Var}(Z_1^2) + \dots + \text{Var}(Z_k^2) \quad (1.8.219)$$

$$= 2k \quad (1.8.220)$$

### Chi Distribution

The chi distribution is the distribution of the square root of the sum of squares of  $k$  standard normal random variables. The density function  $f_R(r)$  can be derived the same way as with the chi-squared distribution. As above, except without substituting  $r$  for  $x$ , we get

$$f_R(r) dr = \frac{e^{-r^2/2}}{(2\pi)^{k/2}} \frac{2r^{k-1}\pi^{k/2}}{\Gamma(k/2)} dr \quad (1.8.221)$$

Hence cancellation yields

$$f_R(r) = \frac{e^{-r^2/2} r^{k-1}}{2^{k/2-1} \Gamma(k/2)} \quad (1.8.222)$$

on support  $r \in [0, \infty)$ . The CDF of the chi-distribution can be obtained from

$$F_R(r) = \frac{1}{\Gamma(k/2)} \int_0^r \frac{e^{t^2/2} t^{k-1}}{2^{k/2-1}} dt \quad (1.8.223)$$

Introducing a change of variables  $s = \frac{t^2}{2}$  so that  $t = (2s)^{1/2}$  and  $ds = tdt$ , we have

$$F_R(r) = \frac{1}{\Gamma(k/2)} \int_0^r \frac{e^{-t^2/2} t^{k-1}}{2^{k/2-1}} dt \quad (1.8.224)$$

$$= \frac{1}{\Gamma(k/2)} \int_0^{r^2/2} \frac{e^{-s} s^{k-1}}{2^{k/2-1}} \frac{ds}{t} \quad (1.8.225)$$

$$= \frac{1}{\Gamma(k/2)} \int_0^{r^2/2} \frac{e^{-s} s^{k-2}}{2^{k/2-1}} ds \quad (1.8.226)$$

$$= \frac{1}{\Gamma(k/2)} \int_0^{r^2/2} \frac{e^{-s} (2s)^{(k-2)/2}}{2^{k/2-1}} ds \quad (1.8.227)$$

$$= \frac{1}{\Gamma(k/2)} \int_0^{r^2/2} e^{-s} s^{k/2-1} ds \quad (1.8.228)$$

which is a regularised incomplete Gamma function with Gamma argument  $k/2$  and upper terminal  $r^2/2$ .

## Rayleigh Distribution

The Rayleigh distribution is the chi-squared distribution with 2 degrees of freedom (i.e. Euclidean norm of two uncorrelated Gaussian random variables). Its PDF is given by

$$f_R(r) = \frac{e^{-r^2/2} r^{k-1}}{2^{k/2-1} \Gamma(k/2)} \Big|_{k=2} \quad (1.8.229)$$

$$= r e^{r^2/2} \quad (1.8.230)$$

A scale parameter  $\sigma > 0$  can be introduced to represent the standard deviation of the two Gaussian random variables.

$$f_R(r) = \frac{r}{\sigma^2} e^{r^2/2\sigma^2} \quad (1.8.231)$$

Integrating this, we find that the CDF has expression

$$F_R(r) = 1 - e^{r^2/2\sigma^2} \quad (1.8.232)$$

Note that the the PDF of the Rayleigh distribution with  $\sigma = 1$  is the vertical reflection (and also the horizontal reflection) of the derivative of the standard Gaussian PDF. That is,

$$f_R(r)|_{\sigma=1} = - \frac{d\mathcal{N}(z; 0, 1^2)}{dz} \Big|_{z=r} \quad (1.8.233)$$

$$= \frac{d\mathcal{N}(z; 0, 1^2)}{dz} \Big|_{z=-r} \quad (1.8.234)$$

### 1.8.10 Student's $t$ Distribution

A Student's  $t$  distribution (or simply  $t$  distribution) with  $k = n - 1$  degrees of freedom can be defined as the distribution of the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (1.8.235)$$

where  $\bar{X}$  is the sample mean from i.i.d.  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$  and

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad (1.8.236)$$

is the sample standard deviation. To derive the density of the  $t$  distribution, we first show how the  $t$  distribution relates to the chi-squared and the normal distribution. The sum of squares of the standardised random variables is given by

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \quad (1.8.237)$$

We can make the following rearrangement:

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \quad (1.8.238)$$

$$= \sum_{i=1}^n (X_i - \bar{X})^2 + 2 \sum_{i=1}^n (X_i - \bar{X})(\bar{X} - \mu) + \sum_{i=1}^n (\bar{X} - \mu)^2 \quad (1.8.239)$$

$$= \sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) + \sum_{i=1}^n (\bar{X} - \mu)^2 \quad (1.8.240)$$

$$= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \quad (1.8.241)$$

Hence

$$\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 \quad (1.8.242)$$

$$= \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \left( \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \right)^2 \quad (1.8.243)$$

We reason that since the chi-squared distribution is the sum i.i.d. standard normal random variables, therefore if  $Q_1 \sim \chi_{n_1}^2$  and  $Q_2 \sim \chi_{n_2}^2$ , then  $Q_3 = Q_1 + Q_2 \sim \chi_{n_1+n_2}^2$ . Thus if given some  $Q_3$  and  $Q_2$ , we deduce that there exists a random variable  $Q_1 \sim \chi_{n_1}^2$  that is independent of  $Q_2$ . Notice that the left hand side of the expression above means  $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$ . Further note that  $\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ . Then the term  $\sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2$  should be chi-squared distributed with  $k = n - 1$  degrees of freedom. Define the random variable

$$V = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \quad (1.8.244)$$

and notice that from the definition of  $S$ :

$$V = k \frac{S^2}{\sigma^2} \quad (1.8.245)$$

Hence rearranging  $T$  yields

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (1.8.246)$$

$$= \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma \sqrt{V/k}} \quad (1.8.247)$$

$$= \frac{Z}{\sqrt{V/k}} \quad (1.8.248)$$

We derive the  $t$  distribution by deriving the distribution of  $T = \frac{Z}{\sqrt{V/k}}$  where  $Z \sim \mathcal{N}(0, 1)$  and  $V \sim \chi_k^2$  independent of  $Z$ . In following the same approach taken to derive the ratio distribution, the density of  $T$  can be expressed as

$$f_T(t) = \int \int_{\{z, v: z/\sqrt{v/k} = t\}} f_Z(z) f_V(v) dz dv \quad (1.8.249)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_Z(z) f_V(v) \delta \left( t - \frac{z}{\sqrt{v/k}} \right) dz dv \quad (1.8.250)$$

where  $\delta(\cdot)$  is the Dirac delta function and the densities  $f_Z(z)$  and  $f_V(v)$  are respectively

$$f_Z(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \quad (1.8.251)$$

$$f_V(v) = \frac{e^{-v/2} v^{k/2-1}}{2^{k/2} \Gamma(k/2)} \quad (1.8.252)$$

The support of  $f_V(v)$  is  $[0, \infty)$ , so

$$f_T(t) = \int_0^\infty \int_{-\infty}^\infty f_Z(z) f_V(v) \delta\left(t - \frac{z}{\sqrt{v/k}}\right) dz dv \quad (1.8.253)$$

Make the substitution  $y = \frac{z}{\sqrt{v/k}}$ , i.e.  $z = y\sqrt{v/k}$  and  $dz = \sqrt{v/k}dy$ .

$$f_T(t) = \int_0^\infty \int_{-\infty}^\infty f_Z\left(y\sqrt{\frac{v}{k}}\right) f_V(v) \delta(t - y) \sqrt{\frac{v}{k}} dy dv \quad (1.8.254)$$

$$= \int_0^\infty \sqrt{\frac{v}{k}} f_V(v) \int_{-\infty}^\infty f_Z\left(y\sqrt{\frac{v}{k}}\right) \delta(t - y) dy dv \quad (1.8.255)$$

Evaluating the inner integral gives

$$\int_{-\infty}^\infty f_Z\left(y\sqrt{\frac{v}{k}}\right) \delta(t - y) dy = f_Z\left(t\sqrt{\frac{v}{k}}\right) \quad (1.8.256)$$

Hence

$$f_T(t) = \int_0^\infty \sqrt{\frac{v}{k}} f_V(v) f_Z\left(t\sqrt{\frac{v}{k}}\right) dv \quad (1.8.257)$$

Note that  $f_Z\left(t\sqrt{\frac{v}{k}}\right) = \frac{\exp\left(-\frac{t^2 v}{2k}\right)}{\sqrt{2\pi}}$ . Substituting the expressions for the densities,

$$f_T(t) = \frac{1}{\sqrt{k}} \frac{1}{\sqrt{2\pi}} \frac{1}{2^{k/2} \Gamma(k/2)} \int_0^\infty \sqrt{v} \cdot v^{k/2-1} \cdot e^{-v/2} \cdot \exp\left(-\frac{t^2 v}{2k}\right) dv \quad (1.8.258)$$

$$= \frac{1}{\sqrt{2\pi k} 2^{k/2} \Gamma(k/2)} \int_0^\infty v^{\frac{k-1}{2}} \cdot \exp\left[-\left(1 + \frac{t^2}{k}\right) \frac{v}{2}\right] dv \quad (1.8.259)$$

Make the substitution  $u = \left(1 + \frac{t^2}{k}\right) \frac{v}{2}$  so  $dv = 2\left(1 + \frac{t^2}{k}\right)^{-1} du$ :

$$f_T(t) = \frac{1}{\sqrt{\pi k} \Gamma(k/2)} \frac{1}{2^{\frac{k+1}{2}}} \int_0^\infty \left[2u\left(1 + \frac{t^2}{k}\right)^{-1}\right]^{\frac{k-1}{2}} \cdot e^{-u} \cdot 2\left(1 + \frac{t^2}{k}\right)^{-1} du \quad (1.8.260)$$

$$= \frac{1}{\sqrt{\pi k} \Gamma(k/2)} \frac{2^{\frac{k-1}{2}} \cdot 2}{2^{\frac{k+1}{2}}} \left(1 + \frac{t^2}{k}\right)^{\frac{-k+1}{2}-1} \int_0^\infty u^{\frac{k+1}{2}-1} \cdot e^{-u} du \quad (1.8.261)$$

Notice the integral is the Gamma function of  $\frac{k+1}{2}$  so

$$f_T(t) = \frac{\left(1 + t^2/k\right)^{-\frac{(k+1)}{2}}}{\sqrt{\pi k} \Gamma(k/2)} \Gamma\left(\frac{k+1}{2}\right) \quad (1.8.262)$$

Applying the property of the Beta function,  $B\left(\frac{1}{2}, \frac{k}{2}\right) = \frac{\Gamma(1/2)\Gamma(k/2)}{\Gamma(\frac{k+1}{2})}$ . The term  $\Gamma(1/2)$  evaluates to  $\sqrt{\pi}$ , so this means the density of the  $t$ -distribution can be alternatively expressed as

$$f_T(t) = \frac{1}{\sqrt{k} B(1/2, k/2)} \left(1 + \frac{t^2}{k}\right)^{-\frac{(k+1)}{2}} \quad (1.8.263)$$

### 1.8.11 F-Distribution

Suppose  $U$  and  $V$  are independent chi-squared random variables, with degrees of freedom  $n$  and  $m$  respectively. The  $F$ -distribution (also known as Snedecor's  $F$  distribution) can be defined as the distribution of the random variable

$$X = \frac{U/m}{V/n} \quad (1.8.264)$$

Alternatively, we may characterise the  $F$ -distribution as follows. Recall as shown in the derivation of the  $t$  distribution that for a  $m+1$  sample of normal random variables  $Y_1, \dots, Y_{m+1}$  with variance  $\sigma^2$ , the standardised sample variance follows a chi-squared distribution with  $m$  degrees of freedom:

$$\frac{1}{\sigma^2} \sum_{i=1}^{m+1} (Y_i - \bar{Y})^2 \sim \chi_m^2 \quad (1.8.265)$$

So letting  $\frac{1}{\sigma^2} \sum_{i=1}^{m+1} (Y_i - \bar{Y})^2 = U$ , then

$$\frac{U}{m} = \frac{1}{\sigma^2} \cdot \frac{1}{m} \sum_{i=1}^{m+1} (Y_i - \bar{Y})^2 \quad (1.8.266)$$

$$= \frac{S_Y^2}{\sigma^2} \quad (1.8.267)$$

where  $S_Y^2$  is the sample variance. Likewise, if we have an  $n+1$  sample of normal random variables  $W_1, \dots, W_{n+1}$  with variance  $\varsigma^2$  and sample variance  $S_W^2$ , then

$$\frac{V}{n} = \frac{S_W^2}{\varsigma^2} \quad (1.8.268)$$

Therefore we may characterise the  $F$ -distribution as the distribution of

$$X = \frac{S_Y^2/\sigma^2}{S_W^2/\varsigma^2} \quad (1.8.269)$$

which is the standardised ratio of sample variances from two independent normal samples of different sizes. To obtain the probability density function, we use the expression for the ratio distribution applied to independent chi-squared random variables. For simplicity, we first find the distribution of  $X' = \frac{U}{V}$  (i.e. suppose that  $m = 1, n = 1$ ). Using the ratio distribution, the density of  $X'$  in terms of the densities  $f_U(u)$  and  $f_V(v)$  for  $U$  and  $V$  respectively is

$$f_{X'}(x) = \int_{-\infty}^{\infty} f_U(uv) f_V(v) |v| dv \quad (1.8.270)$$

$$= \int_0^{\infty} f_U(uv) f_V(v) v dv \quad (1.8.271)$$

since chi-squared random variables are always non-negative. Applying the expression for the chi-squared distribution, we get

$$f_{X'}(x) = \int_0^{\infty} \frac{e^{-xv/2} (xv)^{m/2-1}}{2^{m/2} \Gamma(m/2)} \cdot \frac{-e^{v/2} v^{n/2-1}}{2^{n/2} \Gamma(n/2)} v dv \quad (1.8.272)$$

$$= \frac{x^{m/2-1}}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} \int_0^{\infty} v^{(m+n)/2-2} v e^{-v(x+1)/2} dv \quad (1.8.273)$$

$$= \frac{x^{m/2-1}}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} \int_0^{\infty} v^{(m+n)/2-1} e^{-v(x+1)/2} dv \quad (1.8.274)$$

Make the substitution  $t = v(x+1)/2$  so that  $v = t \left(\frac{x+1}{2}\right)^{-1}$  and  $dv = \left(\frac{x+1}{2}\right)^{-1} dt$ :

$$f_{X'}(x) = \frac{x^{m/2-1}}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} \int_0^\infty t^{(m+n)/2-1} \left(\frac{x+1}{2}\right)^{-(m+n)/2+1} e^{-t} \left(\frac{x+1}{2}\right)^{-1} dt \quad (1.8.275)$$

$$= \frac{x^{m/2-1}}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} \left(\frac{x+1}{2}\right)^{-(m+n)/2} \int_0^\infty t^{(m+n)/2-1} e^{-t} dt \quad (1.8.276)$$

Recognise that the integral becomes a Gamma integral, so using the definition of the Gamma function,

$$f_{X'}(x) = \frac{x^{m/2-1} \Gamma((m+n)/2)}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} \left(\frac{x+1}{2}\right)^{-(m+n)/2} \quad (1.8.277)$$

$$= \frac{x^{m/2-1} \Gamma((m+n)/2)}{\Gamma(m/2) \Gamma(n/2) (x+1)^{(m+n)/2}} \quad (1.8.278)$$

Since  $X = X' \frac{m}{n}$  then by the scaling properties of the density function

$$f_X(x) = \frac{m}{n} f_{X'}\left(\frac{m}{n}x\right) \quad (1.8.279)$$

$$= \frac{m}{n} \cdot \frac{(xm/n)^{m/2-1} \Gamma((m+n)/2)}{\Gamma(m/2) \Gamma(n/2) (xm/n + 1)^{(m+n)/2}} \quad (1.8.280)$$

$$= \frac{x^{m/2-1} (m/n)^{m/2} \Gamma((m+n)/2)}{\Gamma(m/2) \Gamma(n/2) (xm/n + 1)^{(m+n)/2}} \quad (1.8.281)$$

Using the definition of the Beta function, we can write an alternative expression for this density parametrised in two degrees of freedom  $m$  and  $n$ :

$$f_X(x) = \frac{x^{m/2-1} (m/n)^{m/2}}{B(m/2, n/2) (xm/n + 1)^{(m+n)/2}} \quad (1.8.282)$$

An expression can be derived for the cumulative distribution function,  $F_X(x)$ . Firstly,

$$F_X(x) = \int_{-\infty}^x f_X(r) dr \quad (1.8.283)$$

$$= \int_0^x \frac{r^{m/2-1} (m/n)^{m/2}}{B(m/2, n/2) (rm/n + 1)^{(m+n)/2}} dr \quad (1.8.284)$$

Make the change of variables  $s = \frac{mr}{mr+n}$  so that rearranging gives  $r = \frac{n}{m} \cdot \frac{s}{1-s}$  and by the quotient rule  $\frac{dr}{ds} = \frac{n}{m} \cdot \frac{1}{(1-s)^2}$ . Then

$$F_X(x) = \frac{1}{B(m/2, n/2)} \int_0^{\frac{mx}{mx+n}} \frac{s^{m/2-1} (1-s)^{-m/2+1} (n/m)^{m/2-1} (m/n)^{m/2}}{\left(\frac{s+1-s}{1-s}\right)^{(m+n)/2}} \cdot \frac{n}{m} \cdot \frac{ds}{(1-s)^2} \quad (1.8.285)$$

$$= \frac{1}{B(m/2, n/2)} \int_0^{\frac{mx}{mx+n}} \frac{s^{m/2-1} (1-s)^{-m/2+1}}{(1-s)^{-(m+n)/2+2}} ds \quad (1.8.286)$$

$$= \frac{1}{B(m/2, n/2)} \int_0^{\frac{mx}{mx+n}} s^{m/2-1} (1-s)^{n/2-1} ds \quad (1.8.287)$$

$$= \frac{B\left(\frac{mx}{mx+n}; m/2, n/2\right)}{B(m/2, n/2)} \quad (1.8.288)$$

which is a regularised incomplete Beta function.

### 1.8.12 Pareto Distribution

#### Lomax Distribution

### 1.8.13 Gumbel Distribution

The Gumbel distribution has support  $\mathbb{R}$ , scale parameter  $\beta > 0$  and location parameter  $\mu$ , with cumulative distribution function

$$F(x) = \exp\left(-e^{-\frac{x-\mu}{\beta}}\right) \quad (1.8.289)$$

If can be verified that  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ . The probability density function is found by the chain rule:

$$f(x) = \frac{dF(x)}{dx} \quad (1.8.290)$$

$$= \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} \exp\left(-e^{-\frac{x-\mu}{\beta}}\right) \quad (1.8.291)$$

$$= \frac{1}{\beta} \exp\left[-\left(\frac{x-\mu}{\beta} + e^{-\frac{x-\mu}{\beta}}\right)\right] \quad (1.8.292)$$

The mode of the distribution can be found by taking another derivative.

$$f'(x) = \frac{1}{\beta} \left[ -\left(\frac{1}{\beta} - \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}}\right) \exp\left[-\left(\frac{x-\mu}{\beta} + e^{-\frac{x-\mu}{\beta}}\right)\right] \right] \quad (1.8.293)$$

$$= -\left(\frac{1}{\beta} - \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}}\right) f(x) \quad (1.8.294)$$

The mode  $\check{x}$  is found at  $f(\check{x}) = 0$  which gives

$$\frac{1}{\beta} = \frac{1}{\beta} e^{-\frac{\check{x}-\mu}{\beta}} \quad (1.8.295)$$

Hence the mode  $\check{x} = \mu$ .

### 1.8.14 Fréchet Distribution

The Fréchet distribution has cumulative distribution function

$$F(x) = \Pr(X \leq x) \quad (1.8.296)$$

$$= \exp\left[-\left(\frac{x-m}{s}\right)^{-\alpha}\right] \quad (1.8.297)$$

$$= \exp\left[-\left(\frac{s}{x-m}\right)^\alpha\right] \quad (1.8.298)$$

on support  $x > m$ , with parameters  $\alpha > 0$ ,  $s > 0$  and  $m \in \mathbb{R}$ . It can be verified that this is a valid distribution since  $F(x) \rightarrow 0$  as  $x \rightarrow m$  (the exponent tends to  $-\infty$ ) and  $F(x) \rightarrow 1$  as  $x \rightarrow \infty$  (the exponent tends to 0). Differentiating, the probability density function is

$$f(x) = -(-\alpha) \frac{x}{s} \left( \frac{x-m}{s} \right)^{-\alpha-1} \exp \left[ - \left( \frac{x-m}{s} \right)^{-\alpha} \right] \quad (1.8.299)$$

$$= \alpha \frac{x}{s} \left( \frac{x-m}{s} \right)^{-\alpha-1} \exp \left[ - \left( \frac{x-m}{s} \right)^{-\alpha} \right] \quad (1.8.300)$$

### 1.8.15 Weibull Distribution

The Weibull distribution has cumulative distribution function

$$F(x) = \Pr(X \leq x) \quad (1.8.301)$$

$$= 1 - \exp \left[ - \left( \frac{x}{\lambda} \right)^k \right] \quad (1.8.302)$$

on support  $x \geq 0$  with parameters  $k > 0$  and  $\lambda > 0$ . It can be verified that this is a valid distribution since  $F(0) = 0$  and  $F(x) \rightarrow 1$  as  $x \rightarrow \infty$  (the exponent tends to  $-\infty$ ). Differentiating, the probability density function is

$$f(x) = - \left( -\frac{k}{\lambda} \right) \left( \frac{x}{\lambda} \right)^{k-1} \exp \left[ - \left( \frac{x}{\lambda} \right)^k \right] \quad (1.8.303)$$

$$= \frac{k}{\lambda} \left( \frac{x}{\lambda} \right)^{k-1} \exp \left[ - \left( \frac{x}{\lambda} \right)^k \right] \quad (1.8.304)$$

### Reversed Weibull Distribution

If  $g(x)$  has a Weibull distribution, then  $f(x) = g(-x)$  has a distribution known as a reversed Weibull distribution. The cumulative distribution of a ‘standard’ reversed Weibull distribution with parameter  $\lambda = 1$  is

$$F(x) = \begin{cases} \exp \left[ -(-x)^k \right], & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (1.8.305)$$

which by differentiating, we obtain

$$f(x) = \begin{cases} kx^{k-1} \exp \left[ -(-x)^k \right], & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (1.8.306)$$

### 1.8.16 Logistic Distribution

#### Fisk Distribution

### 1.8.17 Cantor Distribution

## 1.9 Families of Discrete Univariate Probability Distributions

#### 1.9.1 Kronecker Delta Distribution

The *unit pulse function*  $\delta_0(i)$  is a function over the integers, such that

$$\delta_0(i) = \begin{cases} 1, & i = 0 \\ 0, & i \neq 0 \end{cases} \quad (1.9.1)$$

This is the probability mass function representation of the Dirac delta distribution, for a degenerate random variable with  $\Pr(X = 0) = 1$ . The Kronecker delta is a generalisation that allows one to shift the unit pulse function. We write

$$\delta_j(i) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad (1.9.2)$$

Note also that

$$\delta_j(i) = \delta_0(i - j) \quad (1.9.3)$$

### 1.9.2 Discrete Uniform Distribution

### 1.9.3 Bernoulli Distribution

### 1.9.4 Rademacher Distribution

### 1.9.5 Binomial Distribution

The binomial distribution is the distribution for the number of successes from  $n$  independent trials, with a success probability of  $p$ . Let this random variable be denoted  $X$ , which will be supported on  $\{0, \dots, n\}$ . We can derive the probability mass function of  $X$  as follows. If we visualise the binary probability tree resulting from  $n$  sequential biased (with probability  $p$ ) ‘coin flips’, there will be  $2^n$  nodes in the final level of the tree. If we consider all the nodes corresponding to there being  $x$  successes, it follows from the binomial coefficient that there will be  $\binom{n}{x}$  such nodes. Moreover, any particular sequence of trials with  $x$  success occurs with probability  $p^x(1-p)^{n-x}$ . Thus

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (1.9.4)$$

and we say that  $X \sim \text{Binom}(n, p)$ . Another way to arrive at the binomial distribution is by using the binomial theorem, which is also a way to verify that the binomial distribution is a valid probability mass function.

$$1 = p + (1-p) \quad (1.9.5)$$

$$= [p + (1-p)]^n \quad (1.9.6)$$

$$= \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} \quad (1.9.7)$$

With  $n = 1$ , the binomial distribution reduces to the Bernoulli distribution. A binomial random variable can in fact be thought of as a sum of  $n$  independent Bernoulli random variables  $\mathbb{I}_1, \dots, \mathbb{I}_n$ :

$$X = \sum_{i=1}^n \mathbb{I}_i \quad (1.9.8)$$

This way, the mean of the binomial distribution can be found as

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[\mathbb{I}_i] \quad (1.9.9)$$

$$= np \quad (1.9.10)$$

and the variance is also found using the variance of sums as

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(\mathbb{I}_i) \quad (1.9.11)$$

$$= np(1-p) \quad (1.9.12)$$

## Cumulative Binomial Distribution

The cumulative binomial distribution can be written as a sum:

$$F(x) = \Pr(X \leq x) \quad (1.9.13)$$

$$= \sum_{i=0}^x \Pr(X = i) \quad (1.9.14)$$

$$= \sum_{i=0}^x f(i) \quad (1.9.15)$$

$$= \sum_{i=0}^x \binom{n}{i} p^i (1-p)^{n-i} \quad (1.9.16)$$

An alternate form of the expression for the cumulative binomial distribution is the regularised incomplete Beta function:

$$F(k) = \frac{B(1-p; n-k, k+1)}{B(n-k, k+1)} \quad (1.9.17)$$

$$= \frac{1}{B(n-k, k+1)} \int_0^{1-p} t^{n-k-1} (1-t)^k dt \quad (1.9.18)$$

$$= \frac{\Gamma(n+1)}{\Gamma(n-k)\Gamma(k+1)} \int_0^{1-p} t^{n-k-1} (1-t)^k dt \quad (1.9.19)$$

$$= \frac{n!}{(n-k-1)!k!} \int_0^{1-p} t^{n-k-1} (1-t)^k dt \quad (1.9.20)$$

$$= (n-k) \binom{n}{k} \int_0^{1-p} t^{n-k-1} (1-t)^k dt \quad (1.9.21)$$

This can be shown by repeated application of integration by parts on the Beta integral. Let  $u = (1-t)^k$  and  $v' = t^{n-k-1}$ , then  $u' = -k(1-t)^{k-1}$  and  $v = \frac{1}{n-k}t^{n-k}$ . So

$$\int_0^{1-p} t^{n-k-1} (1-t)^k dt = [uv]_0^{1-p} - \int_0^{1-p} u'v dt \quad (1.9.22)$$

$$= \frac{1}{n-k} \left[ (1-t)^k t^{n-k} \right]_{t=0}^{t=1-p} + \int_0^{1-p} \frac{k}{n-k} t^{n-k} (1-t)^{k-1} dt \quad (1.9.23)$$

$$= \frac{1}{n-k} p^k (1-p)^{n-k} + \frac{k}{n-k} \int_0^{1-p} (1-t)^{k-1} t^{n-k} dt \quad (1.9.24)$$

Repeating this again on the resultant Beta integral, we will have

$$\begin{aligned} \int_0^{1-p} t^{n-k-1} (1-t)^k dt &= \frac{1}{n-k} p^k (1-p)^{n-k} + \frac{k}{(n-k)(n-k+1)} p^{k-1} (1-p)^{n-k+1} \\ &\quad + \frac{k(k-1)}{(n-k)(n-k+1)} \int_0^{1-p} (1-t)^{k-2} t^{n-k+1} dt \end{aligned} \quad (1.9.25)$$

and we will eventually end up with the series

$$\begin{aligned} \int_0^{1-p} t^{n-k-1} (1-t)^k dt &= \frac{1}{n-k} p^k (1-p)^{n-k} + \frac{k}{(n-k)(n-k+1)} p^{k-1} (1-p)^{n-k+1} \\ &\quad + \frac{k(k-1)}{(n-k)(n-k+1)(n-k+2)} p^{k-2} (1-p)^{n-k+2} + \dots + \frac{k \times \dots \times 2}{(n-k) \times \dots \times (n-1)} p^1 (1-p)^{n-1} \end{aligned}$$

$$+ \frac{k!}{(n-k) \times \cdots \times (n-1)} \int_0^{1-p} (1-t)^0 t^{n-1} dt \quad (1.9.26)$$

with  $\int_0^{1-p} (1-t)^0 t^{n-1} dt = \frac{1}{n} (1-p)^n$ . Note then that this series can be represented compactly as

$$\int_0^{1-p} t^{n-k-1} (1-t)^k dt = \sum_{j=0}^k \frac{k!}{(k-j)!} \frac{(n-k)!}{(n-k)(n-k+j)!} p^{k-j} (1-p)^{n-k+j} \quad (1.9.27)$$

Via cancellations, we see that

$$(n-k) \binom{n}{k} \frac{k!}{(k-j)!} \frac{(n-k)!}{(n-k)(n-k+j)!} = \frac{n!}{(k-j)!(n-k+j)!} \quad (1.9.28)$$

$$= \binom{n}{k-j} \quad (1.9.29)$$

Hence

$$(n-k) \binom{n}{k} \int_0^{1-p} t^{n-k-1} (1-t)^k dt = \sum_{j=0}^k \binom{n}{k-j} p^{k-j} (1-p)^{n-k+j} \quad (1.9.30)$$

Then replace  $k - j$  with  $i$  in the sum to obtain the expression for the cumulative binomial distribution:

$$F(k) = (n-k) \binom{n}{k} \int_0^{1-p} t^{n-k-1} (1-t)^k dt \quad (1.9.31)$$

$$= \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \quad (1.9.32)$$

## Gaussian Approximation to Binomial Distribution

### Continuity Corrections

#### 1.9.6 Categorical Distribution

A categorical distribution is a generic distribution for a discrete, finite set of  $K$  categories which are indexed by  $\{1, \dots, K\}$ . For instance, each category could represent a partition of mutually exclusive events resulting from a random experiment. The categorical distribution can be parametrised by  $K$  parameters, which are the probabilities of occurrence for each category:  $p_1, \dots, p_K$ , with the restriction that each  $p_i > 0$  and the sum of probabilities equals one:

$$\sum_{i=1}^K p_i = 1 \quad (1.9.33)$$

Alternatively, the categorical distribution can be minimally parametrised by  $K - 1$  parameters:  $p_1, \dots, p_{K-1}$ , with the restriction that each  $p_i > 0$  before, and now the sum of probabilities is less than one:

$$\sum_{i=1}^{K-1} p_i < 1 \quad (1.9.34)$$

In this way,  $p_K$  would be automatically defined by

$$p_K = 1 - \sum_{i=1}^{K-1} p_i \quad (1.9.35)$$

The probability mass function for a categorical random variable  $X$  is simply:

$$\Pr(X = i) = p_i \quad (1.9.36)$$

The categorical distribution generalises the Bernoulli distribution (under a shift in support).

### 1.9.7 Poisson Distribution

The Poisson distribution is supported on the non-negative integers  $\{0, 1, 2, \dots\}$  with rate parameter  $\lambda$ , and has the probability mass function

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (1.9.37)$$

A Poisson random variable can be characterised as the distribution of counts for a particular recurring event within a fixed interval of time, where the probability of occurrence at any point in time is small but independent with all other points in time. Formally, the Poisson distribution can be derived as a limiting case of the Binomial distribution. Let  $X_n \sim \text{Binom}(n, p_n)$  where

$$p_n = \frac{\lambda}{n} \quad (1.9.38)$$

Then the mean is  $\mathbb{E}[X] = np_n = \lambda$ . Now consider the limit as  $n \rightarrow \infty$ , which yields the interpretation that we have infinitely many but very small chances at success within some fixed time frame. From the binomial distribution,

$$\Pr(X_n = k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k} \quad (1.9.39)$$

$$= \frac{n!}{k!(n-k)!} \left( \frac{p_n}{1-p_n} \right)^k (1-p_n)^n \quad (1.9.40)$$

$$= \frac{n!}{k!(n-k)!} \left( \frac{\lambda/n}{1-\lambda/n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^n \quad (1.9.41)$$

$$= \frac{n!}{k!(n-k)!} \left( \frac{\lambda}{n-\lambda} \right)^k \left( 1 - \frac{\lambda}{n} \right)^n \quad (1.9.42)$$

$$= \frac{\lambda^k}{k!} \cdot \frac{n!}{(n-k)!(n-\lambda)^k} \left( 1 - \frac{\lambda}{n} \right)^n \quad (1.9.43)$$

Taking the limit,

$$\lim_{n \rightarrow \infty} \Pr(X_n = k) = \frac{\lambda^k}{k!} \lim_{n \rightarrow \infty} \left[ \frac{n!}{(n-k)!(n-\lambda)^k} \left( 1 - \frac{\lambda}{n} \right)^n \right] \quad (1.9.44)$$

$$= \frac{\lambda^k}{k!} \lim_{n \rightarrow \infty} \left[ \frac{n!}{(n-k)!(n-\lambda)^k} \right] \lim_{n \rightarrow \infty} \left[ \left( 1 - \frac{\lambda}{n} \right)^n \right] \quad (1.9.45)$$

Recognising that the middle first limit can be simplified into  $k$  terms in both the numerator and denominator, we evaluate

$$\lim_{n \rightarrow \infty} \left[ \frac{n!}{(n-k)!(n-\lambda)^k} \right] = \lim_{n \rightarrow \infty} \left( \frac{n}{n-\lambda} \times \cdots \times \frac{n-k+1}{n-\lambda} \right) \quad (1.9.46)$$

$$= 1 \quad (1.9.47)$$

Moreover, as one of the definitions of the exponential,  $\lim_{n \rightarrow \infty} \left[ \left(1 - \frac{\lambda}{n}\right)^n \right] = e^{-\lambda}$  so

$$\lim_{n \rightarrow \infty} \Pr(X_n = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (1.9.48)$$

which is the probability mass function of the Poisson distribution with rate  $\lambda$ . Through this characterisation, we can also see that the mean of the Poisson distribution is

$$\mathbb{E}[X] = \lambda \quad (1.9.49)$$

Also, the variance of the Poisson distribution can be found by taking the limit of the binomial variance:

$$\text{Var}(X) = \lim_{n \rightarrow \infty} np_n (1 - p_n) \quad (1.9.50)$$

$$= \lim_{n \rightarrow \infty} \lambda \left(1 - \frac{\lambda}{n}\right) \quad (1.9.51)$$

$$= \lambda \quad (1.9.52)$$

Hence the Poisson distribution has mean equal to variance.

### 1.9.8 Skellam Distribution

### 1.9.9 Geometric Distribution

The geometric distribution is distribution of the number of independent trials required until the first success, with success probability  $p$ . Let  $X$  be the random variable for this number. Then in order for the first success to occur on the  $k^{\text{th}}$  trial, the first  $k - 1$  trials must be failures and the last one must be a success, meaning

$$\Pr(X = k) = (1 - p)^{k-1} p \quad (1.9.53)$$

on support  $\{1, 2, \dots\}$ . The CDF can be reasoned as

$$\Pr(X \leq k) = 1 - (1 - p)^k \quad (1.9.54)$$

because the event  $X \leq k$  is complementary to there being  $k$  failures in a row. The expected number of trials until the first success is reciprocal to the success probability:

$$\mathbb{E}[X] = \frac{1}{p} \quad (1.9.55)$$

To show this,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} p (1 - p)^{k-1} k \quad (1.9.56)$$

$$= p \sum_{k=1}^{\infty} (1 - p)^{k-1} k \quad (1.9.57)$$

$$= -p \sum_{k=1}^{\infty} \frac{d}{dp} (1 - p)^k \quad (1.9.58)$$

$$= -p \cdot \frac{d}{dp} \left( \sum_{k=1}^{\infty} (1 - p)^k \right) \quad (1.9.59)$$

since  $\frac{d}{dp} (1-p)^k = -(1-p)^{k-1} k$ . Then using the formula for the geometric series,

$$\mathbb{E}[X] = -p \cdot \frac{d}{dp} \left( \frac{1}{1-(1-p)} \right) \quad (1.9.60)$$

$$= -p \cdot \frac{d}{dp} \left( \frac{1}{p} \right) \quad (1.9.61)$$

$$= -p \cdot \left( -\frac{1}{p^2} \right) \quad (1.9.62)$$

$$= \frac{1}{p} \quad (1.9.63)$$

An alternative specification to the geometric distribution is for  $X$  to be the number of failures before the first success, in which case it only differs by one less than the previous specification, so that for  $k$  failures before the first success:

$$\Pr(X = k) = (1-p)^k p \quad (1.9.64)$$

on support  $\{0, 1, 2, \dots\}$ .

### Memorylessness of Geometric Distribution

Analogously to the exponential distribution, the geometric distribution can be shown to be memoryless, under essentially the same steps. We have

$$\Pr(X > k_2 | X > k_1) = \frac{1 - \Pr(X \leq k_2)}{1 - \Pr(X \leq k_1)} \quad (1.9.65)$$

$$= \frac{(1-p)^{k_2}}{(1-p)^{k_1}} \quad (1.9.66)$$

$$= (1-p)^{k_2 - k_1} \quad (1.9.67)$$

$$= 1 - \Pr(X \leq k_2 - k_1) \quad (1.9.68)$$

$$= \Pr(X > k_2 - k_1) \quad (1.9.69)$$

### 1.9.10 Hypergeometric Distribution

The hypergeometric distribution at  $x$  with parameters  $N, K, n$  is the probability that of a uniformly random sample of size  $n$  drawn without replacement from a finite population of size  $N$ , for which there are  $K$  successful states/individuals, the number of success drawn is equal to  $x$ . By the structure of this characterisation, we can note that

$$N \in \mathbb{N} \quad (1.9.70)$$

$$K \in \{0, 1, \dots, N\} \quad (1.9.71)$$

$$n \in \{0, 1, \dots, N\} \quad (1.9.72)$$

Let  $X$  be the random variable for the number of successes. Observe that if  $n > N - K$ , then there will be at least one success and the total number of successes will be capped by the lower of  $n$  or  $K$ . Hence the support of  $X$  are the integers between  $\max\{0, n - (N - K)\}$  and  $\min\{n, K\}$ , inclusive. We can derive the probability distribution of  $X$  as follows. Consider a sample with  $x$  successes (or equivalently, a sample with  $n - x$  non-successes); there are  $\binom{n}{x}$  ways to choose such a sample. Each sample has a probability of

$$\frac{K \times (K-1) \times \dots \times (K-x+1) \times (N-K) \times \dots \times [(N-K)-(n-x)+1]}{N \times (N-1) \times \dots \times (N-n+1)}$$

$$= \frac{[K!/(K-x)!] \times (N-K)!/[(N-K)-(n-x)]!}{N!(N-n)!} \quad (1.9.73)$$

of being chosen. Therefore the probability of  $X = x$  is

$$\Pr(X = x) = \binom{n}{x} \frac{[K!/(K-x)!] \times (N-K)!/[(N-K)-(n-x)]!}{N!(N-n)!} \quad (1.9.74)$$

$$= \binom{n}{x} \cdot \frac{K!(N-K)!}{N!} \cdot \frac{(N-n)!}{(K-x)![N-(n-x)]!} \quad (1.9.75)$$

$$= \binom{n}{x} \binom{N-n}{K-x} / \binom{N}{K} \quad (1.9.76)$$

We can arrive at the hypergeometric distribution in a simpler, different manner. There are  $\binom{N}{n}$  ways to choose a sample and of those,  $\binom{K}{x} \times \binom{N-K}{n-x}$  will be samples in which there are  $x$  successes. Thus

$$\Pr(X = x) = \binom{K}{x} \binom{N-n}{K-x} / \binom{N}{n} \quad (1.9.77)$$

We can show that the two distributions above are the same, i.e.

$$\binom{n}{x} \binom{N-n}{K-x} / \binom{N}{K} = \binom{K}{x} \binom{N-n}{K-x} / \binom{N}{n} \quad (1.9.78)$$

from the symmetry of the characterisation, and of the equation itself. Suppose in a data generating process that  $K$  successes are randomly assigned after the sample had been chosen. In this case, we can then view a ‘success’ as whether an individual was selected in a sample or not (because it would have automatically been chosen if it had). Thus,  $n$  and  $K$  are interchangeable parameters.

Also by its characterisation as a finite population version of the binomial distribution, we will see that if  $K$  and  $N$  are large (thus being ‘close’ to an infinite population), the hypergeometric distribution will resemble the binomial distribution with success probability  $K/N$ .

### Mean of Hypergeometric Distribution

To derive the mean of the hypergeometric distribution, first observe the identity

$$x \binom{K}{x} = \frac{K!}{(x-1)!(K-x)!} \quad (1.9.79)$$

$$= K \frac{(K-1)!}{(x-1)!(K-x)!} \quad (1.9.80)$$

$$= K \frac{(K-1)!}{(x-1)![K-(x-1)]!} \quad (1.9.81)$$

$$= K \binom{K-1}{x-1} \quad (1.9.82)$$

Then we can write

$$x \Pr(X = x) = x \binom{K}{x} \binom{N-K}{n-x} / \binom{N}{n} \quad (1.9.83)$$

$$= K \binom{K-1}{x-1} \binom{N-K}{n-x} / \binom{N}{n} \quad (1.9.84)$$

Also note the following:

$$\binom{N-K}{n-x} = \binom{N-1-(K-1)}{n-1-(x-1)} \quad (1.9.85)$$

$$\binom{N}{n} = \frac{N}{n} \binom{N-1}{n-1} \quad (1.9.86)$$

Therefore

$$x \Pr(X = x) = n \frac{K}{N} \binom{K-1}{x-1} \binom{N-1-(K-1)}{n-1-(x-1)} / \binom{N-1}{n-1} \quad (1.9.87)$$

Taking the expectation,

$$\mathbb{E}[X] = \sum_x x \Pr(X = x) \quad (1.9.88)$$

$$= n \frac{K}{N} \sum_x \binom{K-1}{x-1} \binom{N-1-(K-1)}{n-1-(x-1)} / \binom{N-1}{n-1} \quad (1.9.89)$$

where to avoid unnecessary complications over support, the sum is taken over the non-negative integers and it is understood that the binomial coefficient  $\binom{n}{k} = 0$  when  $k \notin \{0, 1, \dots, n\}$ . Then we see that the summands are the masses of the hypergeometric distribution with population size  $N - 1$ , success states  $K - 1$  and sample size  $n - 1$ , so the summation equals 1. Hence

$$\mathbb{E}[X] = n \frac{K}{N} \quad (1.9.90)$$

It is here we see that the mean is the same as the binomial distribution (i.e. sampling with replacement) with success probability  $K/N$ . This may initially seem counterintuitive because the hypergeometric distribution considers sampling without replacement, thus successive samples become dependent, and that  $K, N$  are allowed to be very small (moreso increasing the dependence between successive samples). One possible way to resolve the intuition is to imagine that if one were to draw many successes early on, then subsequent successes will be rarer. Conversely, if one were to draw few successes early on, subsequent successes will become more common. These counteracting effects ‘balance out’ each other in such a way that the expected number of successes is the same as that from an infinite population.

### Vandermonde’s Identity

Vandermonde’s identity can be used to derive the hypergeometric distribution, or show that the hypergeometric distribution is a valid distribution. The identity is

$$\binom{n+m}{K} = \sum_{x=0}^K \binom{n}{x} \binom{m}{K-x} \quad (1.9.91)$$

A combinatorial explanation for this identity is as follows. The number of ways to form a committee of  $K$  members from a group of  $n$  and a group of  $m$  is  $\binom{n+m}{K}$ . We can also count this number by summing over all ways to form the same committee, where  $x$  members are from the first group and  $K - x$  members are from the second group. Substituting  $m = N - n$  and rearranging the identity, we have

$$\sum_{x=0}^K \frac{\binom{n}{x} \binom{N-n}{K-x}}{\binom{N}{K}} = 1 \quad (1.9.92)$$

which shows that the sum of all hypergeometric probabilities is one.

#### 1.9.11 Negative Hypergeometric Distribution

#### 1.9.12 Negative Binomial Distribution

The negative binomial distribution (also called the Pascal distribution) is a discrete distribution on support  $x \in \{0, 1, 2, 3, \dots\}$  for the number of successes of i.i.d. Bernoulli trials (with success

probability  $p$ ) before  $r$  failures occur. For example, if we are interested in 1 success before  $r$  failures, the probability of this is given by

$$\Pr(X = 1) = \binom{r}{1} (1-p)^r p^1 \quad (1.9.93)$$

We know that the sequence will be of length  $n = r + 1$ , and that the final trial in the sequence is fixed as a failure. Hence the term  $\binom{r}{1} = \binom{n-1}{1}$  gives the number of combinations where we can place 1 success among  $n - 1$  trials. Generally, the probability mass function is given by

$$\Pr(X = x) = \binom{r+x-1}{x} (1-p)^r p^x \quad (1.9.94)$$

The negative binomial distribution may be used as a model for count data as an alternative to the Poisson distribution, when the data is too ‘overdispersed’ compared to the Poisson distribution. Matching the means of the negative binomial and Poisson distributions, we get  $\lambda = \frac{rp}{1-p}$  and so rearranging gives

$$p = \frac{\lambda}{r+\lambda} \quad (1.9.95)$$

We can rewrite the probability mass function in terms of  $p$  and  $\lambda$  by

$$\Pr(X = x) = \frac{(x+r-1)!}{x!(r-1)!} p^x (1-p)^r \quad (1.9.96)$$

$$= \frac{(x+r-1)!}{x!(r-1)!} \left(\frac{\lambda}{r+\lambda}\right)^x \left(\frac{r}{r+\lambda}\right)^r \quad (1.9.97)$$

$$= \frac{\lambda^x}{x!} \cdot \frac{(x+r-1)!}{(r-1)!(r+\lambda)^x} \cdot \left(\frac{1}{1+\lambda/r}\right)^r \quad (1.9.98)$$

$$= \frac{\lambda^x}{x!} \cdot \frac{(r+x-1)!}{(r-1)!(r+\lambda)^x} \cdot \frac{1}{(1+\lambda/r)^r} \quad (1.9.99)$$

Now taking the limit as  $r \rightarrow \infty$ , the last factor tends to  $e^{-\lambda}$ , using the fact that  $\lim_{r \rightarrow \infty} (1 + \lambda/r)^r = e^\lambda$ . For the middle factor, similar to deriving the Poisson distribution from the binomial, we can show

$$\lim_{r \rightarrow \infty} \frac{(r+x-1)!}{(r-1)!(r+\lambda)^x} = \lim_{r \rightarrow \infty} \frac{(r+x-1) \times \cdots \times r (r-1)!}{(r-1)!(r+\lambda)^x} \quad (1.9.100)$$

$$= \lim_{r \rightarrow \infty} \frac{(r+x-1) \times \cdots \times r}{(r+\lambda)^x} \div \frac{r^x}{r^x} \quad (1.9.101)$$

$$= \lim_{r \rightarrow \infty} \frac{\left(1 + \frac{x-1}{r}\right) \times \cdots \times 1}{(1 + \lambda/r)^x} \quad (1.9.102)$$

$$= 1 \quad (1.9.103)$$

So in the limit as  $r \rightarrow \infty$ , the probability mass function tends to

$$\lim_{r \rightarrow \infty} \Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (1.9.104)$$

which is the probability mass function of the Poisson distribution. Thus the negative binomial distribution can be thought of as a more ‘flexible’ alternative to the Poisson distribution for modelling, since it has more parameters and allows the mean to not necessarily equal the variance.

The negative binomial can also be seen as a compound Poisson distribution where the rate parameter is itself Gamma distributed. Specifically, if  $X \sim \text{Poisson}(\lambda)$  and  $\lambda \sim \text{Gamma}(\alpha, \beta)$  with shape  $\alpha = r$  and  $\beta = \frac{1-p}{p}$ , then  $X$  will be negative binomial distributed with parameters  $r$  and  $p$ . This can be shown as follows. Using the compound representation,

$$\Pr(X = x) = \int_0^\infty \left( \frac{\lambda^x e^{-\lambda}}{x!} \right) \left( \frac{\beta^\alpha x^{\alpha-1} e^{-\lambda\beta}}{\Gamma(\alpha)} \right) d\lambda \quad (1.9.105)$$

$$= \int_0^\infty \left( \frac{\lambda^x e^{-\lambda}}{x!} \right) \left[ \frac{\left( \frac{1-p}{p} \right)^\alpha \lambda^{\alpha-1} e^{-\lambda(1-p)/p}}{\Gamma(r)} \right] d\lambda \quad (1.9.106)$$

$$= \frac{(1-p)^r p^{-r}}{x! \Gamma(r)} \int_0^\infty \lambda^{r+x-1} \exp \left( -\lambda \left( \frac{1-p}{p} + 1 \right) \right) d\lambda \quad (1.9.107)$$

$$= \frac{(1-p)^r p^{-r}}{x! \Gamma(r)} \int_0^\infty \lambda^{r+x-1} e^{-\lambda/p} d\lambda \quad (1.9.108)$$

$$= \frac{(1-p)^r p^{-r}}{x! \Gamma(r)} p^{r+x-1} \int_0^\infty \left( \frac{\lambda}{p} \right)^{r+x-1} e^{-\lambda/p} d\lambda \quad (1.9.109)$$

Now apply the substitution  $s = \lambda/p$ , so  $d\lambda = pds$  and then using the definition of the Gamma function:

$$\Pr(X = x) = \frac{(1-p)^r p^{-r}}{x! \Gamma(r)} p^{r+x-1} \int_0^\infty s^{r+x-1} e^{-s} p ds \quad (1.9.110)$$

$$= \frac{(1-p)^r p^{-r}}{x! \Gamma(r)} p^{r+x} \Gamma(r+x) \quad (1.9.111)$$

$$= \frac{(r+x-1)!}{x! (r-1)!} (1-p)^r p^x \quad (1.9.112)$$

$$= \binom{r+x-1}{x} (1-p)^r p^x \quad (1.9.113)$$

which is the probability mass function of the negative binomial distribution.

### Mean of Negative Binomial Distribution

Note that if  $r = 1$ , the negative binomial distribution essentially becomes the geometric distribution (except the roles of failure and successes are reversed). With success probability  $p$ , the expected number of failures before the first success is  $\frac{1-p}{p}$ . It follows that the expected number of successes before the first failure is  $\frac{p}{1-p}$ . For  $r$  failures, we then simply multiply this by  $r$  to obtain the mean of the negative binomial distribution as

$$\mathbb{E}[X] = \frac{pr}{1-p} \quad (1.9.114)$$

### Variance of Negative Binomial Distribution

The variance can be obtained in the same approach as the mean, via the connection with the geometric distribution. The variance of the geometric distribution for the number of failures before first success with success probability  $p$  is  $\frac{1-p}{p^2}$ , so accordingly the variance of the negative binomial distribution is

$$\text{Var}(X) = \frac{pr}{(1-p)} \quad (1.9.115)$$

## Polya Distribution

The negative binomial distribution can be extended to the case where  $r$  is no longer integer-valued, but real valued ( $r > 0$ ). This is done by replacing the factorials in the binomial coefficient with the gamma function.

$$\Pr(X = x) = \frac{(r + x - 1)!}{x!(r - 1)!} (1 - p)^r p^x \quad (1.9.116)$$

$$= \frac{\Gamma(r + x)}{x!\Gamma(r)} (1 - p)^r p^x \quad (1.9.117)$$

In this case, we refer to the distribution as the Polya distribution.

### 1.9.13 Beta-Binomial Distribution

### 1.9.14 Benford Distribution

## 1.10 Distribution Relationships [123]

### 1.10.1 Cauchy and Gaussian Distribution

The Cauchy random variable can be defined as a ratio of independent zero-mean Gaussian random variables. Let  $Y_1 \sim \mathcal{N}(0, \sigma_1^2)$  and  $Y_2 \sim \mathcal{N}(0, \sigma_2^2)$ . Then from the property of the ratio distribution, the random variable  $X = \frac{Y_1}{Y_2}$  will have density function

$$f_X(x) = \int_{-\infty}^{\infty} f_{Y_1}(xy) f_{Y_2}(y) |y| dy \quad (1.10.1)$$

$$= \int_{-\infty}^{\infty} \left( \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-x^2 y^2 / (2\sigma_1^2)} \right) \left( \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-y^2 / (2\sigma_2^2)} \right) |y| dy \quad (1.10.2)$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp \left[ -\frac{y^2}{2} \left( \frac{x^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \right] |y| dy \quad (1.10.3)$$

$$= \frac{1}{\pi\sigma_1\sigma_2} \int_0^{\infty} y \exp \left[ -\frac{y^2}{2} \left( \frac{x^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \right] dy \quad (1.10.4)$$

We have that

$$\int_0^{\infty} te^{-at^2} dt = \frac{1}{2} \int_0^{\infty} e^{-au} du \quad (1.10.5)$$

$$= \frac{1}{2} \left[ -\frac{e^{-au}}{a} \right]_0^{\infty} \quad (1.10.6)$$

$$= \frac{1}{2a} \quad (1.10.7)$$

by change of variables  $u = t^2$  (this result can also be obtained from the mean of the half-normal distribution), so this gives

$$f_X(x) = \frac{1}{\pi\sigma_1\sigma_2} \cdot \frac{1}{\frac{2}{2} \left( \frac{x^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)} \quad (1.10.8)$$

$$= \frac{1}{\pi\sigma_1\sigma_2} \cdot \frac{\sigma_1^2\sigma_2^2}{x^2\sigma_2^2 + \sigma_1^2} \quad (1.10.9)$$

$$= \frac{1}{\pi} \cdot \frac{\sigma_1/\sigma_2}{x^2 + (\sigma_1/\sigma_2)^2} \quad (1.10.10)$$

which is the density function of a Cauchy random variable with location parameter zero and scale parameter  $\sigma_1/\sigma_2$ .

### 1.10.2 Box-Muller Transform

Let  $X \sim \mathcal{N}(0, 1)$  and  $Y \sim \mathcal{N}(0, 1)$  be independent standard Gaussian random variables. Suppose these are Cartesian coordinates and consider a polar transform:

$$R = \sqrt{X^2 + Y^2} \quad (1.10.11)$$

$$\Theta = \arctan\left(\frac{Y}{X}\right) \quad (1.10.12)$$

We can firstly show that  $R^2$  will be exponentially distributed with rate parameter  $\lambda = 1/2$ , i.e. a mean of 2. This is shown by recognising that  $R^2$  will have a chi-squared distribution with 2 degrees of freedom, and the form of the density will collapse to the exponential distribution:

$$f_{R^2}(r^2) = \frac{1}{2} \exp\left(-\frac{r^2}{2}\right) \quad (1.10.13)$$

We can then show that  $\Theta$  is uniformly distributed on  $(0, 2\pi)$ . Initially, recognise that the ratio  $X/Y$  will be Cauchy distributed with zero location parameter and scale parameter 1. Thus we should find appropriate distributions for  $\Theta$  such that

$$\tan \Theta \sim \text{Cauchy}(0, 1) \quad (1.10.14)$$

If  $U \sim \text{Uniform}(0, 1)$ , then from inverse transform sampling and the Cauchy quantile function, we know a Cauchy(0, 1) random variate can be generated by

$$\tan\left(\pi\left(U - \frac{1}{2}\right)\right) \sim \text{Cauchy}(0, 1) \quad (1.10.15)$$

Then note that  $\tan$  is  $\pi$ -periodic, so although  $\pi\left(U - \frac{1}{2}\right) \sim \text{Uniform}(-\pi/2, \pi/2)$ , however  $\tan(\pi U)$  will have the same distribution as  $\tan\left(\pi\left(U - \frac{1}{2}\right)\right)$  hence

$$\tan(\pi U) \sim \text{Cauchy}(0, 1) \quad (1.10.16)$$

Moreover,  $\tan(\pi U)$  will have the same distribution as  $\tan(2\pi U)$  since we can effectively view the latter as a mixture of two Cauchy(0, 1) random variables. Therefore we can designate  $\Theta$  to have the uniform density

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi}, & 0 < \theta < 2\pi \\ 0, & \theta \notin (0, 2\pi) \end{cases} \quad (1.10.17)$$

We can also show that  $R$  and  $\Theta$  are independent. To do so, we first assume that  $R$  and  $\Theta$  are independent, and then show that this must imply  $X = R \cos \Theta$  and  $Y = R \sin \Theta$  are independent. The expectation of  $XY$  can be written as

$$\mathbb{E}[XY] = \mathbb{E}[R^2 \cos \Theta \sin \Theta] \quad (1.10.18)$$

$$= \mathbb{E}[R^2] \mathbb{E}[\cos \Theta \sin \Theta] \quad (1.10.19)$$

$$= \mathbb{E}[R^2] \mathbb{E}\left[\frac{1}{2} \sin(2\Theta)\right] \quad (1.10.20)$$

$$= 0 \quad (1.10.21)$$

where we first used independence of  $R$  and  $\Theta$ , followed by the double angle formula for sin. Then,  $\mathbb{E}[\sin(2\Theta)]$  since  $\Theta$  is uniform and the integral of sin over one period is zero. Hence,  $X$  and  $Y$  are orthogonal but since they are also zero-mean, then they are uncorrelated. And for

Gaussian random variables, uncorrelatedness is equivalent to independence. This result leads to the Box-Muller transform, which is a way to generate bivariate Gaussian random variates from uniform random variables. Suppose  $U_1$  and  $U_2$  are independent uniform random variables on  $(0, 1)$ . The CDF of  $R^2$  is

$$F_{R^2}(r^2) = 1 - \exp\left(-\frac{r^2}{2}\right) \quad (1.10.22)$$

Using the inverse transform sampling method, we can set

$$U_1 = 1 - \exp\left(-\frac{R^2}{2}\right) \quad (1.10.23)$$

$$R^2 = -2 \ln(1 - U_1) \quad (1.10.24)$$

But  $1 - U_1$  has the same distribution as  $U_1$ , so alternatively

$$R^2 = -2 \ln U_1 \quad (1.10.25)$$

Thus independent standard Gaussian variates  $X$  and  $Y$  can be generated by converting from polar coordinates to Cartesian:

$$X = \sqrt{-2 \ln U_1} \cos(2\pi U_2) \quad (1.10.26)$$

$$Y = \sqrt{-2 \ln U_1} \sin(2\pi U_2) \quad (1.10.27)$$

### 1.10.3 Exponential and Geometric Distribution

As the exponential and geometric distributions both model waiting times (continuous-valued in the case of the exponential and discrete-valued in the case of the geometric) and they both share the memorylessness property, it turns out that either distribution can be derived from the other. Firstly, the geometric distribution can be derived as the floor of an exponentially distributed random variable. More precisely, if  $X \sim \text{Exp}(\lambda)$ , then  $\lfloor X \rfloor \sim \text{Geom}(1 - e^{-\lambda})$ , i.e. with success probability  $p = 1 - e^{-\lambda}$ , where  $\lfloor X \rfloor$  is under the failures before first success representation. To show this,

$$\Pr(\lfloor X \rfloor = k) = \Pr(k \leq X < k + 1) \quad (1.10.28)$$

$$= \Pr(X < k + 1) - \Pr(X < k) \quad (1.10.29)$$

$$= \left(1 - e^{-\lambda(k+1)}\right) - \left(1 - e^{-\lambda k}\right) \quad (1.10.30)$$

using the form of the CDF for the exponential distribution. It is convenient to let  $\lambda = -\log q$  where  $q = 1 - p$  is the failure probability (this in turn satisfies  $p = 1 - e^{-\lambda}$ ). Then

$$\Pr(\lfloor X \rfloor = k) = \left(1 - e^{(k+1)\log q}\right) - \left(1 - e^{k\log q}\right) \quad (1.10.31)$$

$$= \left(1 - q^{k+1}\right) - \left(1 - q^k\right) \quad (1.10.32)$$

$$= q^k - q^{k+1} \quad (1.10.33)$$

$$= q^k (1 - q) \quad (1.10.34)$$

$$= (1 - p)^k p \quad (1.10.35)$$

which is the geometric probability mass function for then number of failures before the first success.

In the other direction, the exponential distribution can be viewed as the limiting case of a ‘normalised’ geometric random variable. Suppose  $X \sim \text{Geom}\left(\frac{1}{n}\right)$  where  $X$  is the number of

trials until first success, then  $\mathbb{E}[X] = n$ , and  $\frac{X}{n} = \frac{X}{\mathbb{E}[X]}$  is a normalised variable in the sense of the number times the average required in trials for the first success. We are interested in the distribution of  $\frac{X}{n}$  as  $n \rightarrow \infty$ . Immediately however,

$$\Pr\left(\frac{X}{n} > x\right) = \Pr(X > nx) \quad (1.10.36)$$

$$= 1 - \Pr(X \leq nx) \quad (1.10.37)$$

$$= (1-p)^{nx} \quad (1.10.38)$$

$$= \left(1 - \frac{1}{n}\right)^{nx} \quad (1.10.39)$$

$$= \left[\left(1 - \frac{1}{n}\right)^n\right]^x \quad (1.10.40)$$

Then using the fact  $\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1}$ ,

$$\lim_{n \rightarrow \infty} \Pr\left(\frac{X}{n} \leq x\right) = 1 - \lim_{n \rightarrow \infty} \Pr\left(\frac{X}{n} > x\right) \quad (1.10.41)$$

$$= 1 - \lim_{n \rightarrow \infty} \left[\left(1 - \frac{1}{n}\right)^n\right]^x \quad (1.10.42)$$

$$= 1 - \left[\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n\right]^x \quad (1.10.43)$$

$$= 1 - e^{-x} \quad (1.10.44)$$

which is the CDF of an  $\text{Exp}(1)$  random variable. Specialising for  $x = 1$ , the interpretation is that if the probability of success  $p$  is small, then the probability of seeing a success within the expected number of trials is roughly  $1 - e^{-1}$ . More generally, we can consider a success probability  $p = \frac{\lambda}{n}$  where the ‘rate of success’  $\lambda > 0$  is allowed to be arbitrarily high since we are taking  $n \rightarrow \infty$ . Then via the same steps,

$$\lim_{n \rightarrow \infty} \Pr\left(\frac{X}{n} \leq x\right) = 1 - \lim_{n \rightarrow \infty} \Pr\left(\frac{X}{n} > x\right) \quad (1.10.45)$$

$$= 1 - \lim_{n \rightarrow \infty} \left[\left(1 - \frac{\lambda}{n}\right)^n\right]^x \quad (1.10.46)$$

$$= 1 - \left[\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n\right]^x \quad (1.10.47)$$

$$= 1 - e^{-\lambda x} \quad (1.10.48)$$

which is the CDF of an  $\text{Exp}(\lambda)$  random variable.

#### 1.10.4 Beta and Gamma Distribution

If  $X$  and  $Y$  are independent and Gamma distributed with  $X \sim \text{Gamma}(a, \lambda)$  and  $Y \sim \text{Gamma}(b, \lambda)$ , then the random variable

$$Z = \frac{X}{X + Y} \quad (1.10.49)$$

will be Beta distributed with  $Z \sim \text{Beta}(a, b)$ . To show this, first note that we can rearrange

$$\frac{X + Y}{X} = \frac{1}{Z} \quad (1.10.50)$$

$$Y = X \left( \frac{1}{Z} - 1 \right) \quad (1.10.51)$$

and write for the cumulative distribution function of  $Z$ :

$$F_Z(z) = \Pr \left( \frac{X}{X+Y} \leq z \right) \quad (1.10.52)$$

$$= \Pr \left( \frac{X+Y}{X} \geq \frac{1}{z} \right) \quad (1.10.53)$$

$$= \Pr \left( Y \geq X \left( \frac{1}{z} - 1 \right) \right) \quad (1.10.54)$$

so by the law of total probability (and noting the support of the Gamma distribution):

$$F_Z(z) = \int_0^\infty \Pr \left( Y \geq X \left( \frac{1}{z} - 1 \right) \middle| X = x \right) f_X(x) dx \quad (1.10.55)$$

$$= \int_0^\infty [1 - F_Y(x(1/z - 1))] f_X(x) dx \quad (1.10.56)$$

due to independence. Then differentiating the CDF yields

$$f_Z(z) = \frac{d}{dz} \int_0^\infty [1 - F_Y(x(1/z - 1))] f_X(x) dx \quad (1.10.57)$$

$$= \int_0^\infty \frac{x}{z^2} f_Y(x(1/z - 1)) f_X(x) dx \quad (1.10.58)$$

Substituting the Gamma densities, we have

$$f_Z(z) = \int_0^\infty \frac{x}{z^2} \cdot \frac{\lambda^b [x(1/z - 1)]^{b-1} e^{-\lambda[x(1/z - 1)]}}{\Gamma(b)} \cdot \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)} dx \quad (1.10.59)$$

$$= \frac{\lambda^{a+b} (1/z - 1)^{b-1}}{\Gamma(a) \Gamma(b)} \int_0^\infty \frac{x}{z^2} \cdot x^{a+b-2} e^{-\lambda x/z} dx \quad (1.10.60)$$

$$= \frac{\lambda^{a+b} (1-z)^{b-1}}{\Gamma(a) \Gamma(b) z^{b-1}} \int_0^\infty \frac{1}{z^2} \cdot x^{a+b-1} e^{-\lambda x/z} dx \quad (1.10.61)$$

Perform the change of variables  $w = \lambda x/z$  so that  $x = zw/\lambda$  and  $dx = zdw/\lambda$ :

$$f_Z(z) = \frac{\lambda^{a+b} (1-z)^{b-1}}{\Gamma(a) \Gamma(b) z^{b-1}} \int_0^\infty \frac{1}{z^2} \cdot \frac{z^{a+b-1} w^{a+b-1}}{\lambda^{a+b-1}} e^{-w} \frac{z dw}{\lambda} \quad (1.10.62)$$

$$= \frac{\lambda^{a+b} (1-z)^{b-1}}{\Gamma(b) \Gamma(a) z^{b-1}} \int_0^\infty \frac{z^{a+b-2} w^{a+b-1}}{\lambda^{a+b}} e^{-w} dw \quad (1.10.63)$$

$$= \frac{z^{a-1} (1-z)^{b-1}}{\Gamma(a) \Gamma(b)} \int_0^\infty w^{a+b-1} e^{-w} dw \quad (1.10.64)$$

Using the definition of the Gamma function:

$$f_Z(z) = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} z^{a-1} (1-z)^{b-1} \quad (1.10.65)$$

$$= \frac{z^{a-1} (1-z)^{b-1}}{B(a, b)} \quad (1.10.66)$$

which is the required Beta density.

## Chapter 2

# Introductory Statistics

## 2.1 Data Generating Processes

In random experiments involving the collection of data, we call the random experiment the data generating process. This data generating process completely defines how observations of data are made. This includes all characteristics about the underlying ‘population’ of interest as well as how data is sampled. The data generating process is usually not completely known, and the inherent goal of statistics to be able to deduce facts about the data generating process using data for the purposes of interpretation and prediction.

### 2.1.1 Populations

The population is the distribution of the particular subject under study for which data is collected.

#### Population Parameters

If the population distribution takes on a particular structure, then the population parameters are the structural parameters which define this distribution. For example, a commonly studied population parameter is the population mean, typically denoted  $\mu$ , which is the expected value of the population distribution. Another example is the population variance and population standard deviation, typically denoted  $\sigma^2$  and  $\sigma$  respectively, which are respectively the variance and standard deviation of the population distribution.

#### Nuisance Parameters

A nuisance parameter is any parameter not of immediate interest but still must be considered in the analysis of another parameter of interest. For example, the population mean may be of primary interest, but the population variance must be considered to conduct inference.

### 2.1.2 Samples

A sample of size  $n$  is a collection of  $n$  values sampled from population under the data generating process and may be denoted  $X_1, X_2, \dots, X_n$  to indicate the fact that they are random variables. A realisation of a sample of size  $n$  may be denoted using  $x_1, x_2, \dots, x_n$ . A single value of the sample  $X_i$  or  $x_i$  may be referred to as an observation.

#### Sample Statistics

A sample statistic (or simply statistic) is anything that is calculated via the sample. Sample statistics may be considered random variables with respect to the data generating process,

because the sample itself is considered random.

## Estimators

An estimator is a sample statistic that is intended to estimate some population parameter of interest. For a population parameter  $\theta$ , we may denote an estimator for  $\theta$  by  $\hat{\theta}$ .

## Sampling Distribution

The sampling distribution of a sample statistic is the probability distribution of that statistic when considered as a random variable, since the statistic is itself calculated from a random sample.

### Unbiasedness

An unbiased estimator  $\hat{\theta}$  for  $\theta$  satisfies

$$\mathbb{E} [\hat{\theta}] = \theta \quad (2.1.1)$$

If  $\mathbb{E} [\hat{\theta}] \neq \theta$ , then  $\hat{\theta}$  is said to be biased.

### Consistency

An estimator  $\hat{\theta}$  for  $\theta$  is said to be consistent if loosely speaking, the estimate  $\hat{\theta}$  gets closer to the population parameter  $\theta$  as the sample size  $n$  grows large.

## 2.2 Descriptive Statistics

### 2.2.1 Measures of Central Tendency

#### Sample Arithmetic Mean

The sample arithmetic mean  $\bar{x}$  (usually referred to as just the sample mean) of a sample  $x_1, \dots, x_n$  is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.2.1)$$

The sample mean is the most typical estimate of the population mean. Suppose that  $X_1, \dots, X_n$  is a random sample with each  $X_i$  drawn from a probability distribution with mean  $\mathbb{E}[X_i] = \mu$ . We can then show that the estimator for the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2.2.2)$$

is an unbiased estimator for the population parameter  $\mu$  as follows:

$$\mathbb{E} [\bar{X}] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] \quad (2.2.3)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E} [X_i] \quad (2.2.4)$$

$$= \frac{1}{n} n\mu \quad (2.2.5)$$

$$= \mu \quad (2.2.6)$$

Suppose that  $X_1, \dots, X_n$  is a random sample of uncorrelated variables with each  $X_i$  drawn from a probability distribution with population variance  $\text{Var}(X_i) = \sigma^2$ . Then the variance of the sample mean can be found using properties of the variance of sums of uncorrelated random variables (also known as Bienaymé's formula):

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \quad (2.2.7)$$

$$= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \quad (2.2.8)$$

$$= \frac{\sum_{i=1}^n \text{Var}(X_i)}{n^2} \quad (2.2.9)$$

$$= \frac{n\sigma^2}{n^2} \quad (2.2.10)$$

$$= \frac{\sigma^2}{n} \quad (2.2.11)$$

The sample mean of a dataset also has the property that it is the value which minimises the average sum of squared deviations from that value:

$$\bar{x} = \underset{c}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (x_i - c)^2 \quad (2.2.12)$$

To show this, take the derivative:

$$\frac{\partial}{\partial c} \frac{1}{n} \sum_{i=1}^n (x_i - c)^2 = -\frac{1}{n} \sum_{i=1}^n 2(x_i - c) \quad (2.2.13)$$

Setting the derivative to zero and solving gives:

$$-\frac{1}{n} \sum_{i=1}^n 2(x_i - c^*) = 0 \quad (2.2.14)$$

$$\frac{nc^*}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.2.15)$$

$$c^* = \bar{x} \quad (2.2.16)$$

## Median

Let  $F(x)$  be a cumulative distribution function. The median  $\tilde{x}$  of the distribution is defined as the value at which  $F(\tilde{x}) = \frac{1}{2}$  (i.e. the value which splits the distribution ‘in half’). A distribution may have more than one median (i.e. if there are multiple values for which  $F(x) = \frac{1}{2}$ ).

### Sample Median

The sample median  $\tilde{x}$  of a sample  $x_1, \dots, x_n$  is the ‘middle value’ of the sample. To be more precise, suppose we can sort the sample so that

$$x_{(1)} \leq \dots \leq x_{(n)} \quad (2.2.17)$$

Then if  $n$  is odd, the median is defined as

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \quad (2.2.18)$$

If  $n$  is even, then the median is defined as the average of the two ‘middle values’:

$$\tilde{x} = \frac{1}{2} \left( x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)} \right) \quad (2.2.19)$$

The sample median can be useful as a measure of central tendency when there are extreme outliers in the sample, which would otherwise have a greater effect on the sample mean than the sample median.

## Mode

Let  $f(x)$  be a probability density function. The mode  $\check{x}$  of the distribution is the ‘peak’ of the distribution, that is  $\check{x} = \max_x f(x)$ . Some distributions may have multiple local peaks, at which  $\frac{df(x)}{dx} = 0$  (supposing that  $f(x)$  is differentiable). We then say that the distribution is multimodal. If there is only a single local peak, then the distribution is said to be unimodal. This definition is analogous to discrete probability distributions, we take the value which has the highest probability mass associated to it.

### Sample Mode

The sample mode of a sample  $x_1, x_2, \dots, x_n$  is the most frequently occurring value in the sample. A sample may have more than one mode.

### Sample Geometric Mean

The sample geometric mean  $\bar{x}_G$  of a sample  $x_1, \dots, x_n$  is defined as

$$\bar{x}_G = \left( \prod_{i=1}^n x_i \right)^{1/n} \quad (2.2.20)$$

Note that the product  $\prod_{i=1}^n x_i$  should be non-negative in order for the geometric mean to be real-valued. Hence the geometric mean is most applicable for giving a valid indication of central tendency if all values in the sample are positive.

The geometric mean is applicable for averaging over compounding growth rates. To illustrate, suppose we have growth rates  $g_1, \dots, g_n$  such that the overall compounded growth is  $\prod_{i=1}^n (1 + g_i)$ . Taking the geometric mean shall give

$$\left( \prod_{i=1}^n (1 + g_i) \right)^{1/n} = 1 + \bar{g} \quad (2.2.21)$$

and note that

$$\prod_{i=1}^n (1 + g_i) = (1 + \bar{g})^n \quad (2.2.22)$$

so that a growth rate of  $\bar{g}$  over  $n$  periods gives an equivalent compounded growth.

The geometric mean is also applicable for averaging normalised results. To illustrate, suppose we have two data sets  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$ . We have a choice of constructing normalised data sets using either  $x_i$  or  $y_i$  as reference values (i.e.  $y_1/x_1, \dots, y_n/x_n$  or  $x_1/y_1, \dots, x_n/y_n$  respectively). However, we have

$$\frac{\left( \prod_{i=1}^n x_i \right)^{1/n}}{\left( \prod_{i=1}^n y_i \right)^{1/n}} = \left( \prod_{i=1}^n \frac{x_i}{y_i} \right)^{1/n} \quad (2.2.23)$$

and

$$\frac{(\prod_{i=1}^n y_i)^{1/n}}{(\prod_{i=1}^n x_i)^{1/n}} = \left( \prod_{i=1}^n \frac{y_i}{x_i} \right)^{1/n} \quad (2.2.24)$$

Hence, regardless of which sample is used as the reference values, the geometric mean preserves the relative ordering of the averaged data.

### Sample Harmonic Mean

The sample harmonic mean  $\bar{x}_H$  of a sample  $x_1, \dots, x_n$  is defined as

$$\bar{x}_H = \left( \frac{1}{n} \sum_{i=1}^n x_i^{-1} \right)^{-1} \quad (2.2.25)$$

Hence the sample harmonic mean is the reciprocal of the sample arithmetic mean of the reciprocals.

The harmonic mean can be applicable in certain cases for averaging rates and ratios. To illustrate, suppose we have a sample  $x_1, \dots, x_n$  where each  $x_i = \frac{a}{b_i}$  is a ratio of two quantities with the numerator being kept constant for each value. We desire the ‘true’ average of the ratio, being  $na / \sum_{i=1}^n b_i$ . Then taking the harmonic mean gives

$$\left( \frac{1}{n} \sum_{i=1}^n x_i^{-1} \right)^{-1} = \left( \frac{1}{n} \cdot \frac{\sum_{i=1}^n b_i}{a} \right)^{-1} \quad (2.2.26)$$

$$= \left( \frac{\sum_{i=1}^n b_i}{na} \right)^{-1} \quad (2.2.27)$$

$$= \frac{na}{\sum_{i=1}^n b_i} \quad (2.2.28)$$

## 2.2.2 Measures of Dispersion

### Sample Variance

For a sample of observations  $x_1, \dots, x_n$ , the sample variance is an estimator for the population variance, and computed by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.2.29)$$

where  $\bar{x}$  is the sample mean. The sample variance has an alternative expanded form, analogously to the population variance:

$$s^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right] \quad (2.2.30)$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \right] \quad (2.2.31)$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \right] \quad (2.2.32)$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right] \quad (2.2.33)$$

### Bessel's Correction

Suppose we knew the value of the population mean  $\mu$ . To compute an unbiased estimate of the population variance from a random sample  $X_1, \dots, X_n$ , the formula is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (2.2.34)$$

This can be shown to be an unbiased estimator because

$$\mathbb{E} [\hat{\sigma}^2] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \right] \quad (2.2.35)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(X_i - \mu)^2] \quad (2.2.36)$$

$$= \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i) \quad (2.2.37)$$

$$= \text{Var}(X_i) \quad (2.2.38)$$

which gives the population variance  $\sigma^2 = \text{Var}(X_i)$ . However, typically the population mean is not known, and only the sample mean  $\bar{X}$  is known, which can be computed from the sample. We can show that this introduces bias if we were to compute sample variance the same way. Firstly, we show that

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \quad (2.2.39)$$

The left-hand side may be expanded as

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \quad (2.2.40)$$

$$= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \quad (2.2.41)$$

$$= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \quad (2.2.42)$$

where  $\bar{X}$  is treated as constant with respect to the sum, and we have used the fact  $\sum_{i=1}^n X_i = n\bar{X}$  by the definition of the sample mean. On the right-hand side,

$$\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 = \sum_{i=1}^n X_i^2 - 2\mu n\bar{X} + \cancel{n\mu^2} - n\bar{X}^2 + \cancel{2n\bar{X}\mu} - \cancel{n\mu^2} \quad (2.2.43)$$

$$= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \quad (2.2.44)$$

via cancellations. Hence  $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$ . Then to show the bias,

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{n}{n} (\bar{X} - \mu)^2 \right] \quad (2.2.45)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2] - \mathbb{E}[(\bar{X} - \mu)^2] \quad (2.2.46)$$

The term  $\mathbb{E}[(X_i - \mu)^2]$  is  $\text{Var}(X_i) = \sigma^2$ , while the term  $\mathbb{E}[(\bar{X} - \mu)^2]$  is the variance of the sample mean  $\text{Var}(\bar{X})$ , which can be shown to be

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \quad (2.2.47)$$

$$= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \quad (2.2.48)$$

$$= \frac{n \text{Var}(X_i)}{n^2} \quad (2.2.49)$$

$$= \frac{\sigma^2}{n} \quad (2.2.50)$$

assuming that the sample is i.i.d. so that the variance of the sum is the sum of variances. Hence

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n} \sum_{i=1}^n \sigma^2 - \frac{\sigma^2}{n} \quad (2.2.51)$$

$$= \frac{n\sigma^2}{n} - \frac{\sigma^2}{n} \quad (2.2.52)$$

$$= \frac{(n-1)\sigma^2}{n} \quad (2.2.53)$$

which does not equal  $\sigma^2$ , so this estimator is biased. Intuitively, the bias is caused because we are measuring squared deviations of a sample from a mean which is itself determined by the sample. So to ‘compensate’ for this, we divide by  $n-1$  instead of  $n$  in the sample variance to give an unbiased estimate of the population variance. The estimator is corrected by a multiplication of  $\frac{n}{n-1}$  so that

$$\mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{n}{n-1} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] \quad (2.2.54)$$

$$= \frac{n}{n-1} \times \frac{(n-1)\sigma^2}{n} \quad (2.2.55)$$

$$= \sigma^2 \quad (2.2.56)$$

This is known as Bessel’s correction.

### Sample Standard Deviation

The sample standard deviation of a sample  $x_1, \dots, x_n$  is the square root of the sample variance:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.2.57)$$

Although the sample variance is unbiased, taking the square root means that the sample standard deviation is generally biased for the population standard deviation.

## Coefficient of Variation

The coefficient of variation is a standardised measure of dispersion relative to the mean. It is the ratio of standard deviation to the mean.

$$\text{COV} = \frac{\sigma}{\mu} \quad (2.2.58)$$

For samples, the analogue uses the sample standard deviation and sample mean.

$$\text{COV} = \frac{s}{\bar{x}} \quad (2.2.59)$$

The coefficient of variation is dimensionless, so it may be used to compare dispersion between different samples.

## Range

The range of a sample  $x_1, \dots, x_n$  is the difference between the smallest and largest values in the data set. That is,

$$\text{RANGE} = \max_i x_i - \min_j x_j \quad (2.2.60)$$

The range is a simple measure of dispersion using only two observations.

## Percentiles

The percentiles divide a probability distribution into 100. For a probability distribution with cumulative distribution function  $F(x)$ , the  $k^{\text{th}}$  percentile (with  $k \in \{0, 1, \dots, 100\}$ ) may be defined as  $F^{-1}(k/100)$ , and may be interpreted as the value for which  $k\%$  of the distribution lies below (or alternatively, the  $k^{\text{th}}$  percentile is greater than  $k\%$  of values in the distribution).

The percentiles of a population may be estimated from a sample. The nearest-rank method of estimation obtains the value for the  $k^{\text{th}}$  percentile such that no more than  $k\%$  of the data is strictly less than the value, and at least  $k\%$  of the data is less than or equal to the value. That is, we want to find the value  $L_k$  such that for a sample  $x_1, \dots, x_n$ :

$$\frac{|\{x : x_i < L_k, i = 1, \dots, n\}|}{n} \times 100 \leq k \quad (2.2.61)$$

$$\frac{|\{x : x_i \leq L_k, i = 1, \dots, n\}|}{n} \times 100 \geq k \quad (2.2.62)$$

or equivalently

$$|\{x : x_i < L_k, i = 1, \dots, n\}| \leq \frac{k}{100} \times n \quad (2.2.63)$$

$$|\{x : x_i \leq L_k, i = 1, \dots, n\}| \geq \frac{k}{100} \times n \quad (2.2.64)$$

If we order the values in the sample, denoted by

$$x_{(1)} < \dots < x_{(n)} \quad (2.2.65)$$

then the nearest-rank percentile  $L_k$  is given by

$$L_k = x_{(\lceil \frac{k}{100} \times n \rceil)} \quad (2.2.66)$$

This is because

$$x_{(1)} < \dots < x_{(\lfloor \frac{k}{100} \times n \rfloor)} < L_k \quad (2.2.67)$$

so the number of values strictly less than  $L_k$  is definitely less than or equal to  $\frac{k}{100} \times n$ . Also,

$$x_{(1)} < \dots < x_{(\lfloor \frac{k}{100} \times n \rfloor)} < x_{(\lceil \frac{k}{100} \times n \rceil)} = L_k \quad (2.2.68)$$

so the number of values less than or equal to  $L_k$  is at least  $\frac{k}{100} \times n$ .

## Quartiles

The quartiles divide a probability distribution into four.

- The first quartile  $Q_1$  (also known as the lower quartile) is equivalent to the 25<sup>th</sup> percentile.
- The second quartile  $Q_2$  (also known as the median) is equivalent to the 50<sup>th</sup> percentile.
- The third quartile  $Q_3$  (also known as the upper quartile) is equivalent to the 75<sup>th</sup> percentile.

The quartiles of a population may be estimated from a sample. One method of estimation is to divide an ordered dataset  $x_{(1)} \leq \dots \leq x_{(n)}$  into two halves, split at the median; if  $n$  is odd then the median is excluded from the halves. The lower quartile is taken to be the median of the lower half, and the upper quartile is taken to be the median of the upper half.

## Interquartile Range

The interquartile range is defined as the difference between the upper and lower quartiles, i.e.

$$\text{IQR} = Q_3 - Q_1 \quad (2.2.69)$$

The interquartile range gives an indication of the range of values for the ‘middle 50’ of a distribution of dataset.

## Quantiles

The quantiles (sometimes also referred to as *fractiles* [88]) generalise percentiles and quartiles. The  $q$ -quantiles are a set of  $q - 1$  values which split a probability distribution into  $q$  equal sizes. With  $q = 2$ ,  $q = 4$  and  $q = 100$ , we recover the median, quartiles and percentiles respectively. For a random variable  $X$  with cumulative distribution  $F(x)$ , we may define  $L_{k,q}$  as the  $k^{\text{th}}$   $q$ -quantile of  $X$  as the value which

$$F^{-1}\left(\frac{k}{q}\right) = L_{k,q} \quad (2.2.70)$$

There exist methods to estimate quartiles of a population from a sample. One method is to generalise the nearest-rank method of computing percentiles, where for an ordered dataset  $x_{(1)} < \dots < x_{(n)}$  we compute

$$L_{k,q} = x_{\left(\lceil \frac{k}{q} \times n \rceil\right)} \quad (2.2.71)$$

## Frequency Distributions

For a sample  $x_1, \dots, x_n$  taking on discrete values in  $\{a_1, \dots, a_N\}$ , the frequency distribution is the count (or frequency) at which each of the values in  $\{a_1, \dots, a_N\}$  appears in the sample. A frequency distribution may be represented in a table:

| Value    | Frequency                              |
|----------|--|
| $a_1$    | $ \{x : x_i = a_1, i = 1, \dots, n\} $ |
| $a_2$    | $ \{x : x_i = a_2, i = 1, \dots, n\} $ |
| $\vdots$ | $\vdots$                               |
| $a_N$    | $ \{x : x_i = a_N, i = 1, \dots, n\} $ |

If instead the sample took on continuous values, we may replace  $\{a_1, \dots, a_N\}$  with bins  $B_1, \dots, B_M$ , where each bin is an interval. The bins should be disjoint (i.e. there are no ‘overlaps’ of the intervals), however the union of all the bins should contain every value in the sample. It is common practice to choose equally spaced bin widths, and a reasonably large  $M$  to capture the spread of data. The frequency distribution table for continuous data is given by:

| Bin      | Frequency                                |
|----------|--|
| $B_1$    | $ \{x : x_i \in B_1, i = 1, \dots, n\} $ |
| $B_2$    | $ \{x : x_i \in B_2, i = 1, \dots, n\} $ |
| $\vdots$ | $\vdots$                                 |
| $B_M$    | $ \{x : x_i \in B_M, i = 1, \dots, n\} $ |

## Relative Frequency Distributions

In a relative frequency distribution, the counts/frequencies have been divided by the sample size  $n$ , so that the sum of all relative frequencies equals 1. The relative frequency distribution is a way to approximate the population distribution.

## Signal-to-Noise Ratio

While there are many definitions of the signal-to-noise ratio, one such definition is the inverse of the coefficient of variation.

$$\text{SNR} = \frac{\mu}{\sigma} \quad (2.2.72)$$

### 2.2.3 Measures of Dependence

#### Sample Covariance

Given two samples  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$  where each pair  $(x_i, y_i)$  is an observation, the sample covariance is computed by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (2.2.73)$$

An analogous identity to the population covariance exists for the sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (2.2.74)$$

$$= \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{x} y_i + \sum_{i=1}^n \bar{x} \bar{y} \right) \quad (2.2.75)$$

$$= \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} \right) \quad (2.2.76)$$

$$= \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right) \quad (2.2.77)$$

Bessel's correction also applies for computing an unbiased estimate of the population covariance. The proof of unbiasedness is a generalisation of the way unbiasedness is proved for the sample variance. First, we start by showing that for two random samples  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$ :

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y) - n(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \quad (2.2.78)$$

where  $\mu_X$  and  $\mu_Y$  are the population means of  $X_i$  and  $Y_i$  respectively. Via some manipulations the left-hand side is shown to be equal to the right-hand side:

$$\begin{aligned} & \sum_{i=1}^n (X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X} \bar{Y}) \\ &= \sum_{i=1}^n (X_i Y_i - \mu_X Y_i - X_i \mu_Y + \mu_X \mu_Y) - n \bar{X} \bar{Y} + n \bar{X} \mu_Y + n \mu_X \bar{Y} - n \mu_X \mu_Y \end{aligned} \quad (2.2.79)$$

$$\begin{aligned} & \sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} - n \bar{Y} \bar{X} + n \bar{X} \bar{Y} \\ &= \sum_{i=1}^n X_i Y_i - n \mu_X \bar{Y} - n \bar{X} \mu_Y + n \mu_X \mu_Y - n \bar{X} \bar{Y} + n \bar{X} \mu_Y + n \mu_X \bar{Y} - n \mu_X \mu_Y \end{aligned} \quad (2.2.80)$$

$$\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} = \sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \quad (2.2.81)$$

Hence

$$\mathbb{E} \left[ \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \right] = \mathbb{E} \left[ \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y) - \frac{n}{n-1} (\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \right] \quad (2.2.82)$$

$$= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}[(X_i - \mu_X)(Y_i - \mu_Y)] - \frac{n}{n-1} \mathbb{E}[(\bar{X} - \mu_X)(\bar{Y} - \mu_Y)] \quad (2.2.83)$$

$$= \frac{n \text{Cov}(X_i, Y_i)}{n-1} - \frac{n}{n-1} \text{Cov}(\bar{X}, \bar{Y}) \quad (2.2.84)$$

The covariance between the sample means  $\text{Cov}(\bar{X}, \bar{Y})$  is a covariance between linear combinations of random variables, so it can be shown to be

$$\text{Cov}(\bar{X}, \bar{Y}) = \text{Cov} \left( \frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n Y_i \right) \quad (2.2.85)$$

$$= \frac{1}{n^2} \text{Cov}(X_1, Y_1) + \frac{1}{n^2} \text{Cov}(X_1, Y_2) + \cdots + \frac{1}{n^2} \text{Cov} \quad (2.2.86)$$

$$= \frac{1}{n^2} \sum_{i=1}^n \sum_{i=j}^n \text{Cov}(X_i, Y_j) \quad (2.2.87)$$

We assume that each observation  $(X_i, Y_i)$  is independent, so  $\text{Cov}(X_i, Y_j) = 0$  when  $i \neq j$ . Hence

$$\text{Cov}(\bar{X}, \bar{Y}) = \frac{1}{n^2} \sum_{i=j} \text{Cov}(X_i, Y_j) \quad (2.2.88)$$

$$= \frac{n \text{Cov}(X_i, Y_i)}{n^2} \quad (2.2.89)$$

$$= \frac{\text{Cov}(X_i, Y_i)}{n} \quad (2.2.90)$$

This gives

$$\mathbb{E} \left[ \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \right] = \frac{n \text{Cov}(X_i, Y_i)}{n-1} - \frac{n}{n-1} \frac{\text{Cov}(X_i, Y_i)}{n} \quad (2.2.91)$$

$$= \frac{(n-1) \text{Cov}(X_i, Y_i)}{n-1} \quad (2.2.92)$$

$$= \text{Cov}(X_i, Y_i) \quad (2.2.93)$$

### Pearson Correlation Coefficient

The Pearson correlation coefficient is an analogue of the correlation computed for a sample. Given a sample with observations  $(x_1, y_1), \dots, (x_n, y_n)$ , we first compute the sample covariance  $s_{xy}$  and the sample standard deviations  $s_x, s_y$ . Then the Pearson correlation coefficient  $r_{xy}$  is given by

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad (2.2.94)$$

Like the population correlation, the Pearson correlation will always lie in  $[-1, 1]$ . However although the sample covariance and sample standard deviations are unbiased, the Pearson correlation coefficient is in general not unbiased for the population correlation.

### Kendall Rank Correlation Coefficient

The Kendall rank correlation coefficient (also known by the Kendall  $\tau$  coefficient) measures the ordinal association between two sets of observations. Let  $(x_1, y_1), \dots, (x_n, y_n)$  be the sets of observations which are a realised sample from the distributions of  $X$  and  $Y$  respectively. Assume that all values in the observations are unique. A pair of observations  $(x_i, y_i)$  and  $(x_j, y_j)$  for any  $i < j$  is said to be *concordant* if  $x_i > x_j$  and  $y_i > y_j$ , or alternatively if  $x_i < x_j$  and  $y_i < y_j$ . Let  $n_c$  be the number of concordant pairs in the set of observations. A pair of observations  $(x_i, y_i)$  and  $(x_j, y_j)$  for any  $i < j$  is said to be *discordant* if  $x_i > x_j$  and  $y_i < y_j$ , or alternatively if  $x_i < x_j$  and  $y_i > y_j$ . Let  $n_d$  be the number of discordant pairs in the set of observations. The Kendall  $\tau$  coefficient is defined as

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \quad (2.2.95)$$

Note that

- The term  $n(n-1)/2$  is the number of pairs in the observations.
- Like with the Pearson correlation coefficient,  $-1 \leq \tau \leq 1$ .

An explicit expression for  $\tau$  is found by noticing  $n_c - n_d = \sum_{i < j} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j)$ :

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j) \quad (2.2.96)$$

An alternative formula for the Kendall correlation by summing over all  $i$  and  $j$  is

$$\tau = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \text{sign}((x_i - x_j)(y_i - y_j)) \quad (2.2.97)$$

because the summands are zero when  $i = j$ .

## Population Kendall Correlation

Looking at the expression for the Kendall rank correlation coefficient, we can see that an appropriate population version of the statistic from a bivariate population  $(X, Y)$

$$\tau_0 = \mathbb{E} [\text{sign}(X - X') \text{sign}(Y - Y')] \quad (2.2.98)$$

where  $(X', Y')$  is an independent observation from  $(X, Y)$ . Assume that  $(X, Y)$  are both continuous. Then, note that  $\text{sign}(X - X')$  is Rademacher distributed and we can see

$$\mathbb{E} [\text{sign}(X - X')] = 0 \quad (2.2.99)$$

$$\mathbb{E} [\text{sign}(X - X')^2] = 1 \quad (2.2.100)$$

$$\text{Var} (\text{sign}(X - X')) = 1 \quad (2.2.101)$$

Similarly,  $\text{Var} (\text{sign}(Y - Y')) = 1$ . Therefore,

$$\tau_0 = \text{Cov} (\text{sign}(X - X'), \text{sign}(Y - Y')) \quad (2.2.102)$$

$$= \text{Corr} (\text{sign}(X - X'), \text{sign}(Y - Y')) \quad (2.2.103)$$

This characterises the Kendall correlation as the correlation in signs of differences between pairs of observations. Another characterisation in terms of probability is given by

$$\tau_0 = \mathbb{E} [\text{sign}((X - X')(Y - Y'))] \quad (2.2.104)$$

$$= \mathbb{E} [\mathbb{I}_{\{(X-X')(Y-Y')>0\}} - \mathbb{I}_{\{(X-X')(Y-Y')<0\}}] \quad (2.2.105)$$

$$= \Pr ((X - X')(Y - Y') > 0) - \Pr ((X - X')(Y - Y') < 0) \quad (2.2.106)$$

Yet another characterisation can be obtained by noticing

$$\mathbb{E} [\mathbb{I}_{\{(X-X')(Y-Y')>0\}}] = \mathbb{E} [\mathbb{I}_{\{X>X', Y>Y'\}} + \mathbb{I}_{\{X<X', Y<Y'\}}] \quad (2.2.107)$$

$$= \Pr (X > X', Y > Y') + \Pr (X < X', Y < Y') \quad (2.2.108)$$

$$= 2 \Pr (X > X', Y > Y') \quad (2.2.109)$$

because  $(X', Y')$  is an independent copy of  $(X, Y)$ . Then using the fact  $2\mathbb{I}_{\{(X-X')(Y-Y')>0\}} - 1 = \text{sign}((X - X')(Y - Y'))$ , we have

$$\tau_0 = \mathbb{E} [\text{sign}((X - X')(Y - Y'))] \quad (2.2.110)$$

$$= 2\mathbb{E} [\mathbb{I}_{\{(X-X')(Y-Y')>0\}}] - 1 \quad (2.2.111)$$

$$= 4 \Pr (X > X', Y > Y') - 1 \quad (2.2.112)$$

which gives

$$\Pr (X > X', Y > Y') = \frac{\tau_0 + 1}{4} \quad (2.2.113)$$

$$\Pr (X < X', Y < Y') = \frac{\tau_0 + 1}{4} \quad (2.2.114)$$

by symmetry. Also, by using conditional probabilities and the fact that  $\Pr (Y < Y') = \frac{1}{2}$ :

$$\frac{\tau_0 + 1}{4} = \Pr (X < X', Y < Y') \quad (2.2.115)$$

$$= \Pr (X < X' | Y < Y') \Pr (Y < Y') \quad (2.2.116)$$

$$= \Pr (X < X' | Y < Y') \frac{1}{2} \quad (2.2.117)$$

Therefore

$$\Pr (X < X' | Y < Y') = \frac{\tau_0 + 1}{2} \quad (2.2.118)$$

$$\Pr (Y < Y' | X < X') = \frac{\tau_0 + 1}{2} \quad (2.2.119)$$

## Spearman's Rank Correlation Coefficient

The Spearman's rank correlation coefficient (also known as the Spearman's  $\rho$  coefficient) measures the strength of monotonic (not necessarily linear) relationship between two variables. For pairs of observations  $(x_1, y_1), \dots, (x_n, y_n)$ , let the ranked variables  $(r_{xi}, r_{yi})$  denote the ranking of each observation within their respective datasets (i.e.  $r_{xi}$  is the ranking of observation  $x_i$  in the dataset for  $x$ , with 1 being the smallest and  $n$  being the largest). Assume that all values in the observations are unique within each sample. This is a reasonable assumption if the population is continuous. The ranks can be explicitly written using indicator functions by

$$r_{xi} = \sum_{j=1}^n \mathbb{I}_{x_j \leq x_i} \quad (2.2.120)$$

$$r_{yi} = \sum_{j=1}^n \mathbb{I}_{y_j \leq y_i} \quad (2.2.121)$$

or using the unit step function  $u(\cdot)$ :

$$r_{xi} = \sum_{j=1}^n u(x_i - x_j) \quad (2.2.122)$$

$$r_{yi} = \sum_{j=1}^n u(y_i - y_j) \quad (2.2.123)$$

The Spearman's  $\rho$  coefficient is then defined by

$$\rho = \frac{s_{r,xy}}{s_{r,x}s_{r,y}} \quad (2.2.124)$$

where  $s_{r,xy}$  is the sample covariance between the ranked observations, while  $s_{r,x}$  and  $s_{r,y}$  are the sample standard deviations of the ranked  $x$  and  $y$  variables respectively. Essentially, the Spearman's  $\rho$  coefficient is the Pearson correlation coefficient of the ranked variables.

## Alternative Formula for Spearman Correlation

An alternative formula for calculating the Spearman correlation is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (2.2.125)$$

where  $d_i^2 = (r_{xi} - r_{yi})^2$  are the squared differences in ranks. To show this, we first write out the Spearman correlation as

$$\rho = \frac{\sum_{i=1}^n r_{xi}r_{yi} - (\sum_{i=1}^n r_{xi})(\sum_{i=1}^n r_{yi}) / n}{\sqrt{\left[ \sum_{i=1}^n r_{xi}^2 - (\sum_{i=1}^n r_{xi})^2 / n \right] \left[ \sum_{i=1}^n r_{yi}^2 - (\sum_{i=1}^n r_{yi})^2 / n \right]}} \quad (2.2.126)$$

The denominator will just be a function of  $n$ , because each summation involves the ranks from 1 to  $n$ . We have

$$\sum_{i=1}^n r_{xi} = \sum_{i=1}^n i \quad (2.2.127)$$

$$= \frac{n(n+1)}{2} \quad (2.2.128)$$

and

$$\sum_{i=1}^n r_{xi}^2 = \sum_{i=1}^n i^2 \quad (2.2.129)$$

$$= \frac{n(n+1)(2n+1)}{6} \quad (2.2.130)$$

Hence

$$\sum_{i=1}^n r_{xi}^2 - \frac{1}{n} \left( \sum_{i=1}^n r_{xi} \right)^2 = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \quad (2.2.131)$$

$$= \frac{n(n+1)}{12} [2(2n+1) - 3(n+1)] \quad (2.2.132)$$

$$= \frac{n(n+1)(n-1)}{12} \quad (2.2.133)$$

$$= \frac{n(n^2-1)}{12} \quad (2.2.134)$$

This is the same for the ranks  $r_{yi}$ , so the denominator equals

$$\sqrt{\left[ \sum_{i=1}^n r_{xi}^2 - \frac{1}{n} \left( \sum_{i=1}^n r_{xi} \right)^2 \right] \left[ \sum_{i=1}^n r_{yi}^2 - \frac{1}{n} \left( \sum_{i=1}^n r_{yi} \right)^2 \right]} = \frac{n(n^2-1)}{12} \quad (2.2.135)$$

Now expand  $\sum_{i=1}^n d_i^2$  as

$$\sum_{i=1}^n (r_{xi} - r_{yi})^2 = \sum_{i=1}^n r_{xi}^2 - 2 \sum_{i=1}^n r_{xi} r_{yi} + \sum_{i=1}^n r_{yi}^2 \quad (2.2.136)$$

$$= 2 \sum_{i=1}^n i^2 - 2 \sum_{i=1}^n r_{xi} r_{yi} \quad (2.2.137)$$

$$= 2 \cdot \frac{n(n+1)(2n+1)}{6} - 2 \sum_{i=1}^n r_{xi} r_{yi} \quad (2.2.138)$$

So that

$$\sum_{i=1}^n r_{xi} r_{yi} = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2} \sum_{i=1}^n d_i^2 \quad (2.2.139)$$

Putting this into the numerator, we get

$$\sum_{i=1}^n r_{xi} r_{yi} - \frac{1}{n} \left( \sum_{i=1}^n r_{xi} \right) \left( \sum_{i=1}^n r_{yi} \right) = -\frac{1}{2} \sum_{i=1}^n d_i^2 + \frac{n(n+1)(2n+1)}{6} - \frac{1}{n} \sum_{i=1}^n i^2 \quad (2.2.140)$$

$$= -\frac{1}{2} \sum_{i=1}^n d_i^2 + \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \quad (2.2.141)$$

$$= -\frac{1}{2} \sum_{i=1}^n d_i^2 + \frac{n(n^2-1)}{12} \quad (2.2.142)$$

using the same steps as for the denominator. Thus

$$\rho = \frac{n(n^2-1)/12 - \sum_{i=1}^n d_i^2/2}{n(n^2-1)/12} \quad (2.2.143)$$

$$= \frac{n(n^2-1) - 6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \quad (2.2.144)$$

$$= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (2.2.145)$$

### Population Spearman Correlation

We can arrive at an appropriate population version of the Spearman correlation as follows. Consider a bivariate continuous population  $(X, Y)$ . Denote the marginal CDFs by  $F_X(x)$  and  $F_Y(y)$  respectively. Note that

$$\frac{1}{n}r_{xi} = \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{x_j \leq x_i} \quad (2.2.146)$$

approximates  $F_X(x_i)$  (using the empirical distribution function). Similarly,  $\frac{1}{n}r_{yi}$  approximates  $F_Y(y_i)$ . Since the Spearman's  $\rho$  coefficient is the Pearson correlation of the ranks, a natural definition of the population Spearman correlation  $\rho_0$  is

$$\rho_0 = \text{Corr} \left( \frac{1}{n}F_X(X), \frac{1}{n}F_Y(Y) \right) \quad (2.2.147)$$

$$= \text{Corr}(F_X(X), F_Y(Y)) \quad (2.2.148)$$

### Relation Between Spearman and Kendall Correlation

The Spearman correlation  $\rho$  can be written in terms of the Kendall correlation  $\tau$  by

$$\rho = \frac{3}{n+1}\tau + \frac{3}{n(n+1)(n-1)} \sum_{i \neq j \neq k} \text{sign}((x_i - x_j)(y_i - y_k)) \quad (2.2.149)$$

*Proof.* Recall the alternative form of the Spearman correlation:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (2.2.150)$$

$$= \frac{n^3 - n - 6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (2.2.151)$$

In the derivation of the alternative form, it was also shown that

$$\sum_{i=1}^n r_{xi}r_{yi} - \frac{1}{n} \left( \sum_{i=1}^n r_{xi} \right) \left( \sum_{i=1}^n r_{yi} \right) = -\frac{1}{2} \sum_{i=1}^n d_i^2 + \frac{n(n^2 - 1)}{12} \quad (2.2.152)$$

which we can rearrange to

$$12 \left[ \sum_{i=1}^n r_{xi}r_{yi} - n \left( \sum_{i=1}^n \frac{r_{xi}}{n} \right) \left( \sum_{i=1}^n \frac{r_{yi}}{n} \right) \right] = -6 \sum_{i=1}^n d_i^2 + n^3 - n \quad (2.2.153)$$

Hence

$$\rho = \frac{12}{n^3 - n} \left[ \sum_{i=1}^n r_{xi}r_{yi} - n \left( \sum_{i=1}^n \frac{r_{xi}}{n} \right) \left( \sum_{i=1}^n \frac{r_{yi}}{n} \right) \right] \quad (2.2.154)$$

$$= \frac{12}{n^3 - n} \left[ \sum_{i=1}^n \left( r_{xi} - \frac{n+1}{2} \right) \left( r_{yi} - \frac{n+1}{2} \right) \right] \quad (2.2.155)$$

using the identity for the sample covariance, with the fact that the average of all the ranks is  $\frac{n+1}{2}$ . Now introduce a function similar to the step function:

$$\nu(x) = \begin{cases} 1, & x > 0 \\ 1/2, & x = 0 \\ 0, & x < 0 \end{cases} \quad (2.2.156)$$

where we can see that akin to the step function representation of ranks, we can write each of the ranks as

$$r_{xi} = \frac{1}{2} + \sum_{j=1}^n \nu(x_i - x_j) \quad (2.2.157)$$

since the  $\frac{1}{2}$  additional makes up for the case when  $i = j$ , compared to the step function. Note that we can relate  $\nu(x)$  to the sign function by

$$\nu(x) = \frac{1}{2} (\text{sign}(x) + 1) \quad (2.2.158)$$

Thus

$$r_{xi} = \frac{1}{2} + \sum_{j=1}^n \frac{1}{2} (\text{sign}(x_i - x_j) + 1) \quad (2.2.159)$$

$$= \frac{n+1}{2} + \frac{1}{2} \sum_{j=1}^n \text{sign}(x_i - x_j) \quad (2.2.160)$$

Rearranging, we see

$$r_{xi} - \frac{n+1}{2} = \frac{1}{2} \sum_{j=1}^n \text{sign}(x_i - x_j) \quad (2.2.161)$$

so putting this into the form of the Spearman correlation earlier, we get

$$\rho = \frac{3}{n^3 - n} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \text{sign}((x_i - x_j)(y_i - y_k)) \quad (2.2.162)$$

$$= \frac{3}{n(n+1)(n-1)} \sum_{i=1}^n \left[ \sum_{j=k}^n \text{sign}((x_i - x_j)(y_i - y_j)) + \sum_{j \neq k} \text{sign}((x_i - x_j)(y_i - y_k)) \right] \quad (2.2.163)$$

Recognising the alternative form of the Kendall correlation as one of the terms, this yields

$$\rho = \frac{3}{n+1}\tau + \frac{3}{n(n+1)(n-1)} \sum_{i \neq j \neq k} \text{sign}((x_i - x_j)(y_i - y_k)) \quad (2.2.164)$$

since in the second term, the summands would be zero if  $i = j$  or  $i = k$ . □

### Contingency Tables

A contingency table can be used to present the joint probability distribution (or alternatively, relative frequency or absolute frequency distribution) of two variables, and by taking summations across rows and columns to obtain marginal distributions. The layout of a  $2 \times 2$  contingency table for two random variables  $X, Y$  taking on values  $\{0, 1\}$  is given below.

|         | $Y = 0$             | $Y = 1$             |              |
|---------|---------------------|---------------------|--------------|
| $X = 0$ | $\Pr(X = 0, Y = 0)$ | $\Pr(X = 0, Y = 1)$ | $\Pr(X = 0)$ |
| $X = 1$ | $\Pr(X = 1, Y = 0)$ | $\Pr(X = 1, Y = 1)$ | $\Pr(X = 1)$ |
|         | $\Pr(Y = 0)$        | $\Pr(Y = 1)$        | 1            |

## 2.2.4 Measures of Shape

### Sample Moments

The  $k^{\text{th}}$  sample moment  $m_k$  of a sample  $x_1, \dots, x_n$  is defined as

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \quad (2.2.165)$$

where  $\bar{x}$  is the sample mean. The  $k^{\text{th}}$  raw sample moment is defined as

$$\eta_k = \frac{1}{n} \sum_{i=1}^n x_i^k \quad (2.2.166)$$

Sample moments usually appear in the computation of other statistics.

### Population Skewness

The population skewness  $\gamma$  of a random variable  $X$  with population mean  $\mu$  and population standard deviation  $\sigma$  is defined as

$$\gamma = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] \quad (2.2.167)$$

$$= \frac{\mathbb{E} [(X - \mu)^3]}{\sigma^3} \quad (2.2.168)$$

$$= \frac{\mathbb{E} [(X - \mu)^3]}{\left( \mathbb{E} [(X - \mu)^2] \right)^{3/2}} \quad (2.2.169)$$

The population skewness is a measure of asymmetry because it indicates whether there is longer ‘tail’ on the left side or right side of mean. To illustrate, note that  $(X - \mu)^3 < 0$  when  $X < \mu$  and  $(X - \mu)^3 > 0$  when  $X > \mu$ . So a positive skewness suggests that values to the right of the mean are weighted more (i.e. contain more extreme values) than values to the left, which is what causes the expectation  $\mathbb{E} [(X - \mu)^3]$  to be positive. So the shape of the distribution would believably have a longer right-tail. Conversely, a negative skewness suggests that the left of the mean contains more extreme values, and so has a longer left-tail.

### Non-Parametric Skew

For a population with mean  $\mu$ , median  $\nu$  and standard deviation  $\sigma$ , the non-parametric skew  $\tilde{\gamma}$  is defined as

$$\tilde{\gamma} = \frac{\mu - \nu}{\sigma} \quad (2.2.170)$$

### Mode Skewness

Also known as Pearson’s first skewness coefficient, the mode skewness is a sample statistic computed from the sample mean  $\bar{x}$ , sample mode  $\check{x}$  and sample standard deviation  $s$ :

$$\gamma_{\text{mode}} = \frac{\bar{x} - \check{x}}{s} \quad (2.2.171)$$

## Median Skewness

Also known as Pearson's second skewness coefficient, the median skewness is a sample statistic computed from the sample mean  $\bar{x}$ , sample median  $\tilde{x}$  and sample standard deviation  $s$ :

$$\gamma_{\text{median}} = 3 \frac{\bar{x} - \tilde{x}}{s} \quad (2.2.172)$$

Note that the median skewness estimates a scaled version of non-parametric skew.

## Left-Skewed Distributions

A left-skewed (also known as negative-skewed) distribution has a longer left-tail. A sample with mean less than median indicates that the population is left-skewed (and supported by the median skewness being negative), because extreme values to the left are 'pulling' the sample mean downwards relative to the sample median. We also typically associate left-skewed distributions with the median being less than the mode, because the median is being pulled down but to less of an extent compared to the mean.

## Right-Skewed Distributions

A right-skewed (also known as right-skewed) distribution has a longer right-tail. A sample with mean greater than median indicates that the population is right-skewed (and supported by the median skewness being positive), because extreme values to the right are 'pulling' the sample mean upwards relative to the sample median. We also typically associate right-skewed distributions with the mode being less than the median, because the median is being pulled up but to less of an extent compared to the mean.

## Symmetric Distributions

A symmetric distribution is neither left-skewed nor right-skewed. If the mean, median and mode of a distribution are all equal, this is typically associated with a symmetric distribution.

## Fisher-Pearson Coefficient of Skewness

The Fisher-Pearson coefficient of skewness is an estimator for the population skewness. Given a sample  $x_1, \dots, x_n$ , the estimator is

$$\hat{\gamma} = \frac{m_3}{m_2^{3/2}} \quad (2.2.173)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})^3 / n}{\left( \sum_{i=1}^n (x_i - \bar{x})^2 / n \right)^{3/2}} \quad (2.2.174)$$

where  $m_3$  and  $m_2$  are the third and second sample moments respectively. Note that this estimator simply replaces the expectations in the numerator and denominator of the population skewness by their respective sample moments. In general, this estimator is biased.

## Adjusted Fisher-Pearson Standardised Moment Coefficient

The adjusted Fisher-Pearson standardised moment coefficient aims to replace the expectations in the numerator and denominator of the population skewness by unbiased estimates. An unbiased estimator of  $\mathbb{E}[(X - \mu)^3]$  is given by a correction factor on the third sample moment:

$$\widehat{\mathbb{E}}[(X - \mu)^3] = \frac{n^2}{(n-2)(n-1)} m_3 \quad (2.2.175)$$

and an unbiased estimator of  $\mathbb{E}[(X - \mu)^2]$  is simply the sample variance. Hence the adjusted estimator is

$$\hat{\gamma}_{\text{adj}} = \frac{n^2}{(n-2)(n-1)} \frac{m_3}{(s^2)^{3/2}} \quad (2.2.176)$$

$$= \frac{n^2}{(n-2)(n-1)} \frac{m_3}{[nm_2/(n-1)]^{3/2}} \quad (2.2.177)$$

$$= \frac{n^2}{(n-2)(n-1)} \frac{(n-1)^{3/2}}{n^{3/2}} \frac{m_3}{m_2^{3/2}} \quad (2.2.178)$$

$$= \frac{\sqrt{n(n-1)}}{n-2} \frac{\sum_{i=1}^n (x_i - \bar{x})^3 / n}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 / n\right)^{3/2}} \quad (2.2.179)$$

$$= \frac{\sqrt{n(n-1)}}{n-2} \hat{\gamma} \quad (2.2.180)$$

However, in general this estimator will still be biased.

### Population Kurtosis

The population kurtosis  $\kappa$  of a random variable  $X$  with population mean  $\mu$  and population standard deviation  $\sigma$  is

$$\kappa = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] \quad (2.2.181)$$

$$= \frac{\mathbb{E}[(X - \mu)^4]}{\sigma^4} \quad (2.2.182)$$

$$= \frac{\mathbb{E}[(X - \mu)^4]}{\left(\mathbb{E}[(X - \mu)^2]\right)^2} \quad (2.2.183)$$

The kurtosis is a measure of ‘tailedness’ of a distribution. Since the exponent on the deviation  $X - \mu$  is to the fourth order, this heavily weights deviations from the mean, more so than for variance. Hence for two distributions with the same variance, we would interpret the distribution with the higher kurtosis as having ‘fatter’ tails. The distribution with lower kurtosis would be interpreted as being more ‘peaked’ and having thinner tails.

### Sample Kurtosis

A simple estimator for population kurtosis from a sample  $x_1, \dots, x_n$  is to estimate the expectations in the numerator and denominator of population kurtosis, then take the ratio. This means:

$$\hat{\kappa} = \frac{m_4}{m_2^2} \quad (2.2.184)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 / n\right)^2} \quad (2.2.185)$$

Both the numerator and denominator are biased, and this will in general be a biased estimator. One improvement may be to replace the components of the numerator and denominator with

their unbiased counterparts. The unbiased estimator of  $\mathbb{E}[(X - \mu)^2]$  is the sample variance  $s^2$ . The unbiased estimator of  $\mathbb{E}[(X - \mu)^4]$  is given by

$$\hat{\mathbb{E}}[(X - \mu)^2] = \frac{n^2 [(n+1)m_4 - 3(n-1)m_2^2]}{(n-1)(n-2)(n-3)} \quad (2.2.186)$$

Hence the adjusted sample kurtosis in terms of the sample moments is

$$\hat{\kappa}_{\text{adj}} = \frac{n^2 [(n+1)m_4 - 3(n-1)m_2^2]}{(n-1)(n-2)(n-3)} \left(\frac{1}{s^2}\right)^2 \quad (2.2.187)$$

$$= \frac{n^2 [(n+1)m_4 - 3(n-1)m_2^2]}{(n-1)(n-2)(n-3)} \left(\frac{n-1}{nm_2}\right)^2 \quad (2.2.188)$$

$$= \frac{n^2 [(n+1)m_4 - 3(n-1)m_2^2]}{(n-1)(n-2)(n-3)} \cdot \frac{(n-1)}{n^2 m_2^2} \quad (2.2.189)$$

Like the adjusted sample skewness, this adjusted kurtosis will still in general be biased.

## 2.3 Normal Statistics

### 2.3.1 $z$ -Scores

If we have a realisation  $x$  from a normal distribution  $\mathcal{N}(\mu, \sigma^2)$ , then we may compute the so-called  $z$ -score via *standardisation*:

$$z = \frac{x - \mu}{\sigma} \quad (2.3.1)$$

The  $z$ -score gives the number of standard deviations that  $x$  is away from its mean. Additionally, if  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $Z = \frac{X - \mu}{\sigma}$  will be standard normally distributed.

### 2.3.2 Normal Sample Mean

Consider a random i.i.d. sample  $X_1, \dots, X_n$  from  $\mathcal{N}(\mu, \sigma^2)$ . The sampling distribution of the sample mean  $\bar{X}_n$  is given by

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad (2.3.2)$$

We can show this by showing that  $\mathbb{E}[\bar{X}_n] = \mu$  and  $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ , and using the fact that the sum of normally distributed random variables will also be normally distributed. Firstly, by the unbiasedness of the sample mean,

$$\mathbb{E}[\bar{X}_n] = \mu \quad (2.3.3)$$

Next, using properties of variance of sums of uncorrelated random variables (Bienaymé's formula):

$$\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \quad (2.3.4)$$

### 2.3.3 Normal Sample Variance

#### Sampling Distribution of Normal Sample Variance

Consider a random i.i.d. sample  $X_1, \dots, X_n$  from  $\mathcal{N}(\mu, \sigma^2)$ . Consider the sample variance  $S_n^2$ :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (2.3.5)$$

Then the sample statistic

$$\chi^2_{n-1} = \frac{(n-1)S_n^2}{\sigma^2} \quad (2.3.6)$$

$$= \sum_{i=1}^n \frac{(X_i - \mu - \bar{X}_n + \mu)^2}{\sigma^2} \quad (2.3.7)$$

$$= \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} - \frac{\bar{X}_n - \mu}{\sigma} \right)^2 \quad (2.3.8)$$

$$= \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \quad (2.3.9)$$

will follow the chi-squared distribution with  $n - 1$  degrees of freedom.

### Independence of Normal Sample Mean and Sample Variance

Let  $Z_1, \dots, Z_n$  be i.i.d. random variables from the standard normal distribution (i.e. mean of 0 and variance of 1). The sample mean is defined by

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i \quad (2.3.10)$$

We first show that the random variables  $Z_i - \bar{Z}$  and  $\bar{Z}$  are independent for any  $i = 1, \dots, n$ . Because  $Z_i - \bar{Z}$  and  $\bar{Z}$  are jointly normally distributed, it is enough to show that their covariance is zero. Using the definition of covariance and the fact that  $\mathbb{E}[Z_i] = \mathbb{E}[\bar{Z}] = \mathbb{E}[Z_i - \bar{Z}] = 0$ :

$$\text{Cov}(\bar{Z}, Z_i - \bar{Z}) = \mathbb{E}[\bar{Z}(Z_i - \bar{Z})] - \mathbb{E}[\bar{Z}]\mathbb{E}[Z_i - \bar{Z}] \quad (2.3.11)$$

$$= \mathbb{E}[\bar{Z}(Z_i - \bar{Z})] \quad (2.3.12)$$

$$= \mathbb{E}[\bar{Z}Z_i - \bar{Z}^2] \quad (2.3.13)$$

$$= \mathbb{E}\left[\frac{1}{n} \sum_{j=1}^n Z_j Z_i - \frac{1}{n^2} \left(\sum_{j=1}^n Z_j\right) \left(\sum_{k=1}^n Z_k\right)\right] \quad (2.3.14)$$

$$= \mathbb{E}\left[\frac{1}{n} \sum_{j=1}^n Z_j Z_i - \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n Z_j Z_k\right] \quad (2.3.15)$$

$$= \frac{1}{n} \sum_{j=1}^n \mathbb{E}[Z_j Z_i] - \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n \mathbb{E}[Z_j Z_k] \quad (2.3.16)$$

Note that for standard normal random variables  $\mathbb{E}[Z_j Z_i] = \text{Var}(Z_i) = 1$  if  $i = j$  and  $\mathbb{E}[Z_j Z_i] = 0$  otherwise. Also  $\mathbb{E}[Z_j Z_k] = 1$  if  $j = k$  and zero otherwise. Hence

$$\text{Cov}(\bar{Z}, Z_i - \bar{Z}) = \frac{1}{n} - \frac{n}{n^2} \quad (2.3.17)$$

$$= 0 \quad (2.3.18)$$

Since the sample variance given by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2 \quad (2.3.19)$$

contains a sum over  $Z_i - \bar{Z}$  but each  $Z_i - \bar{Z}$  is independent with  $\bar{Z}$ , it follows that  $\bar{Z}$  is independent with  $S^2$ . Then for any i.i.d. normal sample  $X_1, \dots, X_n$  from a population with mean  $\mu$  and

variance  $\sigma^2$ , the sample mean  $\bar{X}$  is independent with the sample variance  $S_X^2$  because the same argument as above can be applied to the standardised  $\frac{X_i - \mu}{\sigma}$  and  $\frac{\bar{X} - \mu}{\sigma}$ .

### 2.3.4 Excess Kurtosis

The normal distribution has a kurtosis of 3 (regardless of mean and variance parameters), which can be obtained by computing:

$$\mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = \mathbb{E} [Z^4] \quad (2.3.20)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^4 e^{-z^2/2} dz \quad (2.3.21)$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} z^4 e^{-z^2/2} dz \quad (2.3.22)$$

by symmetry of the normal distribution. Making a change of variables  $y = z^2$  so that  $dy = 2zdz$ , then:

$$\mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} y^2 e^{-y/2} \frac{dy}{2\sqrt{y}} \quad (2.3.23)$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} y^{3/2} e^{-y/2} dy \quad (2.3.24)$$

By using integration by parts, let  $u = y^{3/2}$  be differentiable and  $v' = e^{-y/2}$  be integrable so that  $u' = 3y^{1/2}/2$  and  $v = -2e^{-y/2}$ . Then

$$\mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = [uv]_{y=0}^{y=\infty} - \int_0^{\infty} u' v dy \quad (2.3.25)$$

$$= \frac{1}{\sqrt{2\pi}} \left[ \left[ -2e^{-y/2} y^{3/2} \right]_0^{\infty} + 3 \int_0^{\infty} y^{1/2} e^{-y/2} dy \right] \quad (2.3.26)$$

Applying integration by parts again, now let  $u = y^{1/2}$  and  $v' = e^{-y/2}$  so that  $u' = y^{-1/2}/2$  and  $v = -2e^{-y/2}$ . Then

$$\int_0^{\infty} y^{1/2} e^{-y/2} dy = \left[ -2y^{1/2} e^{-y/2} \right]_0^{\infty} + \int_0^{\infty} y^{-1/2} e^{-y/2} dy \quad (2.3.27)$$

Hence

$$\mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = \frac{3}{\sqrt{2\pi}} \int_0^{\infty} y^{-1/2} e^{-y/2} dy \quad (2.3.28)$$

$$= \frac{3}{\sqrt{2\pi}} \int_0^{\infty} z^{-1} e^{-z^2/2} 2z dz \quad (2.3.29)$$

$$= \frac{3}{\sqrt{2\pi}} 2 \int_0^{\infty} e^{-z^2/2} dz \quad (2.3.30)$$

$$= 3 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (2.3.31)$$

by reverting the change of variables and symmetry argument. Recognising that  $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1$  for the normal density, we finally have that

$$\mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = 3 \quad (2.3.32)$$

Excess kurtosis is a quantity relative to the kurtocity of the normal distribution, which is simply kurtosis subtracted by 3.

### Mesokurticity

A distribution with excess kurtosis of zero is said to be mesokurtic. The normal distribution itself is mesokurtic.

### Leptokurticity

A distribution with positive excess kurtosis is said to be leptokurtic. Leptokurtic distributions are qualitatively characterised by fatter tails and taller peaks in the centre (relative to a normal distribution).

### Platykurticity

A distribution with negative excess kurtosis is said to be platykurtic. Platykurtic distributions are qualitatively characterised by thinner tails and a broader centre (relative to a normal distribution).

## 2.3.5 Qualitative Central Limit Theorem

The central limit theorem illustrates the ubiquitous nature of the normal distribution. It relates to the sampling distribution of the sample mean from an arbitrary population as the sample size grows. Qualitatively, the sampling distribution of the sample mean resembles a normal distribution for large sample sizes. More formally, let  $X_1, \dots, X_n$  be a random i.i.d. sample from a population with finite mean  $\mu$  and variance  $\sigma^2$ . Then for large  $n$ , the sample mean  $\bar{X}_n$  is approximately distributed as

$$\bar{X}_n \xrightarrow{\text{approx.}} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad (2.3.33)$$

Note that if the population is normally distributed, then the approximation is exact. The central limit theorem allows for inference and estimation to be conducted on unknown population parameters, even when little is known about the underlying distribution.

## 2.4 Inferential Statistics

### 2.4.1 Confidence Intervals

A confidence interval for a population parameter  $\theta$  is an interval estimator for  $\theta$ , constructed from a random sample  $(X_1, \dots, X_n)$  in such a way that there is a prescribed probability  $100(1 - \alpha)\%$  (usually a high number such as 90%, 95% or 99%) that the interval contains  $\theta$  (with respect to the probability distribution of the random sample). That is, the interval consists of endpoints  $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$  which are functions of the random sample such that ideally:

$$\Pr(L(X_1, \dots, X_n) \leq \theta \leq U(X_1, \dots, X_n)) = 1 - \alpha \quad (2.4.1)$$

If we cannot make this probability equal to the confidence level  $1 - \alpha$  exactly (perhaps because the endpoints  $L$  and  $U$  are discrete random variables), then a  $(1 - \alpha)$ -confidence interval should ideally be the smallest interval such that

$$\Pr(L(X_1, \dots, X_n) \leq \theta \leq U(X_1, \dots, X_n)) \geq 1 - \alpha \quad (2.4.2)$$

## Coverage Probabilities

Note that it may not always be the case that we can satisfy  $\Pr(L \leq \theta \leq U) = 1 - \alpha$  or  $\Pr(L \leq \theta \leq U) \geq 1 - \alpha$ , since we sampling distribution may need to be approximated in order to construct the confidence interval. The actual probability of  $\Pr(L \leq \theta \leq U)$  is known as the coverage probability.

### 2.4.2 Confidence Intervals on Population Mean

Suppose for a random variable  $X$ , we know it has a population mean  $\mathbb{E}[X] = \mu$  (we may not actually know the value of the mean itself), however we do know the standard deviation  $\text{sd}(X) = \sigma$ . Also suppose that we do know enough such that the probability of  $X$  being within  $t > 0$  standard deviations within the mean is given by  $1 - \alpha$ , where  $\alpha \in [0, 1]$ . That is,

$$\Pr(\mu - t\sigma \leq X \leq \mu + t\sigma) = 1 - \alpha \quad (2.4.3)$$

We can rearrange the above event as follows.

$$\mu - t\sigma \leq X \leq \mu + t\sigma \quad (2.4.4)$$

$$\mu \leq X + t\sigma \cap X - t\sigma \leq \mu \quad (2.4.5)$$

$$X - t\sigma \leq \mu \leq X + t\sigma \quad (2.4.6)$$

Hence

$$\Pr(X - t\sigma \leq \mu \leq X + t\sigma) = 1 - \alpha \quad (2.4.7)$$

That is, if we obtain a realisation of  $X$  and construct the interval  $[X - t\sigma, X + t\sigma]$ , then there is  $1 - \alpha$  probability that the interval will contain the population mean  $\mu$ . Note that this probability refers to before the random variable  $X$  is realised. Suppose we have already obtained a realisation (call this  $x$ ), then the matter of whether the interval  $[x - t\sigma, x + t\sigma]$  contains  $\mu$  is no longer down to probability (it either will contain  $\mu$  or it will not). Hence this explains why the interval  $[x - t\sigma, x + t\sigma]$  is named a  $(1 - \alpha) \times 100\%$  ‘confidence’ interval.

### Confidence Intervals with Population Standard Deviation

We can construct confidence intervals on the population mean from a independent random sample. First, assume

- The population distribution  $X$  is normally distributed.
- $X$  has unknown mean  $\mu$  and known standard deviation  $\sigma$ .

Then from the central limit theorem, we know that the sample mean  $\bar{X}$  is exactly normally distributed with the sampling distribution  $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ , where  $n$  is the sample size. Denote  $z_{1-\alpha/2}$  as the  $(1 - \alpha/2) \times 100$  percentile of the standard normal distribution, where  $1 - \alpha$  is the confidence level. That is to say,  $\alpha/2$  of the standard normal distribution lies in the upper tail, to the right of  $z_{1-\alpha/2}$ . Due to the symmetry of the normal distribution,  $-z_{1-\alpha/2}$  is the  $\alpha/2 \times 100$  percentile and moreover, the probability within  $z_{1-\alpha/2}$  standard deviations of the mean is  $1 - \alpha$ . Thus, a  $(1 - \alpha)$  confidence interval for  $\mu$  is constructed as  $[\bar{X} - z_{1-\alpha/2}\sigma, \bar{X} + z_{1-\alpha/2}\sigma]$ .

We can relax the assumption that  $X$  is normally distributed, by appealing to the central limit theorem. Now, we only require

- The central limit theorem applies to the population distribution  $X$ .
- $X$  has unknown mean  $\mu$  and known standard deviation  $\sigma$ .

In this case, an approximate  $(1 - \alpha)$  confidence interval for  $\mu$  is constructed as  $[\bar{X} - z_{1-\alpha/2}\sigma, \bar{X} + z_{1-\alpha/2}\sigma]$ .

---

### Confidence Intervals with Sample Standard Deviation

### Confidence Intervals on Population Proportion

#### 2.4.3 Confidence Intervals on Population Variance

#### 2.4.4 Prediction Intervals [137]

Let  $(X_1, \dots, X_n)$  be a random sample from a population. Let  $X_{n+1}$  be the ‘next’ observation from the population (that would be sampled following the first  $n$ ). For a prescribed level of confidence  $100(1 - \alpha)\%$  with  $\alpha \in (0, 1)$ , a prediction interval consists of endpoints  $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$  which are functions of the sample, and constructed in such a way that

$$\Pr(L(X_1, \dots, X_n) \leq X_{n+1} \leq U(X_1, \dots, X_n)) = 1 - \alpha \quad (2.4.8)$$

That is to say,  $100(1 - \alpha)\%$  of the intervals we construct this way will contain the next observation  $X_{n+1}$ . Conversely, in  $100\alpha\%$  of the intervals we construct,  $X_{n+1}$  will be outside it.

#### 2.4.5 Tolerance Intervals [137]

Let  $(X_1, \dots, X_n)$  be an independent random sample from a population with cumulative distribution function  $F(x)$ . For a prescribed level of confidence  $100(1 - \alpha)\%$  with  $\alpha \in (0, 1)$  and a proportion  $p$ , a tolerance interval consists of endpoints  $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$  which are functions of the sample, and constructed in such a way that

$$\Pr(F(U(X_1, \dots, X_n)) - F(L(X_1, \dots, X_n)) \geq p) = 1 - \alpha \quad (2.4.9)$$

That is to say, for  $100(1 - \alpha)\%$  of the intervals we construct this way, at least  $p$  proportion of the population will be contained within the interval. The main distinctions between prediction intervals and tolerance intervals are that:

- Tolerance intervals have a prescribed confidence  $1 - \alpha$  as well as proportion  $p$ , whereas prediction intervals just have a prescribed confidence.
- Tolerance intervals make no explicit reference to a next observation  $X_{n+1}$ .

A tolerance interval could be interpreted as being  $100(1 - \alpha)\%$  confident that a next observation will lie in the interval with probability at least  $p$ . However, this does not necessarily imply that we should be  $100\alpha\%$  confident that the next observation will lie outside the interval with probability no more than  $1 - p$ ; the proper converse is that we are  $100(1 - \alpha)\%$  confident that a next observation will lie outside the interval with probability at least  $1 - p$ . Alternatively, we are  $100 \times \alpha\%$  confident that the probability of the next observation lying in the interval is less than  $p$ .

## 2.4.6 Null Hypothesis Statistical Testing

**Null Hypotheses**

**Alternative Hypotheses**

**Level of Significance**

**Type I Errors**

**Type II Errors**

***p*-values**

**Statistical Size**

**Statistical Power**

If we knew all the information about the population and data-generating process, then in principle it is possible to compute power of a particular test. However, this is highly idealised because if the data-generating process were already known, hypothesis tests would not be needed. Also, if it turns out that the null hypothesis were actually true, then the notion of power becomes ill-defined. Rather, we can still reason about how changing our experimental design will affect power, i.e. increasing the sample size or choosing a larger  $\alpha$  should increase power.

## 2.4.7 Hypothesis Tests for Population Mean

**Hypothesis Tests with Population Standard Deviation**

**Hypothesis Tests with Sample Standard Deviation**

**Hypothesis Tests Population Proportion**

## 2.4.8 Hypothesis Tests for Population Variance

## 2.4.9 Chi-Squared Goodness-of-Fit Testing

We can use a chi-squared statistic to test whether a random sample from a discrete distribution fits/matches a template distribution. The test is also known as Pearson's chi-squared test. Suppose the discrete distribution is over  $K$  categories, and we denote the probabilities in the template distribution by  $p_1, \dots, p_K$ . We observe a random sample of size  $n$ , and denote the respective counts in each of the categories by  $X_1, \dots, X_n$  (so the  $X_i/n$  are the relative frequencies). Define the statistic

$$T = \sum_{i=1}^K \frac{(X_i - np_i)^2}{np_i} \quad (2.4.10)$$

Under the null hypothesis that the data comes from the template distribution, then in large samples,  $T$  can be shown to be approximately chi-squared distributed, using properties of the multinomial distribution. That is,

$$T \xrightarrow{\text{approx.}} \chi_{K-1}^2 \quad (2.4.11)$$

and failure to reject the null is evidence that the data comes from the template distribution (conversely, rejection of the null is statistically significant evidence that the data comes from a different distribution).

### Chi-Squared Goodness-of-Fit Test for Parametric Distributions

Suppose we want to test whether some data comes from a parametric family with cumulative distribution function  $F(y; \theta)$ , parametrised by  $\theta$ . There can be more than one parameter;

generally we will say that there are  $m$  parameters in total. A chi-squared goodness-of-fit can be used, however the main obstacles that need to be addressed are:

1. The distribution may be on infinite support, so we cannot prescribe  $K$  categories in the template distribution.
2. We do not know which value of  $\theta$  to use in the template distribution.

To overcome the first issue, we can divide the distributions into  $K$  non-overlapping bins  $B_1, \dots, B_K$ . Denote  $p_i(\theta)$  the probability of random variable  $Y$  from  $F(y; \theta)$  falling into bin  $B_i$ , which has endpoints  $(b_{i-1}, b_i]$ , i.e.

$$p_i(\theta) = \Pr(Y \in B_i) \quad (2.4.12)$$

$$= \Pr(b_{i-1} < Y \leq b_i) \quad (2.4.13)$$

$$= F(b_i; \theta) - F(b_{i-1}; \theta) \quad (2.4.14)$$

The endpoints  $b_0, \dots, b_K$  may be selected arbitrarily or as desired (although they should ideally not depend on the data). If  $F(y; \theta)$  is on unbounded support, we can take  $b_0$  and  $b_K$  to be  $-\infty$  and  $\infty$  respectively as needed. Now we can perform the usual chi-squared goodness-of-fit test by comparing the relative frequencies in the bins against the template distribution  $p_1(\theta), \dots, p_K(\theta)$ .

Now to address the issue of the value which to use for  $\theta$ , we can take an estimate  $\hat{\theta}$  from the data. For the test to be valid, we should take the maximum likelihood estimate using only the count data from the bins (and not the actual data). However in practice, it may be fine to use another estimator and on the actual data. Overall, the test hypotheses become

- $H_0$ : The data comes from the parametric family of distributions.
- $H_A$ : The data does not come from the parametric family of distributions.

Once  $\hat{\theta}$  has been estimated from the counts  $X_1, \dots, X_K$  in the bins, we can compute the test statistic now given by

$$T = \sum_{i=1}^K \frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \quad (2.4.15)$$

Under the null hypothesis and the appropriate validity conditions satisfied, this statistic is approximately chi-squared distributed with  $K - m - 1$  degrees of freedom in large samples [109]. So compared to the usual test, the degrees of freedom here is reduced by the number of parameters needing to be estimated. The intuition is that a degree of freedom is ‘used up’ to estimate each parameter, and the additional uncertainty associated with having to estimate the parameters makes it harder to reject the null hypothesis.

### Chi-Squared Test of Independence

Let  $(Y, Z)$  be a bivariate population where  $Y$  and  $Z$  are both categorical variables, with number of categories  $K_Y$  and  $K_Z$  respectively. Suppose we wish to test the hypothesis that  $Y$  and  $Z$  are independent, against the alternative that they are dependent. One way to do this is with a chi-squared goodness-of-fit test. If  $Y$  and  $Z$  are truly independent, then

$$\Pr(Y = i, Z = j) = \Pr(Y = i)\Pr(Z = j) \quad (2.4.16)$$

for every  $i$  and  $j$ . This becomes our template distribution, but we must estimate the marginal distributions  $\Pr(Y = i)$  and  $\Pr(Z = j)$ . Once we have count data for both  $Y$  and  $Z$  from a

sample of size  $n$  (which can be represented in a contingency table), let  $X_{i,j}$  denote the counts in category  $i$  for  $Y$  and  $j$  for  $Z$ . The marginal distributions are estimated by

$$\hat{p}_{Y,i} = \widehat{\Pr}(Y = i) \quad (2.4.17)$$

$$:= \frac{1}{n} \sum_{j=1}^{K_Z} X_{i,j} \quad (2.4.18)$$

for  $i = 1, \dots, K_Y - 1$  with the final estimate satisfying

$$\hat{p}_{Y,K_Y} = 1 - \sum_{i=1}^{K_Y-1} \hat{p}_{Y,i} \quad (2.4.19)$$

and similarly

$$\hat{p}_{Z,j} = \widehat{\Pr}(Z = j) \quad (2.4.20)$$

$$:= \frac{1}{n} \sum_{i=1}^{K_Y} X_{i,j} \quad (2.4.21)$$

for  $j = 1, \dots, K_Z - 1$  with

$$\hat{p}_{Z,K_Z} = 1 - \sum_{j=1}^{K_Z-1} \hat{p}_{Z,j} \quad (2.4.22)$$

The test statistic, in the usual way (but now summing over cells in the contingency table and assuming independence) is given by

$$T = \sum_{i=1}^{K_Y} \sum_{j=1}^{K_Z} \frac{(X_{i,j} - n\hat{p}_{Y,i}\hat{p}_{Z,j})^2}{n\hat{p}_{Y,i}\hat{p}_{Z,j}} \quad (2.4.23)$$

The number of parameters we have estimated is

$$(K_Y - 1) + (K_Z - 1) = K_Y + K_Z - 2 \quad (2.4.24)$$

so the appropriate degrees of freedom for the chi-squared statistic is

$$K_Y K_Z - (K_Y + K_Z - 2) - 1 = K_Y K_Z - K_Y - K_Z + 1 \quad (2.4.25)$$

$$= (K_Y - 1)(K_Z - 1) \quad (2.4.26)$$

## 2.5 Two-Sample Inference

### 2.5.1 Pooled Variance

The pooled variance is an estimator for population variance using samples from different populations, under the assumption that the population variances of all the populations are the same and equal to  $\sigma^2$ . Suppose there are two samples with sample sizes  $n_1$  and  $n_2$ . Then the pooled variance estimator  $s^2$  in terms of the individual sample variances  $s_1^2$  and  $s_2^2$  respectively is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (2.5.1)$$

By the assumption that the population variances are equal, then  $\mathbb{E}[s_1^2] = \mathbb{E}[s_2^2] = \sigma^2$  as the sample variance is unbiased. It can then be shown that the pooled variance is also unbiased by

$$\mathbb{E}[s^2] = \mathbb{E}\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right] \quad (2.5.2)$$

$$= \frac{(n_1 - 1)\mathbb{E}[s_1^2] + (n_2 - 1)\mathbb{E}[s_2^2]}{n_1 + n_2 - 2} \quad (2.5.3)$$

$$= \frac{(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2}{n_1 + n_2 - 2} \quad (2.5.4)$$

$$= \sigma^2 \quad (2.5.5)$$

The pooled variance can be generalised for up to an arbitrary  $k$  samples from populations with identical variance, with the estimator (still unbiased) given by

$$s^2 = \frac{\sum_{i=1}^k (n_i - 1)s_i^2}{\sum_{i=1}^k (n_i - 1)} \quad (2.5.6)$$

Note that this is just a weighted average of sample variances, weighted by sample size (with Bessel's correction).

### Pooled Standard Deviation

The pooled standard deviation is the square root of the pooled variance, and estimates the population standard deviation under the same assumption that the variance of all the populations are the same.

$$s = \sqrt{\frac{\sum_{i=1}^k (n_i - 1)s_i^2}{\sum_{i=1}^k (n_i - 1)}} \quad (2.5.7)$$

#### 2.5.2 Matched Pairs $t$ -test

#### 2.5.3 Independent Samples Tests

##### Two-Sample Student's $t$ -test

Let  $(X_1, \dots, X_{n_1})$  and  $(Y_1, \dots, Y_{n_2})$  be independent samples. We would like to test the hypothesis that the population means ( $\mu_1$  and  $\mu_2$  respectively) of the two samples are equal, against the alternative that they are not equal, e.g. for a two-sided test:

$$H_0 : \mu_1 = \mu_2 \quad (2.5.8)$$

$$H_A : \mu_1 \neq \mu_2 \quad (2.5.9)$$

For this test to be valid, we assume that:

- The population variances of both populations are identical and equal to  $\sigma^2$ .
- The sample means  $\bar{X}$  and  $\bar{Y}$  are exactly or approximately normally distributed with means  $\mu_1$ ,  $\mu_2$  respectively, and variances  $\frac{\sigma^2}{n_1}$ ,  $\frac{\sigma^2}{n_2}$  respectively.
- The statistic  $\frac{n_1 + n_2 - 2}{\sigma^2}s^2$  is exactly or approximately chi-squared distributed with  $n_1 + n_2 - 2$  degrees of freedom, and independent of the sample means.

The last two assumptions will hold exactly if both samples are from normally distributed populations. For the latter, notice that

$$\frac{n_1 + n_2 - 2}{\sigma^2}s^2 = \frac{n_1 + n_2 - 2}{\sigma^2} \cdot \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2} \quad (2.5.10)$$

$$= \sum_{i=1}^{n_1} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^{n_2} \left( \frac{Y_i - \bar{Y}}{\sigma} \right)^2 \quad (2.5.11)$$

which if the populations are normal, is the sum of a chi-squared random variable with  $n_1 - 1$  degrees of freedom with another independent chi-squared random variable with  $n_2 - 1$  degrees of freedom. Hence by the characterisation of the chi-squared distribution, their sum is also chi-squared distributed with  $n_1 + n_2 - 2$  degrees of freedom. Even if the samples are not from normal populations, the test can still be used in larger samples since the second assumption can hold by the central limit theorem, and the third assumption is not so important in larger samples. Under the null hypothesis, the statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s\sqrt{1/n_1 + 1/n_2}} \quad (2.5.12)$$

is (or approximately is)  $t$ -distributed with  $n_1 + n_2 - 2$  degrees of freedom. To see why [120], rearrange  $T$  as

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \div \sqrt{\frac{s^2}{\sigma^2}} \quad (2.5.13)$$

$$= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \div \sqrt{\frac{\sum_{i=1}^{n_1} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^{n_2} \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2}{n_1 + n_2 - 2}} \quad (2.5.14)$$

which fits the characterisation of the  $t$ -distribution since the first term is (or approximately is) standard normal distributed (observe that  $\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$ ) and as previously established,  $\sum_{i=1}^{n_1} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^{n_2} \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2$  is (or approximately is) chi-squared distributed, and independent of the first term.

### Welch's $t$ -test

Welch's  $t$ -test is applicable when we have independent samples like in the two-sample Student's  $t$ -test, but we no longer assume that the population variances are identical. Otherwise, we require similar assumptions, the most important being:

- The sample means  $\bar{X}$  and  $\bar{Y}$  are exactly or approximately normally distributed with means  $\mu_1, \mu_2$  respectively, and variances  $\frac{\sigma^2}{n_1}, \frac{\sigma^2}{n_2}$  respectively.

Of course, this assumption is satisfied exactly if the samples are from normal populations. The test is conducted by computing the statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad (2.5.15)$$

which under the null is approximately  $t$ -distributed with degrees of freedom given by the closest integer to

$$r = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2} \quad (2.5.16)$$

This formula for the degrees of freedom is known as Welch-Satterthwaite's formula.

### Welch-Satterthwaite's Formula

Let there be independent samples  $(X_1, \dots, X_{n_1})$  and  $(Y_1, \dots, Y_{n_2})$  from  $\mathcal{N}(\mu_1, \sigma_1)$  and  $\mathcal{N}(\mu_2, \sigma_2)$  respectively. Then consider the ‘*t*-statistic’ given by

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad (2.5.17)$$

which strictly speaking, does not have a *t*-distribution. Note that

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) \quad (2.5.18)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (2.5.19)$$

and we define  $\sigma^2 := \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ . This explains why the expression  $s^2 := \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$  appears in the statistic. Recall that in the one sample case, the statistic  $\frac{(n-1)s^2}{\sigma^2}$  is chi-squared distributed with  $n-1$  degrees of freedom, denoted  $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ . However this will not hold in the two sample case, which is why  $T$  is not *t*-distributed. However, if we wish to approximate this statistic with a *t*-distribution, then all that is left to do is to determine an appropriate degrees of freedom. Because the variance of a  $\chi_r^2$  random variable is  $2r$ , then in the two-sample case we would like to find the degrees of freedom  $r$  such that

$$\text{Var}\left(\frac{rs^2}{\sigma^2}\right) = 2r \quad (2.5.20)$$

Matching the variance in such a way then allows us to reasonably approximate

$$\frac{rs^2}{\sigma^2} \underset{\text{approx}}{\sim} \chi_r^2 \quad (2.5.21)$$

We first compute the actual variance of  $\frac{rs^2}{\sigma^2}$ . Start with

$$\text{Var}\left(\frac{rs^2}{\sigma^2}\right) = \frac{rs^2}{\sigma^4} \text{Var}(s^2) \quad (2.5.22)$$

$$= \frac{rs^2}{\sigma^4} \text{Var}\left(\frac{s_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad (2.5.23)$$

$$= \frac{rs^2}{\sigma^4} \left( \frac{1}{n_1} \text{Var}(s_1^2) + \frac{1}{n_2} \text{Var}(\sigma_2^2) \right) \quad (2.5.24)$$

by independence of the samples. Then using the fact  $\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2$  gives

$$\text{Var}\left(\frac{(n_1-1)s_1^2}{\sigma_1^2}\right) = 2(n_1-1) \quad (2.5.25)$$

$$\frac{(n_1-1)^2}{\sigma_1^4} \text{Var}(s_1^2) = 2(n_1-1) \quad (2.5.26)$$

$$\text{Var}(s_1^2) = \frac{2\sigma_1^4}{n_1-1} \quad (2.5.27)$$

Similarly,

$$\text{Var}(s_2^2) = \frac{2\sigma_2^4}{n_2-1} \quad (2.5.28)$$

Hence

$$\text{Var}\left(\frac{rs^2}{\sigma^2}\right) = \frac{rs^2}{\sigma^4} \left( \frac{1}{n_1} \cdot \frac{2\sigma_1^4}{n_1 - 1} + \frac{1}{n_2} \cdot \frac{2\sigma_2^4}{n_2 - 1} \right) \quad (2.5.29)$$

Matching this to  $2r$  and solving, we find

$$2r = \frac{rs^2}{\sigma^4} \left( \frac{1}{n_1} \cdot \frac{2\sigma_1^4}{n_1 - 1} + \frac{1}{n_2} \cdot \frac{2\sigma_2^4}{n_2 - 1} \right) \quad (2.5.30)$$

$$r = \frac{\sigma^4}{\frac{1}{n_1^2} \cdot \frac{\sigma_1^4}{n_1 - 1} + \frac{1}{n_2^2} \cdot \frac{\sigma_2^4}{n_2 - 1}} \quad (2.5.31)$$

Since the population standard deviations  $\sigma_1$  and  $\sigma_2$  are typically unknown, replace them and  $\sigma$  with their respective sample versions:

$$r = \frac{s^4}{\frac{1}{n_1^2} \cdot \frac{s_1^4}{n_1 - 1} + \frac{1}{n_2^2} \cdot \frac{s_2^4}{n_2 - 1}} \quad (2.5.32)$$

$$= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2} \quad (2.5.33)$$

This gives the Welch-Satterthwaite formula for the degrees of freedom in Welch's  $t$ -test. Rearranging  $T$ , we have

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s} \quad (2.5.34)$$

$$= \frac{[(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)] / \sigma}{\sqrt{\frac{rs^2}{\sigma^2}} / \sqrt{r}} \quad (2.5.35)$$

Note that the numerator has a  $\mathcal{N}(0, 1)$  distribution. So assuming that the numerator and denominator are independent, and that  $\frac{rs^2}{\sigma^2}$  is approximately chi-squared distributed with  $r$  degrees of freedom, then this fits the characterisation of  $T$  being approximately  $t$ -distributed with  $r$  degrees of freedom. Even if  $\frac{rs^2}{\sigma^2}$  is not close to being close to chi-squared distributed, then if the sample sizes  $n_1$  and  $n_2$  are still large enough, the variance of the denominator  $\sqrt{s_1^2/n_1 + s_2^2/n_2}$  should be small enough so that the distribution of the statistic  $T$  is not affected too much.

### McNemar's Test

#### 2.5.4 Fisher's Exact Test

## 2.6 Simple Linear Regression

In statistical modelling, we specify that the *dependent variable*  $y$  is a function of the *independent variable*  $x$ , given by some relationship

$$y = f(x) \quad (2.6.1)$$

However, the “dependent” variable and “independent” variable are unfortunately named so, and need not have anything to do with the probabilistic definitions of the dependence and independence. The word “independent” is used to refer to the ability of the experimenter to choose the values of  $x$  when designing an experiment, although in not all cases will be values of  $x$  be freely available to be chosen by the experimenter. There are many synonyms that may be used in place of the dependent and independent variable.

| $y$                | $x$                  |
|--------------------|----------------------|
| Dependent variable | Independent variable |
| Explained variable | Explanatory variable |
| Label              | Feature              |
| Regressand         | Regressor            |
| Output variable    | Input variable       |
| Response variable  | Covariate            |
| Predicted variable | Control variable     |
| Prediction         | Predictor            |

We also usually model noise in the data generating process so that our collected measurements of  $y$  is also a function of some noise  $\varepsilon$

$$y = f(x) + \varepsilon \quad (2.6.2)$$

where  $\varepsilon$  is called the noise or ‘error’ term, and is usually treated as random.

### 2.6.1 Simple Least Squares Estimator

### 2.6.2 Coefficient of Determination

#### Decomposition of the Total Sum of Squares

In simple linear regression, the total sum of squares SST is given by

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2.6.3)$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \quad (2.6.4)$$

$$= \sum_{i=1}^n \left[ (y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2 \right] \quad (2.6.5)$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (2.6.6)$$

$$= \text{SSR} + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \text{SSE} \quad (2.6.7)$$

which is in terms of the sum of squares of the residuals SSR and the sum of squares of the estimate SSE. We assert that the middle term is zero. To show this, we first take it and write it in terms of the residuals  $e_i$ :

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^n e_i (\hat{y}_i - \bar{y}) \quad (2.6.8)$$

$$= \sum_{i=1}^n e_i \hat{y}_i - \bar{y} \sum_{i=1}^n e_i \quad (2.6.9)$$

Note that the sum of the residuals is zero:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) \quad (2.6.10)$$

$$= \sum_{i=1}^n y_i - \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad (2.6.11)$$

$$= n\bar{y} - n(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) \quad (2.6.12)$$

$$= n\bar{y} - n\bar{y} \quad (2.6.13)$$

$$= 0 \quad (2.6.14)$$

Hence

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^n e_i \hat{y}_i \quad (2.6.15)$$

$$= \sum_{i=1}^n e_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad (2.6.16)$$

$$= \hat{\beta}_0 \sum_{i=1}^n e_i + \hat{\beta}_1 \sum_{i=1}^n e_i x_i \quad (2.6.17)$$

$$= \hat{\beta}_1 \sum_{i=1}^n e_i x_i \quad (2.6.18)$$

We focus on expanding the term  $\sum_{i=1}^n e_i x_i$ :

$$\sum_{i=1}^n e_i x_i = \sum_{i=1}^n (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) x_i \quad (2.6.19)$$

$$= \sum_{i=1}^n x_i y_i - n\hat{\beta}_0 \bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \quad (2.6.20)$$

$$= \sum_{i=1}^n x_i y_i - n(\bar{y} - \hat{\beta}_1 \bar{x}) \bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \quad (2.6.21)$$

$$= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 \left( n\bar{x}^2 - \sum_{i=1}^n x_i^2 \right) \quad (2.6.22)$$

If this expression is zero, then we can rearrange it to

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \quad (2.6.23)$$

which is true by the definition of the simple least squares estimator. This proves the assertion that  $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$  and therefore the total sum of squares can be decomposed into

$$\text{SST} = \text{SSR} + \text{SSE} \quad (2.6.24)$$

### 2.6.3 Inference for Linear Regressions

Standard Errors of Simple Least Squares

Linear Regression Confidence Intervals

Linear Regression  $t$ -tests

Confidence Intervals on the Conditional Mean

Prediction Intervals on the Observed Response

## 2.7 Design of Experiments

### 2.7.1 Survey Methods

Stratified Sampling

Cluster Sampling

### 2.7.2 Factorial Experiments

## 2.8 Statistical Graphics

### 2.8.1 Scatter Plots

A scatter plot of the bivariate data  $(x_1, y_1), \dots, (x_n, y_n)$  gives a way to visualise the joint density of the population. If the data are drawn i.i.d. from a continuous bivariate population  $(X, Y)$ , then regions where points are densely clustered together indicate the density is high over that region, whereas places where points are more sparsely spread indicate regions of lower density.

The scatter plot also allows us to see whether there is a functional relationship between  $X$  and  $Y$ , e.g. for the case of fitting a curve via regression.

### 2.8.2 Histograms

A bar chart of the frequency distribution or relative frequency distribution is called a histogram. This allows us to visualise the shape of the density (or mass) for the underlying population, since the height of the bars directly corresponds to the density/mass of the underlying distribution.

### 2.8.3 Q-Q Plots

## 2.9 Method of Moments [72]

Let  $X$  be a random variable from a population with  $K$  parameters  $\theta_1, \dots, \theta_K$ . The method of moments is an intuitive approach for estimating  $\theta_1, \dots, \theta_K$  from the sample moments. Suppose the first  $K$  population moments  $m_1 = \mathbb{E}[X]$ ,  $m_2 = \mathbb{E}[X^2]$ , etc. can be written in terms of the parameters  $\theta_1, \dots, \theta_K$  by the relations:

$$m_1 = \mu_1(\theta_1, \dots, \theta_K) \quad (2.9.1)$$

$$m_2 = \mu_2(\theta_1, \dots, \theta_K) \quad (2.9.2)$$

$$\vdots \quad (2.9.3)$$

$$m_K = \mu_K(\theta_1, \dots, \theta_K) \quad (2.9.4)$$

Our estimators  $\hat{\theta}_1, \dots, \hat{\theta}_K$  are determined by ‘plugging in’ the sample moments  $\hat{m}_1, \dots, \hat{m}_K$ , so that

$$\hat{m}_1 = \mu_1(\hat{\theta}_1, \dots, \hat{\theta}_K) \quad (2.9.5)$$

$$\hat{m}_2 = \mu_1(\hat{\theta}_1, \dots, \hat{\theta}_K) \quad (2.9.6)$$

$$\vdots \quad (2.9.7)$$

$$\hat{m}_K = \mu_1(\hat{\theta}_1, \dots, \hat{\theta}_K) \quad (2.9.8)$$

where

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n X_i \quad (2.9.9)$$

$$\hat{m}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad (2.9.10)$$

$$\vdots \quad (2.9.11)$$

Then, we solve the system equations to obtain explicit equations for  $\hat{\theta}_1, \dots, \hat{\theta}_K$  in terms of  $\hat{m}_1, \dots, \hat{m}_K$ .

### 2.9.1 Method of Moments for Normal Distribution

Let  $X$  be a normally distributed random variable with parameters being the mean  $\mu$  and variance  $\sigma^2$ . We know that

$$\mu = \mathbb{E}[X] \quad (2.9.12)$$

and

$$\sigma^2 = \text{Var}(X) \quad (2.9.13)$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (2.9.14)$$

Hence the moment equations can be set up as

$$m_1 = \mu \quad (2.9.15)$$

$$m_2 = \mu^2 + \sigma^2 \quad (2.9.16)$$

with  $\mu_1(\mu, \sigma^2) = \mu$  and  $\mu_2(\mu, \sigma^2) = \mu^2 + \sigma^2$ . Replacing these equations by their sample equivalent,

$$\hat{m}_1 = \hat{\mu} \quad (2.9.17)$$

$$\hat{m}_2 = \hat{\mu}^2 + \hat{\sigma}^2 \quad (2.9.18)$$

Rearranging these yields the same mean as the estimator for  $\mu$ :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2.9.19)$$

$$= \bar{X} \quad (2.9.20)$$

and for the estimator of the population variance:

$$\hat{\sigma}^2 = \hat{m}_2 - \hat{\mu}^2 \quad (2.9.21)$$

$$= \frac{1}{n} \sum_{i=1}^n X_i^2 - \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 \quad (2.9.22)$$

$$= \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 \quad (2.9.23)$$

We show how this relates to the traditional sample variance  $s^2$ :

$$\hat{\sigma}^2 = \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \right) \quad (2.9.24)$$

$$= \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - \sum_{i=1}^n 2\bar{X}X_i + \sum_{i=1}^n \bar{X}^2 \right) \quad (2.9.25)$$

$$= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) \quad (2.9.26)$$

$$= \frac{n-1}{n} s^2 \quad (2.9.27)$$

Therefore the moment of methods estimator for the population variance of the normal distribution is similar to the sample variance except with factor  $\frac{1}{n}$  instead of  $\frac{1}{n-1}$  (i.e. without Bessel's correction).

### 2.9.2 Method of Moments for Negative Binomial Distribution

Recall that the negative binomial distribution with parameters  $p$  (probability of success) and  $r$  (number of failures) has mean

$$\mu := \mathbb{E}[X] \quad (2.9.28)$$

$$= \frac{pr}{1-p} \quad (2.9.29)$$

and a variance

$$\sigma^2 := \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (2.9.30)$$

$$= \frac{pr}{(1-p)^2} \quad (2.9.31)$$

We solve the equations above for  $p$  and  $r$ . Recognising that  $\frac{\mu}{\sigma^2} = 1 - p$ , we have

$$p = 1 - \frac{\mu}{\sigma^2} \quad (2.9.32)$$

$$= 1 - \frac{\mathbb{E}[X]}{\mathbb{E}[X^2] - \mathbb{E}[X]^2} \quad (2.9.33)$$

Rearranging the equation for  $\mu$  and substituting this expression for  $p$ , we get

$$r = \frac{\mu(1-p)}{p} \quad (2.9.34)$$

$$= \frac{\mu^2/\sigma^2}{1 - \mu/\sigma^2} \quad (2.9.35)$$

$$= \frac{\mu^2}{\sigma^2 - \mu} \quad (2.9.36)$$

$$= \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2] - \mathbb{E}[X]^2 - \mathbb{E}[X]} \quad (2.9.37)$$

Hence we have the method of moments estimator for  $p$ :

$$\hat{p} = 1 - \frac{\hat{\mu}}{\hat{\sigma}^2} \quad (2.9.38)$$

$$= 1 - \frac{\hat{m}_1}{\hat{m}_2 - \hat{m}_1^2} \quad (2.9.39)$$

where  $\hat{m}_1$  and  $\hat{m}_2$  are the usual estimators for the first and second moments respectively. Similarly for  $r$ :

$$\hat{r} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}} \quad (2.9.40)$$

$$= \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2 - \hat{m}_1} \quad (2.9.41)$$

which will not necessarily be an integer, but if positive, is still a valid parameter for the Polya distribution. If we require  $r$  to be an integer-valued parameter however, we can take the closest integer (via rounding) as an estimate for  $r$ :

$$\hat{r} = \left\lfloor \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}} + \frac{1}{2} \right\rfloor \quad (2.9.42)$$

$$= \left\lfloor \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2 - \hat{m}_1} + \frac{1}{2} \right\rfloor \quad (2.9.43)$$

### 2.9.3 Method of Percentiles [60]

Let  $x_1, \dots, x_n$  be a realised i.i.d. sample from some population with parameters  $\theta_1, \dots, \theta_K$ . Denote its cumulative distribution function by  $F(x; \theta_1, \dots, \theta_K)$ . The realised sample may be ordered, and denoted by

$$x_{(1)} \leq \dots \leq x_{(n)} \quad (2.9.44)$$

Then for any  $i \in \{1, \dots, n\}$ , the value  $x_{(i)}$  is the  $100 \times (i/n)^{\text{th}}$  percentile. A way to estimate  $\theta_1, \dots, \theta_K$  (that is similar in spirit to the method of moments) could then be to take  $K$  arbitrary indices  $i_1, \dots, i_K$ , and equate the empirical quantiles to the population quantiles. This estimation involves solving the following system of equations:

$$F(x_{(i_1)}; \hat{\theta}_1, \dots, \hat{\theta}_K) = \frac{i_1}{n} \quad (2.9.45)$$

$$\vdots \quad (2.9.46)$$

$$F(x_{(i_K)}; \hat{\theta}_1, \dots, \hat{\theta}_K) = \frac{i_K}{n} \quad (2.9.47)$$

for  $\hat{\theta}_1, \dots, \hat{\theta}_K$ .

# Chapter 3

# Intermediate Probability

## 3.1 Random Vectors

A random vector  $\mathbf{X} = (X_1, \dots, X_n)$  is a vector in which each element is a random variable.

### 3.1.1 Multivariate Probability Distributions

#### Multivariate Cumulative Distribution Functions

The cumulative distribution function  $F_{\mathbf{X}}(x_1, \dots, x_n)$  for a random vector  $\mathbf{X} = (X_1, \dots, X_n)$  follows an analogous definition to the univariate case:

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \Pr(X_1 \leq x_1, \dots, X_n \leq x_n) \quad (3.1.1)$$

As with the univariate CDF, the multivariate CDF satisfies the limit properties

$$\lim_{x_1, \dots, x_n \rightarrow -\infty} F_{\mathbf{X}}(x_1, \dots, x_n) = 0 \quad (3.1.2)$$

$$\lim_{x_1, \dots, x_n \rightarrow \infty} F_{\mathbf{X}}(x_1, \dots, x_n) = 1 \quad (3.1.3)$$

Furthermore, the multivariate CDF must be non-decreasing everywhere. Thus if  $F_{\mathbf{X}}(x_1, \dots, x_n)$  is differentiable, then

$$\frac{\partial F_{\mathbf{X}}(x_1, \dots, x_n)}{\partial x_i} \geq 0 \quad (3.1.4)$$

for each  $i \in \{1, \dots, n\}$ .

#### Multivariate Joint Probability Density Functions

If  $\mathbf{X} = (X_1, \dots, X_n)$  is a random vector of continuous variables, then the joint probability density function  $f_{\mathbf{X}}(x_1, \dots, x_n)$  satisfies

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(t_1, \dots, t_n) dt_n \dots dt_1 \quad (3.1.5)$$

Hence to ‘reverse’ this, it follows that the probability density function be obtained from the cumulative distribution function by

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F_{\mathbf{X}}(x_1, \dots, x_n) \quad (3.1.6)$$

## Multivariate Marginal Probability Distributions

If  $\mathbf{X} = (X_1, \dots, X_n)$  is a random vector of continuous variables, then the marginal density for the first  $m < n$  random variables  $X_1, \dots, X_m$ , denoted  $f_{\mathbf{X}_{1:m}}(x_1, \dots, x_m)$ , can be found by ‘integrating out’ the other variables  $x_{m+1}, \dots, x_n$  as so:

$$f_{\mathbf{X}_{1:m}}(x_1, \dots, x_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n) dx_n \dots dx_{m+1} \quad (3.1.7)$$

The marginal cumulative distribution is obtained by allowing

$$F_{\mathbf{X}_{1:m}}(x_1, \dots, x_m) = \lim_{x_{m+1}, \dots, x_n \rightarrow \infty} F_{\mathbf{X}}(x_1, \dots, x_n) \quad (3.1.8)$$

## Multivariate Conditional Probability Distributions

If  $\mathbf{X} = (X_1, \dots, X_n)$  is a random vector of continuous variables, then the density for the first  $m < n$  random variables  $X_1, \dots, X_m$  conditioned on the last  $n - m$  random variables taking on values  $X_{m+1} = x_{m+1}, \dots, X_n = x_n$  is obtained analogously to the bivariate case, where we divide by the marginal density:

$$f_{\mathbf{X}_{1:m}|\mathbf{X}_{(m+1):n}}(x_1, \dots, x_m | x_{m+1}, \dots, x_n) = \frac{f_{\mathbf{X}}(x_1, \dots, x_n)}{f_{\mathbf{X}_{(m+1):n}}(x_{m+1}, \dots, x_n)} \quad (3.1.9)$$

This gives the cumulative distribution conditional on  $X_{m+1} = x_{m+1}, \dots, X_n = x_n$  by integration:

$$\begin{aligned} F_{\mathbf{X}_{1:m}|\mathbf{X}_{(m+1):n}}(x_1, \dots, x_m | x_{m+1}, \dots, x_n) \\ = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} f_{\mathbf{X}_{1:m}|\mathbf{X}_{(m+1):n}}(x_1, \dots, x_m | x_{m+1}, \dots, x_n) dx_m \dots dx_1 \end{aligned} \quad (3.1.10)$$

This conditional CDF can also be arrived at using limits. For simplicity, first recall in the bivariate case with a joint distribution  $F_{X_1, X_2}(x_1, x_2)$  over two random variables:

$$F_{X_2|X_1}(x_2|x_1) = \frac{\frac{\partial}{\partial x_1} F_{X_1, X_2}(x_1, x_2)}{f_{X_1}(x_1)} \quad (3.1.11)$$

Generalising this to multivariate distributions:

$$F_{\mathbf{X}_{1:m}}(x_1, \dots, x_m | \mathbf{X}_{(m+1):n} = (x_{m+1}, \dots, x_n)) = \frac{\frac{\partial^{n-m}}{\partial x_{m+1} \dots \partial x_n} F_{\mathbf{X}_{1:n}}(x_1, \dots, x_n)}{f_{\mathbf{X}_{(m+1):n}}(x_{m+1}, \dots, x_n)} \quad (3.1.12)$$

Alternatively, we also have instead the cumulative distribution conditional on  $X_{m+1} \leq x_{m+1}, \dots, X_n \leq x_n$ :

$$F_{\mathbf{X}_{1:m}}(x_1, \dots, x_m | \mathbf{X}_{(m+1):n} \leq (x_{m+1}, \dots, x_n)) = \frac{F_{\mathbf{X}}(x_1, \dots, x_n)}{F_{\mathbf{X}_{(m+1):n}}(x_{m+1}, \dots, x_n)} \quad (3.1.13)$$

which induces a density function conditional on  $X_{m+1} \leq x_{m+1}, \dots, X_n \leq x_n$ :

$$\begin{aligned} f_{\mathbf{X}_{1:m}}(x_1, \dots, x_m | \mathbf{X}_{(m+1):n} \leq (x_{m+1}, \dots, x_n)) \\ = \frac{\partial^m}{\partial x_1 \dots \partial x_m} F_{\mathbf{X}_{1:m}}(x_1, \dots, x_m | \mathbf{X}_{(m+1):n} \leq (x_{m+1}, \dots, x_n)) \end{aligned} \quad (3.1.14)$$

## Differenced Cumulative Distribution Functions

For a bivariate cumulative distribution  $F(x, y)$  for  $(X, Y)$ , consider the probability that  $x_1 < X \leq x_2$  and  $y_1 < Y \leq y_2$ , where  $x_1 \leq x_2$  and  $y_1 \leq y_2$ . If  $(X, Y)$  are continuous with density  $f(x, y)$ , then this probability can be expressed as the integral

$$\Pr(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y) dx dy \quad (3.1.15)$$

which is why we may refer to this as the  $F$ -volume of the rectangle  $[x_1, x_2] \times [y_1, y_2]$ . Based on the same approach used to derive the bivariate joint mass function from the joint CDF, we can alternatively write this  $F$ -volume in terms of the CDF by

$$\Pr(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \quad (3.1.16)$$

This is also known as taking a second order difference, or 2-difference of the CDF [145]. Now consider a trivariate distribution for  $(X, Y, Z)$ . The third order difference is logically given by

$$\Pr(x_1 < X \leq x_2, y_1 < Y \leq y_2, z_1 < Z \leq z_2) = \Pr(X \leq x_2, Y \leq y_2, Z \leq z_2) - \Pr(A \cup B \cup C) \quad (3.1.17)$$

where

$$A = \{X \leq x_2, Y \leq y_2, Z \leq z_1\} \quad (3.1.18)$$

$$B = \{X \leq x_2, Y \leq y_1, Z \leq z_2\} \quad (3.1.19)$$

$$C = \{X \leq x_1, Y \leq y_2, Z \leq z_2\} \quad (3.1.20)$$

We use the inclusion-exclusion principle for the last term, so that

$$\begin{aligned} \Pr(x_1 < X \leq x_2, y_1 < Y \leq y_2, z_1 < Z \leq z_2) &= \Pr(X \leq x_2, Y \leq y_2, Z \leq z_2) \\ &\quad - \Pr(A) - \Pr(B) - \Pr(C) \\ &\quad + \Pr(A \cap B) + \Pr(A \cap C) + \Pr(B \cap C) - \Pr(A \cap B \cap C) \end{aligned} \quad (3.1.21)$$

Note that

$$\Pr(A \cap B) = \Pr(\{X \leq x_2, Y \leq y_2, Z \leq z_1\} \cap \{X \leq x_2, Y \leq y_1, Z \leq z_2\}) \quad (3.1.22)$$

$$= \Pr(X \leq x_2, Y \leq y_1, Z \leq z_1) \quad (3.1.23)$$

$$= F(x_2, y_1, z_1) \quad (3.1.24)$$

since  $y_1 \leq y_2$ ,  $z_1 \leq z_2$ , and analogously for other pairings. Thus

$$\begin{aligned} \Pr(x_1 < X \leq x_2, y_1 < Y \leq y_2, z_1 < Z \leq z_2) &= F(x_2, y_2, z_2) \\ &\quad - F(x_2, y_2, z_1) - F(x_2, y_1, z_2) - F(x_1, y_2, z_2) \\ &\quad + F(x_2, y_1, z_1) + F(x_1, y_2, z_1) + F(x_1, y_1, z_2) - F(x_1, y_1, z_1) \end{aligned} \quad (3.1.25)$$

Further generalising to  $d$ -variate distributions, we can use the inclusion-exclusion principle to obtain the  $d$ -difference. If for a function  $F$ , the  $F$ -volume of all valid rectangles is non-negative, then we say that  $F$  is  $d$ -non-decreasing. Note that this property is different to (i.e. neither implies, or is implied by) the property of being non-decreasing in a single variable. Yet we require a valid CDF  $F$  to be non-decreasing in a single variable and also  $d$ -non-decreasing, since the  $F$ -volumes are probabilities.

## Degenerate Distributions

A distribution of a random vector (with dimension  $n$ ) is said to be degenerate if the dimension of the support is smaller than  $n$ . Instances of degenerate distributions typically arise when the elements of the random vector are constrained in some way (e.g. they all sum to a particular value), so that the random vector actually takes on values from a  $(n - 1)$ -dimensional hyperplane. Examples of such distributions are the [multinomial](#) and [dirichlet](#) distributions. A Dirac delta distribution is also considered to be a degenerate distribution for a univariate random variable.

### 3.1.2 Mean Vectors

The mean vector of a random vector  $\mathbf{X}$  on support  $\mathcal{X}$  is the expectation  $\mathbb{E}[\mathbf{X}]$  (which will be vector-valued), and is found by integrating over  $\mathcal{X}$  using its density function  $f_{\mathbf{X}}(\mathbf{x})$ :

$$\mathbb{E}[\mathbf{X}] = \int_{\mathcal{X}} \mathbf{x} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (3.1.26)$$

if  $\mathbf{X}$  is a continuous random vector, otherwise if  $\mathbf{X}$  is a discrete random vector then we use the summation over  $\mathcal{X}$  and the probability mass function  $\Pr(\mathbf{X} = \mathbf{x})$ :

$$\mathbb{E}[\mathbf{X}] = \sum_{\mathcal{X}} \Pr(\mathbf{X} = \mathbf{x}) \quad (3.1.27)$$

Alternatively, we have the marginal distributions for  $X_1, \dots, X_n$  with expectations  $\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]$  respectively, then the mean vector is just vector of these marginal means:

$$\mathbb{E}[\mathbf{X}] = [\mathbb{E}[X_1] \ \dots \ \mathbb{E}[X_n]]^\top \quad (3.1.28)$$

### Multivariate Conditional Expectations

Suppose  $\mathbf{Z}$  is a random vector that we can partition as  $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$  with  $\mathbf{Y} \in \mathbb{R}^m$ . Then the conditional expectation of  $\mathbf{Y}$  given  $\mathbf{X}$ , written  $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ , is the mean of the [multivariate conditional distribution](#) of  $\mathbf{Y}$  given  $\mathbf{X}$ . If  $\mathbf{Z}$  is continuous, then  $\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}]$  can be computed by integrating over the conditional density by

$$\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \int_{\mathbb{R}^m} \mathbf{y} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \quad (3.1.29)$$

### 3.1.3 Covariance Matrices

The covariance matrix  $\mathbf{C} = \text{Cov}(\mathbf{X})$  of a random vector  $\mathbf{X}$  generalises the notion of variance and covariance to random vectors and can be defined by

$$\mathbf{C} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \quad (3.1.30)$$

If we designate  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ , then the elements of the covariance matrix will happen to be

$$\mathbf{C} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & & \text{Cov}(X_2, X_n) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{bmatrix} \quad (3.1.31)$$

Note that the covariance matrix will be symmetric since  $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$ . Covariance matrices of real random vectors are also always positive semi-definite. We can show this by considering an arbitrary real vector  $\mathbf{u}$ . We have

$$\mathbf{u}^\top \mathbf{C} \mathbf{u} = \mathbf{u}^\top \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \mathbf{u} \quad (3.1.32)$$

$$= \mathbb{E} \left[ \mathbf{u}^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}]) (\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{u} \right] \quad (3.1.33)$$

Let  $Y := \mathbf{u}^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}])$  be a zero-mean scalar random variable with variance  $\sigma^2 \geq 0$ . Hence

$$\mathbf{u}^\top \mathbf{C} \mathbf{u} = \mathbb{E} \left[ YY^\top \right] \quad (3.1.34)$$

$$= \mathbb{E} \left[ (Y - \mathbb{E}[Y])^2 \right] \quad (3.1.35)$$

$$= \mathbb{E} [Y^2] \quad (3.1.36)$$

$$= \sigma^2 \quad (3.1.37)$$

$$\geq 0 \quad (3.1.38)$$

Therefore  $\mathbf{C}$  satisfies the property of a positive semi-definite matrix.

An alternative formula for the covariance exists, analogous to the variance:

$$\text{Cov}(\mathbf{X}) = \mathbb{E} \left[ (\mathbf{X} - \mathbb{E}[\mathbf{X}]) (\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \right] \quad (3.1.39)$$

$$= \mathbb{E} \left[ \mathbf{X} \mathbf{X}^\top - \mathbf{X} \mathbb{E}[\mathbf{X}]^\top - \mathbb{E}[\mathbf{X}] \mathbf{X}^\top + \mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{X}]^\top \right] \quad (3.1.40)$$

$$= \mathbb{E} [\mathbf{X} \mathbf{X}^\top] - \mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{X}]^\top - \mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{X}]^\top + \mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{X}]^\top \quad (3.1.41)$$

$$= \mathbb{E} [\mathbf{X} \mathbf{X}^\top] - \mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{X}]^\top \quad (3.1.42)$$

## Precision Matrix

The precision matrix of a random vector  $\mathbf{X}$  is defined as the inverse of the covariance matrix, i.e.  $\mathbf{C}^{-1}$ .

## Scalarised Variance of Random Vectors

Another generalisation for the definition of the variance  $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$  to random vectors  $\mathbf{X} = (X_1, \dots, X_n)$  is

$$V(\mathbf{X}) = \mathbb{E} \left[ \|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^2 \right] \quad (3.1.43)$$

which is a scalar term. This can be shown to be equal to the trace of the covariance matrix  $\mathbf{C} = \text{Cov}(\mathbf{X})$  as follows:

$$\mathbb{E} \left[ \|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^2 \right] = \mathbb{E} [(\mathbf{X} - \mathbb{E}[\mathbf{X}]) \cdot (\mathbf{X} - \mathbb{E}[\mathbf{X}])] \quad (3.1.44)$$

$$= \mathbb{E} \left[ \sum_{i=1}^n (X_i - \mathbb{E}[X_i])^2 \right] \quad (3.1.45)$$

$$= \sum_{i=1}^n \mathbb{E} [(X_i - \mathbb{E}[X_i])^2] \quad (3.1.46)$$

$$= \sum_{i=1}^n \text{Var}(X_i) \quad (3.1.47)$$

$$= \text{trace}(\mathbf{C}) \quad (3.1.48)$$

Note that we have  $\sum_{i=1}^n \text{Var}(X_i) = \text{trace}(\mathbf{C})$  because the diagonals of  $\mathbf{C}$  contain the variances of each random element.

### Cross-Covariance Matrices

Suppose we have random vectors  $\mathbf{X} = (X_1, \dots, X_n)$  and  $\mathbf{Y} = (Y_1, \dots, Y_m)$ . Then we can define the cross-covariance between  $\mathbf{X}$  and  $\mathbf{Y}$  as either a  $n \times m$  or  $m \times n$  matrix. Specifically,

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \begin{bmatrix} \text{Cov}(X_1, Y_1) & \dots & \text{Cov}(X_1, Y_m) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, Y_1) & \dots & \text{Cov}(X_n, Y_m) \end{bmatrix} \quad (3.1.49)$$

or

$$\text{Cov}(\mathbf{Y}, \mathbf{X}) = \begin{bmatrix} \text{Cov}(Y_1, X_1) & \dots & \text{Cov}(Y_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(Y_m, X_1) & \dots & \text{Cov}(Y_m, X_n) \end{bmatrix} \quad (3.1.50)$$

This can also be written in terms of expectations as

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top] \quad (3.1.51)$$

$$= \mathbb{E}[\mathbf{XY}^\top] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]^\top \quad (3.1.52)$$

Note that the cross-covariance between a random vector and itself simply equals its covariance matrix:

$$\text{Cov}(\mathbf{X}, \mathbf{X}) = \text{Cov}(\mathbf{X}) \quad (3.1.53)$$

and that the transpose of the cross-covariance is simply the cross-covariance with the ordering reversed:

$$\text{Cov}(\mathbf{X}, \mathbf{Y})^\top = \text{Cov}(\mathbf{Y}, \mathbf{X}) \quad (3.1.54)$$

### Covariance of Sums of Random Vectors

If  $\mathbf{X}$  and  $\mathbf{Y}$  are random vectors of the same dimension, then

$$\text{Cov}(\mathbf{X} + \mathbf{Y}) = \text{Cov}(\mathbf{X}) + \text{Cov}(\mathbf{X}, \mathbf{Y}) + \text{Cov}(\mathbf{Y}, \mathbf{X}) + \text{Cov}(\mathbf{Y}) \quad (3.1.55)$$

This can be shown using the definition of the covariances and the cross-covariances, as well as the linearity of expectation:

$$\text{Cov}(\mathbf{X} + \mathbf{Y}) = \mathbb{E}[(X + Y - \mathbb{E}[X + Y])(X + Y - \mathbb{E}[X + Y])^\top] \quad (3.1.56)$$

$$= \mathbb{E}[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^\top] \quad (3.1.57)$$

$$= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] + \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top] \\ + \mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^\top] + \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^\top] \quad (3.1.58)$$

$$= \text{Cov}(\mathbf{X}) + \text{Cov}(\mathbf{X}, \mathbf{Y}) + \text{Cov}(\mathbf{Y}, \mathbf{X}) + \text{Cov}(\mathbf{Y}) \quad (3.1.59)$$

### Cross-Covariance of Linear Transformations

More general than covariance of sums of random vectors, we consider the cross-covariance between  $a\mathbf{X} + b\mathbf{Y}$  of dimension  $m$  and  $c\mathbf{W} + d\mathbf{Z}$  of dimension  $n$  for scalar constants  $a, b, c$  and  $d$ . A relation analogous to the covariance between scalar random variables can be shown. Firstly,

$$\text{Cov}(a\mathbf{X} + b\mathbf{Y}, c\mathbf{W} + d\mathbf{Z})$$

$$= \begin{bmatrix} \text{Cov}(aX_1 + bY_1, cW_1 + dZ_1) & \dots & \text{Cov}(aX_1 + bY_1, cW_n + dZ_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(aX_m + bY_m, cW_1 + dZ_1) & \dots & \text{Cov}(aX_m + bY_m, cW_n + dZ_n) \end{bmatrix} \quad (3.1.60)$$

Splitting this up:

$$\begin{aligned} \text{Cov}(a\mathbf{X} + b\mathbf{Y}, c\mathbf{W} + d\mathbf{Z}) &= ac \begin{bmatrix} \text{Cov}(X_1, W_1) & \dots & \text{Cov}(X_1, W_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_m, W_1) & \dots & \text{Cov}(X_m, W_n) \end{bmatrix} \\ &+ ad \begin{bmatrix} \text{Cov}(X_1, Z_1) & \dots & \text{Cov}(X_1, Z_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_m, Z_1) & \dots & \text{Cov}(X_m, Z_n) \end{bmatrix} + bc \begin{bmatrix} \text{Cov}(Y_1, W_1) & \dots & \text{Cov}(Y_1, W_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(Y_m, W_1) & \dots & \text{Cov}(Y_m, W_n) \end{bmatrix} \\ &\quad + bd \begin{bmatrix} \text{Cov}(Y_1, Z_1) & \dots & \text{Cov}(Y_1, Z_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(Y_m, Z_1) & \dots & \text{Cov}(Y_m, Z_n) \end{bmatrix} \end{aligned} \quad (3.1.61)$$

Hence

$$\text{Cov}(a\mathbf{X} + b\mathbf{Y}, c\mathbf{W} + d\mathbf{Z}) = ac \text{Cov}(\mathbf{X}, \mathbf{W}) + ad \text{Cov}(\mathbf{X}, \mathbf{Z}) + bc \text{Cov}(\mathbf{Y}, \mathbf{W}) + bd \text{Cov}(\mathbf{Y}, \mathbf{Z}) \quad (3.1.62)$$

### Conditional Covariance Matrices

The conditional covariance matrix of  $\mathbf{Y}$  given  $\mathbf{X}$  can be defined analogously to the covariance matrix, except using conditional expectations.

$$\text{Cov}(\mathbf{Y}|\mathbf{X}) = \mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}|\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}|\mathbf{X}])^\top | \mathbf{X}] \quad (3.1.63)$$

### Conditional Cross-Covariance Matrices

The conditional cross-covariance between  $\mathbf{X}$  and  $\mathbf{Y}$  given  $\mathbf{Z}$  can be defined analogously to the cross-covariance matrix, except using conditional expectations.

$$\text{Cov}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Z}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}|\mathbf{Z}])^\top | \mathbf{Z}] \quad (3.1.64)$$

#### 3.1.4 Independence of Random Vectors

To generalise independence between random variables to independence between random vectors, we may say that random vectors  $\mathbf{X} = (X_1, \dots, X_n)$  and  $\mathbf{Y} = (Y_1, \dots, Y_m)$  are independent if their cumulative distribution functions satisfy

$$\begin{aligned} \Pr(X_1 \leq x_1, \dots, X_n \leq x_n, Y_1 \leq y_1, \dots, Y_m \leq y_m) \\ = \Pr(X_1 \leq x_1, \dots, X_n \leq x_n) \Pr(Y_1 \leq y_1, \dots, Y_m \leq y_m) \end{aligned} \quad (3.1.65)$$

for any  $(x_1, \dots, x_n, y_1, \dots, y_m)$ . This implies that marginalised distributions satisfy the same multiplicative property, for example we have that:

$$\Pr(X_1 \leq x_1, X_n \leq x_n, Y_1 \leq y_1, Y_m \leq y_m) = \Pr(X_1 \leq x_1, X_n \leq x_n) \Pr(Y_1 \leq y_1, Y_m \leq y_m) \quad (3.1.66)$$

for any  $(x_1, x_n, y_1, y_m)$ . It follows that if  $\mathbf{X}$  and  $\mathbf{Y}$  are independent random variables, then any pair  $X_i$  and  $Y_j$  are independent random variables, and also that

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}_{n \times m} \quad (3.1.67)$$

provided the covariance exists.

### 3.1.5 Transformations of Random Vectors

#### Linear Transformations of Random Vectors

Suppose  $\mathbf{X}$  is a continuous random vector with PDF  $f_{\mathbf{X}}(\mathbf{x})$  and the linear transformation  $\mathbf{A}$  is an invertible matrix. Then the PDF of  $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$  is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det(\mathbf{A})|} f_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})) \quad (3.1.68)$$

*Proof.* Let  $B$  be defined by the set  $B = \{\mathbf{y} : \mathbf{y} \leq \mathbf{y}'\}$  for some  $\mathbf{y}'$ , so the definition of the CDF can be written as  $F_{\mathbf{Y}}(\mathbf{y}') = \int_B f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}$ . Define the transformation  $\mathbf{x} = T(\mathbf{y}) = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$  (which is the inverse of the transformation of  $\mathbf{X}$  to  $\mathbf{Y}$ ). So

$$T(B) = \{\mathbf{x} : \mathbf{Ax} + \mathbf{b} \leq \mathbf{y}'\} \quad (3.1.69)$$

is the image of  $B$  under the transformation  $T$ . So the event  $\mathbf{Y} \in B$  can be deemed the same event as  $\mathbf{X} \in T(B)$ . Hence

$$F_{\mathbf{Y}}(\mathbf{y}') = \Pr(\mathbf{Y} \leq \mathbf{y}') \quad (3.1.70)$$

$$= \Pr(\mathbf{Y} \in B) \quad (3.1.71)$$

$$= \Pr(\mathbf{X} \in T(B)) \quad (3.1.72)$$

$$= \int_{T(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (3.1.73)$$

By a change of variables  $\mathbf{x} = T(\mathbf{y}) = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$ , this becomes

$$F_{\mathbf{Y}}(\mathbf{y}') = \int_B f_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})) |\det(\mathbf{A}^{-1})| d\mathbf{y} \quad (3.1.74)$$

where we note that  $d\mathbf{x} = |\det(\mathbf{A}^{-1})| d\mathbf{y}$  which can be rewritten as

$$dx_1 \dots dx_n = |\det(\mathbf{A}^{-1})| dy_1 \dots dy_n \quad (3.1.75)$$

due to the characterisation of the determinant as the hypervolume scale factor under a linear transformation. So here  $|\det(\mathbf{A}^{-1})|$  is the scale factor of the hypervolume element  $dy_1 \dots dy_n$ . Differentiating the CDF, we can see that

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})) |\det(\mathbf{A}^{-1})| \quad (3.1.76)$$

Then applying the fact that  $|\det(\mathbf{A}^{-1})| = |\det(\mathbf{A})|^{-1}$  gives

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det(\mathbf{A})|} f_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})) \quad (3.1.77)$$

as required. Here,  $|\det(\mathbf{A}^{-1})|$  can also be interpreted as acting as the normalising constant.  $\square$

For a linearly transformed random vector  $A\mathbf{X} + b$ , the expectation and covariance in terms of the expectation and covariance  $\mathbf{X}$  (assuming these exist) is given by

$$\mathbb{E}[A\mathbf{X}] = A\mathbb{E}[\mathbf{X}] \quad (3.1.78)$$

$$\text{Cov}(A\mathbf{X}) = A \text{Cov}(\mathbf{X}) A^\top \quad (3.1.79)$$

*Proof.* The relationship  $\mathbb{E}[A\mathbf{X}] = A\mathbb{E}[\mathbf{X}]$  follows from the linearity of expectation. As for  $\text{Cov}(A\mathbf{X})$ , we simply use the definition of the covariance matrix:

$$\text{Cov}(A\mathbf{X}) = \mathbb{E}[(A\mathbf{X} - \mathbb{E}[A\mathbf{X}])(A\mathbf{X} - \mathbb{E}[A\mathbf{X}])^\top] \quad (3.1.80)$$

$$= \mathbb{E}[A(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top A^\top] \quad (3.1.81)$$

$$= A\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] A^\top \quad (3.1.82)$$

$$= A \text{Cov}(\mathbf{X}) A^\top \quad (3.1.83)$$

□

We can also show for the cross-covariance between two linearly transformed random vectors:

$$\text{Cov}(A\mathbf{X}, B\mathbf{Y}) = \mathbb{E}[(A\mathbf{X} - \mathbb{E}[A\mathbf{X}])(B\mathbf{Y} - \mathbb{E}[B\mathbf{Y}])^\top] \quad (3.1.84)$$

$$= \mathbb{E}[A(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top B^\top] \quad (3.1.85)$$

$$= A\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top] B^\top \quad (3.1.86)$$

$$= A \text{Cov}(\mathbf{X}, \mathbf{Y}) B^\top \quad (3.1.87)$$

### Whitening Transformations

A whitening transformation is a special case of a linear transformation which attempts to transform a random vector such that it will have a covariance matrix equal to the identity matrix. Suppose random vector  $\mathbf{X}$  has  $\text{Cov}(\mathbf{X}) = \Sigma$  where  $\Sigma$  is positive definite (i.e. full rank). There is more than one possible whitening transformation, as there different ways to decompose a positive semidefinite matrix into the multiplication of some matrix with its transpose. One such is by eigendecomposition:

$$\Sigma = V\Lambda V^\top \quad (3.1.88)$$

$$= V\Lambda^{1/2}\Lambda^{1/2}V^\top \quad (3.1.89)$$

where  $V$  is orthogonal (i.e.  $V^\top = V^{-1}$ ) and  $\Lambda^{1/2}$  is a diagonal matrix consisting of the square root of the eigenvalues of  $\Sigma$ . Thus if we choose the whitening transformation as  $\mathbf{Y} = \Lambda^{-1/2}V^\top \mathbf{X}$ , then

$$\text{Cov}(\mathbf{Y}) = \Lambda^{-1/2}V^\top \text{Cov}(\mathbf{X}) V\Lambda^{-1/2} \quad (3.1.90)$$

$$= \Lambda^{-1/2}V^\top V\Lambda^{1/2}\Lambda^{1/2}V^\top V\Lambda^{-1/2} \quad (3.1.91)$$

$$= I \quad (3.1.92)$$

Another possibility is with the Cholesky decomposition  $\Sigma = LL^\top$  and the transformation  $\mathbf{Y} = L^{-1}\mathbf{X}$  so that

$$\text{Cov}(\mathbf{Y}) = L^{-1} \text{Cov}(\mathbf{X}) (L^{-1})^\top \quad (3.1.93)$$

$$= L^{-1}LL^\top (L^{-1})^\top \quad (3.1.94)$$

$$= I \quad (3.1.95)$$

Suppose now that  $\Sigma$  is positive semi-definite (i.e. not full rank), so we cannot take the inverse  $\Lambda^{-1/2}$  or the Cholesky decomposition. The best we can do is find the transformation  $\mathbf{Y}$  such that  $\text{Cov}(\mathbf{Y}) = \text{diag}\{1, \dots, 1, 0, \dots, 0\}$ , where the number of ones is equal to the rank of  $\Sigma$ , denoted by  $\text{rank}(\Sigma) = m$ . One possibility is to again use eigendecomposition

$$\Sigma = UDU^\top \quad (3.1.96)$$

$$= UD^{1/2}D^{1/2}U^\top \quad (3.1.97)$$

where in terms of the non-zero eigenvalues  $d_1, \dots, d_m$ :

$$D^{1/2} = \text{diag} \left\{ d_1^{1/2}, \dots, d_m^{1/2}, 0, \dots, 0 \right\} \quad (3.1.98)$$

Choose the transformation  $\mathbf{Y} = (D^{1/2})^\dagger U^\top \mathbf{X}$ , where  $(D^{1/2})^\dagger$  is the Moore-Penrose pseudoinverse of  $D^{1/2}$ , which is given by

$$(D^{1/2})^\dagger = \text{diag} \left\{ d_1^{-1/2}, \dots, d_m^{-1/2}, 0, \dots, 0 \right\} \quad (3.1.99)$$

This may be verified against the definition of the Moore-Penrose pseudoinverse. Moreover, we have

$$D^{1/2} (D^{1/2})^\dagger = \text{diag} \{1, \dots, 1, 0, \dots, 0\} \quad (3.1.100)$$

hence

$$\text{Cov}(\mathbf{Y}) = (D^{1/2})^\dagger U^\top \text{Cov}(\mathbf{X}) U (D^{1/2})^\dagger \quad (3.1.101)$$

$$= (D^{1/2})^\dagger U^\top U D^{1/2} D^{1/2} U^\top U (D^{1/2})^\dagger \quad (3.1.102)$$

$$= \text{diag} \{1, \dots, 1, 0, \dots, 0\} \quad (3.1.103)$$

### Invertible Transformations of Random Vectors

Suppose  $\mathbf{X}$  is a continuous random vector with PDF  $f_{\mathbf{X}}(\mathbf{x})$  and random vector  $\mathbf{Y}$  is related to  $\mathbf{X}$  via the invertible transformation  $\mathbf{Y} = T(\mathbf{X})$ , such that  $\mathbf{X} = T^{-1}(\mathbf{Y})$ . We can derive the density  $f_{\mathbf{Y}}(\mathbf{y})$  in terms of  $f_{\mathbf{X}}(\cdot)$ . The derivation follows in a similar manner to the case with linear transformations, with some slight generalisations. We define  $B$  by the set

$$B = \{\mathbf{y} : \mathbf{y} \leq \mathbf{y}'\} \quad (3.1.104)$$

so that the image of  $B$  under the mapping  $T^{-1}(\cdot)$  is given by

$$T^{-1}(B) = \{\mathbf{x} : T(\mathbf{x}) \leq \mathbf{y}'\} \quad (3.1.105)$$

Hence the CDF of  $\mathbf{Y}$  can be expressed as

$$F_{\mathbf{Y}}(\mathbf{y}') = \Pr(\mathbf{Y} \leq \mathbf{y}') \quad (3.1.106)$$

$$= \Pr(\mathbf{Y} \in B) \quad (3.1.107)$$

$$= \Pr(\mathbf{X} \in T^{-1}(B)) \quad (3.1.108)$$

$$= \int_{T^{-1}(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (3.1.109)$$

$$= \int_B f_{\mathbf{X}}(T^{-1}(\mathbf{y})) \left| \det \left( \frac{\partial T^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| d\mathbf{y} \quad (3.1.110)$$

$$= \int_{\{\mathbf{y}: \mathbf{y} \leq \mathbf{y}'\}} f_{\mathbf{X}}(T^{-1}(\mathbf{y})) \left| \det \left( \frac{\partial T^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| d\mathbf{y} \quad (3.1.111)$$

where we made a change of variables  $\mathbf{x} = T^{-1}(\mathbf{y})$  in the integral. Note that the volume element  $d\mathbf{y}$  is related to the volume element  $d\mathbf{x}$  by

$$d\mathbf{y} = \left| \det \left( \frac{\partial T(\mathbf{x})}{\partial \mathbf{x}} \right) \right| d\mathbf{x} \quad (3.1.112)$$

because the local volume scale factor is given by the Jacobian determinant of the transformation  $T(\cdot)$ . In the reverse direction, we have

$$d\mathbf{x} = \left| \det \left( \frac{\partial T^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| d\mathbf{y} \quad (3.1.113)$$

Hence, differentiating the CDF gives the PDF as

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(T^{-1}(\mathbf{y})) \left| \det \left( \frac{\partial T^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| \quad (3.1.114)$$

### Expectations of Dot Products

For a random vector  $\mathbf{X}$  with mean  $\boldsymbol{\mu}$  and covariance  $\Sigma$ , we consider the dot product  $\mathbf{X}^\top \mathbf{X}$  and particularly its expectation  $\mathbb{E}[\mathbf{X}^\top \mathbf{X}]$ . As it was shown in the scalarised variance of a random vector,  $\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})^\top (\mathbf{X} - \boldsymbol{\mu})] = \text{trace}(\Sigma)$ . Then from this:

$$\text{trace}(\Sigma) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})^\top (\mathbf{X} - \boldsymbol{\mu})] \quad (3.1.115)$$

$$= \mathbb{E}[\mathbf{X}^\top \mathbf{X} - \boldsymbol{\mu}^\top \mathbf{X} - \mathbf{X}^\top \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\mu}] \quad (3.1.116)$$

$$= \mathbb{E}[\mathbf{X}^\top \mathbf{X}] - \boldsymbol{\mu}^\top \boldsymbol{\mu} \quad (3.1.117)$$

Hence

$$\mathbb{E}[\mathbf{X}^\top \mathbf{X}] = \text{trace}(\Sigma) + \boldsymbol{\mu}^\top \boldsymbol{\mu} \quad (3.1.118)$$

This can be generalised to finding  $\mathbb{E}[\mathbf{X}^\top A \mathbf{X}]$  where  $A$  is symmetric positive definite. Then  $A$  has a unique symmetric positive definite square root  $A^{1/2}$ , and we can consider the expectation of  $\mathbb{E}[(A\mathbf{X})^\top (A^{1/2}\mathbf{X})]$ . The random vector  $A^{1/2}\mathbf{X}$  will have expectation  $A^{1/2}\boldsymbol{\mu}$  and covariance  $A^{1/2}\Sigma A^{1/2}$ , so

$$\text{trace}(\text{Cov}(A^{1/2}\mathbf{X})) = \text{trace}(A^{1/2}\Sigma A^{1/2}) \quad (3.1.119)$$

$$= \text{trace}(A^{1/2}A^{1/2}\Sigma) \quad (3.1.120)$$

$$= \text{trace}(A\Sigma) \quad (3.1.121)$$

due to the invariance of the trace to cyclic permutations. Hence

$$\mathbb{E}[\mathbf{X}^\top A \mathbf{X}] = \mathbb{E}[(A\mathbf{X})^\top (A^{1/2}\mathbf{X})] \quad (3.1.122)$$

$$= \text{trace}(A\Sigma) + (A^{1/2}\boldsymbol{\mu})^\top A^{1/2}\boldsymbol{\mu} \quad (3.1.123)$$

$$= \text{trace}(A\Sigma) + \boldsymbol{\mu}^\top A\boldsymbol{\mu} \quad (3.1.124)$$

### Rosenblatt Transformation [94]

The Rosenblatt transformation can be used as a way to transform a continuous random vector  $\mathbf{X} = (X_1, \dots, X_d)$  into a vector of i.i.d. Uniform(0, 1) random variables  $\mathbf{U} = (U_1, \dots, U_d)$ . It is akin to the **probability integral transform**, extended to multiple dimensions. If we let  $F_{\mathbf{X}}(x_1, \dots, x_d)$  be the CDF of  $\mathbf{X}$  with the conditional factorisation:

$$F_{\mathbf{X}}(x_1, \dots, x_d) = F_{X_1}(x_1) F_{X_2|X_1}(x_2|x_1) \times \dots \times F_{X_d|X_1, \dots, X_{d-1}}(x_d|x_1, \dots, x_{d-1}) \quad (3.1.125)$$

Then  $\mathbf{U}$  may be generated by

$$U_1 = F_{X_1}(X_1) \quad (3.1.126)$$

$$U_2 = F_{X_2|X_1}(X_2|X_1) \quad (3.1.127)$$

$$\vdots \quad (3.1.128)$$

$$U_d = F_{X_d|X_1, \dots, X_{d-1}}(X_d|X_1, \dots, X_{d-1}) \quad (3.1.129)$$

To show why this works, firstly for  $U_1$  we have

$$\Pr(U_1 \leq u_1) = \Pr(F_{X_1}(X_1) \leq u_1) \quad (3.1.130)$$

$$= \Pr(X_1 \leq F_{X_1}^{-1}(u_1)) \quad (3.1.131)$$

$$= F_{X_1}(F_{X_1}^{-1}(u_1)) \quad (3.1.132)$$

$$= u_1 \quad (3.1.133)$$

which is the CDF of the Uniform(0, 1) random variable. As for the second random variate conditional on the first, we have

$$\Pr(U_2 \leq u_2|U_1 = u_1) = \Pr\left(F_{X_2|X_1}\left(X_2|F_{X_1}^{-1}(u_1)\right) \leq u_2|U_1 = u_1\right) \quad (3.1.134)$$

$$= \Pr\left(X_2 \leq F_{X_2|X_1}^{-1}\left(u_2|F_{X_1}^{-1}(u_1)\right)|U_1 = u_1\right) \quad (3.1.135)$$

$$= F_{X_2|X_1}\left(F_{X_2|X_1}^{-1}\left(u_2|F_{X_1}^{-1}(u_1)\right)|F_{X_1}^{-1}(u_1)\right) \quad (3.1.136)$$

$$= u_2 \quad (3.1.137)$$

since  $F_{X_2|X_1}(\cdot|F_{X_1}^{-1}(u_1))$  can be treated as just a CDF with the same properties as any other univariate continuous CDF. This implies that  $U_1$  and  $U_2$  are independent, since the conditional CDF of  $U_2$  is not a function of  $u_1$ . Hence we can say  $\Pr(U_2 \leq u_2|U_1 = u_1) = \Pr(U_2 \leq u_2|U_1 \leq u_1)$ . Iterating this in a similar fashion for the other random variates, we can show

$$\Pr(U_1 \leq u_1, \dots, U_d \leq u_d) = \Pr(U_1 \leq u_1) \Pr(U_2 \leq u_2|U_1 \leq u_1) \dots \Pr(U_d \leq u_d|U_1 \leq u_1, U_{d-1} \leq u_{d-1}) \quad (3.1.138)$$

$$= \Pr(U_1 \leq u_1) \Pr(U_2 \leq u_2|U_1 = u_1) \dots \Pr(U_d \leq u_d|U_1 = u_1, U_{d-1} = u_{d-1}) \quad (3.1.139)$$

$$= u_1 u_2 \dots u_d \quad (3.1.140)$$

which is the CDF of the  $d$ -variate uniform distribution on the unit hypercube  $[0, 1]^d$ . Thus  $U_1, \dots, U_d$  are mutually independent Uniform(0, 1) random variables.

### Inverse Rosenblatt Transform

By inverting the process of the [Rosenblatt transformation](#), samples from arbitrary multivariate distributions can be generated (provided the conditional quantile functions are attainable). This generalises the approach of [inverse transform sampling](#) to multivariate distributions. Starting from an i.i.d. Uniform(0, 1) random vector  $(U_1, \dots, U_d)$ , we can generate a sample of a random vector  $\mathbf{X} = (X_1, \dots, X_d)$  by

$$X_1 = F_{X_1}^{-1}(U_1) \quad (3.1.141)$$

$$X_2 = F_{X_2|X_1}^{-1}(U_2|X_1) \quad (3.1.142)$$

$$\vdots \quad (3.1.143)$$

$$X_d = F_{X_d|X_1, \dots, X_{d-1}}^{-1}(U_d|X_1, \dots, X_{d-1}) \quad (3.1.144)$$

where the conditional quantile functions  $F_{X_2|X_1}^{-1}(\cdot|x_1)$ , etc. denote the quantile functions obtained from the respective conditional CDFs (i.e. from  $F_{X_2|X_1}(x_2|x_1)$ , etc.).

## 3.2 Families of Multivariate Probability Distributions

### 3.2.1 Multivariate Gaussian Distribution

#### Multivariate Gaussian Integrals

A multivariate extension of Gaussian integrals is required to evaluate integrals of the form  $\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{x}^\top A\mathbf{x}\right) d\mathbf{x}$  where  $\mathbf{x}$  is  $n$ -dimensional and  $A$  is an  $n \times n$  symmetric positive semi-definite matrix. To obtain the form of this integral, we first evaluate the easier integral:

$$\int_{-\infty}^{\infty} \exp\left(-\mathbf{x}^\top \mathbf{x}\right) d\mathbf{x} = \int_{-\infty}^{\infty} \exp\left(-\sum_{i=1}^n x_i^2\right) d\mathbf{x} \quad (3.2.1)$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^n \exp(-x_i^2) d\mathbf{x} \quad (3.2.2)$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(-x_1^2) \cdots \exp(-x_n^2) dx_1 \cdots dx_n \quad (3.2.3)$$

$$= \int_{-\infty}^{\infty} \exp(-x_n^2) \cdots \int_{-\infty}^{\infty} \exp(-x_1^2) dx_1 \cdots dx_n \quad (3.2.4)$$

$$= \int_{-\infty}^{\infty} \exp(-x_n^2) \cdots \int_{-\infty}^{\infty} \exp(-x_2^2) \sqrt{\pi} dx_2 \cdots dx_n \quad (3.2.5)$$

$$= \sqrt{\pi^n} \quad (3.2.6)$$

where we recall for the univariate Gaussian integral  $\int_{-\infty}^{\infty} \exp(-x^2) dx = \sqrt{\pi}$ . Then because by a change of variables  $\int_{-\infty}^{\infty} \exp(-x^2/2) dx = \sqrt{2\pi}$ , then it follows that

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{x}\right) d\mathbf{x} = \sqrt{2\pi^n} \quad (3.2.7)$$

Because  $A$  is positive semi-definite, it has a unique positive semi-definite square root  $B = B^\top$ , and we may write the multivariate Gaussian integral of interest as

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{x}^\top A\mathbf{x}\right) d\mathbf{x} = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{x}^\top BB\mathbf{x}\right) d\mathbf{x} \quad (3.2.8)$$

Apply the change of variables  $\mathbf{z} = B\mathbf{x}$ ,  $d\mathbf{z} = \det(B) d\mathbf{x}$  so that

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{x}^\top A\mathbf{x}\right) d\mathbf{x} = \frac{1}{\det(B)} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{z}^\top \mathbf{z}\right) d\mathbf{z} \quad (3.2.9)$$

$$= \frac{\sqrt{2\pi^n}}{\det(B)} \quad (3.2.10)$$

Note that  $\det(A) = \det(BB) = \det(B)\det(B) = \sqrt{\det(A)}$ . Hence

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{x}^\top A\mathbf{x}\right) d\mathbf{x} = \sqrt{\frac{2\pi^n}{\det(A)}} \quad (3.2.11)$$

and applying the scaling and reciprocal properties of the determinant, this can be rewritten as

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\mathbf{x}^\top A\mathbf{x}\right) d\mathbf{x} = \sqrt{\det(2\pi A^{-1})} \quad (3.2.12)$$

Even more generally, and since this integral is invariant to a translation in  $\mathbf{x}$ , we have

$$\int_{-\infty}^{\infty} a \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{c})^\top A(\mathbf{x} - \mathbf{c})\right] d\mathbf{x} = a\sqrt{\det(2\pi A^{-1})} \quad (3.2.13)$$

This result is used to establish the normalising constant of a multivariate Gaussian density.

### 3.2.2 Multinomial Distribution

The multinomial distribution generalises both the binomial distribution and the categorical distribution. It characterises the probability for the number of counts in each category in  $n$  independent draws from  $K$  categories. This gives us a  $K$ -variate discrete distribution with parameters  $n, p_1, \dots, p_K$ , and by a straightforward extension in the way the binomial distribution is derived (using the multinomial coefficient), the probability mass function of the multinomial distribution is

$$\Pr(X_1 = x_1, \dots, X_K = x_k) = \frac{n!}{x_1! \times \dots \times x_K!} p_1^{x_1} \times \dots \times p_K^{x_K} \quad (3.2.14)$$

$$= \binom{n}{x_1, \dots, x_K} \prod_{i=1}^K p_i^{x_i} \quad (3.2.15)$$

which is supported on the set where each  $x_i \in \{0, \dots, n\}$ , and additionally  $\sum_{i=1}^K x_i = n$ . If  $n = 1$ , then it reduces to a categorical distribution or if  $K = 2$ , then it reduces to a binomial distribution. Moreover, marginal distributions are binomial distributed.

#### Covariance of Multinomial Distribution

We derive the covariance between  $X_i$  and  $X_j$  for any  $i \neq j$ . The variance of  $X_i$  is  $\text{Var}(X_i) = np_i(1 - p_i)$  because the computation reduces to the case of the binomial distribution. For off-diagonal elements of the covariance matrix, first consider the case with  $n = 1$ , and use indicator variables  $\mathbb{I}_i$  and  $\mathbb{I}_j$ . Then we see that  $\mathbb{E}[\mathbb{I}_i \mathbb{I}_j] = 0$  since  $\mathbb{I}_i$  and  $\mathbb{I}_j$  cannot both be equal to one. Hence

$$\text{Cov}(\mathbb{I}_i, \mathbb{I}_j) = \mathbb{E}[\mathbb{I}_i \mathbb{I}_j] - \mathbb{E}[\mathbb{I}_i] \mathbb{E}[\mathbb{I}_j] \quad (3.2.16)$$

$$= -p_i p_j \quad (3.2.17)$$

Extending to the case where  $n \geq 1$ , since each of the draws is independent, then

$$\text{Cov}(X_i, X_j) = \sum_{k=1}^n \text{Cov}(\mathbb{I}_i, \mathbb{I}_j) \quad (3.2.18)$$

$$= -np_i p_j \quad (3.2.19)$$

The intuition behind why the covariance is negative is that if category  $i$  is drawn, this prevents category  $j$  from being drawn, and vice-versa. The full covariance structure is

$$\text{Cov}(\mathbf{X}) = n \begin{bmatrix} p_1(1 - p_1) & \dots & -p_1 p_K \\ \vdots & \ddots & \vdots \\ -p_K p_1 & \dots & p_K(1 - p_K) \end{bmatrix} \quad (3.2.20)$$

It can be shown that the covariance matrix is not full rank (and moreover, has rank  $K - 1$ ). Let  $\mathbf{1}$  denote a vector of ones, then

$$\text{Cov}(\mathbf{X}) \mathbf{1} = n \begin{bmatrix} p_1(1 - p_1 - \dots - p_K) \\ \vdots \\ p_K(1 - p_1 - \dots - p_K) \end{bmatrix} \quad (3.2.21)$$

$$= \mathbf{0} \quad (3.2.22)$$

which satisfies the definition of matrix singularity, since there exists the non-zero vector  $\mathbf{1}$  such that  $\text{Cov}(\mathbf{X}) \mathbf{1} = \mathbf{0}$ . This rank deficiency property is intuitive, if one reasons that since the

multinomial distribution is overparametrised (i.e.  $\sum_{i=1}^K p_i = 1$  so any one  $p_i$  can be determined from the rest), and moreover  $\sum_{i=1}^K X_i = n$ , then there would be some ‘redundant’ elements in the covariance matrix. Introduce  $\mathbf{p} = [p_1 \ \dots \ p_K]^\top$ , then the covariance matrix can also be expressed as

$$\text{Cov}(\mathbf{X}) = n \begin{bmatrix} p_1 - p_1^2 & \dots & -p_1 p_K \\ \vdots & \ddots & \vdots \\ -p_K p_1 & \dots & p_K - p_K^2 \end{bmatrix} \quad (3.2.23)$$

$$= n (\text{diag}\{\mathbf{p}\} - \mathbf{p}\mathbf{p}^\top) \quad (3.2.24)$$

Because  $\text{diag}\{\mathbf{p}\}$  is full rank and  $\mathbf{p}\mathbf{p}^\top$  has rank one, then the covariance matrix can be viewed as a rank-one update of  $\text{diag}\{\mathbf{p}\}$ . In other words, the rank can change by at most one. This establishes that the rank of  $\text{Cov}(\mathbf{X})$  is  $K - 1$ .

### Multivariate Gaussian Approximation to Multinomial Distribution

In the same way that the binomial distribution can be approximated by the Gaussian distribution with large  $n$ , we can approximate the multinomial distribution with a multivariate Gaussian if  $n$  is large. This is allowed by the multivariate Central Limit Theorem, since each element in a multinomial random vector is the sum of indicators. Thus if  $\mathbf{X}$  is multinomial distributed with parameters  $n$  and  $\mathbf{p}$ , then

$$\mathbf{X} \xrightarrow{\text{approx.}} \mathcal{N}(n\mathbf{p}, n(\text{diag}\{\mathbf{p}\} - \mathbf{p}\mathbf{p}^\top)) \quad (3.2.25)$$

We can also consider an approximation to the statistic

$$\left\| (n \text{diag}\{\mathbf{p}\})^{-1/2} (\mathbf{X} - n\mathbf{p}) \right\|^2 = \sum_{i=1}^K \frac{(X_i - np_i)^2}{np_i} \quad (3.2.26)$$

which arises in chi-squared goodness of fit testing. Applying the multivariate Gaussian approximation,

$$(n \text{diag}\{\mathbf{p}\})^{-1/2} (\mathbf{X} - n\mathbf{p}) \xrightarrow{\text{approx.}} \mathcal{N}(\mathbf{0}, Q) \quad (3.2.27)$$

where the covariance matrix is

$$Q = (n \text{diag}\{\mathbf{p}\})^{-1/2} n (\text{diag}\{\mathbf{p}\} - \mathbf{p}\mathbf{p}^\top) (n \text{diag}\{\mathbf{p}\})^{-1/2} \quad (3.2.28)$$

It can be shown that  $Q$  is idempotent (i.e.  $Q^2 = Q$ ) because

$$Q^2 = \left( I - \text{diag}\{\mathbf{p}\}^{-1/2} \mathbf{p}\mathbf{p}^\top \text{diag}\{\mathbf{p}\}^{-1/2} \right) \left( I - \text{diag}\{\mathbf{p}\}^{-1/2} \mathbf{p}\mathbf{p}^\top \text{diag}\{\mathbf{p}\}^{-1/2} \right) \quad (3.2.29)$$

$$\begin{aligned} &= I - 2 \text{diag}\{\mathbf{p}\}^{-1/2} \mathbf{p}\mathbf{p}^\top \text{diag}\{\mathbf{p}\}^{-1/2} \\ &\quad + \text{diag}\{\mathbf{p}\}^{-1/2} \underbrace{\mathbf{p}\mathbf{p}^\top \text{diag}\{\mathbf{p}\}^{-1/2} \text{diag}\{\mathbf{p}\}^{-1/2} \mathbf{p}\mathbf{p}^\top}_{1} \text{diag}\{\mathbf{p}\}^{-1/2} \end{aligned} \quad (3.2.30)$$

$$= I - \text{diag}\{\mathbf{p}\}^{-1/2} \mathbf{p}\mathbf{p}^\top \text{diag}\{\mathbf{p}\}^{-1/2} \quad (3.2.31)$$

$$= Q \quad (3.2.32)$$

where we have used

$$\mathbf{p}^\top \text{diag}\{\mathbf{p}\}^{-1} \mathbf{p} = \mathbf{p}^\top \mathbf{1} \quad (3.2.33)$$

$$= 1 \quad (3.2.34)$$

Then as  $Q$  is of rank  $K - 1$  and idempotent, it will have exactly  $K - 1$  eigenvalues of 1 and one eigenvalue of 0 (with the same arguments as in the sampling distribution of ordinary least squares residuals). Using the characterisation of the chi-squared distribution, it follows that

$$\left\| (n \text{diag}\{\mathbf{p}\})^{-1/2} (\mathbf{X} - n\mathbf{p}) \right\|^2 \sim \chi_{K-1}^2 \quad (3.2.35)$$

### 3.2.3 Multivariate Hypergeometric Distribution

### 3.2.4 Dirichlet Distribution

The Dirichlet distribution is a multivariate extension of the Beta distribution. To develop the Dirichlet distribution, we first note that although the Beta distribution is a univariate distribution, it can be alternatively defined as a bivariate distribution on support equal to the subset of  $(0, 1) \times (0, 1)$  where  $x_1 + x_2 = 1$ , with parameters  $\alpha_1 > 0, \alpha_2 > 0$ . The density then takes the same form as the univariate Beta:

$$f(x_1, x_2) = \frac{x_1^{\alpha_1-1} x_2^{\alpha_2-1}}{B(\alpha_1, \alpha_2)} \quad (3.2.36)$$

which needs to be integrated over the line  $x_1 + x_2 = 1$  to add up to one. To generalise this to a  $K$ -variate random vector  $\mathbf{X} = (X_1, \dots, X_K)$ , we define the probability simplex as the set where  $\{\mathbf{x} \in \mathbb{R}^K : \mathbf{x} > 0, \|\mathbf{x}\|_1 = 1\}$ , with  $\|\mathbf{x}\|_1 = \sum_{i=1}^K x_i$ . In other words, this set is the positive orthant portion of the hyperplane which intersects each axis at  $x_i = 1$ . This probability simplex is a  $(K - 1)$ -dimensional manifold, which has a probability density function defined over it given by

$$f(\mathbf{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K x_i^{\alpha_i-1} \quad (3.2.37)$$

where we have a vector of parameters  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$  with each  $\alpha_i > 0$  and  $B(\cdot)$  is a multivariate generalisation of the Beta function, given by

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i\right)} \quad (3.2.38)$$

To integrate the density over this  $(K - 1)$ -dimensional simplex, we can take  $x_K = 1 - \sum_{i=1}^{K-1} x_i$ , which we will then find for the  $(K - 1)$ -dimensional integral:

$$\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=1}^{K-2} x_i} \frac{1}{B(\boldsymbol{\alpha})} \left( \prod_{i=1}^{K-1} x_i^{\alpha_i-1} \right) \left( 1 - \sum_{i=1}^{K-1} x_i \right)^{\alpha_K-1} dx_{K-1} \dots dx_2 dx_1 = 1 \quad (3.2.39)$$

The Dirichlet distribution can be used to model a distribution over categorical distributions, where the variates  $x_1, \dots, x_K$  themselves are the parameters (i.e. probabilities) of a  $K$ -variate categorical distribution.

#### Marginal Distributions of Dirichlet Distribution

We can show that marginal distributions (in a single variable) of the Dirichlet distribution are Beta distributed, via a ‘merging property’. Consider without loss of generality the marginal density for  $X_1$ , obtained by integrating out  $X_2, \dots, X_{K-1}$ . This is given by

$$f_{X_1}(x_1) = \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=1}^{K-2} x_i} \frac{1}{B(\boldsymbol{\alpha})} \left( \prod_{i=1}^{K-1} x_i^{\alpha_i-1} \right) \left( 1 - \sum_{i=1}^{K-1} x_i \right)^{\alpha_K-1} dx_{K-1} \dots dx_2 \quad (3.2.40)$$

$$= \frac{x_1^{\alpha_1-1}}{B(\boldsymbol{\alpha})} \int_0^{1-x_1} x_2^{\alpha_2-1} \cdots \int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\alpha_{K-1}-1} \left( 1 - \sum_{i=1}^{K-1} x_i \right)^{\alpha_K-1} dx_{K-1} \dots dx_2 \quad (3.2.41)$$

We focus on the innermost integral. Introduce the change of variables  $z \left(1 - \sum_{i=1}^{K-2} x_i\right) = x_{K-1}$ , which gives  $dx_{K-1} = \left(1 - \sum_{i=1}^{K-2} x_i\right) dz$  and

$$1 - \sum_{i=1}^{K-1} x_i = 1 - \sum_{i=1}^{K-2} x_i - x_{K-1} \quad (3.2.42)$$

$$= 1 - \sum_{i=1}^{K-2} x_i - z \left(1 - \sum_{i=1}^{K-2} x_i\right) \quad (3.2.43)$$

$$= (1 - z) \left(1 - \sum_{i=1}^{K-2} x_i\right) \quad (3.2.44)$$

Hence this yields in the substitution

$$\int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\alpha_{K-1}-1} \left(1 - \sum_{i=1}^{K-1} x_i\right)^{\alpha_K-1} dx_{K-1} = \\ \int_0^1 z^{\alpha_{K-1}-1} \left(1 - \sum_{i=1}^{K-2} x_i\right)^{\alpha_{K-1}-1} \left[(1-z) \left(1 - \sum_{i=1}^{K-2} x_i\right)\right]^{\alpha_K-1} \left(1 - \sum_{i=1}^{K-2} x_i\right) dz \quad (3.2.45)$$

which simplifies to

$$\int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\alpha_{K-1}-1} \left(1 - \sum_{i=1}^{K-1} x_i\right)^{\alpha_K-1} dx_{K-1} = \left(1 - \sum_{i=1}^{K-2} x_i\right)^{\alpha_{K-1}+\alpha_K-1} \int_0^1 z^{\alpha_{K-1}-1} (1-z)^{\alpha_K-1} dz \\ = \left(1 - \sum_{i=1}^{K-2} x_i\right)^{\alpha_{K-1}+\alpha_K-1} B(\alpha_{K-1}, \alpha_K) \quad (3.2.46)$$

Also note that a factorisation of  $\mathbf{B}(\boldsymbol{\alpha})$  occurs in the following way:

$$\mathbf{B}(\alpha_{K-1}, \alpha_K) \mathbf{B}(\alpha_1, \dots, \alpha_{K-1}, \alpha_{K-1} + \alpha_K) = \frac{\Gamma(\alpha_{K-1}) \Gamma(\alpha_K)}{\Gamma(\alpha_{K-1} + \alpha_K)} \times \frac{\Gamma(\alpha_1) \dots \Gamma(\alpha_{K-2}) \Gamma(\alpha_{K-1} + \alpha_K)}{\Gamma(\alpha_1 + \dots + \alpha_{K-2} + \alpha_{K-1} + \alpha_K)} \quad (3.2.48)$$

$$= \frac{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)}{\Gamma(\alpha_1 + \dots + \alpha_K)} \quad (3.2.49)$$

$$= \mathbf{B}(\alpha_1, \dots, \alpha_K) \quad (3.2.50)$$

Therefore fully integrating out the innermost variable yields the form

$$f_{X_1}(x_1) = \frac{x_1^{\alpha_1-1}}{\mathbf{B}(\alpha_1, \dots, \alpha_{K-2}, \alpha_{K-1} + \alpha_K)} \\ \times \int_0^{1-x_1} x_2^{\alpha_2-1} \dots \int_0^{1-\sum_{i=1}^{K-3} x_i} x_{K-2}^{\alpha_{K-2}-1} \left(1 - \sum_{i=1}^{K-2} x_i\right)^{\alpha_{K-1}+\alpha_K-1} dx_{K-2} \dots dx_2 \quad (3.2.51)$$

Recognise this as the same as integrating out over the  $(K-1)$ -variate Dirichlet distribution with parameters  $\alpha_1, \dots, \alpha_{K-2}, \alpha_{K-1} + \alpha_K$ , where we have ‘merged’ the last two variables by their sum and the sum of their respective parameters. Iterating this procedure, we will eventually find that

$$f_{X_1}(x_1) = \frac{x_1^{\alpha_1-1} (1-x_1)^{\alpha_2+\dots+\alpha_K-1}}{\mathbf{B}(\alpha_1, \alpha_2 + \dots + \alpha_K)} \quad (3.2.52)$$

Hence the marginal distribution of  $X_1$  is Beta distributed with parameters  $\alpha_1, \sum_{i=2}^K \alpha_i$ .

### 3.2.5 Dirichlet-Multinomial Distribution

The Dirichlet-multinomial distribution is a compound distribution which is a multivariate generalisation of the Beta-binomial distribution. It express the distribution over draws from a multinomial distribution of size  $n$  with  $K$  categories, where the event probabilities  $\mathbf{p} = (p_1, \dots, p_K)$  themselves are drawn from a Dirichlet distribution with parameters  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ . With  $K = 2$ , it reduces to the Beta-binomial distribution.

### 3.2.6 Multivariate Cauchy Distribution

### 3.2.7 Elliptical Distributions

## 3.3 Inequalities in Probability

### 3.3.1 Boole's Inequality

Also known as the *union bound*, Boole's inequality generalises the addition law of probability. It states that for events  $A_1, A_2, A_3, \dots$ :

$$\Pr \left( \bigcup_{i=1}^n A_i \right) \leq \sum_{i=1}^n \Pr (A_i) \quad (3.3.1)$$

*Proof.* Starting with the base case, we clearly have for  $n = 1$ :

$$\Pr (A_1) \leq \Pr (A_1) \quad (3.3.2)$$

For the induction step, suppose for some  $n \geq 1$ :

$$\Pr \left( \bigcup_{i=1}^n A_i \right) \leq \sum_{i=1}^n \Pr (A_i) \quad (3.3.3)$$

By letting  $A := \bigcup_{i=1}^n A_i$  and  $B := A_{n+1}$  and using the addition law of probability,

$$\Pr \left( \bigcup_{i=1}^{n+1} A_i \right) = \Pr \left( \bigcup_{i=1}^n A_i \right) + \Pr (A_{n+1}) - \Pr \left( \bigcup_{i=1}^n A_i \cap A_{n+1} \right) \quad (3.3.4)$$

Since  $\Pr (\bigcup_{i=1}^n A_i \cap A_{n+1}) \geq 0$ , then

$$\Pr \left( \bigcup_{i=1}^{n+1} A_i \right) \leq \Pr \left( \bigcup_{i=1}^n A_i \right) + \Pr (A_{n+1}) \quad (3.3.5)$$

and because it is given  $\Pr (\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n \Pr (A_i)$ , we have

$$\Pr \left( \bigcup_{i=1}^{n+1} A_i \right) \leq \sum_{i=1}^n \Pr (A_i) + \Pr (A_{n+1}) \quad (3.3.6)$$

$$= \sum_{i=1}^{n+1} \Pr (A_i) \quad (3.3.7)$$

□

### 3.3.2 Comparison of Random Variables

**Theorem 3.1.** *Let  $X$  and  $Y$  be two random variables. Then*

$$\Pr(X \geq Y) \leq \Pr(X \geq \Delta) + \Pr(Y \leq \Delta) \quad (3.3.8)$$

for any  $\Delta \in \mathbb{R}$ .

*Proof.* We have the event  $\{X < \Delta \cap Y > \Delta\} \subseteq \{X < Y\}$ , so

$$\Pr(X < \Delta \cap Y > \Delta) \leq \Pr(X < Y) \quad (3.3.9)$$

By taking complementary probabilities on the right-hand side and applying DeMorgan's laws on the left-hand side,

$$1 - \Pr(X \geq \Delta \cup Y \leq \Delta) \leq 1 - \Pr(X \geq Y) \quad (3.3.10)$$

Rearranging gives

$$\Pr(X \geq Y) \leq \Pr(X \geq \Delta \cup Y \leq \Delta) \quad (3.3.11)$$

and lastly applying Boole's inequality gives

$$\Pr(X \geq Y) \leq \Pr(X \geq \Delta) + \Pr(Y \leq \Delta) \quad (3.3.12)$$

□

By instead starting the proof with  $\{X < \mu_X + \delta \cap Y > \mu_Y - \delta\} \subseteq \{X < Y\}$ , we can have a similar result with strict inequalities.

**Corollary 3.1.** *Let  $X$  and  $Y$  be two random variables. Then*

$$\Pr(X > Y) \leq \Pr(X > \Delta) + \Pr(Y < \Delta) \quad (3.3.13)$$

for all  $\Delta \in \mathbb{R}$ .

While  $\Delta$  can be anything in general, it may sometimes be the case that choosing  $\Delta = \frac{\mathbb{E}[X] + \mathbb{E}[Y]}{2}$  will give less conservative bounds. This gives the following specific result.

**Corollary 3.2.** *Let  $X$  and  $Y$  be two continuous random variables with means  $\mu_X$  and  $\mu_Y$  respectively. Let  $\delta = \frac{\mu_Y - \mu_X}{2}$ . Then*

$$\Pr(X \geq Y) \leq \Pr(X \geq \mu_X + \delta) + \Pr(Y \leq \mu_Y - \delta) \quad (3.3.14)$$

$$\Pr(X > Y) \leq \Pr(X > \mu_X + \delta) + \Pr(Y < \mu_Y - \delta) \quad (3.3.15)$$

The value  $\mu_X + \delta = \mu_Y - \delta$  can be thought of as the ‘midway’ point between  $\mu_X$  and  $\mu_Y$ , although note that this inequality is really only useful when  $\mu_Y \geq \mu_X$ .

### 3.3.3 Jensen's Inequality

A function  $f(x)$  is convex if and only Jensen's inequality holds for any  $x$  and  $y$  in the domain of  $f$ :

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad (3.3.16)$$

where  $\theta \in [0, 1]$  and  $\theta x + (1 - \theta)y$  is termed as a ‘convex combination’ of  $x$  and  $y$ . Intuitively speaking, no part of a secant joining any two points on the graph of  $f(x)$  can lie below the graph. A convex combination can be extended to more than two points, and so can Jensen's inequality:

$$f(\theta_1 x_1 + \dots + \theta_k x_k) \leq \theta_1 f(x_1) + \dots + \theta_k f(x_k) \quad (3.3.17)$$

where  $\theta_1, \dots, \theta_k \geq 0$  and  $\theta_1 + \dots + \theta_k = 1$ . We may extend this to infinite sums, i.e. integrals over probability distributions:

$$f\left(\int p(x) dx\right) \leq \int f(x) p(x) dx \quad (3.3.18)$$

where  $p(x) \geq 0$  and  $\int p(x) dx = 1$ . Recognise that the above inequality pertains to the expected value of a random variable  $X$  with probability density  $p(x)$ , and a convex function  $f(x)$ . We can write

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \quad (3.3.19)$$

If  $f(x)$  is convex then  $g(x) = -f(x)$  is concave, so Jensen's inequality can be alternatively stated for concave  $g(x)$ :

$$g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)] \quad (3.3.20)$$

Jensen's inequality also generalises to real convex functions  $f(\cdot)$  of a random vector:

$$f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})] \quad (3.3.21)$$

### 3.3.4 Markov's Inequality

Let  $X$  be a non-negative random variable (it cannot take on negative values), and let  $a > 0$  be some constant. Then

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad (3.3.22)$$

or alternatively, if we let  $\tilde{a}$  be some constant such that  $a = \tilde{a}\mathbb{E}[X]$ , then making this substitution gives

$$\Pr(X \geq \tilde{a}\mathbb{E}[X]) \leq \frac{1}{\tilde{a}} \quad (3.3.23)$$

*Proof.* For a continuous non-negative random variable, the expectation is defined as

$$\mathbb{E}[X] = \int_0^\infty xf(x) dx \quad (3.3.24)$$

$$= \int_0^a xf(x) dx + \int_a^\infty xf(x) dx \quad (3.3.25)$$

$$\geq \int_a^\infty xf(x) dx \quad (3.3.26)$$

We then have

$$\int_a^\infty af(x) dx \leq \int_a^\infty xf(x) dx \quad (3.3.27)$$

$$a \int_a^\infty f(x) dx \leq \mathbb{E}[X] \quad (3.3.28)$$

$$a \Pr(X \geq a) \leq \mathbb{E}[X] \quad (3.3.29)$$

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad (3.3.30)$$

If  $X$  is a non-negative discrete random variable on support  $\{0, 1, 2, \dots\}$ , then

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \Pr(X = x) \quad (3.3.31)$$

$$= \sum_{x=0}^{a-1} x \Pr(X = x) + \sum_{x=a}^{\infty} x \Pr(X = x) \quad (3.3.32)$$

$$\geq \sum_{x=a}^{\infty} x \Pr(X = x) \quad (3.3.33)$$

Then similar to before,

$$\sum_{x=a}^{\infty} a \Pr(X = x) \leq \sum_{x=a}^{\infty} x \Pr(X = x) \quad (3.3.34)$$

$$a \sum_{x=a}^{\infty} \Pr(X = x) \leq \mathbb{E}[X] \quad (3.3.35)$$

$$a \Pr(X \geq a) \leq \mathbb{E}[X] \quad (3.3.36)$$

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad (3.3.37)$$

□

An alternative proof is given below:

*Proof.* Let  $\mathbb{I}_{X \geq a}$  be an indicator random variable where  $\mathbb{I}_{X \geq a} = 1$  if the event  $X \geq a$  occurs, and  $\mathbb{I}_{X \geq a} = 0$  if the event  $X < a$  occurs. Then we can write

$$a\mathbb{I}_{X \geq a} \leq X \quad (3.3.38)$$

This can be seen as follows:

- Suppose  $X \geq a$ , then  $\mathbb{I}_{X \geq a} = 1$  and the inequality above holds.
- Suppose  $X < a$ , then  $\mathbb{I}_{X \geq a} = 0$  and the inequality above is satisfied because  $X$  is non-negative.

Take the expectation operator of both sides. The expectation operator is monotonically increasing (i.e.  $\mathbb{E}[c] = c$  for some constant  $c$ ) so this preserves the inequality.

$$\mathbb{E}[a\mathbb{I}_{X \geq a}] \leq \mathbb{E}[X] \quad (3.3.39)$$

Taking the constant  $a$  out of the expectation:

$$a\mathbb{E}[\mathbb{I}_{X \geq a}] \leq \mathbb{E}[X] \quad (3.3.40)$$

$$\mathbb{E}[\mathbb{I}_{X \geq a}] \leq \frac{\mathbb{E}[X]}{a} \quad (3.3.41)$$

We can evaluate  $\mathbb{E}[\mathbb{I}_{X \geq a}]$

$$\mathbb{E}[\mathbb{I}_{X \geq a}] = 1 \times \Pr(X \geq a) + 0 \times \Pr(X < a) = \Pr(X \geq a) \quad (3.3.42)$$

Hence

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad (3.3.43)$$

□

We can use this result to bound probabilities of  $X$  being greater than some value in relation to the mean. For example, suppose the population of  $X$  is income (which we assume to be non-negative). Then if we choose  $a$  to be 10 times the average income, we can make the statement that the proportion of people earning at least 10 times the average income is no more than 10%.

### Reverse Markov's Inequality

A lower bound on  $\Pr(X \geq a)$  in terms of  $\mathbb{E}[X]$  can be obtained under special circumstances. Suppose  $X$  is upper bounded by  $b$  such that  $\Pr(X \leq b) = 1$ . Then the random variable  $X' = b - X$  is non-negative so Markov's inequality can be applied to this. We have for some  $a > 0$

$$\Pr(X' \geq a) \leq \frac{\mathbb{E}[X']}{a} \quad (3.3.44)$$

Since  $\mathbb{E}[X'] = b - \mathbb{E}[X]$ :

$$\Pr(X' \geq a) \leq \frac{b - \mathbb{E}[X]}{a} \quad (3.3.45)$$

$$\Pr(b - X \geq a) \leq \frac{b - \mathbb{E}[X]}{a} \quad (3.3.46)$$

$$\Pr(b - a \geq X) \leq \frac{b - \mathbb{E}[X]}{a} \quad (3.3.47)$$

By complements,

$$1 - \Pr(b - a \leq X) \leq \frac{b - \mathbb{E}[X]}{a} \quad (3.3.48)$$

$$\Pr(X \geq b - a) \geq 1 - \frac{b - \mathbb{E}[X]}{a} \quad (3.3.49)$$

$$\Pr(X \geq b - a) \geq \frac{a - b + \mathbb{E}[X]}{a} \quad (3.3.50)$$

Let  $c = b - a$ , then

$$\Pr(X \geq c) \geq \frac{\mathbb{E}[X] - c}{b - c} \quad (3.3.51)$$

Note that this inequality is only useful when  $c < \mathbb{E}[X]$ .

### 3.3.5 Chebychev's Inequality

Let  $X$  be a random variable with finite mean  $\mu$  and non-zero standard deviation  $\sigma$ . Then for any  $k > 0$ , Chebychev's inequality (also known as the Bienaymé-Chebychev inequality) states that

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad (3.3.52)$$

In words, this states that the probability of  $X$  being at least  $k$  standard deviations away from the mean can be no more than  $\frac{1}{k^2}$ .

*Proof.* Let a random variable  $Y$  be defined as  $Y = (X - \mu)^2$  and let  $a = (k\sigma)^2$ . Then applying Markov's inequality

$$\Pr(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a} \quad (3.3.53)$$

$$\Pr((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2\sigma^2} \quad (3.3.54)$$

The event inside the probability is equivalent to  $|X - \mu| \geq k\sigma$  and the numerator on the right hand side is the definition of variance  $\sigma^2$ . So

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad (3.3.55)$$

□

Note that this inequality is only useful for when  $k > 1$ , because when  $0 < k \leq 1$ , the term  $\frac{1}{k^2} \geq 1$  so Chebychev's inequality is trivially satisfied. An alternative form of the inequality is to let  $k = \varepsilon/\sigma$ . Then we have

$$\Pr(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \quad (3.3.56)$$

Chebychev's inequality can be used to bound probabilities within a certain number of standard deviations from the mean, for very general distributions. Obtaining more knowledge of the distribution (such as information the distribution is normal) will often lead to tighter (i.e. smaller) bounds.

### Multivariate Chebychev's Inequality

If  $\mathbf{X}$  is an  $n$ -dimensional random vector with mean  $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$  and covariance matrix  $\text{Cov}(\mathbf{X}) = \Sigma$ , then multivariate Chebychev's inequality states that for any  $k > 0$ :

$$\Pr\left(\sqrt{(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})} > k\right) \leq \frac{n}{k^2} \quad (3.3.57)$$

*Proof.* Define the random variable  $Y = (\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})$ . By positive definiteness of  $\Sigma$  and hence  $\Sigma^{-1}$ , this means  $Z$  is positive so

$$\Pr\left(\sqrt{(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})} > k\right) = \Pr\left(\sqrt{Y} > k\right) \quad (3.3.58)$$

$$= \Pr(Y > k^2) \quad (3.3.59)$$

$$\leq \frac{\mathbb{E}[Y]}{k^2} \quad (3.3.60)$$

where the Markov inequality is applied in the last step. We then have

$$\mathbb{E}[Y] = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})\right] \quad (3.3.61)$$

Since  $Y$  is univariate,

$$\mathbb{E}[Y] = \mathbb{E}\left[\text{trace}\left((\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})\right)\right] \quad (3.3.62)$$

Because the trace is invariant to cyclic permutations,

$$\mathbb{E}[Y] = \mathbb{E}\left[\text{trace}\left(\Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})^\top (\mathbf{X} - \boldsymbol{\mu})\right)\right] \quad (3.3.63)$$

Then use the fact the trace is a linear operator:

$$\mathbb{E}[Y] = \text{trace}\left(\Sigma^{-1} \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})^\top (\mathbf{X} - \boldsymbol{\mu})\right]\right) \quad (3.3.64)$$

And lastly by using the definition of the covariance matrix and the trace of the identity:

$$\mathbb{E}[Y] = \text{trace}(\Sigma^{-1} \Sigma) \quad (3.3.65)$$

$$= \text{trace}(I) \quad (3.3.66)$$

$$= n \quad (3.3.67)$$

Therefore

$$\Pr\left(\sqrt{(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})} > k\right) \leq \frac{n}{k^2} \quad (3.3.68)$$

□

### 3.3.6 Gauss' Inequality [158]

### 3.3.7 Vysochanskij-Petunin Inequality [158]

### 3.3.8 Cantelli's Inequality

For a random variable  $X$  with expectation  $\mu$  and variance  $\sigma^2$ , Cantelli's inequality states that for  $\delta > 0$ :

$$\Pr(|X - \mu| \geq \delta) \leq \frac{2\sigma^2}{\sigma^2 + \delta^2} \quad (3.3.69)$$

*Proof.* Consider the probability  $\Pr(X - \mu > \lambda)$  for the case where  $\lambda > 0$ . Define  $Y := X - \mu$  so that  $\mathbb{E}[Y] = 0$  and  $\text{Var}(Y) = \sigma^2$ . Then for any  $u \geq 0$ :

$$\Pr(X - \mu \geq \lambda) = \Pr(Y \geq \lambda) \quad (3.3.70)$$

$$= \Pr(Y + u \geq \lambda + u) \quad (3.3.71)$$

$$\leq \Pr((Y + u)^2 \geq (\lambda + u)^2) \quad (3.3.72)$$

where the last inequality comes from recognising that  $\lambda + u$  is positive, but  $Y + u$  can generally be negative, so the event  $\{Y + u \geq \lambda + u\} \subseteq \{(Y + u)^2 \geq (\lambda + u)^2\}$ . Then applying Markov's inequality on the non-negative random variable  $(Y + u)^2$ ,

$$\Pr(X - \mu \geq \lambda) \leq \frac{\mathbb{E}[(Y + u)^2]}{(\lambda + u)^2} \quad (3.3.73)$$

$$= \frac{\mathbb{E}[Y^2 + Yu + u^2]}{(\lambda + u)^2} \quad (3.3.74)$$

$$= \frac{\mathbb{E}[Y^2] + \mathbb{E}[Yu] + \mathbb{E}[u^2]}{(\lambda + u)^2} \quad (3.3.75)$$

$$= \frac{\text{Var}(Y) + u^2}{(\lambda + u)^2} \quad (3.3.76)$$

$$= \frac{\sigma^2 + u^2}{(\lambda + u)^2} \quad (3.3.77)$$

To make this bound as tight as possible, we find the value  $u^* \geq 0$  that minimises the left hand side. Differentiating,

$$\frac{d}{du} \left( \frac{\sigma^2 + u^2}{(\lambda + u)^2} \right) = \frac{d}{du} \left( \sigma^2 (\lambda + u)^2 + u^2 (\lambda + u)^{-2} \right) \quad (3.3.78)$$

$$= -2\sigma^2 (\lambda + u)^{-3} + 2u (\lambda + u)^{-2} - 2u^2 (\lambda + u)^{-3} \quad (3.3.79)$$

Setting the derivative to zero and solving:

$$\frac{2u^*}{(\lambda + u^*)^2} - \frac{2u^{*2}}{(\lambda + u^*)^2} - \frac{2\sigma^2}{(\lambda + u^*)^3} = 0 \quad (3.3.80)$$

$$u^*(\lambda + u^*) - u^{*2} - \sigma^2 = 0 \quad (3.3.81)$$

$$u^*\lambda - \sigma^2 = 0 \quad (3.3.82)$$

$$u^* = \frac{\sigma^2}{\lambda} \quad (3.3.83)$$

Hence substituting  $u^*$ :

$$\Pr(X - \mu \geq \lambda) \leq \frac{\sigma^2 + \sigma^4/\lambda^2}{(\lambda + \sigma^2/\lambda)^2} \quad (3.3.84)$$

$$= \frac{\sigma^2 + \sigma^4/\lambda^2}{\lambda^2 + 2\sigma^2 + \sigma^4/\lambda^2} \quad (3.3.85)$$

$$= \frac{\lambda^2\sigma^2 + \sigma^4}{\lambda^4 + 2\sigma^2\lambda^2 + \sigma^4} \quad (3.3.86)$$

$$= \frac{\sigma^2(\lambda^2 + \sigma^2)}{(\lambda^2 + \sigma^2)^2} \quad (3.3.87)$$

$$= \frac{\sigma^2}{\lambda^2 + \sigma^2} \quad (3.3.88)$$

Now suppose  $\lambda < 0$ , then define  $\alpha := -\lambda > 0$  and  $Z = -Y$ . Then as above,

$$\Pr(X - \mu < \lambda) = \Pr(Y < \lambda) \quad (3.3.89)$$

$$= \Pr(Z > \alpha) \quad (3.3.90)$$

$$\leq \frac{\sigma^2}{\alpha^2 + \sigma^2} \quad (3.3.91)$$

$$= \frac{\sigma^2}{\lambda^2 + \sigma^2} \quad (3.3.92)$$

To combine both cases, define  $\delta := |\lambda| > 0$ , then

$$\Pr(|X - \mu| \geq \delta) = \Pr(X - \mu \geq \delta \cup X - \mu \leq -\delta) \quad (3.3.93)$$

$$= \Pr(X - \mu \geq \delta) + \Pr(X - \mu < -\delta) \quad (3.3.94)$$

$$\leq \frac{\sigma^2}{\delta^2 + \sigma^2} + \frac{\sigma^2}{\delta^2 + \sigma^2} \quad (3.3.95)$$

$$= \frac{2\sigma^2}{\delta^2 + \sigma^2} \quad (3.3.96)$$

□

### 3.3.9 Cauchy-Schwarz Inequality

By viewing random variables as objects which assign a number to every outcome in a random experiment, they can be treated as points on a vector space. An inner product between random variables  $X$  and  $Y$  can be defined as

$$\langle X, Y \rangle := \mathbb{E}[XY] \quad (3.3.97)$$

We can show that this satisfies the definition of an inner product space because of the linearity of expectation, and also because  $\mathbb{E}[X^2] = 0$  implies  $\Pr(X = 0) = 1$  (i.e. a value of 0 is assigned to every outcome in the random experiment) and vice-versa. The latter is shown by considering

$$\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2 \quad (3.3.98)$$

$$= 0 \quad (3.3.99)$$

which can only be satisfied if  $\text{Var}(X) = 0$  since we must have  $0 \leq \mathbb{E}[X]^2 = -\text{Var}(X) \leq 0$ . The converse is more obvious to show, just by using the definition of mean and variance. Applying the Cauchy-Schwarz inequality to this inner product gives

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2] \quad (3.3.100)$$

This can also be used to show the inequality that the correlation coefficient is bounded between  $-1$  and  $1$ :

$$\text{Cov}(X, Y)^2 = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \quad (3.3.101)$$

$$\leq \mathbb{E} [(X - \mathbb{E}[X])^2] \mathbb{E} [(Y - \mathbb{E}[Y])^2] \quad (3.3.102)$$

$$= \text{Var}(X) \text{Var}(Y) \quad (3.3.103)$$

Hence

$$-\sqrt{\text{Var}(X) \text{Var}(Y)} \leq \text{Cov}(X, Y) \leq \sqrt{\text{Var}(X) \text{Var}(Y)} \quad (3.3.104)$$

$$-1 \leq \text{Corr}(X, Y) \leq 1 \quad (3.3.105)$$

### 3.3.10 Paley-Zygmund Inequality

For a non-negative random variable  $X \geq 0$  with finite variance,

$$\Pr(X > \theta \mathbb{E}[X]) \geq (1 - \theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \quad (3.3.106)$$

for any  $\theta \in [0, 1]$ . Intuitively, the inequality lower-bounds the probability that the random variable is small, in terms of the first and second moments.

*Proof.* Write  $\mathbb{E}[X]$  using the sum of mutually exclusive indicators:

$$\mathbb{E}[X] = \mathbb{E}[X(\mathbb{I}_{\{X \leq \theta \mathbb{E}[X]\}} + \mathbb{I}_{\{X > \theta \mathbb{E}[X]\}})] \quad (3.3.107)$$

$$= \mathbb{E}[X \mathbb{I}_{\{X \leq \theta \mathbb{E}[X]\}}] + \mathbb{E}[X \mathbb{I}_{\{X > \theta \mathbb{E}[X]\}}] \quad (3.3.108)$$

Note that  $\mathbb{E}[X \mathbb{I}_{\{X \leq \theta \mathbb{E}[X]\}}] \leq \theta \mathbb{E}[X]$  when  $\theta \in [0, 1]$  because  $X \mathbb{I}_{\{X \leq \theta \mathbb{E}[X]\}} = 0$  when  $X > \theta \mathbb{E}[X]$ , otherwise  $X \leq \theta \mathbb{E}[X]$ . Using the Cauchy-Schwarz inequality for the second term,

$$\mathbb{E}[X \mathbb{I}_{\{X > \theta \mathbb{E}[X]\}}]^2 \leq \mathbb{E}[X^2] \mathbb{E}[\mathbb{I}_{\{X > \theta \mathbb{E}[X]\}}^2] \quad (3.3.109)$$

$$= \mathbb{E}[X^2] \mathbb{E}[\mathbb{I}_{\{X > \theta \mathbb{E}[X]\}}] \quad (3.3.110)$$

$$= \mathbb{E}[X^2] \Pr(X > \theta \mathbb{E}[X]) \quad (3.3.111)$$

Hence  $\mathbb{E}[X \mathbb{I}_{\{X > \theta \mathbb{E}[X]\}}] \leq \mathbb{E}[X^2]^{1/2} \Pr(X > \theta \mathbb{E}[X])^{1/2}$  as  $X$  is non-negative. We arrive at the desired inequality via the following rearrangement:

$$\mathbb{E}[X] \leq \theta \mathbb{E}[X] + \mathbb{E}[X^2]^{1/2} \Pr(X > \theta \mathbb{E}[X])^{1/2} \quad (3.3.112)$$

$$\mathbb{E}[X](1 - \theta) \leq \mathbb{E}[X^2]^{1/2} \Pr(X > \theta \mathbb{E}[X])^{1/2} \quad (3.3.113)$$

$$\Pr(X > \theta \mathbb{E}[X])^{1/2} \geq (1 - \theta) \frac{\mathbb{E}[X]}{\mathbb{E}[X^2]^{1/2}} \quad (3.3.114)$$

$$\Pr(X > \theta \mathbb{E}[X]) \geq (1 - \theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \quad (3.3.115)$$

□

### 3.3.11 Hölder's Inequality

For random variables  $X, Y$  and some  $p > 1, q > 1$  such that  $\frac{1}{p} + \frac{1}{q} = 1$ , Hölder's inequality for probability states that

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{1/p} \mathbb{E}[|Y|^q]^{1/q} \quad (3.3.116)$$

*Proof.* If the left hand side is zero (i.e.  $\Pr(X = 0) = 1$  or  $\Pr(Y = 0) = 1$ ), then by applying similar reasoning as for the Cauchy-Schwarz inequality, the right hand side must also be zero, satisfying Hölder's inequality. We then focus on the case where the left hand side is positive. For some positive  $a, b$ , there exist  $s, t$  such that  $a = e^{s/p}$  and  $b = e^{t/q}$ . Since  $\exp(\cdot)$  is a convex function, then by  $p^{-1} + q^{-1} = 1$ :

$$e^{p^{-1}s+q^{-1}t} \leq p^{-1}e^s + q^{-1}e^t \quad (3.3.117)$$

Then applying the definitions of  $a$  and  $b$ , this gives what is known as Young's inequality for products:

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad (3.3.118)$$

which is satisfied when  $a = 0$  or  $b = 0$  as well as for positive  $a, b$ . Then let  $a = \frac{|X|}{\mathbb{E}[|X|^p]^{1/p}}$  and  $b = \frac{|Y|}{\mathbb{E}[|Y|^q]^{1/q}}$  so that

$$\frac{|XY|}{\mathbb{E}[|X|^p]^{1/p} \mathbb{E}[|Y|^q]^{1/q}} \leq \frac{1}{p} \frac{|X|^p}{\mathbb{E}[|X|^p]} + \frac{1}{q} \frac{|Y|^q}{\mathbb{E}[|Y|^q]} \quad (3.3.119)$$

Taking the expectation of both sides,

$$\frac{\mathbb{E}[|XY|]}{\mathbb{E}[|X|^p]^{1/p} \mathbb{E}[|Y|^q]^{1/q}} \leq \frac{1}{p} \frac{\mathbb{E}[|X|^p]}{\mathbb{E}[|X|^p]} + \frac{1}{q} \frac{\mathbb{E}[|Y|^q]}{\mathbb{E}[|Y|^q]} \quad (3.3.120)$$

$$= \frac{1}{p} + \frac{1}{q} \quad (3.3.121)$$

$$= 1 \quad (3.3.122)$$

Therefore

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{1/p} \mathbb{E}[|Y|^q]^{1/q} \quad (3.3.123)$$

□

### 3.3.12 Minkowski's Inequality

For random variables  $X, Y$  and some  $p \geq 1$ , Minkowski's inequality for probability states that

$$\mathbb{E}[|X + Y|^p]^{1/p} \leq \mathbb{E}[|X|^p]^{1/p} + \mathbb{E}[|Y|^p]^{1/p} \quad (3.3.124)$$

*Proof.* It suffices to show that for non-negative  $X$  and  $Y$  that

$$\mathbb{E}\left[\left(X^{1/p} + Y^{1/p}\right)^p\right] \leq \left(\mathbb{E}[X]^{1/p} + \mathbb{E}[Y]\right)^p \quad (3.3.125)$$

because then we can then simply replace  $X$  with  $|X|^p$  and  $Y$  with  $|Y|^p$ , then take both sides to the power of  $1/p$  to obtain Minkowski's inequality. Introduce the function  $f(x, y) = (x^{1/p} + y^{1/p})^p$ , which we will show is concave. Taking the partial derivatives (up to second order):

$$\frac{\partial f(x, y)}{\partial x} = p \cdot \frac{1}{p} x^{1/p-1} \left(x^{1/p} + y^{1/p}\right)^{p-1} \quad (3.3.126)$$

$$= x^{1/p-1} \left(x^{1/p} + y^{1/p}\right)^{p-1} \quad (3.3.127)$$

and

$$\frac{\partial^2 f(x, y)}{\partial x^2} = x^{1/p-1} (p-1) \cdot \frac{1}{p} x^{1/p-1} \left( x^{1/p} + y^{1/p} \right)^{p-2} + \left( \frac{1}{p} - 1 \right) x^{1/p-2} \left( x^{1/p} + y^{1/p} \right)^{p-1} \quad (3.3.128)$$

$$= \frac{p-1}{p} x^{2/p-2} \left( x^{1/p} + y^{1/p} \right)^{p-2} - \frac{p-1}{p} x^{1/p-2} \left( x^{1/p} + y^{1/p} \right) \left( x^{1/p} + y^{1/p} \right)^{p-2} \quad (3.3.129)$$

$$= \frac{p-1}{p} \left( x^{1/p} + y^{1/p} \right)^{p-2} \left[ x^{2/p-2} - x^{1/p-2} \left( x^{1/p} + y^{1/p} \right) \right] \quad (3.3.130)$$

$$= \frac{1-p}{p} \left( x^{1/p} + y^{1/p} \right)^{p-2} x^{1/p-2} y^{1/p} \quad (3.3.131)$$

which is negative since  $p \geq 1$ . By symmetry, an analogous expression can be found for  $\frac{\partial^2 f(x, y)}{\partial y^2}$ .

Also,

$$\frac{\partial^2 f(x, y)}{\partial x \partial y} = \frac{p-1}{p} x^{1/p-1} y^{1/p-1} \left( x^{1/p} + y^{1/p} \right)^{p-2} \quad (3.3.132)$$

We can then show negative semi-definiteness of the Hessian of  $f(x, y)$ , since

$$\left( \frac{\partial^2 f(x, y)}{\partial x \partial y} \right)^2 = \left( \frac{p-1}{p} \right)^2 \left( x^{1/p} + y^{1/p} \right)^{2p-4} x^{2/p-2} y^{2/p-2} \quad (3.3.133)$$

$$\frac{\partial^2 f(x, y)}{\partial x^2} \frac{\partial^2 f(x, y)}{\partial y^2} = \left( \frac{p-1}{p} \right)^2 \left( x^{1/p} + y^{1/p} \right)^{2p-4} x^{2/p-2} y^{2/p-2} \quad (3.3.134)$$

which gives  $\left( \frac{\partial^2 f(x, y)}{\partial x \partial y} \right)^2 = \frac{\partial^2 f(x, y)}{\partial x^2} \frac{\partial^2 f(x, y)}{\partial y^2}$ . Hence  $f(x, y)$  is concave (or equivalently,  $-f(x, y)$  is convex). Using Jensen's inequality,

$$f(\mathbb{E}[X], \mathbb{E}[Y]) \geq \mathbb{E}[f(X, Y)] \quad (3.3.135)$$

therefore

$$\mathbb{E} \left[ \left( X^{1/p} + Y^{1/p} \right)^p \right] \leq \left( \mathbb{E}[X]^{1/p} + \mathbb{E}[Y] \right)^p \quad (3.3.136)$$

as required.  $\square$

### 3.3.13 Lyapunov's Inequality

Lyapunov's inequality for probability can be derived from Hölder's inequality. First let  $0 < \alpha < \beta$  and choose  $p = \beta/\alpha > 1$  and  $q = \beta/(\beta - \alpha) > 1$ , noting that this satisfies  $p^{-1} + q^{-1} = 1$ . Then apply Hölder's inequality to the random variable  $|X|^\alpha$  and the degenerate random variable 1 to yield

$$\mathbb{E}[|X|^\alpha] \leq \mathbb{E}[|X|^\beta]^{\alpha/\beta} \quad (3.3.137)$$

for  $0 < \alpha \leq \beta$ , noting that this inequality is trivially satisfied (with equality) when  $\alpha = \beta$ .

### 3.3.14 Popoviciu's Inequality

Popoviciu's inequality on variances upper bounds the variance of a bounded random variable. Let  $X$  be a random variable such that  $\inf X = a$  and  $\sup X = b$  (so the support could be  $[a, b]$  or  $(a, b)$ , for instance). Then Popoviciu's inequality says

$$\text{Var}(X) \leq \frac{(b-a)^2}{4} \quad (3.3.138)$$

*Proof.* Let  $g(t) = \mathbb{E}[(X - t)^2]$ . We know that this quantity is minimised when  $t = \mathbb{E}[X]$ , giving the definition of the variance, i.e.

$$\operatorname{argmin}_t g(t) = \mathbb{E}[X] \quad (3.3.139)$$

$$\min_t g(t) = \operatorname{Var}(X) \quad (3.3.140)$$

This fact may be verified by setting the derivative of  $g(t)$  to zero. Now put  $t = \frac{a+b}{2}$ , which is the midpoint of the upper and lower bounds. We have

$$\operatorname{Var}(X) \leq g\left(\frac{a+b}{2}\right) \quad (3.3.141)$$

$$= \mathbb{E}\left[\left(X - \frac{a+b}{2}\right)^2\right] \quad (3.3.142)$$

Note that

$$\left|X - \frac{a+b}{2}\right| \leq \frac{b-a}{2} \quad (3.3.143)$$

i.e. the distance of any point in  $[a, b]$  from the midpoint is upper bounded by the half-width. Therefore

$$\operatorname{Var}(X) \leq \mathbb{E}\left[\left(X - \frac{a+b}{2}\right)^2\right] \quad (3.3.144)$$

$$\leq \frac{(b-a)^2}{4} \quad (3.3.145)$$

□

A heuristic proof is also available. Suppose we wanted to construct a random variable with the maximum variance, with the only restriction being it be bounded on support  $[a, b]$ . Intuition says we should construct the random variable as

$$\Pr(X = x) = \begin{cases} 1/2, & x = a \\ 1/2, & x = b \end{cases} \quad (3.3.146)$$

i.e. split the mass evenly at opposite ends of the support, so it has the most ‘spread’. The mean of this random variable is  $\mathbb{E}[X] = \frac{a+b}{2}$ , and its variance can be computed as

$$\operatorname{Var}(X) = \frac{1}{2} \left(a - \frac{a+b}{2}\right)^2 + \frac{1}{2} \left(b - \frac{a+b}{2}\right)^2 \quad (3.3.147)$$

$$= \frac{1}{2} \left(\frac{a-b}{2}\right)^2 + \frac{1}{2} \left(\frac{b-a}{2}\right)^2 \quad (3.3.148)$$

$$= \frac{(b-a)^2}{4} \quad (3.3.149)$$

which is the upper bound in Popoviciu’s inequality.

### 3.3.15 Bhatia-Davis Inequality

The Bhatia-Davis inequality also upper bounds the variance of a bounded random variable, like Popoviciu’s inequality. For a random variable  $X$  supported on  $[a, b]$  (which could be an outer-approximation of the actual support), let its mean be  $\mathbb{E}[X] = \mu_X$ . Then

$$\operatorname{Var}(X) \leq (\mu_X - a)(b - \mu_X) \quad (3.3.150)$$

*Proof.* First suppose  $a = 0$  and  $b = 1$ . For all  $x \in [0, 1]$ , note that  $x^2 \leq x$ , so we can see that  $\mathbb{E}[X^2] \leq \mathbb{E}[X]$ . From the definition of the variance:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (3.3.151)$$

$$\leq \mathbb{E}[X] - \mathbb{E}[X]^2 \quad (3.3.152)$$

$$= \mu_X(1 - \mu_X) \quad (3.3.153)$$

Now suppose we have any  $a, b$  in general such that  $a < b$ , and let  $Z = \frac{X-a}{b-a}$  which is supported on  $[0, 1]$  with  $\mathbb{E}[Z] = \frac{\mu_X - a}{b-a}$ . Via the result above:

$$\text{Var}(Z) \leq \frac{\mu_X - a}{b-a} \left(1 - \frac{\mu_X - a}{b-a}\right) \quad (3.3.154)$$

Hence

$$\text{Var}(X) = (b-a)^2 \text{Var}(Z) \quad (3.3.155)$$

$$\leq (\mu_X - a)[b-a - (\mu_X - a)] \quad (3.3.156)$$

$$= (\mu_X - a)(b - \mu_X) \quad (3.3.157)$$

□

The difference between this bound and Popoviciu's inequality is that we require knowledge about the mean, and so we should think that this bound is stronger. This can indeed be verified as true, because the bound  $(\mu_X - a)(b - \mu_X)$  is quadratic in  $\mu_X$ , which is maximised when  $\mu_X = \frac{a+b}{2}$ . Substituting this value back into  $(\mu_X - a)(b - \mu_X)$  will recover Popoviciu's inequality.

## 3.4 Notions of Probabilistic Convergence

### 3.4.1 Convergence in Distribution

Let  $\{X_1, X_2, \dots, X_n\}$  be a sequence of real-valued random variables. The sequence of random variables is said to converge in distribution to a random variable  $X$  if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x) \quad (3.4.1)$$

where  $F_n(x)$  is the cumulative distribution function of  $X_n$  and  $F(x)$  is the cumulative distribution function of  $X$ .

Let  $\mathbf{X}_1, \mathbf{X}_2, \dots$  be real valued random vectors and let  $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  be a sequence of random vectors. The sequence is said to converge in distribution to a random vector  $\mathbf{X}$  if

$$\lim_{n \rightarrow \infty} \Pr(\mathbf{X}_n \in A) = \Pr(\mathbf{X} \in A) \quad (3.4.2)$$

for every  $A \subset \mathbb{R}^n$  which is a continuity set of  $\mathbf{X}$ . Note that it is stronger condition to say it converges in distribution if the cumulative distribution functions converge. If

$$\lim_{n \rightarrow \infty} F_n(x) = F(x) \quad (3.4.3)$$

then this implies convergence in distribution. Convergence in distribution may be denoted with the  $\xrightarrow{d}$  symbol.

### 3.4.2 Convergence in Mean

A sequence of random variables  $\{X_1, X_2, \dots, X_n\}$  converges in mean (or in expectation) towards random variable  $X$  if

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|] = 0 \quad (3.4.4)$$

A sequence of random vectors  $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  converges in mean (or in expectation) towards random vector  $\mathbf{X}$  if

$$\lim_{n \rightarrow \infty} \mathbb{E}[\|\mathbf{X}_n - \mathbf{X}\|] = 0 \quad (3.4.5)$$

### 3.4.3 Convergence in Mean Square

A sequence of random variables  $\{X_1, X_2, \dots, X_n\}$  converges in mean square towards random variable  $X$  if

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^2] = 0 \quad (3.4.6)$$

A sequence of random vectors  $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  converges in mean (or in expectation) towards random vector  $\mathbf{X}$  if

$$\lim_{n \rightarrow \infty} \mathbb{E}[\|\mathbf{X}_n - \mathbf{X}\|^2] = 0 \quad (3.4.7)$$

### 3.4.4 Convergence in $p$ -Mean

Convergence in  $p$ -mean generalises convergence in mean and convergence in mean square. A sequence of random variables  $\{X_1, X_2, \dots, X_n\}$  converges in  $p$ -mean (or in  $L_p$ ) towards random variable  $X$  if for  $p \geq 1$

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^p] = 0 \quad (3.4.8)$$

A sequence of random vectors  $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  converges in  $p$ -mean (or in  $L_p$ ) towards random vector  $\mathbf{X}$  if for  $p \geq 1$

$$\lim_{n \rightarrow \infty} \mathbb{E}[\|\mathbf{X}_n - \mathbf{X}\|^p] = 0 \quad (3.4.9)$$

For  $p \geq s \geq 1$ , convergence in  $p$ -mean implies convergence in  $s$ -mean.

*Proof.* We can write

$$\mathbb{E}[|X_n - X|^p] = \mathbb{E}\left[(|X_n - X|^s)^{p/s}\right] \quad (3.4.10)$$

Because  $x^{p/s}$  will be convex, then by Jensen's inequality

$$(\mathbb{E}[|X_n - X|^s])^{p/s} \leq \mathbb{E}\left[(|X_n - X|^s)^{p/s}\right] \quad (3.4.11)$$

So if  $\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^p] = 0$  then also

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^s] = 0 \quad (3.4.12)$$

□

### 3.4.5 Convergence in Probability

A sequence of random variables  $\{X_1, X_2, \dots, X_n\}$  converges in probability towards random variable  $X$  if

$$\lim_{n \rightarrow \infty} \Pr(|X_n - X| > \varepsilon) = 0 \quad (3.4.13)$$

for all  $\varepsilon > 0$ . If  $X$  is a constant, then the sequence is said to converge in probability to a constant.

A sequence of random vectors  $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  converges in probability towards random vector  $\mathbf{X}$  if

$$\lim_{n \rightarrow \infty} \Pr(\|\mathbf{X}_n - \mathbf{X}\| > \varepsilon) = 0 \quad (3.4.14)$$

for all  $\varepsilon > 0$ . Convergence in probability may be denoted with the  $\xrightarrow{P}$  symbol.

### Sufficient Conditions for Convergence in Probability

If we have a sequence of random variables  $\{X_1, X_2, \dots, X_n\}$  such that

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \mu \quad (3.4.15)$$

$$\lim_{n \rightarrow \infty} \text{Var}(X_n) = 0 \quad (3.4.16)$$

then  $X_n$  converges in probability to  $\mu$ .

*Proof.* Via Chebychev's inequality, we have for all  $\varepsilon > 0$ :

$$\Pr(|X_n - \mathbb{E}[X_n]| \geq \varepsilon) \leq \frac{\text{Var}(X_n)}{\varepsilon^2} \quad (3.4.17)$$

So taking limits,

$$\lim_{n \rightarrow \infty} \Pr(|X_n - \mathbb{E}[X_n]| > \varepsilon) = \lim_{n \rightarrow \infty} \Pr(|X_n - \mu| > \varepsilon) \quad (3.4.18)$$

$$\leq \lim_{n \rightarrow \infty} \Pr(|X_n - \mu| \geq \varepsilon) \quad (3.4.19)$$

$$\leq \lim_{n \rightarrow \infty} \frac{\text{Var}(X_n)}{\varepsilon^2} \quad (3.4.20)$$

$$= 0 \quad (3.4.21)$$

which satisfies the definition of convergence in probability.  $\square$

### Convergence in Probability Implies Convergence in Distribution

Convergence in probability implies convergence in distribution. That is, if  $X_n \xrightarrow{P} X$ , then this also implies  $X_n \xrightarrow{d} X$ . However, the reverse is not necessarily true. For example, a sequence of i.i.d. random variables of  $X$  will satisfy the definition of convergence in distribution, however it does not satisfy the definition of convergence in probability.

### Convergence in $p$ -Mean Implies Convergence in Probability

For  $p \geq 1$ ,  $p$ -mean convergence implies convergence in probability.

*Proof.* If  $X_n$  converges in  $p$ -mean with  $p > 1$ , this implies convergence in mean so we only consider the case with convergence in mean. From Markov's inequality,

$$\Pr(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{\varepsilon} \quad (3.4.22)$$

Hence if  $X_n$  converges in mean (i.e.  $\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|] = 0$ ), then  $\lim_{n \rightarrow \infty} \Pr(|X_n - X| > \varepsilon) = 0$  which is convergence in probability.  $\square$

### 3.4.6 Almost Sure Convergence

While convergence in probability is a condition on the limit of the probability, almost sure convergence is a condition on the probability of the limit. A sequence of random variables  $\{X_1, X_2, \dots, X_n\}$  converges almost surely towards random variable  $X$  if

$$\Pr \left( \lim_{n \rightarrow \infty} X_n = X \right) = 1 \quad (3.4.23)$$

A sequence of random vectors  $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  converges almost surely towards random vector  $\mathbf{X}$  if

$$\Pr \left( \lim_{n \rightarrow \infty} \mathbf{X}_n = \mathbf{X} \right) = 1 \quad (3.4.24)$$

where the equality is evaluated element-wise. Almost sure convergence may be denoted with the  $\xrightarrow{\text{a.s.}}$  symbol.

#### Almost Sure Convergence Implies Converge in Probability

Almost sure convergence implies convergence in probability, however convergence in probability does not imply almost sure convergence. For example, consider the sequence where  $X_n$  takes the value 1 with probability  $\frac{1}{n}$ , and takes the value of 0 otherwise. Although this sequence converges in probability to 0, since the series  $\sum_{n=1}^{\infty} \Pr(X_n = 1) = \sum_{n=1}^{\infty} \frac{1}{n}$  diverges, and each  $X_n$  is independent, we have by the converse Borel-Cantelli lemma that there is probability 1 that the event  $X_n = 1$  occurs infinitely many times. Therefore the sequence does not almost surely converge to 0.

### 3.4.7 Complete Convergence

A sequence of random variables  $\{X_1, X_2, \dots\}$  is said to be completely convergent to  $X$  if

$$\sum_{n=1}^{\infty} \Pr(|X_n - X| > \varepsilon) < \infty \quad (3.4.25)$$

for all  $\varepsilon > 0$ . If  $X_n$  are independent, then complete convergence is equivalent to almost sure convergence.

*Proof.* To show that complete convergence implies almost sure convergence, let event  $E_n$  be  $\{|X_n - X| > \varepsilon\}$  for some  $\varepsilon > 0$ . Then since  $\sum_{n=1}^{\infty} \Pr(|X_n - X| > \varepsilon) < \infty$  by complete convergence, the Borel-Cantelli lemma tells us that

$$\Pr \left( \limsup_{n \rightarrow \infty} \{|X_n - X| > \varepsilon\} \right) = 0 \quad (3.4.26)$$

Since  $\varepsilon$  can be arbitrarily small, then

$$\Pr \left( \limsup_{n \rightarrow \infty} \{|X_n - X| = 0\} \right) = 1 \quad (3.4.27)$$

$$\Pr \left( \lim_{n \rightarrow \infty} X_n = X \right) = 1 \quad (3.4.28)$$

To show that almost sure convergence implies complete convergence, first suppose we have almost sure convergence but not complete converge. This means there exists an  $\varepsilon > 0$  such that

$$\sum_{n=1}^{\infty} \Pr(|X_n - X| > \varepsilon) = \infty \quad (3.4.29)$$

Given that the sequence  $\{X_n\}$  is independent, we can verify that  $\{X_n - X\}$  is also independent as follows. Since  $X_n \rightarrow X$ , then  $X$  is almost sure constant. Hence  $\{X_n - X\}$  is independent. By the converse Borel-Cantelli lemma, this gives for that particular  $\varepsilon > 0$ :

$$\Pr \left( \limsup_{n \rightarrow \infty} \{|X_n - X| > \varepsilon\} \right) = 1 \quad (3.4.30)$$

This contradicts almost sure convergence, hence almost sure convergence implies complete convergence.  $\square$

### 3.4.8 With High Probability

An event  $E$  occurs with high probability if the probability  $\Pr(E)$  depends on some number  $n$ , and  $\Pr(E) \rightarrow 1$  as  $n \rightarrow \infty$ . That is, we can make the probability as close as desired to 1 by making  $n$  big enough.

### 3.4.9 Continuous Mapping Theorem

The continuous mapping theorem states that limits of convergent sequences of random elements are preserved through a continuous mapping. Formally, let  $\{\mathbf{X}_n\}$  be a sequence of random vectors converging in probability to the random vector  $\mathbf{X}$ . Let  $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$  be an arbitrary mapping that is continuous over the support of  $\mathbf{X}$  (where  $d$  is the dimension of  $\mathbf{X}$  and  $m$  is the dimension being mapped to). Then

$$g(\mathbf{X}_n) \xrightarrow{P} g(\mathbf{X}) \quad (3.4.31)$$

Similarly, if  $\{\mathbf{X}_n\}$  is a sequence of random vectors converging almost surely to  $\mathbf{X}$ , then

$$g(\mathbf{X}_n) \xrightarrow{\text{a.s.}} g(\mathbf{X}) \quad (3.4.32)$$

and analogously if  $\{\mathbf{X}_n\}$  is a sequence of random vectors converging in distribution to  $\mathbf{X}$ , then

$$g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X}) \quad (3.4.33)$$

The claims of convergence in probability and almost surely are intuitive, but convergence in distribution is not so obvious. However, an informal explanation is available for why this holds. An equivalent characterisation for convergence in distribution  $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$  is that  $\mathbb{E}[f(\mathbf{X}_n)] \rightarrow \mathbb{E}[f(\mathbf{X})]$  for any bounded continuous function  $f(\cdot)$ . So in some sense, the set of all  $\mathbb{E}[f(\mathbf{X})]$  characterises the entire distribution of  $\mathbf{X}$ . Hence, it suffices to show that  $\mathbb{E}[f \circ g(\mathbf{X}_n)] \rightarrow \mathbb{E}[f \circ g(\mathbf{X})]$  for any bounded continuous  $f(\cdot)$ . But given some continuous  $g(\cdot)$ , then any bounded continuous  $f(\cdot)$  will make  $f \circ g$  also bounded and continuous. Since we already have that  $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ , this shows that  $\mathbb{E}[f \circ g(\mathbf{X}_n)] \rightarrow \mathbb{E}[f \circ g(\mathbf{X})]$ .

### 3.4.10 Slutsky's Theorem

Slutsky's theorem describes properties of algebraic operations on convergent sequences of random variables. Let  $\{X_n\}$ ,  $\{Y_n\}$  be sequences of random variables. Suppose  $\{X_n\}$  converges in distribution to random variable  $X$ , while  $\{Y_n\}$  converges in probability to a constant  $c$ . Then

$$X_n + Y_n \xrightarrow{d} X + c \quad (3.4.34)$$

$$X_n Y_n \xrightarrow{d} cX \quad (3.4.35)$$

$$X_n / Y_n \xrightarrow{d} X/c \quad (3.4.36)$$

where the last property requires that  $c \neq 0$ . Analogous results hold if  $\{X_n\}$  were to converge in probability to  $X$ .

### 3.4.11 Cramér-Wold Theorem [24]

Suppose  $\{\mathbf{X}_n\}$  is a sequence of random vectors converging in distribution to  $\mathbf{X} \in \mathbb{R}^k$ . The continuous mapping theorem allows us to see that

$$\mathbf{a}^\top \mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{X} \quad (3.4.37)$$

for any vector  $\mathbf{a}$ . The Cramér-Wold Theorem is a stronger result which states that  $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$  if and only if  $\mathbf{a}^\top \mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{X}$  for all  $\mathbf{a} \in \mathbb{R}^k$ .

## 3.5 Moments

The  $n^{\text{th}}$  moment of a random variable  $X$  is defined as  $\mathbb{E}[X^n]$ . Thus the 1st moment is simply the expected value while the 2nd moment of a zero-mean random variable is the same as its variance. The integral for the  $n^{\text{th}}$  moment of a continuous random variable  $X$  with probability density function  $f_X(x)$  is given by

$$\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx \quad (3.5.1)$$

If  $X$  is instead a discrete random variable with probability mass function  $\Pr(X = x)$ , then

$$\mathbb{E}[X^n] = \sum_{x=-\infty}^{\infty} x^n \Pr(X = x) \quad (3.5.2)$$

### 3.5.1 Central Moments

The  $n^{\text{th}}$  central moment of a random variable  $X$  is defined as  $\mathbb{E}[(X - \mathbb{E}[X])^n]$ . It is equivalent to the  $n^{\text{th}}$  moment of a random variable defined as  $Y := X - \mu$  where  $\mu := \mathbb{E}[X]$ . Thus for a continuous random variable,

$$\mathbb{E}[(X - \mathbb{E}[X])^n] = \int_{-\infty}^{\infty} (x - \mu)^n f_X(x) dx \quad (3.5.3)$$

Note that the second central moment is identical to the definition of the variance.

### 3.5.2 Standardised Moments

The  $n^{\text{th}}$  standardised moment of a random variable  $X$  is defined as the  $n^{\text{th}}$  central moment of  $X$  divided by the  $n^{\text{th}}$  power of the standard deviation  $\sigma$  of  $X$ . That is,

$$\frac{\mathbb{E}[(X - \mu)^n]}{\sigma^n} = \frac{\mathbb{E}[(X - \mu)^n]}{\left(\sqrt{\mathbb{E}[(X - \mu)^2]}\right)^n} \quad (3.5.4)$$

$$= \frac{\mathbb{E}[(X - \mu)^n]}{\mathbb{E}[(X - \mu)^2]^{n/2}} \quad (3.5.5)$$

By definition, the second standardised moment is 1. The third standardised moment is used to define the skewness of  $X$ , while the fourth standardised moment is used to define the kurtosis of  $X$ .

### 3.5.3 Moment Generating Functions

The moment generating function of a random variable  $X$  is defined as

$$\phi_X(s) = \mathbb{E}[e^{sX}] \quad (3.5.6)$$

That is, it is the expectation of the random variable  $e^{sX}$  as a function of  $s$ . If  $X$  has a probability density function  $f_X(x)$ , then

$$\phi_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx \quad (3.5.7)$$

Notice that

$$\phi_X(-s) = \int_{-\infty}^{\infty} e^{-sx} f_X(x) dx \quad (3.5.8)$$

which is the Laplace transform of  $f_X(x)$ .

The moment generating function can be used to compute the  $n^{\text{th}}$  moment of a random variable. The random variable  $X$  with MGF  $\phi_X(s)$  has  $n^{\text{th}}$  moment

$$\mathbb{E}[X^n] = \left. \frac{d^n \phi_X(s)}{ds^n} \right|_{s=0} \quad (3.5.9)$$

That is, we take the  $n^{\text{th}}$  derivative of the MGF and evaluate it at  $s = 0$ .

*Proof.* Assuming  $X$  is a continuous random variable, the  $n^{\text{th}}$  derivative of the MGF is

$$\frac{d^n \phi_X(s)}{ds^n} = \frac{d^n}{ds^n} \int_{-\infty}^{\infty} e^{sx} f_X(x) dx \quad (3.5.10)$$

$$= \int_{-\infty}^{\infty} \frac{d^n}{ds^n} e^{sx} f_X(x) dx \quad (3.5.11)$$

$$= \int_{-\infty}^{\infty} x^n e^{sx} f_X(x) dx \quad (3.5.12)$$

Evaluating at  $s = 0$  gives

$$\left. \frac{d^n \phi_X(s)}{ds^n} \right|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(x) dx \quad (3.5.13)$$

$$= \mathbb{E}[X^n] \quad (3.5.14)$$

This can be analogously proved for a discrete random variable in the same manner.  $\square$

Note that a random variable is not always guaranteed to have a moment generating function (i.e. when the moments do not exist, such as for a Cauchy random variable).

#### Moment Generating Function of Binomial Distribution

The calculation of the moment generating function for the binomial distribution is given by

$$\phi(s) = \mathbb{E}[e^{sX}] \quad (3.5.15)$$

$$= \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} e^{sx} \quad (3.5.16)$$

$$= \sum_{x=0}^n \binom{n}{x} (pe^s)^x (1-p)^{n-x} \quad (3.5.17)$$

Notice by the binomial theorem that

$$(pe^s + 1 - p)^n = \sum_{x=0}^n \binom{n}{x} (pe^s)^x (1 - p)^{n-x} \quad (3.5.18)$$

Hence

$$\phi(s) = (pe^s + 1 - p)^n \quad (3.5.19)$$

### Moment Generating Function of Gamma Distribution

The moment generating function of the gamma distribution with PDF

$$f(x; a, b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b} \quad (3.5.20)$$

can be calculated by

$$\mathbb{E}[e^{sX}] = \int_0^\infty e^{sx} f(x; a, b) dx \quad (3.5.21)$$

$$= \frac{1}{\Gamma(a)b^a} \int_0^\infty x^{a-1} e^{-x/b+sx} dx \quad (3.5.22)$$

Introducing the substitution  $u = x/b - sx$  so that  $x = \frac{u}{1/b - s}$  and assuming that  $1/b - s > 0$  (so the terminals are unchanged), we have

$$\mathbb{E}[e^{sX}] = \int_0^\infty \left( \frac{u}{1/b - s} \right)^{a-1} e^{-u} \frac{du}{1/b - s} \quad (3.5.23)$$

$$= \frac{1}{(1/b - s)^a} \cdot \frac{1}{\Gamma(a)b^a} \int_0^\infty u^{a-1} e^{-u} du \quad (3.5.24)$$

$$= \frac{1}{(1/b - s)^a} \cdot \frac{1}{\Gamma(a)b^a} \Gamma(a) \quad (3.5.25)$$

where we have used the definition of the [gamma function](#). Simplifying:

$$\mathbb{E}[e^{sX}] = \frac{1}{(1/b - s)^a} \cdot \frac{1}{b^a} \quad (3.5.26)$$

$$= \frac{1}{(1 - bs)^a} \quad (3.5.27)$$

If we however assumed that  $1/b - s \leq 0$ , then  $-x/b + tx \geq 0$  and the integral  $\int_0^\infty x^{a-1} e^{-x/b+sx} dx$  blows up. Therefore, the moment generating function of the gamma distribution is only defined for  $s < 1/b$ .

This moment generating function can be used to compute the mean and variance of the gamma distribution. For the mean,

$$\mathbb{E}[X] = \frac{d}{dt} (1 - bs)^{-a} \Big|_{s=0} \quad (3.5.28)$$

$$= ab(1 - bs)^{-a-1} \Big|_{s=0} \quad (3.5.29)$$

$$= ab \quad (3.5.30)$$

For the variance, start with the second moment:

$$\mathbb{E}[X^2] = \frac{d^2}{dt^2} (1 - bs)^{-a} \Big|_{s=0} \quad (3.5.31)$$

$$= \frac{d}{dt} ab (1 - bs)^{-a-1} \Big|_{s=0} \quad (3.5.32)$$

$$= a(a+1)b^2 (1 - bs)^{-a-2} \Big|_{s=0} \quad (3.5.33)$$

$$= a^2 b^2 + ab^2 \quad (3.5.34)$$

Thus

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (3.5.35)$$

$$= a^2 b^2 + ab^2 - (ab)^2 \quad (3.5.36)$$

$$= ab^2 \quad (3.5.37)$$

### 3.5.4 Sums of Random Variables with Moment Generating Functions

Let  $X = X_1 + \dots + X_n$  be a sum of independent random variables. Then

$$\phi_X(s) = \mathbb{E}\left[e^{s(X_1 + \dots + X_n)}\right] \quad (3.5.38)$$

$$= \mathbb{E}\left[e^{sX_1} \times \dots \times e^{sX_n}\right] \quad (3.5.39)$$

By independence,

$$\phi_X(t) = \mathbb{E}\left[e^{sX_1}\right] \times \dots \times \mathbb{E}\left[e^{sX_n}\right] \quad (3.5.40)$$

$$= \phi_{X_1}(s) \times \dots \times \phi_{X_n}(s) \quad (3.5.41)$$

Hence the moment generating function of the sum of independent random variables is just the product of the moment generating functions.

#### Random Sums with Moment Generating Functions

Given a positive integer valued random variable  $N$  and a sequence of i.i.d. random variables  $X_1, X_2, \dots$ , let the random sum  $R$  be defined by

$$R = X_1 + X_2 + \dots + X_N \quad (3.5.42)$$

**Theorem 3.2.** *If each term  $X$  in a random sum has moment generating function  $\phi_X(s)$  and the number of terms  $N$  has moment generating function  $\phi_N(s)$  which is independent of all the terms, then the random sum  $R$  has moment generating function*

$$\phi_R(s) = \phi_N(\ln \phi_X(s)) \quad (3.5.43)$$

*Proof.* The MGF of  $R$  is given by  $\phi_R(s) = \mathbb{E}[e^{sR}]$ . Using the law of iterated expectations, this is then

$$\phi_R(s) = \sum_{n=0}^{\infty} \mathbb{E}\left[e^{s(X_1 + \dots + X_n)} \mid N = n\right] \Pr(N = n) \quad (3.5.44)$$

By independence of  $N$  from  $X_1, \dots, X_N$ ,

$$\phi_R(s) = \sum_{n=0}^{\infty} \mathbb{E}\left[e^{s(X_1 + \dots + X_n)}\right] \Pr(N = n) \quad (3.5.45)$$

$$= \sum_{n=0}^{\infty} \mathbb{E}\left[e^{sX_1}\right] \times \dots \times \mathbb{E}\left[e^{sX_n}\right] \Pr(N = n) \quad (3.5.46)$$

$$= \sum_{n=0}^{\infty} [\phi_X(s)]^n \Pr(N = n) \quad (3.5.47)$$

Note that we can write  $[\phi_X(x)]^n = [\exp(\ln \phi_X(x))]^n = \exp(\ln \phi_X(s) \cdot n)$ . Hence

$$\phi_R(s) = \sum_{n=0}^{\infty} \exp(\ln \phi_X(s) \cdot n) \Pr(N=n) \quad (3.5.48)$$

This is just the MGF of  $N$  evaluated at  $\ln \phi_X(s)$ . Therefore

$$\phi_R(s) = \phi_N(\ln \phi_X(s)) \quad (3.5.49)$$

□

**Theorem 3.3.** *For the random sum of i.i.d. random variables  $R = X_1 + \dots + X_N$ , we have*

$$\mathbb{E}[R] = \mathbb{E}[X] \mathbb{E}[N] \quad (3.5.50)$$

$$\text{Var}(R) = \mathbb{E}[N] \text{Var}(X) + \text{Var}(N) \mathbb{E}[X]^2 \quad (3.5.51)$$

*Proof.* The MGF of  $R$  is  $\phi_R(s) = \phi_N(\ln \phi_X(s))$ . By the moment generating property of the MGF,  $\mathbb{E}[R]$  is given by the first derivative of  $\phi_R(s)$  evaluated at  $s = 0$ . By applying the chain rule, this yields

$$\phi'_R(s) = \phi'_N(\ln \phi_X(s)) \frac{d}{ds} \ln \phi_X(s) \quad (3.5.52)$$

$$= \phi'_N(\ln \phi_X(s)) \frac{\phi'_X(s)}{\phi_X(s)} \quad (3.5.53)$$

Since  $\phi_X(0) = \mathbb{E}[e^0] = 1$  and  $\phi'_N(0) = \mathbb{E}[N]$  and  $\phi'_X(0) = \mathbb{E}[X]$ , then

$$\mathbb{E}[R] = \phi'_N(0) \phi'_X(0) \quad (3.5.54)$$

$$= \mathbb{E}[X] \mathbb{E}[N] \quad (3.5.55)$$

To find  $\text{Var}(R)$ , we first need to find the second moment  $\mathbb{E}[R^2]$ . The second derivative of  $\phi_R(s)$  using the product and quotient rules is

$$\phi''_R(s) = \phi''_N(\ln \phi_X(s)) \left( \frac{\phi'_X(s)}{\phi_X(s)} \right)^2 + \phi'_N(\ln \phi_X(s)) \frac{d}{ds} \left( \frac{\phi'_X(s)}{\phi_X(s)} \right) \quad (3.5.56)$$

$$= \phi''_N(\ln \phi_X(s)) \left( \frac{\phi'_X(s)}{\phi_X(s)} \right)^2 + \phi'_N(\ln \phi_X(s)) \frac{\phi_X(s) \phi''_X(s) - \phi'_X(s)^2}{\phi_X(s)^2} \quad (3.5.57)$$

Evaluating this at  $s = 0$  gives

$$\mathbb{E}[R^2] = \phi''_N(0) \left( \frac{\phi'_X(0)}{1} \right)^2 + \phi'_N(0) \frac{\phi''_X(0) - \phi'_X(0)^2}{1} \quad (3.5.58)$$

$$= \mathbb{E}[N^2] \mathbb{E}[X]^2 + \mathbb{E}[N] (\mathbb{E}[X^2] - \mathbb{E}[X]^2) \quad (3.5.59)$$

$$= \mathbb{E}[N^2] \mathbb{E}[X]^2 + \mathbb{E}[N] \text{Var}(X) \quad (3.5.60)$$

Then to get  $\text{Var}(R)$ , subtract  $\mathbb{E}[R]^2 = \mathbb{E}[N]^2 \mathbb{E}[X]^2$  from  $\mathbb{E}[R^2]$ .

$$\text{Var}(R) = \mathbb{E}[R^2] - \mathbb{E}[R]^2 \quad (3.5.61)$$

$$= \mathbb{E}[N^2] \mathbb{E}[X]^2 + \mathbb{E}[N] \text{Var}(X) - \mathbb{E}[N]^2 \mathbb{E}[X]^2 \quad (3.5.62)$$

$$= \mathbb{E}[N] \text{Var}(X) + (\mathbb{E}[N^2] - \mathbb{E}[N]^2) \mathbb{E}[X]^2 \quad (3.5.63)$$

$$= \mathbb{E}[N] \text{Var}(X) + \text{Var}(N) \mathbb{E}[X]^2 \quad (3.5.64)$$

□

## Wald's Equation

In the expectation of random sums, the identity  $\mathbb{E}[R] = \mathbb{E}[X]\mathbb{E}[N]$  is known as Wald's equation.

### 3.5.5 Chernoff Bound

Let  $X$  be a random variable and let  $\phi_X(s)$  be the moment generating function of  $X$ . Then for all  $s \geq 0$ ,

$$\Pr(X \geq a) \leq e^{-sa}\phi_X(s) \quad (3.5.65)$$

which also means that

$$\Pr(X \geq a) \leq \min_{s \geq 0} \{e^{-sa}\phi_X(s)\} \quad (3.5.66)$$

*Proof.* From the Markov inequality, we have for some random variable  $Y$  and any  $b > 0$ ,

$$\Pr(Y \geq b) \leq \frac{\mathbb{E}[Y]}{b} \quad (3.5.67)$$

Let  $b = e^{sa} > 0$  for any  $a \in \mathbb{R}$ , and let  $Y = e^{sX}$ . Hence

$$\Pr(e^{sX} \geq e^{sa}) \leq \frac{\mathbb{E}[e^{sX}]}{e^{sa}} \quad (3.5.68)$$

Since  $s \geq 0$ , then  $e^{sX} \geq e^{sa}$  is equivalent to  $X \geq a$ . Also,  $\mathbb{E}[e^{sX}]$  is the definition of the moment generating function of  $X$ . Therefore

$$\Pr(X \geq a) \leq e^{-sa}\phi_X(s) \quad (3.5.69)$$

□

## Okamoto Bound

The Chernoff bound applied to the binomial distribution is known as the Okamoto bound. Letting  $X$  be a binomial distribution random variable, use the moment generating function of the binomial distribution to obtain

$$\Pr(X \geq a) \leq e^{-sa}(pe^s + 1 - p)^n \quad (3.5.70)$$

We can optimise this bound with respect to  $s$ . Let  $u = e^s$ . Differentiating via the product rule:

$$\frac{d}{du}(u^{-a}(pu + 1 - p)^n) = -au^{-a-1}(pu + 1 - p)^n + npu^{-a}(pu + 1 - p)^{n-1} \quad (3.5.71)$$

Setting the derivative equal to zero, this becomes

$$\frac{npu^{-a}}{au^{-a-1}} = \frac{(pu + 1 - p)^n}{(pu + 1 - p)^{n-1}} \quad (3.5.72)$$

from which we simplify and rearrange to get

$$\frac{np}{a}u = pu + 1 - p \quad (3.5.73)$$

$$u\left(\frac{np}{a} - p\right) = 1 - p \quad (3.5.74)$$

$$u = \frac{a}{n-a} \cdot \frac{1-p}{p} \quad (3.5.75)$$

Putting  $u$  back into the bound gives

$$u^{-a} (pu + 1 - p)^n = \left( \frac{a}{n-a} \cdot \frac{1-p}{p} \right)^{-a} \left[ (1-p) \frac{a}{n-a} + 1 - p \right]^n \quad (3.5.76)$$

$$= \left[ \frac{a}{n-a} \cdot \frac{n(1-p)}{np} \right]^{-a} \left[ (1-p) \frac{a+n-a}{n-a} \right]^n \quad (3.5.77)$$

$$= \left( \frac{a}{np} \right)^{-a} \left[ \frac{n(1-p)}{n-a} \right]^{-a} \left[ (1-p) \frac{n}{n-a} \right]^n \quad (3.5.78)$$

$$= \left( \frac{np}{a} \right)^a \left[ \frac{(1-p)n}{n-a} \right]^{n-a} \quad (3.5.79)$$

Thus

$$\Pr(X \geq a) \leq \left( \frac{np}{a} \right)^a \left[ \frac{(1-p)n}{n-a} \right]^{n-a} \quad (3.5.80)$$

Now putting  $a = \mu + t$  where  $\mu = np$  is the mean of the binomial distribution, this yields

$$\Pr(X \geq \mu + t) = \Pr(X - \mu \geq t) \quad (3.5.81)$$

$$\leq \left( \frac{\mu}{\mu+t} \right)^{\mu+t} \left( \frac{n-\mu}{n-\mu-t} \right)^{n-\mu-t} \quad (3.5.82)$$

which is valid for all  $t$  such that  $0 \leq \mu + t \leq n$ . Otherwise, the bound becomes zero, as the binomial distribution is supported on  $\{0, \dots, n\}$ .

### Gaussian Tail Bounds [209]

Consider a Gaussian random variable  $X \sim \mathcal{N}(\mu, \sigma^2)$ . This random variable has a univariate moment generating function (which can be found as a special case of the joint moment generating function):

$$\mathbb{E}[e^{\lambda X}] = \exp \left( \mu\lambda + \frac{1}{2}\sigma^2\lambda^2 \right) \quad (3.5.83)$$

with any  $\lambda \in \mathbb{R}$ . Using the ‘Chernoff technique’ (the same technique used to derive the Chernoff bound) on the zero-mean  $X - \mu$ , we get

$$\Pr(X - \mu > t) = \Pr(e^{\lambda(X-\mu)} > e^{\lambda t}) \quad (3.5.84)$$

$$\leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} \quad (3.5.85)$$

$$= \exp \left( \frac{\sigma^2\lambda^2}{2} - \lambda t \right) \quad (3.5.86)$$

using Markov’s inequality along the way. To optimise this bound with choice of  $\lambda$ , observe that

$$\operatorname{argmin}_{\lambda \geq 0} \left\{ \frac{\sigma^2\lambda^2}{2} - \lambda t \right\} = \frac{t}{\sigma^2} \quad (3.5.87)$$

for all  $t \geq 0$ . Thus the minimum becomes

$$\min_{\lambda \geq 0} \left\{ \frac{\sigma^2\lambda^2}{2} - \lambda t \right\} = \frac{\sigma^2\lambda^2}{2} - \lambda t \Big|_{\lambda=t/\sigma^2} \quad (3.5.88)$$

$$= \frac{t^2}{2\sigma^2} - \frac{t^2}{\sigma^2} \quad (3.5.89)$$

$$= -\frac{t^2}{2\sigma^2} \quad (3.5.90)$$

and the one-sided Gaussian upper tail bound is

$$\Pr(X - \mu > t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad (3.5.91)$$

for all  $t \geq 0$ . We can exploit the symmetry of the Gaussian (i.e.  $X - \mu$  has the same distribution as  $-(X - \mu)$ ) to arrive at a similar bound for the lower tail, and then combine the two bounds using the union bound (Boole's inequality) to give the two-sided Gaussian tail bound

$$\Pr(|X - \mu| > t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad (3.5.92)$$

for all  $t \geq 0$ .

### 3.5.6 Hoeffding's Lemma

Hoeffding's lemma bounds the moment generating function  $\mathbb{E}[e^{\lambda X}]$  of a bounded random variable  $X$ .

**Lemma 3.1.** *Let  $X$  be a random variable with expectation  $\mathbb{E}[X] = 0$  and is bounded by  $a \leq X \leq b$  almost surely. Then for all  $\lambda \in \mathbb{R}$ ,*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left[\frac{\lambda^2(b-a)^2}{8}\right] \quad (3.5.93)$$

*Proof.* Since  $e^{\lambda X}$  is convex in  $X$ , taking a convex combination of  $e^{\lambda a}$  and  $e^{\lambda b}$  gives for all  $a \leq X \leq b$ :

$$e^{\lambda X} \leq \frac{b-X}{b-a}e^{\lambda a} + \frac{X-a}{b-a}e^{\lambda b} \quad (3.5.94)$$

Taking the expectation of both sides:

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b-\mathbb{E}[X]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[X]-a}{b-a}e^{\lambda b} \quad (3.5.95)$$

Since  $\mathbb{E}[X] = 0$ , then

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} \quad (3.5.96)$$

Now by letting  $h = \lambda(b-a)$ ,  $p = \frac{-a}{b-a}$  and  $L(h) = -hp + \ln(1-p+pe^h)$ , we can show that

$\frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{L(h)}$  as follows:

$$e^{L(h)} = e^{-hp+\ln(1-p+pe^h)} \quad (3.5.97)$$

$$= e^{-hp}e^{\ln(1-p+pe^h)} \quad (3.5.98)$$

$$= (1-p+pe^h)e^{-hp} \quad (3.5.99)$$

$$= \left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)}\right) \exp\left[-\lambda(b-a)\frac{-a}{b-a}\right] \quad (3.5.100)$$

$$= \left(\frac{b}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)}\right)e^{\lambda a} \quad (3.5.101)$$

$$= \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} \quad (3.5.102)$$

Differentiating  $L(h)$ , we find

$$L'(h) = -p + \frac{pe^h}{1-p+pe^h} \quad (3.5.103)$$

and by the product rule

$$L''(h) = \frac{pe^h}{1-p+pe^h} + \frac{pe^h(pe^h)}{-(1-p+pe^h)^2} \quad (3.5.104)$$

$$= \frac{pe^h}{1-p+pe^h} \left( 1 - \frac{pe^h}{1-p+pe^h} \right) \quad (3.5.105)$$

Note that  $L(0) = 0$  and  $L'(0) = 0$ . Letting  $t = \frac{pe^h}{1-p+pe^h}$ , we also see that

$$L''(h) = t(1-t) \quad (3.5.106)$$

$$\leq \frac{1}{4} \quad (3.5.107)$$

which can be deduced by considering the graph of the parabola  $t(1-t)$ . By applying Taylor's theorem about zero, we have for some  $\theta \in (0, 1)$ :

$$L(h) = L(0) + hL'(0) + \frac{1}{2}h^2L''(\theta h) \quad (3.5.108)$$

$$= \frac{1}{2}h^2L''(\theta h) \quad (3.5.109)$$

$$\leq \frac{1}{8}\lambda^2(b-a)^2 \quad (3.5.110)$$

Therefore

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} \quad (3.5.111)$$

$$= e^{L(h)} \quad (3.5.112)$$

$$= \exp\left[\frac{\lambda^2(b-a)^2}{8}\right] \quad (3.5.113)$$

□

### 3.5.7 Hoeffding's Inequality

Hoeffding's lemma can be used to derive Hoeffding's inequality for a sum of bounded independent random variables (they need not necessarily be identical).

**Theorem 3.4.** *Let  $X_1, \dots, X_n$  be independent random variables bounded by  $a_i \leq X_i \leq b_i$  almost surely for all  $i = 1, \dots, n$ . Let  $S = \sum_{i=1}^n X_i$ . Then for all  $t > 0$ ,*

$$\Pr(|S - \mathbb{E}[S]| \geq t) \leq 2 \exp\left[-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right] \quad (3.5.114)$$

*Proof.* For any random variable  $Z$  and  $s > 0$ ,

$$\Pr(Z \geq t) = \Pr(e^{sZ} \geq e^{st}) \quad (3.5.115)$$

since  $e^{sz}$  is monotonically increasing in  $z$ . Then apply Markov's inequality to the non-negative random variable  $e^{sz}$ :

$$\Pr(e^{sZ} \geq e^{st}) \leq \frac{\mathbb{E}[e^{sZ}]}{e^{st}} \quad (3.5.116)$$

$$= e^{-st} \mathbb{E}[e^{sZ}] \quad (3.5.117)$$

Choose  $Z = S - \mathbb{E}[S]$ , so

$$\Pr\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t\right) \leq e^{-st} \mathbb{E}\left[\exp\left[s \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right]\right] \quad (3.5.118)$$

$$= e^{-st} \mathbb{E}\left[\prod_{i=1}^n \exp[s(X_i - \mathbb{E}[X_i])] \right] \quad (3.5.119)$$

$$= e^{-st} \prod_{i=1}^n \mathbb{E}[\exp[s(X_i - \mathbb{E}[X_i])]] \quad (3.5.120)$$

where the latter equality is due to independence of all the  $X_i$ . Note that  $\mathbb{E}[\exp[s(X_i - \mathbb{E}[X_i])]]$  is the moment generating function of  $X_i - \mathbb{E}[X_i]$ . Hence applying Hoeffding's lemma,

$$\Pr\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t\right) \leq e^{-st} \prod_{i=1}^n \exp\left[\frac{s^2(b_i - a_i)^2}{8}\right] \quad (3.5.121)$$

$$= \exp\left[-st + s^2 \sum_{i=1}^n \frac{(b_i - a_i)^2}{8}\right] \quad (3.5.122)$$

We can choose the value of  $s > 0$  which gives the best bound. This amounts to minimising the term  $-st + s^2 \sum_{i=1}^n \frac{(b_i - a_i)^2}{8}$ , which is a quadratic in  $s$  with coefficients  $\sum_{i=1}^n \frac{(b_i - a_i)^2}{8}$  and  $-t$ . Hence the choice of

$$s = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2} \quad (3.5.123)$$

minimises the upper bound. Substituting this value of  $s$  gives

$$\Pr(S - \mathbb{E}[S] \geq t) = \Pr\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq e^{st}\right) \quad (3.5.124)$$

$$\leq \exp\left[-\frac{4t^2}{\sum_{i=1}^n (b_i - a_i)^2} + \frac{16t^2}{\left[\sum_{i=1}^n (b_i - a_i)^2\right]^2} \sum_{i=1}^n \frac{(b_i - a_i)^2}{8}\right] \quad (3.5.125)$$

$$= \exp\left[-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right] \quad (3.5.126)$$

Instead if we chose  $Z = -S + \mathbb{E}[S]$ , we get

$$\Pr(S - \mathbb{E}[S] \leq -t) = \Pr(-S + \mathbb{E}[S] \geq t) \quad (3.5.127)$$

$$= \Pr(Z \geq t) \quad (3.5.128)$$

$$\leq \exp\left[-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right] \quad (3.5.129)$$

therefore combining these two inequalities using Boole's inequality gives

$$\Pr(|S - \mathbb{E}[S]| \geq t) \leq 2 \exp\left[-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right] \quad (3.5.130)$$

□

A special case of Hoeffding's inequality can also be stated in terms of the deviation of the i.i.d. sample mean from the expectation.

**Corollary 3.3.** *Let  $X_1, \dots, X_n$  be i.i.d. random variables bounded by  $a \leq X_i \leq b$  almost surely for all  $i = 1, \dots, n$ . Let  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Then for all  $t > 0$ ,*

$$\Pr(|\bar{X} - \mathbb{E}[X]| \geq t) \leq 2 \exp\left[-\frac{2nt^2}{(b-a)^2}\right] \quad (3.5.131)$$

*Proof.* Let  $t' = nt$ , then use Hoeffding's inequality stated for sums:

$$\Pr(|\bar{X} - \mathbb{E}[X]| \geq t) = \Pr\left(\frac{1}{n}|S - \mathbb{E}[S]| \geq t\right) \quad (3.5.132)$$

$$= \Pr(|S - \mathbb{E}[S]| \geq t') \quad (3.5.133)$$

$$\leq 2 \exp\left[-\frac{2t'^2}{n(b-a)^2}\right] \quad (3.5.134)$$

$$= 2 \exp\left[-\frac{2nt^2}{(b-a)^2}\right] \quad (3.5.135)$$

□

### 3.5.8 Joint Moment Generating Functions

Joint moment generating functions extend moment generating functions to random vectors. For a random vector  $\mathbf{X}$ , the joint moment generating function (if it exists) is defined by

$$\phi_{\mathbf{X}}(\mathbf{s}) = \mathbb{E}\left[\exp(\mathbf{s}^\top \mathbf{X})\right] \quad (3.5.136)$$

which is a real-valued function in the multivariate  $\mathbf{s}$ .

#### Joint Moment Generating Function of Independent Random Variables

Suppose  $X_1, \dots, X_n$  are independent random variables. Then the joint moment generating function of the random vector  $\mathbf{X} = (X_1, \dots, X_n)$  is given by

$$\phi_{\mathbf{X}}(\mathbf{s}) = \mathbb{E}\left[\exp(\mathbf{s}^\top \mathbf{X})\right] \quad (3.5.137)$$

$$= \mathbb{E}\left[\exp\left(\sum_{i=1}^n s_i X_i\right)\right] \quad (3.5.138)$$

$$= \mathbb{E}\left[\prod_{i=1}^n e^{s_i X_i}\right] \quad (3.5.139)$$

$$= \prod_{i=1}^n \mathbb{E}[e^{s_i X_i}] \quad (3.5.140)$$

$$= \prod_{i=1}^n \phi_{X_i}(s_i) \quad (3.5.141)$$

Hence this shows that the joint MGF of independent random variables is given by the product of the MGFs of the individual random variables.

### Joint Moment Generating Function of Linear Transformations

Suppose  $\mathbf{X}$  has moment generating function  $\phi_{\mathbf{X}}(\mathbf{s})$ . Consider the linear transformation  $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ . Then  $\mathbf{Y}$  has joint moment generating function given by

$$\phi_{\mathbf{Y}}(\mathbf{s}) = \mathbb{E} \left[ \exp \left( \mathbf{s}^\top \mathbf{Y} \right) \right] \quad (3.5.142)$$

$$= \mathbb{E} \left[ \exp \left( \mathbf{s}^\top (A\mathbf{X} + \mathbf{b}) \right) \right] \quad (3.5.143)$$

$$= \mathbb{E} \left[ \exp \left( \mathbf{s}^\top \mathbf{b} \right) \exp \left( \mathbf{s}^\top A\mathbf{X} \right) \right] \quad (3.5.144)$$

$$= \mathbb{E} \left[ \exp \left( \mathbf{s}^\top \mathbf{b} \right) \exp \left( (A^\top \mathbf{s})^\top \mathbf{X} \right) \right] \quad (3.5.145)$$

$$= \exp \left( \mathbf{s}^\top \mathbf{b} \right) \phi_{\mathbf{X}} \left( A^\top \mathbf{s} \right) \quad (3.5.146)$$

### Cross-Moments from Joint Moment Generating Functions

Cross-moments of a random vector can be computed from the joint moment generating function. Suppose random vector  $\mathbf{X} = (X_1, \dots, X_n)$  possesses a joint moment generating function  $\phi_{\mathbf{X}}(s_1, \dots, s_n)$ , and suppose the  $k^{\text{th}}$  order cross moment of interest is

$$\mu_{\mathbf{X}}(k_1, \dots, k_n) = \mathbb{E} \left[ X_1^{k_1} \cdot X_2^{k_2} \cdots \cdots X_n^{k_n} \right] \quad (3.5.147)$$

with  $k_1 + \cdots + k_n = k$ . By taking a  $k^{\text{th}}$  order partial derivative of the joint moment generating function, we can show:

$$\frac{\partial^k \phi_{\mathbf{X}}(s_1, \dots, s_n)}{\partial s_1^{k_1} \cdots \partial s_n^{k_n}} = \frac{\partial^k}{\partial s_1^{k_1} \cdots \partial s_n^{k_n}} (\mathbb{E} [e^{s_1 X_1 + \cdots + s_n X_n}]) \quad (3.5.148)$$

$$= \mathbb{E} \left[ \frac{\partial^k}{\partial s_1^{k_1} \cdots \partial s_n^{k_n}} (e^{s_1 X_1 + \cdots + s_n X_n}) \right] \quad (3.5.149)$$

$$= \mathbb{E} \left[ X_1^{k_1} \cdot X_2^{k_2} \cdots \cdots X_n^{k_n} e^{s_1 X_1 + \cdots + s_n X_n} \right] \quad (3.5.150)$$

Then evaluating this partial derivative at  $s_1 = \cdots = s_n = 0$ :

$$\left. \frac{\partial^k \phi_{\mathbf{X}}(s_1, \dots, s_n)}{\partial s_1^{k_1} \cdots \partial s_n^{k_n}} \right|_{s_1=\cdots=s_n=0} = \mathbb{E} \left[ X_1^{k_1} \cdot X_2^{k_2} \cdots \cdots X_n^{k_n} e^{0 \cdot X_1 + \cdots + 0 \cdot X_n} \right] \quad (3.5.151)$$

$$= \mathbb{E} \left[ X_1^{k_1} \cdot X_2^{k_2} \cdots \cdots X_n^{k_n} \right] \quad (3.5.152)$$

$$= \mu_{\mathbf{X}}(k_1, \dots, k_n) \quad (3.5.153)$$

which gives the desired cross-moment.

## 3.6 Probability Generating Functions

The probability generating function of a discrete random variable  $X$  taking on values in the non-negative integers with probability mass function  $\Pr(X = x)$  is defined by

$$\psi_X(z) = \mathbb{E} [z^X] \quad (3.6.1)$$

$$= \sum_{x=0}^{\infty} z^x \Pr(X = x) \quad (3.6.2)$$

which is the  $z$ -transform of the probability mass function.

### 3.6.1 Sums of Random Variables with Probability Generating Functions

**Theorem 3.5.** Let  $X = X_1 + \dots + X_n$  be a sum of independent non-negative random variables. Then

$$\psi_X(z) = \psi_{X_1}(z) \times \dots \times \psi_{X_n}(z) \quad (3.6.3)$$

*Proof.* By definition,

$$\psi_X(z) = \mathbb{E}[z^{X_1+\dots+X_n}] \quad (3.6.4)$$

$$= \mathbb{E}[z^{X_1} \times \dots \times z^{X_n}] \quad (3.6.5)$$

Using independence,

$$\psi_X(z) = \mathbb{E}[z^{X_1}] \times \dots \times \mathbb{E}[z^{X_n}] \quad (3.6.6)$$

$$= \psi_{X_1}(z) \times \dots \times \psi_{X_n}(z) \quad (3.6.7)$$

□

### 3.6.2 Probability Generating Function of Poisson Distribution

For a Poisson random variable  $X$  with parameter  $\lambda$ , its probability generating function is given by

$$\psi_X(z) = \sum_{x=0}^{\infty} z^x \Pr(X=x) = \sum_{x=0}^{\infty} z^x \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(z\lambda)^x}{x!} = e^{-\lambda} e^{z\lambda} \quad (3.6.8)$$

$$= e^{\lambda(z-1)} \quad (3.6.9)$$

where we used the power series expansion for the exponential. This can be used to show that the Poisson distribution is indeed a stable distribution. Let  $X \sim \text{Poisson}(\lambda_X)$  and  $Y \sim \text{Poisson}(\lambda_Y)$ , where  $X$  and  $Y$  are independent. Then consider the sum  $W = X + Y$ , which has probability generating function given by

$$\psi_W(z) = \psi_X(z) \psi_Y(z) \quad (3.6.10)$$

$$= e^{\lambda_X(z-1)} e^{\lambda_Y(z-1)} \quad (3.6.11)$$

$$= e^{(\lambda_X + \lambda_Y)(z-1)} \quad (3.6.12)$$

which is the probability generating function of a Poisson( $\lambda_X + \lambda_Y$ ) random variable. Therefore, the sum of independent Poisson random variables is also Poisson distributed, with their parameters added together.

## 3.7 Characteristic Functions

The characteristic function of a random variable  $X$  is a complex-valued function in a real-valued argument  $t$ , defined as

$$\varphi_X(t) = \mathbb{E}[e^{-itX}] \quad (3.7.1)$$

where  $i = \sqrt{-1}$ . By Euler's formula, we can see that it is complex-valued.

$$\varphi_X(t) = \mathbb{E}[\cos(tx)] + i\mathbb{E}[\sin(tx)] \quad (3.7.2)$$

If  $X$  has a probability density function  $f_X(x)$ , then the characteristic function is

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{-itx} f_X(x) dx \quad (3.7.3)$$

which is the Fourier transform of  $f_X(x)$ . Hence the characteristic function encodes the full distribution of  $X$ . The characteristic function of a random variable always exists. We can show this by considering a change of variables  $x = Q(p)$  where  $Q(p)$  is the quantile function. So

$$p = Q^{-1}(x) = F_X(x) \quad (3.7.4)$$

$$\frac{dp}{dx} = f_X(x) \quad (3.7.5)$$

$$dp = f_X(x) dx \quad (3.7.6)$$

Hence

$$\int_{-\infty}^{\infty} e^{-itx} f_X(x) dx = \int_0^1 e^{-itQ(p)} dp \quad (3.7.7)$$

This is a proper integral, and since the magnitude of the integrand  $|e^{-itQ(p)}|$  is bounded from looking at Euler's formula, this integral exists.

If the moment generating function  $\phi_X(t)$  of  $X$  exists, then the relationship between the characteristic function and the moment generating function is given by

$$\varphi_X(t) = \phi_X(-it) \quad (3.7.8)$$

$$\phi_X(t) = \varphi_X(-it) \quad (3.7.9)$$

### 3.7.1 Sums of Random Variables with Characteristic Functions

Let  $X = X_1 + \dots + X_n$  be a sum of independent random variables. Then

$$\varphi_X(t) = \mathbb{E} [e^{-it(X_1 + \dots + X_n)}] \quad (3.7.10)$$

$$= \mathbb{E} [e^{-itX_1} \times \dots \times e^{-itX_n}] \quad (3.7.11)$$

By independence,

$$\varphi_X(t) = \mathbb{E} [e^{-itX_1}] \times \dots \times \mathbb{E} [e^{-itX_n}] \quad (3.7.12)$$

$$= \varphi_{X_1}(t) \times \dots \times \varphi_{X_n}(t) \quad (3.7.13)$$

Hence the characteristic function of the sum of independent random variables is just the product of the characteristic functions. This property is shared with the moment generating function.

### 3.7.2 Subindependence

Using characteristic functions, ‘subindependence’ can be defined as a weaker version of independence and uncorrelatedness. Two random variables  $X$  and  $Y$  are said to be subindependent if their characteristic functions satisfy

$$\varphi_{X+Y}(t) = \varphi_X(t) \varphi_Y(t) \quad (3.7.14)$$

Hence if  $X$  and  $Y$  are independent, then they are also subindependent. However if  $X$  and  $Y$  are subindependent, they are not necessarily independent, nor even uncorrelated (because the covariance may not exist, for example in the case of Cauchy random variables). However if  $X$  and  $Y$  are subindependent and the covariance exists, then they are uncorrelated.

### 3.7.3 Characteristic Functions of Gaussians

The characteristic function of a Gaussian is another (unnormalised) Gaussian. We can show this as follows for a standard Gaussian random variable  $X \sim \mathcal{N}(0, 1^2)$ . The characteristic function is given by

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-itx} e^{-x^2/2} dx \quad (3.7.15)$$

A slight (although not necessary) simplification is to write

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} (\cos(tx) - i \sin(tx)) dx \quad (3.7.16)$$

$$= \frac{1}{\sqrt{2\pi}} \left( \int_{-\infty}^{\infty} e^{-x^2/2} \cos(tx) dx - i \int_{-\infty}^{\infty} e^{-x^2/2} \sin(tx) dx \right)^0 \quad (3.7.17)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \cos(tx) dx \quad (3.7.18)$$

since  $e^{-x^2/2} \sin(tx)$  is an odd function. We then apply the trick of differentiating  $\varphi_X(t)$ :

$$\varphi'_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -e^{-x^2/2} x \sin(tx) dx \quad (3.7.19)$$

Let  $v' = -xe^{-x^2/2}$  and  $u = \sin(tx)$  so  $v = e^{-x^2/2}$  and  $u' = t \cos(tx)$ . Differentiating by parts, this gives

$$\varphi'_X(t) = \frac{1}{\sqrt{2\pi}} \left( [uv]_{x=-\infty}^{x=\infty} - \int_{-\infty}^{\infty} u' v dx \right) \quad (3.7.20)$$

$$= \frac{1}{\sqrt{2\pi}} \left( \left[ e^{-x^2/2} \sin(tx) \right]_{x=-\infty}^{x=\infty} - \int_{-\infty}^{\infty} t \cos(tx) e^{-x^2/2} dx \right) \quad (3.7.21)$$

$$= \frac{-t}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \cos(tx) dx \quad (3.7.22)$$

$$= -t \varphi_X(t) \quad (3.7.23)$$

This is a differential equation with solution  $\varphi_X(t) = e^{-t^2/2}$ , which can be shown through direct verification.

## 3.8 Cumulants

### 3.8.1 Cumulant Generating Functions [186]

The cumulant generating function  $K_X(t)$  of a random variable  $X$  is defined as the natural logarithm of the moment generating function.

$$K_X(t) = \log \phi_X(t) \quad (3.8.1)$$

$$= \log \mathbb{E}[e^{tX}] \quad (3.8.2)$$

By a Taylor series expansion about  $t = 0$ , the cumulant generating function can be written as

$$K_X(t) = \kappa_0 + \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \dots \quad (3.8.3)$$

The  $n^{\text{th}}$  cumulant of  $X$  is then defined as the  $n^{\text{th}}$  coefficient of the series expansion for  $K_X(t)$ , which is  $\kappa_n$ . This is the same as taking the  $n^{\text{th}}$  derivative of  $K(t)$  and evaluating at zero. Note that  $\kappa_0 = 0$  since  $\phi_X(0) = 1$ .

**Theorem 3.6.** Let  $X = X_1 + \dots + X_n$  be a sum of independent random variables. Then

$$K_X(t) = K_{X_1}(t) + \dots + K_{X_n}(t) \quad (3.8.4)$$

*Proof.* Using the definition of the cumulant generating function and the property of the moment generating function for the sum  $X$ ,

$$K_X(t) = \log \phi_X(t) \quad (3.8.5)$$

$$= \log \prod_{i=1}^n \phi_{X_i}(t) \quad (3.8.6)$$

$$= \sum_{i=1}^n \log \phi_{X_i}(t) \quad (3.8.7)$$

$$= \sum_{i=1}^n K_{X_i}(t) \quad (3.8.8)$$

□

### Cumulant Generating Function of Gamma Distribution

By taking the log of the moment generating function of the gamma distribution, we get

$$\log \mathbb{E}[e^{tX}] = \log(1 - bt)^{-a} \quad (3.8.9)$$

$$= -a \log(1 - bt) \quad (3.8.10)$$

Now consider the cumulant generating function of the centered gamma random variable  $X - \mathbb{E}[X]$ . Since  $\mathbb{E}[X] = ab$  (which can be computed through the moment generating function), then

$$\log \mathbb{E}[e^{t(X - \mathbb{E}[X])}] = \log \mathbb{E}[e^{t(X - ab)}] \quad (3.8.11)$$

$$= \log e^{-abt} + \log \mathbb{E}[e^{tX}] \quad (3.8.12)$$

$$= -abt - a \log(1 - bt) \quad (3.8.13)$$

An upper bound can be derived for this cumulant generating function. First, we claim that for  $z \in (0, 1)$ , we have

$$-\log(1 - z) = \sum_{n=1}^{\infty} \frac{z^n}{n} \quad (3.8.14)$$

$$= x + \frac{x^2}{2} + \frac{x^3}{3} + \dots \quad (3.8.15)$$

This can be verified by differentiating both sides

$$\frac{d}{dz}(-\log(1 - z)) = \frac{d}{dz} \left( \sum_{n=1}^{\infty} \frac{z^n}{n} \right) \quad (3.8.16)$$

Thus

$$\frac{1}{1 - z} = \sum_{n=1}^{\infty} z^{n-1} = \sum_{n=0}^{\infty} z^n \quad (3.8.17)$$

which is the geometric series. So for  $z \in (0, 1)$ ,

$$-z - \log(1 - z) = \frac{z^2}{2} + \frac{z^3}{3} + \dots \quad (3.8.18)$$

$$\leq \frac{z^2}{2} + \frac{z^3}{2} + \dots \quad (3.8.19)$$

$$= \frac{z^2}{2} (1 + z + z^2 + \dots) \quad (3.8.20)$$

$$= \frac{z^2}{2} \cdot \frac{1}{1-z} \quad (3.8.21)$$

Now putting  $z = bt$ , where  $z \in (0, 1)$  implies that  $t \in (0, 1/b)$  (which the moment generating function is defined over), we get

$$\log \mathbb{E} [e^{t(X - \mathbb{E}[X])}] = -a(bt - \log(1 - bt)) \quad (3.8.22)$$

$$\leq a \frac{(bt)^2}{2(1 - bt)} \quad (3.8.23)$$

$$= \frac{t^2 ab^2}{2(1 - bt)} \quad (3.8.24)$$

$$= \frac{t^2 v}{2(1 - bt)} \quad (3.8.25)$$

where  $v = ab^2$  is the variance of the gamma distribution (which can also be computed via the moment generating function).

### Relation Between Moments and Cumulants

The relation between moments and cumulants can be analysed as follows. Denote the  $n^{\text{th}}$  moment of  $X$  by  $\mu_n$ . A Taylor series expansion of the moment generating function can be written as

$$\phi_X(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \dots \quad (3.8.26)$$

then writing  $\phi_X(t) = \exp K(t)$  gives

$$1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \dots = \exp \left( \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \dots \right) \quad (3.8.27)$$

$$= \exp(\kappa_1 t) \exp \left( \kappa_2 \frac{t^2}{2!} \right) \exp \left( \kappa_3 \frac{t^3}{3!} \right) \dots \quad (3.8.28)$$

$$= \prod_{i=1}^n \exp \left( \kappa_i \frac{t^i}{i!} \right) \quad (3.8.29)$$

Using the series expansion for  $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$ ,

$$1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \dots = \prod_{i=1}^n \left( 1 + \kappa_i \frac{t^i}{i!} + \kappa_i^2 \frac{t^{2i}}{2!(i!)^2} + \dots \right) \quad (3.8.30)$$

Collecting powers of  $t$  and equating, we can see that

$$\mu_1 t = \kappa_1 t \quad (3.8.31)$$

$$\mu_2 \frac{t^2}{2!} = \frac{(\kappa_1^2 + \kappa_2) t^2}{2!} \quad (3.8.32)$$

$$\mu_3 \frac{t^3}{3!} = \frac{\kappa_3 t^3}{3!} + \frac{\kappa_1^3 t^3}{3!} + \kappa_1 t \frac{\kappa_2 t^2}{2!} \quad (3.8.33)$$

Hence

$$\mu_1 = \kappa_1 \quad (3.8.34)$$

$$\mu_2 = \kappa_1^2 + \kappa_2 \quad (3.8.35)$$

$$\mu_3 = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3 \quad (3.8.36)$$

From this, we get  $\kappa_2 = \mu_2 - \kappa_1^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$  which is the same as the variance of  $X$ . The third cumulant is given by

$$\kappa_3 = \mu_3 - 3\kappa_1\kappa_2 - \kappa_1^3 \quad (3.8.37)$$

which is identical to the third central moment because

$$\mathbb{E}[(X - \mu_1)^3] = \mathbb{E}[X^3 - 3X^2\mu_1 + 3X\mu_1^2 - \mu_1^3] \quad (3.8.38)$$

$$= \mathbb{E}[X^3] - 3\mu_1(\mathbb{E}[X^2] - \mu_1^2) - \mu_1^3 \quad (3.8.39)$$

$$= \mathbb{E}[X^3] - 3\mu_1 \text{Var}(X) - \mu_1^3 \quad (3.8.40)$$

$$= \mu_3 - 3\kappa_1\kappa_2 - \kappa_1^3 \quad (3.8.41)$$

However,  $n^{\text{th}}$  cumulants higher than the third are generally not equal to the  $n^{\text{th}}$  central moments.

### 3.8.2 Law of Total Cumulance

The Law of Total Cumulance generalises the Law of Iterated Expectations and the Law of Total Covariance.

## 3.9 Exponential Families

The exponential families are a parametric class of distributions, which have probability density or mass functions that can be written in a particular way involving the exponential function.

### 3.9.1 Single-Parameter Exponential Families

Let  $f(x; \theta)$  be the density or mass function of a random variable  $X$  with a single parameter  $\theta$ . Then the distribution is part of the exponential family if it can be written in the form

$$f(x; \theta) = h(x) \exp(\eta(\theta) \cdot T(x) - \psi(\theta)) \quad (3.9.1)$$

where  $h(x)$  is known as a *carrier function* [75]. The function  $T(x)$  is a sufficient statistic for  $\theta$ . To see why, factorise  $f(x; \theta)$  as

$$f(x; \theta) = h(x) g(T(x), \theta) \quad (3.9.2)$$

where  $g(T(x), \theta) = \exp(\eta(\theta) \cdot T(x) - \psi(\theta))$ . Then by the Fisher-Neyman factorisation theorem, this shows that  $T(x)$  is a sufficient statistic for  $\theta$ . An equivalent representation of the exponential family is

$$f(x; \theta) = \frac{h(x)}{\phi(\theta)} \exp(\eta(\theta) \cdot T(x)) \quad (3.9.3)$$

where  $\phi(\theta) := \exp(\psi(\theta))$  is known as the *partition function*. The ratio  $\frac{1}{\phi(\theta)}$  essentially determines the normalising constant of the distribution, since we must have

$$\phi(\theta) = \int h(x) \exp(\eta(\theta) \cdot T(x)) dx \quad (3.9.4)$$

in order for  $\int f(x; \theta) dx = 1$  if  $f(x; \theta)$  is a probability density function, and analogously if  $f(x; \theta)$  were a probability mass function. Another property of exponential families is that the support of  $X$  does not depend on  $\theta$ .

## Binomial Exponential Family

The Binomial  $(n, p)$  distribution is generally not a member of the exponential family, because the parameter  $n$  affects the support of the distribution. However, if  $n$  is a fixed and known value (i.e. it is no longer considered a parameter), then the binomial distribution is part of the exponential family. To show this, let  $\theta = p$  and write

$$f(x; p) = \binom{n}{x} p^x (1-p)^{n-x} \quad (3.9.5)$$

$$= \binom{n}{x} (1-p)^n \left( \frac{p}{1-p} \right)^x \quad (3.9.6)$$

$$= \underbrace{\binom{n}{x}}_{h(x)} \underbrace{(1-p)^n}_{1/\phi(\theta)} \exp \left( x \log \left( \frac{p}{1-p} \right) \right) \quad (3.9.7)$$

with  $T(x)$  and  $\eta(\theta) = \log \left( \frac{p}{1-p} \right)$ .

## Geometric Exponential Family

The geometric distribution with single parameter  $\theta = p$  is a member of the exponential family, since we can write its probability mass function as

$$f(x; p) = (1-p)^x p \quad (3.9.8)$$

$$= \underbrace{p}_{1/\phi(\theta)} \exp(x \log(1-p)) \quad (3.9.9)$$

with  $h(x) = 1$ ,  $T(x) = x$  and  $\eta(\theta) = \log(1-p)$ .

### 3.9.2 Multiple-Parameter Exponential Families

Parametric distributions with multiple parameters (represented by the vector  $\boldsymbol{\theta} \in \mathbb{R}^k$ ) are part of the exponential family if they can be written as

$$f(x; \boldsymbol{\theta}) = h(x) \exp \left( \eta(\boldsymbol{\theta})^\top T(x) - \psi(\boldsymbol{\theta}) \right) \quad (3.9.10)$$

where the vector-valued  $\eta : \mathbb{R}^k \rightarrow \mathbb{R}^k$  is an invertible function, and  $T : \mathbb{R} \rightarrow \mathbb{R}^k$  is a vector of sufficient statistics for  $\boldsymbol{\theta}$ . If  $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$ , then we can express the distribution as

$$f(x; \boldsymbol{\theta}) = h(x) \exp \left( \boldsymbol{\theta}^\top T(x) - \psi(\boldsymbol{\theta}) \right) \quad (3.9.11)$$

This is known as the *canonical form*. By transforming the parameters, any exponential family can be converted into canonical form.

## Gaussian Exponential Family

We can show that the Gaussian distribution with parameters  $\boldsymbol{\theta} = (\mu, \sigma)$  is a member of the exponential family. The density can be written as

$$f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \quad (3.9.12)$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \right) \quad (3.9.13)$$

$$= \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)}_{1/\phi(\boldsymbol{\theta})} \exp\left(\begin{bmatrix} \mu & -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} & 1 \end{bmatrix} \begin{bmatrix} x \\ x^2 \end{bmatrix}\right) \quad (3.9.14)$$

with  $h(x) = 1$  and

$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad (3.9.15)$$

$$\eta(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ 1 \\ -\frac{1}{2\sigma^2} \end{bmatrix} \quad (3.9.16)$$

### 3.9.3 Multiple-Parameter Multivariate Exponential Families

A multivariate distribution for a random vector  $\mathbf{X} \in \mathbb{R}^d$  with multiple parameters (represented by the vector  $\boldsymbol{\theta} \in \mathbb{R}^k$ ) is a member of the exponential family if its distribution can be written as

$$f(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x}) \exp\left(\eta(\boldsymbol{\theta})^\top T(\mathbf{x}) - \psi(\boldsymbol{\theta})\right) \quad (3.9.17)$$

where now  $T : \mathbb{R}^d \rightarrow \mathbb{R}^k$  is a vector of sufficient statistics for  $\boldsymbol{\theta}$ , and  $\eta : \mathbb{R}^k \rightarrow \mathbb{R}^k$  as before.

## Chapter 4

# Intermediate Statistics

## 4.1 Multivariate Statistics

### 4.1.1 Multivariate Sample Mean

For a multivariate sample  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , the sample mean is computed by

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (4.1.1)$$

which is an unbiased estimator for the population mean vector, because each element is simply the univariate sample mean.

### 4.1.2 Multivariate Medians

The notion of median and sample median can be generalised to multivariate domains in several ways.

#### Marginal Median

Suppose that the random vector  $\mathbf{X} \in \mathbb{R}^d$  has marginal cumulative distribution functions  $F_1(x_1), \dots, F_d(x_d)$ . Then the marginal median  $\mathbf{m} = [m_1 \ \dots \ m_d]^\top$  can be defined as the value(s) for which

$$F_1(m_1) = \dots = F_d(m_d) = \frac{1}{2} \quad (4.1.2)$$

Likewise, if we have a random sample of vectors  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , then the sample marginal median can be defined as the the vector which has as each of its components the univariate median of the sample in the corresponding dimension.

#### Medoid

For a random sample of vectors  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , the medoid  $\mathbf{c}^*$  on a distance function  $d(\mathbf{a}, \mathbf{b})$  (e.g. the Euclidean norm between two vectors) is defined as the vector from the sample that is closest on average to all others:

$$\mathbf{c}^* = \underset{\mathbf{c} \in \{\mathbf{X}_1, \dots, \mathbf{X}_n\}}{\operatorname{argmin}} \sum_{i=1}^d d(\mathbf{X}_i, \mathbf{c}) \quad (4.1.3)$$

## Spatial Median

For a random sample of vectors  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , the spatial median  $\mathbf{a}^*$  on a norm  $\|\cdot\|$  (e.g. the Euclidean norm or Manhattan norm) is defined as the vector which minimises the average distance to each point:

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \|\mathbf{X}_i - \mathbf{a}\| \quad (4.1.4)$$

### 4.1.3 Sample Variance as Quadratic Forms

The computation of sample variance of a sample  $x_1, \dots, x_n$  can be written as a quadratic form. Introduce the matrix

$$M = I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \quad (4.1.5)$$

where  $\mathbf{1}$  is a vector of  $n$  ones. Hence  $\mathbf{1} \mathbf{1}^\top$  is an  $n \times n$  matrix of all ones and  $M$  has the form

$$M = \begin{bmatrix} 1 - 1/n & -1/n & \dots & -1/n \\ -1/n & 1 - 1/n & \dots & -1/n \\ \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & \dots & 1 - 1/n \end{bmatrix} \quad (4.1.6)$$

Note that  $M$  is symmetric, can we can also verify that it is idempotent (i.e. the multiplication with itself equals itself) by

$$MM = \left( I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \right) \left( I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \right) \quad (4.1.7)$$

$$= I - \frac{2}{n} \mathbf{1} \mathbf{1}^\top + \frac{1}{n^2} n \mathbf{1} \mathbf{1}^\top \quad (4.1.8)$$

$$= I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \quad (4.1.9)$$

$$= M \quad (4.1.10)$$

where we have used the fact that  $(\mathbf{1} \mathbf{1}^\top)^2$  is a matrix containing all  $n$ . Denote

$$\mathbf{x} = [x_1 \ \dots \ x_n]^\top \quad (4.1.11)$$

Then we can see that

$$\begin{bmatrix} x_1 - \bar{x}_n \\ \vdots \\ x_n - \bar{x}_n \end{bmatrix} = \mathbf{x} - \frac{1}{n} \sum_{i=1}^n x_i \mathbf{1} \quad (4.1.12)$$

$$= \mathbf{x} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x} \quad (4.1.13)$$

$$= M \mathbf{x} \quad (4.1.14)$$

since  $\sum_{i=1}^n x_i = \mathbf{1}^\top \mathbf{x}$ . Then by applying the properties of symmetry and idempotence for  $M$ , the formula for sample variance can be rewritten as the quadratic form:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \quad (4.1.15)$$

$$= \frac{1}{n-1} (M \mathbf{x})^\top M \mathbf{x} \quad (4.1.16)$$

$$= \frac{1}{n-1} \mathbf{x}^\top M^\top M \mathbf{x} \quad (4.1.17)$$

$$= \frac{1}{n-1} \mathbf{x}^\top M M \mathbf{x} \quad (4.1.18)$$

$$= \frac{1}{n-1} \mathbf{x}^\top M \mathbf{x} \quad (4.1.19)$$

$M$  is sometimes referred to the centering matrix, since  $M\mathbf{x}$  has the same effect as subtracting the mean of the components of  $\mathbf{x}$  from each component.

#### 4.1.4 Sample Covariance Matrix

For a multivariate sample  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , the canonical sample covariance  $C$  is computed by

$$C = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \quad (4.1.20)$$

where  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$  is the sample mean. We see that this estimator also contains Bessel's correction (factor of  $n-1$  in the denominator), so each element of the sample covariance matrix is an unbiased estimator of the corresponding population covariance. So it follows that this estimator for the sample covariance is overall unbiased.

We can derive an identity for the sample covariance analogous to the bivariate sample covariance and the population covariance. This is done by expanding the formula above (leaving out the  $\frac{1}{n-1}$  factor for simplicity):

$$\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^n \mathbf{x}_i \bar{\mathbf{x}}^\top - \sum_{i=1}^n \bar{\mathbf{x}} \mathbf{x}_i^\top + \sum_{i=1}^n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \quad (4.1.21)$$

$$= \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top - n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top + n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \quad (4.1.22)$$

$$= \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \quad (4.1.23)$$

Hence

$$C = \frac{1}{n-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - \frac{n}{n-1} \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \quad (4.1.24)$$

#### Sample Correlation Matrix

The sample correlation matrix is the matrix of sample correlations between pairs of variables. Thus the sample correlation matrix will have all diagonal elements equal to 1 and all off-diagonal elements will be between  $-1$  and  $1$ . The sample correlation matrix can be computed in the following way. Let  $D$  be a diagonal matrix of sample standard deviations for each component of  $\mathbf{x} \in \mathbb{R}^d$ :

$$D = \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_d \end{bmatrix} \quad (4.1.25)$$

Hence  $D^{-1} = \text{diag}\{s_1^{-1}, \dots, s_d^{-1}\}$  and the transformed data  $D^{-1}\mathbf{x}$  will have unit sample variance for each component. Therefore the sample correlation matrix  $P$  can be represented by

$$P = \frac{1}{n-1} \sum_{i=1}^n [D^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})] [D^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})]^\top \quad (4.1.26)$$

$$= D^{-1} C D^{-1} \quad (4.1.27)$$

### Scatter Matrix

For a  $n$  samples  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , the scatter matrix  $S$  is defined as

$$S = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \quad (4.1.28)$$

where  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$  is the sample mean. Note that this is just the sample covariance without the  $\frac{1}{n-1}$  factor. The scatter matrix generalises the sum of squared deviations from the mean in univariate data.

### Pooled Covariance Matrix

The pooled covariance matrix extends the pooled variance to estimating the covariance of multiple covariances, under the assumption that the covariances of all the populations are identical. The formula takes the same form as the pooled variance, and in the same way is an unbiased estimate, i.e. for  $k$  samples:

$$\hat{\Sigma} = \frac{\sum_{i=1}^k (n_i - 1) \hat{\Sigma}_i}{\sum_{i=1}^k (n_i - 1)} \quad (4.1.29)$$

### Pairwise Characterisation of Sample Covariance

An alternative way to calculate the sample covariance matrix is by a summation involving all pairs of points:

$$C = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \quad (4.1.30)$$

and because each summand is replicated throughout the double sum (when  $i$  and  $j$  ‘switch around’), we can consider only half the terms (and the summand is zero when  $i = j$ ), so this is also equal to

$$C = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \quad (4.1.31)$$

*Proof.* We expand the expression  $\sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$  and then simplify:

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top &= \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i \mathbf{x}_j^\top - \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_i^\top + \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^\top \\ &\quad (4.1.32) \end{aligned}$$

$$\begin{aligned} &= n \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^n \mathbf{x}_i \sum_{j=1}^n \mathbf{x}_j^\top - \sum_{i=1}^n \mathbf{x}_j \sum_{j=1}^n \mathbf{x}_i^\top + n \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^\top \\ &\quad (4.1.33) \end{aligned}$$

$$\begin{aligned} &= 2n \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - \left( \sum_{i=1}^n \mathbf{x}_i \right) \left( \sum_{j=1}^n \mathbf{x}_j^\top \right) - \left( \sum_{i=1}^n \mathbf{x}_j \right) \left( \sum_{j=1}^n \mathbf{x}_i^\top \right) \\ &\quad (4.1.34) \end{aligned}$$

$$= 2n \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - 2n^2 \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \quad (4.1.35)$$

$$= 2n \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \right) \quad (4.1.36)$$

Hence the sample covariance matrix is

$$C = \frac{1}{n-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top - n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \right) \quad (4.1.37)$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \quad (4.1.38)$$

□

In the bivariate special case, we can analogously show

$$\sum_{i=1}^n \sum_{j=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n \sum_{j=1}^n x_i y_i - \sum_{i=1}^n \sum_{j=1}^n x_i y_j - \sum_{i=1}^n \sum_{j=1}^n x_j y_i + \sum_{i=1}^n \sum_{j=1}^n x_j y_j \quad (4.1.39)$$

$$= 2n \sum_{i=1}^n x_i y_i - 2n^2 \bar{x} \bar{y} \quad (4.1.40)$$

$$= 2n \left( \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right) \quad (4.1.41)$$

Thus an alternative formula for the sample covariance is

$$s_{xy} = \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right) \quad (4.1.42)$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (4.1.43)$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (4.1.44)$$

Notice that this is a sample-version of the characterisation of the population covariance using independent copies.

### Pairwise Characterisation of Sample Covariance

Analogously to their population versions, we can define a generalised scalar sample variance of multivariate data by the trace of the sample covariance matrix, which is equivalent to a sum of inner products representation:

$$\text{trace} \left( \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \right) = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (4.1.45)$$

$$= \frac{1}{n-1} \sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \quad (4.1.46)$$

We can derive the following identity which relates this quantity to the sum of pairwise distances  $\|\mathbf{x}_i - \mathbf{x}_j\|$  over the dataset:

$$\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \frac{1}{2n} \sum_{i,j \in \{1, \dots, n\}} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (4.1.47)$$

*Proof.* We expand:

$$\sum_{i,j \in \{1, \dots, n\}} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j) \quad (4.1.48)$$

$$= \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i^\top \mathbf{x}_i - 2 \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_j^\top \mathbf{x}_j \quad (4.1.49)$$

$$= n \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i - 2 \sum_{i=1}^n \mathbf{x}_i^\top \sum_{j=1}^n \mathbf{x}_j + n \sum_{j=1}^n \mathbf{x}_j^\top \mathbf{x}_j \quad (4.1.50)$$

$$= 2n \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i - 2 \sum_{i=1}^n \mathbf{x}_i^\top (n \bar{\mathbf{x}}) \quad (4.1.51)$$

$$= 2n \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i - 2n^2 \bar{\mathbf{x}}^\top \bar{\mathbf{x}} \quad (4.1.52)$$

$$= 2n \left( \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i - n \bar{\mathbf{x}}^\top \bar{\mathbf{x}} \right) \quad (4.1.53)$$

Hence  $\sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j) = 2n \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}})$ .  $\square$

This result is a sample version of the characterisation of the variance using independent copies. An alternative identity involves all distinct pairs and the sample mean:

$$\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \sum_{i \neq j} (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}} - \mathbf{x}_j) \quad (4.1.54)$$

*Proof.* We write the summation over all the pairs and subtract terms involving the pairs of repeated elements:

$$\sum_{i \neq j} (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}} - \mathbf{x}_j) = \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}} - \mathbf{x}_j) - \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}} - \mathbf{x}_i) \quad (4.1.55)$$

$$= \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}} - \mathbf{x}_j) + \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (4.1.56)$$

$$= \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}} - \mathbf{x}_j) + \sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \quad (4.1.57)$$

Thus, we just need to expand the first term and show it is equal to zero.

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}} - \mathbf{x}_j) &= \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i^\top \bar{\mathbf{x}} - \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^n \sum_{j=1}^n \bar{\mathbf{x}}^\top \bar{\mathbf{x}} - \sum_{i=1}^n \sum_{j=1}^n \bar{\mathbf{x}}^\top \mathbf{x}_j \\ &\quad (4.1.58) \end{aligned}$$

$$= n^2 \bar{\mathbf{x}}^\top \bar{\mathbf{x}} - n^2 \bar{\mathbf{x}}^\top \bar{\mathbf{x}} + n^2 \bar{\mathbf{x}}^\top \bar{\mathbf{x}} - n^2 \bar{\mathbf{x}}^\top \bar{\mathbf{x}} \quad (4.1.59)$$

$$= 0 \quad (4.1.60)$$

$\square$

This means we have several alternative expressions for the sample variance, which we can also directly compare to the **U-statistic for the variance**. In the scalar special case, they are

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4.1.61)$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2 \quad (4.1.62)$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n (x_i - x_j)^2 \quad (4.1.63)$$

$$= \frac{1}{n-1} \sum_{i \neq j}^n (x_i - \bar{x})(\bar{x} - x_j) \quad (4.1.64)$$

#### 4.1.5 Partial Correlation

Partial correlation is a closely related concept to the conditional correlation. Let  $\mathbf{V} = (V_1, V_2, \dots, V_n)$  be a random vector with some joint distribution. For two random variables of interest  $Y_i \equiv V_i$  and  $Y_j \equiv V_j$ , their partial correlation may be thought of as the correlation between  $Y_i$  and  $Y_j$  while controlling for all the other random variables, denoted  $\mathbf{X} = \mathbf{V} \setminus \{Y_i, Y_j\}$ . Suppose  $\mathbf{V}$  has covariance  $\text{Cov}(\mathbf{V}) = \Sigma$  and precision matrix:

$$\Sigma^{-1} = P \quad (4.1.65)$$

$$= \begin{bmatrix} p_{11} & \dots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nn} \end{bmatrix} \quad (4.1.66)$$

Then the partial correlation between  $Y_i$  and  $Y_j$  can be related to the elements of  $P$  via the following formula:

$$\rho_{Y_i, Y_j | \mathbf{X}} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}} \quad (4.1.67)$$

To derive this relation, we first introduce the notion of a ‘best linear approximation’ of a random variable. Suppose there is a random vector which can be partitioned into two random vectors:  $(\mathbf{Y}, \mathbf{X})$ . By best linear approximation of  $\mathbf{Y}$  given  $\mathbf{X}$ , we mean that suppose there exists a random variable  $\hat{\mathbf{Y}} \approx \mathbf{Y}$  which can be written as

$$\hat{\mathbf{Y}} = \boldsymbol{\mu}_{\mathbf{Y}} + A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \mathbf{Z} \quad (4.1.68)$$

where  $\boldsymbol{\mu}_{\mathbf{Y}} := \mathbb{E}[\mathbf{Y}]$ ,  $\boldsymbol{\mu}_{\mathbf{X}} := \mathbb{E}[\mathbf{X}]$  while  $\mathbf{Z}$  is another random vector and  $A$  is a matrix of appropriate dimensions. Additionally, the following conditions are imposed:

$$\mathbb{E}[\mathbf{Z}] = \mathbf{0} \quad (4.1.69)$$

$$\mathbb{E}[\mathbf{Z}|\mathbf{X}] = \mathbb{E}[\mathbf{Z}] \quad (4.1.70)$$

which implies that  $\mathbb{E}[\mathbf{Z}|\mathbf{X}] = \mathbf{0}$ , and recall that the second condition is something in between independence and uncorrelatedness. Intuitively, this condition means that  $\mathbf{Z}$  is not very useful for predicting  $\mathbf{Y}$ . Also, for ‘closeness’ between  $\hat{\mathbf{Y}}$  and  $\mathbf{Y}$ , we should also require:

$$\mathbb{E}[\hat{\mathbf{Y}}] = \mathbb{E}[\mathbf{Y}] \quad (4.1.71)$$

$$\text{Cov}(\hat{\mathbf{Y}}) = \text{Cov}(\mathbf{Y}) \quad (4.1.72)$$

$$\text{Cov}(\hat{\mathbf{Y}}, \mathbf{X}) = \text{Cov}(\mathbf{Y}, \mathbf{X}) \quad (4.1.73)$$

For simplicity of notation, define

$$\text{Cov}\left(\begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix}\right) = \begin{bmatrix} \text{Cov}(\mathbf{Y}) & \text{Cov}(\mathbf{Y}, \mathbf{X}) \\ \text{Cov}(\mathbf{X}, \mathbf{Y}) & \text{Cov}(\mathbf{X}) \end{bmatrix} \quad (4.1.74)$$

$$:= \begin{bmatrix} \Sigma_{\mathbf{YY}} & \Sigma_{\mathbf{YX}} \\ \Sigma_{\mathbf{XY}} & \Sigma_{\mathbf{XX}} \end{bmatrix} \quad (4.1.75)$$

Under these conditions, we can go about finding the matrix  $A$ . Denote the mean-centered  $\widehat{\mathbf{Y}} := \widehat{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}}$  and  $\widetilde{\mathbf{X}} := \mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}$  so that

$$\widehat{\mathbf{Y}} = A\widetilde{\mathbf{X}} + \mathbf{Z} \quad (4.1.76)$$

Taking expectations of both sides conditional on  $\mathbf{X}$ :

$$\mathbb{E} \left[ \widehat{\mathbf{Y}} \mid \mathbf{X} \right] = \mathbb{E} \left[ A\widetilde{\mathbf{X}} + \mathbf{Z} \mid \mathbf{X} \right] \quad (4.1.77)$$

$$= A\widetilde{\mathbf{X}} \quad (4.1.78)$$

since  $\mathbf{X}$  contains as much ‘information’ as  $\widetilde{\mathbf{X}}$  and also using  $\mathbb{E}[\mathbf{Z}|\mathbf{X}] = \mathbf{0}$ . Post-multiply both sides by  $\widetilde{\mathbf{X}}^\top$  giving:

$$\mathbb{E} \left[ \widehat{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \mid \mathbf{X} \right] = A\widetilde{\mathbf{X}} \widetilde{\mathbf{X}}^\top \quad (4.1.79)$$

Taking unconditional expectations and using the Law of Iterated Expectations:

$$\mathbb{E} \left[ \mathbb{E} \left[ \widehat{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \mid \mathbf{X} \right] \right] = A\mathbb{E} \left[ \widetilde{\mathbf{X}} \widetilde{\mathbf{X}}^\top \right] \quad (4.1.80)$$

$$\mathbb{E} \left[ \widehat{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \right] = A\mathbb{E} \left[ \widetilde{\mathbf{X}} \widetilde{\mathbf{X}}^\top \right] \quad (4.1.81)$$

So  $A$  can be found as

$$A = \mathbb{E} \left[ \widehat{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \right] \mathbb{E} \left[ \widetilde{\mathbf{X}} \widetilde{\mathbf{X}}^\top \right]^{-1} \quad (4.1.82)$$

$$= \mathbb{E} \left[ (\widehat{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}}) (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top \right] \mathbb{E} \left[ (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top \right]^{-1} \quad (4.1.83)$$

$$= \text{Cov}(\widehat{\mathbf{Y}}, \mathbf{X}) \text{Cov}(\mathbf{X})^{-1} \quad (4.1.84)$$

$$= \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \quad (4.1.85)$$

Note that since the condition  $\mathbb{E}[\mathbf{Z}|\mathbf{X}] = \mathbb{E}[\mathbf{Z}]$  implies  $\mathbf{X}$  and  $\mathbf{Z}$  should be uncorrelated, this can be verified using the found  $A$ :

$$\text{Cov}(\mathbf{Z}, \mathbf{X}) = \text{Cov}(\widehat{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}} - A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}), \mathbf{X}) \quad (4.1.86)$$

$$= \text{Cov}(\widehat{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}}, \mathbf{X}) - \text{Cov}(A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}), \mathbf{X}) \quad (4.1.87)$$

$$= \text{Cov}(\widehat{\mathbf{Y}}, \mathbf{X}) - A \text{Cov}(\mathbf{X}) \quad (4.1.88)$$

$$= \Sigma_{\mathbf{YX}} - \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XX}} \quad (4.1.89)$$

$$= \mathbf{0} \quad (4.1.90)$$

We now focus of the conditional covariance of  $\widehat{\mathbf{Y}}$  given  $\mathbf{X}$ :

$$\text{Cov}(\widehat{\mathbf{Y}} \mid \mathbf{X}) = \text{Cov}(\boldsymbol{\mu}_{\mathbf{Y}} + A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \mathbf{Z} \mid \mathbf{X}) \quad (4.1.91)$$

$$= \text{Cov}(A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \mathbf{Z} \mid \mathbf{X}) \quad (4.1.92)$$

$$= \text{Cov}(A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) \mid \mathbf{X}) + \text{Cov}(A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}), \mathbf{Z} \mid \mathbf{X}) + \text{Cov}(\mathbf{Z}, A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) \mid \mathbf{X}) + \text{Cov}(\mathbf{Z} \mid \mathbf{X}) \quad (4.1.93)$$

$$= A \operatorname{Cov}(\mathbf{X}|\mathbf{X}) + A \operatorname{Cov}(\mathbf{X}, \mathbf{Z}|\mathbf{X}) + \operatorname{Cov}(\mathbf{Z}, \mathbf{X}|\mathbf{X}) A^\top + \operatorname{Cov}(\mathbf{Z}|\mathbf{X}) \quad (4.1.94)$$

Note that  $\operatorname{Cov}(\mathbf{X}|\mathbf{X})$  and we can show  $\operatorname{Cov}(\mathbf{Z}, \mathbf{X}|\mathbf{X}) = \operatorname{Cov}(\mathbf{X}, \mathbf{Z}|\mathbf{X})^\top = \mathbf{0}$ :

$$\operatorname{Cov}(\mathbf{Z}, \mathbf{X}|\mathbf{X}) = \mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top | \mathbf{X}] \quad (4.1.95)$$

$$= \mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])|\mathbf{X}] (\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \quad (4.1.96)$$

$$= \mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])] (\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \quad (4.1.97)$$

$$= \mathbf{0} \quad (4.1.98)$$

due to  $\mathbb{E}[\mathbf{Z}|\mathbf{X}] = \mathbb{E}[\mathbf{Z}]$ . Hence

$$\operatorname{Cov}(\hat{\mathbf{Y}}|\mathbf{X}) = \operatorname{Cov}(\mathbf{Z}|\mathbf{X}) \quad (4.1.99)$$

and also because of  $\mathbb{E}[\mathbf{Z}|\mathbf{X}] = \mathbb{E}[\mathbf{Z}]$ , this is equal to the unconditional covariance:

$$\operatorname{Cov}(\hat{\mathbf{Y}}|\mathbf{X}) = \operatorname{Cov}(\mathbf{Z}) \quad (4.1.100)$$

$$= \operatorname{Cov}(\hat{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}} - A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})) \quad (4.1.101)$$

$$= \operatorname{Cov}(\hat{\mathbf{Y}} - A\mathbf{X}) \quad (4.1.102)$$

$$= \operatorname{Cov}(\hat{\mathbf{Y}}) - \operatorname{Cov}(\hat{\mathbf{Y}}, \mathbf{X}) A^\top - A \operatorname{Cov}(\mathbf{X}, \hat{\mathbf{Y}}) + A \operatorname{Cov}(\mathbf{X}) A^\top \quad (4.1.103)$$

$$= \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} - \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} + \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XX}} \Sigma_{\mathbf{XY}}^{-1} \Sigma_{\mathbf{XY}} \quad (4.1.104)$$

$$= \Sigma_{\mathbf{YY}} - 2\Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} + \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \quad (4.1.105)$$

$$= \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \quad (4.1.106)$$

We can now formally characterise the partial correlation as follows. The partial correlation between  $Y_i$  and  $Y_j$  for  $i \neq j$  given  $\mathbf{X} = \mathbf{V} \setminus \{Y_i, Y_j\}$  is the conditional correlation between  $\hat{Y}_i$  and  $\hat{Y}_j$  given  $\mathbf{X}$ , where  $\hat{\mathbf{Y}} = (\hat{Y}_i, \hat{Y}_j)$  is deemed as the best linear approximation of  $\mathbf{Y} = (Y_i, Y_j)$  given  $\mathbf{X}$  [183]. Without loss of generality, assume  $i = 1$  and  $j = 2$  so that  $\operatorname{Cov}(\mathbf{V}) = \Sigma$  which can be partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{YY}} & \Sigma_{\mathbf{YX}} \\ \Sigma_{\mathbf{XY}} & \Sigma_{\mathbf{XX}} \end{bmatrix} \quad (4.1.107)$$

Then the precision matrix  $P = \Sigma^{-1}$  can be partitioned as

$$P = \begin{bmatrix} P_{\mathbf{YY}} & P_{\mathbf{YX}} \\ P_{\mathbf{XY}} & P_{\mathbf{XX}} \end{bmatrix} \quad (4.1.108)$$

with  $P_{\mathbf{YY}}$  of the same dimension  $2 \times 2$  as  $\Sigma_{\mathbf{YY}}$ . Block matrix inversion formulae gives the relationship

$$P_{\mathbf{YY}}^{-1} = \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \quad (4.1.109)$$

where we recognise that  $\Sigma_{\mathbf{Y}|\mathbf{X}} := \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}}$  is the conditional covariance of  $\hat{\mathbf{Y}}$  given  $\mathbf{X}$ , which means that the precision of  $\hat{\mathbf{Y}}$  given  $\mathbf{X}$  is equal to the block  $P_{\mathbf{YY}}$  of the unconditional precision of  $(\hat{\mathbf{Y}}, \mathbf{X})$ . By our definition of the partial correlation, we write

$$\rho_{Y_i, Y_j | \mathbf{X}} = \frac{[\Sigma_{\mathbf{Y}|\mathbf{X}}]_{12}}{\sqrt{[\Sigma_{\mathbf{Y}|\mathbf{X}}]_{11} [\Sigma_{\mathbf{Y}|\mathbf{X}}]_{22}}} \quad (4.1.110)$$

Since  $\Sigma_{\mathbf{Y}|\mathbf{X}}$  is  $2 \times 2$ , the elements of  $P_{\mathbf{YY}}^{-1} = \Sigma_{\mathbf{Y}|\mathbf{X}}$  can be easily expressed in terms of the elements of  $P_{\mathbf{YY}}$ :

$$P_{\mathbf{YY}} = \begin{bmatrix} p_{ii} & p_{ij} \\ p_{ji} & p_{jj} \end{bmatrix} \quad (4.1.111)$$

by

$$P_{\mathbf{YY}}^{-1} = \frac{1}{\det(P_{\mathbf{YY}})} \begin{bmatrix} p_{jj} & -p_{ij} \\ -p_{ji} & p_{ii} \end{bmatrix} \quad (4.1.112)$$

So equating:

$$[\Sigma_{\mathbf{Y}|\mathbf{X}}]_{11} = \frac{p_{jj}}{\det(P_{\mathbf{YY}})} \quad (4.1.113)$$

$$[\Sigma_{\mathbf{Y}|\mathbf{X}}]_{22} = \frac{p_{ii}}{\det(P_{\mathbf{YY}})} \quad (4.1.114)$$

$$[\Sigma_{\mathbf{Y}|\mathbf{X}}]_{12} = -\frac{p_{ij}}{\det(P_{\mathbf{YY}})} \quad (4.1.115)$$

So the partial correlation can be alternatively expressed as

$$\rho_{Y_i, Y_j | \mathbf{X}} = \frac{-\frac{p_{ij}}{\det(P_{\mathbf{YY}})}}{\sqrt{\frac{p_{jj}}{\det(P_{\mathbf{YY}})} \frac{p_{ii}}{\det(P_{\mathbf{YY}})}}} \quad (4.1.116)$$

$$= -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}} \quad (4.1.117)$$

as originally asserted.

### Sample Partial Correlation Coefficient

The definition of the partial correlation coefficient as defined above motivates the following approach to compute the sample partial coefficient from multivariate data. For sample partial correlation between the  $i^{\text{th}}$  and  $j^{\text{th}}$  variables, we first obtain all the residuals  $\mathbf{e}_i$  from regressing the  $i^{\text{th}}$  variables on all other variables except the  $j^{\text{th}}$  (including an intercept) using ordinary least squares. Then we do the same on the  $j^{\text{th}}$  variable to obtain the residuals  $\mathbf{e}_j$ . The sample covariance between  $\mathbf{e}_i$  and  $\mathbf{e}_j$  is intended to estimate the covariance  $\text{Cov}(\mathbf{Z}) = \text{Cov}(\widehat{\mathbf{Y}} | \mathbf{X})$ , so then the sample partial correlation coefficient is taken to be the sample correlation between  $\mathbf{e}_i$  and  $\mathbf{e}_j$ .

Alternatively, the sample partial correlation can be computed using the elements of the full sample covariance matrix in the same way as above.

### Diagonal Elements of Inverted Correlation Matrix

Let  $\mathbf{V} = (V_1, V_2, \dots, V_n)$  be a random vector with some joint distribution. Consider the best linear approximation  $\widehat{Y}_i = \mu_{Y_i} + A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + Z$  for one random variable of interest  $Y_i \equiv V_i$  given  $\mathbf{X} = \mathbf{V} \setminus \{Y_i\}$ . Suppose  $\mathbf{V}$  has correlation matrix  $\text{Corr}(\mathbf{V}) = \Omega$  and inverted correlation matrix  $\Omega^{-1} = \Psi$ . Then the correlation between  $\widehat{Y}_i$  and  $\widehat{Y}_i - Z$  can be written in terms of the  $i^{\text{th}}$  diagonal element of  $\Psi$  by

$$\text{Corr}(\widehat{Y}_i, \widehat{Y}_i - Z) = \sqrt{1 - \frac{1}{\psi_{ii}}} \quad (4.1.118)$$

*Proof.* Without loss of generality, assume  $i = 1$ . Also without loss of generality, assume that all of  $V_1, V_2, \dots, V_n$  have zero mean and variances of 1. This is because we can always scale and center  $\mathbf{V}$  via an invertible linear transform so that this is true, and by virtue of  $\widehat{Y}_i$  being a linear approximation, then this linear transform will subsequently carry through and have no

effect on the correlation. This means to say that  $\text{Corr} \left( a\widehat{Y}_i + b, c(\widehat{Y}_i - Z) + d \right)$  for non-zero  $a, c$ . As a result, the correlation matrix is the same as the covariance matrix and as determined above, the element  $\psi_{ii} = \psi_{11}$  of the inverted correlation matrix will be given via the relation

$$\psi_{ii}^{-1} = \Omega_Y - \Omega_{YX}\Omega_{XX}^{-1}\Omega_{XY} \quad (4.1.119)$$

where the correlation matrix is partitioned as

$$\Omega = \begin{bmatrix} \Omega_Y & \Omega_{YX} \\ \Omega_{XY} & \Omega_{XX} \end{bmatrix} \quad (4.1.120)$$

Note that  $\Omega_Y = 1$  since it is a correlation matrix, and since  $\psi_{ii}$  is a scalar, then inverting yields

$$\frac{1}{\psi_{ii}} = \frac{1}{1 - \Omega_{YX}\Omega_{XX}^{-1}\Omega_{XY}} \quad (4.1.121)$$

$$1 - \frac{1}{\psi_{ii}} = \Omega_{YX}\Omega_{XX}^{-1}\Omega_{XY} \quad (4.1.122)$$

By the zero-mean assumption, note that  $\widehat{Y}_i - Z = A\mathbf{X}$  and that

$$\Omega_{YX}\Omega_{XX}^{-1}\Omega_{XY} = \Omega_{YX}\Omega_{XX}^{-1}\Omega_{XX}\Omega_{XX}^{-1}\Omega_{XY} \quad (4.1.123)$$

$$= A\Omega_{XX}A^\top \quad (4.1.124)$$

$$= A \text{Cov}(\mathbf{X}) A^\top \quad (4.1.125)$$

$$= \text{Var}(A\mathbf{X}) \quad (4.1.126)$$

$$= \text{Var}(\widehat{Y}_i - Z) \quad (4.1.127)$$

Also note that

$$\text{Cov}(\widehat{Y}_i, \widehat{Y}_i - Z) = \text{Cov}(A\mathbf{X} + Z, A\mathbf{X}) \quad (4.1.128)$$

$$= \text{Cov}(A\mathbf{X}, A\mathbf{X}) + \text{Cov}(Z, A\mathbf{X}) \quad (4.1.129)$$

$$= \text{Cov}(A\mathbf{X}, A\mathbf{X}) \quad (4.1.130)$$

$$= \text{Var}(A\mathbf{X}) \quad (4.1.131)$$

since  $\mathbf{X}$  and  $Z$  are uncorrelated because of the property of the best linear approximation. Hence

$$\text{Cov}(\widehat{Y}_i, \widehat{Y}_i - Z) = \text{Var}(\widehat{Y}_i - Z) \quad (4.1.132)$$

$$= 1 - \frac{1}{\psi_{ii}} \quad (4.1.133)$$

By the definition of the correlation and the fact that  $\text{Var}(\widehat{Y}_i) = \text{Var}(Y_1) = 1$ ,

$$\text{Corr}(\widehat{Y}_i, \widehat{Y}_i - Z) = \frac{\text{Cov}(\widehat{Y}_i, \widehat{Y}_i - Z)}{\sqrt{\text{Var}(\widehat{Y}_i) \text{Var}(\widehat{Y}_i - Z)}} \quad (4.1.134)$$

$$= \frac{1 - 1/\psi_{ii}}{\sqrt{1 - 1/\psi_{ii}}} \quad (4.1.135)$$

$$= \sqrt{1 - \frac{1}{\psi_{ii}}} \quad (4.1.136)$$

or

$$\text{Corr}(\widehat{Y}_i, \widehat{Y}_i - Z)^2 = 1 - \frac{1}{\psi_{ii}} \quad (4.1.137)$$

□

The analogy of this to when we have multivariate data is that if we compute the diagonals of the inverse of the sample correlation matrix, then the  $i^{\text{th}}$  diagonal  $\widehat{\psi}_{ii}$  should be related to the R-squared value  $R^2$  obtained by regressing the  $i^{\text{th}}$  variable on all other variables (including an intercept) by

$$R^2 = 1 - \frac{1}{\widehat{\psi}_{ii}} \quad (4.1.138)$$

#### 4.1.6 Mahalanobis Distance

The Mahalanobis distance generalises the notion of standard deviations from the mean. For a random vector  $\mathbf{X}$  with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\mathbf{C}$ , the Mahalanobis distance  $d_M$  is defined as

$$d_M = \sqrt{(\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{C}^{-1} (\mathbf{X} - \boldsymbol{\mu})} \quad (4.1.139)$$

#### 4.1.7 Higher Co-Moments [139]

##### Coskewness

Coskewness generalises skewness of a single random variable to skewness as a function of up to three random variables. For three random variables  $X, Y, Z$ , with population means  $\mu_X, \mu_Y, \mu_Z$  and population standard deviations  $\sigma_X, \sigma_Y, \sigma_Z$  respectively, the population coskewness is defined as

$$\gamma_{XYZ} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)(Z - \mu_Z)]}{\sigma_X \sigma_Y \sigma_Z} \quad (4.1.140)$$

We develop some intuition for interpreting coskewness between two random variables and their joint distribution. For random variables  $X$  and  $Y$ , there are two relevant coskewness parameters to consider:

$$\gamma_{XXY} = \frac{\mathbb{E}[(X - \mu_X)^2(Y - \mu_Y)]}{\sigma_X^2 \sigma_Y} \quad (4.1.141)$$

$$\gamma_{XYY} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)^2]}{\sigma_X \sigma_Y^2} \quad (4.1.142)$$

noting that  $\gamma_{XXX}$  and  $\gamma_{YYY}$  are just the skewness of  $X$  and  $Y$  respectively, so only provide information about the marginal distributions. Using the Law of Iterated Expectations, we can write  $\gamma_{XXY}$  as

$$\gamma_{XXY} \propto \mathbb{E}\left[(Y - \mu_Y) \mathbb{E}\left[(X - \mu_X)^2 \mid Y\right]\right] \quad (4.1.143)$$

This can be viewed as taking an expectation over  $Y$ , weighted by the squared deviation of  $X$  from its mean conditional on  $Y$ . So if the coskewness  $\gamma_{XXY}$  is positive, this indicates that  $X$  is taking on more extreme values when  $Y$  is above the mean. Similarly,  $\gamma_{XYY}$  can be written as

$$\gamma_{XYY} \propto \mathbb{E}\left[(X - \mu_X) \mathbb{E}\left[(Y - \mu_Y)^2 \mid X\right]\right] \quad (4.1.144)$$

so a positive coskewness  $\gamma_{XYY} > 0$  indicates that  $Y$  is taking on more extreme values when  $X$  is above the mean. So if both coskewness parameters  $\gamma_{XXY}, \gamma_{XYY} > 0$ , we would anticipate that extreme positive events on  $X$  and  $Y$  occur together. By similar reasoning, if both coskewness parameters  $\gamma_{XXY}, \gamma_{XYY} < 0$ , we would anticipate that extreme negative events on  $X$  and  $Y$  occur together.

## Cokurtosis

Cokurtosis generalises kurtosis of a single random variable to kurtosis between up to four different random variables. For random variables  $W, X, Y, Z$  with population means  $\mu_W, \mu_X, \mu_Y, \mu_Z$  and population standard deviations  $\sigma_W, \sigma_X, \sigma_Y, \sigma_Z$ , the population cokurtosis is defined as

$$\kappa_{WXYZ} = \frac{\mathbb{E}[(W - \mu_W)(X - \mu_X)(Y - \mu_Y)(Z - \mu_Z)]}{\sigma_W \sigma_X \sigma_Y \sigma_Z} \quad (4.1.145)$$

The cokurtosis parameter between two random variables  $X$  and  $Y$  may be interpreted to describe the shape of their joint distribution. We consider the three cokurtosis parameters:

$$\kappa_{XXXY} = \frac{\mathbb{E}[(X - \mu_X)^3(Y - \mu_Y)]}{\sigma_X^3 \sigma_Y} \quad (4.1.146)$$

$$\kappa_{XXYY} = \frac{\mathbb{E}[(X - \mu_X)^2(Y - \mu_Y)^2]}{\sigma_X^2 \sigma_Y^2} \quad (4.1.147)$$

$$\kappa_{XYYY} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)^3]}{\sigma_X \sigma_Y^3} \quad (4.1.148)$$

while noting that the parameters  $\kappa_{XXX}$  and  $\kappa_{YYY}$  are special cases while become the population kurtosis of  $X$  and  $Y$  respectively, so only provide information about their marginal distributions. Each of the parameters above can be analysed as follows:

- If the parameter  $\kappa_{XXXY}$  is large and positive, this indicates that extreme positive events of  $X$  typically occur when  $Y$  is above the mean. Likewise extreme negative events of  $X$  typically occur when  $Y$  is below the mean.
- The more frequently extreme events of  $X$  and  $Y$  occur, the higher  $\kappa_{XXYY}$  will be.
- If the parameter  $\kappa_{XYYY}$  is large and positive, this indicates that extreme positive events of  $Y$  typically occur when  $X$  is above the mean. Likewise extreme negative events of  $Y$  typically occur when  $X$  is below the mean.

Therefore if the parameters  $\kappa_{XXXY}$ ,  $\kappa_{XXYY}$  and  $\kappa_{XYYY}$  are all positive and large, we would anticipate that  $X$  and  $Y$  both undergo extreme positive and negative events at the same time.

### 4.1.8 Confidence Regions

## 4.2 Statistical Decision Theory

### 4.2.1 Optimal Prediction

Suppose we have to make a prediction about the realisation of a random variable  $Y$ , which has the probability density function  $f_Y(y)$ . One reasonable choice of the prediction  $\hat{Y}$  is the value which minimises the expected squared deviation of the realisation of the prediction, i.e.

$$\hat{Y} = \operatorname{argmin}_c \mathbb{E}[(Y - c)^2] \quad (4.2.1)$$

$$= \operatorname{argmin}_c \int_{-\infty}^{\infty} f_Y(y) (y - c)^2 dy \quad (4.2.2)$$

We can derive  $\hat{Y}$  by taking the derivative (supposing sufficient regularity to do so):

$$\frac{\partial}{\partial c} \int_{-\infty}^{\infty} f_Y(y) (y - c)^2 dy = \int_{-\infty}^{\infty} f_Y(y) \frac{\partial}{\partial c} (y - c)^2 dy \quad (4.2.3)$$

$$= - \int_{-\infty}^{\infty} f_Y(y) 2(y - c) dy \quad (4.2.4)$$

Setting the derivative to zero:

$$\int_{-\infty}^{\infty} f_Y(y) (y - \hat{Y}) dy = 0 \quad (4.2.5)$$

$$\int_{-\infty}^{\infty} f_Y(y) y dy = \hat{Y} \int_{-\infty}^{\infty} f_Y(y) dy \quad (4.2.6)$$

$$\hat{Y} = \int_{-\infty}^{\infty} f_Y(y) y dy \quad (4.2.7)$$

$$\hat{Y} = \mathbb{E}[Y] \quad (4.2.8)$$

Hence the expectation of  $Y$  (if it exists) is the optimal prediction in this sense. The same can be analogously shown if  $Y$  is a discrete random variable. In most cases arising in statistics, the true distribution of  $Y$  will not be known, nor even its population mean. If however we have a sample drawn from the distribution of  $Y$ , we can take the sample mean to give an unbiased estimate of  $\mathbb{E}[Y]$ . These optimal predictions can also be conditional on predictors  $X$ , in which case we take the conditional expectation. It can be analogously shown that the optimal prediction conditional on predictors  $X$  defined as

$$\hat{Y}|X = \operatorname{argmin}_c \mathbb{E}[(Y - c)^2 | X] \quad (4.2.9)$$

$$= \operatorname{argmin}_c \int_{-\infty}^{\infty} f_{Y|X}(y) (y - c)^2 dy \quad (4.2.10)$$

is equal to

$$\hat{Y}|X = \mathbb{E}[Y|X] \quad (4.2.11)$$

Hence this justifies estimating the conditional mean from a sample for predictive or modelling purposes. We choose the model regression function to ideally give a good estimate of

$$f(x) = \mathbb{E}[Y|X = x] \quad (4.2.12)$$

If we instead change the prediction criterion to minimising the expected absolute deviation:

$$\hat{Y} = \operatorname{argmin}_c \mathbb{E}[|Y - c|] \quad (4.2.13)$$

then we can derive an alternative prediction rule by first writing (supposing that  $Y$  is a continuous random variable):

$$\mathbb{E}[|Y - c|] = \int_{-\infty}^{\infty} f_Y(y) |y - c| dy \quad (4.2.14)$$

$$= \int_{-\infty}^c f_Y(y) (c - y) dy + \int_c^{\infty} f_Y(y) (y - c) dy \quad (4.2.15)$$

We seek to find the value for which the right derivative of the first term equals the negative of the left derivative of the second term (the whole term is not differentiable everywhere). We can further split this into

$$\int_{-\infty}^c f_Y(y) (c - y) dy = c \int_{-\infty}^c f_Y(y) dy - \int_{-\infty}^c f_Y(y) y dy \quad (4.2.16)$$

$$\int_c^{\infty} f_Y(y) (y - c) dy = \int_c^{\infty} f_Y(y) y dy - c \int_c^{\infty} f_Y(y) dy \quad (4.2.17)$$

Note that by the Fundamental Theorem of Calculus, we have

$$\frac{\partial}{\partial c} \int_c^\infty f_Y(y) dy = \frac{\partial}{\partial c} \left( \lim_{y \rightarrow \infty} F_Y(y) - F_Y(c) \right) \quad (4.2.18)$$

$$= \frac{\partial}{\partial c} (1 - F_Y(c)) \quad (4.2.19)$$

$$= -f_Y(c) \quad (4.2.20)$$

where  $F_Y(y)$  is the cumulative distribution function of  $Y$ . Similarly,  $\frac{\partial}{\partial c} \int_c^\infty f_Y(y) y dy = -f_Y(c)c$  and by the product rule,  $\frac{\partial}{\partial c} c \int_c^\infty f_Y(y) dy = \int_c^\infty f_Y(y) dy - cf_Y(c)$ . Using these to take the derivative the terms above we obtain

$$\frac{\partial}{\partial c} \int_{-\infty}^c f_Y(y) (c - y) dy = cf_Y(c) + \int_{-\infty}^c f_Y(y) dy - f_Y(c)c \quad (4.2.21)$$

$$= \int_{-\infty}^c f_Y(y) dy \quad (4.2.22)$$

and

$$\frac{\partial}{\partial c} \int_c^\infty f_Y(y) (y - c) dy = -f_Y(c)c - \int_c^\infty f_Y(y) dy + cf_Y(c) \quad (4.2.23)$$

$$= - \int_c^\infty f_Y(y) dy \quad (4.2.24)$$

Therefore we find the value of  $c$  for which

$$\int_{-\infty}^c f_Y(y) dy = \int_c^\infty f_Y(y) dy \quad (4.2.25)$$

This value happens to be the median  $m_Y$  of  $Y$ , since at the median,

$$\int_{-\infty}^{m_Y} f_Y(y) dy = \frac{1}{2} \quad (4.2.26)$$

$$\int_{m_Y}^\infty f_Y(y) dy = \frac{1}{2} \quad (4.2.27)$$

So the optimal prediction is the population median

$$\hat{Y} = m_Y \quad (4.2.28)$$

If we have a sample drawn from the population of  $Y$ , then we may estimate the population median using the sample median. Likewise, if given a predictor  $x$ , it can be shown via roughly the same steps that the ideal predictive function is the conditional median (i.e. median of the conditional distribution):

$$f(x) = \operatorname{argmin}_c \mathbb{E}[|Y - c| | X = x] \quad (4.2.29)$$

$$= m_{Y|X=x} \quad (4.2.30)$$

## 4.2.2 Binary Hypothesis Testing

In binary hypothesis testing, we have the null hypothesis  $H_0$  and alternative hypothesis  $H_1$  like in **null hypothesis statistical testing**. Unlike null hypothesis statistical testing however, we do not “reject” or “fail to reject” the null. Rather, we “decide” on accepting either  $H_0$  or  $H_1$  (which comes off as a slightly stronger interpretation than rejecting or not rejecting). Binary hypothesis testing is suitable when both  $H_0$  and  $H_1$  can be well-modelled probabilistically.

## Maximum Likelihood Hypothesis Testing [220]

Consider a binary hypothesis test with hypotheses  $H_0$  and  $H_1$ . After conducting an experiment to test the hypotheses, we observe a decision statistic  $X$ . Then the maximum likelihood decision rule based on a realisation of the decision statistic  $x$  is defined by:

- Decide  $H_0$  if  $p(x|H_0) \geq p(x|H_1)$ .
- Otherwise, decide  $H_1$ .

Here,  $p(x|H_j)$  is either the probability density or mass (depending on whether  $X$  is continuous or discrete) of  $x$  given hypothesis  $H_j$ . Intuitively speaking, we decide on the hypothesis which was the most likely.

## Maximum a Posteriori Hypothesis Testing [220]

Consider a binary hypothesis test with hypotheses  $H_0$  and  $H_1$ . After conducting an experiment to test the hypotheses, we observe a decision statistic  $X$ . Then the maximum a posteriori decision rule based on a realisation of the decision statistic  $x$  is defined by:

- Decide  $H_0$  if  $p(H_0|x) \geq p(H_1|x)$ .
- Otherwise, decide  $H_1$ .

Here,  $p(H_j|x)$  is the probability of observing hypothesis  $H_j$  given  $x$ . Using Bayes' rule, the decision rule can be written as

$$\frac{p(x|H_0)p(H_0)}{p(x)} \geq \frac{p(x|H_1)p(H_1)}{p(x)} \quad (4.2.31)$$

which is equivalent to

$$p(x|H_0)p(H_0) \geq p(x|H_1)p(H_1) \quad (4.2.32)$$

due to the common denominator. We can view this approach as comparing the *posterior* probabilities (likelihood multiplied by prior) of each hypothesis, where  $p(H_0)$  and  $p(H_1)$  are our prior probabilities of each hypothesis being true. Let  $A_0$  denote the event of deciding  $H_0$  and let  $A_1$  denote the event of deciding  $H_1$ . The Type I error probability (otherwise known as the probability of ‘false alarm’) is

$$P_{\text{FA}} = \Pr(A_1|H_0) \quad (4.2.33)$$

whereas the Type II error probability (also known as the probability of ‘miss’) is

$$P_{\text{miss}} = \Pr(A_0|H_1) \quad (4.2.34)$$

Hence the total probability of making a decision error is (by the Law of Total Probability):

$$P_{\text{error}} = \Pr(A_1|H_0)p(H_0) + \Pr(A_0|H_1)p(H_1) \quad (4.2.35)$$

We can illustrate that the maximum a posteriori decision rule minimises the probability of error. Let  $\mathcal{X}$  be the support of  $X$ . Assume that  $X$  is discrete, but analogous arguments apply if  $X$  were continuous. A decision rule essentially means partitioning  $\mathcal{X}$  into  $\mathcal{X}_0$  and  $\mathcal{X}_1$  such that

$$\Pr(A_0) = \Pr(x \in \mathcal{X}_0) \quad (4.2.36)$$

$$\Pr(A_1) = \Pr(x \in \mathcal{X}_1) \quad (4.2.37)$$

The probability of error can be rewritten as

$$P_{\text{error}} = \sum_{\{x:x \in \mathcal{X}_1\}} p(x|H_0)p(H_0) + \sum_{\{x:x \in \mathcal{X}_0\}} p(x|H_1)p(H_1) \quad (4.2.38)$$

For each  $x \in \mathcal{X}$ , we can imagine placing it into either  $\mathcal{X}_0$ , in which case its contribution to  $P_{\text{error}}$  becomes  $p(x|H_1)p(H_1)$ , or we can place it into  $\mathcal{X}_1$ , in which case its contribution to  $P_{\text{error}}$  becomes  $p(x|H_0)p(H_0)$ . To minimise  $P_{\text{error}}$ , we should place it into the partition which minimises the contribution. Thus for each  $x \in \mathcal{X}$ , we should place it into  $\mathcal{X}_0$  if

$$p(x|H_1)p(H_1) \leq p(x|H_0)p(H_0) \quad (4.2.39)$$

which gives the maximum a posteriori decision rule. Intuitively, the maximum a posteriori decision rule decides based on which hypothesis is most likely, after taking into account prior information. The maximum likelihood decision rule can be considered a special case of the maximum a posteriori decision rule, when all prior probabilities are the same.

### Minimum Cost Hypothesis Testing [220]

Consider a binary hypothesis test with hypotheses  $H_0$  and  $H_1$ . After conducting an experiment to test the hypotheses, we observe a decision statistic  $X$ . Let  $c_{\text{FA}}$  be the cost incurred by a Type I error (i.e. false alarm) and let  $c_{\text{miss}}$  be the cost incurred by a Type II error (i.e. miss). Then the minimum cost decision rule based on a realisation of the decision statistic  $x$  is defined by:

- Decide  $H_0$  if  $p(x|H_0)p(H_0)c_{\text{FA}} \geq p(x|H_1)p(H_1)c_{\text{miss}}$ .
- Otherwise, decide  $H_1$ .

Note that upon deciding  $H_0$ , we avoid false alarms and expose ourselves to misses. Thus,

$$p(x|H_0)p(H_0)c_{\text{FA}} \propto \frac{p(x|H_0)p(H_0)}{p(x)}c_{\text{FA}} \quad (4.2.40)$$

$$= p(H_0|x)c_{\text{FA}} \quad (4.2.41)$$

is proportional to the expected cost  $p(H_0|x)c_{\text{FA}}$  avoided by deciding  $H_0$ , while

$$p(x|H_1)p(H_1)c_{\text{miss}} \propto \frac{p(x|H_1)p(H_1)}{p(x)}c_{\text{miss}} \quad (4.2.42)$$

$$= p(H_1|x)c_{\text{miss}} \quad (4.2.43)$$

is the expected cost exposed to  $p(H_1|x)c_{\text{miss}}$  by deciding  $H_0$ . Hence the minimum cost decision rule can be justified as avoiding the larger cost and exposing ourselves to the smaller cost, given  $X$ . We can also show that this decision rule minimises the expected cost,  $C$  (before the experiment has taken place). The expect cost can be written as

$$\mathbb{E}[C] = \Pr(A_1|H_0)p(H_0)c_{\text{FA}} + \Pr(A_0|H_1)p(H_1)c_{\text{miss}} \quad (4.2.44)$$

$$= \sum_{\{x:x \in \mathcal{X}_1\}} p(x|H_0)p(H_0)c_{\text{FA}} + \sum_{\{x:x \in \mathcal{X}_0\}} p(x|H_1)p(H_1)c_{\text{miss}} \quad (4.2.45)$$

with notation as per the maximum a posteriori decision rule. And in the same way, we can show that partitioning  $\mathcal{X}$  so that it leads to the smallest contribution to  $\mathbb{E}[C]$  yields the minimum cost decision rule. The maximum a posteriori decision rule can be viewed as a special case of the minimum cost decision rule, where  $c_{\text{FA}}$  and  $c_{\text{miss}}$  are the same.

### Neyman-Pearson Hypothesis Testing [220]

Consider a binary hypothesis test with hypotheses  $H_0$  and  $H_1$ . After conducting an experiment to test the hypotheses, we observe a decision statistic  $X$ . From  $X$ , we can form another statistic called the likelihood ratio, defined by

$$\Lambda(x) = \frac{p(x|H_0)}{p(x|H_1)} \quad (4.2.46)$$

Then a class of decision rules can be formed by putting a threshold  $\eta$  on the likelihood ratio, according to:

- Decide  $H_0$  if  $\Lambda(x) \geq \eta$ .
- Otherwise, decide  $H_1$ .

The maximum likelihood, maximum a posterior and minimum cost decision rules can then be stated as special cases within this class of tests, with choice of  $\eta$  according to the table below.

| Decision rule        | $\eta$  |
|----------------------|---|
| Maximum likelihood   | 1   |
| Maximum a posteriori | $\frac{p(H_1)}{p(H_0)}$                             |
| Minimum cost         | $\frac{p(H_1)c_{\text{miss}}}{p(H_0)c_{\text{FA}}}$ |

The Neyman-Pearson decision rule is to set a threshold  $\eta$  to minimise  $P_{\text{miss}}$  such that  $P_{\text{FA}} \leq \alpha$ , where  $\alpha$  is some level of significance. We heuristically explain how to find such an  $\eta$ . Firstly, note that

$$P_{\text{FA}} = \Pr(\Lambda(X) < \eta | H_0) \quad (4.2.47)$$

is weakly increasing in  $\eta$ , whereas

$$P_{\text{miss}} = \Pr(\Lambda(X) \geq \eta | H_1) \quad (4.2.48)$$

is weakly decreasing in  $\eta$ . So to make  $P_{\text{miss}}$  as small as possible, we should choose the largest  $\eta$  that still satisfies  $P_{\text{FA}} \leq \alpha$ . This amounts to finding the  $\eta$  such that

$$\Pr(\Lambda(X) < \eta | H_0) = \alpha \quad (4.2.49)$$

or if this equality cannot be satisfied exactly (for example if  $X$  is discrete), then choose the largest  $\eta$  such that

$$\Pr(\Lambda(X) < \eta | H_0) \leq \alpha \quad (4.2.50)$$

### Neyman-Pearson Lemma

The Neyman-Pearson Lemma ensures that thresholding the likelihood ratio  $\Lambda(X)$  (and no other statistic derived from  $X$ , where  $X$  is continuous) using the Neyman-Pearson decision rule delivers the most powerful test (i.e. least  $P_{\text{miss}}$ ) at level of significance  $P_{\text{FA}} = \alpha$ .

*Proof.* For a threshold  $\eta$ , we can partition the support of  $X$  into:

$$\mathcal{X}_0 = \{x \in \mathcal{X} : \Lambda(x) \geq \eta\} \quad (4.2.51)$$

$$\mathcal{X}_1 = \{x \in \mathcal{X} : \Lambda(x) < \eta\} \quad (4.2.52)$$

Let  $\mathcal{X}_0, \mathcal{X}_1$  be the decision regions from using the Neyman-Pearson decision rule with likelihood ratio, and let  $\mathcal{X}'_0, \mathcal{X}'_1$  be the decision regions from another arbitrary decision rule with  $P_{\text{FA}} \leq \alpha$ . The power of the Neyman-Pearson hypothesis test is  $\Pr(X \in \mathcal{X}_1 | H_1)$ , and we seek to prove that

$$\Pr(X \in \mathcal{X}_1 | H_1) \geq \Pr(X \in \mathcal{X}'_1 | H_1) \quad (4.2.53)$$

for any other  $\mathcal{X}'_1 \subseteq \mathcal{X}$ . In the Neyman-Pearson hypothesis test we require  $\Pr(X \in \mathcal{X}_1 | H_0) = \alpha$ , while our other test must satisfy  $\Pr(X \in \mathcal{X}'_1 | H_0) \leq \alpha$ . Hence

$$\Pr(X \in \mathcal{X}'_1 | H_0) \leq \Pr(X \in \mathcal{X}_1 | H_0) \quad (4.2.54)$$

Notice due to mutual exclusivity that

$$\Pr(X \in \mathcal{X}'_1 | H_0) = \Pr(X \in \mathcal{X}'_1 \cap X \in \mathcal{X}_1 | H_0) + \Pr(X \in \mathcal{X}'_1 \cap X \notin \mathcal{X}_1 | H_0) \quad (4.2.55)$$

$$\Pr(X \in \mathcal{X}_1 | H_0) = \Pr(X \in \mathcal{X}_1 \cap X \in \mathcal{X}'_1 | H_0) + \Pr(X \in \mathcal{X}_1 \cap X \notin \mathcal{X}'_1 | H_0) \quad (4.2.56)$$

As the first term is common to both, this implies that

$$\Pr(X \in \mathcal{X}_1 \cap X \notin \mathcal{X}'_1 | H_0) \geq \Pr(X \in \mathcal{X}'_1 \cap X \notin \mathcal{X}_1 | H_0) \quad (4.2.57)$$

As the powers  $\Pr(X \in \mathcal{X}_1 | H_1)$  and  $\Pr(X \in \mathcal{X}'_1 | H_1)$  can be decomposed in the same way, it suffices to show

$$\Pr(X \in \mathcal{X}_1 \cap X \notin \mathcal{X}'_1 | H_1) \geq \Pr(X \in \mathcal{X}'_1 \cap X \notin \mathcal{X}_1 | H_1) \quad (4.2.58)$$

Beginning from the left-hand side,

$$\Pr(X \in \mathcal{X}_1 \cap X \notin \mathcal{X}'_1 | H_1) = \int_{\mathcal{X}_1 \cap \mathcal{X}'_0} p(x | H_1) dx \quad (4.2.59)$$

$$> \frac{1}{\eta} \int_{\mathcal{X}_1 \cap \mathcal{X}'_0} p(x | H_0) dx \quad (4.2.60)$$

$$= \frac{1}{\eta} \Pr(X \in \mathcal{X}_1 \cap X \notin \mathcal{X}'_1 | H_0) \quad (4.2.61)$$

since  $p(x | H_1) > \frac{1}{\eta} p(x | H_0)$  or equivalently  $\frac{p(x | H_0)}{p(x | H_1)} < \eta$  for all  $x \in \mathcal{X}_1$ . Continuing, we apply the inequality obtained above to give

$$\Pr(X \in \mathcal{X}_1 \cap X \notin \mathcal{X}'_1 | H_1) > \frac{1}{\eta} \Pr(X \in \mathcal{X}'_1 \cap X \notin \mathcal{X}_1 | H_0) \quad (4.2.62)$$

$$= \frac{1}{\eta} \int_{\mathcal{X}'_1 \cap \mathcal{X}_0} p(x | H_0) dx \quad (4.2.63)$$

$$\geq \int_{\mathcal{X}'_1 \cap \mathcal{X}_0} p(x | H_1) dx \quad (4.2.64)$$

$$= \Pr(X \in \mathcal{X}'_1 \cap X \notin \mathcal{X}_1 | H_1) \quad (4.2.65)$$

as required, since  $\frac{1}{\eta} p(x | H_0) \geq p(x | H_1)$  or equivalently  $\frac{p(x | H_0)}{p(x | H_1)} \geq \eta$  for all  $x \in \mathcal{X}_0$ .  $\square$

### 4.2.3 Admissible Decision Rules [17]

Admissible decision rules are a way to define the ‘best’ decision rules. Let  $\theta \in \Theta$  denote the ‘true state of nature’ (e.g. it could be the value of a population parameter). Let  $x \in \mathcal{X}$  be some data that we observe. Assume that  $x$  is generated according to some conditional probability distribution given  $\theta$ , denoted  $p(x | \theta)$ . Based on this observation, we make a non-random decision from the action set  $\mathcal{A}$ , using the decision rule  $\delta : \mathcal{X} \rightarrow \mathcal{A}$ . For example, the decision could be an estimate of  $\theta$ . Based on the quality of our decision, we will incur some loss  $L : \Theta \times \mathcal{A} \rightarrow \mathbb{R}$ . For instance, if the decision is to estimate  $\theta$  from the data, then a squared error loss could be defined by

$$L(\theta, \delta(x)) = \|\theta - \delta(x)\|^2 \quad (4.2.66)$$

The loss is a random quantity since  $x$  is random. To derive a deterministic quantity of the quality of decision rule, we define the risk  $R : \Theta \times \delta$  as the expectation of the loss over the distribution  $p(x | \theta)$ :

$$R(\theta, \delta) = \mathbb{E}[L(\theta, \delta(x))] \quad (4.2.67)$$

Hence whether the decision rule has a low risk also depends on the true value of  $\theta$ . We say that a decision rule  $\delta^*$  dominates another decision rule  $\delta$  if and only if

$$R(\theta, \delta^*) \leq R(\theta, \delta) \quad (4.2.68)$$

for all  $\theta \in \Theta$ , with strict inequality for at least some  $\theta$ . Then a decision rule  $\delta^*$  is said to be *admissible* (with respect to the risk  $R$ ) if no other decision rule dominates it. Conversely if a decision rule is dominated by at least one other decision rule, then it is *inadmissible* (with respect to the risk  $R$ ).

#### 4.2.4 Unbiased Tests [126]

Let  $\Theta_0$  be a set of null hypotheses and  $\Theta_1$  be a set of alternative hypotheses. Let  $\beta(\theta)$  denote the *power function* of a test, which is the probability of rejecting the null under  $\theta \in \Theta_0 \cup \Theta_1$ . Note that this departs from the traditional definition of the power of a test; we allow for the notion of the power function even the null is true (in which case the rejection probability would be termed the size). Suppose the test has a nominal level of significance  $\alpha$ . Then the test is said to be unbiased if

$$\beta(\theta) \leq \alpha \quad (4.2.69)$$

for all  $\theta \in \Theta_0$ , and

$$\beta(\theta) \geq \alpha \quad (4.2.70)$$

for all  $\theta \in \Theta_1$ . Put another way, this implies that when the alternative is true, the probability of rejection is never smaller than when the null is true. An ideal test will have  $\beta(\theta) = \alpha$  for all  $\theta \in \Theta_0$  and  $\beta(\theta) = 1$  for all  $\theta \in \Theta_1$ , so an ideal test will of course also be unbiased.

#### 4.2.5 Consistent Tests [126]

Let  $\Theta_1$  be a set of alternative hypotheses. Then a test is said to be consistent (or more precisely, *pointwise consistent in power*) if the power function  $\beta_n(\theta)$  indexed in the sample size  $n$  satisfies

$$\lim_{n \rightarrow \infty} \beta_n(\theta) = 1 \quad (4.2.71)$$

for all  $\theta \in \Theta_1$ . In words, it means that if the alternative is true, the test will end up always rejecting the null as  $n \rightarrow \infty$ .

#### 4.2.6 Uniformly Most Powerful Tests

#### 4.2.7 Uncertainty Quantification

Uncertainty quantification is the science of characterising uncertainty (i.e. the distribution of errors, in a statistical sense) for a model.

##### Aleatoric Uncertainty

Aleatoric uncertainty refers to noise which can cause the outcome of an observation to be different each time we run the same experiment. This can be caused by, say, the explanatory variables not capturing all the information required to precisely predict a dependent variable. Suppose we prescribe the data generating process with a model

$$Y = f(x) + \varepsilon \quad (4.2.72)$$

Then the noise term  $\varepsilon$  causes the aleatoric uncertainty, which can be quantified by  $\text{Var}(Y) = \text{Var}(\varepsilon)$ . Aleatoric uncertainty can be decomposed into homoskedastic uncertainty and heteroskedastic uncertainty, with the latter being the uncertainty which depends on the input values. For instance, suppose we are able to write

$$\text{Var}(Y|X = x) = \sigma^2 + \varsigma^2(x) \quad (4.2.73)$$

Then the term  $\sigma^2$  quantifies the homoskedastic uncertainty and the term  $\varsigma^2(x)$  which is a function of  $x$  quantifies the heteroskedastic uncertainty.

### Epistemic Uncertainty

Epistemic uncertainty refers to error due to uncertainty in model parameters - values which are difficult to know in practice. Suppose (in the absence of aleatoric uncertainties), that we have the model parametrised by  $\theta$ :

$$y = f(x; \theta) \quad (4.2.74)$$

However we estimate  $\hat{\theta}$  for  $\theta$  with some variance  $\text{Var}(\hat{\theta})$ , and thus predict

$$\hat{Y} = f(x; \hat{\theta}) \quad (4.2.75)$$

Then the epistemic uncertainty can be quantified by  $\text{Var}(f(x; \hat{\theta}))$ , which we can crudely relate to  $\text{Var}(\hat{\theta})$  via a first-order Taylor approximation, assuming  $f(x; \theta)$  is differentiable with respect to  $\theta$ :

$$f(x; \hat{\theta}) \approx f\left(x; \mathbb{E}[\hat{\theta}]\right) + (\hat{\theta} - \mathbb{E}[\hat{\theta}]) \frac{\partial f(x; \theta)}{\partial \theta} \Big|_{\theta=\mathbb{E}[\hat{\theta}]} \quad (4.2.76)$$

So

$$\text{Var}(f(x; \hat{\theta})) \approx \text{Var}(\hat{\theta}) \frac{\partial f(x; \theta)}{\partial \theta} \Big|_{\theta=\mathbb{E}[\hat{\theta}]} \quad (4.2.77)$$

### Propagation of Errors

Consider a generally nonlinear model  $f(x)$  in the explanatory variables  $x$ , where  $x$  is a vector of dimension  $n$ . However, it is known that we obtain or measure  $x$  with some error  $\varepsilon$  with  $\text{Cov}(\varepsilon) = \Sigma$ , so that we have the random vector  $\mathbf{X} := x + \varepsilon$ . We then ask how this error propagates through to the estimate  $f(\mathbf{X})$ . Suppose  $\varepsilon$  is a continuous random variable and  $f(x)$  is differentiable. Then by a first order Taylor series approximation,

$$f(x + \varepsilon) \approx f(x) + \sum_{i=1}^n \frac{\partial f}{\partial x_i} \varepsilon_i \quad (4.2.78)$$

$$= f(x) + \nabla f(x)^\top \varepsilon \quad (4.2.79)$$

So then we can approximate the variance of  $f(x + \varepsilon)$  by

$$\text{Var}(f(x + \varepsilon)) \approx \text{Var}\left(f(x) + \nabla f(x)^\top \varepsilon\right) \quad (4.2.80)$$

$$= \text{Cov}\left(\nabla f(x)^\top \varepsilon\right) \quad (4.2.81)$$

$$= \nabla f(x)^\top \text{Cov}(\varepsilon) \nabla f(x) \quad (4.2.82)$$

$$= \nabla f(x)^\top \Sigma \nabla f(x) \quad (4.2.83)$$

More generally, suppose we have a vector-valued nonlinear model  $\mathbf{f}(x)$ , then let  $\mathbf{J} = \frac{\partial \mathbf{f}}{\partial x}$  denote the Jacobian matrix. Then the first order Taylor series approximation is

$$\mathbf{f}(x + \varepsilon) \approx \mathbf{f}(x) + \mathbf{J}\varepsilon \quad (4.2.84)$$

and by the same vein, an approximation of the error propagation is

$$\text{Cov}(\mathbf{f}(x + \varepsilon)) \approx \mathbf{J}\Sigma\mathbf{J}^\top \quad (4.2.85)$$

Even more generally, suppose we have a composition of nonlinear mappings  $\mathbf{g} \circ \mathbf{f}(x)$ . Then taking the Jacobian matrix  $\frac{\partial \mathbf{g}}{\partial \mathbf{f}}$  we come up with the approximation

$$\text{Cov}(\mathbf{g} \circ \mathbf{f}(x + \varepsilon)) \approx \frac{\partial \mathbf{g}}{\partial \mathbf{f}} \mathbf{J}\Sigma\mathbf{J}^\top \frac{\partial \mathbf{g}}{\partial \mathbf{f}}^\top \quad (4.2.86)$$

We can in this way approximate the error propagation through an arbitrary amount of nested nonlinear models by chaining Jacobians, keeping in mind that the approximation may get worse and worse through multiple compositions.

## 4.3 Least Squares

### 4.3.1 Ordinary Least Squares

Suppose we have  $n$  data pairs consisting of regressors  $\mathbf{x}_i$  (vectors) and scalar responses  $y_i$ . The task is to fit a linear model of the form  $\hat{y}_i = \hat{\beta}^\top \mathbf{x}_i$  where  $\hat{\beta}$  is an estimated  $d$ -dimensional vector of parameters. We use the least squares cost function

$$V(\beta) = \frac{1}{2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4.3.1)$$

$$= \frac{1}{2} \sum_{i=1}^n (y_i - \beta^\top \mathbf{x}_i)^2 \quad (4.3.2)$$

Using matrix notation, this problem can be written down more compactly. Denote by the matrices  $\mathbf{X}$  and  $\mathbf{Y}$ :

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \quad (4.3.3)$$

$$\mathbf{Y} = [y_1 \ \dots \ y_n]^\top \quad (4.3.4)$$

Hence  $\mathbf{X}$  is a ‘tall’ matrix and  $\mathbf{Y}$  is actually a column vector. Note that if we wish to include a constant in the regression function, i.e.  $\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{d-1} x_{i,d-1}$ , then we may simply let the first element of  $\mathbf{x}_i$  be equal to 1. Using this matrix notation, we can rewrite the least squares problem as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} V(\beta) \quad (4.3.5)$$

$$= \underset{\beta}{\operatorname{argmin}} \frac{1}{2} (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) \quad (4.3.6)$$

Expanding out the quadratic cost function,

$$V(\beta) = \frac{1}{2} \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\beta + \frac{1}{2} \beta^\top \mathbf{X}^\top \mathbf{X}\beta \quad (4.3.7)$$

Taking the derivative:

$$\nabla V(\beta) = -\mathbf{X}^\top \mathbf{Y} + \mathbf{X}^\top \mathbf{X}\beta \quad (4.3.8)$$

Setting this to zero and solving obtains the ordinary least squares estimator.

$$-\mathbf{X}^\top \mathbf{Y} + \mathbf{X}^\top \mathbf{X}\hat{\beta} = 0 \quad (4.3.9)$$

$$\mathbf{X}^\top \mathbf{X}\hat{\beta} = \mathbf{X}^\top \mathbf{Y} \quad (4.3.10)$$

This is known as the least squares normal equation, which has solution

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \quad (4.3.11)$$

An alternative way to compute the derivative is via the chain rule (on the un-expanded cost function), using the facts  $\nabla_\beta(\mathbf{X}\beta) = \mathbf{X}^\top$  and  $\nabla_\beta(\beta^\top \beta) = 2\beta$  to give

$$\nabla V(\beta) = \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta) \quad (4.3.12)$$

which can be set to zero and rearranged to arrive at the least squares estimator. Note that the matrix  $\mathbf{X}^\dagger = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  is the left Moore-Penrose pseudoinverse of  $\mathbf{X}$ . So we can alternatively write the least squares estimator as

$$\hat{\beta} = \mathbf{X}^\dagger \mathbf{Y} \quad (4.3.13)$$

which is the value of  $\beta$  which ‘solves’  $\mathbf{X}\beta = \mathbf{Y}$ .

### Existence of Ordinary Least Squares Estimator

In order for the ordinary least squares estimator to exist, the matrix  $\mathbf{X}$  must be full column-rank. To see why, it suffices to show that  $\mathbf{X}^\top \mathbf{X}$  and  $\mathbf{X}$  have the same rank. We claim that the linear systems of equations  $\mathbf{X}v = 0$  and  $\mathbf{X}^\top \mathbf{X}v = 0$  for some vector  $v$  share the same solutions. Firstly, if  $\mathbf{X}v = 0$  then we have

$$\mathbf{X}^\top \mathbf{X}v = \mathbf{X}^\top (\mathbf{X}v) \quad (4.3.14)$$

$$= 0 \quad (4.3.15)$$

For the reverse direction, if  $\mathbf{X}^\top \mathbf{X}v = 0$  then

$$v^\top \mathbf{X}^\top \mathbf{X}v = 0 \quad (4.3.16)$$

$$(\mathbf{X}v)^\top \mathbf{X}v = 0 \quad (4.3.17)$$

which implies  $\mathbf{X}v = 0$  because by definition of the dot product,  $(\mathbf{X}v)^\top \mathbf{X}v = 0$  if and only if  $\mathbf{X}v = 0$ . Hence the null spaces (set of solutions to  $\mathbf{X}v = 0$  and  $\mathbf{X}^\top \mathbf{X}v = 0$ ) are the same, and therefore the nullity (dimension of null space) of  $\mathbf{X}$  and  $\mathbf{X}^\top \mathbf{X}$  are identical. By the Rank-Nullity Theorem,  $\text{rank}(\mathbf{X}) + \text{null}(\mathbf{X}) = k$  so it follows that  $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^\top \mathbf{X})$ . We now know that if  $\text{rank}(\mathbf{X}) < k$ , then  $\text{rank}(\mathbf{X}^\top \mathbf{X}) < k$  and thus  $\mathbf{X}^\top \mathbf{X}$ , whose inverse appears in the ordinary least squares estimator, is not invertible.

### Sampling Distribution of Ordinary Least Squares Residuals

Assume the data generating process (which we call the *general linear model*):

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (4.3.18)$$

where parameter vector  $\beta \in \mathbb{R}^p$ , error terms  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$  and we assume that the regressors  $\mathbf{X}$  are non-random. Then the residuals  $\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}$  under the ordinary least squares estimate is given by

$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta} \quad (4.3.19)$$

$$= \mathbf{Y} - \mathbf{X} \left[ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \right] \quad (4.3.20)$$

$$= \left[ I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \mathbf{Y} \quad (4.3.21)$$

$$= \left[ I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] (\mathbf{X}\beta + \varepsilon) \quad (4.3.22)$$

$$= \left[ \mathbf{X} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) \right] \beta + \left[ I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \varepsilon \quad (4.3.23)$$

$$= \underbrace{\left[ I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right]}_Q \varepsilon \quad (4.3.24)$$

where we may call  $Q$  the ‘residual-maker’ matrix since  $\hat{\varepsilon} = Q\mathbf{Y}$ . It is also sometimes called the ‘annihilator’ matrix because  $Q\mathbf{X} = \mathbf{0}$ . Note that  $Q$  is symmetric ( $Q^\top = Q$ ) and idempotent ( $Q^2 = Q$ ). We may verify the latter property by

$$Q^2 = \left[ I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \left[ I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \quad (4.3.25)$$

$$= I - 2\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \quad (4.3.26)$$

$$= 1 - 2\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \quad (4.3.27)$$

$$= I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \quad (4.3.28)$$

$$= Q \quad (4.3.29)$$

We then use the property of idempotent matrices that their eigenvalues can only either be 0 or 1. To show this, we may write for some  $\mathbf{v} \neq 0$ :

$$Q\mathbf{v} = \lambda\mathbf{v} \quad (4.3.30)$$

$$QQ\mathbf{v} = \lambda Q\mathbf{v} \quad (4.3.31)$$

$$Q\mathbf{v} = \lambda(\lambda\mathbf{v}) \quad (4.3.32)$$

$$\lambda\mathbf{v} = \lambda^2\mathbf{v} \quad (4.3.33)$$

which only admits solutions of either  $\lambda = 1$  or  $\lambda = 0$ . Furthermore, we can compute (using the invariance of trace to cyclic permutations):

$$\text{trace}(Q) = \text{trace}(I_{n \times n}) - \text{trace}\left(\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right) \quad (4.3.34)$$

$$= n - \text{trace}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\right) \quad (4.3.35)$$

$$= n - \text{trace}(I_{p \times p}) \quad (4.3.36)$$

$$= n - p \quad (4.3.37)$$

An eigendecomposition of  $Q = V\Lambda V^\top$  where  $V^\top = V^{-1}$  then shows

$$\text{trace}(Q) = \text{trace}(V\Lambda V^\top) \quad (4.3.38)$$

$$= \text{trace} (V \Lambda V^\top) \quad (4.3.39)$$

$$= \text{trace} (\Lambda V^\top V) \quad (4.3.40)$$

$$= \text{trace} (\Lambda) \quad (4.3.41)$$

which yields  $\text{trace} (\Lambda) = n - p$ . As  $\Lambda$  contains the eigenvalues of  $Q$ , then this establishes that  $\Lambda$  must contain exactly  $n - p$  ones and  $p$  zeros in its main diagonal. Without loss of generality, we can assume that all the ones appear at the beginning of the main diagonal (as this aspect of eigendecomposition is arbitrary) so

$$\Lambda = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \quad (4.3.42)$$

$$= \text{diag} \left\{ \underbrace{1, \dots, 1}_{n-p}, \underbrace{0, \dots, 0}_p \right\} \quad (4.3.43)$$

Now let  $K = V^\top \hat{\varepsilon}$ . By  $\hat{\varepsilon} = Q\varepsilon$ , we have that

$$\hat{\varepsilon} \sim \mathcal{N} (0, \sigma^2 Q I Q^\top) \quad (4.3.44)$$

$$\sim \mathcal{N} (0, \sigma^2 Q) \quad (4.3.45)$$

So

$$K \sim \mathcal{N} (0, \sigma^2 V^\top Q V) \quad (4.3.46)$$

$$\sim \mathcal{N} (0, \sigma^2 \Lambda) \quad (4.3.47)$$

Hence the last  $p$  elements of  $K$  are zero. By the characterisation of the chi-squared distribution as the sum of squared independent standard normal random variables, this means

$$\frac{\|K\|^2}{\sigma^2} \sim \chi_{n-p}^2 \quad (4.3.48)$$

The residual sum of squares is given by  $\text{RSS} = \|\hat{\varepsilon}\|^2$ . Also as  $V^\top$  is an orthogonal matrix, then  $\|K\|^2 = \|\hat{\varepsilon}\|^2$  and therefore the residual sum of squares has distribution

$$\frac{\text{RSS}}{\sigma^2} \sim \chi_{n-p}^2 \quad (4.3.49)$$

Using the fact that the mean of the chi-squared distribution is equal to its degrees of freedom, we have

$$\mathbb{E} \left[ \frac{\text{RSS}}{\sigma^2} \right] = n - p \quad (4.3.50)$$

Rearranging, we also arrive at an unbiased estimator for the error variance:

$$\mathbb{E} \left[ \frac{\text{RSS}}{n - p} \right] = \sigma^2 \quad (4.3.51)$$

Note that we do not require normality of the error terms for the above unbiased estimator for  $\sigma^2$  to be valid, which we show as follows. A relaxation is that they are zero-mean and uncorrelated, i.e.  $\mathbb{E}[\varepsilon] = 0$  and  $\text{Cov}(\varepsilon) = \sigma^2 I$ . Then the residual sum of squares is given by

$$\mathbb{E}[\text{RSS}] = \mathbb{E}[\hat{\varepsilon}^\top \hat{\varepsilon}] \quad (4.3.52)$$

$$= \mathbb{E}[\varepsilon^\top Q^\top Q \varepsilon] \quad (4.3.53)$$

$$= \mathbb{E}[\varepsilon^\top Q \varepsilon] \quad (4.3.54)$$

Note that  $Q$  is necessarily positive definite if we assume that always  $\text{RSS} > 0$ . Then using the result on expectations of inner products:

$$\mathbb{E}[\text{RSS}] = \text{trace}(\sigma^2 Q) + \mathbb{E}[\varepsilon^\top] Q \mathbb{E}[\varepsilon] \quad (4.3.55)$$

$$= \sigma^2 \text{trace}(Q) \quad (4.3.56)$$

$$= \sigma^2(n - p) \quad (4.3.57)$$

which yields the same expectation as above.

### Standard Errors in Ordinary Least Squares

If we assume the data generating process

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (4.3.58)$$

where  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$  and we assume that the regressors  $\mathbf{X}$  are non-random, then

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I) \quad (4.3.59)$$

By the ordinary least squares estimator  $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ ; this is just a linear transformation of the multivariate normally distributed  $\mathbf{Y}$ , so the sampling distribution of  $\hat{\beta}$  is

$$\hat{\beta} \sim \mathcal{N}\left(\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\beta, \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \sigma^2 I \left[\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top\right]^\top\right) \quad (4.3.60)$$

$$\sim \mathcal{N}\left(\beta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right) \quad (4.3.61)$$

$$\sim \mathcal{N}\left(\beta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right) \quad (4.3.62)$$

Hence

$$\text{Cov}(\hat{\beta}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \quad (4.3.63)$$

The term  $\sigma^2$  may be estimated by the sample variance of the residuals (unbiased estimator), so that the estimate of  $\text{Cov}(\hat{\beta})$  is

$$\widehat{\text{Cov}}(\hat{\beta}) = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{\beta}^\top x_i)^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \quad (4.3.64)$$

where  $p$  is the dimension of  $\hat{\beta}$ . To obtain the standard error  $\text{se}(\hat{\beta}_k)$  for the  $k^{\text{th}}$  coefficient estimate, we take the square root of the  $k^{\text{th}}$  diagonal of  $\widehat{\text{Cov}}(\hat{\beta})$ . To perform coefficients on the

$k^{\text{th}}$  coefficient of interest however, it is not possible to use  $z$ -statistics if the population variance  $\sigma^2$  is not known. We consider the distribution of the standardised score:

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)} \quad (4.3.65)$$

From the distribution of  $\hat{\beta}$ , we have that

$$\hat{\beta} - \beta \sim \mathcal{N}\left(0, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right) \quad (4.3.66)$$

Let  $s_{kk}$  denote the  $k^{\text{th}}$  diagonal of  $(\mathbf{X}^\top \mathbf{X})^{-1}$ . We then have that

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{s_{kk}}} \quad (4.3.67)$$

$$\sim \mathcal{N}(0, 1) \quad (4.3.68)$$

From the distribution of the residuals, we know that  $V := \frac{\|\hat{\varepsilon}\|^2}{\sigma^2}$  has a chi-squared distribution with  $n - p$  degrees of freedom. It can then be shown that

$$\text{Cov}(\hat{\beta}, \hat{\varepsilon}) = \mathbb{E}[\hat{\beta} \hat{\varepsilon}^\top] - \mathbb{E}[\hat{\beta}] \mathbb{E}[\hat{\varepsilon}]^\top \quad (4.3.69)$$

$$= \mathbb{E}[\hat{\beta} \hat{\varepsilon}^\top] \quad (4.3.70)$$

where we have used  $\mathbb{E}[\hat{\varepsilon}] = \mathbb{E}[Q\varepsilon] = 0$ . Then from

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \quad (4.3.71)$$

$$= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \varepsilon) \quad (4.3.72)$$

$$= \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon \quad (4.3.73)$$

and the residual-maker matrix for  $\varepsilon = [I - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top] \hat{\varepsilon}$ , it follows that

$$\text{Cov}(\hat{\beta}, \hat{\varepsilon}) = \mathbb{E}\left[\left(\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon\right) \varepsilon^\top \left(I - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right)\right] \quad (4.3.74)$$

$$= \mathbb{E}\left[\beta \varepsilon^\top I - \beta \varepsilon^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon \varepsilon^\top - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon \varepsilon^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right] \quad (4.3.75)$$

$$= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\varepsilon \varepsilon^\top] - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\varepsilon \varepsilon^\top] \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \quad (4.3.76)$$

$$= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 I - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 I \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \quad (4.3.77)$$

$$= \sigma^2 \left[ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \quad (4.3.78)$$

$$= \mathbf{0}_{p \times n} \quad (4.3.79)$$

Since  $\hat{\beta}$  and  $\hat{\varepsilon}$  are uncorrelated and they are both multivariate normally distributed, then this implies that  $\hat{\beta}$  and  $\hat{\varepsilon}$  are independent. Hence  $z_k$  and  $V$  are also independent. Thus

$$t_k = \frac{z_k}{\sqrt{V/(n-p)}} \quad (4.3.80)$$

fits the characterisation of a  $t$ -distributed random variable with  $n - p$  degrees of freedom. We are then left to show that this is equal to the standardised statistic from above:

$$t_k = \frac{z_k}{\sqrt{V/(n-p)}} \quad (4.3.81)$$

$$= \frac{(\hat{\beta}_k - \beta_k) / \sqrt{\sigma^2 s_{kk}}}{\sqrt{\frac{\|\hat{\varepsilon}\|^2}{\sigma^2} / (n-p)}} \quad (4.3.82)$$

$$= \frac{\hat{\beta}_k - \beta_k}{\sqrt{\frac{\|\hat{\varepsilon}\|^2}{n-p} \cdot s_{kk}}} \quad (4.3.83)$$

$$= \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{s_{kk}}} \quad (4.3.84)$$

$$= \frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)} \quad (4.3.85)$$

Hence the fact that this statistic is  $t$ -distributed with  $n - p$  degrees of freedom may be used to conduct hypothesis tests and construct confidence intervals for  $\beta_k$ .

### Confidence Ellipses in Ordinary Least Squares

### Prediction Intervals in Ordinary Least Squares

### Decomposition of the Total Sum of Squares in Multiple Regression

As is the case in simple linear regression, we can decompose the total sum of squares TSS as

$$\text{TSS} = \text{ESS} + \text{RSS} \quad (4.3.86)$$

where ESS is the ‘explained’ sum of squares and RSS is the residual sum of squares. To show this, we begin with the form of the general linear model  $\mathbf{y} = \mathbf{X}\beta + \varepsilon$  (including intercept coefficient) and write the total sum of squares as

$$\text{TSS} = \|\mathbf{y} - \bar{\mathbf{y}}\|^2 \quad (4.3.87)$$

$$= \|\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 \quad (4.3.88)$$

$$= (\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \bar{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \bar{\mathbf{y}}) \quad (4.3.89)$$

where  $\bar{\mathbf{y}}$  represents a vector where each element is  $\bar{y}$ , which is the sample mean across the elements in  $\mathbf{y}$ . Expanding,

$$\text{TSS} = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) + 2(\mathbf{y} - \hat{\mathbf{y}})^\top (\hat{\mathbf{y}} - \bar{\mathbf{y}}) + (\hat{\mathbf{y}} - \bar{\mathbf{y}})^\top (\hat{\mathbf{y}} - \bar{\mathbf{y}}) \quad (4.3.90)$$

$$= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + 2\mathbf{y}^\top \hat{\mathbf{y}} - 2\mathbf{y}^\top \bar{\mathbf{y}} - 2\hat{\mathbf{y}}^\top \hat{\mathbf{y}} + 2\hat{\mathbf{y}}^\top \bar{\mathbf{y}} + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 \quad (4.3.91)$$

We will use the following result involving the residuals:

$$\mathbf{X}^\top \hat{\varepsilon} = \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) \quad (4.3.92)$$

$$= \mathbf{X}^\top \left[ I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \mathbf{y} \quad (4.3.93)$$

$$= (\mathbf{X}^\top - \mathbf{X}^\top) \mathbf{y} \quad (4.3.94)$$

$$= \mathbf{0} \quad (4.3.95)$$

Hence the red terms becomes

$$2\mathbf{y}^\top \hat{\mathbf{y}} - 2\hat{\mathbf{y}}^\top \hat{\mathbf{y}} = 2(\mathbf{y} - \hat{\mathbf{y}})^\top \hat{\mathbf{y}} \quad (4.3.96)$$

$$= 2\hat{\varepsilon}^\top X\hat{\beta} \quad (4.3.97)$$

$$= 0 \quad (4.3.98)$$

As for the blue terms,

$$2\hat{\mathbf{y}}^\top \bar{\mathbf{y}} - 2\mathbf{y}^\top \bar{\mathbf{y}} = -2(\mathbf{y} - \hat{\mathbf{y}})^\top \bar{\mathbf{y}} \quad (4.3.99)$$

$$= -2\hat{\varepsilon}^\top \bar{\mathbf{y}} \quad (4.3.100)$$

$$= -2 \sum_{i=1}^n \hat{\varepsilon}_i \bar{y} \quad (4.3.101)$$

$$= -2\bar{y} \sum_{i=1}^n \hat{\varepsilon}_i \quad (4.3.102)$$

Since  $\mathbf{X}^\top \hat{\varepsilon} = \mathbf{0}$  and a column of  $\mathbf{X}$  will be all ones due to the intercept coefficient, it follows that  $\sum_{i=1}^n \hat{\varepsilon}_i = 0$  so

$$2\hat{\mathbf{y}}^\top \bar{\mathbf{y}} - 2\mathbf{y}^\top \bar{\mathbf{y}} = 0 \quad (4.3.103)$$

Therefore,

$$\text{TSS} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 \quad (4.3.104)$$

$$= \text{RSS} + \text{ESS} \quad (4.3.105)$$

### Multiple Coefficient of Determination

The multiple coefficient of determination is the generalisation of  $R^2$  when there are multiple regressors. It may still be computed as the ratio of explained variation to total variation (or the complement of the ratio of unexplained variation to total variation):

$$R^2 = \frac{\text{ESS}}{\text{TSS}} \quad (4.3.106)$$

$$= 1 - \frac{\text{RSS}}{\text{TSS}} \quad (4.3.107)$$

Other equivalent formulas for  $R^2$  exist. One such is

$$R^2 = \mathbf{c}^\top R_{xx}^{-1} \mathbf{c} \quad (4.3.108)$$

where  $R_{xx}$  is the sample correlation matrix of the regressors, and  $\mathbf{c}$  is the sample cross-correlation vector between the regressors and the dependent variable. To derive this formula, we consider the general linear model  $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$  (intercept can be included) where each row is given by the equation

$$Y_i = \beta^\top X_i + \varepsilon_i \quad (4.3.109)$$

For convenience of notation, we will let  $\widehat{\text{Var}}(\cdot)$  denote the sample estimate of the variance, and analogously for  $\widehat{\text{Cov}}(\cdot)$  and  $\widehat{\mathbb{E}}[\cdot]$ . The reason for doing this is that many properties of the population versions of these quantities will apply or are analogous to the sample versions as well. Using this notation, we can write  $R^2$  (in the ratio of explained to total variation characterisation) as

$$R^2 = \frac{\widehat{\text{Var}}(\beta^\top X_i)}{\widehat{\text{Var}}(Y_i)} \quad (4.3.110)$$

which is equal to  $\frac{\text{ESS}}{\text{TSS}}$  since the numerator can be expressed as

$$\widehat{\text{Var}}(\beta^\top X_i) = \widehat{\mathbb{E}} \left[ (\beta^\top X_i - \widehat{\mathbb{E}}[\beta^\top X_i])^2 \right] \quad (4.3.111)$$

$$= \widehat{\mathbb{E}} \left[ (\beta^\top X_i - \bar{Y})^2 \right] \quad (4.3.112)$$

$$= \frac{1}{n} \sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2 \quad (4.3.113)$$

$$= \frac{1}{n} \text{ESS} \quad (4.3.114)$$

with  $\widehat{Y}_i = \widehat{\beta}^\top X_i$  as the fitted value and  $\bar{Y}$  as the sample mean of  $Y_i$ . Similarly for the denominator,

$$\widehat{\text{Var}}(Y_i) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (4.3.115)$$

$$= \frac{1}{n} \text{TSS} \quad (4.3.116)$$

We assume that the data are centered so  $\bar{Y} = 0$  and  $\widehat{\mathbb{E}}[X_i] = \mathbf{0}$ . This can be done without loss of generality since translating the data to be centered will not affect the  $R^2$ , but it will simplify the calculations. The  $R^2$  as proposed above can be manipulated to become

$$R^2 = \frac{\widehat{\beta}^\top \widehat{\text{Cov}}(X_i) \widehat{\beta}}{\widehat{\text{sd}}(Y_i)^2} \quad (4.3.117)$$

$$= \frac{\mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \widehat{\text{Cov}}(X_i) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}{\widehat{\text{sd}}(Y_i)^2} \quad (4.3.118)$$

$$= \frac{\mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} [\mathbf{X}^\top \mathbf{X}/(n-1)] (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}{\widehat{\text{sd}}(Y_i)^2} \quad (4.3.119)$$

$$= \frac{\mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}{\widehat{\text{sd}}(Y_i)^2 / (n-1)} \quad (4.3.120)$$

where we used  $\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$  by the definition of the OLS estimator and  $\widehat{\text{Cov}}(X_i) = [\mathbf{X}^\top \mathbf{X}/(n-1)]$  for centered data. Note that the latter does not necessarily mean that  $X_i$  needs to be random; the notation just means to take the sample covariance of the data in  $\mathbf{X}$ , even if  $\mathbf{X}$  is a fixed design matrix. Introduce  $D$  as the diagonal matrix of sample standard deviations of each regressor in  $\mathbf{X}$ , such that  $\mathbf{X}D^{-1}$  standardises each column to have unit variance. Then

$$R^2 = \frac{\mathbf{Y}^\top \mathbf{X} D^{-1} D (\mathbf{X}^\top \mathbf{X})^{-1} D D^{-1} \mathbf{X}^\top \mathbf{Y}}{\widehat{\text{sd}}(Y_i)^2 / (n-1)} \quad (4.3.121)$$

$$= \frac{1}{n-1} \cdot \frac{\mathbf{Y}^\top}{\widehat{\text{sd}}(Y_i)} \mathbf{X} D^{-1} \cdot (n-1) D (\mathbf{X}^\top \mathbf{X})^{-1} D \cdot \frac{1}{n-1} D^{-1} \mathbf{X}^\top \frac{\mathbf{Y}}{\widehat{\text{sd}}(Y_i)} \quad (4.3.122)$$

$$= \frac{1}{n-1} \cdot \frac{\mathbf{Y}^\top}{\widehat{\text{sd}}(Y_i)} \mathbf{X} D^{-1} \cdot \left( \frac{D^{-1} \mathbf{X}^\top \mathbf{X} D^{-1}}{n-1} \right)^{-1} \cdot \frac{1}{n-1} D^{-1} \mathbf{X}^\top \frac{\mathbf{Y}}{\widehat{\text{sd}}(Y_i)} \quad (4.3.123)$$

Recognise that

$$\frac{1}{n-1} D^{-1} \mathbf{X}^\top \frac{\mathbf{Y}}{\widehat{\text{sd}}(Y_i)} = \widehat{\text{Corr}}(X_i, Y_i) \quad (4.3.124)$$

gives the sample cross-correlation between  $X_i$  and  $Y_i$  since it is the sample covariance of the standardised data. Also,

$$\frac{D^{-1}\mathbf{X}^\top \mathbf{X} D^{-1}}{n-1} = D^{-1} \widehat{\text{Cov}}(X_i) D^{-1} \quad (4.3.125)$$

$$= \widehat{\text{Corr}}(X_i) \quad (4.3.126)$$

analogously to the population correlation matrix. Hence

$$R^2 = \widehat{\text{Corr}}(X_i, Y_i)^\top \widehat{\text{Corr}}(X_i) \widehat{\text{Corr}}(X_i, Y_i) \quad (4.3.127)$$

which is the result claimed.

### Adjusted $R^2$

The adjusted  $R^2$  (denoted  $\bar{R}^2$ ) is an adjustment of the multiple coefficient of determination, adjusted for degrees of freedom. The standard formula for multiple  $R^2$  can be written as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (4.3.128)$$

$$= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.3.129)$$

$$= 1 - \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.3.130)$$

In this form, the numerator  $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$  can be interpreted as an estimate

$$\widehat{\text{Var}}(\varepsilon) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4.3.131)$$

for the residuals, while the denominator  $\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$  can be interpreted as an estimate

$$\widehat{\text{Var}}(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4.3.132)$$

for the variance of the dependent variable. These estimators will be biased, because they lack the degrees of freedom correction. In the adjusted  $R^2$ , we instead replace these quantities with their unbiased versions, giving

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.3.133)$$

$$= 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \quad (4.3.134)$$

with  $p$  regressors (including intercept). This has the effect of reducing the  $R^2$  value. Since  $p \geq 1$ , then the adjusted  $R^2$  will be less than or equal to the unadjusted  $R^2$ :

$$\bar{R}^2 \leq R^2 \quad (4.3.135)$$

This can be thought of as applying a ‘penalty’ to the measure of regression fit when too many regressors are added. In fact, it is possible for the adjusted  $R^2$  to become negative.

### 4.3.2 Weighted Least Squares

In the case where we ‘trust’ some observations more than others, we can modify the least squares criterion with weightings. Introduce weights  $w_i > 0$  for each observation  $i = 1, \dots, n$ . The weighted linear least squares cost function becomes

$$V(\beta) = \frac{1}{2} \sum_{i=1}^n w_i (y_i - \beta^\top \mathbf{x}_i)^2 \quad (4.3.136)$$

Note that only the relative values of the weights matter. For example, we may set all  $w_i = 1$  by default and adjust  $w_i < 1$  for observations we do not trust or  $w_i > 1$  for observations we really trust. Using matrix notation, denote by  $\mathbf{W}$  the diagonal weighting matrix:

$$\mathbf{W} = \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix} \quad (4.3.137)$$

Then

$$V(\beta) = \frac{1}{2} (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta) \quad (4.3.138)$$

Solving for the estimator  $\hat{\beta}$  takes similar steps as for ordinary least squares, and yields

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y} \quad (4.3.139)$$

Note that if  $w_1 = \dots = w_n$ , then the problem effectively reduces to ordinary least squares.

### 4.3.3 Generalised Least Squares [205]

Generalised least squares (GLS) may be thought of as generalisations to both ordinary and weighted least squares. The major difference is that instead of assuming that the error covariance is  $\sigma^2 I$ , it is instead some positive definite matrix  $\sigma^2 \Sigma$ . For now we suppose that  $\sigma^2$  and  $\Sigma$  are known. In a linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (4.3.140)$$

with  $\text{Cov}(\varepsilon) = \Sigma$ , computing standard errors using the ordinary least squares method will be incorrect. What can be done instead is to firstly note that  $\Sigma$  can be written as

$$\Sigma^{-1} = LL^\top \quad (4.3.141)$$

where  $L$  is an invertible square matrix which is a square root of  $\Sigma^{-1}$  (for example, it could be from the Cholesky decomposition). Then we can observe

$$L^\top \Sigma L = L^\top (LL^\top)^{-1} L \quad (4.3.142)$$

$$= L^\top \left[ (L^\top)^{-1} L^{-1} \right] L \quad (4.3.143)$$

$$= I \quad (4.3.144)$$

From this we can see

$$\text{Cov}(L^\top \varepsilon) = L^\top \text{Cov}(\varepsilon) L \quad (4.3.145)$$

$$= \sigma^2 L^\top \Sigma L \quad (4.3.146)$$

$$= \sigma^2 I \quad (4.3.147)$$

Thus pre-multiplying the linear model by  $L^\top$  yields a modified linear model:

$$L^\top \mathbf{Y} = L^\top \mathbf{X}\beta + L^\top \varepsilon \quad (4.3.148)$$

in which the errors  $L^\top \varepsilon$  have covariance  $\sigma^2 I$ , and so the covariance computation of ordinary least squares can be applied to this. For convenience of notation denote this modified linear model by

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}} + \tilde{\varepsilon} \quad (4.3.149)$$

The generalised least squares estimator is then given by

$$\hat{\beta}_{\text{GLS}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}} \quad (4.3.150)$$

$$= (\mathbf{X}^\top LL^\top \mathbf{X})^{-1} \mathbf{X}^\top LL^\top \mathbf{Y} \quad (4.3.151)$$

$$= (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{Y} \quad (4.3.152)$$

And the covariance can be obtained from the ordinary least squares expression:

$$\text{Cov}(\hat{\beta}_{\text{GLS}}) = \sigma^2 (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \quad (4.3.153)$$

$$= \sigma^2 (\mathbf{X}^\top LL^\top \mathbf{X})^{-1} \quad (4.3.154)$$

$$= \sigma^2 (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \quad (4.3.155)$$

It can be shown that the generalised least squares estimator is the least squares estimator with the following cost function:

$$V(\beta) = \frac{1}{2} (\mathbf{Y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{Y} - \mathbf{X}\beta) \quad (4.3.156)$$

where if  $\Sigma^{-1}$  is set to a diagonal matrix of positive weights  $\mathbf{W}$ , reduces to weighted least squares. And if  $\Sigma^{-1}$  is set to a positive scaled identity matrix, it reduces to ordinary least squares.

### Estimated Generalised Least Squares

In practice the matrix  $\Sigma$  and/or  $\sigma^2$  may not be known, so we can instead replace these with estimates. If  $\Sigma$  is known and  $\sigma^2$  is not known, then  $\sigma^2$  can be estimated by  $\hat{\sigma}^2$  in the standard way from the OLS estimate of  $\mathbf{Y}$  on  $\mathbf{X}$ . If  $\Sigma$  is also not known, then  $\Sigma$  must be replaced by an estimate  $\hat{\Sigma}$  from the residuals of a ‘first-pass’ using OLS. To do this, typically some structure may be imposed on  $\Sigma$  depending on the type of model. This estimator is known as estimated generalised least squares (EGLS) or also feasible generalised least squares (FGLS).

#### 4.3.4 Total Least Squares

#### Orthogonal Regression

#### Deming Regression

#### Errors-in-Variables Regression

#### 4.3.5 Regularised Least Squares

#### Ridge Regression

In ridge regression (also known as Tikhonov regularisation or  $\ell_2$  regularisation), a regularisation term which is some multiple of the squared  $\ell_2$  norm of the parameters is added to the least squares cost function:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} V(\beta) \quad (4.3.157)$$

$$= \underset{\beta}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda \|\beta\|_2^2 \right\} \quad (4.3.158)$$

where  $\lambda \geq 0$  is a regularisation hyperparameter. Conceptually,  $\lambda$  trades off fitting the data (lower  $\lambda$  against keeping the parameters as ‘small’ as possible). Having higher  $\lambda$  is intended to prevent ‘overfitting’ where the estimated parameters may fit well to the data, but perform poorly when predicting new observations. To solve the least squares problem, we write the cost function as

$$V(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta \quad (4.3.159)$$

which has gradient

$$\nabla V(\beta) = -\mathbf{X}^T \cdot 2(\mathbf{Y} - \mathbf{X}\beta) + 2\lambda\beta \quad (4.3.160)$$

and solving for the gradient equal to zero:

$$0 = -\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\hat{\beta} + 2\lambda\hat{\beta} \quad (4.3.161)$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \mathbf{X}^T\mathbf{Y} \quad (4.3.162)$$

Another reason for using ridge regression is to better ‘condition’ the  $\mathbf{X}^T\mathbf{X}$  matrix. For instance, if the problem is overparametrised (i.e. the dimension of  $\beta$  is larger than the number of data points), then  $\mathbf{X}^T\mathbf{X}$  will not be invertible. But then the addition of  $\lambda I$  will add a constant to the diagonals of  $\mathbf{X}^T\mathbf{X}$ , so that it can be inverted. This allows us to still find solutions to overparametrised least squares problems.

### Bias of Ridge Regression

The ridge regression estimator with hyperparameter  $\lambda$ , denoted

$$\hat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \mathbf{X}^T\mathbf{Y} \quad (4.3.163)$$

can be shown to be **biased** for  $\beta$ . For the data-generating process, assume a general linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (4.3.164)$$

with fixed design matrix  $\mathbf{X}$ , and the errors are zero-mean  $\mathbb{E}[\varepsilon] = \mathbf{0}$  and covariance  $\text{Cov}(\varepsilon) = \sigma^2 I$ . It is usual to assume that  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ , but this is not strictly necessary. Then the expectation of the ridge regression estimator is

$$\mathbb{E}[\hat{\beta}_\lambda] = \mathbb{E} \left[ (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \mathbf{X}^T\mathbf{Y} \right] \quad (4.3.165)$$

$$= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbb{E}[\mathbf{X}\beta + \varepsilon] \quad (4.3.166)$$

$$= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{X}\beta \quad (4.3.167)$$

In OLS, this is the same as ridge regression with  $\lambda = 0$ , and so the expectation of the estimator would be equal to  $\beta$ , showing OLS is unbiased. With  $\lambda > 0$  however, the bias is

$$\mathbb{E}[\hat{\beta}_\lambda] - \beta = \left[ (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{X} - I \right] \beta \quad (4.3.168)$$

with is generally non-zero.

## Variance of Ridge Regression

Under the general linear model assumption, the covariance matrix of the ridge regression estimator  $\hat{\beta}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \mathbf{X}^\top \mathbf{Y}$  is

$$\text{Cov}(\hat{\beta}_\lambda) = \text{Cov}\left(\left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{Y}\right) \quad (4.3.169)$$

$$= \text{Cov}\left(\left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \varepsilon)\right) \quad (4.3.170)$$

$$= \text{Cov}\left(\left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \varepsilon\right) \quad (4.3.171)$$

$$= \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \text{Cov}(\varepsilon) \left[\left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top\right]^\top \quad (4.3.172)$$

Since  $\text{Cov}(\varepsilon) = \sigma^2 I$ , then

$$\text{Cov}(\hat{\beta}_\lambda) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \quad (4.3.173)$$

So if we take  $\lambda = 0$ , we get a covariance of  $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ , which is the covariance of the OLS estimator. It can be shown that the effect of regularisation in ridge regression is that the variance is lower than OLS, in an appropriate sense.

**Theorem 4.1.** *If  $\lambda > 0$  and the OLS estimator exists, then*

$$\text{Cov}(\hat{\beta}) - \text{Cov}(\hat{\beta}_\lambda) \succ \mathbf{0} \quad (4.3.174)$$

*Proof.* Since the OLS estimator exists, then implicitly  $\mathbf{X}^\top \mathbf{X}$  is full-rank and invertible, and also positive definite. Let

$$\Omega = \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \quad (4.3.175)$$

which is also invertible, with inverse

$$\Omega^{-1} = \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \quad (4.3.176)$$

Write the covariance of the ridge regression estimator as

$$\text{Cov}(\hat{\beta}_\lambda) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \quad (4.3.177)$$

$$= \sigma^2 \Omega^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \Omega \quad (4.3.178)$$

With this, expresses the difference between covariances as

$$\Xi := \text{Cov}(\hat{\beta}) - \text{Cov}(\hat{\beta}_\lambda) \quad (4.3.179)$$

$$= \sigma^2 \left[ \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} - \Omega^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \Omega \right] \quad (4.3.180)$$

$$= \sigma^2 \left[ \Omega^\top \left(\Omega^\top\right)^{-1} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \Omega^{-1} \Omega - \Omega^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \Omega \right] \quad (4.3.181)$$

$$= \sigma^2 \Omega^\top \left[ \left(\Omega^\top\right)^{-1} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \Omega^{-1} - \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \right] \Omega \quad (4.3.182)$$

Note that  $(\Omega^\top)^{-1} = (\Omega^{-1})^\top$ , so

$$(\Omega^\top)^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda I) \quad (4.3.183)$$

Substituting this, we have

$$\Xi = \sigma^2 \Omega^\top \left[ (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda I) (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda I) (\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \right] \Omega \quad (4.3.184)$$

$$= \sigma^2 \Omega^\top \left[ \left( I + \lambda (\mathbf{X}^\top \mathbf{X})^{-1} \right) (\mathbf{X}^\top \mathbf{X})^{-1} \left( I + \lambda (\mathbf{X}^\top \mathbf{X})^{-1} \right) - (\mathbf{X}^\top \mathbf{X})^{-1} \right] \Omega \quad (4.3.185)$$

Expanding the quadratic term:

$$\Xi = \sigma^2 \Omega^\top \left[ \left( (\mathbf{X}^\top \mathbf{X})^{-1} + \lambda (\mathbf{X}^\top \mathbf{X})^{-2} \right) \left( I + \lambda (\mathbf{X}^\top \mathbf{X})^{-1} \right) - (\mathbf{X}^\top \mathbf{X})^{-1} \right] \Omega \quad (4.3.186)$$

$$= \sigma^2 \Omega^\top \left[ (\mathbf{X}^\top \mathbf{X})^{-1} + \lambda (\mathbf{X}^\top \mathbf{X})^{-2} + \lambda (\mathbf{X}^\top \mathbf{X})^{-2} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-3} - (\mathbf{X}^\top \mathbf{X})^{-1} \right] \Omega \quad (4.3.187)$$

$$= \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \mathbf{X}^\top \mathbf{X} \left[ 2\lambda (\mathbf{X}^\top \mathbf{X})^{-2} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-3} \right] \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \quad (4.3.188)$$

$$= \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \left[ 2\lambda I + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-1} \right] (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \quad (4.3.189)$$

We now show that  $\Xi$  is positive definite by showing that  $v^\top \Xi v > 0$  for any vector  $v \neq \mathbf{0}$ . Since  $(\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1}$  is invertible, then also  $z := (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} v \neq 0$  for any non-zero  $v$ . Therefore

$$v^\top \Xi v = \sigma^2 \left[ (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} v \right]^\top \left[ 2\lambda I + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-1} \right] (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} v \quad (4.3.190)$$

$$= \sigma^2 z^\top \left[ 2\lambda I + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-1} \right] z \quad (4.3.191)$$

$$= \sigma^2 \lambda z^\top z + \sigma^2 \lambda^2 z^\top (\mathbf{X}^\top \mathbf{X})^{-1} z \quad (4.3.192)$$

$$> 0 \quad (4.3.193)$$

since  $\sigma^2 > 0$  (implicitly),  $\lambda > 0$ , and because  $(\mathbf{X}^\top \mathbf{X})^{-1}$  is positive definite.  $\square$

Also, the scalarised variance (equal to the trace of the covariance) will also be smaller for the ridge regression estimator:

$$\text{trace} \left( \text{Cov} \left( \hat{\beta} \right) \right) - \text{trace} \left( \text{Cov} \left( \hat{\beta}_\lambda \right) \right) = \text{trace} (\Xi) \quad (4.3.194)$$

$$> 0 \quad (4.3.195)$$

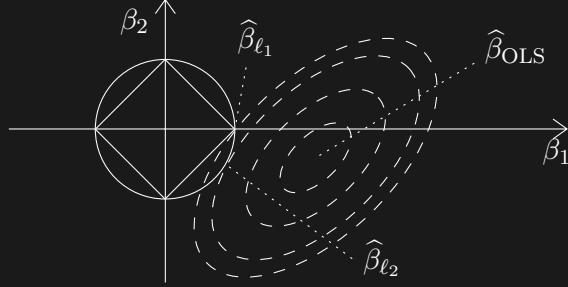
since the trace of a positive definite matrix is positive.

## Lasso Regression

In lasso regression (also known as  $\ell_1$  regularisation), a multiple of the  $\ell_1$  norm of the parameters  $\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$  is added to the least squares cost function, giving the estimator:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) + \lambda \|\beta\|_1 \right\} \quad (4.3.196)$$

with regularisation hyperparameter  $\lambda \geq 0$ . Unlike ridge regression, there is no closed-form for the estimator, however a solution can be obtained via numerical iterative optimisation algorithms, since the problem is convex. Like with ridge regression, lasso regression may be used when there are more parameters than observations. A key feature of lasso regression is that it promotes *sparsity* of the solution. That is, the estimate will have relatively few non-zero coefficients. To see why this might be true, we can visualise a case in two dimensions.



We can equivalently treat the solution to a regularised least squares problem as a constrained least squares problem. Recall that the  $\ell_1$  norm ball is diamond-shaped, so the constraint set is also diamond-shaped. The unregularised least squares problem (which has elliptical level sets) has a solution  $\hat{\beta}_{OLS}$  with non-zero  $\beta_1$  and  $\beta_2$ . The ridge regression solution  $\hat{\beta}_{\ell_2}$  also has non-zero coefficients, whereas the lasso solution  $\hat{\beta}_{\ell_1}$  is more sparse because the second coefficient is zero.

#### 4.3.6 Constrained Least Squares

##### Non-Negative Least Squares

Non-negative least squares solves the linear least squares problem under the constraint that all the coefficients are non-negative, i.e.

$$\begin{aligned} \min_{\beta} \quad & (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) \\ \text{s.t.} \quad & \beta \geq \mathbf{0} \end{aligned} \tag{4.3.197}$$

This is a convex optimisation problem, so iterative methods can be used to find the solution. Non-negative least squares may be applied as a parametric form of *isotonic regression*.

##### Equivalence Between Constrained and Regularised Least Squares

Consider the constrained least squares problem

$$\begin{aligned} \min_{\beta} \quad & \sum_{i=1}^n (y_i - \beta^\top x_i)^2 \\ \text{s.t.} \quad & \sum_{j=1}^d |\beta_j|^q \leq \eta \end{aligned} \tag{4.3.198}$$

with  $q \geq 1$ . Denote the inequality constraint function

$$c(\beta) = \sum_{j=1}^d |\beta_j|^q - \eta \tag{4.3.199}$$

$$\leq 0 \tag{4.3.200}$$

Using the method of Lagrange multipliers, we can write the Lagrangian function as

$$\mathcal{L}(\beta, \lambda) = \sum_{i=1}^n (y_i - \beta^\top x_i)^2 + \lambda c(\beta) \quad (4.3.201)$$

$$= \sum_{i=1}^n (y_i - \beta^\top x_i)^2 + \lambda \left( \sum_{j=1}^d |\beta_j|^q - \eta \right) \quad (4.3.202)$$

where  $\lambda > 0$  is a Lagrange multiplier. The Lagrange dual function is defined as

$$\mathcal{L}^*(\lambda) = \inf_{\beta} \mathcal{L}(\beta, \lambda) \quad (4.3.203)$$

$$= \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta^\top x_i)^2 + \lambda \left( \sum_{j=1}^d |\beta_j|^q - \eta \right) \right\} \quad (4.3.204)$$

Note the minimiser of this with respect to  $\beta$  is actually

$$\operatorname{argmin}_{\beta} \mathcal{L}(\beta, \lambda) = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta^\top x_i)^2 + \lambda \sum_{j=1}^d |\beta_j|^q \right\} \quad (4.3.205)$$

because the  $\lambda\eta$  term does not depend on  $\beta$ . This looks like a **regularised least squares problem**, where  $q = 1$  is the case of lasso regression and  $q = 2$  is the case of ridge regression. Now suppose for some choice of  $\lambda > 0$ , we solved the regularised least squares problem

$$\hat{\beta}_{\ell_q} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta^\top x_i)^2 + \lambda \sum_{j=1}^d |\beta_j|^q \right\} \quad (4.3.206)$$

which has a minimum given by the Lagrangian dual  $\mathcal{L}^*(\lambda)$ . Because of the complementary slackness condition,  $\lambda > 0$  implies  $c(\hat{\beta}_{\ell_q}) = 0$ , which corresponds to having solved a constrained least squares problem with

$$\eta = \sum_{j=1}^d |\hat{\beta}_{\ell_q, j}|^q \quad (4.3.207)$$

By duality, if we solve a constrained least squares problem with the constraint active (i.e.  $c(\hat{\beta}) = 0$ ), this corresponds to having solved a regularised least squares problem with  $\lambda > 0$ .

Alternately, if the constraint is inactive at the solution (i.e.  $c(\hat{\beta}) < 0$ ), this implies  $\lambda = 0$  by complementary slackness. This demonstrates (on both sides) that the constrained solution is then equal to the unconstrained solution.

### 4.3.7 Recursive Least Squares

Suppose we have  $n$  observations, which we denote with subscripts in the matrices  $\mathbf{X}_n$  and  $\mathbf{Y}_n$ . The ordinary least squares estimate indexed by  $n$  is given by

$$\hat{\beta}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n \quad (4.3.208)$$

Consider a new datum  $(\mathbf{x}_{n+1}, y_{n+1})$ . The new least squares estimate can be computed from the old estimate follows. First note that to obtain the new data matrices from the old matrices we can concatenate them by

$$\mathbf{X}_{n+1} = \begin{bmatrix} \mathbf{X}_n \\ \mathbf{x}_{n+1}^\top \end{bmatrix} \quad (4.3.209)$$

and

$$\mathbf{Y}_{n+1} = \begin{bmatrix} \mathbf{Y}_n \\ y_{n+1} \end{bmatrix} \quad (4.3.210)$$

Hence

$$\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1} = [\mathbf{X}_n^\top \ \mathbf{x}_{n+1}] \begin{bmatrix} \mathbf{X}_n \\ \mathbf{x}_{n+1}^\top \end{bmatrix} \quad (4.3.211)$$

$$= \mathbf{X}_n^\top \mathbf{X}_n + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \quad (4.3.212)$$

and

$$\mathbf{X}_{n+1}^\top \mathbf{Y}_{n+1} = [\mathbf{X}_n^\top \ \mathbf{x}_{n+1}] \begin{bmatrix} \mathbf{Y}_n \\ y_{n+1} \end{bmatrix} \quad (4.3.213)$$

$$= \mathbf{X}_n^\top \mathbf{Y}_n + \mathbf{x}_{n+1} y_{n+1} \quad (4.3.214)$$

So from the least squares estimate with  $n + 1$  observations, we can write

$$\hat{\beta}_{n+1} = (\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1})^{-1} \mathbf{X}_{n+1}^\top \mathbf{Y}_{n+1} \quad (4.3.215)$$

$$(\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1}) \hat{\beta}_{n+1} = \mathbf{X}_n^\top \mathbf{Y}_n + \mathbf{x}_{n+1} y_{n+1} \quad (4.3.216)$$

$$(\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1}) \hat{\beta}_{n+1} = (\mathbf{X}_n^\top \mathbf{X}_n) \hat{\beta}_n + \mathbf{x}_{n+1} y_{n+1} \quad (4.3.217)$$

$$(\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1}) \hat{\beta}_{n+1} = (\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1} - \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top) \hat{\beta}_n + \mathbf{x}_{n+1} y_{n+1} \quad (4.3.218)$$

$$\hat{\beta}_{n+1} = (\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1})^{-1} [(\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1} - \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top) \hat{\beta}_n + \mathbf{x}_{n+1} y_{n+1}] \quad (4.3.219)$$

which then becomes

$$\hat{\beta}_{n+1} = \hat{\beta}_n + (\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1})^{-1} (\mathbf{x}_{n+1} y_{n+1} - \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \hat{\beta}_n) \quad (4.3.220)$$

$$= \hat{\beta}_n + (\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1})^{-1} \mathbf{x}_{n+1} (y_{n+1} - \mathbf{x}_{n+1}^\top \hat{\beta}_n) \quad (4.3.221)$$

This gives a formula for recursive update of  $\hat{\beta}$ . Letting  $\mathbf{K}_{n+1} = (\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1})^{-1} \mathbf{x}_{n+1}$  and  $e_{n+1} = y_{n+1} - \mathbf{x}_{n+1}^\top \hat{\beta}_n$ , we have

$$\hat{\beta}_{n+1} = \hat{\beta}_n + \mathbf{K}_{n+1} e_{n+1} \quad (4.3.222)$$

Here,  $e_{n+1}$  may be interpreted as an error term for the new datum, and  $\mathbf{K}_{n+1}$  may be interpreted as a gain matrix on the error to correct the current estimate. An alternative form of the recursive update may be derived which does not require the inversion of  $\mathbf{X}_{n+1}^\top \mathbf{X}_{n+1}$  (which may be more computationally expensive the larger  $n$  becomes). First, it is convenient to denote  $\mathbf{R}_n = \mathbf{X}_n^\top \mathbf{X}_n$ , so that

$$\mathbf{R}_{n+1}^{-1} = (\mathbf{R}_n + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top)^{-1} \quad (4.3.223)$$

By the matrix inversion lemma  $(B^{-1} + C^\top A^{-1} C)^{-1} = B - BC^\top (A + CBC^\top)^{-1} CB$  and letting  $B = \mathbf{R}_n^{-1}$ ,  $C = \mathbf{x}_{n+1}^\top$  and  $A = I$ , we have

$$\mathbf{R}_{n+1}^{-1} = \mathbf{R}_n^{-1} - \frac{\mathbf{R}_n^{-1} \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1}}{1 + \mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1} \mathbf{x}_{n+1}} \quad (4.3.224)$$

so  $\mathbf{R}_{n+1}^{-1}$  can be computed recursively from  $\mathbf{R}_n^{-1}$ . For convenience, introduce  $\mathbf{P}_n = \mathbf{R}_n^{-1}$ . Then gain matrix is given by

$$\mathbf{K}_{n+1} = \mathbf{P}_{n+1} \mathbf{x}_{n+1} \quad (4.3.225)$$

$$= \left( \mathbf{P}_n - \frac{\mathbf{P}_n \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{P}_n}{1 + \mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}} \right) \mathbf{x}_{n+1} \quad (4.3.226)$$

$$= \mathbf{P}_n \mathbf{x}_{n+1} - \frac{\mathbf{P}_n \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}}{1 + \mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}} \quad (4.3.227)$$

$$= \mathbf{P}_n \mathbf{x}_{n+1} \left( 1 - \frac{\mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}}{1 + \mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}} \right) \quad (4.3.228)$$

$$= \frac{\mathbf{P}_n \mathbf{x}_{n+1}}{1 + \mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}} \quad (4.3.229)$$

where we can also see the recursive formula for  $\mathbf{P}_{n+1}$ :

$$\mathbf{P}_{n+1} = \mathbf{P}_n - \frac{\mathbf{P}_n \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{P}_n}{1 + \mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}} \quad (4.3.230)$$

#### 4.3.8 Partial Least Squares [175]

#### 4.3.9 Stochastic Least Squares [206]

#### 4.3.10 Multiple Output Least Squares [80]

We now consider least squares in a setting where there are multiple outputs (i.e. response variables). Suppose there are  $K$  response variables. We write the general linear model as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (4.3.231)$$

where  $\mathbf{X}$  is a  $n \times p$  data matrix as before, while  $\mathbf{Y}$  is a  $n \times K$  matrix structured as

$$\mathbf{Y} = [\mathbf{y}_1 \ \dots \ \mathbf{y}_K] \quad (4.3.232)$$

$$= \begin{bmatrix} y_{11} & \dots & y_{1K} \\ \vdots & \ddots & \vdots \\ y_{n1} & \dots & y_{nK} \end{bmatrix} \quad (4.3.233)$$

and  $\mathbf{B}$  is a  $p \times K$  matrix structured as

$$\mathbf{B} = [\beta_1 \ \dots \ \beta_K] \quad (4.3.234)$$

Lastly,  $\mathbf{E}$  is a  $n \times K$  matrix of errors. A suitable least squares criterion to minimise is the ‘overall’ sum of squares

$$V(\mathbf{B}) = \sum_{k=1}^K \sum_{i=1}^n \left( y_{ik} - \beta_k^\top x_i \right)^2 \quad (4.3.235)$$

Noticing that the  $k^{\text{th}}$  diagonal element of  $(\mathbf{Y} - \mathbf{X}\mathbf{B})^\top (\mathbf{Y} - \mathbf{X}\mathbf{B})$  is  $(\mathbf{y}_k - \mathbf{X}\beta_k)^\top (\mathbf{y}_k - \mathbf{X}\beta_k)$  which is the univariate criterion applied to the  $k^{\text{th}}$  response variable, we can write the overall criterion as

$$V(\mathbf{B}) = \text{trace} \left( (\mathbf{Y} - \mathbf{X}\mathbf{B})^\top (\mathbf{Y} - \mathbf{X}\mathbf{B}) \right) \quad (4.3.236)$$

We can differentiate this criterion using properties of the trace and the chain rule. Firstly,

$$\frac{\partial}{\partial \mathbf{B}} \text{trace}(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \frac{\partial}{\partial \mathbf{B}} \text{trace}(\mathbf{Y}) - \frac{\partial}{\partial \mathbf{B}} \text{trace}(\mathbf{X}\mathbf{B}) \quad (4.3.237)$$

$$= -\mathbf{X}^\top \quad (4.3.238)$$

Next, we use the fact that

$$\frac{\partial}{\partial U} \text{trace} (UAU^\top C) = \frac{\partial}{\partial U} \text{trace} (U^\top CUA) \quad (4.3.239)$$

$$= CUA + C^\top UA^\top \quad (4.3.240)$$

Letting  $U = \mathbf{Y} - \mathbf{X}\mathbf{B}$  and applying this with  $A = I$  and  $C = I$ , we get

$$\frac{\partial V(\mathbf{B})}{\partial \mathbf{B}} = \frac{\partial U}{\partial \mathbf{B}} \cdot \frac{\partial V}{\partial U} \quad (4.3.241)$$

$$= -\mathbf{X}^\top \cdot 2(\mathbf{Y} - \mathbf{X}\mathbf{B}) \quad (4.3.242)$$

Thus the least squares estimator  $\hat{\mathbf{B}}$  satisfies

$$\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = 0 \quad (4.3.243)$$

$$\mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}} \quad (4.3.244)$$

$$\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \quad (4.3.245)$$

which takes the same form of the estimator for the univariate response case. Notably, we can write this out as

$$[\hat{\beta}_1 \ \dots \ \hat{\beta}_K] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top [\mathbf{y}_1 \ \dots \ \mathbf{y}_K] \quad (4.3.246)$$

which shows that the  $k^{\text{th}}$  response coefficients are just the same as regressing the  $k^{\text{th}}$  response variables  $\mathbf{y}_k$  on  $\mathbf{X}$ . In other words, having multiple outputs does not affect one another's least squares estimates. Hence we can perform multiple output least squares by running  $K$  separate regressions.

#### 4.3.11 Nonlinear Least Squares

Suppose we have  $n$  data pairs  $(x_i, y_i)$  and would like to fit a nonlinear model  $f(x_i, \theta)$  with respect to a  $d$ -dimensional vector of parameters  $\theta$ . Then define the residuals

$$r_i(\theta) = y_i - f(x_i, \theta) \quad (4.3.247)$$

and the residual vector

$$\mathbf{r}(\theta) = [r_1(\theta) \ \dots \ r_n(\theta)]^\top \quad (4.3.248)$$

Then the nonlinear least squares problem is to minimise (half) the sum of squared residuals  $V(\theta)$ :

$$\theta^* = \underset{\theta}{\operatorname{argmin}} V(\theta) \quad (4.3.249)$$

$$= \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^n r_i(\theta)^2 \quad (4.3.250)$$

$$= \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{r}(\theta)\|^2 \quad (4.3.251)$$

#### Gauss-Newton Algorithm

The Gauss-Newton algorithm is a method for solving nonlinear least squares problems. It is based on the Newton method of finding local optima, however it does not require computation

of the Hessian (which can be unwieldy for particular nonlinear functions). Define the Jacobian matrix as

$$\mathbf{J}(\theta) = \frac{\partial \mathbf{r}(\theta)}{\partial \theta} \quad (4.3.252)$$

$$= \begin{bmatrix} \partial r_1 / \partial \theta_1 & \dots & \partial r_1 / \partial \theta_d \\ \vdots & \ddots & \vdots \\ \partial r_n / \partial \theta_1 & \dots & \partial r_n / \partial \theta_d \end{bmatrix} \quad (4.3.253)$$

$$= \begin{bmatrix} \nabla r_1(\theta)^\top \\ \vdots \\ \nabla r_n(\theta)^\top \end{bmatrix} \quad (4.3.254)$$

Taking the derivative of the nonlinear least squares cost function, we have by the chain rule

$$\nabla V(\theta) = \frac{1}{2} \sum_{i=1}^n 2r_i(\theta) \nabla r_i(\theta) \quad (4.3.255)$$

$$= [\nabla r_1(\theta) \dots \nabla r_n(\theta)]^\top \begin{bmatrix} r_1(\theta) \\ \vdots \\ r_n(\theta) \end{bmatrix} \quad (4.3.256)$$

$$= \mathbf{J}(\theta)^\top \mathbf{r}(\theta) \quad (4.3.257)$$

We also obtain the Hessian of the nonlinear least squares cost function by the product rule

$$\nabla^2 V(\theta) = \sum_{i=1}^n \nabla r_i(\theta) \nabla r_i(\theta)^\top + \sum_{i=1}^n r_i(\theta) \nabla^2 r_i(\theta) \quad (4.3.258)$$

$$= [\nabla r_1(\theta) \dots \nabla r_n(\theta)]^\top \begin{bmatrix} \nabla r_1(\theta)^\top \\ \vdots \\ \nabla r_n(\theta)^\top \end{bmatrix} + \sum_{i=1}^n r_i(\theta) \nabla^2 r_i(\theta) \quad (4.3.259)$$

$$= \mathbf{J}(\theta)^\top \mathbf{J}(\theta) + \sum_{i=1}^n r_i(\theta) \nabla^2 r_i(\theta) \quad (4.3.260)$$

In the Gauss-Newton algorithm, the approximation is made on the Hessian

$$\nabla^2 V(\theta) \approx \mathbf{J}(\theta)^\top \mathbf{J}(\theta) \quad (4.3.261)$$

This approximation is reasonable if the residuals  $r_i(\theta)$  are small, which are what we are trying to make small with the objective of least squares. Using this approximation to an update step identical to the Newton algorithm, the parameter update  $\theta_{k+1}$  after  $k$  iterations is given by

$$\theta_{k+1} = \theta_k - (\mathbf{J}(\theta_k)^\top \mathbf{J}(\theta_k))^{-1} \mathbf{J}(\theta_k)^\top \mathbf{r}(\theta_k) \quad (4.3.262)$$

### Levenberg-Marquardt Algorithm

Also known as damped least squares, the Levenberg-Marquardt algorithm finds a middle ground between gradient descent and the Gauss-Newton algorithm. The update step is given by

$$\theta_{k+1} = \theta_k - (\mathbf{J}(\theta_k)^\top \mathbf{J}(\theta_k) + \lambda I)^{-1} \mathbf{J}(\theta_k)^\top \mathbf{r}(\theta_k) \quad (4.3.263)$$

where  $\lambda$  is known as the damping parameter. Note that if  $\lambda$  is small, the update will be closer to the Levenberg-Marquardt algorithm. If  $\lambda$  is relatively large, then the update direction will

be closer to the (steepest) gradient descent direction. The choice of parameter  $\lambda$  can be made adaptive, so that it starts off larger (where the Hessian approximation made by the Levenberg-Marquardt algorithm may not be as valid), and reduces over the number of iterations. An alternative update is given by

$$\theta_{k+1} = \theta_k - \left( \mathbf{J}(\theta_k)^\top \mathbf{J}(\theta_k) + \lambda \text{diag} \left\{ \mathbf{J}(\theta_k)^\top \mathbf{J}(\theta_k) \right\} \right)^{-1} \mathbf{J}(\theta_k)^\top \mathbf{r}(\theta_k) \quad (4.3.264)$$

where  $\text{diag} \left\{ \mathbf{J}(\theta_k)^\top \mathbf{J}(\theta_k) \right\}$  is a diagonal matrix containing only the main diagonal of  $\mathbf{J}(\theta_k)^\top \mathbf{J}(\theta_k)$ . This version of the update makes it so that the damping term is of the same scale as  $\mathbf{J}(\theta_k)^\top \mathbf{J}(\theta_k)$ .

## 4.4 Estimation Theory

### 4.4.1 Asymptotic Consistency

A sequence of estimators  $\hat{\theta}_n$  for a population parameter  $\theta$  indexed by the sample size  $n$  is said to be (asymptotically) consistent for  $\theta$  if  $\hat{\theta}_n$  converges in probability to  $\theta$ .

#### Asymptotic Unbiasedness

A sequence of estimators  $\hat{\theta}_n$  for a population parameter  $\theta$  indexed by the sample size  $n$  is said to be asymptotically unbiased for  $\theta$  if  $\hat{\theta}_n$  is unbiased in the limit, i.e.

$$\lim_{n \rightarrow \infty} \mathbb{E} [\hat{\theta}_n] = \theta \quad (4.4.1)$$

Sufficient conditions for consistency are asymptotic unbiasedness as well as the variance of the estimator decreasing to zero:

$$\lim_{n \rightarrow \infty} \text{Var} (\hat{\theta}_n) = 0 \quad (4.4.2)$$

This is because using Chebychev's inequality, we have:

$$\lim_{n \rightarrow \infty} \Pr \left( \left| \hat{\theta}_n - \mathbb{E} [\hat{\theta}_n] \right| > \varepsilon \right) \leq \frac{\lim_{n \rightarrow \infty} \text{Var} (\hat{\theta}_n)}{\varepsilon^2} \quad (4.4.3)$$

which with the two sufficient conditions, gives convergence in probability:

$$\lim_{n \rightarrow \infty} \Pr \left( \left| \hat{\theta}_n - \theta \right| > \varepsilon \right) = 0 \quad (4.4.4)$$

#### Consistency Does Not Necessarily Imply Asymptotic Unbiasedness

Consider the sequence of estimators defined by

$$\Pr \left( \hat{\theta}_n = y \right) = \begin{cases} \frac{n-1}{n}, & y = 0 \\ \frac{1}{n}, & y = n \end{cases} \quad (4.4.5)$$

Thus  $\hat{\theta}_n$  is consistent for 0 because for all  $\varepsilon > 0$ :

$$\lim_{n \rightarrow \infty} \Pr \left( \left| \hat{\theta}_n \right| > \varepsilon \right) = \lim_{n \rightarrow \infty} \Pr \left( \hat{\theta}_n = n \right) \quad (4.4.6)$$

$$= 0 \quad (4.4.7)$$

However, we compute  $\mathbb{E}[\hat{\theta}_n] = 1$  for all  $n$ . Therefore  $\hat{\theta}_n$  is consistent for 0, but not asymptotically unbiased for 0. The required condition in order for consistency to imply asymptotic unbiasedness is a uniform bound on the variance, that is

$$\text{Var}(\hat{\theta}_n - \theta) \leq C \quad (4.4.8)$$

for all  $n$ . The counter-example above does not satisfy this because we can show that  $\text{Var}(\hat{\theta}_n) = n - 1$ . On the other hand, if we are given consistency and have a uniform bound on the variance, then asymptotic unbiasedness can be shown in the following way. By an indicator variable expansion:

$$\mathbb{E}[|\hat{\theta}_n - \theta|] = \mathbb{E}[|\hat{\theta}_n - \theta| (\mathbb{I}_{\{|\hat{\theta}_n - \theta| \leq \varepsilon\}} + \mathbb{I}_{\{|\hat{\theta}_n - \theta| > \varepsilon\}})] \quad (4.4.9)$$

$$= \mathbb{E}[|\hat{\theta}_n - \theta| \mathbb{I}_{\{|\hat{\theta}_n - \theta| \leq \varepsilon\}}] + \mathbb{E}[|\hat{\theta}_n - \theta| \mathbb{I}_{\{|\hat{\theta}_n - \theta| > \varepsilon\}}] \quad (4.4.10)$$

Note that since  $|\hat{\theta}_n - \theta| \mathbb{I}_{\{|\hat{\theta}_n - \theta| \leq \varepsilon\}} \leq \varepsilon$  always holds via the indicator, we have

$$\mathbb{E}[(\hat{\theta}_n - \theta) \mathbb{I}_{\{|\hat{\theta}_n - \theta| \leq \varepsilon\}}] \leq \varepsilon \quad (4.4.11)$$

As for the other term, using the Cauchy-Schwarz inequality yields

$$\mathbb{E}[|\hat{\theta}_n - \theta| \mathbb{I}_{\{|\hat{\theta}_n - \theta| > \varepsilon\}}] \leq \sqrt{\mathbb{E}[|\hat{\theta}_n - \theta|^2] \mathbb{E}[\mathbb{I}_{\{|\hat{\theta}_n - \theta| > \varepsilon\}}^2]} \quad (4.4.12)$$

$$= \sqrt{\mathbb{E}[|\hat{\theta}_n - \theta|^2] \mathbb{E}[\mathbb{I}_{\{|\hat{\theta}_n - \theta| > \varepsilon\}}]} \quad (4.4.13)$$

$$= \sqrt{\mathbb{E}[|\hat{\theta}_n - \theta|^2] \Pr(|\hat{\theta}_n - \theta| > \varepsilon)} \quad (4.4.14)$$

The bounded variance condition ensures that  $\mathbb{E}[|\hat{\theta}_n - \theta|^2]$  stays bounded, hence due to  $\hat{\theta}_n$  being consistent (meaning  $\lim_{n \rightarrow \infty} \Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0$  for any  $\varepsilon > 0$ ), we see that

$$\lim_{n \rightarrow \infty} \sqrt{\mathbb{E}[|\hat{\theta}_n - \theta|^2] \Pr(|\hat{\theta}_n - \theta| > \varepsilon)} = 0 \quad (4.4.15)$$

Thus

$$\lim_{n \rightarrow \infty} \mathbb{E}[|\hat{\theta}_n - \theta|] \leq \varepsilon \quad (4.4.16)$$

Then as  $\varepsilon$  can be made arbitrarily small, this implies asymptotic unbiasedness  $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\theta}_n] = \theta$ .

#### 4.4.2 Weak Law of Large Numbers

Let  $X_1, \dots, X_n$  be a sequence of independent and identically distributed random variables, with mean  $\mathbb{E}[X_i] = \mu$  and standard deviation  $\sigma$ . Then the sample mean  $\bar{X}$  is defined as

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \quad (4.4.17)$$

The Weak Law of Large Numbers states that  $\bar{X} \xrightarrow{\text{P}} \mu$  as  $n \rightarrow \infty$ . That is, the sample mean converges in probability to the population mean as the sample size increases to infinity. We can also write this as

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X} - \mu| \geq \varepsilon) = 0 \quad (4.4.18)$$

for all  $\varepsilon > 0$ .

*Proof.* From Chebychev's inequality

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\text{Var}(\bar{X})}{\varepsilon^2} \quad (4.4.19)$$

Using  $\text{Var}(\bar{X}) = \sigma^2/n$

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \quad (4.4.20)$$

Taking the limit

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X} - \mu| \geq \varepsilon) = 0 \quad (4.4.21)$$

□

### 4.4.3 Strong Law of Large Numbers

The Strong Law of Large Numbers is a stronger version of the Weak Law of Large Numbers which says that the sample mean converges almost surely to the population mean. That is,

$$\Pr\left(\lim_{n \rightarrow \infty} \bar{X} = \mu\right) = 1 \quad (4.4.22)$$

**Theorem 4.2** ([23], [59]). *Let  $\{X_n\}$  be a sequence of i.i.d. copies of  $X$ . Assume that  $\mathbb{E}[|X|] < \infty$ . Then*

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} \mathbb{E}[X] \quad (4.4.23)$$

*Proof.* If we denote  $X^+ = \max\{X, 0\}$  and  $X^- = \max\{-X, 0\}$  then  $X = X^+ - X^-$ . Then without loss of generality assume  $X \geq 0$ , otherwise we can just apply the theorem to  $X^+$  and  $X^-$  separately. Let  $Y_i = X_i \mathbb{I}_{X_i \leq i}$  (i.e.  $Y_i$  is constructed to be a random variable such that the higher  $i$  is, the more likely it is to be equal to  $X_i$ ). Then from the characterisation of this indicator function:

$$\sum_{i=1}^{\infty} \Pr(X_i \neq Y_i) = \sum_{i=1}^{\infty} \Pr(X_i \neq X_i \mathbb{I}_{X_i \leq i}) \quad (4.4.24)$$

$$= \sum_{i=1}^{\infty} \Pr(X_i > i) \quad (4.4.25)$$

$$= \sum_{i=1}^{\infty} \Pr(X > i) \quad (4.4.26)$$

$$\leq \int_0^{\infty} \Pr(X > x) dx \quad (4.4.27)$$

$$= \int_0^{\infty} \mathbb{E}[\mathbb{I}_{X>x}] dx \quad (4.4.28)$$

$$= \mathbb{E}\left[\int_0^{\infty} \mathbb{I}_{X>x} dx\right] \quad (4.4.29)$$

$$= \mathbb{E}\left[\int_0^X 1 dx\right] \quad (4.4.30)$$

$$= \mathbb{E}[X] \quad (4.4.31)$$

$$< \infty \quad (4.4.32)$$

where the inequality between the sum and the integral follows since the sum can be thought of as a left-rectangle approximation of the integral using rectangle widths of 1, where the integrand is weakly decreasing in  $x$  (hence the sum lower bounds the integral). Now because  $\sum_{i=1}^{\infty} \Pr(X_i \neq Y_i) < \infty$ , by the Borel-Cantelli lemma the event  $\{X_i \neq Y_i\}$  occurs finitely many times with probability 1. Therefore it suffices to show that

$$\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{\text{a.s.}} \mathbb{E}[X] \quad (4.4.33)$$

because

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n X_i \mathbb{I}_{X_i \leq i} \quad (4.4.34)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (X_i - X_i \mathbb{I}_{X_i > i}) \quad (4.4.35)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i - \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n X_i \mathbb{I}_{X_i > i}}{n} \quad (4.4.36)$$

and as the numerator of the second term will be finite, the limit will be zero. Therefore

$$\frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n X_i \quad (4.4.37)$$

almost surely. Introduce  $Z_n = \frac{1}{n} \sum_{i=1}^n Y_i$  and observe that

$$\text{Var}(Z_n) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) \quad (4.4.38)$$

$$= \frac{1}{n^2} \sum_{i=1}^n (\mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2) \quad (4.4.39)$$

$$\leq \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[Y_i^2] \quad (4.4.40)$$

$$= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[X_i^2 \mathbb{I}_{X_i \leq i}] \quad (4.4.41)$$

$$= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[X^2 \mathbb{I}_{X_i \leq i}] \quad (4.4.42)$$

$$= \frac{1}{n^2} \cdot n \mathbb{E}[X^2 \mathbb{I}_{X_i \leq n}] \quad (4.4.43)$$

$$= \frac{\mathbb{E}[X^2 \mathbb{I}_{X_i \leq n}]}{n} \quad (4.4.44)$$

For any  $\alpha > 1$  and  $\varepsilon > 0$  and noting that  $\lfloor \alpha^n \rfloor$  will be a natural number, we have using Chebychev's inequality:

$$\sum_{n=1}^{\infty} \Pr(|Z_{\lfloor \alpha^n \rfloor} - \mathbb{E}[Z_{\lfloor \alpha^n \rfloor}]| > \varepsilon) \leq \sum_{n=1}^{\infty} \frac{\text{Var}(Z_{\lfloor \alpha^n \rfloor})}{\varepsilon^2} \quad (4.4.45)$$

$$\leq \sum_{n=1}^{\infty} \frac{\mathbb{E}[X^2 \mathbb{I}_{X_i \leq \lfloor \alpha^n \rfloor}]}{\varepsilon^2 \lfloor \alpha^n \rfloor} \quad (4.4.46)$$

$$= \frac{1}{\varepsilon^2} \mathbb{E}\left[X^2 \sum_{n=1}^{\infty} \frac{\mathbb{I}_{X_i \leq \lfloor \alpha^n \rfloor}}{\lfloor \alpha^n \rfloor}\right] \quad (4.4.47)$$

where the second inequality uses the preceding result. Let  $K$  be a random variable for the smallest natural number such that  $X \leq \lfloor \alpha^K \rfloor$ . Then use the fact  $\lfloor \alpha^n \rfloor > \frac{\alpha^n}{2}$  for all  $n \geq 1$  to obtain:

$$\sum_{n=1}^{\infty} \frac{\mathbb{I}_{X_i \leq \lfloor \alpha^n \rfloor}}{\lfloor \alpha^n \rfloor} \leq 2 \sum_{n=1}^{\infty} \frac{\mathbb{I}_{X_i \leq \lfloor \alpha^n \rfloor}}{\alpha^n} \quad (4.4.48)$$

$$= 2 \sum_{n=K}^{\infty} \frac{1}{\alpha^n} \quad (4.4.49)$$

$$= \frac{2}{\alpha^K} \sum_{n=0}^{\infty} \frac{1}{\alpha^n} \quad (4.4.50)$$

$$\leq \frac{2}{X} \sum_{n=0}^{\infty} \left(\frac{1}{\alpha}\right)^n \quad (4.4.51)$$

$$= \frac{2}{X} \cdot \frac{1}{1 - 1/\alpha} \quad (4.4.52)$$

$$= \frac{2\alpha}{(\alpha - 1)X} \quad (4.4.53)$$

Hence multiplying both sides by  $X^2$  and taking expectations:

$$\mathbb{E}\left[X^2 \sum_{n=1}^{\infty} \frac{\mathbb{I}_{X_i \leq \lfloor \alpha^n \rfloor}}{\lfloor \alpha^n \rfloor}\right] = \mathbb{E}\left[\frac{2X^2\alpha}{(\alpha - 1)X}\right] \quad (4.4.54)$$

$$\leq \frac{2\alpha}{\alpha - 1} \mathbb{E}[X] \quad (4.4.55)$$

$$< \infty \quad (4.4.56)$$

From using the Borel-Cantelli lemma with arbitrarily small  $\varepsilon$ , this implies that  $Z_{\lfloor \alpha^n \rfloor} \neq \mathbb{E}[Z_{\lfloor \alpha^n \rfloor}]$  finitely many times with probability 1, hence

$$\Pr\left(\lim_{n \rightarrow \infty} Z_{\lfloor \alpha^n \rfloor} = \lim_{n \rightarrow \infty} \mathbb{E}[Z_{\lfloor \alpha^n \rfloor}]\right) = 1 \quad (4.4.57)$$

or equivalently,  $Z_{\lfloor \alpha^n \rfloor} \xrightarrow{\text{a.s.}} \lim_{n \rightarrow \infty} \mathbb{E}[Z_{\lfloor \alpha^n \rfloor}]$ . Also since  $\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{\text{a.s.}} \frac{1}{n} \sum_{i=1}^n X_i$ , then by the Dominated Convergence Theorem

$$\lim_{n \rightarrow \infty} \mathbb{E}[Z_n] = \mathbb{E}\left[\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Y_i\right] \quad (4.4.58)$$

$$= \mathbb{E}\left[\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i\right] \quad (4.4.59)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \quad (4.4.60)$$

$$= \mathbb{E}[X] \quad (4.4.61)$$

which establishes  $Z_{\lfloor \alpha^n \rfloor} \xrightarrow{\text{a.s.}} \mathbb{E}[X]$ . All that remains to show is that  $Z_n \xrightarrow{\text{a.s.}} \mathbb{E}[X]$ . For every natural number  $n > \alpha$ , we can find another natural number  $k_n$  (possibly dependent on  $n$ ) such that

$$\lfloor \alpha^{k_n} \rfloor \leq n \leq \lfloor \alpha^{k_n+1} \rfloor \quad (4.4.62)$$

(which we can verify via choice of  $k_n = \lceil \log_\alpha n - 1 \rceil$ ). Then as  $\frac{\lfloor \alpha^{k_n} \rfloor}{\lfloor \alpha^{k_n+1} \rfloor} < 1$  and  $\frac{\lfloor \alpha^{k_n+1} \rfloor}{\lfloor \alpha^{k_n} \rfloor} > 1$  while  $Z_n$  is monotonic in  $n$  (i.e. each summand is non-negative), we have

$$\frac{\lfloor \alpha^{k_n} \rfloor}{\lfloor \alpha^{k_n+1} \rfloor} Z_{\lfloor \alpha^{k_n} \rfloor} \leq Z_n \leq \frac{\lfloor \alpha^{k_n+1} \rfloor}{\lfloor \alpha^{k_n} \rfloor} Z_{\lfloor \alpha^{k_n+1} \rfloor} \quad (4.4.63)$$

Taking the limit as  $n \rightarrow \infty$  and using the characterisation of  $\liminf$  and  $\limsup$  as the greatest lower bounds and least upper bounds respectively, we have

$$\frac{1}{\alpha} \mathbb{E}[X] \leq \liminf_{n \rightarrow \infty} Z_n \leq \limsup_{n \rightarrow \infty} Z_n \leq \alpha \mathbb{E}[X] \quad (4.4.64)$$

almost surely. Thus for  $\alpha$  arbitrarily close to 1, and with probability 1:

$$\lim_{n \rightarrow \infty} Z_n = \mathbb{E}[X] \quad (4.4.65)$$

or

$$\Pr \left( \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mathbb{E}[X] \right) = 1 \quad (4.4.66)$$

which completes the proof.  $\square$

#### 4.4.4 Sufficient Statistics

Loosely speaking, a statistic  $T(\mathbf{X})$  of a random sample  $\mathbf{X}$  is said to be sufficient for parameter  $\theta$  if  $T(\mathbf{X})$  contains all ‘information’ required for estimating  $\theta$ . To follow with a more formal definition, let  $\mathbf{X}$  be a random vector from a distribution parametrised by  $\theta$ . We allow for  $\theta$  to be a vector-valued parameter. Then the statistic  $T(\mathbf{X})$  is also a generally vector-valued statistic from  $\mathbf{X}$ . We say that  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if and only the conditional distribution of  $\mathbf{X}$  given  $T(\mathbf{X})$  does not depend on (which is to say, it is not a function of)  $\theta$ .

To illustrate further in a discrete setting, if  $T(\mathbf{X})$  were sufficient it would mean that while the probability  $\Pr_\theta(\mathbf{X} = \mathbf{x})$  depends on the value of  $\theta$  and  $\Pr_\theta(T(\mathbf{X}) = T(\mathbf{x}))$  also depends on the value of theta, the conditional probability  $\Pr_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$  no longer depends on  $\theta$ . So sufficiency characterises that somehow, just having  $T(\mathbf{X})$  takes out any extra ambiguity about  $\theta$ , compared to having all of  $\mathbf{X}$ .

#### Fisher-Neyman Factorisation Theorem

Another formal characterisation of sufficient statistics is provided by the Fisher-Neyman Factorisation Theorem. Denote the probability density or mass function of  $\mathbf{X}$  parametrised in  $\theta$  as  $f(\mathbf{x}; \theta)$ . Then  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if and only if there exist non-negative functions  $g(t, \theta)$  and  $h(\mathbf{x})$  such that

$$f(\mathbf{x}; \theta) = h(\mathbf{x}) g(T(\mathbf{x}), \theta) \quad (4.4.67)$$

for all parameter points  $\theta$ .

*Proof.* We first consider case for only discrete distributions [6, 41]. Suppose  $T(\mathbf{X})$  is sufficient; we seek to show the factorisation exists. Because  $T(\mathbf{X})$  is a function of  $\mathbf{X}$ :

$$\Pr_{\theta}(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})) = \Pr_{\theta}(\mathbf{X} = \mathbf{x}) \quad (4.4.68)$$

Hence

$$f(\mathbf{x}; \theta) = \Pr_{\theta}(\mathbf{X} = \mathbf{x}) \quad (4.4.69)$$

$$= \Pr_{\theta}(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})) \quad (4.4.70)$$

$$= \Pr_{\theta}(T(\mathbf{X}) = T(\mathbf{x})) \Pr_{\theta}(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \quad (4.4.71)$$

so by choice of  $g(t, \theta) = \Pr_{\theta}(T(\mathbf{X}) = t)$  and  $h(\mathbf{x}) = \Pr_{\theta}(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$  (which does not depend on  $\theta$  due to sufficiency), the factorisation holds. Now suppose the factorisation exists; we seek to show  $T(\mathbf{X})$  is sufficient. We write the conditional probability of  $\mathbf{X}$  given  $T(\mathbf{X})$  as

$$\Pr_{\theta}(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) = \frac{\Pr_{\theta}(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x}))}{\Pr_{\theta}(T(\mathbf{X}) = T(\mathbf{x}))} \quad (4.4.72)$$

$$= \frac{\Pr_{\theta}(\mathbf{X} = \mathbf{x})}{\Pr_{\theta}(T(\mathbf{X}) = T(\mathbf{x}))} \quad (4.4.73)$$

$$= \frac{f(\mathbf{x}; \theta)}{\Pr_{\theta}(T(\mathbf{X}) = T(\mathbf{x}))} \quad (4.4.74)$$

$$= \frac{h(\mathbf{x}) g(T(\mathbf{x}), \theta)}{\Pr_{\theta}(T(\mathbf{X}) = T(\mathbf{x}))} \quad (4.4.75)$$

using the factorisation. There is a one-to-many correspondence between  $T(\mathbf{x})$  and  $\mathbf{x}$ . That is, there can be many values of  $\mathbf{x}$  which can correspond to a single value of  $T(\mathbf{x})$ . Hence we compute  $\Pr_{\theta}(T(\mathbf{X}) = T(\mathbf{x}))$  by taking the Law of Total Probability over the many values of  $\mathbf{x}$  which lead to  $T(\mathbf{x})$ .

$$\frac{f(\mathbf{x}; \theta)}{\Pr_{\theta}(T(\mathbf{X}) = T(\mathbf{x}))} = \frac{h(\mathbf{x}) g(T(\mathbf{x}), \theta)}{\sum_{\{\mathbf{y}: T(\mathbf{y})=T(\mathbf{x})\}} f(\mathbf{y}; \theta)} \quad (4.4.76)$$

$$= \frac{h(\mathbf{x}) g(T(\mathbf{x}), \theta)}{\sum_{\{\mathbf{y}: T(\mathbf{y})=T(\mathbf{x})\}} h(\mathbf{y}) g(T(\mathbf{x}), \theta)} \quad (4.4.77)$$

$$= \frac{h(\mathbf{x}) g(T(\mathbf{x}), \theta)}{g(T(\mathbf{x}), \theta) \sum_{\{\mathbf{y}: T(\mathbf{y})=T(\mathbf{x})\}} h(\mathbf{y})} \quad (4.4.78)$$

$$= \frac{h(\mathbf{x})}{\sum_{\{\mathbf{y}: T(\mathbf{y})=T(\mathbf{x})\}} h(\mathbf{y})} \quad (4.4.79)$$

This ratio is not dependent on  $\theta$ , satisfying the definition of sufficiency. As we move to the continuous case [87, 124], the reason why it cannot be simply analogous to the discrete case is because generally for the density of  $T$ :

$$f_T(t; \theta) \neq \int_{\{\mathbf{y}: T(\mathbf{y})=t\}} f(\mathbf{y}; \theta) d\mathbf{y} \quad (4.4.80)$$

because we also need to take into account the Jacobian determinant in the differential, however the transformation  $T(\mathbf{x})$  is not necessarily an invertible transformation. Thus, we need to construct a transformation that is invertible. Suppose  $\mathbf{x} \in \mathbb{R}^n$  and  $T(\mathbf{x}) \in \mathbb{R}^r$  with  $r \leq n$ . Introduce the function  $Y(\mathbf{x}) \in \mathbb{R}^{n-r}$  and assume that  $\mathbf{x}$  and  $(T(\mathbf{x}), Y(\mathbf{x}))$  is one-to-one on a suitable domain. This assumption is valid if some regularity conditions are satisfied. Hence by the relation of densities between invertible transformations,

$$f(\mathbf{x}; \theta) = f_{T,Y}(T(\mathbf{x}), Y(\mathbf{x})) |\det(\mathbf{J})| \quad (4.4.81)$$

where  $\mathbf{J}$  is the Jacobian of  $(T(\mathbf{x}), Y(\mathbf{x}))$  with respect to  $\mathbf{x}$ . Now suppose the factorisation exists; we seek to show that  $T(\mathbf{X})$  is sufficient. As  $(T(\mathbf{x}), Y(\mathbf{x}))$  is one-to-one with  $\mathbf{x}$ , it is satisfactory to show that the conditional distribution of  $(T(\mathbf{X}), Y(\mathbf{X}))$  given  $T(\mathbf{X})$  does not depend on  $\theta$ . This conditional distribution can be written as just the conditional distribution of  $Y(\mathbf{X})$  given  $T(\mathbf{X})$ :

$$f_{T,Y|T}(t, y|t; \theta) = \frac{f_{T,Y,T}(t, y, t; \theta)}{f_T(t; \theta)} \quad (4.4.82)$$

$$= \frac{f_{T,Y}(t, y; \theta)}{f_T(t; \theta)} \quad (4.4.83)$$

$$= f_{Y|T}(y|t; \theta) \quad (4.4.84)$$

Thus we end up showing that the conditional distribution of  $Y(\mathbf{X})$  given  $T(\mathbf{X})$  does not depend on  $\theta$ . This conditional distribution is

$$f_{Y|T}(y|t; \theta) = \frac{f_{T,Y}(t, y; \theta)}{\int f_{T,Y}(t, y'; \theta) dy'} \quad (4.4.85)$$

Using the relation between densities and the factorisation:

$$f_{Y|T}(y|t; \theta) = \frac{h(\mathbf{x}) g(T(\mathbf{x}), \theta) / |\det(\mathbf{J})|}{\int h(\mathbf{x}') g(T(\mathbf{x}), \theta) / |\det(\mathbf{J})| dy'} \quad (4.4.86)$$

$$= \frac{h(\mathbf{x}) g(T(\mathbf{x}), \theta)}{g(T(\mathbf{x}), \theta) \int h(\mathbf{x}') dy'} \quad (4.4.87)$$

$$= \frac{h(\mathbf{x})}{\int h(\mathbf{x}') dy'} \quad (4.4.88)$$

where  $\mathbf{x}'(t, y')$  is some function of  $t$  and  $y'$ . Hence this does not depend on  $\theta$ , so sufficiency is shown. How suppose  $T(\mathbf{X})$  is sufficient; we seek to show the factorisation exists. From  $f(\mathbf{x}; \theta)$ :

$$f(\mathbf{x}; \theta) = f_{T,Y}(T(\mathbf{x}), Y(\mathbf{x})) |\det(\mathbf{J})| \quad (4.4.89)$$

$$= f_T(T(\mathbf{x}); \theta) f_{Y|T}(Y(\mathbf{x})|T(\mathbf{x})) |\det(\mathbf{J})| \quad (4.4.90)$$

where  $f_{Y|T}(Y(\mathbf{x})|T(\mathbf{x}))$  does not depend on  $\theta$  due to sufficiency. Also note that  $\mathbf{J}$  does not depend on  $\theta$ . Therefore choosing  $h(\mathbf{x}) = f_{Y|T}(Y(\mathbf{x})|T(\mathbf{x})) |\det(\mathbf{J})|$  and  $g(T(\mathbf{x}), \theta) = f_T(T(\mathbf{x}); \theta)$  realises the factorisation and completes the proof.  $\square$

The Fisher-Neyman Factorisation Theorem can help illustrate a ‘likelihood interpretation’ of sufficient statistics [212]. The left hand side of the factorisation,  $f(\mathbf{x}; \theta)$ , is also known as the likelihood of  $\theta$  given  $\mathbf{X}$ . Conditioning on  $\mathbf{X} = \mathbf{x}$ , then in the right-hand side,  $h(\mathbf{X})$  becomes a constant and we have proportionality

$$f(\mathbf{x}; \theta) \propto g(T(\mathbf{x}), \theta) \quad (4.4.91)$$

with respect to  $\theta$ . As the right-hand side now only contains  $T(\mathbf{x})$ , this means we can recover the shape of the likelihood using only  $T(\mathbf{x})$ . Thus  $T(\mathbf{x})$  contains all the information we need to infer  $\theta$  based on a likelihood approach.

### Minimally Sufficient Statistics [41]

There can be many different sufficient statistics for a parameter. A minimally sufficient statistic characterises in some sense how some sufficient statistics may be better than others. A sufficient statistic  $T(\mathbf{X})$  is said to be *minimally sufficient* if for any other sufficient statistic  $T'(\mathbf{X})$ , then  $T(\mathbf{x})$  is a function of  $T'(\mathbf{x})$ . That is to say,  $T(\mathbf{x})$  can be computed from  $T'(\mathbf{x})$ . As an example, suppose the sample mean  $T(\mathbf{X}) = \bar{\mathbf{X}}$  and the entire sample  $T'(\mathbf{X}) = \mathbf{X}$  are both sufficient statistics. Then  $T'(\mathbf{X})$  cannot be minimally sufficient (and  $T(\mathbf{X})$  is not precluded from being minimally sufficient) because although we can compute the sample mean from the entire sample, we cannot do it in reverse.

## Sufficiency Principle [152]

Let  $\mathbf{x}$  and  $\mathbf{y}$  be two different observations from an experiment, and let  $T(\cdot)$  be a sufficient statistic for a parameter  $\theta$ . The sufficiency principle says that if  $T(\mathbf{x}) = T(\mathbf{y})$ , then roughly speaking, we will end up drawing the same conclusions about  $\theta$  for either observation. For example, two different samples can lead to the same sample mean, and the sufficiency principle says that we should draw the same conclusions about the population mean from either sample.

### 4.4.5 Ancillary Statistics [41]

Ancillary statistics have a complementary property to sufficient statistics. A statistic  $T(\mathbf{X})$  of a random sample  $\mathbf{X}$  from a population parametrised by  $\theta$  is said to be *ancillary* for  $\theta$ , if the sampling distribution for  $T(\mathbf{X})$  does not depend on  $\theta$  (i.e.  $\theta$  does not appear in the functional form of the distribution). This is to say, an ancillary statistic contains no ‘information’ about  $\theta$ .

For example, suppose a random sample  $\mathbf{X}$  is drawn from a population parametrised by its location parameter  $\mu$ . The sample variance is then an ancillary statistic for  $\mu$  because shifting the location parameter does not affect the sampling distribution of the sample variance.

### 4.4.6 Complete Statistics [41]

A statistic  $T(\mathbf{X})$  of a random sample  $\mathbf{X}$  from a population parametrised by  $\theta$  is said to be *complete* if for any function  $g(\cdot)$ , then the property  $\mathbb{E}_\theta[g(T)] = 0$  over all values of  $\theta$  implies that  $g(T) = 0$  almost surely, irrespective of  $\theta$ . The latter is to say that:

$$\Pr_\theta(g(T) = 0) = 1 \quad (4.4.92)$$

for all  $\theta$ . To put it another way,  $T(\mathbf{X})$  is not complete if we can find a  $g(\cdot)$  such that  $\mathbb{E}_\theta[g(T)] = 0$  over all  $\theta$ , but  $g(T)$  is not almost surely 0. So we can tell if a statistic is complete or not by seeing, for any choice of  $g(\cdot)$ , whether the only way we can have  $\mathbb{E}_\theta[g(T)] = 0$  over all  $\theta$  is that because  $\Pr_\theta(g(T) = 0) = 1$  for all  $\theta$ . Also note that if we do have  $\Pr_\theta(g(T) = 0) = 1$ , does not necessarily mean that  $g(\cdot) = 0$  for all  $t$ . It is only required that  $g(\cdot) = 0$  over the support of  $T$ .

As an example, suppose  $T$  is a Rademacher-distributed random variable for all  $\theta$  (either by construction from some arbitrary population and statistic, or taking  $T$  to be a Rademacher random variable as a defintion). Pick  $g(t) = t$ . Then

$$\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[T] \quad (4.4.93)$$

$$= 0 \quad (4.4.94)$$

irrespective of  $\theta$ . However,  $g(T) = T$  is also Rademacher-distributed, hence not almost surely 0. Therefore,  $T$  is not a complete statistic. Note that the property of completeness applies to a family of distributions for  $T$  (parametrised by  $\theta$ ), hence the definition of completeness is always with respect to some parameter.

In the definition for completeness, there is nothing particular about the equality with zero for the expectation. We could come up with an alternative definition whereby the condition is  $\mathbb{E}_\theta[g(T)] = c$  for some constant  $c$ . However this will not be of any serious consequence, since for every  $g(\cdot)$  that satisfies  $\mathbb{E}_\theta[g(T)] = c$ , we will have that  $g'(t) := g(t) - c$  satisfies  $\mathbb{E}_\theta[g'(T)] = 0$ . Rather, the idea behind the completeness property is to characterise whether the sampling distribution of  $T$  is ‘distinct’ enough corresponding to different values of  $\theta$  for us

to be able to infer  $\theta$ , or whether the sampling distribution of  $T$  is concentrated at a single point. To develop this principle, start in the simple case of  $\mathbb{E}_\theta[T]$ , i.e. the first moment of  $T$ . If  $T$  is ‘distinct’ enough, then reasonably  $\mathbb{E}_\theta[T]$  should vary as  $\theta$  is varied. Otherwise  $\mathbb{E}_\theta[T]$  will be concentrated at a single point for all  $\theta$ . Extending this concept further, we take functions of  $T$ , these being  $g(\cdot)$ . Realise that this then includes higher moments  $\mathbb{E}_\theta[T^2]$ ,  $\mathbb{E}_\theta[T^3]$ , etc. So then  $\mathbb{E}_\theta[g(T)]$  should vary as  $\theta$  is varied, and because the moments contain all the information about the distribution, this means the whole distribution of  $T$  is ‘distinct’ enough. Or else if  $\mathbb{E}_\theta[g(T)]$  is concentrated at a single point, then this should only be because that this particular  $g(\cdot)$  is such that  $g(T) = 0$  almost surely. Otherwise then,  $T$  is not considered complete.

### Boundedly Complete Statistics

A statistic is said to be boundedly complete if it satisfies the definition of completeness, but for every function  $g(\cdot)$  which is bounded.

#### 4.4.7 Basu’s Theorem [41, 176]

Let  $T(\mathbf{X})$  be a boundedly complete and sufficient statistic, and let  $S(\mathbf{X})$  be an ancillary statistic for parameter  $\theta$ . Then  $T(\mathbf{X})$  is independent of  $S(\mathbf{X})$ .

*Proof.* Let  $S^{-1}(B) = \{\mathbf{x} : S(\mathbf{x}) \in B\}$  denote the preimage of set  $B$  under  $S(\cdot)$ . We write  $\Pr_\theta(S \in B)$  as

$$\Pr_\theta(S \in B) = \int \Pr_\theta(\mathbf{X} \in S^{-1}(B) | T = t) dF_T(t; \theta) \quad (4.4.95)$$

where  $F_T(t; \theta)$  is the CDF of  $T$ . Note that we can write the integral this way (and do so in order to facilitate discrete distributions) because  $dF_T(t; \theta) = f_T(t; \theta) dt$  where  $f_T(t; \theta)$  is the PDF of  $T$ . Otherwise, in only the discrete case, we could write this probability as

$$\Pr_\theta(S \in B) = \sum \Pr_\theta(\mathbf{X} \in S^{-1}(B) | T = t) \Pr_\theta(T = t) \quad (4.4.96)$$

We continuing on from the former. Since  $S(\mathbf{X})$  is ancillary, we can drop the  $\theta$  subscript in  $\Pr_\theta(S \in B)$ . Also since  $T(\mathbf{X})$  is sufficient, we can drop the  $\theta$  subscript in  $\Pr_\theta(\mathbf{X} \in S^{-1}(B) | T = t)$ . Then we have

$$\Pr_\theta(S \in B) \int dF_T(t; \theta) = \int \Pr_\theta(\mathbf{X} \in S^{-1}(B) | T = t) dF_T(t; \theta) \quad (4.4.97)$$

since  $\int dF_T(t; \theta) = 1$ . Rearrange this as

$$\int (\Pr_\theta(\mathbf{X} \in S^{-1}(B) | T = t) - \Pr_\theta(S \in B)) dF_T(t; \theta) = 0 \quad (4.4.98)$$

Let

$$g(t) = \Pr_\theta(\mathbf{X} \in S^{-1}(B) | T = t) - \Pr_\theta(S \in B) \quad (4.4.99)$$

So that

$$\int g(t) dF_T(t; \theta) = 0 \quad (4.4.100)$$

or equivalently,

$$\mathbb{E}_\theta[g(T)] = 0 \quad (4.4.101)$$

for any  $\theta$ . Since  $T(\mathbf{X})$  is boundedly complete (and  $g(t)$  is a bounded function), then this implies that  $g(T) = 0$  almost surely, i.e.

$$\Pr_\theta(\mathbf{X} \in S^{-1}(B) | T = t) = \Pr_\theta(S \in B) \quad (4.4.102)$$

almost surely. This implies that

$$\Pr_\theta(S \in B | T = t) = \Pr_\theta(S \in B) \quad (4.4.103)$$

which shows that  $T(\mathbf{X})$  and  $S(\mathbf{X})$  are independent.  $\square$

#### 4.4.8 U-Statistics [122]

The theory of U-statistics allows for symmetric unbiased estimators to be derived. Consider a parameter  $\theta$  of a population, and a sample of independent copies of  $X$  from the population. Suppose that  $k$  is the minimum sample size in order to conduct an unbiased estimate of  $\theta$  via the estimator  $\psi(X_1, \dots, X_k)$ , such that

$$\mathbb{E}[\psi(X_1, \dots, X_k)] = \theta \quad (4.4.104)$$

A U-statistic for  $\theta$  can be constructed as follows. Let  $\bar{\psi}(X_1, \dots, X_k)$  denote a ‘symmetrised’ version of  $\psi(X_1, \dots, X_k)$ , given by

$$\bar{\psi}(X_1, \dots, X_k) = \frac{1}{k!} \sum_{(i_1, \dots, i_k) \in \{1, \dots, k\}} \psi(X_{i_1}, \dots, X_{i_k}) \quad (4.4.105)$$

where the sum is taken over all  $k!$  permutations of indices in  $\{1, \dots, k\}$ . Clearly, if  $\psi(X_1, \dots, X_k)$  were not symmetric before (with symmetric meaning that the function is invariant to a permutation of its arguments), then  $\bar{\psi}(X_1, \dots, X_k)$  will be now. Also,  $\bar{\psi}(X_1, \dots, X_k)$  will retain its unbiasedness property due to the linearity of expectation. We extend this ‘minimal’ estimator to estimators on samples of size  $n$ . The U-statistic  $\hat{\theta}$  is defined as the average over all  $n$  choose  $k$  sub-samples evaluated on  $\bar{\psi}(X_1, \dots, X_k)$ . As a formula, this is expressed as

$$\hat{\theta} = \binom{n}{k}^{-1} \sum_{\mathcal{I}} \bar{\psi}(X_{i_1}, \dots, X_{i_k}) \quad (4.4.106)$$

where the set  $\mathcal{I}$  is the set of all  $n$  choose  $k$  subsets  $1 \leq i_1 < \dots < i_k \leq n$  in  $\{1, \dots, n\}$ . Due to the linearity of expectation,  $\hat{\theta}$  is also going to be an unbiased estimator.

#### U-Statistic of Mean

We require a minimal  $k = 1$  samples to make an unbiased estimate of the population mean  $\mu$ , since  $\mathbb{E}[X_1] = \mu$  with  $\psi(X_1) = \bar{\psi}(X_1) = X_1$ . Hence the U-statistic of the mean is

$$\hat{\mu} = n^{-1} \sum_{i=1}^n X_i \quad (4.4.107)$$

which is the usual unbiased sample mean.

#### U-Statistic of Variance

Recall that if  $X_1$  and  $X_2$  are independent copies, the population variance  $\sigma^2$  can be characterised as

$$\sigma^2 = \mathbb{E}\left[\frac{1}{2}(X_1 - X_2)^2\right] \quad (4.4.108)$$

Hence  $\psi(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2$  is a minimal unbiased estimator for the population variance, with  $k = 2$ . We have that

$$\bar{\psi}(X_1, X_2) = \frac{1}{2!} \left[ \frac{1}{2}(X_1 - X_2)^2 + \frac{1}{2}(X_2 - X_1)^2 \right] \quad (4.4.109)$$

$$= \frac{1}{2}(X_1 - X_2)^2 \quad (4.4.110)$$

Hence the U-statistic of the variance is

$$\hat{\sigma}^2 = \binom{n}{2}^{-1} \sum_{\{i,j:1 \leq i < j \leq n\}} \frac{1}{2}(X_i - X_j)^2 \quad (4.4.111)$$

Note that the summation can be equivalently computed by

$$\sum_{\{i,j:1 \leq i < j \leq n\}} \frac{1}{2} (X_i - X_j)^2 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{2} (X_i - X_j)^2 \quad (4.4.112)$$

because we can imagine a symmetric table of  $n \times n$ ; then iterating through all  $i$  and  $j$  amounts to iterating through  $\{i, j : 1 \leq i < j \leq n\}$  since  $X_i - X_j = 0$  when  $i = j$ . Hence proceeding with the computation:

$$\hat{\sigma}^2 = \frac{2}{n(n-1)} \left( \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n X_i^2 - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n X_i X_j + \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n X_j^2 \right) \quad (4.4.113)$$

$$= \frac{2}{n(n-1)} \left[ \frac{n}{4} \sum_{i=1}^n X_i^2 - \frac{1}{2} \left( \sum_{i=1}^n X_i \right) \left( \sum_{j=1}^n X_j \right) + \frac{n}{4} \sum_{j=1}^n X_j^2 \right] \quad (4.4.114)$$

$$= \frac{2}{n(n-1)} \left( \frac{n}{2} \sum_{i=1}^n X_i^2 - \frac{n^2}{2} \bar{X} \right) \quad (4.4.115)$$

$$= \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n \bar{X} \right) \quad (4.4.116)$$

where  $\bar{X}$  is the sample mean, which we know equals

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (4.4.117)$$

because this is the usual unbiased estimator for the population variance. In principle, symmetric unbiased estimators for the population skewness and kurtosis using a sample of size  $n$  can be derived this way.

### ***k*-Statistics**

The  $k$ -statistics are name given to the unique symmetric unbiased estimators of the cumulants.

#### **4.4.9 Minimum Variance Unbiased Estimators**

Consider  $X_1, \dots, X_n$  to be a sample of i.i.d. data from some underlying distribution, in which  $\theta$  is a parameter. Let  $T(X_1, \dots, X_n)$  be an unbiased estimator for  $\theta$ . Then the unbiased estimator  $T^*(X_1, \dots, X_n)$  is said to be the minimum variance unbiased estimator for  $\theta$  if

$$\text{Var}(T^*(X_1, \dots, X_n)) \leq \text{Var}(T(X_1, \dots, X_n)) \quad (4.4.118)$$

for any other unbiased estimator  $T(X_1, \dots, X_n)$ . For vector-valued estimators, this can be generalised to the matrix inequality involving the covariance matrices of the estimators, or a scalar inequality involving the trace of the covariance matrices (equal to the sum of the variances).

#### **4.4.10 Best Linear Unbiased Estimators**

Best linear unbiased estimators are a special class of minimum variance unbiased estimators, for when the estimator can be written as a linear combination of the random data:

$$T(X_1, \dots, X_n) = c_1 X_1 + c_2 X_2 + \dots + c_n X_n \quad (4.4.119)$$

where  $c_1, \dots, c_n$  are not allowed to depend on the true value of the parameter  $\theta$  being estimated.

#### 4.4.11 Gauss-Markov Theorem

The Gauss-Markov theorem states that ordinary least squares is the best linear unbiased estimator for linear regression under some standard assumptions. Consider the general linear model:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (4.4.120)$$

where  $\mathbf{X}$  is a fixed design matrix. The assumptions required are:

- Zero-mean residuals  $\mathbb{E}[\varepsilon] = 0$ .
- Homoskedastic errors:  $\text{Var}(\varepsilon_i) = \sigma^2$ .
- Uncorrelatedness between distinct error terms:  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  for  $i \neq j$ .

These three conditions together imply that  $\text{Cov}(\varepsilon) = \mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I$ . To prove the Gauss-Markov theorem, suppose  $\tilde{\beta} = C\mathbf{Y}$  is a linear estimator with

$$C = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + D \quad (4.4.121)$$

where  $D$  is an arbitrary matrix such that  $\tilde{\beta}$  can be made into any linear estimator by choice of  $D$ . To obtain the necessary conditions for this estimator to be unbiased, we have

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E}[C\mathbf{Y}] \quad (4.4.122)$$

$$= \mathbb{E}\left[\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + D\right)(\mathbf{X}\beta + \varepsilon)\right] \quad (4.4.123)$$

$$= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta + D\mathbf{X}\beta \quad (4.4.124)$$

$$= (I + D\mathbf{X})\beta \quad (4.4.125)$$

where we have used  $\mathbb{E}[\varepsilon] = 0$  and the fact that  $\mathbf{X}$  is deterministic with respect to the data generating process. Hence  $D\mathbf{X} = \mathbf{0}$  necessarily in order for the estimator to be unbiased. Then

$$\text{Cov}(\tilde{\beta}) = \text{Cov}(C\mathbf{Y}) \quad (4.4.126)$$

$$= C \text{Cov}(\mathbf{Y}) C^\top \quad (4.4.127)$$

$$= C \text{Cov}(\mathbf{X}\beta + \varepsilon) C^\top \quad (4.4.128)$$

$$= C \text{Cov}(\varepsilon) C^\top \quad (4.4.129)$$

$$= \sigma^2 C C^\top \quad (4.4.130)$$

$$= \sigma^2 \left[ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + D \right] \left[ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + D \right]^\top \quad (4.4.131)$$

$$= \sigma^2 \left[ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top D^\top + D\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} + D D^\top \right] \quad (4.4.132)$$

As  $D\mathbf{X} = \mathbf{0}$ , then

$$\text{Cov}(\tilde{\beta}) = \sigma^2 \left[ (\mathbf{X}^\top \mathbf{X})^{-1} + D D^\top \right] \quad (4.4.133)$$

$$= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} + \sigma^2 D D^\top \quad (4.4.134)$$

Recognise that  $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$  is the covariance of the ordinary least squares estimator  $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$  so

$$\text{Cov}(\hat{\beta}) = \text{Cov}(\hat{\beta}) + \sigma^2 DD^\top \quad (4.4.135)$$

$$\succeq \text{Cov}(\hat{\beta}) \quad (4.4.136)$$

since  $\sigma^2 DD^\top$  will be some symmetric positive semidefinite matrix. Additionally since  $\text{trace}(\sigma^2 DD^\top) \geq 0$ , this implies

$$\text{trace}(\text{Cov}(\tilde{\beta})) \geq \text{trace}(\text{Cov}(\hat{\beta})) \quad (4.4.137)$$

so the total variance of any other linear unbiased estimator will be lower bounded by the total variance of the OLS estimator. However, it is possible for a biased estimator to give a lower total variance, for example the ridge regression estimator.

#### 4.4.12 Rao-Blackwell Theorem [41]

The Rao-Blackwell theorem gives a way to improve an estimator. Let  $\hat{\theta}$  be an unbiased estimator of  $\theta$ , which for regularity, we shall require to have finite second moment. Let  $T$  be a sufficient statistic for  $\theta$ . Define the estimator

$$\hat{\theta}^* = \mathbb{E}[\hat{\theta}|T] \quad (4.4.138)$$

Then  $\hat{\theta}^*$  is also unbiased, and is also uniformly (i.e. for all  $\theta$ ) ‘better’ than  $\hat{\theta}$  (in the sense of lower variance), i.e.

$$\text{Var}(\hat{\theta}^*) \leq \text{Var}(\theta) \quad (4.4.139)$$

*Proof.* We use the Law of Iterated Expectations to show

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\mathbb{E}[\hat{\theta}|T]] \quad (4.4.140)$$

$$= \mathbb{E}[\hat{\theta}] \quad (4.4.141)$$

$$= \theta \quad (4.4.142)$$

irrespective of the value of  $\theta$ , where the first equality follows by definition, the second equality by the Law of Iterated Expectations, and the last equality by the condition that  $\hat{\theta}$  be unbiased. Then using the Law of Total Variance, we can show for any  $\theta$  that

$$\text{Var}(\hat{\theta}) = \text{Var}(\mathbb{E}[\hat{\theta}|T]) + \mathbb{E}[\text{Var}(\hat{\theta}|T)] \quad (4.4.143)$$

$$= \text{Var}(\hat{\theta}^*) + \mathbb{E}[\text{Var}(\hat{\theta}|T)] \quad (4.4.144)$$

$$\geq \text{Var}(\hat{\theta}^*) \quad (4.4.145)$$

where the second equality is by definition and the final inequality is due to the fact that  $\mathbb{E}[\text{Var}(\hat{\theta}|T)] \geq 0$ .  $\square$

We should also establish that the quantity  $\hat{\theta}^*$  can be computed. Since  $T(\mathbf{X})$  of a random sample  $\mathbf{X}$  is sufficient for  $\theta$ , then the conditional distribution of  $\mathbf{X}$  given  $T(\mathbf{X})$  does not contain  $\theta$ . As  $\hat{\theta}$  is also computed from the sample, by extension the conditional distribution of  $\hat{\theta}$  given  $T(\mathbf{X})$  also does not contain  $\theta$ . It follows that the conditional expectation  $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|T]$  does not contain  $\theta$ .

### Rao-Blackwell Estimators

A Rao-Blackwell estimator is any estimator  $\widehat{\theta}^*$  that has been improved from  $\widehat{\theta}$  using the Rao-Blackwell theorem. We also say that  $\widehat{\theta}^*$  has been ‘Rao-Blackwellised’. Note that even if  $\widehat{\theta}^*$  or  $\widehat{\theta}$  are biased, we will still have the variance reduction property; it holds that  $\text{Var}(\widehat{\theta}^*) \leq \text{Var}(\widehat{\theta})$ .

To illustrate an unbiased estimator being improved by the Rao-Blackwell theorem, suppose  $\mathbf{X}$  is a random sample of  $n$  i.i.d. Bernoulli random variables with success parameter  $\theta$ . Let a ‘crude’ estimator simply be the first result, i.e.  $\widehat{\theta} = X_1$ . Clearly, this estimator is unbiased as  $\mathbb{E}[X_1] = \theta$ . We claim that  $T(\mathbf{X}) = \sum_{i=1}^n X_i$  is a sufficient statistic for  $\theta$ . To show this, recognise that  $T(\mathbf{X})$  is the number of successes (i.e. it will be binomially distributed), and by the Fisher-Neyman Factorisation Theorem, the factorisation

$$f(\mathbf{x}; \theta) = \binom{n}{T(\mathbf{x})} \theta^{T(\mathbf{x})} (1 - \theta)^{n - T(\mathbf{x})} \quad (4.4.146)$$

$$= g(T(\mathbf{x}), \theta) h(\mathbf{x}) \quad (4.4.147)$$

holds, with  $g(T(\mathbf{x}), \theta) = \binom{n}{T(\mathbf{x})} \theta^{T(\mathbf{x})} (1 - \theta)^{n - T(\mathbf{x})}$  and  $h(\mathbf{x}) = 1$ . By Rao-Blackwellisation, we take

$$\widehat{\theta}^* = \mathbb{E}[\widehat{\theta} | T(\mathbf{X})] \quad (4.4.148)$$

$$= \mathbb{E}[X_1 | X_1 + \dots + X_n] \quad (4.4.149)$$

To compute this conditional expectation, we firstly use symmetry:

$$\mathbb{E}[X_1 | X_1 + \dots + X_n] = \mathbb{E}[X_2 | X_1 + \dots + X_n] \quad (4.4.150)$$

$$\vdots \quad (4.4.151)$$

$$= \mathbb{E}[X_n | X_1 + \dots + X_n] \quad (4.4.152)$$

and then

$$\mathbb{E}[X_1 | X_1 + \dots + X_n] + \dots + \mathbb{E}[X_n | X_1 + \dots + X_n] = \mathbb{E}[X_1 + \dots + X_n | X_1 + \dots + X_n] \quad (4.4.153)$$

$$= X_1 + \dots + X_n \quad (4.4.154)$$

to yield

$$\mathbb{E}[X_1 | X_1 + \dots + X_n] = \frac{X_1 + \dots + X_n}{n} \quad (4.4.155)$$

Hence we have recovered  $\widehat{\theta}^*$  as the sample mean, which is a better estimator than  $\widehat{\theta}$ . This is verified by applying the variance of the Bernoulli and binomial distributions:

$$\text{Var}(\widehat{\theta}) = \theta(1 - \theta) \quad (4.4.156)$$

$$\text{Var}(\widehat{\theta}^*) = \frac{1}{n} \theta(1 - \theta) \quad (4.4.157)$$

so we see that  $\text{Var}(\widehat{\theta}^*) \leq \text{Var}(\widehat{\theta})$ .

#### 4.4.13 Lehmann-Scheffé Theorem

Suppose  $T$  is a complete and sufficient statistic for  $\theta$ . Let  $\widehat{\theta}(T)$  be an estimator for  $\theta$ , which can be computed from  $T$ . If  $\widehat{\theta}(T)$  is unbiased, i.e.

$$\mathbb{E}[\widehat{\theta}(T)] = \theta \quad (4.4.158)$$

then  $\widehat{\theta}(T)$  is the unique minimum variance unbiased estimator for  $\theta$ .

*Proof.* The minimum variance property can be established via the Rao-Blackwell theorem. Note that Rao-Blackwellisation of  $\hat{\theta}(T)$  gives itself:

$$\hat{\theta}^*(T) = \mathbb{E} [\hat{\theta}(T) | T] \quad (4.4.159)$$

$$= \hat{\theta}(T) \quad (4.4.160)$$

hence  $\hat{\theta}(T)$  must achieve the lowest variance. To show uniqueness, consider another candidate minimum variance unbiased estimator, denoted  $\hat{\theta}'(T)$ . We can assume that  $\hat{\theta}'$  is a function of  $T$ , because otherwise we could just Rao-Blackwellise to obtain another estimator with the same minimum variance that does depend on  $T$ . By unbiasedness, we have that

$$\mathbb{E} [\hat{\theta}(T) - \hat{\theta}'(T)] = 0 \quad (4.4.161)$$

Let  $g(T) = \hat{\theta}(T) - \hat{\theta}'(T)$ . Since  $T$  is complete, this implies  $g(T) = 0$  almost surely, i.e.

$$\hat{\theta}(T) = \hat{\theta}'(T) \quad (4.4.162)$$

almost surely. Hence  $\hat{\theta}(T)$  is a unique estimator.  $\square$

#### 4.4.14 Minimum Mean Square Error Estimators

Let  $Y$  be a random vector and let  $\hat{Y}$  be its estimate. Then the mean square error is given by

$$\text{MSE} = \mathbb{E} [(\mathbf{Y} - \hat{\mathbf{Y}})^2] \quad (4.4.163)$$

The estimator  $\hat{Y}$  which minimises this is called the minimum mean square error (MMSE) estimator. If we are considering random vector  $\mathbf{Y}$  being estimated by  $\hat{\mathbf{Y}}$  instead, the mean square error may be defined as the trace of the expected outer product of estimation error:

$$\text{MSE} = \text{trace} \left( \mathbb{E} [(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})^\top] \right) \quad (4.4.164)$$

In the same way as defining the scalarised variance, this can be shown to be the same as

$$\text{MSE} = \mathbb{E} [\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2] \quad (4.4.165)$$

#### Projection Theorem [197]

Let  $\mathbf{X}$  be a random vector, and let  $\hat{Y}$  be a linear estimator for  $Y$  from  $\mathbf{X}$  by  $\hat{Y} = \hat{\theta}^\top \mathbf{X}$ . Denote the estimation error  $\varepsilon = Y - \hat{Y}$ . If  $\hat{\theta}$  is chosen such that

$$\mathbb{E} [\varepsilon \mathbf{X}] = \mathbf{0} \quad (4.4.166)$$

i.e. the estimation error is orthogonal to  $\mathbf{X}$ , then  $\hat{\theta}$  is the minimum mean square error estimator.

*Proof.* Let  $\theta^*$  be the weighting vector which makes  $\mathbb{E} [\varepsilon \mathbf{X}] = \mathbf{0}$ , and let  $\hat{\theta}$  be any arbitrary weighting vector. We have

$$\varepsilon = Y - \hat{\theta} \mathbf{X} \quad (4.4.167)$$

$$= Y - \mathbf{X}^\top \theta^* + \mathbf{X}^\top (\theta^* - \hat{\theta}) \quad (4.4.168)$$

$$= \varepsilon^* + \mathbf{X}^\top (\theta^* - \hat{\theta}) \quad (4.4.169)$$

where  $\varepsilon^* = Y - \mathbf{X}^\top \theta^*$  denotes the estimation error from using  $\theta^*$ . Then the mean squared estimation from using  $\hat{\theta}$  is

$$\mathbb{E} [\varepsilon^2] = \mathbb{E} \left[ (\varepsilon^* + \mathbf{X}^\top (\theta^* - \hat{\theta})) (\varepsilon^* + \mathbf{X}^\top (\theta^* - \hat{\theta}))^\top \right] \quad (4.4.170)$$

$$= \mathbb{E} [(\varepsilon^*)^2] + 2\mathbb{E} \left[ \varepsilon^* \mathbf{X}^\top \right] (\theta^* - \hat{\theta}) + \mathbb{E} \left[ \mathbf{X}^\top (\theta^* - \hat{\theta}) (\theta^* - \hat{\theta})^\top \mathbf{X} \right] \quad (4.4.171)$$

$$= \mathbb{E} [(\varepsilon^*)^2] + \mathbb{E} \left[ \left\| \mathbf{X}^\top (\theta^* - \hat{\theta}) \right\|^2 \right] \quad (4.4.172)$$

where  $\mathbb{E} [\varepsilon^* \mathbf{X}^\top] = 0$  by definition of  $\theta^*$ . We see that in order to minimise  $\mathbb{E} [\varepsilon^2]$  in terms of  $\hat{\theta}$ , we should set  $\hat{\theta} = \theta^*$ .  $\square$

Moreover, the minimum mean square error itself is given by

$$\mathbb{E} [(\varepsilon^*)^2] = \mathbb{E} \left[ (Y - \mathbf{X}^\top \theta^*) \varepsilon \right] \quad (4.4.173)$$

$$= \mathbb{E} [Y \varepsilon] - \mathbb{E} [\varepsilon \mathbf{X}^\top] \theta^* \quad (4.4.174)$$

$$= 0 \quad (4.4.175)$$

#### 4.4.15 James-Stein Estimation

**Stein's Lemma**

**Stein's Example**

**James-Stein Estimator**

### 4.5 Maximum Likelihood Estimation

#### 4.5.1 Likelihood Function

Suppose we have a random sample  $\mathbf{X}$  from a parametrised family of distributions, parametrised in  $\theta \in \Theta \subseteq \mathbb{R}^p$ , where  $\Theta$  is the parameter space. If  $\mathbf{X}$  is a discrete random vector, then it is drawn from a parametrised joint probability mass function  $p_{\mathbf{X}}(\mathbf{x}; \theta)$ . This can be thought of as the probability of  $\mathbf{x}$ , given  $\theta$ . The likelihood function can be imagined as a ‘reversal’ of this characterisation - it is a function of  $\theta$ , given  $\mathbf{x}$  (the data). The value of the likelihood loosely means how likely it was that the parameter  $\theta$  generated the data  $\mathbf{x}$ . We write the likelihood function by

$$\mathcal{L}(\theta; \mathbf{x}) = p_{\mathbf{X}}(\mathbf{x}; \theta) \quad (4.5.1)$$

If  $\mathbf{X}$  is a continuous random vector with parametrised joint density  $f_{\mathbf{X}}(\mathbf{x}; \theta)$ , then we use the density for the likelihood instead:

$$\mathcal{L}(\theta; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \theta) \quad (4.5.2)$$

Note that likelihoods should not always be interpreted as probabilities. In the case  $\mathbf{X}$  is a continuous random variable, then the density hence likelihood may be greater than one. Also the likelihood function will generally not sum/integrate to one over  $\Theta$ .

## Likelihood Function of Independent and Identically Distributed Samples

If the random sample  $\mathbf{X}$  of size  $n$  is an i.i.d. sample from the density  $f_X(x)$ , then the likelihood function can be written as a product of densities:

$$\mathcal{L}(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i; \theta) \quad (4.5.3)$$

This analogously holds if  $\mathbf{X}$  is a discrete random variable.

### Log-Likelihood Function

The log-likelihood function is simply the log of the likelihood function, denoted  $\log \mathcal{L}(\theta; \mathbf{x})$ . If the sample is i.i.d., then taking the log-likelihood means taking the sum of the log of the individual likelihoods:

$$\log \mathcal{L}(\theta; x_1, \dots, x_n) = \log \left( \prod_{i=1}^n f_X(x_i; \theta) \right) \quad (4.5.4)$$

$$= \sum_{i=1}^n \log f_X(x_i; \theta) \quad (4.5.5)$$

### Likelihood Principle [18, 164]

The sufficiency principle can be thought of as a weaker version of the likelihood principle [152], which says that if we have two different observations  $\mathbf{x}$  and  $\mathbf{y}$  such that  $\mathcal{L}(\theta; \mathbf{x}) = \mathcal{L}(\theta; \mathbf{y})$ , then either observation should lead to the same conclusions about  $\theta$ . A stronger statement (given by the likelihood principle) is to assert that the likelihood function contains all the ‘evidence’ about  $\theta$ . Moreover, suppose we consider two different likelihoods  $\mathcal{L}_1(\theta; \mathbf{x})$  and  $\mathcal{L}_2(\theta; \mathbf{y})$  pertaining to the same parameter  $\theta$ , but arising from two different experiments. If we observe  $\mathbf{x}$  and  $\mathbf{y}$  such that there exists a some constant  $c$  whereby

$$\mathcal{L}_1(\theta; \mathbf{x}) = c \mathcal{L}_2(\theta; \mathbf{y}) \quad (4.5.6)$$

for all  $\theta$ , then either experiment will have brought about identical inferences about  $\theta$ . For example, consider a binomial experiment with  $n$  trials and  $x$  observed successes, with the success probability parameter  $\theta$ . The likelihood here is  $\mathcal{L}_1(\theta; x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$ . Now consider an alternative experiment to count the number of trials  $y$  until the first success. This has a geometric likelihood  $\mathcal{L}_2(\theta; y) = \theta (1 - \theta)^{y-1}$ . If both experiments were conducted simultaneously and it happened to be that we observe  $x = 1$  and  $y = n$ , then

$$\underbrace{n \theta (1 - \theta)^{n-1}}_{\mathcal{L}_1(\theta; x)} = c \underbrace{\theta (1 - \theta)^{n-1}}_{\mathcal{L}_2(\theta; y)} \quad (4.5.7)$$

with  $c = n$ . Thus under the likelihood principle, we would end up drawing the same conclusions about  $\theta$  from the results of either experiment.

#### 4.5.2 Maximum Likelihood Estimator

The maximum likelihood estimator for a parameter  $\theta$  given the sample  $\mathbf{x}$  finds the value of  $\theta$  which ‘most likely’ explains  $\mathbf{x}$ . It is expressed as

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta; \mathbf{x}) \quad (4.5.8)$$

Since  $\log$  is a strictly positive monotonic transformation, then this is equivalent of maximising the log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log \mathcal{L}(\theta; \mathbf{x}) \quad (4.5.9)$$

This is also equivalent to minimising the negative log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\operatorname{argmin}} \{-\log \mathcal{L}(\theta; \mathbf{x})\} \quad (4.5.10)$$

If the negative log-likelihood is twice-differentiable, then the sufficient conditions that a local minimum is found at the maximum likelihood estimate is that the gradient is zero and the Hessian is positive semi-definite at the maximum likelihood estimate:

$$-\nabla_{\theta} \log \mathcal{L}(\theta; \mathbf{x})|_{\theta=\hat{\theta}_{\text{MLE}}} = 0 \quad (4.5.11)$$

$$-\nabla_{\theta}^2 \log \mathcal{L}(\theta; \mathbf{x})|_{\theta=\hat{\theta}_{\text{MLE}}} \succeq 0 \quad (4.5.12)$$

### Maximum Likelihood Estimator for Normal Distribution

Given an i.i.d. sample  $\mathbf{X} = (X_1, \dots, X_n)$  where each  $X_i$  is normally distributed with  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ , maximum likelihood estimates for the parameters  $\mu$  and  $\sigma^2$  can be derived as follows. First writing the likelihood of the sample using the normal density:

$$\mathcal{L}(\mu, \sigma^2; \mathbf{X}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(X_i - \mu)^2}{2\sigma^2} \right] \quad (4.5.13)$$

Taking the log-likelihood yields:

$$\log \mathcal{L}(\mu, \sigma^2; \mathbf{X}) = \log \left[ \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(X_i - \mu)^2}{2\sigma^2} \right] \right] \quad (4.5.14)$$

$$= \sum_{i=1}^n \log \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(X_i - \mu)^2}{2\sigma^2} \right] \right] \quad (4.5.15)$$

$$= -\frac{1}{2} \sum_{i=1}^n \log(2\pi\sigma^2) + \sum_{i=1}^n \log \left[ \exp \left[ -\frac{(X_i - \mu)^2}{2\sigma^2} \right] \right] \quad (4.5.16)$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \quad (4.5.17)$$

The partial derivative with respect to  $\mu$  is

$$\frac{\partial}{\partial \mu} \log \mathcal{L}(\mu, \sigma^2; \mathbf{X}) = -\frac{1}{2\sigma^2} \sum_{i=1}^n [-2(X_i - \mu)] \quad (4.5.18)$$

Setting the derivative to zero and solving for  $\mu$  yields the maximum likelihood estimator the the population mean:

$$\sum_{i=1}^n (X_i - \hat{\mu}) = 0 \quad (4.5.19)$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \quad (4.5.20)$$

which is the usual sample mean. Now taking the partial derivative with respect to  $\sigma^2$  gives

$$\frac{\partial}{\partial \sigma^2} \log \mathcal{L}(\mu, \sigma^2; \mathbf{X}) = -\frac{n}{2} \cdot \frac{2\pi}{2\pi\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2 \quad (4.5.21)$$

Setting the derivative to zero and solving for  $\sigma^2$  yields the maximum likelihood estimator for the population variance:

$$\frac{n}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^4} \sum_{i=1}^n (X_i - \mu)^2 \quad (4.5.22)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (4.5.23)$$

We have previously found that  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$  causes  $\frac{\partial}{\partial \mu} \log \mathcal{L}(\mu, \sigma^2; \mathbf{X}) = 0$ , which is necessary for the maximum likelihood estimate. Hence replacing  $\mu$  with  $\hat{\mu}$  gives

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2 \quad (4.5.24)$$

Note that this estimate of the population variance does not contain Bessel's correction, hence it is a biased estimate.

### Maximum Likelihood Estimator for Multivariate Normal Distribution

Given an i.i.d. sample  $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$  where each  $\mathbf{X}_i \in \mathbb{R}^p$  is multivariate normally distributed with  $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ , maximum likelihood estimates for the mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\Sigma$  can be derived as follows. First writing the likelihood of the sample using the multivariate normal density:

$$\mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbb{X}) = \prod_{i=1}^n (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) \right] \quad (4.5.25)$$

Taking the log-likelihood yields:

$$\log \mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbb{X}) = \log \left[ \prod_{i=1}^n (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) \right] \right] \quad (4.5.26)$$

$$= \sum_{i=1}^n \log \left[ (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) \right] \right] \quad (4.5.27)$$

$$= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) \quad (4.5.28)$$

$$= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{i=1}^n \text{trace} \left[ (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) \right] \quad (4.5.29)$$

$$= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{i=1}^n \text{trace} \left[ \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})^\top \right] \quad (4.5.30)$$

$$= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \text{trace} \left[ \Sigma^{-1} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})^\top \right] \quad (4.5.31)$$

where we took the trace of a scalar and applied the property of invariance to cyclic permutations for the trace. Then taking the derivative of the log-likelihood with respect to  $\boldsymbol{\mu}$ :

$$\frac{\partial}{\partial \boldsymbol{\mu}} \log \mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbb{X}) = \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^n \text{trace} \left[ (\mathbf{X}_i - \boldsymbol{\mu})^\top (\mathbf{X}_i - \boldsymbol{\mu}) \Sigma^{-1} \right] \quad (4.5.32)$$

$$= \sum_{i=1}^n \left[ \frac{\partial}{\partial \boldsymbol{\mu}} (\mathbf{X}_i - \boldsymbol{\mu})^\top (\mathbf{X}_i - \boldsymbol{\mu}) \right] \Sigma^{-1} \quad (4.5.33)$$

$$= - \sum_{i=1}^n 2 (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} \quad (4.5.34)$$

where we have used the fact  $\frac{\partial \text{trace}(AZB)}{\partial Z} = BA$ , followed by the chain rule. Setting the derivative to zero and rearranging, we will recover the usual multivariate sample mean estimator for the population mean:

$$- \sum_{i=1}^n 2 (\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top \Sigma^{-1} = 0 \quad (4.5.35)$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \quad (4.5.36)$$

For the covariance, we use the fact that the differential of a determinant is given by:

$$d \det(\Sigma) = \det(\Sigma) \text{trace}(\Sigma^{-1} d\Sigma) \quad (4.5.37)$$

Hence by the chain rule

$$d \log \det(\Sigma) = \frac{\det(\Sigma) \text{trace}(\Sigma^{-1} d\Sigma)}{\det(\Sigma)} \quad (4.5.38)$$

$$= \text{trace}(\Sigma^{-1} d\Sigma) \quad (4.5.39)$$

The differential of a matrix inverse is

$$d\Sigma^{-1} = -\Sigma^{-1} (d\Sigma) \Sigma^{-1} \quad (4.5.40)$$

Combining this with the chain rule and using the fact that the trace is a linear operator and so commutes with the derivative:

$$d \text{trace} \left[ \Sigma^{-1} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})^\top \right] = \text{trace} \left[ d\Sigma^{-1} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})^\top \right] \quad (4.5.41)$$

$$= \text{trace} \left[ -\Sigma^{-1} (d\Sigma) \Sigma^{-1} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})^\top \right] \quad (4.5.42)$$

Therefore the differential of the log-likelihood (using our earlier result of the derivative with respect to the mean):

$$\begin{aligned} d \log \mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbb{X}) &= -\frac{n}{2} \text{trace}(\Sigma^{-1} d\Sigma) - \frac{1}{2} \text{trace} \left[ -\Sigma^{-1} (d\Sigma) \Sigma^{-1} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})^\top \right] \\ &\quad - \sum_{i=1}^n 2 (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} d\boldsymbol{\mu} \end{aligned} \quad (4.5.43)$$

For the first order condition  $d \log \mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbb{X}) = 0$ , we require the terms multiplying  $d\boldsymbol{\mu}$  and  $d\Sigma$  be zero. We solved for  $\hat{\boldsymbol{\mu}}$  earlier, so substituting  $\boldsymbol{\mu}$  for this yields

$$d \log \mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbb{X}) = -\frac{n}{2} \text{trace}(\Sigma^{-1} d\Sigma) - \frac{1}{2} \text{trace} \left[ -\Sigma^{-1} (d\Sigma) \Sigma^{-1} \sum_{i=1}^n (\mathbf{X}_i - \hat{\boldsymbol{\mu}}) (\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top \right] \quad (4.5.44)$$

Let  $S = \sum_{i=1}^n (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top$  represent the scatter matrix, so we simplify by

$$d \log \mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbb{X}) = -\frac{n}{2} \text{trace}(\Sigma^{-1} d\Sigma) - \frac{1}{2} \text{trace}[-\Sigma^{-1}(d\Sigma)\Sigma^{-1}S] \quad (4.5.45)$$

$$= -\frac{1}{2} \text{trace}[\Sigma^{-1} d\Sigma (nI - \Sigma^{-1}S)] \quad (4.5.46)$$

The first order condition holds when  $nI - \Sigma^{-1}S = 0$ , so solving for the maximum likelihood estimator:

$$nI = \hat{\Sigma}^{-1}S \quad (4.5.47)$$

$$\hat{\Sigma} = \frac{1}{n}S \quad (4.5.48)$$

which is written out in full as

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top \quad (4.5.49)$$

This shows that the maximum likelihood estimator for the covariance of a multivariate normal sample is of the same form as the sample covariance, but without Bessel's correction (hence it will be a biased estimate).

### Maximum Likelihood Estimator from Binned Data

Consider a parametric family of distributions  $f(y; \theta)$ , and suppose we wish to estimate  $m$ -dimensional  $\theta$  from a sample of size  $n$ , using only the counts of the data which have fallen in  $K$  bins  $B_1, \dots, B_K$  with endpoints:

$$B_1 = (b_0, b_1] \quad (4.5.50)$$

$$\vdots \quad (4.5.51)$$

$$B_K = (b_{K-1}, b_K] \quad (4.5.52)$$

If the distribution is supported on the real line, we can take  $b_0 = -\infty$  and  $b_K = \infty$ . Let  $X_1, \dots, X_K$  denote the counts such that  $\sum_{i=1}^K X_i = n$ . Then the likelihood is given by the multinomial distribution:

$$\mathcal{L}(\theta; X_1, \dots, X_K) \propto p_1(\theta)^{X_1} \times \cdots \times p_K(\theta)^{X_K} \quad (4.5.53)$$

where

$$p_1(\theta) = F(b_1; \theta) - F(b_0; \theta) \quad (4.5.54)$$

$$\vdots \quad (4.5.55)$$

$$p_K(\theta) = F(b_K; \theta) - F(b_{K-1}; \theta) \quad (4.5.56)$$

and we can ignore the multinomial coefficient because it does not affect maximisation of the likelihood (i.e. due to the likelihood principle). Taking the log likelihood:

$$\log \mathcal{L}(\theta; X_1, \dots, X_K) = X_1 \log p_1(\theta) + \cdots + X_K \log p_K(\theta) \quad (4.5.57)$$

and differentiating with respect to parameter  $\theta_j$  for each  $j = 1, \dots, m$ :

$$\frac{\partial \log \mathcal{L}(\theta; X_1, \dots, X_K)}{\partial \theta_j} = \sum_{i=1}^K X_i \frac{\partial p_i(\theta)}{\partial \theta_j} \cdot \frac{1}{p_i(\theta)} \quad (4.5.58)$$

Thus setting derivatives to zero, the maximum likelihood estimate  $\hat{\theta}$  satisfies for each  $j = 1, \dots, m$ :

$$\sum_{i=1}^K \frac{X_i}{p_i(\hat{\theta})} \left. \frac{\partial p_i(\theta)}{\partial \theta_j} \right|_{\theta=\hat{\theta}} = 0 \quad (4.5.59)$$

This property is required to derive the degrees of freedom in the chi-squared goodness-of-fit test for parametric distributions. If only the linear constraint  $\sum_i^K X_i = n$  is imposed on the random vector  $\mathbf{X} = (X_1, \dots, X_K)$ , then it was shown that the corresponding chi-squared statistic had  $K - 1$  degrees of freedom. In the case that the  $p_i$  must be estimated from data, the chi-squared statistic  $T$  now becomes

$$T = \sum_{i=1}^K \frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \quad (4.5.60)$$

Since we impose  $m$  additional linear restrictions on  $\mathbf{X}$  from setting the derivative of the log-likelihood to zero, then the covariance matrix of  $\mathbf{X}$  will be of rank  $K - 1 - m$ , and therefore the appropriate degrees of freedom of  $T$  is  $K - 1 - m$ . In effect, these linear restrictions mean that the standardised vector  $(n \operatorname{diag}\{\mathbf{p}(\hat{\theta})\})^{-1/2} (\mathbf{X} - n\mathbf{p}(\hat{\theta}))$  lies on a reduced dimension subspace, which reduces the degrees of freedom of the chi-squared random variable [73].

### Conditional Maximum Likelihood Estimation

Conditional maximum likelihood estimation is when we maximise the likelihood conditioned on some other random variables. The usual scenario is a regression problem with random sample  $(\mathbf{y}, \mathbf{X})$ . For a model with parameter  $\theta$ , we may wish to maximise the likelihood of  $\theta$  given  $\mathbf{y}$ , all conditioned on  $\mathbf{X}$ . The notation for this problem can be expressed as

$$\hat{\theta} = \max_{\theta} \mathcal{L}(\theta; \mathbf{y} | \mathbf{X}) \quad (4.5.61)$$

to show explicit conditioning on  $\mathbf{X}$ . With this, we are effectively treating  $\mathbf{X}$  as non-random, and can essentially solve the problem in the same way as in maximum likelihood, where  $\mathbf{X}$  is considered deterministic.

#### 4.5.3 Asymptotic Consistency of Maximum Likelihood Estimator

Suppose  $\theta_0$  is the ‘true’ parameter of the parametrised distribution from which a sample of  $X_i$  is generated. To derive the asymptotic consistency of the maximum likelihood estimator, we first note the (weak) Law of Large Numbers implies

$$\frac{1}{n} \sum_{i=1}^n \log \mathcal{L}(\theta; X_i) \xrightarrow{p} \mathbb{E}[\log f_X(X; \theta)] \quad (4.5.62)$$

We assume the following regularity condition:

$$\max_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \log \mathcal{L}(\theta; X_i) \right\} \xrightarrow{p} \max_{\theta \in \Theta} \mathbb{E}[\log f_X(X; \theta)] \quad (4.5.63)$$

which holds under ‘nice enough’ conditions, i.e. roughly speaking  $\theta_0$  is the unique maximiser of the likelihood,  $\Theta$  should be compact,  $\log \mathcal{L}(\theta; x)$  should be continuous in  $\theta$  for all  $x$ , and the expectation  $\mathbb{E}[\log f_X(X; \theta)]$  should be always finite. We then claim that  $\operatorname{argmax}_{\theta \in \Theta} \mathbb{E}[\log f_X(X; \theta)] = \theta_0$ , which is shown as follows. For any  $\theta \in \Theta$ ,

$$\mathbb{E}[\log f_X(X; \theta)] - \mathbb{E}[\log f_X(X; \theta_0)] = \mathbb{E} \left[ \log \frac{f_X(X; \theta)}{f_X(X; \theta_0)} \right] \quad (4.5.64)$$

Since  $\log$  is concave, then using Jensen's inequality we have

$$\mathbb{E} \left[ \log \frac{f_X(X; \theta)}{f_X(X; \theta_0)} \right] \leq \log \mathbb{E} \left[ \frac{f_X(X; \theta)}{f_X(X; \theta_0)} \right] \quad (4.5.65)$$

$$= \log \left( \int \frac{f_X(x; \theta)}{f_X(x; \theta_0)} f(x; \theta_0) dx \right) \quad (4.5.66)$$

$$= \log \left( \int f_X(x; \theta) dx \right) \quad (4.5.67)$$

$$= \log 1 \quad (4.5.68)$$

$$= 0 \quad (4.5.69)$$

Note that although we have treated  $X$  as continuous, it can analogously shown if  $X$  were discrete. Hence

$$\mathbb{E} [\log f_X(X; \theta)] - \mathbb{E} [\log f_X(X; \theta_0)] \leq 0 \quad (4.5.70)$$

or

$$\mathbb{E} [\log f_X(X; \theta)] \leq \mathbb{E} [\log f_X(X; \theta_0)] \quad (4.5.71)$$

for all  $\theta \in \Theta$ , which means  $\operatorname{argmax}_{\theta \in \Theta} \mathbb{E} [\log f_X(X; \theta)] = \theta_0$ . Therefore

$$\operatorname{argmax}_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \log \mathcal{L}(\theta; X_i) \right\} \xrightarrow{\text{P}} \theta_0 \quad (4.5.72)$$

This can also analogously be shown with almost sure convergence using the Strong Law of Large numbers (yielding strong consistency).

#### 4.5.4 Maximum Likelihood Justification of Least Squares

Suppose we have data consisting of  $n$  data pairs with regressors  $x_i$  and observations  $y_i$  along with a model  $f(x_i; \theta)$  which is parametrised by the parameter vector  $\theta$  to be estimated. If we assume the data generating process

$$Y_i = f(x_i; \theta) + \varepsilon_i \quad (4.5.73)$$

with Gaussian errors  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ , then  $Y_i \sim \mathcal{N}(f(x_i; \theta), \sigma^2)$  and the likelihood under an i.i.d. assumption on each  $Y_i$  is given by

$$\mathcal{L}(\theta | Y_1, \dots, Y_n) = p(Y_1, \dots, Y_n | \theta) \quad (4.5.74)$$

$$= p(Y_1 | \theta) \dots p(Y_n | \theta) \quad (4.5.75)$$

$$= \prod_{i=1}^n p(Y_i | \theta) \quad (4.5.76)$$

$$= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(y_i - f(x_i; \theta))^2}{2\sigma^2} \right] \quad (4.5.77)$$

The log-likelihood is given by

$$\log \mathcal{L}(\theta | Y_1, \dots, Y_n) = - \sum_{i=1}^n \frac{1}{2\sigma^2} (y_i - f(x_i; \theta))^2 - \sum_{i=1}^n \log (\sigma \sqrt{2\pi}) \quad (4.5.78)$$

and maximising the log-likelihood with respect to the parameters,

$$\operatorname{argmax}_{\theta} \log \mathcal{L}(\theta | Y_1, \dots, Y_n) = \operatorname{argmax}_{\theta} \left\{ - \sum_{i=1}^n \frac{1}{2\sigma^2} (y_i - f(x_i; \theta))^2 - \sum_{i=1}^n \log (\sigma \sqrt{2\pi}) \right\} \quad (4.5.79)$$

$$= \operatorname{argmin}_{\theta} \left\{ \sum_{i=1}^n \frac{1}{2\sigma^2} (y_i - f(x_i; \theta))^2 + \sum_{i=1}^n \log(\sigma\sqrt{2\pi}) \right\} \quad (4.5.80)$$

$$= \operatorname{argmin}_{\theta} \sum_{i=1}^n \frac{1}{2\sigma^2} (y_i - f(x_i; \theta))^2 \quad (4.5.81)$$

$$= \operatorname{argmin}_{\theta} \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i; \theta))^2 \quad (4.5.82)$$

so we arrive at the least squares cost function. Hence the least squares fitting problem can be interpreted as a maximum likelihood problem under an i.i.d. assumption on the observations  $Y_i$  and Gaussian errors  $\varepsilon_i$ . If the standard deviation on the errors is allowed to vary such that  $\text{sd}(\varepsilon_i) = \sigma_i$ , then the estimation problem becomes

$$\theta^* = \operatorname{argmin}_{\theta} \frac{1}{2} \sum_{i=1}^n \frac{(y_i - f(x_i; \theta))^2}{\sigma_i^2} \quad (4.5.83)$$

and the formulation becomes equivalent to weighted least squares.

### Maximum Likelihood Estimator for Linear Regression

Beginning with the general linear model  $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ , if we assume errors are normally distributed, ‘spherical’ (i.e. uncorrelated) and ‘homoskedastic’ (i.e. constant conditional covariance  $\text{Cov}(\varepsilon) = \sigma^2 I$ ), then the conditional distribution of  $\mathbf{y}$  given  $\mathbf{X}$  is

$$\mathbf{y} | \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I) \quad (4.5.84)$$

with parameters  $\beta$  and  $\sigma^2$ . Thus the likelihood of parameters  $\beta, \sigma^2$  given data  $\mathbf{X}, \mathbf{y}$  can be written using the multivariate normal density

$$\mathcal{L}(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(\sigma^2 I)}} \exp \left[ -\frac{(\mathbf{y} - \mathbf{X}\beta)^\top (\sigma^2 I)^{-1} (\mathbf{y} - \mathbf{X}\beta)}{2} \right] \quad (4.5.85)$$

$$= \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} \exp \left[ -\frac{(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \right] \quad (4.5.86)$$

Thus taking the log-likelihood:

$$\log \mathcal{L}(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \quad (4.5.87)$$

Using the chain rule, the gradient with respect to  $\beta$  is

$$\nabla_{\beta} \log \mathcal{L}(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) = -\frac{-\mathbf{X}^\top \cdot 2(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \quad (4.5.88)$$

$$= \frac{1}{\sigma^2} (\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\beta) \quad (4.5.89)$$

The maximum likelihood estimate for  $\beta$  is obtained by solving  $\nabla_{\beta} \log \mathcal{L}(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) = 0$ , which yields

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (4.5.90)$$

This is the same as the ordinary least squares estimate. To find the maximum likelihood estimate, we take the partial derivative of the log-likelihood with respect to  $\sigma^2$ :

$$\frac{\partial \log \mathcal{L}(\beta, \sigma^2 | \mathbf{X}, \mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)}{2\sigma^4} \quad (4.5.91)$$

The maximum likelihood estimate is when both  $\nabla_{\beta} \log \mathcal{L}(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) = 0$  and  $\frac{\partial \log \mathcal{L}(\beta, \sigma^2 | \mathbf{X}, \mathbf{y})}{\partial \sigma^2} = 0$ . Hence the latter equation will involve the estimate  $\hat{\sigma}^2$  and the already-obtained  $\hat{\beta}$ .

$$-\frac{n}{2\hat{\sigma}^2} + \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})}{2\hat{\sigma}^4} = 0 \quad (4.5.92)$$

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) \quad (4.5.93)$$

We see that  $\hat{\sigma}^2$  is not taken as the sample variance of the residuals with Bessel's correction as is usually done in OLS, but rather the MLE for  $\sigma^2$  has factor  $\frac{1}{n}$  which is the same factor in the MLE for variance from normal samples.

#### 4.5.5 Maximum Likelihood Justification of Least Absolute Deviations

Suppose we have data consisting of  $n$  data pairs  $(x_i, y_i)$  along with a model  $f(x_i; \theta)$  parametrised in the parameter vector  $\theta$  to be estimated. If we assume the data generating process

$$Y_i = f(x_i; \theta) + \epsilon_i \quad (4.5.94)$$

with zero-mean Laplace-distributed errors  $\epsilon \sim \frac{\sqrt{2}}{\sigma} \exp\left(-\frac{|\epsilon|}{\sigma/\sqrt{2}}\right)$  with variance  $\sigma^2$ , then  $Y_i \sim \frac{\sqrt{2}}{\sigma} \exp\left(-\frac{|y_i - f(x_i; \theta)|}{\sigma/\sqrt{2}}\right)$  and the likelihood under an i.i.d. assumption on each  $Y_i$  is given by

$$\mathcal{L}(\theta | Y_1, \dots, Y_n) = p(Y_1, \dots, Y_n | \theta) \quad (4.5.95)$$

$$= p(Y_1 | \theta) \dots p(Y_n | \theta) \quad (4.5.96)$$

$$= \prod_{i=1}^n p(Y_i | \theta) \quad (4.5.97)$$

$$= \prod_{i=1}^n \frac{\sqrt{2}}{\sigma} \exp\left(-\frac{|y_i - f(x_i; \theta)|}{\sigma/\sqrt{2}}\right) \quad (4.5.98)$$

$$\propto \prod_{i=1}^n \exp\left(-\frac{|y_i - f(x_i; \theta)|}{\sigma/\sqrt{2}}\right) \quad (4.5.99)$$

Maximising the log-likelihood with respect to the parameters:

$$\operatorname{argmax}_{\theta} \log \mathcal{L}(\theta | Y_1, \dots, Y_n) = \operatorname{argmax}_{\theta} \left\{ - \sum_{i=1}^n \left( \frac{|y_i - f(x_i; \theta)|}{\sigma/\sqrt{2}} \right) \right\} \quad (4.5.100)$$

$$= \operatorname{argmin}_{\theta} \left\{ \sum_{i=1}^n |y_i - f(x_i; \theta)| \right\} \quad (4.5.101)$$

which gives the least absolute deviation cost function. Hence solving the least absolute deviation fitting problem can be interpreted as solving a maximum likelihood problem under an i.i.d. assumption on the observations  $Y_i$  and Laplace-distributed errors  $\epsilon_i$ .

#### 4.5.6 Log Concavity of Likelihoods

Sometimes, it is useful to know whether the maximum likelihood problem has a unique solution. Establishing log concavity of the likelihood function can be used to determine this. A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is said to be log concave if  $f(x) > 0$  and  $\log f(x)$  is concave in  $x$ . This implies

$-\log f(x)$  is convex in  $x$ . Note that in the context of maximum likelihood, we look for log concavity of the likelihood function in the parameters. A necessary and sufficient condition for log concavity can be obtained by the second derivative of  $g(x) := \log f(x)$ . In the simple case  $x \in \mathbb{R}$  and assuming  $f(x)$  is twice differentiable:

$$g'(x) = \frac{f'(x)}{f(x)} \quad (4.5.102)$$

Using the product rule:

$$g''(x) = -\left[\frac{f'(x)}{f(x)}\right]^2 + \frac{f''(x)}{f(x)} \quad (4.5.103)$$

$$= \frac{f''(x)f(x) - f'(x)^2}{f(x)^2} \quad (4.5.104)$$

The condition for concavity of  $g(x)$  is  $g''(x) \leq 0$ , so this gives the condition for log concavity of  $f(x)$  as

$$f'(x)^2 \geq f''(x)f(x) \quad (4.5.105)$$

while  $f'(x)^2 > f''(x)f(x)$  for strict log concavity. Another useful result is that log concavity of the cumulative distribution function can be established from log concavity of the density function. This can be useful in cases where observations consist of ‘greater than’ or ‘less than’ some parameter, since the CDF expresses the probability of an observation being less than its argument.

**Theorem 4.3.** *The cumulative distribution function of a log concave differentiable probability density function is also log concave.*

*Proof.* If  $f(x)$  is a log concave density then  $h(x) := -\log f(x)$  is convex. The cumulative distribution is

$$F(x) = \int_{-\infty}^x f(t) dt \quad (4.5.106)$$

$$= \int_{-\infty}^x e^{-h(t)} dt \quad (4.5.107)$$

The derivatives of  $F(x)$  are

$$F'(x) = f(x) \quad (4.5.108)$$

$$= e^{-h(x)} \quad (4.5.109)$$

and

$$F''(x) = -h'(x)e^{-h(x)} \quad (4.5.110)$$

$$= -h'(x)f(x) \quad (4.5.111)$$

Consider the case  $h'(x) \geq 0$ . Then because  $F(x) > 0$  and  $f(x) > 0$ :

$$F(x)F''(x) = -F(x)h'(x)f(x) \leq 0 \quad (4.5.112)$$

and

$$F'(x)^2 \geq 0 \quad (4.5.113)$$

which implies  $F'(x)^2 \geq F(x)F''(x)$ , satisfying the log concavity condition. Now consider the case  $h'(x) < 0$ . By the property that tangents of a convex function never lie above the graph, we get

$$h(t) \geq h(x) + h'(x)(t - x) \quad (4.5.114)$$

Hence

$$\int_{-\infty}^x e^{-h(t)} dt \leq \int_{-\infty}^x e^{-h(x)-h'(x)(t-x)} dt \quad (4.5.115)$$

We show for the right hand side:

$$\int_{-\infty}^x e^{-h(x)-h'(x)(t-x)} dt = e^{-h(x)+xh'(x)} \int_{-\infty}^x e^{-th'(x)} dt \quad (4.5.116)$$

$$= e^{-h(x)+xh'(x)} \frac{e^{-th'(x)}}{h'(x)} \quad (4.5.117)$$

$$= \frac{e^{-h(x)}}{h'(x)} \quad (4.5.118)$$

Multiplying both sides of the inequality by  $-h'(x) e^{-h(x)}$ :

$$-h'(x) e^{-h(x)} \int_{-\infty}^x e^{-h(t)} dt \leq e^{-2h(x)} \quad (4.5.119)$$

which is the same as the log concavity condition:

$$F''(x) F(x) \leq F(x)^2 \quad (4.5.120)$$

□

#### 4.5.7 Maximum Likelihood of Exponential Families [143]

Recall that we can express the distribution of an exponential family in canonical form as

$$f(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x}) - \psi(\boldsymbol{\theta})\right) \quad (4.5.121)$$

$$= \frac{h(\mathbf{x})}{\phi(\boldsymbol{\theta})} \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x})\right) \quad (4.5.122)$$

where  $\phi(\boldsymbol{\theta}) = \exp(\psi(\boldsymbol{\theta}))$ . We assume that  $f(\mathbf{x}; \boldsymbol{\theta})$  is a continuous distribution; everything is analogous if  $f(\mathbf{x}; \boldsymbol{\theta})$  were a probability mass function. Since the density integrates to one, we have

$$\phi(\boldsymbol{\theta}) = \int h(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x})\right) d\mathbf{x} \quad (4.5.123)$$

We can also refer to

$$\psi(\boldsymbol{\theta}) = \log \phi(\boldsymbol{\theta}) \quad (4.5.124)$$

as the *log-partition* function. We first demonstrate the the log-partition function is convex in parameters  $\boldsymbol{\theta}$ . Taking the first partial derivative with respect to the  $j^{\text{th}}$  parameter using the chain rule:

$$\frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} (\log \phi(\boldsymbol{\theta})) \quad (4.5.125)$$

$$= \frac{1}{\phi(\boldsymbol{\theta})} \cdot \frac{\partial \phi(\boldsymbol{\theta})}{\partial \theta_j} \quad (4.5.126)$$

$$= \frac{1}{\exp(\psi(\boldsymbol{\theta}))} \cdot \frac{\partial}{\partial \theta_j} \left( \int h(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x})\right) d\mathbf{x} \right) \quad (4.5.127)$$

$$= \frac{1}{\exp(\psi(\boldsymbol{\theta}))} \cdot \int h(\mathbf{x}) \frac{\partial}{\partial \theta_j} \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x})\right) d\mathbf{x} \quad (4.5.128)$$

$$= \frac{1}{\exp(\psi(\boldsymbol{\theta}))} \cdot \int h(\mathbf{x}) \cdot T_j(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x})\right) d\mathbf{x} \quad (4.5.129)$$

where  $T_j(\mathbf{x})$  is the  $j^{\text{th}}$  sufficient statistic. Continuing,

$$\frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_j} = \int h(\mathbf{x}) \cdot T_j(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x}) - \psi(\boldsymbol{\theta})\right) d\mathbf{x} \quad (4.5.130)$$

$$= \int T_j(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \quad (4.5.131)$$

$$= \mathbb{E}_{f(\mathbf{x}; \boldsymbol{\theta})}[T_j(\mathbf{X})] \quad (4.5.132)$$

Thus the gradient vector is

$$\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) = \mathbb{E}_{f(\mathbf{x}; \boldsymbol{\theta})}[T(\mathbf{X})] \quad (4.5.133)$$

Taking the second partial derivative:

$$\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} = \frac{\partial}{\partial \theta_k} \left( \int h(\mathbf{x}) \cdot T_j(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x}) - \psi(\boldsymbol{\theta})\right) d\mathbf{x} \right) \quad (4.5.134)$$

$$= \int h(\mathbf{x}) \cdot T_j(\mathbf{x}) \frac{\partial}{\partial \theta_k} \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x}) - \psi(\boldsymbol{\theta})\right) d\mathbf{x} \quad (4.5.135)$$

$$= \int h(\mathbf{x}) \cdot T_j(\mathbf{x}) \left( T_k(\mathbf{x}) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_k} \right) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x}) - \psi(\boldsymbol{\theta})\right) d\mathbf{x} \quad (4.5.136)$$

$$= \int T_j(\mathbf{x}) T_k(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_k} \int T_j(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \quad (4.5.137)$$

$$= \mathbb{E}_{f(\mathbf{x}; \boldsymbol{\theta})}[T_j(\mathbf{X}) T_k(\mathbf{X})] - \mathbb{E}_{f(\mathbf{x}; \boldsymbol{\theta})}[T_j(\mathbf{X})] \mathbb{E}_{f(\mathbf{x}; \boldsymbol{\theta})}[T_k(\mathbf{X})] \quad (4.5.138)$$

$$= \text{Cov}_{f(\mathbf{x}; \boldsymbol{\theta})}(T_j(\mathbf{X}), T_k(\mathbf{X})) \quad (4.5.139)$$

Hence the Hessian of the log-partition function is

$$\nabla_{\boldsymbol{\theta}}^2 \psi(\boldsymbol{\theta}) = \text{Cov}_{f(\mathbf{x}; \boldsymbol{\theta})}(T(\mathbf{X})) \quad (4.5.140)$$

$$\succeq \mathbf{0} \quad (4.5.141)$$

Since the Hessian is positive semi-definite, this shows that the log-partition function is convex. Now consider the log-likelihood for an exponential family distribution:

$$\log \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \log f(\mathbf{x}; \boldsymbol{\theta}) \quad (4.5.142)$$

$$= \boldsymbol{\theta}^\top T(\mathbf{x}) - \psi(\boldsymbol{\theta}) + \log h(\mathbf{x}) \quad (4.5.143)$$

Suppose we have data  $\mathcal{D} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ , then the log-likelihood of the parameters given the data is

$$\log \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \boldsymbol{\theta}^\top \sum_{i=1}^n T(\mathbf{x}_i) - n\psi(\boldsymbol{\theta}) + \sum_{i=1}^n \log h(\mathbf{x}_i) \quad (4.5.144)$$

Of the terms which depend on  $\boldsymbol{\theta}$ , the term  $\boldsymbol{\theta}^\top \sum_{i=1}^n T(\mathbf{x}_i)$  is linear while  $-n\psi(\boldsymbol{\theta})$  is concave (as the log-partition function is convex). Hence the log-likelihood is concave in  $\boldsymbol{\theta}$ , meaning that it has a unique global maximum. Moreover, if we take the gradient of the log-likelihood, we get

$$\nabla_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{i=1}^n T(\mathbf{x}_i) - n\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) \quad (4.5.145)$$

$$= \sum_{i=1}^n T(\mathbf{x}_i) - n\mathbb{E}_{f(\mathbf{x}; \boldsymbol{\theta})}[T(\mathbf{X})] \quad (4.5.146)$$

Setting the gradient to zero, we see that at the maximum likelihood estimate  $\hat{\boldsymbol{\theta}}$  that:

$$\mathbb{E}_{f(\mathbf{x}; \hat{\boldsymbol{\theta}})}[T(\mathbf{X})] = \frac{1}{n} \sum_{i=1}^n T(\mathbf{x}_i) \quad (4.5.147)$$

So finding the maximum likelihood estimate for an exponential family amounts to setting  $\theta$  such that the theoretical expectation of the sufficient statistics under  $\theta$  is equal to the empirical average of the sufficient statistics. This is similar in idea to [moment matching](#).

#### 4.5.8 Expectation Maximisation

The expectation maximisation (EM) algorithm is suited for solving more difficult maximum likelihood problems, where the maximum likelihood estimator does not necessarily have a closed-form solution, so iterative methods must be used to maximise the likelihood. Suppose the observed data  $Y$  has the parametric distribution  $p(y; \theta)$ , and the goal as usual is to find

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(y; \theta) \quad (4.5.148)$$

If this is difficult to solve directly (e.g.  $\log \mathcal{L}(\theta) := \log p(y; \theta)$  may not even be differentiable in  $\theta$ ), then we can introduce the latent (unobserved) variable  $z$ , with the joint probability model  $p(z, y; \theta)$ . The motivation is that it may be easier to maximise the likelihood given  $z$ , instead of  $y$ . Note that the underlying problem may or may not need to consider an unobserved variable;  $z$  could be artificially introduced as a way to make maximum likelihood estimation tractable. For problems that do consider latent variables however (e.g.  $z$  could represent missing data), then the EM algorithm provides a natural framework for estimation. The idea is to locally approximate the log-likelihood function about some value  $\theta'$ . Then for another arbitrary  $\theta$ ,

$$\log \mathcal{L}(\theta) - \log \mathcal{L}(\theta') = \log \frac{p(y; \theta)}{p(y; \theta')} \quad (4.5.149)$$

$$= \log \left( \frac{\int p(z, y; \theta) dz}{p(y; \theta')} \right) \quad (4.5.150)$$

$$= \log \left( \int \frac{p(z, y; \theta)}{p(y; \theta')} dz \right) \quad (4.5.151)$$

From  $p(z, y; \theta') = p(z|y; \theta') p(y; \theta')$  we have  $p(y; \theta') = \frac{p(z, y; \theta')}{p(z|y; \theta')}$ , and substituting this gives

$$\log \mathcal{L}(\theta) - \log \mathcal{L}(\theta') = \log \left( \int \frac{p(z, y; \theta)}{p(z, y; \theta')} p(z|y; \theta') dz \right) \quad (4.5.152)$$

$$\geq \underbrace{\int \log \left( \frac{p(z, y; \theta)}{p(z, y; \theta')} \right) p(z|y; \theta') dz}_{\Delta(\theta; \theta')} \quad (4.5.153)$$

where the inequality follows from [Jensen's inequality](#), since the logarithm is concave. We can split up  $\Delta(\theta; \theta')$  as

$$\Delta(\theta; \theta') = \int \log(p(z, y; \theta) p(z|y; \theta')) dz - \int \log(p(z, y; \theta') \cdot p(z|y; \theta')) dz \quad (4.5.154)$$

where only the first term depends on  $\theta$ , so define

$$Q(\theta; \theta') = \int \log(p(z, y; \theta) \cdot p(z|y; \theta')) dz \quad (4.5.155)$$

whereby if  $\theta'$  is fixed and we want to maximise the local log-likelihood approximation about  $\theta'$  with respect to  $\theta$ , we can maximise  $Q(\theta; \theta')$ . Note that at worst,  $\Delta(\theta'; \theta') = 0$ , so  $\max_{\theta} \Delta(\theta; \theta') \geq 0$ , and from Jensen's inequality this ensures

$$\log \mathcal{L}(\theta) - \log \mathcal{L}(\theta') \geq 0 \quad (4.5.156)$$

whenever  $\theta = \operatorname{argmax}_\theta \Delta(\theta; \theta')$ . Thus the log-likelihood is guaranteed to never decrease. We can also interpret the quantity  $Q(\theta; \theta')$  as the conditional expected log-likelihood:

$$Q(\theta; \theta') = \mathbb{E}_{Z|y; \theta'} [\log p(Z, y; \theta)] \quad (4.5.157)$$

We can further assume that we are able to write

$$p(z, y; \theta) = p(y|z) p(z; \theta) \quad (4.5.158)$$

which asserts that  $Y$  could be determined fully if we knew  $Z$ , without knowing  $\theta$ . If  $\theta$  were treated as a random variable, then  $Y$  and  $\theta$  would be conditionally independent, given  $Z$ . Since we are introducing  $Z$  for the purpose of making  $Q(\theta; \theta')$  easier to maximise than  $\log \mathcal{L}(\theta)$  (i.e. inferring  $\theta$  is easier if we knew  $Z$ , compared to knowing  $Y$ ), then this is a reasonable assumption. Since the term  $p(y|z)$  does not contain  $\theta$ , then under this scenario it means that  $Q(\theta; \theta')$  can be written as

$$Q(\theta; \theta') = \mathbb{E}_{Z|y; \theta'} [\log p(Z; \theta)] \quad (4.5.159)$$

### Expectation Maximisation Algorithm

We can use successive local log-likelihood approximations to formulate an iterative algorithm. Starting from some initial guess  $\hat{\theta}_0$  of  $\theta \in \Theta$  and until some termination condition, the  $k^{\text{th}}$  iteration of the algorithm consists of the following steps [76]:

1. Formulate the conditional probability distribution  $p(z|y; \hat{\theta}_k)$  for  $Z$ .
2. For the expectation step, calculate the conditional expected log-likelihood:

$$Q(\theta; \hat{\theta}_k) = \int \log(p(z, y; \theta)) \cdot p(z|y; \hat{\theta}_k) dz \quad (4.5.160)$$

$$= \mathbb{E}_{Z|y; \hat{\theta}_k} [\log p(Z, y; \theta)] \quad (4.5.161)$$

or if it can be simplified:

$$Q(\theta; \hat{\theta}_k) = \int \log(p(z; \theta)) \cdot p(z|y; \hat{\theta}_k) dz \quad (4.5.162)$$

$$= \mathbb{E}_{Z|y; \hat{\theta}_k} [\log p(Z; \theta)] \quad (4.5.163)$$

3. For the maximisation step, find

$$\hat{\theta}_{k+1} = \operatorname{argmax}_{\theta \in \Theta} Q(\theta; \hat{\theta}_k) \quad (4.5.164)$$

Intuitively, the EM algorithm can be thought to be iteratively performing inference about  $Z$  (with the distribution  $p(z|y, \hat{\theta}_k)$ ), then this can be used to make an updated guess about  $\theta$ , which can be used to perform better inference about  $Z$ , etc.

### Point Estimate Expectation Maximisation

A simplified variant of the EM algorithm involves:

1. Analogous to the expectation step, calculate

$$\hat{z}_k = \operatorname{argmax}_z \left\{ p(z|y, \hat{\theta}_k) \right\} \quad (4.5.165)$$

2. For the maximisation step, find

$$\hat{\theta}_{k+1} = \operatorname{argmax}_{\theta \in \Theta} \{ \log p(\hat{z}_k; \theta) \} \quad (4.5.166)$$

This algorithm is also known as the *classification EM algorithm*. We can think of this algorithm as performing point estimates for  $Z$  using  $\hat{z}_k$ , instead of using  $p(z|y, \hat{\theta}_k)$  like with the original EM algorithm.

## Data Imputation

### 4.5.9 M-Estimators [84]

M-estimators generalise maximum likelihood estimators. M-estimators are a class of estimators that are obtained by minimisation of a sample average of some function with respect to a parameter:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n m(X_i; \theta) \right\} \quad (4.5.167)$$

In the maximum likelihood estimation, we would have  $m(X_i; \theta) = -n \log \mathcal{L}(\theta|X_i)$  (note that the factor of  $n$  does not make a difference in the minimisation). Non-linear least squares can also be considered a special case of M-estimation.

### 4.5.10 Extremum Estimators [84]

Extremum estimators are a class of estimators which further generalise M-estimators. Extremum estimates for  $\theta$  are obtained by the minimisation (or maximisation) of some objective function which depends on the data  $\mathbf{X}$ :

$$\hat{\theta} = \operatorname{argmin}_{\theta} Q(\theta; \mathbf{X}) \quad (4.5.168)$$

## 4.6 Maximum Likelihood Inference

### 4.6.1 Score Function

Let  $f(x; \theta)$  be a probability density function parametrised by  $\theta$  for the random variable  $X$  on support  $\mathcal{X}$ . Assume that the support  $\mathcal{X}$  does not change with  $\theta$ . Note that  $f(X; \theta)$  is a random variable for the density of  $X$ , and that  $f(x; \theta)$  is also a likelihood function for  $\theta$ . The partial derivative of the log likelihood with respect to  $\theta$  is called the score  $\frac{\partial}{\partial \theta} \log f(X; \theta)$ . It can be shown that the expected value of the score with respect to  $X$  is zero:

$$\mathbb{E}_X \left[ \frac{\partial}{\partial \theta} \log f(X; \theta) \right] = \mathbb{E}_X \left[ \frac{1}{f(X; \theta)} \frac{\partial}{\partial \theta} f(X; \theta) \right] \quad (4.6.1)$$

by the chain rule

$$= \int_{\mathcal{X}} \frac{1}{f(x; \theta)} \cdot \frac{\partial}{\partial \theta} f(x; \theta) \cdot f(x; \theta) dx \quad (4.6.2)$$

$$= \int_{\mathcal{X}} \frac{\partial}{\partial \theta} f(x; \theta) dx \quad (4.6.3)$$

$$= \frac{\partial}{\partial \theta} \left( \int_{\mathcal{X}} f(x; \theta) dx \right) \quad (4.6.4)$$

$$= \frac{\partial}{\partial \theta} (1) \quad (4.6.5)$$

$$= 0 \quad (4.6.6)$$

The reason why we require the support  $\mathcal{X}$  to change with  $\theta$  is because otherwise, passing the differentiation through the integration would not be straightforward.

#### 4.6.2 Fisher Information

The Fisher information  $\mathcal{I}(\theta)$  is defined as the variance of the score

$$\mathcal{I}(\theta) = \text{Var} \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right) \quad (4.6.7)$$

$$= \mathbb{E}_X \left[ \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \right] - \mathbb{E}_X \left[ \frac{\partial}{\partial \theta} \log f(X; \theta) \right]^2 \xrightarrow{0} \quad (4.6.8)$$

$$= \mathbb{E}_X \left[ \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \right] \quad (4.6.9)$$

We can show that

$$\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) = \frac{\partial}{\partial \theta} \left( \frac{1}{f(X; \theta)} \cdot \frac{\partial}{\partial \theta} f(X; \theta) \right) \quad (4.6.10)$$

by the chain rule

$$= -\frac{\partial}{\partial \theta} f(X; \theta) \cdot \frac{\partial}{\partial \theta} f(X; \theta) + \frac{1}{f(X; \theta)} \cdot \frac{\partial^2}{\partial \theta^2} f(X; \theta) \quad (4.6.11)$$

by the product rule

$$= -\left( \frac{\partial}{\partial \theta} f(X; \theta) \right)^2 + \frac{\frac{\partial^2}{\partial \theta^2} f(X; \theta)}{f(X; \theta)} \quad (4.6.12)$$

$$= \frac{\frac{\partial^2}{\partial \theta^2} f(X; \theta)}{f(X; \theta)} - \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \quad (4.6.13)$$

using  $\frac{\partial}{\partial \theta} \log f(X; \theta) = \frac{\frac{\partial}{\partial \theta} f(X; \theta)}{f(X; \theta)}$

Hence

$$\left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 = \frac{\frac{\partial^2}{\partial \theta^2} f(X; \theta)}{f(X; \theta)} - \frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \quad (4.6.14)$$

Additionally, we can show

$$\mathbb{E}_X \left[ \frac{\frac{\partial^2}{\partial \theta^2} f(X; \theta)}{f(X; \theta)} \right] = \int_{\mathcal{X}} \frac{\frac{\partial^2}{\partial \theta^2} f(x; \theta)}{f(x; \theta)} \cdot f(x; \theta) dx \quad (4.6.15)$$

$$= \int_{\mathcal{X}} \frac{\partial^2}{\partial \theta^2} f(x; \theta) dx \quad (4.6.16)$$

$$= \frac{\partial^2}{\partial \theta^2} \left( \int_{\mathcal{X}} f(x; \theta) dx \right) \quad (4.6.17)$$

$$= \frac{\partial^2}{\partial \theta^2} (1) \quad (4.6.18)$$

$$= 0 \quad (4.6.19)$$

Returning to the Fisher information

$$\mathcal{I}(\theta) = \mathbb{E}_X \left[ \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \right] \quad (4.6.20)$$

$$= \mathbb{E}_X \left[ \frac{\frac{\partial^2}{\partial \theta^2} f(X; \theta)}{f(X; \theta)} - \frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right] \quad (4.6.21)$$

$$= \mathbb{E}_X \left[ \frac{\frac{\partial^2}{\partial \theta^2} f(X; \theta)}{f(X; \theta)} \right] - \mathbb{E}_X \left[ \frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right] \quad (4.6.22)$$

$$= -\mathbb{E}_X \left[ \frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right] \quad (4.6.23)$$

This shows that the Fisher information is equal to the negative of the curvature of the log likelihood, assuming the log likelihood is twice differentiable with respect to  $\theta$ .

### Fisher Information Matrix

If  $\theta$  is a vector, then the Fisher information is a matrix given by

$$\mathcal{I}(\theta) = \mathbb{E}_X \left[ (\nabla_\theta \log f(X; \theta)) (\nabla_\theta \log f(X; \theta))^\top \right] \quad (4.6.24)$$

$$= -\mathbb{E}_X [\nabla_\theta^2 \log f(X; \theta)] \quad (4.6.25)$$

#### 4.6.3 Observed Fisher Information

The observed Fisher information (sometimes also referred to as the empirical Fisher information) is a sample-based version of Fisher information. As the Fisher information is the expectation of the negative Hessian of the log-likelihood of a single sample, the observed Fisher information is the sample mean of the negative Hessian of the log-likelihood for a sample of size  $n$ . The observed information evaluated at  $\theta^*$  can be computed by

$$\mathcal{J}(\theta^*) = -\frac{1}{n} \sum_{i=1}^n \nabla_\theta^2 \log \mathcal{L}(\theta; X_i) \Big|_{\theta=\theta^*} \quad (4.6.26)$$

#### 4.6.4 Fisher Information in Linear Regression

In a linear regression model with  $y_i = x_i^\top \beta + \sigma \varepsilon_i$  where  $\varepsilon \sim \mathcal{N}(0, 1)$ , the log likelihood of an observation  $(x_i, y_i)$  is given by

$$\log \mathcal{L}(\beta; x_i, y_i) = \log \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2} \cdot \frac{(y_i - x_i^\top \beta)^2}{\sigma^2} \right) \right] \quad (4.6.27)$$

$$= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2} \cdot \frac{(y_i - x_i^\top \beta)^2}{\sigma^2} \quad (4.6.28)$$

The gradient (using the chain rule) is given by

$$\nabla_\beta \log \mathcal{L}(\beta; x_i, y_i) = -\frac{1}{2\sigma^2} \left[ - (y_i - x_i^\top \beta) x_i \right] \quad (4.6.29)$$

$$= \frac{(y_i - x_i^\top \beta) x_i}{2\sigma^2} \quad (4.6.30)$$

Which gives the Hessian

$$\nabla_\beta^2 \log \mathcal{L}(\beta; x_i, y_i) = -\frac{2x_i x_i^\top}{2\sigma^2} \quad (4.6.31)$$

$$= -\frac{x_i x_i^\top}{\sigma^2} \quad (4.6.32)$$

The Fisher information is the negative of the expected Hessian, so

$$\mathcal{I}(\beta) = \frac{1}{\sigma^2} \mathbb{E} [x_i x_i^\top] \quad (4.6.33)$$

Suppose there are  $N$  i.i.d. observations  $(x_i, y_i)$ , then the log likelihood is simply the sum of individual log likelihoods, and so the Fisher information is also just the sum:

$$\mathcal{I}(\beta) = \frac{1}{\sigma^2} \mathbb{E} \left[ \sum_{i=1}^N x_i x_i^\top \right] \quad (4.6.34)$$

We can also define  $X$  as a (tall) matrix given by  $X := [x_1 \ \dots \ x_N]^\top$ , then this lets us write the Fisher information as

$$\mathcal{I}(\beta) = \frac{1}{\sigma^2} \mathbb{E} \left[ [x_1 \ \dots \ x_N] \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} \right] \quad (4.6.35)$$

$$= \frac{\mathbb{E}[X^\top X]}{\sigma^2} \quad (4.6.36)$$

Note that if  $X$  is a ‘design matrix’, that is the factors in  $X$  can be freely chosen, then the expectation is redundant and the Fisher information (also known as simply the *information matrix* [10]) is

$$\mathcal{I}(\beta) = \frac{X^\top X}{\sigma^2} \quad (4.6.37)$$

#### 4.6.5 Cramer-Rao Bound

Let  $\hat{\theta}$  be an unbiased estimator for the scalar parameter  $\theta$ . Let  $V$  be the score of the log likelihood  $\log f(X; \theta)$ . Then the covariance between  $V$  and  $\hat{\theta}$  satisfies

$$\text{Cov}(V, \hat{\theta}) = \mathbb{E}[V\hat{\theta}] \quad (4.6.38)$$

since we know the expected value of the score  $\mathbb{E}[V] = 0$ . Then

$$\mathbb{E}[V\hat{\theta}] = \int_{\mathcal{X}} f(x; \theta) V \hat{\theta} dx \quad (4.6.39)$$

$$= \int_{\mathcal{X}} f(x; \theta) \frac{\partial f(x; \theta)}{\partial \theta} \hat{\theta} dx \quad (4.6.40)$$

$$= \int_{\mathcal{X}} \frac{\partial f(x; \theta)}{\partial \theta} \hat{\theta} dx \quad (4.6.41)$$

$$= \frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(x; \theta) \hat{\theta} dx \quad (4.6.42)$$

$$= \frac{\partial}{\partial \theta} \mathbb{E}[\hat{\theta}] \quad (4.6.43)$$

$$= \frac{\partial}{\partial \theta} \theta \quad (4.6.44)$$

$$= 1 \quad (4.6.45)$$

Hence  $\text{Cov}(V, \hat{\theta}) = \mathbb{E}[V\hat{\theta}] = 1$ . Then from  $\text{Corr}(V, \hat{\theta}) \leq 1$ :

$$\text{Cov}(V, \hat{\theta}) \leq \sqrt{\text{Var}(V)} \sqrt{\text{Var}(\hat{\theta})} \quad (4.6.46)$$

$$1 \leq \sqrt{\text{Var}(V)} \sqrt{\text{Var}(\hat{\theta})} \quad (4.6.47)$$

$$\text{Var}(V) \text{Var}(\hat{\theta}) \geq 1 \quad (4.6.48)$$

Then from the fact that the Fisher information  $\mathcal{I}(\theta)$  is the variance of the score, this gives the Cramer-Rao bound:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta)} \quad (4.6.49)$$

In the case  $\theta$  is a vector, then  $\text{Cov}(\hat{\theta})$  is a covariance matrix and the Cramer-Rao bound generalises to

$$\text{Cov}(\hat{\theta}) \succeq \mathcal{I}(\theta)^{-1} \quad (4.6.50)$$

#### 4.6.6 Efficient Estimators

An estimator  $\hat{\theta}$  that achieves the Cramer-Rao bound with equality, i.e.

$$\text{Cov}(\hat{\theta}) = \mathcal{I}(\theta)^{-1} \quad (4.6.51)$$

is said to be efficient. This implies that it is an unbiased estimator with the ‘least variance’. The maximum likelihood estimator can be shown to be asymptotically efficient (i.e. as data size  $n \rightarrow \infty$ ).

#### Efficiency of Maximum Likelihood Estimator

#### 4.6.7 Asymptotic Normality of Maximum Likelihood Estimator

The maximum likelihood estimate  $\hat{\theta}$  is defined as maximising the log-likelihood  $\log \mathcal{L}(\theta)$ , where we have suppressed the notation of including the sample  $\mathbf{X}$  in the log-likelihood. Hence the gradient of the log-likelihood satisfies

$$\nabla_{\theta} \log \mathcal{L}(\hat{\theta}) = 0 \quad (4.6.52)$$

As established with the asymptotic consistency of the maximum likelihood estimator, for large  $n$ ,  $\hat{\theta}$  will be ‘close’ to the true parameter  $\theta_0$  with high probability. So by applying a first order Taylor expansion of  $\nabla_{\theta} \log \mathcal{L}(\hat{\theta})$  about  $\theta_0$ , we have

$$\nabla_{\theta} \log \mathcal{L}(\hat{\theta}) \approx \nabla_{\theta} \log \mathcal{L}(\theta_0) + \nabla_{\theta}^2 \log \mathcal{L}(\theta_0) (\hat{\theta} - \theta_0) \quad (4.6.53)$$

Hence

$$\nabla_{\theta} \log \mathcal{L}(\theta_0) + \nabla_{\theta}^2 \log \mathcal{L}(\theta_0) (\hat{\theta} - \theta_0) \approx 0 \quad (4.6.54)$$

$$\hat{\theta} - \theta_0 \approx -[\nabla_{\theta}^2 \log \mathcal{L}(\theta_0)]^{-1} \nabla_{\theta} \log \mathcal{L}(\theta_0) \quad (4.6.55)$$

Multiplying out both sides by  $\sqrt{n}$  (as will be convenient for later on):

$$\sqrt{n} (\hat{\theta} - \theta_0) \approx - \left[ \frac{1}{n} \nabla_{\theta}^2 \log \mathcal{L}(\theta_0) \right]^{-1} \frac{1}{\sqrt{n}} \nabla_{\theta} \log \mathcal{L}(\theta_0) \quad (4.6.56)$$

For an i.i.d. sample  $X_1, \dots, X_n$ , we have

$$-\frac{1}{n} \nabla_{\theta}^2 \log \mathcal{L}(\theta_0) = -\frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 \log f(X_i; \theta) \Big|_{\theta=\theta_0} \quad (4.6.57)$$

Recognising that this term is observed Fisher information evaluated at  $\theta_0$ , we have by the Law of Large numbers

$$-\frac{1}{n} \nabla_\theta^2 \log \mathcal{L}(\theta_0) \xrightarrow{\text{P}} \mathcal{I}(\theta_0) \quad (4.6.58)$$

where recall that  $\mathcal{I}(\theta) = -\mathbb{E} [\nabla_\theta^2 \log f(X; \theta)]$ . Next for the term

$$\frac{1}{\sqrt{n}} \nabla_\theta \log \mathcal{L}(\theta_0) = -\frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\theta \log f(X_i; \theta) \Big|_{\theta=\theta_0} \quad (4.6.59)$$

recall that the expectation and covariance of the score (derivative of the log-likelihood for a random  $X$ ) satisfy

$$\mathbb{E} [\nabla_\theta \log f(X; \theta)] = 0 \quad (4.6.60)$$

$$\text{Cov} (\nabla_\theta \log f(X; \theta)) = \mathcal{I}(\theta) \quad (4.6.61)$$

So by the multivariate Central Limit Theorem, this gives

$$\frac{1}{\sqrt{n}} \nabla_\theta \log \mathcal{L}(\theta_0) = \sqrt{n} \left( \frac{1}{n} \nabla_\theta \log \mathcal{L}(\theta_0) \right) \quad (4.6.62)$$

$$\xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{I}(\theta_0)) \quad (4.6.63)$$

Now by the continuous mapping theorem,

$$\left[ -\frac{1}{n} \nabla_\theta^2 \log \mathcal{L}(\theta_0) \right]^{-1} \xrightarrow{\text{P}} \mathcal{I}(\theta_0)^{-1} \quad (4.6.64)$$

and combining this with Slutsky's theorem yields

$$-\left[ \frac{1}{n} \nabla_\theta^2 \log \mathcal{L}(\theta_0) \right]^{-1} \frac{1}{\sqrt{n}} \nabla_\theta \log \mathcal{L}(\theta_0) \xrightarrow{\text{P}} \mathcal{I}(\theta_0)^{-1} \left( \frac{1}{\sqrt{n}} \nabla_\theta \log \mathcal{L}(\theta_0) \right) \quad (4.6.65)$$

$$\xrightarrow{\text{d}} \mathcal{N} \left( 0, \mathcal{I}(\theta_0)^{-1} \underbrace{\mathcal{I}(\theta_0)}_{\mathcal{I}(\theta_0)^{-1}} \left[ \mathcal{I}(\theta_0)^{-1} \right]^\top \right) \quad (4.6.66)$$

And for large  $n$ , we can ignore the higher order terms in the Taylor approximation and say that

$$\sqrt{n} (\hat{\theta} - \theta_0) \xrightarrow{\text{d}} \mathcal{N} \left( 0, \mathcal{I}(\theta_0)^{-1} \right) \quad (4.6.67)$$

Therefore this shows the asymptotic normality of the maximum likelihood estimator.

#### 4.6.8 Standard Errors for Maximum Likelihood Estimator

The asymptotic normality of the maximum likelihood estimator can be used to obtain standard errors for  $\hat{\theta}$ , which can be used to construct confidence regions and conduct hypothesis tests. For large  $n$ , we will have

$$\hat{\theta} \xrightarrow{\text{approx.}} \mathcal{N} \left( \theta_0, \frac{1}{n} \mathcal{I}(\theta_0)^{-1} \right) \quad (4.6.68)$$

Since the population parameter  $\theta_0$  is treated as unknown and because  $\hat{\theta}$  is consistent for  $\theta_0$  then it is reasonable to replace  $\theta_0$  with the maximum likelihood estimate  $\hat{\theta}$  so that

$$\hat{\theta} \xrightarrow{\text{approx.}} \mathcal{N} \left( \theta_0, \frac{1}{n} \mathcal{I}(\hat{\theta})^{-1} \right) \quad (4.6.69)$$

It may be that calculating the Fisher information (which requires expectation of the negative Hessian of the log-likelihood) can be tricky, so we can instead replace the Fisher information by the observed Fisher information, which will be ‘close’ for large  $n$ :

$$-\frac{1}{n} \nabla_{\theta}^2 \log \mathcal{L}(\hat{\theta}) \approx \mathcal{I}(\hat{\theta}) \quad (4.6.70)$$

and moreover we can rearrange this to

$$n \left( -\nabla_{\theta}^2 \log \mathcal{L}(\hat{\theta}) \right)^{-1} \approx \mathcal{I}(\hat{\theta})^{-1} \quad (4.6.71)$$

$$\left( -\nabla_{\theta}^2 \log \mathcal{L}(\hat{\theta}) \right)^{-1} \approx \frac{1}{n} \mathcal{I}(\hat{\theta})^{-1} \quad (4.6.72)$$

Substituting this in the covariance gives

$$\hat{\theta} \stackrel{\text{approx.}}{\sim} \mathcal{N}\left(\theta_0, \left( -\nabla_{\theta}^2 \log \mathcal{L}(\hat{\theta}) \right)^{-1} \right) \quad (4.6.73)$$

This means we can take the inverse of the second derivative (i.e. curvature) of the objective function being minimised at the maximum likelihood estimate itself, and use this to approximate the covariance of the estimate. Intuitively, a ‘flatter’ curvature means the inverse is ‘larger’, so the uncertainty is greater. Conversely, a more ‘pointed’ curvature means the inverse is ‘smaller’ indicating less uncertainty of the estimate. Furthermore, it could be that computing the analytical second derivative itself can be tricky. In this case, the second derivative can be replaced by a numerical estimate of the second derivative (via a method such as numerical differentiation). This yields

$$\hat{\theta} \stackrel{\text{approx.}}{\sim} \mathcal{N}\left(\theta_0, \left( -\hat{\nabla}_{\theta}^2 \log \mathcal{L}(\hat{\theta}) \right)^{-1} \right) \quad (4.6.74)$$

One possible way to approximate the Hessian is via the ‘outer-product of gradients’ from:

$$\nabla_{\theta}^2 \log \mathcal{L}(\hat{\theta}; \mathbf{X}) \approx \sum_{i=1}^n \left( \nabla_{\theta} \log \mathcal{L}(\hat{\theta}; X_i) \right) \left( \nabla_{\theta} \log \mathcal{L}(\hat{\theta}; X_i) \right)^{\top} \quad (4.6.75)$$

The rationale behind this is because of the property that the Hessian is equal to the covariance of the score; thus we attempt to estimate the covariance of the score. Another justification for doing this is the same as for using the outer-product to approximate the Hessian in the Gauss-Newton algorithm. In all the above, we have reached an asymptotic approximation of the sampling distribution for the maximum likelihood estimator which takes the form  $\hat{\theta} \stackrel{\text{approx.}}{\sim} \mathcal{N}(\theta_0, \Sigma_{\hat{\theta}})$ . Regardless of whichever method is used to obtain  $\Sigma_{\hat{\theta}}$ , the standard error of the estimate for the  $j^{\text{th}}$  parameter is the usual square root of the  $j^{\text{th}}$  diagonal of  $\Sigma_{\hat{\theta}}$ :

$$\text{se}(\hat{\theta}_j) = \sqrt{\Sigma_{\hat{\theta},jj}} \quad (4.6.76)$$

### 4.6.9 Hypothesis Testing for Maximum Likelihood Estimation [136]

**Wilks' Theorem**

**Likelihood Ratio Test**

**Lagrange Multiplier Test**

## 4.7 Multiple Hypothesis Testing

### 4.7.1 Multiple Competing Hypothesis Testing

#### Maximum Likelihood Multiple Hypothesis Testing

The maximum likelihood binary hypothesis testing rule can be extended to the multiple hypothesis case. Consider multiple competing (i.e. mutually exclusive) hypotheses  $H_0, \dots, H_{M-1}$ . We decide  $H_m$  if

$$p(x|H_m) \geq p(x|H_j) \quad (4.7.1)$$

for all  $j \neq m$ , where  $p(x|H_j)$  is the likelihood of the  $j^{\text{th}}$  hypothesis. At a discrete level, this is essentially the idea behind maximum likelihood estimation, where each hypothesis could represent the parameter taking on some value from a finite set.

#### Maximum a Posteriori Multiple Hypothesis Testing

To extend the maximum a posteriori hypothesis testing decision rule to multiple competing hypotheses  $H_0, \dots, H_{M-1}$ , we decide  $H_m$  if

$$p(H_m|x) \geq p(H_j|x) \quad (4.7.2)$$

for all  $j \neq m$ , where  $p(H_j|x)$  is the posterior of the  $j^{\text{th}}$  hypothesis.

### 4.7.2 Multiple Simultaneous Hypothesis Testing

Consider a family of hypotheses  $H_1, \dots, H_M$ . These need not be mutually exclusive hypotheses (allowing for them all to be true simultaneously). We wish to test the null hypothesis that all of  $H_1, \dots, H_M$  are true, against the alternative that at least one of them is false. Suppose we have a method for testing each individual hypothesis separately, at the  $\alpha$  level of significance. However, this  $\alpha$  is not applicable for testing the family of hypothesis together, i.e. if we reject at least one of the hypothesis individually each at the  $\alpha$  level, this does not mean we reject the family-wise null at the  $\alpha$  level. Intuitively, this is because the more hypotheses we test, the more likely that at least one will be rejected, so we cannot control for the probability of a Type I error. Therefore, we need to adjust the  $\alpha$  when testing the family-wise hypotheses.

#### Šidák Correction

The Šidák correction is a method of adjusting  $\alpha$  for testing a family-wise null hypothesis. This is done by assuming that the result of testing each individual hypothesis  $H_1, \dots, H_M$  is mutually independent. Let  $p_1, \dots, p_M$  denote the respective  $p$ -values, so that under the null, the probability of rejection at the  $\alpha$  level of significance is given by

$$\Pr(p_j \leq \alpha|H_j) = \alpha \quad (4.7.3)$$

so the probability of non-rejection is

$$\Pr(p_j > \alpha|H_j) = 1 - \alpha \quad (4.7.4)$$

With the independence assumption, this means

$$\Pr \left( \bigcap_{j=1}^M \{p_j > \alpha\} \middle| H_1, \dots, H_M \right) = (1 - \alpha)^M \quad (4.7.5)$$

and applying DeMorgan's laws, the probability that at least one null is rejected (also known as the *family-wise error rate*) is

$$\Pr \left( \bigcup_{j=1}^M \{p_j \leq \alpha\} \middle| H_1, \dots, H_M \right) = 1 - (1 - \alpha)^M \quad (4.7.6)$$

If we wish to control the family-wise error rate at level  $\alpha_M$ , we equate

$$\alpha_M = 1 - (1 - \alpha)^M \quad (4.7.7)$$

and solve for  $\alpha$  to obtain

$$\alpha = 1 - (1 - \alpha_M)^{1/M} \quad (4.7.8)$$

Thus to test the hypothesis that all of  $H_1, \dots, H_M$  at level  $\alpha_M$  using the Šidák correction, we test each hypothesis individually at level  $\alpha = 1 - (1 - \alpha_M)^{1/M}$ , and reject the family-wise null if at least one of the individual null are rejected. Note that for  $M = 1$ , we have  $\alpha = \alpha_M$  and for  $M > 1$ , we have  $\alpha < \alpha_M$ . Hence we need to test each individual at a more strict level if we want to control the family-wise error rate.

### Bonferroni Correction

The Bonferroni correction is another method for adjusting  $\alpha$  when testing all the hypotheses  $H_1, \dots, H_M$  simultaneously, which unlike the Šidák correction does not require any assumption about the nature of dependence between each of the tests. Let  $p_1, \dots, p_M$  denote the respective *p*-values. The family-wise error rate (probability of rejecting at least one null) can be upper-bounded with Boole's inequality by

$$\Pr \left( \bigcup_{j=1}^M \{p_j \leq \alpha\} \middle| H_1, \dots, H_M \right) \leq \sum_{j=1}^M \Pr(p_j \leq \alpha | H_j) \quad (4.7.9)$$

$$= M\alpha \quad (4.7.10)$$

If we wish to control the family-wise error rate at level  $\alpha_M$ , we equate

$$\alpha_M = M\alpha \quad (4.7.11)$$

and solve for  $\alpha$  to obtain

$$\alpha = \frac{\alpha_M}{M} \quad (4.7.12)$$

Thus to test the hypothesis that all of  $H_1, \dots, H_M$  at level  $\alpha_M$  using the Bonferroni correction, we test each hypothesis individually at level  $\alpha = \alpha_M/M$ , and reject the family-wise null if at least one of the individual null are rejected. Although this correction procedure is quite general, it can also be conservative (especially as the number of hypotheses increases) because we ensure that the family-wise error rate is at most  $\alpha_M$ , but in reality it could be much smaller than that. Performing each individual test at level  $\alpha_M/M$  means that the statistical power will also be reduced.

### 4.7.3 Multiple Comparisons [189]

Let  $\theta_1, \dots, \theta_K$  be different parameters, with the goal of identifying which of the parameters, if any, is greater than the others (or smaller than the others, but we can assume the former without loss of generality). Suppose we can conduct inference on each the parameters, and compare them pairwise via two-sample inference. Then to compare the family of parameters, we can conduct a series of two-sample tests, and use appropriate  $\alpha$  adjustment.

#### Many-to-Many Comparisons

Let the null hypothesis be that all the parameters are the same  $\theta_1 = \dots = \theta_K$  against the alternative that there are at least two parameters that are not equal. Then we can perform  $M = \binom{K}{2} = K(K - 1)/2$  simultaneous two-sample two-tailed tests given by the hypotheses:

$$H_1 : \theta_1 = \theta_2 \quad (4.7.13)$$

$$\vdots \quad (4.7.14)$$

$$H_M : \theta_{K-1} = \theta_K \quad (4.7.15)$$

and use  $\alpha$  adjustment, such as the Bonferroni correction, whereas the Šidák correction will not be suitable because the outcome of each test will not be independent.

#### Many-to-One Comparisons

Suppose we have a ‘hunch’ that one particular parameter, say  $\theta_\kappa$ , is greater than the others. We test the joint hypothesis that  $\theta_\kappa = \theta_k$  for all  $k \neq \kappa$ , against the alternative that there is at least one  $k$  such that  $\theta_\kappa > \theta_k$ . Instead of  $K(K - 1)/2$  tests in many-to-many comparisons, we can test the  $M = K - 1$  hypotheses

$$H_1 : \theta_\kappa = \theta_1 \quad (4.7.16)$$

$$\vdots \quad (4.7.17)$$

$$H_M : \theta_\kappa = \theta_K \quad (4.7.18)$$

using one-tailed tests (i.e. each test is against the alternative  $\theta_\kappa > \theta_K$ ). We also use  $\alpha$  adjustment to control for the family-wise error rate.

### 4.7.4 Analysis of Variance [169]

#### Linear Contrasts

#### Multivariate Analysis of Variance

### 4.7.5 False Discovery Rate Control

## 4.8 Generalised Linear Models

### 4.8.1 Poisson Regression

### 4.8.2 Logistic Regression

When the response variable  $Y$  can take on a 0 or 1 (i.e. success or fail), the probability that the response equals 1 can be modelled using logistic regression. For a parameter vector  $\beta$  and predictors  $x$ , this probability takes the form

$$\Pr(Y = 1|X = x) = \frac{1}{1 + \exp(-\beta^\top x)} \quad (4.8.1)$$

An alternative form is

$$\Pr(Y = 1|X = x) = \frac{\exp(\beta^\top x)}{\exp(\beta^\top x) + 1} \quad (4.8.2)$$

Notice that this probability will be bounded between 0 and 1. Denoting  $p(x) := \Pr(Y = 1|X = x)$ , we can rearrange for a ‘linear form’, given as

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta^\top x \quad (4.8.3)$$

To fit a logistic regression to some data, we first construct the likelihood (assuming independent samples) as

$$\mathcal{L}(\beta; x) = \Pr(Y|X = x, \beta) \quad (4.8.4)$$

$$= \prod_{i=1}^n \Pr(y_i|X = x_i, \beta) \quad (4.8.5)$$

$$= \prod_{i=1}^n \left( \frac{1}{1 + \exp(-\beta^\top x_i)} \right)^{y_i} \left( 1 - \frac{1}{1 + \exp(-\beta^\top x_i)} \right)^{1-y_i} \quad (4.8.6)$$

Note that since  $y_i$  can only be 0 or 1, each multiplicand ‘selects’ either the probability of occurrence of  $\Pr(Y = 1|X = x, \beta)$  if  $y_i = 1$  or  $\Pr(Y = 0|X = x, \beta)$  if  $y_i = 0$ . The log-likelihood is

$$\log \mathcal{L}(\beta; x) = \sum_{i=1}^n y_i \log\left(\frac{1}{1 + \exp(-\beta^\top x_i)}\right) + \sum_{i=1}^n (1 - y_i) \log\left(1 - \frac{1}{1 + \exp(-\beta^\top x_i)}\right) \quad (4.8.7)$$

Hence the maximum likelihood estimate is the solution to

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{-\log \mathcal{L}(\beta; x)\} \quad (4.8.8)$$

### 4.8.3 Probit Regression

### 4.8.4 Multinomial Logistic Regression

### 4.8.5 Ordinal Regression

## 4.9 Quantile Regression

## Chapter 5

# Advanced Probability

## 5.1 Multivariate Gaussian Properties

### 5.1.1 Joint Moment Generating Function of Multivariate Gaussian

To obtain the expression for the joint moment generating function of a multivariate Gaussian, we first require the moment generating function of a standard Gaussian random variable  $Z$ . This is given by

$$\phi_Z(s) = \mathbb{E}[e^{sZ}] \quad (5.1.1)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sz} e^{-z^2/2} dz \quad (5.1.2)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z^2-2sz)/2} dz \quad (5.1.3)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-[(z-s)^2-s^2]/2} dz \quad (5.1.4)$$

$$= \frac{e^{s^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-s)^2/2} dz \quad (5.1.5)$$

$$= e^{s^2/2} \quad (5.1.6)$$

where we have recognised that  $\frac{1}{\sqrt{2\pi}}e^{-(z-s)^2/2}$  is the density of a  $\mathcal{N}(s, 1)$  random variable. Then a random vector of independent standard Gaussian random variables  $\mathbf{Z} = (Z_1, \dots, Z_n)$  has joint moment generating function

$$\phi_{\mathbf{Z}}(\mathbf{s}) = \prod_{i=1}^n e^{s_i^2/2} \quad (5.1.7)$$

$$= \exp\left(\frac{1}{2} \sum_{i=1}^n s_i^2\right) \quad (5.1.8)$$

$$= \exp\left(\frac{1}{2} \mathbf{s}^\top \mathbf{s}\right) \quad (5.1.9)$$

Using the characterisation that a multivariate Gaussian random vector  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  has the same distribution as  $\boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$ , then using properties for joint MGFs of linear transformations:

$$\phi_{\mathbf{X}}(\mathbf{s}) = \exp(\mathbf{s}^\top \boldsymbol{\mu}) \phi_{\mathbf{Z}}\left(\left(\boldsymbol{\Sigma}^{1/2}\right)^\top \mathbf{s}\right) \quad (5.1.10)$$

$$= \exp(\mathbf{s}^\top \boldsymbol{\mu}) \exp\left(\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Sigma}^{1/2} \left(\boldsymbol{\Sigma}^{1/2}\right)^\top \mathbf{s}\right) \quad (5.1.11)$$

$$= \exp(\mathbf{s}^\top \boldsymbol{\mu}) \exp\left(\frac{1}{2}\mathbf{s}^\top \boldsymbol{\Sigma}\mathbf{s}\right) \quad (5.1.12)$$

Then a special case for univariate Gaussian  $X \sim \mathcal{N}(\mu, \sigma^2)$  is  $\phi_X(s) = e^{s\mu + s^2\sigma^2/2}$ .

### 5.1.2 Uncorrelatedness and Independence of Multivariate Gaussians

For multivariate Gaussians, uncorrelatedness and independence are equivalent. Suppose that  $\mathbf{X} = (X_1, \dots, X_n)$  is a multivariate Gaussian vector with mean vector  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$  and with  $\text{Cov}(X_i, X_j) = 0$  for any  $i \neq j$ . This means that  $\mathbf{X}$  has a covariance matrix  $\mathbf{K}$  which only consists of a main diagonal:

$$\mathbf{K} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix} \quad (5.1.13)$$

Then  $X_1, \dots, X_n$  are independent.

*Proof.* The joint moment generating function of  $\mathbf{X}$  is given by

$$\phi_{\mathbf{X}}(\mathbf{s}) = \exp(\mathbf{s}^\top \boldsymbol{\mu}) \exp\left(\frac{1}{2}\mathbf{s}^\top \mathbf{K}\mathbf{s}\right) \quad (5.1.14)$$

Since  $\mathbf{K}$  is a diagonal matrix, this can be written as

$$\phi_{\mathbf{X}}(\mathbf{s}) = \exp\left(\sum_{i=1}^n s_i \mu_i\right) \exp\left(\frac{1}{2} \sum_{i=1}^n \sigma_i^2 s_i^2\right) \quad (5.1.15)$$

$$= \prod_{i=1}^n \exp(s_i \mu_i) \exp\left(\frac{1}{2} \sigma_i^2 s_i^2\right) \quad (5.1.16)$$

$$= \prod_{i=1}^n \phi_{X_i}(s_i) \quad (5.1.17)$$

which is the product of the individual MGFs of  $X_1, \dots, X_n$ . Hence this implies that  $X_1, \dots, X_n$  are independent.  $\square$

It also follows that bivariate Gaussian random variables  $(X, Y)$  being uncorrelated is equivalent to mean independence both ways, i.e.  $\mathbb{E}[X|Y] = \mathbb{E}[X]$  and  $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ .

### 5.1.3 Marginal Gaussians and Joint Gaussians

If  $X_1, \dots, X_n$  are dependent Gaussian random variables (i.e. the marginal distributions are Gaussian densities), this does not necessarily imply that  $\mathbf{X} = (X_1, \dots, X_n)$  is a multivariate Gaussian random vector.

*Proof.* Consider the following counterexample. Let  $X$  be a zero-mean Gaussian random variable and let  $S$  have the distribution

$$\Pr(S = s) = \begin{cases} 1/2, & s = -1, 1 \\ 0, & \text{elsewhere} \end{cases} \quad (5.1.18)$$

Now define  $Y = SX$ , i.e.  $Y$  will either be  $X$  or the negative of  $X$ . Since  $X$  is symmetric and zero-centered, then the marginal distribution of  $Y$  will also be Gaussian. However, the joint distribution of  $X$  and  $Y$  (written using Dirac delta distributions) is given by

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) f_X(x) \quad (5.1.19)$$

$$= \left( \frac{1}{2} \delta(y-x) + \frac{1}{2} \delta(y+x) \right) f_X(x) \quad (5.1.20)$$

which is not a bivariate Gaussian density.  $\square$

Other counterexamples can be constructed using **copulae**. A multivariate distribution with Gaussian marginals is multivariate Gaussian if and only if it has a **Gaussian copula**. So any multivariate distribution with Gaussian marginals but a non-Gaussian copula will not be multivariate Gaussian.

However of course, if  $X_1, \dots, X_n$  are independent Gaussian random variables then  $\mathbf{X} = (X_1, \dots, X_n)$  is going to be a multivariate Gaussian random vector.

### 5.1.4 Conditional Gaussian Densities

Recall that a multivariate  $D$ -dimensional Gaussian distribution for random vector  $X$  can be denoted by the density function

$$p(x; \mu, \Sigma) = (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp \left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right) \quad (5.1.21)$$

where  $\mu$  and  $\Sigma$  are the mean and covariance respectively. Here,  $|\cdot|$  denotes determinant. We can also denote a multivariate Gaussian using  $X \sim \mathcal{N}(\mu, \Sigma)$ . For a multivariate distribution between two Gaussian random vectors  $X$  and  $Y$ , we can denote it as

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_{YY} \end{bmatrix} \right) \quad (5.1.22)$$

$$\sim \mathcal{N} \left( \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \tilde{\Sigma}_{XX} & \tilde{\Sigma}_{XY} \\ \tilde{\Sigma}_{XY}^\top & \tilde{\Sigma}_{YY} \end{bmatrix}^{-1} \right) \quad (5.1.23)$$

where  $X \sim \mathcal{N}(\mu_X, A)$ ,  $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$  and the cross-covariance  $\text{Cov}(X, Y) = \Sigma_{XY}$ . The inverse of the block partitioned matrix can be expressed using the block inversion formula:

$$\begin{aligned} & \begin{bmatrix} \tilde{\Sigma}_{XX} & \tilde{\Sigma}_{XY} \\ \tilde{\Sigma}_{XY}^\top & \tilde{\Sigma}_{YY} \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{XX}^{-1} + \Sigma_{XX}^{-1} \Sigma_{XY} (\Sigma_{YY} - \Sigma_{XY}^\top \Sigma_{XX}^{-1} \Sigma_{XY})^{-1} \Sigma_{XY}^\top \Sigma_{XX}^{-1} & -\Sigma_{XX}^{-1} \Sigma_{XY} (\Sigma_{YY} - \Sigma_{XY}^\top \Sigma_{XX}^{-1} \Sigma_{XY})^{-1} \\ -(\Sigma_{YY} - \Sigma_{XY}^\top \Sigma_{XX}^{-1} \Sigma_{XY})^{-1} \Sigma_{XY}^\top \Sigma_{XX}^{-1} & (\Sigma_{YY} - \Sigma_{XY}^\top \Sigma_{XX}^{-1} \Sigma_{XY})^{-1} \end{bmatrix} \end{aligned} \quad (5.1.24)$$

or alternatively:

$$\begin{aligned} & \begin{bmatrix} \tilde{\Sigma}_{XX} & \tilde{\Sigma}_{XY} \\ \tilde{\Sigma}_{XY}^\top & \tilde{\Sigma}_{YY} \end{bmatrix} \\ &= \begin{bmatrix} (\Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top)^{-1} & -(\Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top)^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \\ -\Sigma_{XY}^\top \Sigma_{YY}^{-1} (\Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top)^{-1} & \Sigma_{YY}^{-1} + \Sigma_{YY}^{-1} \Sigma_{XY}^\top (\Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top)^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \end{bmatrix} \end{aligned} \quad (5.1.25)$$

If we desire the conditional distribution of  $X$  given  $Y$ , this is another multivariate Gaussian with:

$$[X | Y = y] \sim \mathcal{N} \left( \mu_X + \Sigma_{XY} \Sigma_{YY}^{-1} (y - \mu_Y), \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right) \quad (5.1.26)$$

$$\sim \mathcal{N} \left( \mu_X - \tilde{\Sigma}_{XX}^{-1} \tilde{\Sigma}_{XY} (y - \mu_Y), \tilde{\Sigma}_{XX}^{-1} \right) \quad (5.1.27)$$

*Proof.* The random vector  $(X, Y)$  has joint density:

$$f_{X,Y}(x, y) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} Q(x, y) \right] \quad (5.1.28)$$

with

$$Q(x, y) = [x - \mu_X \ y - u_Y] \Sigma^{-1} \begin{bmatrix} x - \mu_X \\ y - u_Y \end{bmatrix} \quad (5.1.29)$$

$$= [x - \mu_X \ y - u_Y] \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_{YY} \end{bmatrix}^{-1} \begin{bmatrix} x - \mu_X \\ y - u_Y \end{bmatrix} \quad (5.1.30)$$

$$= [x - \mu_X \ y - u_Y] \begin{bmatrix} \tilde{\Sigma}_{XX} & \tilde{\Sigma}_{XY} \\ \tilde{\Sigma}_{XY}^\top & \tilde{\Sigma}_{YY} \end{bmatrix} \begin{bmatrix} x - \mu_X \\ y - u_Y \end{bmatrix} \quad (5.1.31)$$

$$= (x - \mu_X)^\top \tilde{\Sigma}_{XX} (x - \mu_X) + 2(y - u_Y)^\top \tilde{\Sigma}_{XY}^\top (x - \mu_X) + (y - u_Y)^\top \tilde{\Sigma}_{YY} (y - u_Y) \quad (5.1.32)$$

Substitute the above expressions from the block matrix inversion:

$$\tilde{\Sigma}_{XX} = \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} \quad (5.1.33)$$

$$\tilde{\Sigma}_{YY} = \Sigma_{YY}^{-1} + \Sigma_{YY}^{-1} \Sigma_{XY}^\top \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \quad (5.1.34)$$

$$\tilde{\Sigma}_{XY}^\top = -\Sigma_{XY}^\top \Sigma_{YY}^{-1} \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} \quad (5.1.35)$$

This yields:

$$\begin{aligned} Q(x, y) &= (x - \mu_X)^\top \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} (x - \mu_X) \\ &\quad - 2(y - u_Y)^\top \left[ \Sigma_{XY}^\top \Sigma_{YY}^{-1} \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} \right] (x - \mu_X) \\ &\quad + (y - u_Y)^\top \left[ \Sigma_{YY}^{-1} + \Sigma_{YY}^{-1} \Sigma_{XY}^\top \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \right] (y - u_Y) \end{aligned} \quad (5.1.36)$$

Splitting out the last term:

$$\begin{aligned} Q(x, y) &= (x - \mu_X)^\top \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} (x - \mu_X) \\ &\quad - 2(y - u_Y)^\top \left[ \Sigma_{XY}^\top \Sigma_{YY}^{-1} \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} \right] (x - \mu_X) \\ &\quad + (y - u_Y)^\top \left[ \Sigma_{YY}^{-1} \Sigma_{XY}^\top \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \right] (y - u_Y) \\ &\quad + (y - u_Y)^\top \Sigma_{YY}^{-1} (y - u_Y) \end{aligned} \quad (5.1.37)$$

Note that we can factorise the first three terms so that:

$$\begin{aligned} Q(x, y) &= [(x - \mu_X) - \Sigma_{XY} \Sigma_{YY}^{-1} (y - u_Y)]^\top \left( \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top \right)^{-1} [(x - \mu_X) - \Sigma_{XY} \Sigma_{YY}^{-1} (y - u_Y)] \\ &\quad + (y - u_Y)^\top \Sigma_{YY}^{-1} (y - u_Y) \end{aligned} \quad (5.1.38)$$

and write this as

$$Q(x, y) = Q_1(x, y) + Q_2(x, y) \quad (5.1.39)$$

where

$$Q_1(x, y) = [(x - \mu_X) - \Sigma_{XY}\Sigma_{YY}^{-1}(y - u_Y)]^\top (\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top)^{-1} [(x - \mu_X) - \Sigma_{XY}\Sigma_{YY}^{-1}(y - u_Y)] \quad (5.1.40)$$

$$Q_2(x, y) = (y - u_Y)^\top \Sigma_{YY}^{-1}(y - u_Y) \quad (5.1.41)$$

Now suppose that  $X$  has dimension  $p$  and  $Y$  has dimension  $q$  so that  $p + q = D$ . Then the joint density can be written as

$$f_{X,Y}(x, y) = \frac{1}{(2\pi)^{(p+q)/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} Q_1(x, y) \right] \exp \left[ -\frac{1}{2} Q_2(x, y) \right] \quad (5.1.42)$$

The determinant of  $\Sigma$  can also be split out as

$$|\Sigma| = \begin{vmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_{YY} \end{vmatrix} \quad (5.1.43)$$

$$= \begin{vmatrix} I & \Sigma_{XY} \\ 0 & \Sigma_{YY} \end{vmatrix} \begin{vmatrix} \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top & 0 \\ \Sigma_{YY}^{-1}\Sigma_{XY}^\top & I \end{vmatrix} \quad (5.1.44)$$

$$= \begin{vmatrix} I & \Sigma_{XY} \\ 0 & \Sigma_{YY} \end{vmatrix} \times \begin{vmatrix} \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top & 0 \\ \Sigma_{YY}^{-1}\Sigma_{XY}^\top & I \end{vmatrix} \quad (5.1.45)$$

$$= |\Sigma_{YY}| \times \left| \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top \right| \quad (5.1.46)$$

Hence splitting up the joint density:

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{(2\pi)^{p/2} |\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top|^{1/2}} \exp \left[ -\frac{1}{2} Q_1(x, y) \right] \\ &\quad \times \frac{1}{(2\pi)^{q/2} |\Sigma_{YY}|^{1/2}} \exp \left[ -\frac{1}{2} (y - u_Y)^\top \Sigma_{YY}^{-1}(y - u_Y) \right] \end{aligned} \quad (5.1.47)$$

Recognise the second term is the marginal density of  $Y$  so

$$f_{X,Y}(x, y) = \frac{1}{(2\pi)^{p/2} |\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top|^{1/2}} \exp \left[ -\frac{1}{2} Q_1(x, y) \right] f_Y(y) \quad (5.1.48)$$

By  $f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y)$ , it must be that the first density is the conditional density:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{\exp \left[ -\frac{1}{2} [(x - \mu_X) - \Sigma_{XY}\Sigma_{YY}^{-1}(y - u_Y)]^\top (\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top)^{-1} [(x - \mu_X) - \Sigma_{XY}\Sigma_{YY}^{-1}(y - u_Y)] \right]}{(2\pi)^{p/2} |\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top|^{1/2}} \end{aligned} \quad (5.1.49)$$

which is a multivariate Gaussian density with mean  $\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y)$  and covariance  $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top$ , therefore

$$[X|Y = y] \sim \mathcal{N} \left( \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top \right) \quad (5.1.50)$$

By also using  $\tilde{\Sigma}_{XX} = (\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top)^{-1}$  to give  $\tilde{\Sigma}_{XX}^{-1} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top$  and substituting this into  $\tilde{\Sigma}_{XY}^\top = -\Sigma_{XY}^\top\Sigma_{YY}^{-1}(\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^\top)^{-1}$  to obtain

$$-\tilde{\Sigma}_{XY}^\top\tilde{\Sigma}_{XX}^{-1} = -\Sigma_{XY}^\top\Sigma_{YY}^{-1} \quad (5.1.51)$$

then we also have

$$[X|Y = y] \sim \mathcal{N} \left( \mu_X - \tilde{\Sigma}_{XX}^{-1} \tilde{\Sigma}_{XY} (y - \mu_Y), \tilde{\Sigma}_{XX}^{-1} \right) \quad (5.1.52)$$

□

This result shows that conditional distributions of multivariate Gaussians are also multivariate Gaussians, and if equipped with this fact, it is much easier to derive the conditional density just by deriving the conditional mean and covariance, in the same way that was done when deriving the expression for partial correlations. In fact, it so happens that for the multivariate Gaussian family of distributions, the ‘best linear approximation’ conditional on a partition happens to be a Gaussian, so that the partial correlation coincides exactly with the conditional correlation. Also note that the conditional covariance  $\Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top$  is the Schur complement of the block  $\Sigma_{YY}$  in  $\Sigma$ .

### Degenerate Conditional Gaussian Distributions

Suppose we have a zero-mean (for simplicity) multivariate Gaussian random vector  $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$  where  $\mathbf{X}$  is  $n$ -dimensional and  $\Sigma$  is positive definite. Let  $A$  be some full rank  $m \times n$  matrix with  $m < n$  (i.e. it will be of rank  $m$ ). If we condition the distribution of  $\mathbf{X}$  on the event that  $A\mathbf{X} = 0$ , then it can be shown that the resulting conditional distribution will be **degenerate**. To do this, first recall that by using the properties of **linear transformations of random vectors**, we have  $\text{Cov}(A\mathbf{X}) = A\Sigma A^\top$  and  $\text{Cov}(\mathbf{X}, A\mathbf{X}) = \Sigma A^\top$ . Thus we may write the following augmented Gaussian random vector as

$$\begin{bmatrix} \mathbf{X} \\ A\mathbf{X} \end{bmatrix} \sim \mathcal{N} \left( 0, \begin{bmatrix} \Sigma & \Sigma A^\top \\ A\Sigma & A\Sigma A^\top \end{bmatrix} \right) \quad (5.1.53)$$

Conditioning on  $A\mathbf{X} = 0$  and applying conditioning formulae (but we could condition on values other than zero accordingly), we get

$$[\mathbf{X}|A\mathbf{X} = 0] \sim \mathcal{N} \left( 0, \Sigma - \Sigma A^\top (A\Sigma A^\top)^{-1} A\Sigma \right) \quad (5.1.54)$$

Consider this conditional covariance. The matrix  $A^\top (A\Sigma A^\top)^{-1} A$  will be of rank  $m$ , thus  $\Sigma A^\top (A\Sigma A^\top)^{-1} A\Sigma$  will also be of rank  $m$ . Therefore the conditional covariance will be of reduced rank  $n - m$ , since it is a rank  $m$  update applied to  $\Sigma$ . To further see why the rank is reduced, consider using the matrix inversion lemma to find the inverse of the conditional covariance:

$$(\Sigma - VCV^\top)^{-1} = \Sigma^{-1} - \Sigma^{-1}V(C^{-1} + V^\top \Sigma^{-1}V)^{-1}V^\top \Sigma^{-1} \quad (5.1.55)$$

with  $V = \Sigma A^\top$  and  $C = -(A\Sigma A^\top)^{-1}$ . However, observe that

$$C^{-1} + V^\top \Sigma^{-1}V = -A\Sigma A^\top + A\Sigma \Sigma^{-1} \Sigma A^\top \quad (5.1.56)$$

$$= 0 \quad (5.1.57)$$

which is not invertible, so we cannot take the inverse of conditional covariance, implying it is not full rank. Instead, the support lies on the  $(n - m)$ -dimensional hyperplane  $A\mathbf{X} = 0$ , hence the conditional distribution is degenerate. Furthermore, notice that since the inverse of the covariance appears in the multivariate Gaussian density, yet the conditional covariance cannot be inverted, this means that the conditional distribution  $[\mathbf{X}|A\mathbf{X} = 0]$  does not have a joint density. However, we can still draw from this distribution, e.g. by randomly drawing the first  $n - m$  elements of  $\mathbf{X}$ , then determining the remaining  $m$  elements by solving  $A\mathbf{X} = 0$

### 5.1.5 Product of Gaussian Densities

Suppose we take the product of two multivariate Gaussian densities. Note that this is different from the distribution of the product of multiple Gaussian random variables, which will not be Gaussian distributed. The product of two Gaussian densities will be another (un-normalised) Gaussian density:

$$\mathcal{N}(x; a, A) \mathcal{N}(x; b, B) = Z^{-1} \mathcal{N}(x; c, C) \quad (5.1.58)$$

where

$$C = (A^{-1} + B^{-1})^{-1} \quad (5.1.59)$$

$$c = C(A^{-1}a + B^{-1}b) \quad (5.1.60)$$

and the normalising factor looks like a Gaussian:

$$Z^{-1} = (2\pi)^{-D/2} |A + B|^{-1/2} \exp\left(-\frac{1}{2}(a - b)^\top (A + B)^{-1} (a - b)\right) \quad (5.1.61)$$

We can show this as follows. First replace the distributions using the definition of their density functions, and substitute the definitions of  $Z$ ,  $C$  and  $c$ . We get

$$\begin{aligned} & \frac{\exp\left[-\frac{1}{2}(x - a)^\top A^{-1}(x - a)\right]}{(2\pi)^{D/2}|A|^{1/2}} \cdot \frac{\exp\left[-\frac{1}{2}(x - b)^\top B^{-1}(x - b)\right]}{(2\pi)^{D/2}|B|^{1/2}} = \\ & \frac{\exp\left[-\frac{1}{2}(a - b)^\top (A + B)^{-1}(a - b)\right]}{(2\pi)^{D/2}|A + B|^{1/2}} \times \\ & \frac{\exp\left[-\frac{1}{2}\left(x - (A^{-1} + B^{-1})^{-1}(A^{-1}a + B^{-1}b)\right)^\top (A^{-1} + B^{-1})(x - (A^{-1} + B^{-1})^{-1}(A^{-1}a + B^{-1}b))\right]}{(2\pi)^{D/2}\left|(A^{-1} + B^{-1})^{-1}\right|^{1/2}} \end{aligned} \quad (5.1.62)$$

First we show equivalence over the denominators. Equating the denominators gives

$$(2\pi)^{D/2}|A|^{1/2}(2\pi)^{D/2}|B|^{1/2} = (2\pi)^{D/2}|A + B|^{1/2}(2\pi)^{D/2}\left|(A^{-1} + B^{-1})^{-1}\right|^{1/2} \quad (5.1.63)$$

$$|A|^{1/2}|B|^{1/2} = |A + B|^{1/2}\left|(A^{-1} + B^{-1})^{-1}\right|^{1/2} \quad (5.1.64)$$

$$|A||B| = |A + B|\left|(A^{-1} + B^{-1})^{-1}\right| \quad (5.1.65)$$

We can use a property of determinants that the determinant of the inverse is the inverse of the determinant. Hence

$$|A + B| = |A||B|\left|A^{-1} + B^{-1}\right| \quad (5.1.66)$$

We can apply the determinant form of the matrix inversion lemma:

$$\left|Z + UWV^\top\right| = |Z||W|\left|W^{-1} + V^\top Z^{-1}U\right| \quad (5.1.67)$$

(note that this  $Z$  is not the same as the one defined above). Letting  $Z = A$ ,  $W = B$ ,  $U = I$ ,  $V = I$ , we have

$$|A + B| = |A||B|\left|A^{-1} + B^{-1}\right| \quad (5.1.68)$$

Thus we get the same as above, hence the denominators are equal. We now begin the more tedious process of proving equivalence over the numerators. Grouping exponentials, the terms inside should be equal, i.e. (after taking out the  $-1/2$ )

$$\begin{aligned} & (x - a)^\top A^{-1} (x - a) + (x - b)^\top B^{-1} (x - b) \\ &= (a - b)^\top (A + B)^{-1} (a - b) + \\ & \left( x - (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \right)^\top (A^{-1} + B^{-1}) \left( x - (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \right) \end{aligned} \quad (5.1.69)$$

Expand the quadratics to get lengthy expressions for the LHS and RHS

$$\begin{aligned} RHS &= a^\top (A + B)^{-1} a - b^\top (A + B)^{-1} a - a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b + x^\top (A^{-1} + B^{-1}) x \\ &- \left[ (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \right]^\top (A^{-1} + B^{-1}) x - x^\top (A^{-1} + B^{-1}) (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \\ &+ \left[ (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \right]^\top (A^{-1} + B^{-1}) (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \end{aligned} \quad (5.1.70)$$

$$LHS = x^\top A^{-1} x - a^\top A^{-1} x - x^\top A^{-1} a + a^\top A^{-1} a + x^\top B^{-1} x - b^\top B^{-1} x - x^\top B^{-1} b + b^\top B^{-1} b \quad (5.1.71)$$

The highlighted terms in red cancel out. So we have (in addition to grouping some similar terms such as  $b^\top (A + B)^{-1} a$  and  $a^\top (A + B)^{-1} b$  together since they are scalar and  $A, B$  are symmetric):

$$\begin{aligned} RHS &= a^\top (A + B)^{-1} a - 2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b \\ &\quad - 2x^\top (A^{-1} + B^{-1}) (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \\ &+ (A^{-1}a + B^{-1}b)^\top (A^{-1} + B^{-1})^{-1} (A^{-1} + B^{-1}) (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \end{aligned} \quad (5.1.72)$$

$$LHS = -a^\top A^{-1} x - x^\top A^{-1} a + a^\top A^{-1} a - b^\top B^{-1} x - x^\top B^{-1} b + b^\top B^{-1} b \quad (5.1.73)$$

The highlighted terms in blue cancel out because they are the inverses of each other. Also by grouping similar terms in the LHS, we get

$$\begin{aligned} RHS &= a^\top (A + B)^{-1} a - 2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b \\ &\quad - 2x^\top (A^{-1}a + B^{-1}b) + (A^{-1}a + B^{-1}b)^\top (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \end{aligned} \quad (5.1.74)$$

$$LHS = -2x^\top A^{-1} a + a^\top A^{-1} a - 2x^\top B^{-1} b + b^\top B^{-1} b \quad (5.1.75)$$

Note now that another set of highlighted terms in green cancel out on both sides. So

$$\begin{aligned} RHS &= a^\top (A + B)^{-1} a - 2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b + \\ &\quad (A^{-1}a + B^{-1}b)^\top (A^{-1} + B^{-1})^{-1} (A^{-1}a + B^{-1}b) \end{aligned} \quad (5.1.76)$$

$$LHS = a^\top A^{-1} a + b^\top B^{-1} b \quad (5.1.77)$$

We deal with the highlighted term in red above using the matrix inversion lemma:

$$(A^{-1} + B^{-1})^{-1} = A - A(A + B)^{-1} A \quad (5.1.78)$$

Substituting this into the RHS:

$$\begin{aligned} RHS &= a^\top (A + B)^{-1} a - 2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b \\ &\quad + \left( a^\top A^{-1} + b^\top B^{-1} \right) \left( A - A(A + B)^{-1} A \right) (A^{-1} a + B^{-1} b) \end{aligned} \quad (5.1.79)$$

Expanding out the quadratic, this gives the very length expression

$$\begin{aligned} RHS &= a^\top (A + B)^{-1} a - 2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b + a^\top A^{-1} A A^{-1} a \\ &\quad + a^\top A^{-1} A B^{-1} b + b^\top B^{-1} A A^{-1} a + b^\top B^{-1} A B^{-1} b - a^\top A^{-1} A (A + B)^{-1} A A^{-1} a \\ &\quad - b^\top B^{-1} A (A + B)^{-1} A A^{-1} a - a^\top A^{-1} A (A + B)^{-1} A B^{-1} b - b^\top B^{-1} A (A + B)^{-1} A B^{-1} b \end{aligned} \quad (5.1.80)$$

We can notice instances where  $A^{-1}A$  or  $AA^{-1}$  appears and cancel them out.

$$\begin{aligned} RHS &= a^\top (A + B)^{-1} a - 2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b + a^\top A^{-1} a \\ &\quad + a^\top B^{-1} b + b^\top B^{-1} a + b^\top B^{-1} A B^{-1} b - a^\top (A + B)^{-1} a - b^\top B^{-1} A (A + B)^{-1} a \\ &\quad - a^\top (A + B)^{-1} A B^{-1} b - b^\top B^{-1} A (A + B)^{-1} A B^{-1} b \end{aligned} \quad (5.1.81)$$

We can cancel out some blue terms within the RHS, as well as cancel the  $a^\top A^{-1} a$  from the LHS. We are left with

$$\begin{aligned} RHS &= -2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b + a^\top B^{-1} b + b^\top B^{-1} a + b^\top B^{-1} A B^{-1} b \\ &\quad - b^\top B^{-1} A (A + B)^{-1} a - a^\top (A + B)^{-1} A B^{-1} b - b^\top B^{-1} A (A + B)^{-1} A B^{-1} b \end{aligned} \quad (5.1.82)$$

$$LHS = b^\top B^{-1} b \quad (5.1.83)$$

We can also group similar terms in the RHS, giving

$$\begin{aligned} RHS &= -2a^\top (A + B)^{-1} b + b^\top (A + B)^{-1} b + 2a^\top B^{-1} b + \\ &\quad b^\top B^{-1} A B^{-1} b - 2a^\top (A + B)^{-1} A B^{-1} b - b^\top B^{-1} A (A + B)^{-1} A B^{-1} b \end{aligned} \quad (5.1.84)$$

Next we subtract the LHS from the RHS and factorise out  $a^\top (\cdot) b$  and  $b^\top (\cdot) b$ :

$$\begin{aligned} RHS - LHS &= a^\top \underbrace{\left( -2(A + B)^{-1} + 2B^{-1} - 2(A + B)^{-1} A B^{-1} \right)}_{=0} b \\ &\quad + b^\top \underbrace{\left( (A + B)^{-1} + B^{-1} A B^{-1} - B^{-1} A (A + B)^{-1} A B^{-1} \right)}_{=0} b \end{aligned} \quad (5.1.85)$$

Thus to show equivalence, we need to show that the terms inside the brackets are equal to zero. This requires carrying out some manipulations. Starting from the first term:

$$-2(A + B)^{-1} + 2B^{-1} - 2(A + B)^{-1} A B^{-1} = 0 \quad (5.1.86)$$

$$-(A + B)^{-1} + B^{-1} - (A + B)^{-1} A B^{-1} = 0 \quad (5.1.87)$$

$$B^{-1} = (A + B)^{-1} (I + A B^{-1}) \quad (5.1.88)$$

$$A + B = (I + A B^{-1}) B \quad (5.1.89)$$

$$A + B = B + A \quad (5.1.90)$$

For the second term:

$$(A + B)^{-1} + B^{-1} A B^{-1} - B^{-1} A (A + B)^{-1} A B^{-1} = 0 \quad (5.1.91)$$

$$B^{-1} \left( AB^{-1} - A(A+B)^{-1} AB^{-1} - I \right) = -(A+B)^{-1} \quad (5.1.92)$$

$$\left( AB^{-1} - A(A+B)^{-1} AB^{-1} - I \right) (A+B) = -B \quad (5.1.93)$$

$$AB^{-1}A - A(A+B)^{-1}AB^{-1}A - A + A - A(A+B)^{-1}A - B = -B \quad (5.1.94)$$

$$AB^{-1}A - A(A+B)^{-1}AB^{-1}A - A(A+B)^{-1}A = 0 \quad (5.1.95)$$

$$A \left( B^{-1} - (A+B)^{-1}AB^{-1} - (A+B)^{-1} \right) A = 0 \quad (5.1.96)$$

Since  $A \neq 0$  (we have already taken the inverse of  $A$  up until this stage):

$$B^{-1} - (A+B)^{-1}AB^{-1} - (A+B)^{-1} = 0 \quad (5.1.97)$$

$$B^{-1} = (A+B)^{-1}(AB^{-1} + I) \quad (5.1.98)$$

$$A+B = (AB^{-1} + I)B \quad (5.1.99)$$

$$A+B = A+B \quad (5.1.100)$$

Hence the terms inside the brackets are zero, so the LHS equals the RHS. This finally shows equivalence over the numerators.

### 5.1.6 Quotient of Gaussian Densities

A corollary from the result on the product of Gaussian identities is that we can take the quotient between Gaussian densities as follows:

$$\frac{\mathcal{N}(x; c, C)}{\mathcal{N}(x; b, B)} = Z\mathcal{N}(x; a, A) \quad (5.1.101)$$

where

$$A = (C^{-1} - B^{-1})^{-1} \quad (5.1.102)$$

$$a = A(C^{-1}c - B^{-1}b) \quad (5.1.103)$$

$$Z = (2\pi)^{D/2} |A+B|^{1/2} \exp \left( \frac{1}{2} (a-b)^\top (A+B)^{-1} (a-b) \right) \quad (5.1.104)$$

Note that this implicitly requires that  $C^{-1} - B^{-1}$  is positive definite.

### 5.1.7 Marginalisation of Gaussians

Suppose we have distributions  $p(x)$  and  $p(y|x)$ , and we want to obtain the distribution  $p(y)$ . We can do this by first computing the joint distribution  $p(x,y) = p(x)p(y|x)$  and integrating over  $x$  as follows

$$p(y) = \int p(x,y) dx \quad (5.1.105)$$

This is known as marginalisation. Suppose  $p(x)$  and  $p(y|x)$  are Gaussians in the sense that

$$p(x) = \mathcal{N}_x(a, A^{-1}) \quad (5.1.106)$$

$$p(y|x) = \mathcal{N}_y(C^\top x, B^{-1}) \quad (5.1.107)$$

Note that  $x$  and  $y$  do not have to be of the same dimension. The general technique to evaluate the integral is from the joint distribution, decompose it into  $p(x,y) = p(y)p(x|y)$ . Then

$$\int p(x,y) dx = \int p(y)p(x|y) dx \quad (5.1.108)$$

$$= p(y) \int p(x|y) dx \quad (5.1.109)$$

$$= p(y) \quad (5.1.110)$$

since the integral of  $p(x|y)$  evaluates to 1. Without worrying about normalising constants, we substitute the exponential expressions for the Gaussian distributions

$$p(y) = \int \mathcal{N}_x(a, A^{-1}) \mathcal{N}_y(C^\top x, B^{-1}) dx \quad (5.1.111)$$

$$\propto \int \exp \left[ -\frac{1}{2} \left( (x-a)^\top A(x-a) + (y-C^\top x)^\top B(y-C^\top x) \right) \right] dx \quad (5.1.112)$$

$$= \int \exp \left[ -\frac{1}{2} \left( x^\top Ax - 2a^\top Ax + a^\top Aa + y^\top By - 2y^\top BC^\top x + x^\top CBC^\top x \right) \right] dx \quad (5.1.113)$$

$$= \int \exp \left[ -\frac{1}{2} \left( x^\top (A + CBC^\top) x - 2(a^\top A + y^\top BC^\top) x + a^\top Aa + y^\top By \right) \right] dx \quad (5.1.114)$$

To decompose the distributions, we complete the square:

**Lemma 5.1** (Completing the Square).

$$\frac{1}{2}x^\top Mx + d^\top x + e = \frac{1}{2}(x-m)^\top M(x-m) + v \quad (5.1.115)$$

where

$$m = -C^{-1}d \quad (5.1.116)$$

$$v = e - \frac{1}{2}d^\top M^{-1}d \quad (5.1.117)$$

So by letting  $M = A + CBC^\top$ ,  $d = -Aa - CBy$ , and  $e = \frac{1}{2}a^\top Aa + \frac{1}{2}y^\top By$ , this gives

$$m = (A + CBC^\top)^{-1}(Aa + CBy) \quad (5.1.118)$$

$$v = \frac{1}{2}a^\top Aa + \frac{1}{2}y^\top By - \frac{1}{2}(Aa + CBy)^\top (A + CBC^\top)^{-1}(Aa + CBy) \quad (5.1.119)$$

So we can write

$$p(y) \propto \int \exp \left[ -\frac{1}{2} \left( (x-m)^\top M(x-m) + v \right) \right] dx \quad (5.1.120)$$

and from this we can see that

$$p(x|y) = \mathcal{N}_x(m, M^{-1}) \quad (5.1.121)$$

$$= \mathcal{N}_x \left( (A + CBC^\top)^{-1}(Aa + CBy), (A + CBC^\top)^{-1} \right) \quad (5.1.122)$$

$$(5.1.123)$$

For the remaining terms, these can be taken out of the integral

$$p(y) \propto \exp \left( -\frac{1}{2}v \right) \int \mathcal{N}_x \left( (A + CBC^\top)^{-1}(Aa + CBy), (A + CBC^\top)^{-1} \right) dx \quad (5.1.124)$$

$$= \exp \left( -\frac{1}{2}v \right) \quad (5.1.125)$$

$$= \exp \left[ -\frac{1}{2} \left( a^\top Aa + y^\top By - (Aa + CBCy)^\top (A + CBC^\top)^{-1} (Aa + CBCy) \right) \right] \quad (5.1.126)$$

$$\begin{aligned} &= \exp \left[ -\frac{1}{2} \left( y^\top By - y^\top BC^\top (A + CBC^\top)^{-1} CBCy - 2a^\top A (A + CBC^\top)^{-1} CBCy \right. \right. \\ &\quad \left. \left. + a^\top Aa - a^\top A (A + CBC^\top)^{-1} Aa \right) \right] \end{aligned} \quad (5.1.127)$$

Getting rid of the terms which do not depend on  $y$

$$p(y) \propto \exp \left[ -\frac{1}{2} \left( y^\top By - y^\top BC^\top (A + CBC^\top)^{-1} CBCy - 2a^\top A (A + CBC^\top)^{-1} CBCy \right) \right] \quad (5.1.128)$$

$$= \exp \left[ -\frac{1}{2} \left( y^\top \left( B - BC^\top (A + CBC^\top)^{-1} CB \right) y - 2a^\top A (A + CBC^\top)^{-1} CBCy \right) \right] \quad (5.1.129)$$

Note by the matrix inversion lemma,  $B - BC^\top (A + CBC^\top)^{-1} CB = (B^{-1} + C^\top A^{-1} C)^{-1}$ . This yields the simplification

$$p(y) \propto \exp \left[ -\frac{1}{2} \left( y^\top \left( B^{-1} + C^\top A^{-1} C \right)^{-1} y - 2a^\top A (A + CBC^\top)^{-1} CBCy \right) \right] \quad (5.1.130)$$

Once again we can complete the square in  $y$ . Doing so (and ignoring the terms which don't depend on  $y$  due to proportionality) gives

$$\begin{aligned} p(y) \propto \exp & \left[ -\frac{1}{2} \left( y - \left( B^{-1} + C^\top A^{-1} C \right) BC^\top (A + CBC^\top)^{-1} Aa \right)^\top \cdot \left( B^{-1} + C^\top A^{-1} C \right)^{-1} \right. \\ & \left. \cdot \left( y - \left( B^{-1} + C^\top A^{-1} C \right) BC^\top (A + CBC^\top)^{-1} Aa \right) \right] \end{aligned} \quad (5.1.131)$$

This gives the distribution for  $p(y)$

$$p(y) = \mathcal{N}_y \left( \left( B^{-1} + C^\top A^{-1} C \right) BC^\top (A + CBC^\top)^{-1} Aa, B^{-1} + C^\top A^{-1} C \right) \quad (5.1.132)$$

To simplify the mean, we can show that

$$\left( B^{-1} + C^\top A^{-1} C \right) BC^\top (A + CBC^\top)^{-1} Aa = C^\top a \quad (5.1.133)$$

Expanding the LHS gives

$$C^\top (A + CBC^\top)^{-1} Aa + C^\top A^{-1} CBC^\top (A + CBC^\top)^{-1} Aa = C^\top a \quad (5.1.134)$$

$$\underbrace{C^\top \left[ (A + CBC^\top)^{-1} A + A^{-1} CBC^\top (A + CBC^\top)^{-1} A \right]}_{=I} a = C^\top a \quad (5.1.135)$$

So we require  $(A + CBC^\top)^{-1} A + A^{-1} CBC^\top (A + CBC^\top)^{-1} A = I$  by hypothesis. Some manipulation yields

$$(I + A^{-1} CBC^\top) (A + CBC^\top)^{-1} A = I \quad (5.1.136)$$

$$I + A^{-1} CBC^\top = A^{-1} (A + CBC^\top) \quad (5.1.137)$$

$$I + A^{-1} CBC^\top = I + A^{-1} CBC^\top \quad (5.1.138)$$

Therefore we finally have

$$p(y) = \mathcal{N}_y \left( C^\top a, B^{-1} + C^\top A^{-1} C \right) \quad (5.1.139)$$

### 5.1.8 Exchangeability of Gaussian Sequences

Let  $(X_1, \dots, X_n)$  be a multivariate Gaussian sequence, with a correlation matrix such that

$$R = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix} \quad (5.1.140)$$

Permuting any of the rows and columns together does not change the correlation matrix, so the joint distribution remains unchanged. Therefore an equi-correlated standard Gaussian sequence is an exchangeable sequence.

### 5.1.9 Bivariate Gaussian Properties

#### Bivariate Gaussian Conditioning and Skew Normality [11]

**Theorem 5.1.** *Let  $Y, W$  be independent standard Gaussian random variables. Then for any  $\lambda \in \mathbb{R}$ , the conditional distribution of  $Y$  given  $\lambda Y > W$  is equal in law to a skew normal distribution with shape parameter  $\lambda$ .*

*Proof.* Use  $f_Y(y)$  to denote the marginal density of  $Y$  and  $f_{Y|\lambda Y > W}(y; \lambda)$  to denote the conditional density of  $Y$  given  $\lambda Y > W$ . Using Bayes' theorem,

$$f_{Y|\lambda Y > W}(y; \lambda) = \frac{\Pr(\lambda Y > W | Y = y) f_Y(y)}{\Pr(\lambda Y > W)} \quad (5.1.141)$$

Note for the denominator:

$$\Pr(\lambda Y > W) = \Pr(\lambda Y - W \leq 0) \quad (5.1.142)$$

$$= \Phi(0) \quad (5.1.143)$$

$$= \frac{1}{2} \quad (5.1.144)$$

since  $\lambda Y - W$  is zero-mean and symmetric. As for the numerator:

$$\Pr(\lambda Y > W | Y = y) f_Y(y) = \Pr(W \leq \lambda y) f_Y(y) \quad (5.1.145)$$

$$= \Phi(\lambda y) \phi(y) \quad (5.1.146)$$

Hence

$$f_{Y|\lambda Y > W}(y; \lambda) = 2\Phi(\lambda y) \phi(y) \quad (5.1.147)$$

which matches the form of the skew normal distribution.  $\square$

**Corollary 5.1.** *If  $X, Y$  are bivariate standard Gaussian with correlation  $\rho$ , then the conditional distribution of  $Y$  given  $X < 0$  is equal in law to a skew normal distribution with shape parameter  $\rho/\sqrt{1 - \rho^2}$ .*

*Proof.* Using the result above, put  $X = (\lambda Y - W)/\sqrt{1 + \lambda^2}$ , which we can verify as being standard Gaussian since  $\text{Var}(\lambda Y - W) = \lambda^2 + 1$ . Thus we have  $\lambda Y - W < 0$  whenever  $X < 0$ . It immediately follows that  $Y$  given  $X < 0$  is skew normally distributed with shape parameter  $\lambda$ . To derive  $\lambda$  in terms of  $\rho$ , we can show

$$\text{Cov}(Y, X) = \text{Cov}\left(Y, \frac{\lambda Y - W}{\sqrt{1 + \lambda^2}}\right) \quad (5.1.148)$$

$$= \text{Cov}\left(Y, \frac{\lambda}{\sqrt{1 + \lambda^2}} Y\right) \quad (5.1.149)$$

$$= \frac{\lambda}{\sqrt{1+\lambda^2}} \operatorname{Var}(Y) \quad (5.1.150)$$

$$= \frac{\lambda}{\sqrt{1+\lambda^2}} \quad (5.1.151)$$

Hence

$$\rho = \operatorname{Corr}(Y, X) \quad (5.1.152)$$

$$= \frac{\lambda}{\sqrt{1+\lambda^2}} \quad (5.1.153)$$

Since  $X$  and  $Y$  have unit variance. Upon rearranging for  $\lambda$ , we get

$$\lambda = \frac{\rho}{\sqrt{1-\rho^2}} \quad (5.1.154)$$

□

### Bivariate Gaussian Quadrant Probability

**Theorem 5.2.** Let  $(Z_1, Z_2)$  be a bivariate Gaussian distribution with

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \quad (5.1.155)$$

i.e.  $Z_1$  and  $Z_2$  are both standard Gaussian, with correlation  $\rho$ . Then

$$\Pr(Z_1 > 0, Z_2 > 0) = \frac{1}{4} + \frac{\arcsin \rho}{2\pi} \quad (5.1.156)$$

*Proof.* By using conditioning formulae, first note that

$$[Z_2|Z_1 = z] \sim \mathcal{N}(\rho z, 1 - \rho^2) \quad (5.1.157)$$

Because  $(Z_1, Z_2)$  are zero-mean, it is identically distributed with  $(-Z_1, -Z_2)$  and thus

$$\Pr(Z_1 > 0, Z_2 > 0) = \Pr(Z_1 < 0, Z_2 < 0) \quad (5.1.158)$$

$$= \int_{-\infty}^0 \phi(z_1) \Pr(Z_2 < 0|Z_1 = z_1) dz_1 \quad (5.1.159)$$

$$= \int_{-\infty}^0 \phi(z_1) \Phi\left(\frac{-\rho z_1}{\sqrt{1-\rho^2}}\right) dz_1 \quad (5.1.160)$$

Differentiating both sides with respect to  $\rho$ ,

$$\frac{d}{d\rho} \Pr(Z_1 > 0, Z_2 > 0) = \int_{-\infty}^0 \phi(z_1) \frac{d}{d\rho} \Phi\left(\frac{-\rho z_1}{\sqrt{1-\rho^2}}\right) dz_1 \quad (5.1.161)$$

Before using the chain rule, we calculate

$$\frac{d}{d\rho} \rho (1 - \rho^2)^{-1/2} = (1 - \rho^2)^{-1/2} + \rho \left(-\frac{1}{2}\right) (-2\rho) (1 - \rho^2)^{-3/2} \quad (5.1.162)$$

$$= \frac{\rho^2}{\left(\sqrt{1-\rho^2}\right)^3} + \frac{1}{\sqrt{1-\rho^2}} \quad (5.1.163)$$

$$= \frac{\rho^2 + 1 - \rho^2}{\left(\sqrt{1-\rho^2}\right)^3} \quad (5.1.164)$$

$$= (1 - \rho^2)^{-3/2} \quad (5.1.165)$$

Hence applying the chain rule

$$\frac{d}{d\rho} \Pr(Z_1 > 0, Z_2 > 0) = \int_{-\infty}^0 \phi(z_1) \left[ -z_1 (1 - \rho^2)^{-3/2} \right] \phi\left(\frac{-\rho z_1}{\sqrt{1 - \rho^2}}\right) dz_1 \quad (5.1.166)$$

$$= (1 - \rho^2)^{-3/2} \int_0^\infty z \phi(z) \phi\left(\frac{-\rho z}{\sqrt{1 - \rho^2}}\right) dz \quad (5.1.167)$$

$$= (1 - \rho^2)^{-3/2} \int_0^\infty z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\rho^2 z^2}{2(1 - \rho^2)}\right) dz \quad (5.1.168)$$

$$= \frac{(1 - \rho^2)^{-3/2}}{2\pi} \int_0^\infty z \exp\left(-\frac{z^2}{2} - \frac{\rho^2 z^2}{2(1 - \rho^2)}\right) dz \quad (5.1.169)$$

$$= \frac{(1 - \rho^2)^{-3/2}}{2\pi} \int_0^\infty z \exp\left(-\frac{1}{2} \cdot \frac{(1 - \rho^2) z^2 + \rho^2 z^2}{1 - \rho^2}\right) dz \quad (5.1.170)$$

$$= \frac{(1 - \rho^2)^{-3/2}}{2\pi} \int_0^\infty z \exp\left(-\frac{1}{2} \cdot \frac{z^2}{1 - \rho^2}\right) dz \quad (5.1.171)$$

Now use the substitution  $x = \frac{z}{\sqrt{1 - \rho^2}}$  so  $dz = (1 - \rho^2)^{1/2} dx$  and

$$\frac{d}{d\rho} \Pr(Z_1 > 0, Z_2 > 0) = \frac{(1 - \rho^2)^{-3/2}}{2\pi} \int_0^\infty (1 - \rho^2)^{1/2} x \exp\left(-\frac{1}{2} \cdot \frac{z^2}{1 - \rho^2}\right) (1 - \rho^2)^{1/2} dx \quad (5.1.172)$$

$$= \frac{(1 - \rho^2)^{-1/2}}{2\pi} \int_0^\infty x \exp\left(-\frac{x^2}{2}\right) dx \quad (5.1.173)$$

To evaluate the integral, use another change of variables  $y = x^2$  and  $dy = 2x dx$  so

$$\int_0^\infty x \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{2} \int_0^\infty \exp\left(-\frac{y}{2}\right) dy \quad (5.1.174)$$

$$= \frac{1}{2} \left[ -2 \exp\left(-\frac{y}{2}\right) \right]_0^\infty \quad (5.1.175)$$

$$= 1 \quad (5.1.176)$$

Therefore

$$\frac{d}{d\rho} \Pr(Z_1 > 0, Z_2 > 0) = \frac{(1 - \rho^2)^{-1/2}}{2\pi} \quad (5.1.177)$$

Take the antiderivative (recognising that it is the derivative of arcsin) to give

$$\Pr(Z_1 > 0, Z_2 > 0) = \frac{1}{2\pi} \arcsin \rho + c \quad (5.1.178)$$

To solve for the constant  $c$ , we use the fact that when  $\rho = 0$  (i.e. independent Gaussians), we have  $\Pr(Z_1 > 0, Z_2 > 0) = \frac{1}{4}$ . Then

$$\frac{1}{4} = \frac{1}{2\pi} \arcsin 0 + c \quad (5.1.179)$$

$$c = \frac{1}{4} \quad (5.1.180)$$

□

### Population Kendall Correlation of Bivariate Gaussian

Recall that the population Kendall correlation  $\tau$  of  $(X, Y)$  can be characterised as

$$\Pr(X > X', Y > Y') = \frac{\tau + 1}{4} \quad (5.1.181)$$

where  $(X', Y')$  is an independent copy. Suppose  $(X, Y)$  are bivariate standard Gaussian with covariance and correlation  $\rho$ . Then  $X - X'$  and  $Y - Y'$  are both zero-mean Gaussian with  $\text{Var}(X - X') = 2$ ,  $\text{Var}(Y - Y') = 2$  and

$$\text{Cov}(X - X', Y - Y') = \text{Cov}(X, Y) + \text{Cov}(X', Y') \quad (5.1.182)$$

$$= 2\rho \quad (5.1.183)$$

Thus the correlation is still

$$\text{Corr}(X - X', Y - Y') = \frac{\text{Cov}(X - X', Y - Y')}{\sqrt{\text{Var}(X - X') \text{Var}(Y - Y')}} \quad (5.1.184)$$

$$= \rho \quad (5.1.185)$$

Then using the bivariate Gaussian quadrant probability, the population Kendall correlation is related to  $\rho$  by

$$\frac{\tau + 1}{4} = \Pr(X > X', Y > Y') \quad (5.1.186)$$

$$= \Pr(X - X' > 0, Y - Y' > 0) \quad (5.1.187)$$

$$= \frac{1}{4} + \frac{\arcsin \rho}{2\pi} \quad (5.1.188)$$

Therefore

$$\tau + 1 = 1 + \frac{2}{\pi} \arcsin \rho \quad (5.1.189)$$

$$\tau = \frac{2}{\pi} \arcsin \rho \quad (5.1.190)$$

or

$$\rho = \sin\left(\frac{\pi}{2}\tau\right) \quad (5.1.191)$$

Because positive affine transformations of random variables do not affect the correlation, it follows that this relationship also holds for any bivariate Gaussian with correlation  $\rho$ .

## 5.2 Stochastic Processes

A continuous time stochastic process  $X(t)$  assigns a function of time  $x(t)$  to each outcome in the sample space. The function  $x(t)$  is said to be a sample path, i.e. a realisation of the stochastic process. At some fixed time  $t_0$ , the variable  $X(t_0)$  is a random variable. A discrete time stochastic process  $X_k$  assigns a sequence to each outcome in the sample space, where similarly for some fixed integer  $k_0$ , the variable  $X_{k_0}$  is a random variable. Note that  $X(t_0)$  and  $X_{k_0}$  can be discrete or continuous-valued random variables (excluding mixed-valued random variables), so we can consider four archetypal classes of stochastic processes:

- Continuous-time and continuous-valued
- Continuous-time and discrete-valued
- Discrete-time and continuous-valued
- Discrete-time and discrete-valued

### 5.2.1 Properties of Stochastic Processes

#### $n^{\text{th}}$ Order Distribution Functions

The  $n^{\text{th}}$  order distribution function of a continuous-time stochastic process  $X(t)$  for times  $t_1, \dots, t_n$  is given by the joint CDF of the random variables  $X(t_1), \dots, X(t_n)$ :

$$F_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) = \Pr(X(t_1) \leq x_1, \dots, X(t_n) \leq x_n) \quad (5.2.1)$$

If the stochastic process is continuous-valued, then the  $n^{\text{th}}$  order probability density function may be defined as

$$f_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) \quad (5.2.2)$$

The  $n^{\text{th}}$  order cumulative distribution functions and probability density functions of a discrete-time stochastic process may be analogously defined.

#### Mean Function

The mean function  $\mu_X(t)$  for a continuous time stochastic process  $X(t)$  is defined by

$$\mu_X(t) = \mathbb{E}[X(t)] \quad (5.2.3)$$

and similarly for a discrete time stochastic process.

#### Autocovariance Function

The autocovariance function of a stochastic process  $X(t)$  is the covariance between the process at time  $t$  and the process at time shifted by some  $\tau$ . Formally,

$$C_X(t, \tau) = \text{Cov}(X(t), X(t + \tau)) \quad (5.2.4)$$

$$= \mathbb{E}[X(t)X(t + \tau)] - \mu_X(t)\mu_X(t + \tau) \quad (5.2.5)$$

An alternative notation instead expresses the covariance at two times  $t$  and  $s$ , in which case

$$C_X(t, s) = \mathbb{E}[X(t)X(s)] - \mu_X(t)\mu_X(s) \quad (5.2.6)$$

Analogously for a random sequence at time  $m$  with time shift  $k$ ,

$$C_X(m, k) = \mathbb{E}[X_m X_{m+k}] - \mu_X(m)\mu_X(m+k) \quad (5.2.7)$$

#### Autocorrelation Function

The autocorrelation function of a stochastic process  $X(t)$  for time  $t$  and difference  $\tau$  is defined as

$$R_X(t, \tau) = \mathbb{E}[X(t)X(t + \tau)] \quad (5.2.8)$$

and analogously for a discrete random sequence with time  $m$  and difference  $k$ :

$$R_X(m, k) = \mathbb{E}[X_m X_{m+k}] \quad (5.2.9)$$

The autocorrelation function can alternatively be specified with two times, i.e.  $R_X(t, s) = \mathbb{E}[X(t)X(s)]$ . It is clear that a zero-mean stochastic process has autocovariance function equal to autocorrelation function. Note that the autocorrelation function is distinct from the autocorrelation coefficient, the latter which refers to a function for the traditional (normalised) correlation coefficient.

## Autocorrelation Coefficient

The autocorrelation coefficient for a stochastic process  $X(t)$  with times  $t$  and  $s$  is defined as

$$r_X(t, s) = \frac{C_X(t, s)}{\sqrt{C_X(t, t) C_X(s, s)}} \quad (5.2.10)$$

which is the correlation coefficient between random variables  $X(t)$  and  $X(s)$ . Analogously for a random sequence

$$r_X(m, n) = \frac{C_X(m, n)}{\sqrt{C_X(m, m) C_X(n, n)}} \quad (5.2.11)$$

## Cross-Correlation Function

For two continuous stochastic processes  $X(t)$  and  $Y(t)$ , the cross-correlation function with time  $t$  and difference  $\tau$  is defined as

$$R_{XY}(t, \tau) = \mathbb{E}[X(t)Y(t + \tau)] \quad (5.2.12)$$

For two continuous random sequences  $X_n$  and  $Y_n$ ,

$$R_{XY}(m, k) = \mathbb{E}[X_m Y_{m+k}] \quad (5.2.13)$$

The autocorrelation function can be thought of as a special case of the cross-correlation function where  $X$  and  $Y$  are identical processes.

## Second-Order Processes

Second-order processes have finite second moments, i.e. for a continuous time process:

$$\mathbb{E}[X(t)^2] < \infty \quad (5.2.14)$$

for all  $t$ .

## Independent Increment Processes

Consider a continuous-time stochastic process  $X(t)$  and some increasing time indices  $t_1 < t_2 < \dots < t_k$ . The process is said to be an independent increment process if the random variables  $X(t_k) - X(t_{k-1}), \dots, X(t_2) - X(t_1)$  are all independent for any  $t_1 < t_2 < \dots < t_k$ . Independent increment processes can be analogously defined for discrete-time processes.

## Orthogonal Increment Processes

Orthogonal increment processes are weaker than independent increment processes, where the random variables only need be uncorrelated.

## Quadratic Variation

For a continuous-time stochastic process  $X(t)$ , suppose the interval  $[0, t]$  has been partitioned at equidistant points  $t_0, t_1, \dots, t_n$ , i.e.

$$t_0 = 0 \quad (5.2.15)$$

$$t_1 = \frac{1}{n}t \quad (5.2.16)$$

$$\vdots \quad (5.2.17)$$

$$t_n = t \quad (5.2.18)$$

Then consider the sum

$$V_n := \sum_{k=1}^n (X(t_k) - X(t_{k-1}))^2 \quad (5.2.19)$$

The quadratic variation of  $X(t)$ , denoted  $[X](t)$ , is itself a stochastic process, and defined at time  $t$  by

$$[X](t) = \lim_{n \rightarrow \infty} V_n \quad (5.2.20)$$

$$= \lim_{n \rightarrow \infty} \sum_{k=1}^n (X(t_k) - X(t_{k-1}))^2 \quad (5.2.21)$$

This quantity intuitively measures the cumulative ‘roughness’ of sample paths, as note that  $[X](t) = 0$  for smooth functions.

## 5.2.2 Stationarity

### Strict Stationarity

Let  $X(t)$  be a continuous time stochastic process where  $f_{X(t_1), \dots, X(t_m)}(x_1, \dots, x_m)$  is the joint probability density function for the process at the set of time instants  $t_1, \dots, t_m$ . The process  $X(t)$  is said to be strictly stationary if and only if for all sets of time instants and any time difference  $\tau$ ,

$$f_{X(t_1), \dots, X(t_m)}(x_1, \dots, x_m) = f_{X(t_1+\tau), \dots, X(t_m+\tau)}(x_1, \dots, x_m) \quad (5.2.22)$$

Let  $X_n$  be a discrete time random sequence where  $p_{X_{n_1}, \dots, X_{n_m}}(x_1, \dots, x_m)$  is the joint probability mass function for the process at the set of integer time instants  $n_1, \dots, n_m$ . The process  $X_n$  is said to be strictly stationary if and only if for all sets of time instants and any time difference  $k$ ,

$$p_{X_{n_1}, \dots, X_{n_m}}(x_1, \dots, x_m) = p_{X_{n_1+k}, \dots, X_{n_m+k}}(x_1, \dots, x_m) \quad (5.2.23)$$

### Wide Sense Stationarity

Wide sense stationarity (WSS), also known as weak stationarity, is notionally a ‘weaker’ form of strict stationarity. A wide sense stationary process requires that the mean function be a constant, i.e.

$$\mathbb{E}[X(t)] = \mu_X \quad (5.2.24)$$

and the autocorrelation function only depend on the time difference, i.e.

$$R_X(t, \tau) = R_X(\tau) \quad (5.2.25)$$

One property this implies is that  $R_X(-\tau) = R_X(\tau)$  (i.e. the autocorrelation function is symmetric about zero) since  $\mathbb{E}[X(t)X(t-\tau)] = \mathbb{E}[X(t-\tau)X(t)]$ .

A wide sense stationary process does not need to be strictly stationary. For example, a sequence of uncorrelated random variables with the same mean and variance is wide-sense stationary by definition, but it will not be strictly stationary if these random variables are not identically distributed.

A strictly stationary process also does not necessarily imply it is wide-sense stationary. For example, a sequence of i.i.d Cauchy random variables is strictly stationary, but not wide-sense stationary because the Cauchy distribution does not have a finite second moment. Generally however, second-order processes which are strictly stationary are also wide-sense stationary.

## Joint Wide Sense Stationarity

Two stochastic processes are jointly wide sense stationary if both processes are wide sense stationary, and additionally the cross-correlation between the two only depends on the time difference, i.e.

$$R_{XY}(t, \tau) = R_{XY}(\tau) \quad (5.2.26)$$

A property is then  $R_{XY}(-\tau) = R_{YX}(\tau)$  because with the substitution  $s = t - \tau$ :

$$R_{XY}(-\tau) = \mathbb{E}[X(t)Y(t-\tau)] \quad (5.2.27)$$

$$= \mathbb{E}[Y(t-\tau)X(t)] \quad (5.2.28)$$

$$= \mathbb{E}[Y(s)X(s+\tau)] \quad (5.2.29)$$

$$= R_{YX}(\tau) \quad (5.2.30)$$

## Stationary Increment Processes

If for an independent increment process the distribution of  $X(t) - X(s)$  only depends on the time difference  $t - s$ , then the process is said to be a stationary increment process.

## Partial Autocorrelation Function

The partial autocorrelation function is defined for wide-sense stationary discrete-time stochastic processes. For lag length  $h$ , the partial autocorrelation function of a stochastic process  $X_t$  is the partial correlation between  $X_t$  and  $X_{t+h}$ , given  $(X_{t+1}, \dots, X_{t+h-1})$ :

$$\varphi_X(h) = \rho_{X_t, X_{t+h}|(X_{t+1}, \dots, X_{t+h-1})} \quad (5.2.31)$$

## Trend Stationarity

A continuous-time stochastic process  $Y(t)$  is said to be trend stationary if it can be expressed as

$$Y(t) = f(t) + X(t) \quad (5.2.32)$$

where  $f(t)$  is a deterministic function of time, and  $X(t)$  is a stationary process. Thus, after the subtracting the trend,  $Y(t) - f(t)$  is stationary. An analogous definition exists for discrete-time stochastic processes.

## Asymptotic Stationarity [151]

An asymptotically stationary stochastic process is characterised by the process becoming stationary after a long time. Formally, we require the  $n^{\text{th}}$  order distribution function  $F_{X(t_1+\tau), \dots, X(t_n+\tau)}$  to tend to a limit as  $\tau \rightarrow \infty$ . So we can say that approximately

$$F_{X(t_1+\tau), \dots, X(t_n+\tau)}(x_1, \dots, x_n) \approx F_{X(t_1+s), \dots, X(t_n+s)}(x_1, \dots, x_n) \quad (5.2.33)$$

for large  $\tau$  and  $s$ .

## Cyclostationarity [151]

A cyclostationary stochastic process can be characterised as having statistical properties which vary cyclically with time. For a continuous-time stochastic process  $X(t)$ , we may define *strict cyclostationarity* with period  $T$  if its  $n^{\text{th}}$  order distribution satisfies

$$F_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) = F_{X(t_1+mT), \dots, X(t_n+mT)}(x_1, \dots, x_n) \quad (5.2.34)$$

for any integer  $m$ . We may also define *wide-sense cyclostationarity* with period  $T$  if its mean function satisfies

$$\mu_X(t) = \mu_X(t + mT) \quad (5.2.35)$$

and its autocorrelation function  $R_X(t, s) = \mathbb{E}[X(t)X(s)]$  satisfies

$$R_X(t, s) = R_X(t + mT, s + mT) \quad (5.2.36)$$

for any integer  $m$ .

### Time-Reversible Processes [52]

A stationary stochastic process is said to be time-reversible if, roughly speaking, the statistical properties of the process would be the same if it were ‘played’ in reverse. In the continuous-time case, we require that for a set of time increments  $\tau_1 < \dots < \tau_n$  that its  $n^{\text{th}}$  order distribution satisfies

$$F_{X(t+\tau_1), \dots, X(t+\tau_n)}(x_1, \dots, x_n) = F_{X(t-\tau_1), \dots, X(t-\tau_n)}(x_1, \dots, x_n) \quad (5.2.37)$$

### Mean-Square Periodicity [151]

A stochastic process  $X(t)$  is said to be mean-square periodic with period  $T$  if for every  $t$ :

$$\mathbb{E} [|X(t+T) - X(t)|^2] = 0 \quad (5.2.38)$$

Note that since  $\text{Var}(X(t+T) - X(t)) \leq \mathbb{E} [|X(t+T) - X(t)|^2]$ , this implies that

$$\text{Var}(X(t+T) - X(t)) = 0 \quad (5.2.39)$$

so  $X(t+T) = X(t)$  with probability one. An example of a mean-square periodic stochastic process is any periodic function with a random phase shift.

### 5.2.3 Ergodicity

Ergodicity can informally be thought of as a ‘stronger’ form of stationarity that expresses in some sense a notion of ‘repetition’ in the stochastic process - that a single sample path is sufficiently ‘rich’ in information because its properties repeat themselves over and over again enough times that we can determine the properties of the entire stochastic process just by looking at the sample path.

### Mean-Ergodicity [151]

A stationary (or wide-sense stationary) process is said to be mean-ergodic (or usually simply shortened to “ergodic”) if the ensemble average equals the time average of almost all sample paths. What this means is that if we take the mean function (which is necessarily a constant  $\mu_X$  due to stationarity), and compare it to the time average of a sample path (for example the time-average of the continuous-time sample path  $x(t)$ ), then they will converge to each other in an appropriate sense. If we take this to be in the mean-squared sense, then

$$\lim_{T \rightarrow \infty} \mathbb{E} \left[ \left| \mu_X - \frac{1}{T} \int_0^T X(t) dt \right|^2 \right] = 0 \quad (5.2.40)$$

Note the we could alternatively define the time-average as  $\frac{1}{T} \int_{T/2}^{-T/2} X(t) dt$ . For a discrete-time stochastic process starting at time  $k = 0$ , the mean-ergodicity condition is

$$\lim_{N \rightarrow \infty} \mathbb{E} \left[ \left| \mu_X - \frac{1}{N} \sum_{k=0}^{N-1} X_k \right|^2 \right] = 0 \quad (5.2.41)$$

We can also consider stronger forms of mean-ergodicity, such as in the almost sure sense. In discrete-time, this requires that

$$\mu_X = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} x_k \quad (5.2.42)$$

for almost all sample paths. Or in other words,

$$\Pr \left( \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} X_k = \mu_X \right) = 1 \quad (5.2.43)$$

As an example, the strong law of large numbers immediately establishes that any sequence of i.i.d. random variables with finite mean is a mean-ergodic stochastic process. The mean-ergodicity property can be thought of as the strongest possible characterisation of a sequence which satisfies a law of large numbers.

### Mean-Ergodicity of Wide-Sense Stationary Processes

A wide-sense stationary continuous-time process  $X(t)$  is mean-ergodic in the mean-square sense if the time-average of a sample path:

$$\bar{\mu} = \frac{1}{T} \int_{T/2}^{-T/2} X(t) dt \quad (5.2.44)$$

converges in mean to the ensemble average  $\mu = \mathbb{E}[X(t)]$ , and the variance of  $\bar{\mu}$  converges to zero. Convergence in mean is guaranteed because

$$\mathbb{E}[\bar{\mu}] = \mathbb{E} \left[ \frac{1}{T} \int_{T/2}^{-T/2} X(t) dt \right] \quad (5.2.45)$$

$$= \frac{1}{T} \int_{T/2}^{-T/2} \mathbb{E}[X(t)] dt \quad (5.2.46)$$

$$= \mu \quad (5.2.47)$$

For the variance of  $\bar{\mu}$ , we write

$$\text{Var}(\bar{\mu}) = \mathbb{E}[(\bar{\mu} - \mu)^2] \quad (5.2.48)$$

$$= \mathbb{E} \left[ \left( \frac{1}{T} \int_{T/2}^{-T/2} (X(t) - \mu) dt \right) \left( \frac{1}{T} \int_{T/2}^{-T/2} (X(s) - \mu) ds \right) \right] \quad (5.2.49)$$

$$= \frac{1}{T^2} \int_{T/2}^{-T/2} \int_{T/2}^{-T/2} \mathbb{E}[(X(t) - \mu)(X(s) - \mu)] dt ds \quad (5.2.50)$$

$$= \frac{1}{T^2} \int_{T/2}^{-T/2} \int_{T/2}^{-T/2} C(t-s) dt ds \quad (5.2.51)$$

where  $C(\tau)$  is the autocovariance function of the wide-sense stationary process. Performing a change of variables  $\tau = t - s$ , it can be shown (as in the Wiener-Khintchine theorem) that this integral equals

$$\text{Var}(\bar{\mu}) = \frac{1}{T} \int_T^{-T} \left( 1 - \frac{|\tau|}{T} \right) C(\tau) d\tau \quad (5.2.52)$$

$$\leq \frac{1}{T} \int_{-\infty}^{\infty} \left| 1 - \frac{|\tau|}{T} \right| |C(\tau)| d\tau \quad (5.2.53)$$

$$\leq \frac{1}{T} \int_{-\infty}^{\infty} |C(\tau)| d\tau \quad (5.2.54)$$

If  $\int_{-\infty}^{\infty} |C(\tau)| d\tau$  is bounded (meaning  $C(\tau)$  is absolutely integrable), then

$$\lim_{T \rightarrow \infty} \text{Var}(\bar{\mu}) = 0 \quad (5.2.55)$$

which satisfies the mean-square convergence criterion. Hence a sufficient condition for a wide-sense stationary process to be mean-ergodic in the mean-square sense is for the covariance function to be absolutely integrable. Roughly speaking,  $C(\tau)$  should decay ‘quickly’ in  $\tau$ .

### Autocovariance-Ergodicity

Analogously to mean-ergodicity, a stationary (or wide-sense stationary) process  $X(t)$  is said to be autocovariance-ergodic (in an appropriate sense) if the time-average estimate of the autocovariance function converges (in that sense) to the actual autocovariance  $C_X(\tau)$ . Note that the autocovariance function may be written in terms of the time difference  $\tau$  since the process is wide-sense stationary. In the mean-square sense, this condition requires

$$\lim_{T \rightarrow \infty} \mathbb{E} \left[ \left| C_X(\tau) - \frac{1}{T} \int_0^T (X(t + \tau) - \mu_X)(X(t) - \mu_X) dt \right|^2 \right] = 0 \quad (5.2.56)$$

for any  $\tau$ .

### Wide-Sense Ergodicity

A stochastic process which is both mean-ergodic and autocovariance-ergodic can be said to be wide-sense ergodic.

### Stationarity Does Not Necessarily Imply Ergodicity

Although ergodicity is usually characterised as being ‘stronger’ than stationarity, not all stationary processes are ergodic. Consider the stochastic process defined by  $X(t) = Y$  where  $Y$  is a random variable with finite mean. That is, a realisation of  $X(t)$  is a constant function with value drawn from the distribution of  $Y$ . Then  $X(t)$  is stationary because first-order distribution is the same for all time (namely, the distribution of  $Y$ ) and moreover, the  $n^{\text{th}}$ -order distributions do not change with time. However,  $X(t)$  is not ergodic because it is not possible to reconstruct the mean function  $\mu_X = \mathbb{E}[Y]$  from a single realisation  $x(t) = y$ .

#### 5.2.4 Karhunen-Loëve Theorem

### 5.3 Families of Stochastic Processes

#### 5.3.1 Bernoulli Processes

A Bernoulli process is a discrete-time discrete-valued (in  $\{0, 1\}$ ) process in which each value of the process is an i.i.d. Bernoulli ( $p$ ) random variable, i.e.

$$\Pr(\mathbb{I}_k = x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases} \quad (5.3.1)$$

### 5.3.2 Binomial Processes

A binomial process is a counting process (i.e. discrete-valued on the non-negative integers) that is the summation along a Bernoulli process, so that  $X_n$  is a Binomial  $(n, p)$  random variable. That is,

$$\Pr(X_n = x) = \sum_{k=1}^n \mathbb{I}_k \quad (5.3.2)$$

#### Memorylessness of Binomial Process

The waiting time between increments in a binomial process will be geometrically distributed, using the natural characterisation of the geometric distribution. Because of the memorylessness property of the geometric distribution, we can also state a memorylessness property of the binomial process:

$$\Pr(X_{n_2} = k_2 | X_{n_1} = k_1) = \Pr(X_{n_2} - X_{n_1} = k_2 - k_1) \quad (5.3.3)$$

### 5.3.3 Poisson Processes

The Poisson point process is a counting process based on the Poisson distribution. It is a discrete-valued, continuous-time process with rate parameter  $\lambda$ , where  $\lambda$  is the average number of occurrences per unit time. Thus a Poisson point process may be specified by  $X(t) \sim \text{Poisson}(\lambda t)$  and

$$\Pr(X(t) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \quad (5.3.4)$$

with independent increments, i.e. for any  $t_1, \dots, t_n$ , the increments  $X(t_2) - X(t_1), \dots, X(t_n) - X(t_{n-1})$  are independent. We can also relate arrivals in a Poisson point process with exponentially distributed waiting times. Since a Poisson distribution is characterised by a small independent chance of arrival at each time instant, while the exponential distribution is memoryless, then the distribution of time between arrivals in a Poisson point process should be exponentially distributed (by analogy to the binomial process and geometric distribution). To confirm this intuition, we recall the Erlang distribution as the sum of  $k$  i.i.d.  $\text{Exp}(\lambda)$  random variables, with PDF

$$f(t; k, \lambda) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!} \quad (5.3.5)$$

$$= \lambda \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} \quad (5.3.6)$$

This already resembles a Poisson distribution. But to make this connection explicit, consider the probability for a Poisson point process  $X(t)$ , that when the  $k^{\text{th}}$  arrival occurs (call this time  $T_k$ ), it occurs within the infinitesimally small interval  $(t, t + dt]$ . Via the Poisson process, this is given by the probability that there are  $k-1$  arrivals by time  $t$ , followed by the probability that there is exactly one arrival in  $(t, t + dt]$ :

$$\Pr(t < T_k \leq t + dt) = \Pr(X(t) = k-1) \Pr(X(t+dt) - X(t) = 1) \quad (5.3.7)$$

$$= \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} \cdot \Pr(X(dt) = 1) \quad (5.3.8)$$

$$= \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} \cdot \frac{(\lambda dt)^1 e^{-\lambda dt}}{1!} \quad (5.3.9)$$

$$= \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} \cdot \lambda dt \quad (5.3.10)$$

where we are allowed to take  $e^{-\lambda dt} = 1$  since  $dt$  is infinitesimally small. Via the Erlang approach, the probability is the probability that the sum of  $k$  waiting times is in the interval  $(t, t + dt]$ :

$$\Pr(t < T_k \leq t + dt) = f(t; k, \lambda) dt \quad (5.3.11)$$

$$= \lambda \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} dt \quad (5.3.12)$$

which is the same expression as above. Hence the Poisson point process is consistent with exponentially distributed waiting times in between arrivals.

### Memorylessness of Poisson Process

Since the Poisson point process has exponentially distribution waiting times, we can state a memoryless property for Poisson point processes:

$$\Pr(X(t_2) = k_2 | X(t_1) = k_1) = \Pr(X(t_2) - X(t_1) = k_2 - k_1) \quad (5.3.13)$$

### Inhomogeneous Poisson Processes [212]

In an inhomogeneous Poisson process, rather than a constant rate parameter  $\lambda$ , we have an *intensity function*  $\lambda(t)$ . This interpretation here is now that the probability of arrival in the small interval  $(t, t + dt]$  is

$$\Pr(X(t + dt) - X(t) = 1) = \lambda(t) dt \quad (5.3.14)$$

Thus, the average rate between the interval  $[0, t]$  is

$$\bar{\lambda}(t) = \frac{1}{t} \int_0^t \lambda(s) ds \quad (5.3.15)$$

and the mean function extends  $\mathbb{E}[X(t)] = \lambda t$  in the case of homogeneous Poisson processes by

$$m(t) := \mathbb{E}[X(t)] \quad (5.3.16)$$

$$= \bar{\lambda}(t) t \quad (5.3.17)$$

$$= \int_0^t \lambda(s) ds \quad (5.3.18)$$

such that  $X(t) \sim \text{Poisson}(m(t))$ .

#### 5.3.4 Random Walks

Let  $X_1, \dots, X_n$  be a sequence of i.i.d. random variables. We usually assume that these variables have been standardised, so that they have mean zero and variance one. Define the cumulative sum

$$S_n = \sum_{i=1}^n X_i \quad (5.3.19)$$

We call the process  $S_n$  (indexed in  $n$ ) a random walk. There is also a recursive form of the random walk:

$$S_{n+1} = S_n + X_{n+1} \quad (5.3.20)$$

with initial value  $S_0 = 0$ .

### 5.3.5 Gaussian Processes

Gaussian processes can be discrete-time or continuous-time. In discrete-time, we say that  $X_n$  is a Gaussian process if the random vector

$$\mathbf{X} = [X_{n_1} \ \dots \ X_{n_k}]^\top \quad (5.3.21)$$

is a multivariate Gaussian for any indices  $n_1, \dots, n_k$ . Likewise in continuous-time, we say that  $X(t)$  is a Gaussian process if the random vector

$$\mathbf{X} = [X(t_1) \ \dots \ X(t_k)]^\top \quad (5.3.22)$$

is a multivariate Gaussian for any times  $t_1, \dots, t_k$ . That is, the finite-order distributions of a Gaussian process are always multivariate Gaussian.

#### Gaussian White Noise

A simple example of a Gaussian process is called Gaussian white noise. In discrete-time, Gaussian white noise is when each  $X_i$  is i.i.d. (hence uncorrelated)  $\mathcal{N}(0, 1)$ .

### 5.3.6 Wiener Process

The Wiener process is a continuous-time Gaussian process which is a formalisation of Brownian motion, and the two terms may be used interchangeably. The (standard) Wiener process satisfies:

- $W(0) = 0$
- $W(t) - W(s) \sim \mathcal{N}(0, |t-s|)$
- Increments are independent, i.e.  $W(t_2) - W(t_1)$  is independent of  $W(t_1)$ .

By this definition, the standard Wiener process has mean function

$$\mathbb{E}[W(t)] = \mathbb{E}[W(t) - W(0)] \quad (5.3.23)$$

$$= 0 \quad (5.3.24)$$

and variance function

$$\text{Var}(W(t)) = \text{Var}(W(t) - W(0)) \quad (5.3.25)$$

$$= t \quad (5.3.26)$$

The autocovariance function for  $t > s$  is

$$\text{Cov}(W(s), W(t)) = \mathbb{E}[W(s)W(t)] \quad (5.3.27)$$

$$= \mathbb{E}[W(s)(W(t) - W(s))] + \mathbb{E}[W(s)^2] \quad (5.3.28)$$

$$= \text{Cov}(W(s), W(t) - W(s)) + \text{Var}(W(s)) \quad (5.3.29)$$

$$= s \quad (5.3.30)$$

since  $\text{Cov}(W(s), W(t) - W(s)) = 0$ . Or more generally, we have

$$\text{Cov}(W(s), W(t)) = \min\{s, t\} \quad (5.3.31)$$

By this, we can see that the Wiener process is not stationary. The second-order joint distribution is given by

$$\begin{bmatrix} W(t) \\ W(s) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} t & \min\{s, t\} \\ \min\{s, t\} & s \end{bmatrix}\right) \quad (5.3.32)$$

and by applying the conditioning formula when  $t > s$ , we have the conditional distribution

$$[W(t)|W(s)] \sim \mathcal{N}(W(s), t-s) \quad (5.3.33)$$

which is consistent with the definition of the Wiener process.

The Wiener process can be considered as a limit of a random walk, as we ‘squish’ points together in an appropriate way. Indeed, we can write for some small time increment  $\tau$ :

$$W(t+\tau) = W(t) + Z \quad (5.3.34)$$

where  $Z$  is an independent  $\mathcal{N}(0, \tau)$  increment. Because a random walk is formed by the cumulative sum of an i.i.d. sequence, this gives rise to Gaussian distributed increments in the Wiener process, due to the Central Limit Theorem.

A non-standard Wiener process can be obtained by a scaling  $X(t) = \sigma W(t)$ , where

$$X(t) - X(s) = \sigma(W(t) - W(s)) \quad (5.3.35)$$

$$\sim \mathcal{N}(0, \sigma^2 |t-s|) \quad (5.3.36)$$

### Quadratic Variation of Wiener Process

The quadratic variation of the Wiener process can be calculated to be

$$[W](t) = t \quad (5.3.37)$$

To show this, denote the sum

$$V_n = \sum_{k=1}^n (W(t_k) - W(t_{k-1}))^2 \quad (5.3.38)$$

Taking expectations, we have, using properties of the Wiener process:

$$\mathbb{E}[V_n] = \sum_{k=1}^n \mathbb{E}[(W(t_k) - W(t_{k-1}))^2] \quad (5.3.39)$$

$$= \sum_{k=1}^n \text{Var}(W(t_k) - W(t_{k-1})) \quad (5.3.40)$$

$$= \sum_{k=1}^n (t_k - t_{k-1}) \quad (5.3.41)$$

$$= t_n - t_0 \quad (5.3.42)$$

$$= t \quad (5.3.43)$$

Taking the variance instead, and since the Wiener process has independent increments:

$$\text{Var}(V_n) = \sum_{k=1}^n \text{Var}((W(t_k) - W(t_{k-1}))^2) \quad (5.3.44)$$

To compute this, note from the properties of the Wiener process that

$$\frac{W(t_k) - W(t_{k-1})}{\sqrt{t_k - t_{k-1}}} \sim \mathcal{N}(0, 1) \quad (5.3.45)$$

Thus

$$\frac{(W(t_k) - W(t_{k-1}))^2}{t_k - t_{k-1}} \sim \chi_1^2 \quad (5.3.46)$$

and from the variance of the chi-squared distribution, we have

$$\text{Var} \left( \frac{(W(t_k) - W(t_{k-1}))^2}{t_k - t_{k-1}} \right) = 2 \quad (5.3.47)$$

hence

$$\text{Var}((W(t_k) - W(t_{k-1}))^2) = 2(t_k - t_{k-1})^2 \quad (5.3.48)$$

Therefore

$$\text{Var}(V_n) = 2 \sum_{k=1}^n (t_k - t_{k-1})^2 \quad (5.3.49)$$

$$= 2 \sum_{k=1}^n \left( \frac{k}{n}t - \frac{k-1}{n}t \right)^2 \quad (5.3.50)$$

$$= 2 \sum_{k=1}^n \left( \frac{t}{n} \right)^2 \quad (5.3.51)$$

$$= \frac{2t^2}{n} \quad (5.3.52)$$

In the limit, we see

$$\lim_{n \rightarrow \infty} \mathbb{E}[V_n] = t \quad (5.3.53)$$

$$\lim_{n \rightarrow \infty} \text{Var}(V_n) = 0 \quad (5.3.54)$$

These are sufficient conditions for  $V_n \xrightarrow{\text{P}} t$ , so we can say that the quadratic variation of  $W(t)$  on the interval  $[0, T]$  is  $T$ .

### Exponential Brownian Motion

Let  $W(t)$  be the Wiener process. Then the process  $X(t) = e^{W(t)}$  will have a lognormal distribution at any  $t$ , and is called exponential Brownian motion. Since  $W(t) \sim \mathcal{N}(0, t)$  at any  $t$ , we can compute the moments of  $X(t)$  using the moment generating function of the Gaussian distribution:

$$\mathbb{E}[e^{sW(t)}] = e^{s^2t/2} \quad (5.3.55)$$

In particular,

$$\mathbb{E}[X(t)] = \mathbb{E}[e^{W(t)}] \quad (5.3.56)$$

$$= e^{t/2} \quad (5.3.57)$$

and

$$\mathbb{E}[X(t)^2] = \mathbb{E}[e^{2W(t)}] \quad (5.3.58)$$

$$= e^{2t} \quad (5.3.59)$$

So the mean grows over time, and the variance of exponential Brownian motion which is

$$\text{Var}(X(t)) = \mathbb{E}[X(t)^2] - \mathbb{E}[X(t)]^2 \quad (5.3.60)$$

$$= e^{2t} - e^t \quad (5.3.61)$$

also grows over time. In the way that the Wiener process can be characterised as the continuous-time limit of random walk (i.e. a sum with infinitesimally small random increments), we can characterise exponential Brownian motion as the continuous-time limit of a product of infinitesimally small percentage changes.

### Geometric Brownian Motion

Related to exponential Brownian motion is geometric Brownian motion, which may defined as

$$X(t) = \exp\left(\sigma W(t) + \left(\mu - \frac{\sigma^2}{2}\right)t\right) \quad (5.3.62)$$

where  $\mu$  is called the drift and  $\sigma$  is called the volatility. The exact relation can be seen by

$$X(t) = e^{(\mu - \sigma^2/2)t} \cdot (e^{W(t)})^\sigma \quad (5.3.63)$$

where notice that  $e^{W(t)}$  is the usual exponential Brownian motion, thus exponential Brownian motion can be considered a special case of geometric Brownian motion when  $\sigma = 1$  and  $\mu = 1/2$ . We can determine the mean and variance in the same way, since the exponent is Gaussian with

$$\sigma W(t) + \left(\mu - \frac{\sigma^2}{2}\right)t \sim \mathcal{N}\left(\left(\mu - \frac{\sigma^2}{2}\right)t, \sigma^2 t\right) \quad (5.3.64)$$

Hence its moment generating function is

$$\mathbb{E}\left[\exp\left(s\left[\sigma W(t) + \left(\mu - \frac{\sigma^2}{2}\right)t\right]\right)\right] = \exp\left(s\left(\mu - \frac{\sigma^2}{2}\right)t + s^2 \frac{\sigma^2}{2}t\right) \quad (5.3.65)$$

and we have

$$\mathbb{E}[X(t)] = \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \frac{\sigma^2}{2}t\right) \quad (5.3.66)$$

$$= e^{\mu t} \quad (5.3.67)$$

as well as

$$\mathbb{E}[X(t)^2] = \exp\left(2\left(\mu - \frac{\sigma^2}{2}\right)t + 4\frac{\sigma^2}{2}t\right) \quad (5.3.68)$$

$$= e^{2\mu t + \sigma^2 t} \quad (5.3.69)$$

Therefore

$$\text{Var}(X(t)) = \mathbb{E}[X(t)^2] - \mathbb{E}[X(t)] \quad (5.3.70)$$

$$= e^{2\mu t + \sigma^2 t} - e^{2\mu t} \quad (5.3.71)$$

$$= e^{2\mu t} (e^{\sigma^2 t} - 1) \quad (5.3.72)$$

So in geometric Brownian motion, whether the mean and variance grows or shrinks over time depends on the drift  $\mu$ , while the volatility  $\sigma$  controls how quickly the variance grows/shrinks.

## Brownian Bridge

The Brownian bridge  $B(t)$  is a stochastic process on  $[0, T]$  that can be characterised as the Wiener process conditioned on the value of  $W(T)$  to be zero:

$$B(t) = [W(t)|W(T) = 0] \quad (5.3.73)$$

Since  $W(0) = 0$ , this means that for the Brownian bridge,  $B(0) = B(T) = 0$ . In other words, the Brownian bridge looks like a Wiener process, except it has been ‘pinned’ or ‘anchored’ to zero at the endpoints 0 and  $T$ .

Another way to characterise the Brownian bridge is via

$$B(t) = W(t) - \frac{t}{T}W(T) \quad (5.3.74)$$

So to realise a Brownian bridge, we can first realise a Wiener process, and then subtract  $\frac{t}{T}W(T)$  from the function. Note that this formula evaluates to  $B(0) = 0$  and  $B(T) = 0$  as well. To demonstrate why this characterisation and the above one via conditioning are the same, first consider the third-order joint distribution of the Wiener process

$$\begin{bmatrix} W(t_1) \\ W(t_2) \\ W(T) \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} t_1 & t_1 & t_1 \\ t_1 & t_2 & t_2 \\ t_1 & t_2 & T \end{bmatrix} \right) \quad (5.3.75)$$

where  $0 \leq t_1 \leq t_2 \leq T$ . Using the [conditioning formula for Gaussians](#), the conditional distribution of  $W(t_1), W(t_2)$  given  $W(T) = 0$  is

$$\begin{bmatrix} W(t_1) \\ W(t_2) \end{bmatrix} \Big| W(T) = 0 \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} t_1 & t_1 \\ t_1 & t_2 \end{bmatrix} - \frac{1}{T} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \begin{bmatrix} t_1 & t_2 \end{bmatrix} \right) \quad (5.3.76)$$

$$\sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} t_1 - t_1^2/T & t_1 - t_1 t_2/T \\ t_1 - t_1 t_2/T & t_2 - t_2^2/T \end{bmatrix} \right) \quad (5.3.77)$$

Then consider second-order distribution of the other characterisation, which is given by

$$\begin{bmatrix} B(t_1) \\ B(t_2) \end{bmatrix} = \begin{bmatrix} W(t_1) - \frac{t_1}{T}W(T) \\ W(t_2) - \frac{t_2}{T}W(T) \end{bmatrix} \quad (5.3.78)$$

$$= \begin{bmatrix} 1 & 0 & -t_1/T \\ 0 & 1 & -t_2/T \end{bmatrix} \begin{bmatrix} W(t_1) \\ W(t_2) \\ W(T) \end{bmatrix} \quad (5.3.79)$$

This is a linear transformation of a Gaussian, thus  $(B(t_1), B(t_2))$  is also Gaussian with mean

$$\mathbb{E} \left[ \begin{bmatrix} B(t_1) \\ B(t_2) \end{bmatrix} \right] = \mathbf{0} \quad (5.3.80)$$

and covariance

$$\text{Cov} \left( \begin{bmatrix} B(t_1) \\ B(t_2) \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 & -t_1/T \\ 0 & 1 & -t_2/T \end{bmatrix} \begin{bmatrix} t_1 & t_1 & t_1 \\ t_1 & t_2 & t_2 \\ t_1 & t_2 & T \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -t_1/T & -t_2/T \end{bmatrix} \quad (5.3.81)$$

$$= \begin{bmatrix} 1 & 0 & -t_1/T \\ 0 & 1 & -t_2/T \end{bmatrix} \begin{bmatrix} t_1 - t_1^2/T & t_1 - t_1 t_2/T \\ t_1 - t_1 t_2/T & t_2 - t_2^2/T \\ 0 & 0 \end{bmatrix} \quad (5.3.82)$$

$$= \begin{bmatrix} t_1 - t_1^2/T & t_1 - t_1 t_2/T \\ t_1 - t_1 t_2/T & t_2 - t_2^2/T \end{bmatrix} \quad (5.3.83)$$

This matches the mean and covariance derived earlier via conditioning.

The variance of the Brownian bridge may be written as

$$\text{Var}(B(t)) = t \left(1 - \frac{t}{T}\right) \quad (5.3.84)$$

so we can see that the variance is zero at  $t = 0$  and  $t = T$ , while it is maximised in the middle at  $t = T/2$ .

### 5.3.7 Dirichlet Processes

## 5.4 Branching Processes

### 5.4.1 Galton-Watson Processes

## 5.5 Renewal Theory

## 5.6 Central Limit Theorems

### 5.6.1 Lindberg-Levy Central Limit Theorem [106]

Let  $X_1, X_2, \dots, X_n$  be random variables that are i.i.d. with  $X$ , where  $\mathbb{E}[X] = \mu$  and  $\text{Var}(X) = \sigma^2$ . Define the sum  $S_n := X_1 + \dots + X_n$ . Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\text{d}} Z \sim \mathcal{N}(0, 1^2) \quad (5.6.1)$$

as  $n \rightarrow \infty$ . Equivalently, if we define the ‘normalised’ sequence  $\tilde{X}_n := \frac{X_n - \mu}{\sigma}$ , then

$$\frac{1}{\sqrt{n}} (\tilde{X}_1 + \dots + \tilde{X}_n) \xrightarrow{\text{d}} Z \sim \mathcal{N}(0, 1^2) \quad (5.6.2)$$

as  $n \rightarrow \infty$ . Note that

$$\mathbb{E}[\tilde{X}] = 0 \quad (5.6.3)$$

$$\text{Var}(\tilde{X}) = \mathbb{E}[\tilde{X}^2] = 1 \quad (5.6.4)$$

$$\mathbb{E}\left[\frac{\tilde{X}}{\sqrt{n}}\right] = 0 \quad (5.6.5)$$

$$\text{Var}\left(\frac{\tilde{X}}{\sqrt{n}}\right) = \mathbb{E}\left[\frac{\tilde{X}^2}{n}\right] = \frac{1}{n} \quad (5.6.6)$$

*Proof.* It suffices to show that the characteristic function of the sum  $\frac{\tilde{X}_1}{\sqrt{n}} + \dots + \frac{\tilde{X}_n}{\sqrt{n}}$  converges to the characteristic function of the standard Gaussian distribution as  $n \rightarrow \infty$ . Firstly, an  $n^{\text{th}}$  order Taylor expansion of the characteristic function of an arbitrary random variable  $Y$  about  $t = 0$  is given by

$$\varphi_Y(t) = \sum_{k=0}^n \frac{t^k \varphi_Y^{(k)}(0)}{k!} + o(t^n) \quad (5.6.7)$$

where  $\varphi_Y^{(k)}(t)$  denotes the  $k^{\text{th}}$  derivative of  $\varphi_Y(t)$ . Note that the Little-o notation means that the term  $o(t^n)$  goes to zero faster than  $t^n$  as  $t \rightarrow 0$ . Evaluating the  $k^{\text{th}}$  derivative gives

$$\varphi_Y^{(k)}(t) = \mathbb{E} [(-iY)^k e^{-itY}] \quad (5.6.8)$$

and in particular  $\varphi_Y^{(k)}(0) = \mathbb{E} [(-iY)^k]$ . Hence

$$\varphi_Y(t) = \sum_{k=0}^n \frac{(-it)^k \mathbb{E}[Y^k]}{k!} + o(t^n) \quad (5.6.9)$$

and for the random variable  $\tilde{X}/\sqrt{n}$ , the second order Taylor expansion is:

$$\varphi_{\tilde{X}/\sqrt{n}}(t) = 1 + \frac{-it\mathbb{E}[\tilde{X}/\sqrt{n}]}{1} - \frac{t^2\mathbb{E}[\tilde{X}^2/n]}{2} + o\left(\frac{t^2}{n}\right) \quad (5.6.10)$$

$$= 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \quad (5.6.11)$$

Since the characteristic function of a sum of random variables is the product of the characteristic functions, we apply this and take  $n \rightarrow \infty$ .

$$\lim_{n \rightarrow \infty} \left[ \varphi_{\tilde{X}/\sqrt{n}}(t) \right]^n = \lim_{n \rightarrow \infty} \left[ 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \right]^n \quad (5.6.12)$$

$$= e^{-t^2/2} \quad (5.6.13)$$

which gives the characteristic function of the standard Gaussian distribution, where we have also used the fact  $\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = e^x$ .  $\square$

The Central Limit Theorem also implies another equivalent statement, which can be obtained by scaling the random variables:

$$\sqrt{n} \left( \frac{S_n}{n} - \mu \right) \xrightarrow{\text{d}} \sigma Z \sim \mathcal{N}(0, \sigma^2) \quad (5.6.14)$$

We can also make large-sample approximations of sums of random variables, by making appropriate scalings and shiftings:

$$S_n - n\mu \approx \sigma\sqrt{n}Z \sim \mathcal{N}(0, \sigma^2 n) \quad (5.6.15)$$

$$S_n \approx \sigma\sqrt{n}Z + n\mu \sim \mathcal{N}(n\mu, \sigma^2 n) \quad (5.6.16)$$

$$\frac{S_n}{n} \approx \frac{\sigma\sqrt{n}Z + n\mu}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad (5.6.17)$$

### 5.6.2 De Moivre-Laplace Theorem [171]

The De Moivre-Laplace Theorem is a special case of the Central Limit Theorem to Bernoulli random variables (and hence is a formal result on the normal approximation to the binomial distribution). For a sequence of i.i.d. Bernoulli trials  $X_1, \dots, X_n$  with  $\Pr(X_i = 1) = p$ , then

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \xrightarrow{\text{d}} \mathcal{N}(0, 1) \quad (5.6.18)$$

as  $n \rightarrow \infty$ .

### 5.6.3 Lindberg-Feller Central Limit Theorem

### 5.6.4 Multivariate Central Limit Theorem [125]

Let  $\mathbf{X}_1, \mathbf{X}_2, \dots$  be i.i.d. copies of  $\mathbf{X} \in \mathbb{R}^k$ , with mean vector  $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$  and positive definite covariance matrix  $\text{Cov}(\mathbf{X}) = \Sigma$ . Let

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \quad (5.6.19)$$

Then

$$\sqrt{n} (\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \quad (5.6.20)$$

as  $n \rightarrow \infty$ .

*Proof.* For convenience, denote a random vector  $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ . By appealing to the Cramér-Wold theorem, it suffices to show that

$$\mathbf{a}^\top \sqrt{n} (\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{a}^\top \mathbf{Y} \quad (5.6.21)$$

for all vectors  $\mathbf{a}^\top \in \mathbb{R}^k$ . Now note that  $\mathbb{E}[\mathbf{a}^\top \mathbf{X}] = \mathbf{a}^\top \boldsymbol{\mu}$ ,  $\text{Var}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \Sigma \mathbf{a}$  and  $\text{Var}(\mathbf{a}^\top \mathbf{Y}) = \mathbf{a}^\top \Sigma \mathbf{a}$ . Since  $\mathbf{a}^\top \bar{\mathbf{X}}_n = \sum_{i=1}^n \mathbf{a}^\top \mathbf{X}_i / n$  is a scalar, by applying the univariate central limit theorem we get

$$\sqrt{n} (\mathbf{a}^\top \bar{\mathbf{X}}_n - \mathbf{a}^\top \boldsymbol{\mu}) \xrightarrow{d} \mathbf{a}^\top \mathbf{Y} \quad (5.6.22)$$

$$\sim \mathcal{N}(0, \mathbf{a}^\top \Sigma \mathbf{a}) \quad (5.6.23)$$

for any  $\mathbf{a}^\top \in \mathbb{R}^k$ , which completes the proof.  $\square$

### 5.6.5 Finite Population Central Limit Theorem

#### Finite Population Sample Mean

Consider a finite population of size  $N$ , and sample of size  $n$  denoted  $\{X_1, X_2, \dots, X_n\}$ . Then the sampling distribution of the sample mean  $\bar{X}$  must be developed separately to the infinite population case. Suppose the population consists of the set  $\{y_1, y_2, \dots, y_N\}$ , then the population mean and variance are given by

$$\mu = \frac{1}{N} \sum_{i=1}^N y_i \quad (5.6.24)$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 \quad (5.6.25)$$

Due to the linearity of expectation, we can show for the unbiasedness of the sample mean:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \quad (5.6.26)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \quad (5.6.27)$$

$$= \mu \quad (5.6.28)$$

To derive the variance of the sampling distribution, first introduce the  $N$  indicator random variables  $Z_1, Z_2, \dots, Z_N$  for whether the respective population elements are in the sample. Then  $\bar{X} = \sum_{i=1}^N y_i Z_i / n$  and

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^N y_i Z_i\right) \quad (5.6.29)$$

$$= \frac{1}{n^2} \left[ \sum_{i=1}^N y_i^2 \text{Var}(Z_i) + 2 \sum_{i,j:i < j} y_i y_j \text{Cov}(Z_i Z_j) \right] \quad (5.6.30)$$

Each  $Z_i$  has a marginal distribution which is a Bernoulli distribution with mean  $\mathbb{E}[Z_i] = n/N$  and variance  $\text{Var}(Z_i) = (1 - n/N) n/N$  and also for any  $i \neq j$ :

$$\mathbb{E}[Z_i Z_j] = \Pr(Z_i = 1, Z_j = 1) \quad (5.6.31)$$

$$= \frac{n}{N} \cdot \frac{n-1}{N-1} \quad (5.6.32)$$

Hence this gives

$$\text{Cov}(Z_i Z_j) = \mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i] \mathbb{E}[Z_j] \quad (5.6.33)$$

$$= \frac{n}{N} \cdot \frac{n-1}{N-1} - \frac{n^2}{N^2} \quad (5.6.34)$$

$$= \frac{Nn^2 - Nn - n^2 N + n^2}{N^2(N-1)} \quad (5.6.35)$$

$$= -\frac{n(N-n)}{N^2(N-1)} \quad (5.6.36)$$

$$= -\frac{1}{N-1} \cdot \frac{n}{N} \left(1 - \frac{n}{N}\right) \quad (5.6.37)$$

Computing the variance of the sampling distribution:

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \left[ \sum_{i=1}^N y_i^2 \frac{n}{N} \left(1 - \frac{n}{N}\right) - 2 \sum_{i,j:i < j} y_i y_j \frac{1}{N-1} \cdot \frac{n}{N} \left(1 - \frac{n}{N}\right) \right] \quad (5.6.38)$$

$$= \frac{1}{n^2} \cdot \frac{n}{N} \left(1 - \frac{n}{N}\right) \left[ \sum_{i=1}^N y_i^2 - \frac{2}{N-1} \sum_{i,j:i < j} y_i y_j \right] \quad (5.6.39)$$

$$(5.6.40)$$

Note from the population variance that

$$\sum_{i=1}^N (y_i^2 - \mu)^2 = \sum_{i=1}^N y_i^2 - N\mu^2 \quad (5.6.41)$$

$$= \sum_{i=1}^N y_i^2 - N \left( \frac{1}{N} \sum_{i=1}^N y_i \right)^2 \quad (5.6.42)$$

$$= \sum_{i=1}^N y_i^2 - \frac{1}{N} \left( \sum_{i=1}^N y_i^2 + 2 \sum_{i,j:i < j} y_i y_j \right) \quad (5.6.43)$$

$$= \frac{N-1}{N} \sum_{i=1}^N y_i^2 - \frac{2}{N} \sum_{i,j:i < j} y_i y_j \quad (5.6.44)$$

So multiplying out by  $\frac{N}{N-1}$ :

$$\frac{N}{N-1} \sum_{i=1}^N (y_i^2 - \mu)^2 = \sum_{i=1}^N y_i^2 - \frac{2}{N-1} \sum_{i,j:i < j} y_i y_j \quad (5.6.45)$$

and therefore

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \cdot \frac{n}{N} \left(1 - \frac{n}{N}\right) \left( \frac{N}{N-1} \sum_{i=1}^N (y_i^2 - \mu)^2 \right) \quad (5.6.46)$$

$$= \frac{1}{n} \cdot \left(\frac{N-n}{N}\right) \left(\frac{N}{N-1} \sigma^2\right) \quad (5.6.47)$$

$$= \left(\frac{N-n}{N-1}\right) \frac{\sigma^2}{n} \quad (5.6.48)$$

which is the infinite population variance multiplied by  $\frac{N-n}{N-1}$ . Hence the  $\frac{N-n}{N-1}$  is known as a finite population correction factor.

### Approximate Normality of Sample Mean from Finite Population

Under further technical assumptions [215], the sampling distribution of the sample mean from a finite population converges in distribution to a normal distribution as the population size  $N \rightarrow \infty$ . This is not too absurd to imagine, since a finite population closely resembles an infinite population when  $N$  is large. Thus for large  $N$  and  $n$ , we can approximate the sampling distribution of the sample mean from a finite population with

$$\bar{X} \xrightarrow{\text{approx.}} \mathcal{N}\left(\mu, \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}\right) \quad (5.6.49)$$

#### 5.6.6 Functional Central Limit Theorem

For the random walk  $S_n = \sum_{i=1}^n X_i$  where the  $X_i$  are standardised, consider the rescaled random walk

$$W_n(t) := \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \quad (5.6.50)$$

$$= \sum_{i=1}^{\lfloor nt \rfloor} \frac{X_i}{\sqrt{n}} \quad (5.6.51)$$

for  $t \in [0, 1]$ . With  $t = 1$ , the usual central limit theorem gives us  $W_n(1) \xrightarrow{d} \mathcal{N}(0, 1)$  as  $n \rightarrow \infty$ . However, the functional central limit theorem (also known as Donsker's theorem) asserts something stronger, which is that as a function,

$$W_n(t) \xrightarrow{d} W(t) \quad (5.6.52)$$

where  $W(t)$  is the standard Wiener process on  $[0, 1]$ . We can show this holds pointwise, because

$$W_n(t) = \sum_{i=1}^{\lfloor nt \rfloor} \frac{X_i}{\sqrt{n}} \quad (5.6.53)$$

$$= \sum_{i=1}^{\lfloor nt \rfloor} \frac{X_i}{\sqrt{\lfloor nt \rfloor}} \cdot \frac{\sqrt{\lfloor nt \rfloor}}{\sqrt{n}} \quad (5.6.54)$$

but we have  $\sum_{i=1}^{\lfloor nt \rfloor} X_i / \sqrt{\lfloor nt \rfloor} \xrightarrow{d} \mathcal{N}(0, 1)$  and moreover  $\sqrt{\lfloor nt \rfloor / n} \rightarrow \sqrt{t}$ , thus pointwise, using Slutsky's theorem,

$$W_n(t) \xrightarrow{d} \mathcal{N}(0, t) \quad (5.6.55)$$

A further argument for why the entire function converges is that for any pair  $s, t \in [0, 1]$  with  $t > s$ , we have

$$W_n(t) - W_n(s) = \sum_{i=1}^{\lfloor nt \rfloor} \frac{X_i}{\sqrt{n}} - \sum_{i=1}^{\lfloor ns \rfloor} \frac{X_i}{\sqrt{n}} \quad (5.6.56)$$

$$= \sum_{i=\lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \frac{X_i}{\sqrt{n}} \quad (5.6.57)$$

and since the  $X_i$  are i.i.d., we have equality in law:

$$\sum_{i=\lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \frac{X_i}{\sqrt{n}} \stackrel{\text{st}}{=} \sum_{i=1}^{\lfloor nt \rfloor - \lfloor ns \rfloor} \frac{X_i}{\sqrt{n}} \quad (5.6.58)$$

$$= \sum_{i=1}^{\lfloor nt \rfloor - \lfloor ns \rfloor} \frac{X_i}{\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}} \cdot \frac{\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}}{\sqrt{n}} \quad (5.6.59)$$

and applying similar reasoning as above,

$$\sum_{i=1}^{\lfloor nt \rfloor - \lfloor ns \rfloor} \frac{X_i}{\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}} \xrightarrow{d} \mathcal{N}(0, 1) \quad (5.6.60)$$

$$\frac{\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}}{\sqrt{n}} \rightarrow t - s \quad (5.6.61)$$

so

$$W_n(t) - W_n(s) \xrightarrow{d} \mathcal{N}(0, t - s) \quad (5.6.62)$$

This, along with  $W_n(t)$  having independent increments (due to it being a random walk), satisfies the definition of the Wiener process. With the functional central limit theorem, we are able to formally characterise the Wiener process as a continuous-time random walk with infinitesimal increments.

### 5.6.7 Stationary Process Central Limit Theorem [2]

#### Linear Process Central Limit Theorem

## 5.7 Concentration Inequalities

A concentration inequality quantifies how a random variable  $X$  deviates about its mean  $\mu$ . Concentration inequalities usually take the form of an upper bound on  $\Pr(|X - \mu| > t)$ , in terms of some  $t \geq 0$ . Chebychev's inequality is an example of a concentration inequality.

### 5.7.1 Sub-Gaussian Random Variables

Recall the two-sided tail bounds for a Gaussian:

$$\Pr(|X - \mu| > t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad (5.7.1)$$

for all  $t \geq 0$ . This was derived due to the moment generating function of the zero-mean Gaussian  $X - \mu$  satisfying

$$\mathbb{E} [e^{\lambda(X-\mu)}] = \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \quad (5.7.2)$$

We can use this property to describe a whole class of random variables that generalise Gaussian random variables, called sub-Gaussian random variables. The random variable  $X$  with mean  $\mu$  is said to be sub-Gaussian with parameter  $\sigma$  if the moment generating function of the centered random variable  $X - \mu$  satisfies

$$\mathbb{E} [e^{\lambda(X-\mu)}] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \quad (5.7.3)$$

for all  $\lambda \in \mathbb{R}$ . The quantity  $\sigma^2$  is also referred to as the *variance factor*. Naturally, we can apply the same Chernoff technique as for bounding the Gaussian tail to obtain first the upper deviation inequality:

$$\Pr(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad (5.7.4)$$

for all  $t \geq 0$ . Now note the symmetry of the definition for sub-Gaussian random variables. By considering both  $\lambda$  and  $-\lambda$ , this means that  $X$  is sub-Gaussian if and only if  $-X$  is sub-Gaussian. Thus we can similarly arrive at the lower deviation inequality:

$$\Pr(X - \mu \leq -t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad (5.7.5)$$

for all  $\lambda \in \mathbb{R}$ . Putting these together, we see that any sub-Gaussian random variable must satisfy the same tail inequality for Gaussians (albeit with weak inequality inside the probability, since discrete random variables may also be sub-Gaussian):

$$\Pr(|X - \mu| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad (5.7.6)$$

Intuitively, we can think of sub-Gaussian random variables as having tails which ‘decay’ at least as fast as Gaussian random variables (of comparable variance factor), since the moment generating function is upper bounded by that of a Gaussian. The idea is similar to a distribution being platykurtic.

### Characterisation of Sub-Gaussian Random Variables [30]

One way to characterise a sub-Gaussian random variable  $X$  with variance factor  $v$  is by taking the maximum of both tail probabilities:

$$\max \{\Pr(X - \mu \geq t), \Pr(X - \mu \leq -t)\} \leq \exp\left(-\frac{t^2}{2v}\right) \quad (5.7.7)$$

Another characterisation of sub-Gaussian random variables is in terms of upper bounds on the even central moments (e.g. variance, kurtosis, etc.) is as follows.

**Theorem 5.3.** *Consider a centered random variable  $X$  (i.e.  $\mathbb{E}[X] = 0$ ) with variance factor  $v$ . Then for every integer  $q \geq 1$ , we have*

$$\mathbb{E}[X^{2q}] \leq q! (4v)^q \quad (5.7.8)$$

*Proof.* Without loss of generality, assume  $v = 1$  since its only role is as a scaling. Then we have

$$\mathbb{E}[X^{2q}] = \mathbb{E}\left[\int_0^{X^{2q}} 1 \cdot dx\right] \quad (5.7.9)$$

$$= \mathbb{E}\left[\int_0^\infty \mathbb{I}_{\{X^{2q} > x\}} dx\right] \quad (5.7.10)$$

$$= \int_0^\infty \mathbb{E}[\mathbb{I}_{\{X^{2q} > x\}}] dx \quad (5.7.11)$$

$$= \int_0^\infty \Pr(X^{2q} > x) dx \quad (5.7.12)$$

Now apply the change of variables  $x = y^{2q}$  so that  $dx = 2qy^{2q-1}dy$ . Then

$$\mathbb{E}[X^{2q}] = 2q \int_0^\infty y^{2q-1} \Pr(X^{2q} > y^{2q}) dy \quad (5.7.13)$$

$$= 2q \int_0^\infty y^{2q-1} \Pr(|X| > y) dy \quad (5.7.14)$$

$$\leq 2q \int_0^\infty y^{2q-1} 2 \exp\left(-\frac{y^2}{2}\right) dy \quad (5.7.15)$$

from  $X$  being sub-Gaussian. Apply another change of variables  $y = \sqrt{2}t^{1/2}$  so that  $dy = \sqrt{2}t^{-1/2}/2dt$  and

$$\mathbb{E}[X^{2q}] \leq 4q \int_0^\infty \left(\sqrt{2}t^{1/2}\right)^{2q-1} \exp(-t) \frac{\sqrt{2}}{2} t^{-1/2} dt \quad (5.7.16)$$

$$= 2^2 q \int_0^\infty 2^q \cdot 2^{-1/2} \cdot t^q \cdot t^{-1/2} \cdot 2^{-1/2} \cdot t^{-1/2} e^{-t} dt \quad (5.7.17)$$

$$= 2^{q+1} q \int_0^\infty (t)^{q-1} e^{-t} dt \quad (5.7.18)$$

$$= 2^{q+1} q (q-1)! \quad (5.7.19)$$

$$= 2^{q+1} q! \quad (5.7.20)$$

using the definition of the Gamma function. With the fact  $2^{q+1} q! = 2 \cdot 2^q q! \leq 4^q q!$ , this implies

$$\mathbb{E}[X^{2q}] \leq q! 4^q \quad (5.7.21)$$

□

It turns out that this type of bound on the even moments is an equivalent characterisation of sub-Gaussian random variables, as we can show the following converse result (albeit with a worse constant).

**Theorem 5.4.** Consider a centered random variable  $X$  (i.e.  $\mathbb{E}[X] = 0$ ) that for every integer  $q \geq 1$  satisfies

$$\mathbb{E}[X^{2q}] \leq q! C^q \quad (5.7.22)$$

for some constant  $C$ . Then  $X$  is sub-Gaussian with variance factor  $v = 4C$ .

*Proof.* Introduce  $X'$  as an independent copy of  $X$ , and consider the moment generating function of  $X - X'$ , which can be written as

$$\mathbb{E}[e^{\lambda X}] \mathbb{E}[e^{-\lambda X}] = \mathbb{E}[e^{\lambda(X-X')}] \quad (5.7.23)$$

$$= \mathbb{E} \left[ \sum_{q=0}^{\infty} \frac{[\lambda(X - X')]^q}{q!} \right] \quad (5.7.24)$$

$$= \sum_{q=0}^{\infty} \frac{\lambda^q \mathbb{E}[(X - X')^q]}{q!} \quad (5.7.25)$$

using the series definition of the exponential. Then since  $X - X'$  is symmetric about zero, then  $\mathbb{E}[(X - X')^q] = 0$  for all odd  $q$ . Thus

$$\mathbb{E}[e^{\lambda X}] \mathbb{E}[e^{-\lambda X}] = \sum_{q=0}^{\infty} \frac{\lambda^{2q} \mathbb{E}[(X - X')^{2q}]}{(2q)!} \quad (5.7.26)$$

Using the fact that  $x^{2q}$  is convex in  $x$ ,

$$(X - X')^{2q} = \left( \frac{1}{2} \cdot 2X - \frac{1}{2} \cdot 2X' \right)^{2q} \quad (5.7.27)$$

$$\leq \frac{1}{2} (2X)^{2q} - \frac{1}{2} (2X')^{2q} \quad (5.7.28)$$

$$\leq \frac{1}{2} (2X)^{2q} + \frac{1}{2} (2X')^{2q} \quad (5.7.29)$$

$$= 2^{2q-1} (X^{2q} - X'^{2q}) \quad (5.7.30)$$

so

$$\mathbb{E}[(X - X')^{2q}] \leq 2^{2q-1} (\mathbb{E}[X^{2q}] + \mathbb{E}[X'^{2q}]) \quad (5.7.31)$$

$$= 2^{2q} (\mathbb{E}[X^{2q}]) \quad (5.7.32)$$

as  $\mathbb{E}[X^{2q}] = \mathbb{E}[X'^{2q}]$ . Now by the assumption that  $\mathbb{E}[X^{2q}] \leq q!C^q$ , this implies  $\mathbb{E}[(X - X')^{2q}] \leq 2^{2q}q!C^q$  and we have the bound

$$\mathbb{E}[e^{\lambda X}] \mathbb{E}[e^{-\lambda X}] \leq \sum_{q=0}^{\infty} \frac{\lambda^{2q} 2^{2q} q! C^q}{(2q)!} \quad (5.7.33)$$

From Jensen's inequality (with convex  $e^{-\lambda x}$ ) and the fact that  $\mathbb{E}[X] = 0$ , this gives  $\mathbb{E}[e^{-\lambda X}] \geq e^{-\lambda \mathbb{E}[X]} = 1$ . Also,

$$\frac{(2q)!}{q!} = \prod_{j=1}^q (q+j) \quad (5.7.34)$$

$$\geq \prod_{j=1}^q (j+j) \quad (5.7.35)$$

$$= \prod_{j=1}^q 2j \quad (5.7.36)$$

$$= 2^q q! \quad (5.7.37)$$

Applying these inequalities yields

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[e^{\lambda X}] \mathbb{E}[e^{-\lambda X}] \quad (5.7.38)$$

$$\leq \sum_{q=0}^{\infty} \frac{\lambda^{2q} 2^{2q} q! C^q}{(2q)!} \quad (5.7.39)$$

$$\leq \sum_{q=0}^{\infty} \frac{\lambda^{2q} 2^{2q} C^q}{2^q q!} \quad (5.7.40)$$

$$= \sum_{q=0}^{\infty} \frac{(2\lambda^2 C)^q}{q!} \quad (5.7.41)$$

$$= e^{2\lambda^2 C} \quad (5.7.42)$$

Let  $C = v/4$ , then

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\lambda^2 v}{4}\right) \quad (5.7.43)$$

which is the condition for  $X$  to be sub-Gaussian with variance factor  $v = 4C$ .  $\square$

Note that since  $e^{\lambda^2 v/2} \leq e^{\lambda^2 v'/2}$  for  $v \leq v'$ , then  $X$  being sub-Gaussian with variance factor  $v$  also implies that  $X$  is sub-Gaussian with variance factor  $v'$ .

### Hoeffding's Inequality for Sub-Gaussian Random Variables

Recall that Hoeffding's inequality bounds the probability of deviation for sums of independent bounded random variables. Hoeffding's inequality can be extended to sum of independent sub-Gaussian random variables (of which bounded random variables are a part of). First we consider two sub-Gaussian random variables  $X_1$  and  $X_2$  with means  $\mu_1$  and  $\mu_2$  respectively, and with variance factors  $\sigma_1$  and  $\sigma_2$ . We can demonstrate the sum of independent sub-Gaussian random variables  $X_1 + X_2$  will also be sub-Gaussian, and additive in the variance factor. From the property that the moment generating function becomes multiplicative for sums of independent random variables, then by applying the property of sub-Gaussianity for each of  $X_1$  and  $X_2$ , we have

$$\mathbb{E}[e^{\lambda(X_1+X_2)}] = \mathbb{E}[e^{\lambda X_1}] \mathbb{E}[e^{\lambda X_2}] \quad (5.7.44)$$

$$\leq \exp\left(\frac{\sigma_1^2 \lambda^2}{2}\right) \exp\left(\frac{\sigma_2^2 \lambda^2}{2}\right) \quad (5.7.45)$$

$$= \exp\left(\frac{(\sigma_1^2 + \sigma_2^2) \lambda^2}{2}\right) \quad (5.7.46)$$

Thus  $X_1 + X_2$  also satisfies the requirements to be sub-Gaussian, with variance factor  $\sigma_1^2 + \sigma_2^2$ . From applying the Chernoff technique, it also follows that

$$\Pr(X_1 + X_2 - \mu_1 - \mu_2 \geq t) \leq \exp\left(-\frac{t^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \quad (5.7.47)$$

for all  $t \geq 0$ . Generalising, we can say that the sum of independent sub-Gaussian random variables  $X_1, \dots, X_n$  satisfies the concentration inequality

$$\Pr\left(\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right) \quad (5.7.48)$$

for all  $t \geq 0$ . Analogously to the traditional Hoeffding inequality, we can also denote  $S = \sum_{i=1}^n X_i$ ,  $\mathbb{E}[S] = \sum_{i=1}^n \mu_i$  and give the two-sided concentration inequality

$$\Pr(|S - \mathbb{E}[S]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right) \quad (5.7.49)$$

as well as the concentration inequality for the sample mean  $\bar{X} = \sum_{i=1}^n X_i/n$  of i.i.d. random variables:

$$\Pr(|\bar{X} - \mu| \geq t) \leq 2 \exp\left(-\frac{nt^2}{2\sigma^2}\right) \quad (5.7.50)$$

Therefore if we wanted to use this to arrive back at the traditional Hoeffding inequality, it suggests that we should show that a bounded random variable bounded between  $[a, b]$  has variance factor no greater than  $(b - a)^2 / 4$ .

This result is also somewhat like a non-asymptotic analogue of the central limit theorem applicable for sub-Gaussian random variables, which says that the tail probability of a sample mean is bounded by that of a Gaussian. Unlike the central limit theorem, this gives a way to obtain to obtain exact (i.e. not approximate) confidence intervals for the population mean (provided  $\sigma^2$  is known or assumed) valid for finite  $n$ , since

$$\Pr(|\bar{X} - \mu| < t) \geq 1 - 2 \exp\left(-\frac{nt^2}{2\sigma^2}\right) \quad (5.7.51)$$

So putting  $2 \exp\left(-\frac{nt^2}{2\sigma^2}\right) = \alpha$ , this is rearranged as  $t = \sqrt{\frac{2\sigma^2}{n} \log\left(\frac{2}{\alpha}\right)}$  so

$$\Pr\left(|\bar{X} - \mu| < \sqrt{\frac{2\sigma^2}{n} \log\left(\frac{2}{\alpha}\right)}\right) \geq 1 - \alpha \quad (5.7.52)$$

and

$$\bar{X} - \sqrt{\frac{2\sigma^2}{n} \log\left(\frac{2}{\alpha}\right)} < \mu < \bar{X} + \sqrt{\frac{2\sigma^2}{n} \log\left(\frac{2}{\alpha}\right)} \quad (5.7.53)$$

is a confidence interval for  $\mu$  with confidence level at least  $1 - \alpha$ . However, when comparing against asymptotic confidence intervals constructed at the same sample size  $n$ , these tend to be more conservative (i.e. wider). Also, hypothesis tests performed using the same idea will result in low power. Regardless, Hoeffding's inequality is still useful for providing a qualitative description of the rate at which the confidence interval shrinks as  $n$  increases, which is of the order  $O(n^{-1/2})$ .

### Sub-Gaussian Random Vectors [207]

Note that if  $\mathbf{X}$  is a multivariate Gaussian random vector, then  $\mathbf{a}^\top \mathbf{X}$  will be a univariate Gaussian random variable for any vector  $\mathbf{a}$ . Thus this gives a natural way to generalise sub-Gaussian random variables to sub-Gaussian random vectors. We say that  $d$ -dimensional random vector  $\mathbf{X}$  is sub-Gaussian if  $\mathbf{a}^\top \mathbf{X}$  is a sub-Gaussian random variable for any  $\mathbf{a} \in \mathbb{R}^d$ . By letting  $\mathbf{a}$  be each of the unit basis vectors, this also means that the marginal random variables of a sub-Gaussian random vector will each be sub-Gaussian random variables.

We define the variance factor of  $\mathbf{X}$  as the supremum of the variance factor of  $\mathbf{a}^\top \mathbf{X}$  over the hypersphere where  $\|\mathbf{a}\|_2 = 1$ . That is, letting  $v_{\mathbf{X}}$  denote the variance factor of  $\mathbf{X}$ , we explicitly write

$$v_{\mathbf{X}} = \sup_{\|\mathbf{a}\|_2=1} v_{\mathbf{a}^\top \mathbf{X}} \quad (5.7.54)$$

### 5.7.2 Sub-Exponential Random Variables

### 5.7.3 Sub-Gamma Random Variables [30]

The cumulant generating function of a centered gamma distribution  $\psi(\lambda)$  satisfies the upper bound

$$\psi(\lambda) \leq \frac{\lambda^2 v}{2(1 - c\lambda)} \quad (5.7.55)$$

valid for  $\lambda \in (0, 1/c)$ , where  $v$  is the variance of the Gamma distribution and  $c$  is the scale parameter. Using this characterisation, we can consider a more general class of distributions also satisfying this property, which roughly speaking, describes that the right tail of the distribution decays as fast as that of a Gamma distribution. We say that a centered random variable  $X$  is sub-Gamma on the right with variance factor  $v$  and scale factor  $c$  if for all  $\lambda \in (0, 1/c)$ , its cumulant generating function satisfies

$$\psi_X(\lambda) \leq \frac{\lambda^2 v}{2(1 - c\lambda)} \quad (5.7.56)$$

In that case, we denote  $X \in \Gamma_+(v, c)$ . Note that sub-Gaussian random variables can be considered special cases of sub-Gamma random variables because if we let  $c \rightarrow 0$ , then  $\psi_X(\lambda) \leq \frac{\lambda^2 v}{2}$  which is the property required by sub-Gaussian random variables.

A random variable that is sub-Gamma on the right will satisfy the concentration inequality

$$\Pr(X \geq \sqrt{2vt} + ct) \leq e^t \quad (5.7.57)$$

for all  $t > 0$ .

*Proof.* Via the Chernoff bound, we have for any  $\lambda \in (0, 1/c)$ :

$$\Pr(X > t) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda X}] = \exp(-\lambda t + \psi_X(\lambda)) \quad (5.7.58)$$

$$= \exp(-\lambda t + \psi_X(\lambda)) \quad (5.7.59)$$

$$\leq \exp(-\lambda t + \bar{\psi}_X(\lambda)) \quad (5.7.60)$$

where

$$\bar{\psi}_X(\lambda) = \frac{\lambda^2 v}{2(1 - c\lambda)} \quad (5.7.61)$$

is the upper bound for the cumulant generating function. Optimising the bound with respect to  $\lambda$ ,

$$\Pr(X > t) \leq \inf_{\lambda \in (0, 1/c)} \{\exp(-\lambda t + \bar{\psi}_X(\lambda))\} \quad (5.7.62)$$

$$= \exp\left(-\sup_{\lambda \in (0, 1/c)} \{\lambda t - \bar{\psi}_X(\lambda)\}\right) \quad (5.7.63)$$

$$= \exp(-\psi_X^*(t)) \quad (5.7.64)$$

where we call

$$\psi_X^*(t) = \sup_{\lambda \in (0, 1/c)} \{\lambda t - \bar{\psi}_X(\lambda)\} \quad (5.7.65)$$

the Cramer transform. We proceed to optimise this quantity. Differentiating  $\bar{\psi}_X(\lambda)$  using the product rule, we have

$$\frac{d\bar{\psi}_X(\lambda)}{d\lambda} = \frac{d}{d\lambda} \left( \frac{\lambda^2 v}{2} \cdot \frac{1}{1 - c\lambda} \right) \quad (5.7.66)$$

$$= \frac{\lambda v}{1 - c\lambda} + \frac{\lambda^2 v}{2} \cdot \frac{c}{(1 - c\lambda)^2} \quad (5.7.67)$$

Setting the derivative to zero,

$$\frac{d}{d\lambda} (t\lambda - \bar{\psi}_X(\lambda)) = 0 \quad (5.7.68)$$

$$t = \frac{\lambda v}{1 - c\lambda} + \frac{\lambda^2 v c}{2(1 - c\lambda)^2} \quad (5.7.69)$$

This can be rearranged into a quadratic equation in  $\lambda$ :

$$(1 - c\lambda)^2 t = \lambda v (1 - c\lambda) + \frac{\lambda^2 v c}{2} \quad (5.7.70)$$

$$t - 2c\lambda t + c^2\lambda^2 t = \lambda v - cv\lambda^2 + \frac{\lambda^2 v c}{2} \quad (5.7.71)$$

$$\left(\frac{vc}{2} - c^2 t - cv\right) \lambda^2 + (v + 2ct) \lambda - t = 0 \quad (5.7.72)$$

$$\left(-\frac{vc}{2} - c^2 t\right) \lambda^2 + (v + 2ct) \lambda - t = 0 \quad (5.7.73)$$

The general solution from the quadratic formula is

$$\lambda = \frac{-(v + 2ct) \pm \sqrt{(v + 2ct)^2 - 4(-\frac{vc}{2} - c^2 t)(-t)}}{2(-\frac{vc}{2} - c^2 t)} \quad (5.7.74)$$

$$= \frac{-(v + 2ct) \pm \sqrt{v^2 + 4vct + 4c^2 t^2 - 2vct - 4c^2 t^2}}{-vc - 2c^2 t} \quad (5.7.75)$$

$$= \frac{-(v + 2ct) \pm \sqrt{v^2 + 2vct}}{-vc - 2c^2 t} \quad (5.7.76)$$

$$= \frac{1}{c} \left( \frac{-(v + 2ct) \pm \sqrt{v^2 + 2vct}}{-(v + 2ct)} \right) \quad (5.7.77)$$

$$= \frac{1}{c} \left( 1 \pm \frac{\sqrt{v}\sqrt{v + 2ct}}{(v + 2ct)} \right) \quad (5.7.78)$$

$$= \frac{1}{c} \left( 1 \pm \sqrt{\frac{v}{v + 2ct}} \right) \quad (5.7.79)$$

However since  $\lambda$  is constrained to  $(0, 1/c)$ , we must take the solution:

$$\lambda = \frac{1}{c} \left( 1 - \sqrt{\frac{v}{v + 2ct}} \right) \quad (5.7.80)$$

$$= \frac{1}{c} \left( 1 - \frac{1}{\sqrt{1 + 2ct/v}} \right) \quad (5.7.81)$$

Substituting this back into  $\bar{\psi}_X(\lambda)$ ,

$$\bar{\psi} \left( \frac{1}{c} \left( 1 - \frac{1}{\sqrt{1 + 2ct/v}} \right) \right) = \frac{v}{2c^2} \left( 1 - \frac{1}{\sqrt{1 + 2ct/v}} \right)^2 \frac{1}{1 - c \cdot \frac{1}{c} \left( 1 - \frac{1}{\sqrt{1 + 2ct/v}} \right)} \quad (5.7.82)$$

$$= \frac{v}{2c^2} \left( 1 - \frac{2}{\sqrt{1 + 2ct/v}} + \frac{1}{1 + 2ct/v} \right) \frac{1}{\frac{1}{\sqrt{1 + 2ct/v}}} \quad (5.7.83)$$

$$= \frac{v}{2c^2} \left( \frac{1+2ct/v}{1+2ct/v} - \frac{2}{\sqrt{1+2ct/v}} + \frac{1}{1+2ct/v} \right) \sqrt{1+2\frac{ct}{v}} \quad (5.7.84)$$

$$= \frac{v}{2c^2} \left( \frac{2+2ct/v}{1+2ct/v} - \frac{2}{\sqrt{1+2ct/v}} \right) \sqrt{1+2\frac{ct}{v}} \quad (5.7.85)$$

$$= \frac{v}{2c^2} \left( \frac{2+2ct/v}{\sqrt{1+2ct/v}} - 2 \right) \quad (5.7.86)$$

Thus the Cramer transform is

$$\psi_X^*(t) = \frac{t}{c} \left( 1 - \frac{1}{\sqrt{1+2ct/v}} \right) - \frac{v}{2c^2} \left( \frac{2+2ct/v}{\sqrt{1+2ct/v}} - 2 \right) \quad (5.7.87)$$

$$= \frac{t}{c} - \frac{t/c}{\sqrt{1+2ct/v}} - \frac{v+ct}{c^2\sqrt{1+2ct/v}} + \frac{v}{c^2} \quad (5.7.88)$$

$$= \frac{t}{c} + \frac{v}{c^2} - \frac{t/c+v/c^2+t/c}{\sqrt{1+2ct/v}} \quad (5.7.89)$$

$$= \frac{t}{c} + \frac{v}{c^2} - \frac{2t/c+v/c^2}{\sqrt{1+2ct/v}} \quad (5.7.90)$$

$$= \frac{v}{c^2} \left( 1 + \frac{ct}{v} - \frac{2ct/v+1}{\sqrt{1+2ct/v}} \right) \quad (5.7.91)$$

$$= \frac{v}{c^2} \left( 1 + \frac{ct}{v} - \sqrt{1+2\frac{ct}{v}} \right) \quad (5.7.92)$$

Introduce the strictly increasing function

$$h(u) = 1 + u - \sqrt{1+2u} \quad (5.7.93)$$

so that by letting  $u = ct/v$ ,

$$\psi_X^*(t) = \frac{v}{c^2} h\left(\frac{ct}{v}\right) \quad (5.7.94)$$

Therefore

$$\Pr(X > t) \leq \exp\left(-\frac{v}{c^2} h\left(\frac{ct}{v}\right)\right) \quad (5.7.95)$$

To make the upper bound  $e^{-t}$ , we seek the inverse of  $h(\cdot)$ . Let

$$w = h^{-1}(u) \quad (5.7.96)$$

so

$$u = h(w) \quad (5.7.97)$$

$$= 1 + w - \sqrt{1+2w} \quad (5.7.98)$$

Solving for  $w$ , we find

$$(1+w-u)^2 = 1+2w \quad (5.7.99)$$

$$1+w-u+w+w^2-wu-u-wu+u^2 = 1+2w \quad (5.7.100)$$

$$w^2 - 2wu - 2u + u^2 = 0 \quad (5.7.101)$$

which is a quadratic equation in  $w$ . Then applying the quadratic formula,

$$w = \frac{-(-2u) \pm \sqrt{4u^2 - 4(-2u + u^2)}}{2} \quad (5.7.102)$$

$$= \frac{2u \pm 2\sqrt{2u}}{2} \quad (5.7.103)$$

$$= u \pm \sqrt{2u} \quad (5.7.104)$$

Hence

$$h^{-1}(u) = u + \sqrt{2u} \quad (5.7.105)$$

because the inverse must also be strictly increasing. Now to find the inverse of  $\psi_X^*(t)$ , we rearrange:

$$\frac{c^2}{v}\psi_X^*(t) = h\left(\frac{ct}{v}\right) \quad (5.7.106)$$

$$h^{-1}\left(\frac{c^2}{v}\psi_X^*(t)\right) = \frac{ct}{v} \quad (5.7.107)$$

$$\frac{c^2}{v}\psi_X^*(t) + \sqrt{2\frac{c^2}{v}\psi_X^*(t)} = \frac{ct}{v} \quad (5.7.108)$$

$$t = \sqrt{2v\psi_X^*(t)} + c\psi_X^*(t) \quad (5.7.109)$$

Thus

$$\psi_X^{*-1}(t) = \sqrt{2vt} + ct \quad (5.7.110)$$

From  $\Pr(X > t) \leq \exp(-\psi_X^*(t))$  we can instead equivalently write

$$\Pr(X > \psi_X^{*-1}(t)) \leq \exp(-\psi_X^{*-1}(\psi_X^*(t))) \quad (5.7.111)$$

which gives

$$\Pr(X > \sqrt{2vt} + ct) \leq e^{-t} \quad (5.7.112)$$

□

#### 5.7.4 Azuma-Hoeffding Inequality [55]

The Azuma-Hoeffding inequality is closely related to the Hoeffding inequality, except it relaxes the requirement of independence on the random variables.

**Theorem 5.5.** Consider a sequence of random variables  $X_0, X_1, \dots$  such that

$$\mathbb{E}[X_i | X_0, \dots, X_{i-1}] = X_{i-1} \quad (5.7.113)$$

for all  $i \geq 1$ . Or succinctly using random vector notation  $\mathbf{X}_{i-1} = (X_0, \dots, X_{i-1})$ ,

$$\mathbb{E}[X_i | X_0, \dots, X_{i-1}] = X_{i-1} \quad (5.7.114)$$

for all  $i \geq 1$ . This sequence is known as a ‘martingale’. Suppose that the martingale satisfies the ‘bounded difference’ condition:

$$a_i \leq X_i - X_{i-1} \leq b_i \quad (5.7.115)$$

for each  $i \geq 1$ . Then for any  $t > 0$  we have

$$\Pr(X_n > X_0 + t) \leq \exp\left[-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right] \quad (5.7.116)$$

$$\Pr(X_n < X_0 - t) \leq \exp \left[ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right] \quad (5.7.117)$$

and additionally

$$\Pr(|X_n - X_0| > t) \leq 2 \exp \left[ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right] \quad (5.7.118)$$

*Proof.* Assume without loss of generality that  $X_0 = 0$ , otherwise we can just apply the same steps to the translated sequence  $X'_i = X_i - X_0$ . Introduce the martingale difference sequence (i.e. sequence formed by the difference in a martingale):

$$Y_i = X_i - X_{i-1} \quad (5.7.119)$$

which by the bounded difference condition implies  $a_i \leq Y_i \leq b_i$ . Note that the random variable  $\mathbb{E}[Y_i | \mathbf{X}_{i-1}]$  satisfies

$$\mathbb{E}[Y_i | \mathbf{X}_{i-1}] = \mathbb{E}[X_i | \mathbf{X}_{i-1}] - \mathbb{E}[X_{i-1} | \mathbf{X}_{i-1}] \quad (5.7.120)$$

$$= X_{i-1} - X_{i-1} \quad (5.7.121)$$

$$= 0 \quad (5.7.122)$$

using the property of the martingale. As it satisfies the prerequisites, we can apply Hoeffding's lemma to  $[Y_i | \mathbf{X}_{i-1}]$ , which is a random variable with the same distribution as the conditional distribution of  $Y_i$  given  $\mathbf{X}_{i-1}$ . This yields

$$\mathbb{E}[e^{\lambda Y_i} | \mathbf{X}_{i-1}] \leq \exp \left[ \frac{\lambda^2 (b_i - a_i)^2}{8} \right] \quad (5.7.123)$$

for any  $\lambda \in \mathbb{R}$ , where we know  $a_i \leq [Y_i | \mathbf{X}_{i-1}] \leq b_i$  since the bound  $a_i \leq Y_i \leq b_i$  holds unconditionally. Now consider applying the Law of Iterated Expectations to the moment generating function of  $X_i$ :

$$\mathbb{E}[e^{\lambda X_i}] = \mathbb{E}[e^{\lambda(X_{i-1} + Y_i)}] \quad (5.7.124)$$

$$= \mathbb{E}[\mathbb{E}[e^{\lambda(X_{i-1} + Y_i)} | \mathbf{X}_{i-1}]] \quad (5.7.125)$$

$$= \mathbb{E}[e^{\lambda X_{i-1}} \mathbb{E}[e^{\lambda Y_i} | \mathbf{X}_{i-1}]] \quad (5.7.126)$$

$$\leq \mathbb{E}[e^{\lambda X_{i-1}}] \exp \left[ \frac{\lambda^2 (b_i - a_i)^2}{8} \right] \quad (5.7.127)$$

By induction, we have  $\mathbb{E}[e^{\lambda X_0}] = 1$  by the assumption  $X_0 = 0$ , and we have also shown directly above that  $\mathbb{E}[e^{\lambda X_i}] \leq \mathbb{E}[e^{\lambda X_{i-1}}] \exp \left[ \lambda^2 (b_i - a_i)^2 / 8 \right]$ , hence

$$\mathbb{E}[e^{\lambda X_n}] \leq \prod_{i=1}^n \exp \left[ \frac{\lambda^2 (b_i - a_i)^2}{8} \right] \quad (5.7.128)$$

$$= \exp \left[ \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2 \right] \quad (5.7.129)$$

$$= e^{\lambda^2 c / 8} \quad (5.7.130)$$

where  $c = \sum_{i=1}^n (b_i - a_i)^2$ . As in the proof of the Hoeffding inequality, we state  $\Pr(X_n > t) = \Pr(e^{\lambda X_n} > e^{\lambda t})$  for any  $\lambda > 0$ . Then using Markov's inequality for the non-negative random variable  $e^{\lambda X_n}$ :

$$\Pr(X_n > t) = \Pr(e^{\lambda X_n} > e^{\lambda t}) \quad (5.7.131)$$

$$\leq e^{-\lambda t} \mathbb{E} [e^{\lambda X_n}] \quad (5.7.132)$$

$$\leq e^{-\lambda t} e^{\lambda^2 c/8} \quad (5.7.133)$$

Following the same steps as in the Hoeffding inequality, this bound is optimised with choice of  $\lambda = 4t/c > 0$ , which yields

$$\Pr (X_n > t) \leq \exp \left[ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right] \quad (5.7.134)$$

Returning  $X_0 = 0$  to the inequality gives us

$$\Pr (X_n > X_0 + t) \leq \exp \left[ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right] \quad (5.7.135)$$

If we apply the same steps to the negated sequence  $-X_i$  we get

$$\Pr (-X_n > t) \leq \exp \left[ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right] \quad (5.7.136)$$

which rearranges to

$$\Pr (X_n < X_0 - t) \leq \exp \left[ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right] \quad (5.7.137)$$

Finally putting the bounds together using the union bound (Boole's inequality), we arrive at

$$\Pr (\{X_n > X_0 + t\} \cup \{X_n < X_0 - t\}) = \Pr (|X_n - X_0| > t) \quad (5.7.138)$$

$$\leq 2 \exp \left[ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right] \quad (5.7.139)$$

□

### 5.7.5 McDiarmid's Inequality [140]

McDiarmid's inequality builds on the Azuma-Hoeffding inequality and can be used to bound the deviation of a function of independent random variables from its mean.

**Theorem 5.6.** *Let  $X_1, \dots, X_m$  be mutually independent random variables. Suppose for a scalar-valued function  $f(X_1, \dots, X_m)$ , and for each  $i = 1, \dots, m$  we have almost surely*

$$|f(X_1, \dots, X_i, \dots, X_m) - f(X_1, \dots, x'_i, \dots, X_m)| \leq c_i \quad (5.7.140)$$

for any  $x'_i$  in the support of  $X_i$ . For succinctness, denote  $\mathbf{X}_m = (X_1, \dots, X_m)$ . Then for any  $t > 0$ , we have

$$\Pr (f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)] > t) \leq \exp \left[ -\frac{2t^2}{\sum_{i=1}^m c_i^2} \right] \quad (5.7.141)$$

$$\Pr (f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)] < -t) \leq \exp \left[ -\frac{2t^2}{\sum_{i=1}^m c_i^2} \right] \quad (5.7.142)$$

and additionally

$$\Pr (|f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)]| > t) \leq 2 \exp \left[ -\frac{2t^2}{\sum_{i=1}^m c_i^2} \right] \quad (5.7.143)$$

Note that the condition  $|f(X_1, \dots, X_i, \dots, X_m) - f(X_1, \dots, x'_i, \dots, X_m)| \leq c_i$  can be interpreted as saying that there is bounded deviation in  $f(\cdot)$  by altering one of its arguments.

*Proof.* Introduce the zero-mean random variable

$$Y_m = f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)] \quad (5.7.144)$$

and for given  $m$  let

$$Z_1 := \mathbb{E}[Y_m | X_1] - \mathbb{E}[Y_m] \quad (5.7.145)$$

and

$$Z_k := \mathbb{E}[Y_m | \mathbf{X}_k] - \mathbb{E}[Y_m | \mathbf{X}_{k-1}] \quad (5.7.146)$$

for  $k = 2, \dots, m$ . Using the Law of Iterated Expectations, we see

$$\mathbb{E}[\mathbb{E}[Y_m | \mathbf{X}_k] | \mathbf{X}_{k-1}] = \mathbb{E}[Y_m | \mathbf{X}_{k-1}] \quad (5.7.147)$$

so

$$\mathbb{E}[Z_k | \mathbf{X}_{k-1}] = \mathbb{E}[\mathbb{E}[Y_m | \mathbf{X}_k] - \mathbb{E}[Y_m | \mathbf{X}_{k-1}] | \mathbf{X}_{k-1}] \quad (5.7.148)$$

$$= \mathbb{E}[Y_m | \mathbf{X}_{k-1}] - \mathbb{E}[Y_m | \mathbf{X}_{k-1}] \quad (5.7.149)$$

$$= 0 \quad (5.7.150)$$

Moreover, we can show

$$Z_k = \mathbb{E}[Y_m | \mathbf{X}_k] - \mathbb{E}[Y_m | \mathbf{X}_{k-1}] \quad (5.7.151)$$

$$= \mathbb{E}[f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)] | \mathbf{X}_k] - \mathbb{E}[f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)] | \mathbf{X}_{k-1}] \quad (5.7.152)$$

$$= \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_k] - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] + \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] \quad (5.7.153)$$

$$= \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_k] - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] \quad (5.7.154)$$

Hence  $Z_k$  can be considered a martingale difference sequence for the martingale sequence  $\mathbb{E}[Y_m | \mathbf{X}_k]$  (indexed in  $k$ ). Also note that

$$\mathbb{E}[Y_m | \mathbf{X}_m] = \mathbb{E}[f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)] | \mathbf{X}_m] \quad (5.7.155)$$

$$= f(\mathbf{X}_m) - \mathbb{E}[f(\mathbf{X}_m)] \quad (5.7.156)$$

We now establish bounds for the differences of the form  $a_k \leq Z_k \leq b_k$ . Firstly we have

$$Z_k = \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_k] - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] \quad (5.7.157)$$

$$\leq \sup_x \{\mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x]\} - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] \quad (5.7.158)$$

and then

$$Z_k = \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_k] - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] \quad (5.7.159)$$

$$\geq \inf_x \{\mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x]\} - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] \quad (5.7.160)$$

Combining these, we get

$$b_k - a_k = \sup_x \{\mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x]\} - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}] \quad (5.7.161)$$

$$- \inf_x \{\mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x]\} + \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}]$$

$$= \sup_x \{\mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x]\} + \sup_x \{-\mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x]\} \quad (5.7.162)$$

$$= \sup_{x,x'} \{\mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x] - \mathbb{E}[f(\mathbf{X}_m) | \mathbf{X}_{k-1}, X_k = x']\} \quad (5.7.163)$$

$$\leq c_k \quad (5.7.164)$$

from the condition  $|f(X_1, \dots, X_i, \dots, X_m) - f(X_1, \dots, x'_i, \dots, X_m)| \leq c_i$  which always surely must hold. Finally applying the Azuma-Hoeffding inequality for the martingale sequence  $\mathbb{E}[Y_m | \mathbf{X}_k]$  up to  $k = m$ , we arrive at the inequalities claimed.  $\square$

### 5.7.6 Bernstein's Inequality [30]

In words, Bernstein's inequality says that a sum of independent random variables is sub-Gamma, under a moment boundedness condition of the **left-truncated distribution** similar to that of a **sub-Gaussian**. More precisely, let  $X_1, \dots, X_n$  be independent random variables. For the moment boundedness conditions, assume there exist constants  $c, v > 0$  such that

$$\sum_{i=1}^n \mathbb{E}[X_i^2] \leq v \quad (5.7.165)$$

and for all integers  $q \geq 3$ ,

$$\sum_{i=1}^n \mathbb{E}[(X_i)_+^q] \leq \frac{q!}{2} vc^{q-2} \quad (5.7.166)$$

where  $(X_i)_+ := \max\{X_i, 0\}$  denotes the left-truncation at zero. If we let the sum

$$S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \quad (5.7.167)$$

then for all  $\lambda \in (0, 1/c)$  we have for the **cumulant generating function**

$$\psi_S(\lambda) := \log \mathbb{E}[e^{\lambda S}] \quad (5.7.168)$$

$$\leq \frac{v\lambda^2}{2(1-c\lambda)} \quad (5.7.169)$$

and in particular for all  $t > 0$ ,

$$\Pr(S \geq \sqrt{2vt} + ct) \leq e^{-t} \quad (5.7.170)$$

*Proof.* Denote  $\phi(u) = e^u - u - 1$ . A graph will reveal that

$$\phi(u) \leq \frac{u^2}{2} \quad (5.7.171)$$

for all  $u \leq 0$ . To confirm this, note  $\phi(0) = 0^2/2$  and we can compare the second derivative  $\phi''(u) = e^u \leq 1$  for all  $u \leq 0$ , so  $u^2/2$  has a ‘stronger’ curvature than  $\phi(u)$  for  $u < 0$ . Then consider  $\phi(\lambda X_i)$  for  $\lambda > 0$ . If  $X_i \leq 0$ , then it follows that

$$\phi(\lambda X_i) \leq \frac{\lambda^2 X_i^2}{2} \quad (5.7.172)$$

If  $X_i > 0$  on the other hand, then using the series expansion of the exponential we have

$$\phi(u) = -1 - u + \sum_{q=0}^{\infty} \frac{u^q}{q!} \quad (5.7.173)$$

$$= -1 - u + 1 + u + \frac{u^2}{2} + \sum_{q=3}^{\infty} \frac{u^q}{q!} \quad (5.7.174)$$

$$= \frac{u^2}{2} + \sum_{q=3}^{\infty} \frac{u^q}{q!} \quad (5.7.175)$$

so that

$$\phi(\lambda X_i) \leq \frac{\lambda^2 X_i^2}{2} + \sum_{q=3}^{\infty} \frac{\lambda^q X_i^q}{q!} \quad (5.7.176)$$

Combining both cases, we can generally write

$$\phi(\lambda X_i) \leq \frac{\lambda^2 X_i^2}{2} + \sum_{q=3}^{\infty} \frac{\lambda^q (X_i)_+^q}{q!} \quad (5.7.177)$$

Taking the expectation of both sides gives

$$\mathbb{E}[\phi(\lambda X_i)] \leq \frac{\lambda^2 \mathbb{E}[X_i^2]}{2} + \sum_{q=3}^{\infty} \frac{\lambda^q \mathbb{E}[(X_i)_+^q]}{q!} \quad (5.7.178)$$

Taking the sum over  $1, \dots, n$  and applying the moment boundedness conditions:

$$\sum_{i=1}^n \mathbb{E}[\phi(\lambda X_i)] = \frac{\lambda^2}{2} \sum_{i=1}^n \mathbb{E}[X_i^2] + \sum_{i=1}^n \sum_{q=3}^{\infty} \frac{\lambda^q \mathbb{E}[(X_i)_+^q]}{q!} \quad (5.7.179)$$

$$\leq \frac{\lambda^2 v}{2} + \sum_{q=3}^{\infty} \frac{\lambda^q}{q!} \sum_{i=1}^n \mathbb{E}[(X_i)_+^q] \quad (5.7.180)$$

$$\leq \frac{\lambda^2 v}{2} + \sum_{q=3}^{\infty} \frac{\lambda^q}{q!} \cdot \frac{q!}{2} v c^{q-2} \quad (5.7.181)$$

$$= \sum_{q=2}^{\infty} \frac{\lambda^q}{2} v c^{q-2} \quad (5.7.182)$$

$$= \frac{v \lambda^2}{2} \sum_{q=0}^{\infty} (\lambda c)^q \quad (5.7.183)$$

As we restrict  $\lambda \in (0, 1/c)$ , then  $\lambda c \in (0, 1)$  hence the infinite sum is a geometric series evaluating to  $(1 - \lambda c)^{-1}$ , giving

$$\sum_{i=1}^n \mathbb{E}[\phi(\lambda X_i)] \leq \frac{v \lambda^2}{2(1 - c\lambda)} \quad (5.7.184)$$

Now via independence, the cumulant generating function of  $S$  can be expressed as

$$\psi_S(\lambda) = \log \mathbb{E}[e^{\lambda S}] \quad (5.7.185)$$

$$= \log \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)\right] \quad (5.7.186)$$

$$= \log \mathbb{E}\left[\frac{\prod_{i=1}^n e^{\lambda X_i}}{\prod_{i=1}^n e^{\lambda \mathbb{E}[X_i]}}\right] \quad (5.7.187)$$

$$= \log\left(\frac{\prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}]}{\prod_{i=1}^n \mathbb{E}[e^{\lambda \mathbb{E}[X_i]}]}\right) \quad (5.7.188)$$

$$= \log\left(\frac{\prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}]}{\prod_{i=1}^n e^{\lambda \mathbb{E}[X_i]}} e^{\lambda \mathbb{E}[X_i]}\right) \quad (5.7.189)$$

$$= \sum_{i=1}^n \left(\log \mathbb{E}[e^{\lambda X_i}] - \log e^{\lambda \mathbb{E}[X_i]}\right) \quad (5.7.190)$$

$$= \sum_{i=1}^n \left(\log \mathbb{E}[e^{\lambda X_i}] - \lambda \mathbb{E}[X_i]\right) \quad (5.7.191)$$

Using the fact that  $\log u \leq u - 1$  for  $u > 0$  (which can for example be shown via concavity of  $\log u$ ), then

$$\log \mathbb{E}[e^{\lambda X_i}] \leq \mathbb{E}[e^{\lambda X_i}] - 1 \quad (5.7.192)$$

and we have for the cumulant generating function

$$\psi_S(\lambda) \leq \sum_{i=1}^n \left( \log \mathbb{E} [e^{\lambda X_i}] - \lambda \mathbb{E} [X_i] \right) \quad (5.7.193)$$

$$\leq \sum_{i=1}^n \left( \mathbb{E} [e^{\lambda X_i}] - 1 - \lambda \mathbb{E} [X_i] \right) \quad (5.7.194)$$

$$= \sum_{i=1}^n \left( \mathbb{E} [e^{\lambda X_i} - \lambda X_i - 1] \right) \quad (5.7.195)$$

$$= \sum_{i=1}^n \mathbb{E} [\phi(\lambda X_i)] \quad (5.7.196)$$

$$\leq \frac{v\lambda^2}{2(1-c\lambda)} \quad (5.7.197)$$

where in the last inequality we have used our previously derived bound on  $\sum_{i=1}^n \mathbb{E} [\phi(\lambda X_i)]$  from above. This is exactly the characterisation of a random variable that is sub-Gamma on the right. Therefore  $S$  is sub-Gamma on the right, and it further follows that

$$\Pr \left( S \geq \sqrt{2vt} + ct \right) \leq e^{-t} \quad (5.7.198)$$

for all  $t > 0$ . □

## 5.8 Large Deviations Theory

### 5.9 Random Matrices

Not to be confused with stochastic matrices, a random matrix is a matrix whose elements are random variables.

#### 5.9.1 Matrix Gaussian Distribution

The matrix Gaussian distribution generalises the multivariate Gaussian distribution to distributions of matrices. A  $n \times m$  random matrix  $\mathbf{X}$  is matrix Gaussian distributed with location parameter  $M \in \mathbb{R}^{n \times m}$  and scale parameters  $U \in \mathbb{R}^{n \times n}$ ,  $U \succ \mathbf{0}$ ,  $V \in \mathbb{R}^{m \times m}$ ,  $V \succ \mathbf{0}$  if it has density

$$p(X|M, U, V) = \frac{1}{(2\pi)^{nm/2} |V|^{n/2} |U|^{m/2}} \exp \left[ -\frac{1}{2} \text{trace} \left( V^{-1} (X - M)^\top U^{-1} (X - M) \right) \right] \quad (5.9.1)$$

and denoted  $\mathcal{MN}(M, U, V)$ . There is an equivalence between the matrix Gaussian distribution and the multivariate Gaussian distribution. We have  $\mathbf{X} \sim \mathcal{MN}(M, U, V)$  if and only if  $\text{vec}(\mathbf{X}) \sim \mathcal{N}(\text{vec}(M), V \otimes U)$ .

*Proof.* Consider the exponent in the matrix Gaussian distribution. We show equivalence to the exponent in the multivariate Gaussian distribution using several properties of vectorisation and the Kronecker product.

$$\text{trace} \left( V^{-1} (X - M)^\top U^{-1} (X - M) \right) = \text{trace} \left( (X - M)^\top U^{-1} (X - M) V^{-1} \right) \quad (5.9.2)$$

$$= \text{vec}(X - M)^\top \text{vec}(U^{-1}(X - M)V^{-1}) \quad (5.9.3)$$

using the verifiable property  $\text{trace}(A^\top B) = \text{vec}(A)^\top \text{vec}(B)$ . Then

$$\text{vec}(X - M)^\top \text{vec}(U^{-1}(X - M)V^{-1}) = \text{vec}(X - M)^\top ((V^{-1})^\top \otimes U^{-1}) \text{vec}(X - M) \quad (5.9.4)$$

because of the ‘vec trick’  $\text{vec}(ABC) = (C^\top \otimes A) \text{vec}(B)$ . Continuing:

$$\text{vec}(X - M)^\top ((V^{-1})^\top \otimes U^{-1}) \text{vec}(X - M) = (\text{vec}(X) - \text{vec}(M))^\top (V^{-1} \otimes U^{-1}) (\text{vec}(X) - \text{vec}(M)) \quad (5.9.5)$$

$$= (\text{vec}(X) - \text{vec}(M))^\top (V \otimes U)^{-1} (\text{vec}(X) - \text{vec}(M)) \quad (5.9.6)$$

with the identity  $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ . Now for equivalence the normalising factor, we use the property for determinants of Kronecker products:

$$|V \otimes U| = |V|^n |U|^m \quad (5.9.7)$$

noting that  $V$  is  $m \times m$  while  $U$  is  $n \times n$ . Therefore

$$p(X|M, U, V) = \frac{1}{(2\pi)^{nm/2} |V \otimes U|^{1/2}} \exp \left[ -\frac{1}{2} (\text{vec}(X) - \text{vec}(M))^\top (V \otimes U)^{-1} (\text{vec}(X) - \text{vec}(M)) \right] \quad (5.9.8)$$

which is equivalent to the density of  $\text{vec}(\mathbf{X}) \sim \mathcal{N}(\text{vec}(M), V \otimes U)$ .  $\square$

### 5.9.2 Gaussian Ensembles

### 5.9.3 Wishart Distribution

## Chapter 6

# Bayesian Probability & Statistics

### 6.1 Bayes' Theorem Extensions

#### 6.1.1 Bayes' Theorem for Multiple Events

For three events  $A$ ,  $B$  and  $C$ . There are six ways of writing the chain rule of probability:

$$\Pr(A, B, C) = \Pr(A, B|C) \Pr(C) \quad (6.1.1)$$

$$= \Pr(A, C|B) \Pr(B) \quad (6.1.2)$$

$$= \Pr(B, C|A) \Pr(A) \quad (6.1.3)$$

$$= \Pr(A|B, C) \Pr(B, C) \quad (6.1.4)$$

$$= \Pr(B|A, C) \Pr(A, C) \quad (6.1.5)$$

$$= \Pr(C|A, B) \Pr(A, B) \quad (6.1.6)$$

This means that there are  $\binom{6}{2} = 15$  different ways to write Bayes' theorem for three events by choosing any two and rearranging, for example:

$$\Pr(A, B|C) = \frac{\Pr(B, C|A) \Pr(A)}{\Pr(C)} \quad (6.1.7)$$

or

$$\Pr(A|B, C) = \frac{\Pr(B, C|A) \Pr(A)}{\Pr(B, C)} \quad (6.1.8)$$

To extend this to when there are  $n$  events  $A_1, \dots, A_n$ , we can consider the same approach. To write down a chain rule expression, we can start by selecting any  $k$  events, where  $1 < k < n$ . For example, with the first  $k$  events we would have

$$\Pr(A_1, \dots, A_n) = \Pr(A_{k+1}, \dots, A_n | A_1, \dots, A_k) \Pr(A_1, \dots, A_k) \quad (6.1.9)$$

The total number of ways to do this for all  $1 < k < n$  is

$$\binom{n}{1} + \dots + \binom{n}{n-1} = 2^n - 2 \quad (6.1.10)$$

since  $\binom{n}{0} + \dots + \binom{n}{n} = 2^n$  and  $\binom{n}{0} = \binom{n}{n} = 1$ . Therefore, there are a total of  $\binom{2^n - 2}{2}$  ways to write a Bayes' theorem expression involving  $n$  events.

#### 6.1.2 Bayes' Theorem for Distributions

Let  $X$  and  $Y$  be discrete random variables with some joint probability mass  $p_{XY}(x, y)$ . Using Bayes' theorem, we can write for events  $\{X = x\}$  and  $\{Y = y\}$ :

$$p_{X|Y}(x|y) := \Pr(X = x | Y = y) \quad (6.1.11)$$

$$= \frac{\Pr(Y = y|X = x)\Pr(X = x)}{\Pr(Y = y)} \quad (6.1.12)$$

$$= \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \quad (6.1.13)$$

which is valid when  $p_Y(y) > 0$ . In this context, the conditional mass function  $p_{X|Y}(x|y)$  is called the *posterior distribution*,  $p_{Y|X}(y|x)$  is called the *likelihood*, and  $p_X(x)$  is called the *prior distribution*. The denominator  $p_Y(y)$  may be called the *marginal likelihood* because to obtain it, we usually have to marginalise over the distribution of  $X \in \mathcal{X}$  with a sum:

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_{XY}(x, y) \quad (6.1.14)$$

$$= \sum_{x \in \mathcal{X}} p_{Y|X}(y|x)p_X(x) \quad (6.1.15)$$

This gives the version of Bayes' theorem:

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x \in \mathcal{X}} p_{Y|X}(y|x)p_X(x)} \quad (6.1.16)$$

Now suppose  $X$  and  $Y$  be discrete random variables with some joint probability density  $f_{XY}(x, y)$ . This joint density can be factorised as

$$f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y) \quad (6.1.17)$$

or

$$f_{XY}(x, y) = f_{Y|X}(y|x)f_X(x) \quad (6.1.18)$$

so by equating and rearranging, we have a form of Bayes' theorem involving densities

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} \quad (6.1.19)$$

which is valid when  $f_Y(y) > 0$ . Like in the discrete case, the marginal likelihood  $f_Y(y)$  in the denominator can be obtained by integrating the numerator over  $x \in \mathcal{X}$ :

$$f_Y(y) = \int_{\mathcal{X}} f_{XY}(x, y) dx \quad (6.1.20)$$

$$= \int_{\mathcal{X}} f_{Y|X}(y|x)f_X(x) dx \quad (6.1.21)$$

This gives the version of Bayes' theorem

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{\mathcal{X}} f_{Y|X}(y|x)f_X(x) dx} \quad (6.1.22)$$

where we can see that in both the discrete and continuous case, the denominator plays the role of a normalising constant, so that the posterior sums or integrates to one.

### Bayes' Theorem with Discrete Prior and Continuous Likelihood

Suppose that  $X$  is a discrete random variable, while  $Y$  is a continuous random variable. We derive the form of the posterior distribution of  $X$  given  $Y$ , written  $p_{X|Y}(x|y)$ . Although  $Y$  is continuous, and thus  $\Pr(Y = y) = 0$ , we can define  $p_{X|Y}(x|y)$  by taking the limit

$$p_{X|Y}(x|y) = \lim_{\varepsilon \rightarrow 0} \Pr(X = x|y \leq Y < y + \varepsilon) \quad (6.1.23)$$

To derive this, first factorise the joint probability

$$\Pr(X = x, y \leq Y < y + \varepsilon) = \Pr(y \leq Y < y + \varepsilon | X = x) \Pr(X = x) \quad (6.1.24)$$

$$= (F_{Y|X}(y + \varepsilon | x) - F_{Y|X}(y | x)) \Pr(X = x) \quad (6.1.25)$$

where  $F_{Y|X}(y | x)$  is the conditional CDF of  $Y$  given  $X = x$ . Then taking the limit,

$$\lim_{\varepsilon \rightarrow 0} \Pr(X = x, y \leq Y < y + \varepsilon) = \lim_{\varepsilon \rightarrow 0} (F_{Y|X}(y + \varepsilon | x) - F_{Y|X}(y | x)) \Pr(X = x) \quad (6.1.26)$$

$$= \lim_{\varepsilon \rightarrow 0} \left( \varepsilon \frac{F_{Y|X}(y + \varepsilon | x) - F_{Y|X}(y | x)}{\varepsilon} \right) p_X(x) \quad (6.1.27)$$

$$= \lim_{\varepsilon \rightarrow 0} \varepsilon \frac{dF_{Y|X}(y | x)}{dy} p_X(x) \quad (6.1.28)$$

$$= \lim_{\varepsilon \rightarrow 0} \varepsilon f_{Y|X}(y | x) p_X(x) \quad (6.1.29)$$

Via the alternative factorisation of the joint probability,

$$\lim_{\varepsilon \rightarrow 0} \Pr(X = x, y \leq Y < y + \varepsilon) = \lim_{\varepsilon \rightarrow 0} \Pr(X = x | y \leq Y < y + \varepsilon) \Pr(y \leq Y < y + \varepsilon) \quad (6.1.30)$$

$$= p_{X|Y}(x | y) \lim_{\varepsilon \rightarrow 0} \Pr(y \leq Y < y + \varepsilon) \quad (6.1.31)$$

$$= p_{X|Y}(x | y) \lim_{\varepsilon \rightarrow 0} \left( \varepsilon \frac{F_Y(y + \varepsilon) - F_Y(y)}{\varepsilon} \right) \quad (6.1.32)$$

$$= \lim_{\varepsilon \rightarrow 0} \varepsilon p_{X|Y}(x | y) f_Y(y) \quad (6.1.33)$$

where  $p_{X|Y}(x | y)$  comes by our definition of the posterior distribution. Then from equating these expressions and rearranging, we have

$$p_{X|Y}(x | y) = \frac{f_{Y|X}(y | x) p_X(x)}{f_Y(y)} \quad (6.1.34)$$

This essentially takes the usual form of Bayes' theorem for distributions, except that continuous distributions have density functions, while discrete distributions have mass functions. The marginal likelihood  $f_Y(y)$  can also be obtained by taking a sum:

$$f_Y(y) = \sum_{x \in \mathcal{X}} f_{Y|X}(y | x) p_X(x) \quad (6.1.35)$$

### Bayes' Theorem with Continuous Prior and Discrete Likelihood

Suppose that  $X$  is a continuous random variable, while  $Y$  is a discrete random variable. The derivation for the form of the posterior distribution of  $X$  given  $Y$  essentially follows the case where  $X$  is discrete and  $Y$  is continuous, except  $X$  and  $Y$  are swapped around. We can in the limit obtain

$$f_{X|Y}(x | y) p_Y(y) = p_{Y|X}(y | x) f_X(x) \quad (6.1.36)$$

and rearrange to get the conditional density

$$f_{X|Y}(x | y) = \frac{p_{Y|X}(y | x) f_X(x)}{p_Y(y)} \quad (6.1.37)$$

where the marginal likelihood  $p_Y(y)$  is obtained via an integral:

$$p_Y(y) = \int_{\mathcal{X}} p_{Y|X}(y | x) f_X(x) dx \quad (6.1.38)$$

### 6.1.3 Bayes' Theorem for Mixed Distributions

Suppose we want to obtain a posterior distribution for  $X$  given  $Y$ , where  $X$  and  $Y$  are both mixed random variables. That is, they both have discrete and continuous components in their distributions. One way to express the distributions is by using probability densities, where Dirac delta impulses are used whenever there are point masses. The prior distribution of  $X$ , write

$$f_X(x) = \left(1 - \sum_j q_j\right) h(x) + \sum_j q_j \gamma_j(x) \quad (6.1.39)$$

where  $q' := \sum_j q_j$  and the  $q_j$  are mixing coefficients,  $h(x)$  is the density for the continuous component, and the  $\gamma_j(x)$  are impulses at the respective locations with probability masses. Similarly, write the likelihood as

$$f_{Y|X}(y|x) = \left(1 - \sum_i p_i\right) g(y|x) + \sum_i p_i \delta_i(y|x) \quad (6.1.40)$$

where  $p' := 1 - \sum_i p_i$  are mixing coefficients,  $g(y|x)$  is the conditional density for the continuous component, and the  $\delta_i(y|x)$  are impulses at the respective locations with conditional probability masses. With the form of Bayes' theorem, the posterior density is given by

$$f_{X|Y}(x|y) \propto f_{Y|X}(y|x) f_X(x) \quad (6.1.41)$$

$$= p' q' g(y|x) h(x) + p' \sum_j q_j g(y|x) \gamma_j(x) \quad (6.1.42)$$

$$+ q' \sum_i p_i \delta_i(y|x) h(x) + \sum_i p_i \sum_j q_j \delta_i(y|x) \gamma_j(x)$$

We have written a proportionality by ignoring the denominator in Bayes' theorem, since the latter only acts as a normalisation constant. However, we do assume that the denominator is positive at the  $y$  of interest. Also assume without loss of generality that  $p' > 0$ ,  $p_i > 0$ ,  $q' > 0$ ,  $q_j > 0$ , otherwise it just becomes a special case of Bayes' theorem with conventional distributions. Thus, we can see that the posterior is itself a mixture consisting of four components, each of which can be interpreted in a more conventional sense.

- If  $h(x) > 0$  and  $g(y|x) > 0$ , then this component is like the continuous-continuous Bayes' theorem, and the posterior will have positive density at  $x$  given  $y$ .
- If  $\sum_j \gamma_j(x) > 0$  and  $g(y|x) > 0$ , then this component is like the discrete prior continuous likelihood Bayes' theorem, so the posterior will have positive mass at  $x$  given  $y$ .
- If  $h(x) > 0$  and  $\sum_i \delta_i(y|x) > 0$ , then this component is like the continuous prior discrete likelihood Bayes' theorem. so the posterior will have positive density at  $x$  given  $y$ .
- If  $\sum_j \gamma_j(x) > 0$  and  $\sum_i \delta_i(y|x) > 0$ , then this is like the discrete-discrete Bayes' theorem, and the posterior will have positive mass at  $x$  given  $y$ .

### 6.1.4 Bayes' Theorem for Multivariate Distributions

Consider a multivariate distribution, denoted  $p(x_1, \dots, x_n)$ . If the variables are continuous, then this is the probability density function, or if the variables are discrete then this is the probability mass function. However, it does not overly matter whether the variables are continuous or discrete (or even mixed), as the form of Bayes' theorem still holds. As in Bayes' theorem for multiple events, this joint distribution can be factorised in several ways, e.g.

$$p(x_1, \dots, x_n) = p(x_{i_{k+1}}, \dots, x_{i_n} | x_{i_1}, \dots, x_{i_k}) p(x_{i_1}, \dots, x_{i_k}) \quad (6.1.43)$$

where  $i_1, \dots, i_n$  is an arbitrary permutation of  $\{1, \dots, n\}$ , and  $1 \leq k < n$ . Note that for simplicity of notation, we can denote all distributions with the same symbol  $p(\cdot)$ , but with context as to which distributions given in the parentheses. Another way to factorise the joint distribution is

$$p(x_1, \dots, x_n) = p(x_{j_{m+1}}, \dots, x_{j_n} | x_{j_1}, \dots, x_{j_m}) p(x_{j_1}, \dots, x_{j_m}) \quad (6.1.44)$$

where  $j_1, \dots, j_n$  is another permutation of  $\{1, \dots, n\}$ , and  $1 \leq m < n$  also. Then for any two possible factorisations, we may equate and rearrange to arrive at the form of Bayes' theorem

$$p(x_{i_{k+1}}, \dots, x_{i_n} | x_{i_1}, \dots, x_{i_k}) = \frac{p(x_{j_{m+1}}, \dots, x_{j_n} | x_{j_1}, \dots, x_{j_m}) p(x_{j_1}, \dots, x_{j_m})}{p(x_{i_1}, \dots, x_{i_k})} \quad (6.1.45)$$

## 6.2 Bayesian Priors

### 6.2.1 Principle of Indifference

The principle of indifference says that if an experiment yields  $n$  mutually exclusive outcomes, and nothing else is known, then the prior probability of each outcome should be assigned  $1/n$ .

### 6.2.2 Cromwell's Rule

Cromwell's rule asserts that prior probabilities of zero should be avoided, except for statements which are logically true or false.

### 6.2.3 Improper Priors

An improper prior is a prior distribution whose sum or integral does not necessarily add to 1. From Bayes' theorem, it can be seen that scaling all prior probabilities or densities by some constant will still give the same result for the posterior up to that proportionality constant. So it is still possible to use improper priors in Bayesian inference. An example is a continuous uniform distribution over an infinite interval, known as a 'flat' prior. This is done by taking the prior  $p(x) = c$  for some positive constant  $c$  over the interval (even though a continuous uniform distribution over an infinite interval is not well-defined). This expresses that the posterior should be proportional to the likelihood.

### 6.2.4 Uninformative Priors

An uninformative prior aims to give no subjective 'information'. For a discrete random variable, the discrete uniform distribution is the canonical choice of uninformative prior. Sometimes an improper prior may be used as an uninformative prior, for example the uniform distribution over an infinite interval.

### 6.2.5 Jeffrey's Priors

One way to define whether a prior is 'truly' uninformative is to see whether the distribution is invariant to monotonic (i.e. invertible) transformations. That is, if we have a uniform prior over random variable  $X$  and then take  $h(X)$  where  $h(\cdot)$  is some invertible monotonic transformation, then probabilities are preserved. In this sense, a uniform distribution over discrete support is uninformative, because each  $x$  and  $h(x)$  share identical probability masses, which means 'no information' about  $X$  implies 'no information' about  $h(X)$ . However, this will not be the case for a continuous uniform distribution. To illustrate, consider the uniform random variable  $X$  over  $[0, 1]$  with density  $f_X(x) = 1$  for  $x \in [0, 1]$ . Then suppose we take the coordinate transform

$Z = e^X$  with inverse transformation  $X = \log Z$ . Using the rules for monotonic transformations of random variables, we can derive

$$f_Z(z) = f_X(\log z) \cdot \left| \frac{d}{dz} \log z \right| \quad (6.2.1)$$

$$= \frac{1}{z} \quad (6.2.2)$$

for  $z \in [1, e]$ . However, if we evaluate the density at  $x = 1$  and the corresponding transform  $z = e$ , we get  $f_X(1) = 1$  and  $f_Z(e) = e^{-1}$ . Hence, the prior uncertainty assigned to each value changes with respect to the coordinate transform.

The Jeffrey's prior seeks to find a prior for continuous variables that is invariant to these coordinate transforms (so that the uncertainty is the same no matter which parametrisation is used to represent the variable). We consider the more general multivariate case, where the generalisation of a monotonic transform is an invertible transform. Let  $p(\boldsymbol{\theta})$  denote a multivariate continuous prior distribution on  $\boldsymbol{\theta} \in \Theta$ . To be a suitable uninformative prior, then for some parametrisation given by the invertible transformation  $\boldsymbol{\varphi} = h(\boldsymbol{\theta})$ , then  $p(\boldsymbol{\theta})$  should satisfy

$$p(\boldsymbol{\theta}) = p(h(\boldsymbol{\theta})) \quad (6.2.3)$$

$$= p(\boldsymbol{\varphi}) \quad (6.2.4)$$

for all  $\boldsymbol{\theta} \in \Theta$ . Note that we can weaken this condition so that they are equal up to a positive multiplicative constant (which then means the prior will be an improper prior). Given some data, let  $\mathcal{L}(\boldsymbol{\theta})$  denote the likelihood of generating that data (with dependence on the data suppressed in the notation). We claim that choosing the prior

$$p(\boldsymbol{\theta}) \propto \sqrt{\det(\mathcal{I}(\boldsymbol{\theta}))} \quad (6.2.5)$$

satisfies this condition, where  $\mathcal{I}(\boldsymbol{\theta})$  is the Fisher information matrix. To verify this claim, it suffices to show that  $p(\boldsymbol{\varphi}) \propto \sqrt{\det(\mathcal{I}(\boldsymbol{\varphi}))}$  also holds. Beginning from the form of the density  $p(\boldsymbol{\varphi})$  in terms of  $\boldsymbol{\varphi}$  as an invertible transformation of  $\boldsymbol{\theta}$ , we have

$$p(\boldsymbol{\varphi}) = p(h^{-1}(\boldsymbol{\varphi})) \left| \det \left( \frac{\partial h^{-1}(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right) \right| \quad (6.2.6)$$

$$= p(\boldsymbol{\theta}) \left| \det \left( \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}} \right) \right| \quad (6.2.7)$$

where we have replaced  $h^{-1}(\boldsymbol{\varphi})$  with  $\boldsymbol{\theta}$ . Applying our claim  $p(\boldsymbol{\theta}) \propto \sqrt{\det(\mathcal{I}(\boldsymbol{\theta}))}$ , then

$$p(\boldsymbol{\varphi}) \propto \sqrt{\det(\mathcal{I}(\boldsymbol{\theta}))} \left| \det \left( \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}} \right) \right| \quad (6.2.8)$$

$$= \sqrt{\det(\mathcal{I}(\boldsymbol{\theta}))} \cdot \sqrt{\det \left( \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}} \right)^2} \quad (6.2.9)$$

$$= \sqrt{\det \left( \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}} \right) \det(\mathcal{I}(\boldsymbol{\theta})) \det \left( \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\varphi}} \right)} \quad (6.2.10)$$

since the determinant is unchanged after taking the transpose. Then apply the definition of the information matrix and the property that the determinant commutes with the product to give

$$p(\boldsymbol{\varphi}) \propto \sqrt{\det \left( \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}} \right) \det \left( \mathbb{E} \left[ \left( \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \left( \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \right] \right) \det \left( \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\varphi}} \right)} \quad (6.2.11)$$

$$= \sqrt{\det \left( \mathbb{E} \left[ \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}}^\top \left( \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \left( \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}} \right] \right)} \quad (6.2.12)$$

$$= \sqrt{\det \left( \mathbb{E} \left[ \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}}^\top \nabla_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}) [\nabla_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta})]^\top \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}} \right] \right)} \quad (6.2.13)$$

where  $\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = [\nabla_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta})]^\top$  is a row vector. We remind ourselves of the chain rule that  $\nabla_{\boldsymbol{\varphi}} \log \mathcal{L}(\boldsymbol{\varphi}) = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\varphi}}^\top \nabla_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta})$ , so

$$p(\boldsymbol{\varphi}) \propto \sqrt{\det \left( \mathbb{E} \left[ \nabla_{\boldsymbol{\varphi}} \log \mathcal{L}(\boldsymbol{\varphi}) [\nabla_{\boldsymbol{\varphi}} \log \mathcal{L}(\boldsymbol{\varphi})]^\top \right] \right)} \quad (6.2.14)$$

$$= \sqrt{\det(\mathcal{I}(\boldsymbol{\varphi}))} \quad (6.2.15)$$

as required. The prior  $p(\boldsymbol{\theta}) \propto \sqrt{\det(\mathcal{I}(\boldsymbol{\theta}))}$  is known as a Jeffrey's prior. In the special case that  $\boldsymbol{\theta} = \theta$  is univariate, then the Jeffrey's prior is

$$p(\theta) \propto \sqrt{\mathcal{I}(\theta)} \quad (6.2.16)$$

### 6.2.6 Conjugate Priors

If a particular family of prior distributions and another family of likelihoods is such that the posterior belongs to the same family as the prior, then we say that the prior is a *conjugate family* for the likelihood. Moreover, we say that the prior and likelihood families form a *conjugate pair*.

#### Beta Conjugate Prior for Binomial Likelihood

#### Gaussian Conjugate Prior for Gaussian Likelihood

#### Dirichlet Conjugate Prior for Multinomial Likelihood

If the prior distribution on some parameters is Dirichlet distributed, and the likelihood is categorical distributed, then the posterior will also be Dirichlet. This fits with the characterisation of the Dirichlet distribution being a suitable model for categorical distributions. Suppose  $\boldsymbol{\theta}$  are the parameters of a categorical distribution with prior

$$\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha}) \quad (6.2.17)$$

For a single observation  $X$  from a categorical distribution, the likelihood of observing  $X = j$  is

$$\Pr(X = j | \boldsymbol{\theta}) = \theta_j \quad (6.2.18)$$

Hence the posterior for  $\boldsymbol{\theta}$  takes the form

$$p(\boldsymbol{\theta} | X = j) \propto \Pr(X = j | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \quad (6.2.19)$$

$$\propto \theta_j \times \frac{\prod_{i=1}^K \theta_i^{\alpha_i - 1}}{\mathbf{B}(\boldsymbol{\alpha})} \quad (6.2.20)$$

Hence without worrying about the normalising constant, we can see that the posterior is Dirichlet distributed, with

$$\boldsymbol{\theta} | X = j \sim \text{Dir}(\alpha_1, \dots, \alpha_j + 1, \dots, \alpha_K) \quad (6.2.21)$$

So the concentration parameter for the  $j^{\text{th}}$  category gets incremented by one. Extending this to the case where we have a sample of  $n$  i.i.d. categorical random variables, our likelihood

becomes a multinomial distribution. Letting  $Y_1, \dots, Y_K$  denote the counts for each category, the posterior now takes the form

$$p(\boldsymbol{\theta}|Y_1 = y_1, \dots, Y_K = y_K) \propto \theta_1^{y_1} \dots \theta_K^{y_K} \frac{\prod_{i=1}^K \theta_i^{\alpha_i-1}}{B(\boldsymbol{\alpha})} \quad (6.2.22)$$

so the posterior is also Dirichlet distributed with

$$\boldsymbol{\theta}|Y_1 = y_1, \dots, Y_K = y_K \sim \text{Dir}(\alpha_1 + y_1, \dots, \alpha_K + y_K) \quad (6.2.23)$$

Thus the Dirichlet distribution is the conjugate prior for the multinomial likelihood (with a categorical likelihood being a special case).

### Conjugate Priors for Exponential Family Likelihoods [65]

Consider a likelihood from an exponential family. That is, we can express the likelihood of parameters  $\theta$  given data  $y$  as

$$p(y|\theta) \propto g(\theta)^n \exp\left(\eta(\theta)^\top T(y)\right) \quad (6.2.24)$$

where the proportionality is with respect to  $\theta$ . We allow both  $y$  and  $\theta$  to generally be multi-dimensional. Hence we write  $g(\theta)^n$  as  $n$  will typically represent the sample size for an i.i.d. sample, and  $g(\theta)$  is the reciprocal of the partition function in the likelihood of a single observation. If the likelihood is any member of the exponential family, then we can show that it has a conjugate prior, and the conjugate prior will also be a member of the exponential family. Specifically, choose the prior

$$p(\theta) \propto g(\theta)^m \exp\left(\eta(\theta)^\top \mathbf{s}\right) \quad (6.2.25)$$

for some  $m$  (which could represent a ‘previous’ sample size) and  $\mathbf{s}$ . Although the form appears to be similar to the likelihood, note that the likelihood  $p(y|\theta)$  is a distribution in  $y$ , while  $p(\theta)$  is a distribution in  $\theta$ . Hence the  $g(\theta)^m$  now plays the role of the carrier function, and  $\eta(\theta)^\top$  plays the role of the sufficient statistics for the hyperparameters of  $\theta$ , appearing in  $\mathbf{s}$ . Thus this prior does not necessarily need to be the same family as the likelihood. However, we can still see that  $p(\theta)$  fits the definition of an exponential family. In computing the posterior, we find

$$p(\theta|y) \propto p(y|\theta) p(\theta) \quad (6.2.26)$$

$$\propto g(\theta)^{m+n} \exp\left(\eta(\theta)^\top (\mathbf{s} + T(y))\right) \quad (6.2.27)$$

which has the same form as the prior, hence belongs in the same family as the prior. Therefore this family is the conjugate family for the likelihood.

#### 6.2.7 Empirical Bayes

### 6.3 Bayesian Updating

#### 6.3.1 Rule of Succession

The rule of succession is the application of Bayes’ theorem to the probability of success/fail after repeated trials. Suppose  $X_1, \dots, X_{n+1}$  are independent binary random variables. Then suppose that after  $n$  trials,  $s$  successes were observed. Then under a uniform prior on the success probability  $\theta$ , the probability of the next trial being a success is given by

$$\Pr(X_{n+1} = 1|X_1, \dots, X_{n+1}) = \frac{s+1}{n+2} \quad (6.3.1)$$

To show this, first write the likelihood of  $\theta$  given  $s$  successes as

$$\mathcal{L}(\theta|s) = \Pr(X_1 + \cdots + X_n = s|\theta) \quad (6.3.2)$$

$$= \theta^s (1 - \theta)^{n-s} \quad (6.3.3)$$

The prior density  $p(\theta)$  has function

$$p(\theta) = \begin{cases} 1, & 0 \leq \theta \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (6.3.4)$$

The posterior density  $p(\theta|X_1 + \cdots + X_n = s)$  takes the form

$$p(\theta|X_1 + \cdots + X_n = s) = \frac{\theta^s (1 - \theta)^{n-s}}{\int_0^1 \theta^s (1 - \theta)^{n-s} d\theta} \quad (6.3.5)$$

Note that the denominator is a Beta function, and in fact the posterior density is a Beta distribution. Since  $n$  and  $s$  are integers, then posterior distribution can be written in terms of factorials as

$$p(\theta|X_1 + \cdots + X_n = s) = \frac{(n+1)!}{s!(n-s)!} \theta^s (1 - \theta)^{n-s} \quad (6.3.6)$$

where in terms of the traditional parameters of the Beta distribution we have  $\alpha = s+1$  and  $\beta = n-s+1$ . Hence the mean of the Beta posterior density is

$$\mathbb{E}[\theta|X_1 + \cdots + X_n = s] = \frac{\alpha}{\alpha + \beta} = \frac{s+1}{n+2} \quad (6.3.7)$$

As  $X_{n+1}$  is a binary variable, the probability of success is simply given by this mean. Note that when  $s = 0$  and  $n = 0$  (i.e. no trials have been conducted), the expectation becomes  $\frac{1}{2}$ , in agreement with having used a uniform prior on the probability of success.

An alternative, more intuitive guess for the probability of success is the observed sample proportion of success  $\frac{s}{n}$ . We can show that this results in using the improper prior

$$p(\theta) \propto \frac{1}{\theta(1-\theta)} \quad (6.3.8)$$

over  $0 \leq \theta \leq 1$ . The posterior density is then proportional to

$$p(\theta|X_1 + \cdots + X_n = s) \propto \theta^{s-1} (1 - \theta)^{n-s-1} \quad (6.3.9)$$

where the normalising constant is the Beta function  $B(s, n)$ , hence (provided  $s \neq 0$  and  $s \neq n$ ) the posterior density is a Beta distribution with parameters  $\alpha = s$  and  $\beta = n-s$ , which has a mean of  $\frac{s}{n}$ .

### 6.3.2 Odds Ratio Updating

The odds ratio of an event  $A$  is defined as  $\frac{\Pr(A)}{\Pr(\bar{A})}$ . The odds ratio can be used as a way to

simplify Bayesian updating. From some evidence  $B$ , the likelihood ratio is defined as  $\frac{\Pr(B|A)}{\Pr(B|\bar{A})}$ .

Then from Bayes' theorem the posterior probabilities are:

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)} \quad (6.3.10)$$

$$\Pr(\overline{A}|B) = \frac{\Pr(B|\overline{A})\Pr(\overline{A})}{\Pr(B)} \quad (6.3.11)$$

Dividing these two probabilities gives

$$\frac{\Pr(A|B)}{\Pr(\overline{A}|B)} = \frac{\Pr(B|A)}{\Pr(B|\overline{A})} \cdot \frac{\Pr(A)}{\Pr(\overline{A})} \quad (6.3.12)$$

Hence this shows that the posterior odds ratio can be expressed as a multiplication between the likelihood ratio and the prior odds ratio.

### 6.3.3 Log Odds Updating

The log odds is the logarithm of the odds ratio, also known as the logit.

$$\text{logit}(\Pr(A)) = \log\left(\frac{\Pr(A)}{\Pr(\overline{A})}\right) \quad (6.3.13)$$

$$= \log\left(\frac{\Pr(A)}{1 - \Pr(A)}\right) \quad (6.3.14)$$

$$= \log \Pr(A) - \log(1 - \Pr(A)) \quad (6.3.15)$$

The posterior log odds can be obtain by addition of the log likelihood ratio and the prior log odds.

$$\log\left(\frac{\Pr(A|B)}{\Pr(\overline{A}|B)}\right) = \log\left(\frac{\Pr(B|A)}{\Pr(B|\overline{A})} \cdot \frac{\Pr(A)}{\Pr(\overline{A})}\right) \quad (6.3.16)$$

$$\text{logit}(\Pr(A|B)) = \log\left(\frac{\Pr(B|A)}{\Pr(B|\overline{A})}\right) + \text{logit}(\Pr(A)) \quad (6.3.17)$$

### 6.3.4 Bayes Filters [35, 198]

Consider a discrete time stochastic process for a state variable  $x_k$ , where the evolution of the state can be expressed with the distribution  $p(x_{k+1}|x_k)$ . The state  $x_k$  is not directly measured but instead another variable  $z_k$  is observed with the distribution  $p(z_k|x_k)$ . The task in Bayes filtering is to recover the distribution for  $x_k$  given only measurements up to and including  $z_k$ . We denote the measurement set up to and including time  $k$  by

$$\mathbf{z}_{1:k} = (z_1, \dots, z_k) \quad (6.3.18)$$

Thus we denote the posterior distribution as  $p(x_k|\mathbf{z}_{1:k})$ . From Bayes' theorem we can write

$$p(x_k|\mathbf{z}_{1:k}) = \frac{p(z_k|x_k, \mathbf{z}_{1:(k-1)}) p(x_k|\mathbf{z}_{1:(k-1)})}{p(z_k|\mathbf{z}_{1:(k-1)})} \quad (6.3.19)$$

$$\propto p(z_k|x_k, \mathbf{z}_{1:(k-1)}) p(x_k|\mathbf{z}_{1:(k-1)}) \quad (6.3.20)$$

It is evident that we need only be concerned with finding  $p(z_k|x_k, \mathbf{z}_{1:(k-1)}) p(x_k|\mathbf{z}_{1:(k-1)})$ , because it is the same as the posterior  $p(x_k|\mathbf{z}_{1:k})$  up to a normalising factor (constant with respect to  $x_k$ ). We assume that  $z_k$  and  $\mathbf{z}_{1:(k-1)}$  are conditionally independent given  $x_k$ , meaning that

$$p(z_k|x_k, \mathbf{z}_{1:(k-1)}) = p(z_k|x_k) \quad (6.3.21)$$

which is provided in the model. Note that it is sometimes typical to write

$$p(x_k|\mathbf{z}_{1:k}) = \kappa_k p(z_k|x_k) p(x_k|\mathbf{z}_{1:(k-1)}) \quad (6.3.22)$$

where  $\kappa_k$  is referred to as the gain. The distribution  $p(x_k | \mathbf{z}_{1:(k-1)})$  is the only remaining distribution required to be found. We can perform marginalisation:

$$p(x_k | \mathbf{z}_{1:(k-1)}) = \int p(x_k | x_{k-1}, \mathbf{z}_{1:(k-1)}) p(x_{k-1} | \mathbf{z}_{1:(k-1)}) dx_{k-1} \quad (6.3.23)$$

We further assume that  $x_k$  and  $\mathbf{z}_{1:(k-1)}$  are conditionally independent given  $x_{k-1}$ , so that

$$p(x_k | x_{k-1}, \mathbf{z}_{1:(k-1)}) = p(x_k | x_{k-1}) \quad (6.3.24)$$

Hence

$$p(x_k | \mathbf{z}_{1:(k-1)}) = \int p(x_k | x_{k-1}) p(x_{k-1} | \mathbf{z}_{1:(k-1)}) dx_{k-1} \quad (6.3.25)$$

and the filtering distribution is given by

$$p(x_k | \mathbf{z}_{1:k}) = \kappa_k p(z_k | x_k) \int p(x_k | x_{k-1}) p(x_{k-1} | \mathbf{z}_{1:(k-1)}) dx_{k-1} \quad (6.3.26)$$

We see that this integral contains the posterior from the previous time step,  $p(x_{k-1} | \mathbf{z}_{1:(k-1)})$ . Therefore at the beginning the algorithm requires an initial guess of the distribution of the initial state given by  $p(x_0)$ .

### Bayes Filter with Input

When a known input (or *control*)  $u_k$  is involved in the state evolution, this allows for the update distribution to be specified as  $p(x_{k+1} | x_k, u_k)$ . The task is now to obtain the posterior distribution  $p(x_k | \mathbf{z}_{1:k}, \mathbf{u}_{1:k})$ , where  $\mathbf{u}_{1:k}$  denotes the sequence of inputs up to and including time  $k$ :

$$\mathbf{u}_{1:k} = (u_1, \dots, u_k) \quad (6.3.27)$$

This filter is very similar to before, with a few modified assumptions. From Bayes' theorem we require

$$p(x_k | \mathbf{z}_{1:k}, \mathbf{u}_{1:k}) = \frac{p(z_k | x_k, \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) p(x_k | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k})}{p(z_k | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k})} \quad (6.3.28)$$

$$\propto p(z_k | x_k, \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) p(x_k | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) \quad (6.3.29)$$

$$= p(z_k | x_k) p(x_k | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) \quad (6.3.30)$$

$$= \kappa_k p(z_k | x_k) p(x_k | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) \quad (6.3.31)$$

where we have assumed that  $p(z_k | x_k, \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) = p(z_k | x_k)$  by conditional independence of  $z_k$  and  $(\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k})$  given  $x_k$ . Then to obtain the prior we marginalise:

$$p(x_k | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) = \int p(x_k | x_{k-1}, \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) p(x_{k-1} | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:(k-1)}) dx_{k-1} \quad (6.3.32)$$

$$= \int p(x_k | x_{k-1}, u_k) p(x_{k-1} | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:(k-1)}) dx_{k-1} \quad (6.3.33)$$

where  $p(x_{k-1} | \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:(k-1)})$  is the posterior at time  $k-1$  and we have additionally assumed that  $x_k$  and  $(\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:(k-1)})$  are conditionally independent given  $x_{k-1}$  and  $u_k$ .

## Gaussian Filters

Gaussian filters are Bayes Filters for when the distributions  $p(x_{k+1}|x_k, u_k)$  and  $p(z_k|x_k)$  are Gaussian (for concreteness, we can consider a controlled linear Gaussian model). In this case, if the posterior distribution is Gaussian and the conditional expectation  $\mathbb{E}[x_{k+1}|x_k, u_k]$  is linear in  $x_k$  and  $u_k$ , the posterior distributions will also be Gaussian. This result follows from the fact that products of Gaussian densities will also be Gaussian (after normalisation), and that marginalisation involving Gaussians and linear conditional expectations will also produce a Gaussian. Here, we consider the computation of

$$p(x_k|\mathbf{z}_{1:k}, \mathbf{u}_{1:k}) = \frac{p(z_k|x_k, \mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) p(x_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k})}{p(z_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k})} \quad (6.3.34)$$

Converting the numerator into a joint density:

$$p(x_k|\mathbf{z}_{1:k}, \mathbf{u}_{1:k}) = \frac{p(z_k, x_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k})}{p(z_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k})} \quad (6.3.35)$$

we assume that the densities on the right-hand side are available to us. For ease notation we write the conditional expectations and covariances of each density using the notation:

$$\hat{x}_k = \mathbb{E}[x_k|\mathbf{z}_{1:k}, \mathbf{u}_{1:k}] \quad (6.3.36)$$

$$\hat{x}_k^- = \mathbb{E}[x_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}] \quad (6.3.37)$$

$$\hat{z}_k^- = \mathbb{E}[z_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}] \quad (6.3.38)$$

$$\mathbf{P}_{xx,k} = \text{Cov}(x_k|\mathbf{z}_{1:k}, \mathbf{u}_{1:k}) \quad (6.3.39)$$

$$\mathbf{P}_{xx,k}^- = \text{Cov}(x_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) \quad (6.3.40)$$

$$\mathbf{P}_{zz,k}^- = \text{Cov}(z_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) \quad (6.3.41)$$

and the cross-covariances

$$\mathbf{P}_{xz,k}^- = \text{Cov}(x_k, z_k|\mathbf{z}_{1:(k-1)}, \mathbf{u}_{1:k}) \quad (6.3.42)$$

$$\mathbf{P}_{zx,k}^- = (\mathbf{P}_{xz,k}^-)^\top \quad (6.3.43)$$

Therefore the posterior can be written with notation

$$\mathcal{N}_{x_k}(\hat{x}_k, \mathbf{P}_{xx,k}) = \mathcal{N}_{x_k, z_k}\left(\begin{bmatrix}\hat{x}_k^- \\ \hat{z}_k^-\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{xx,k}^- & \mathbf{P}_{xz,k}^- \\ \mathbf{P}_{zx,k}^- & \mathbf{P}_{zz,k}^-\end{bmatrix}\right) \div \mathcal{N}_{z_k}(\hat{z}_k^-, \mathbf{P}_{zz,k}^-) \quad (6.3.44)$$

where  $\mathcal{N}_{x_k}(\cdot, \cdot)$  denotes a (multivariate) Gaussian density in  $x_k$ , etc. For further notational simplicity, we drop the time index  $k$ , and denote

$$\tilde{x} = x - \hat{x} \quad (6.3.45)$$

$$\tilde{x}^- = x - \hat{x}^- \quad (6.3.46)$$

$$\tilde{z}^- = z - \hat{z}^- \quad (6.3.47)$$

Explicitly writing out the exponentials in the Gaussian densities, we have by proportionality:

$$\exp(\tilde{x}^\top \mathbf{P}_{xx}^{-1} \tilde{x}) \propto \exp\left(\begin{bmatrix}\tilde{x}^- \\ \tilde{z}^-\end{bmatrix}^\top \begin{bmatrix}\mathbf{P}_{xx}^- & \mathbf{P}_{xz}^- \\ \mathbf{P}_{zx}^- & \mathbf{P}_{zz}^-\end{bmatrix}^{-1} \begin{bmatrix}\tilde{x}^- \\ \tilde{z}^-\end{bmatrix}\right) \div \exp((\tilde{z}^-)^\top \mathbf{P}_{xx}^{-1} \tilde{z}^-) \quad (6.3.48)$$

and equating the exponents gives

$$\tilde{x}^\top \mathbf{P}_{xx}^{-1} \tilde{x} = \begin{bmatrix}\tilde{x}^- \\ \tilde{z}^-\end{bmatrix}^\top \begin{bmatrix}\mathbf{P}_{xx}^- & \mathbf{P}_{xz}^- \\ \mathbf{P}_{zx}^- & \mathbf{P}_{zz}^-\end{bmatrix}^{-1} \begin{bmatrix}\tilde{x}^- \\ \tilde{z}^-\end{bmatrix} - (\tilde{z}^-)^\top \mathbf{P}_{xx}^{-1} \tilde{z}^- \quad (6.3.49)$$

Using the block matrix inversion formula:

$$\begin{bmatrix} \mathbf{P}_{xx}^- & \mathbf{P}_{xz}^- \\ \mathbf{P}_{zx}^- & \mathbf{P}_{zz}^- \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \\ -(\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1} & (\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} + (\mathbf{P}_{zz}^-)^{-1} \end{bmatrix} \quad (6.3.50)$$

where  $\mathbf{B}^{-1} = \mathbf{P}_{xx}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-$ . Then expanding out the quadratic form on the right-hand side:

$$\begin{bmatrix} \tilde{x}^- \\ \tilde{z}^- \end{bmatrix}^\top \begin{bmatrix} \mathbf{P}_{xx}^- & \mathbf{P}_{xz}^- \\ \mathbf{P}_{zx}^- & \mathbf{P}_{zz}^- \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x}^- \\ \tilde{z}^- \end{bmatrix} - (\tilde{z}^-)^\top \mathbf{P}_{xx}^- \tilde{z}^- \quad (6.3.51)$$

$$= \begin{bmatrix} \tilde{x}^- \\ \tilde{z}^- \end{bmatrix}^\top \begin{bmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \\ -(\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1} & (\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} + (\mathbf{P}_{zz}^-)^{-1} \end{bmatrix} \begin{bmatrix} \tilde{x}^- \\ \tilde{z}^- \end{bmatrix} - (\tilde{z}^-)^\top \mathbf{P}_{xx}^- \quad (6.3.52)$$

$$= (\tilde{x}^-)^\top \mathbf{B}^{-1} \tilde{x}^- - (\tilde{x}^-)^\top \mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^- - (\tilde{z}^-)^\top (\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1} \tilde{x}^- + (\tilde{z}^-)^\top [(\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} + (\mathbf{P}_{zz}^-)^{-1}] \tilde{z}^- - (\tilde{z}^-)^\top \mathbf{P}_{xx}^- \tilde{z}^- \quad (6.3.53)$$

$$= (\tilde{x}^-)^\top \mathbf{B}^{-1} \tilde{x}^- - (\tilde{x}^-)^\top \mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^- - (\tilde{z}^-)^\top (\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1} \tilde{x}^- + (\tilde{z}^-)^\top (\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^-\mathbf{B}^{-1}\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^- \quad (6.3.54)$$

$$= [\tilde{x}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^-]^\top \mathbf{B}^{-1} [\tilde{x}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^-] \quad (6.3.55)$$

where in the last step we have factorised the quadratic. Equating this to the left-hand side, we have

$$\tilde{x}^\top \mathbf{P}_{xx}^- \tilde{x} = [\tilde{x}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^-]^\top \mathbf{B}^{-1} [\tilde{x}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^-] \quad (6.3.56)$$

We can equate the vector and matrix factors in the equation. First equating the vector factors:

$$\tilde{x} = \tilde{x}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} \tilde{z}^- \quad (6.3.57)$$

$$x - \hat{x} = x - \hat{x}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} (z - \hat{z}^-) \quad (6.3.58)$$

$$\hat{x} = \hat{x}^- + \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1} (z - \hat{z}^-) \quad (6.3.59)$$

Let  $\mathbf{K} = \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1}$  be known as the gain and returning the time index  $k$ :

$$\hat{x}_k = \hat{x}_k^- + \mathbf{K}_k (z_k - \hat{z}_k^-) \quad (6.3.60)$$

This gives an update equation for the posterior mean  $\hat{x}_k$  in terms of the prior mean  $\hat{x}_k^-$ , prior output  $\hat{z}_k^-$  and prior covariances  $\mathbf{P}_{xz}^-$ ,  $\mathbf{P}_{zz}^-$  (which use information only up to including time  $k-1$ , and the realisation of the measurement at time  $k$ , which is  $z_k$ ). Equating the Hessian matrix in the equation:

$$\mathbf{P}_{xx}^{-1} = \mathbf{B}^{-1} \quad (6.3.61)$$

$$\mathbf{P}_{xx} = \mathbf{P}_{xx}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^- \quad (6.3.62)$$

Inserting a factor  $I = (\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zz}^-$ :

$$\mathbf{P}_{xx} = \mathbf{P}_{xx}^- - \mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zz}^- (\mathbf{P}_{zz}^-)^{-1}\mathbf{P}_{zx}^- \quad (6.3.63)$$

$$= \mathbf{P}_{xx}^- - [\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1}] \mathbf{P}_{zz}^- [\mathbf{P}_{xz}^-(\mathbf{P}_{zz}^-)^{-1}]^\top \quad (6.3.64)$$

Recognising the presence of  $\mathbf{K}$  and returning the time index, this gives the update equation for the posterior covariance:

$$\mathbf{P}_{xx,k} = \mathbf{P}_{xx,k}^- - \mathbf{K}_k \mathbf{P}_{zz,k}^- \mathbf{K}_k^\top \quad (6.3.65)$$

## 6.4 Bayesian Inference

### 6.4.1 Maximum a Posteriori Estimation

Maximum a posteriori estimation extends and generalises the concept of maximum likelihood estimation. Suppose we are trying to estimate a parameter  $\theta \in \Theta$ , and have observed data  $\mathcal{D}$ . We may setup the likelihood

$$p(\mathcal{D}|\theta) = \mathcal{L}(\theta; \mathcal{D}) \quad (6.4.1)$$

Traditionally in maximum likelihood estimation, we would aim to maximise the likelihood with respect to  $\theta$ . If a prior distribution  $p(\theta)$  is provided however, then using Bayes' theorem, the posterior distribution for  $\theta$  is given by

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \quad (6.4.2)$$

$$= \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_{\Theta} p(\mathcal{D}|\theta)p(\theta)d\theta} \quad (6.4.3)$$

where the denominator  $p(\mathcal{D})$  is known as the *marginal likelihood*, i.e. to obtain it, we marginalise over  $\theta$  with the integral  $\int_{\Theta} p(\mathcal{D}|\theta)p(\theta)d\theta$ . The maximum a posteriori estimate is defined as the mode of the posterior distribution:

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} p(\theta|\mathcal{D}) \quad (6.4.4)$$

$$= \underset{\theta}{\operatorname{argmax}} \left\{ \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \right\} \quad (6.4.5)$$

Since the denominator does not depend on  $\theta$ , then equivalently

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} \{p(\mathcal{D}|\theta)p(\theta)\} \quad (6.4.6)$$

Like with maximum likelihood estimation, this can be transformed into a minimisation problem, and logs can be taken:

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmin}} \{-\log p(\mathcal{D}|\theta) - \log p(\theta)\} \quad (6.4.7)$$

We can also see that if a **flat prior** is used for  $\theta$ , i.e.  $p(\theta)$  is proportional to a constant, then the maximum a priori estimate becomes identical to the maximum likelihood estimate. In this way maximum likelihood estimation is a special case of maximum a posteriori estimation.

### 6.4.2 Bayes Estimators

### 6.4.3 Credible Intervals

### 6.4.4 Posterior Predictive Distributions

### 6.4.5 Bayes Factors [19]

A Bayesian approach to hypothesis testing can be performed using Bayes factors. Let  $H_1$  and  $H_2$  be competing hypotheses, and we observe data  $\mathcal{D}$ . Let  $\ell_{ij}$  denote the loss incurred from choosing hypothesis  $i$  when in fact  $j$  was the correct hypothesis. Suppose that  $\ell_{11} = \ell_{22} = 0$  (i.e. there is no loss incurred in choosing the correct model) and  $\ell_{12}, \ell_{21} > 0$ . Introduce the utility  $U_i|\mathcal{D}$  in choosing model  $i$ . Then the expected utility of choosing model  $i$  is

$$\mathbb{E}[U_i|\mathcal{D}] = -\ell_{i1}\Pr(H_1|\mathcal{D}) - \ell_{i2}\Pr(H_2|\mathcal{D}) \quad (6.4.8)$$

So the expected utility of choosing  $H_2$  is greater than the expected utility of choosing the  $H_1$  if

$$\mathbb{E}[U_2|\mathcal{D}] > \mathbb{E}[U_1|\mathcal{D}] \quad (6.4.9)$$

$$-\ell_{21} \Pr(H_1|\mathcal{D}) - \ell_{22} \Pr(H_2|\mathcal{D}) > -\ell_{11} \Pr(H_1|\mathcal{D}) - \ell_{12} \Pr(H_2|\mathcal{D}) \quad (6.4.10)$$

$$-\ell_{21} \Pr(H_1|\mathcal{D}) > -\ell_{12} \Pr(H_2|\mathcal{D}) \quad (6.4.11)$$

$$\frac{\Pr(H_1|\mathcal{D})}{\Pr(H_2|\mathcal{D})} < \frac{\ell_{12}}{\ell_{21}} \quad (6.4.12)$$

Note that the left-hand side is a posterior odds ratio. This gives the basis for the Bayes test. Given data  $\mathcal{D}$ , we compute the Bayes factor  $B_{12}|\mathcal{D}$  in favour of  $H_1$  against  $H_2$  by the likelihood ratio:

$$B_{12}|\mathcal{D} = \frac{\Pr(\mathcal{D}|H_1)}{\Pr(\mathcal{D}|H_2)} \quad (6.4.13)$$

$$= \frac{\Pr(H_1|\mathcal{D}) \Pr(\mathcal{D}) / \Pr(H_1)}{\Pr(H_2|\mathcal{D}) \Pr(\mathcal{D}) / \Pr(H_2)} \quad (6.4.14)$$

$$= \frac{\Pr(H_1|\mathcal{D})}{\Pr(H_2|\mathcal{D})} \div \frac{\Pr(H_1)}{\Pr(H_2)} \quad (6.4.15)$$

which can be written as a ratio of the posterior odds ratio to the prior odds ratio. Rearranging for the posterior odds ratio, we get

$$\frac{\Pr(H_1|\mathcal{D})}{\Pr(H_2|\mathcal{D})} = B_{12}|\mathcal{D} \times \frac{\Pr(H_1)}{\Pr(H_2)} \quad (6.4.16)$$

Thus we reject  $H_1$  in favour of  $H_2$  if and only if

$$B_{12}|\mathcal{D} \times \frac{\Pr(H_1)}{\Pr(H_2)} < \frac{\ell_{12}}{\ell_{21}} \quad (6.4.17)$$

or equivalently

$$B_{12}|\mathcal{D} < \frac{\ell_{12}}{\ell_{21}} \frac{\Pr(H_2)}{\Pr(H_1)} \quad (6.4.18)$$

A strong resemblance can be seen between this decision rule and that from **minimum cost binary hypothesis testing**. Note that if we are performing model selection, and hypotheses  $H_1$  and  $H_2$  consist of model structures parametrised by parameters  $\theta_1$  and  $\theta_2$  respectively, we may compute the Bayes factor by integrating over the priors on the parameter values like so:

$$B_{12}|\mathcal{D} = \frac{\int p(\mathcal{D}|\theta_1) p(\theta_1)}{\int p(\mathcal{D}|\theta_2) p(\theta_2)} \div \frac{\Pr(H_1)}{\Pr(H_2)} \quad (6.4.19)$$

#### 6.4.6 Bayesian Regularisation

Consider fitting the parameters of a model, where the conditional distribution of the output  $y$  given the model parameters  $\theta$  and inputs  $x$  are Gaussian:

$$y|\theta, x \sim \mathcal{N}(f(\theta, x), \sigma^2) \quad (6.4.20)$$

That is, observations are generated by the process

$$y = f(\theta, x) + \varepsilon \quad (6.4.21)$$

where  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ . After a dataset of  $n$  observations has been collected:

$$\mathbf{y} = (y_1, \dots, y_n) \quad (6.4.22)$$

$$\mathbf{x} = (x_1, \dots, x_n) \quad (6.4.23)$$

and assuming the  $\varepsilon_i$  are i.i.d., we can form the likelihood as

$$p(\mathbf{y}|\theta, \mathbf{x}) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - f(\theta, x_i))^2}{2\sigma^2}\right) \quad (6.4.24)$$

Furthermore, suppose we impose a Gaussian prior on the parameter vector  $\theta$ , with independent components:

$$\theta \sim \mathcal{N}(\mathbf{0}, \sigma_\theta^2 I) \quad (6.4.25)$$

The prior distribution can then be written as

$$p(\theta) = \prod_{j=1}^d \frac{1}{\sigma_\theta\sqrt{2\pi}} \exp\left(-\frac{\theta_j^2}{2\sigma_\theta^2}\right) \quad (6.4.26)$$

Formulating the posterior for  $\theta$  given the data, we have

$$p(\theta|\mathbf{y}, \mathbf{x}) = \frac{p(\mathbf{y}|\theta, \mathbf{x}) p(\theta)}{p(\mathbf{y}|\mathbf{x})} \quad (6.4.27)$$

$$= \frac{p(\mathbf{y}|\theta, \mathbf{x}) p(\theta)}{p(\mathbf{y}|\mathbf{x})} \quad (6.4.28)$$

where we have treated  $\theta$  and  $\mathbf{x}$  as being independent. The maximum a posteriori estimate may be found through

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} \{p(\mathbf{y}|\theta, \mathbf{x}) p(\theta)\} \quad (6.4.29)$$

$$= \underset{\theta}{\operatorname{argmin}} \{-\log p(\mathbf{y}|\theta, \mathbf{x}) - \log p(\theta)\} \quad (6.4.30)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ -\log \left( \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - f(\theta, x_i))^2}{2\sigma^2}\right) \right) - \log \left( \prod_{j=1}^d \frac{1}{\sigma_\theta\sqrt{2\pi}} \exp\left(-\frac{\theta_j^2}{2\sigma_\theta^2}\right) \right) \right\} \quad (6.4.31)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^n \log \left( \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - f(\theta, x_i))^2}{2\sigma^2}\right) \right) - \sum_{i=1}^d \log \left( \frac{1}{\sigma_\theta\sqrt{2\pi}} \exp\left(-\frac{\theta_j^2}{2\sigma_\theta^2}\right) \right) \right\} \quad (6.4.32)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \frac{(y_i - f(\theta, x_i))^2}{2\sigma^2} + \sum_{i=1}^d \frac{\theta_j^2}{2\sigma_\theta^2} \right\} \quad (6.4.33)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \frac{(y_i - f(\theta, x_i))^2}{2\sigma^2} + \sum_{i=1}^d \frac{\theta_j^2}{2\sigma_\theta^2} \right\} \quad (6.4.34)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - f(\theta, x_i))^2 + \frac{\sigma^2}{2\sigma_\theta^2} \sum_{i=1}^d \theta_j^2 \right\} \quad (6.4.35)$$

By substituting  $\lambda = \frac{\sigma^2}{2\sigma_\theta^2}$ , we get

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - f(\theta, x_i))^2 + \lambda \sum_{i=1}^d \theta_j^2 \right\} \quad (6.4.36)$$

From this, we can see that this is effectively solving a nonlinear least squares problem with  $\ell_2$  regularisation. In this sense, a prior can be thought of as providing regularisation, or alternatively,  $\ell_2$  regularisation can be thought of as placing a Gaussian prior on the parameters. We can also intuitively see how the ratio  $\sigma^2/\sigma_\theta^2$  affects the regularisation, which reflects the amount of ‘trust’ in the data relative to the prior. If  $\sigma_\theta^2$  is small relative to  $\sigma^2$ , this indicates that we are more certain about the zero-mean prior, and thus the estimator will be heavily regularised to zero.

Instead of using a Gaussian prior, a Laplace prior can be used instead:

$$p(\theta) = \prod_{j=1}^d \frac{\sqrt{2}}{\sigma_\theta} \exp\left(-\frac{|\theta_j|}{\sigma_\theta/\sqrt{2}}\right) \quad (6.4.37)$$

The log-prior now becomes:

$$-\log p(\theta) = -\log \left( \prod_{j=1}^d \frac{\sqrt{2}}{\sigma_\theta} \exp\left(-\frac{|\theta_j|}{\sigma_\theta/\sqrt{2}}\right) \right) \quad (6.4.38)$$

$$= \sum_{i=1}^d \frac{|\theta_j|}{\sigma_\theta/\sqrt{2}} - \sum_{i=1}^d \log\left(\frac{\sqrt{2}}{\sigma_\theta}\right) \quad (6.4.39)$$

Thus the maximum a posteriori estimator is now

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \frac{(y_i - f(\theta, x_i))^2}{2\sigma^2} + \sum_{i=1}^d \frac{|\theta_j|}{\sigma_\theta/\sqrt{2}} \right\} \quad (6.4.40)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - f(\theta, x_i))^2 + \frac{\sigma^2 \sqrt{2}}{\sigma_\theta} \sum_{i=1}^d |\theta_j| \right\} \quad (6.4.41)$$

which after letting  $\lambda = \frac{\sigma^2 \sqrt{2}}{\sigma_\theta}$ , is effectively  $\ell_1$  regularisation with a similar interpretation as above. Here, switching from a Gaussian to a Laplace prior yields an analogous outcome to switching from a Gaussian likelihood to a Laplace likelihood.

#### 6.4.7 Bayesian Classifiers

##### Bayes-Optimal Classifier [80]

A classification problem is to predict the discrete random variable  $Y$  (i.e. categorical variable) from the random element  $\mathbf{X}$  (which could be a random vector or simply a random variable). Let  $K$  be the number of classes which  $Y$  can take. A classification rule is a function  $h : \mathcal{X} \rightarrow \{0, \dots, K-1\}$ , where  $\mathcal{X}$  is the domain (or support) of  $\mathbf{X}$ . We introduce a *loss function*, which can be thought of as the cost of a single prediction for the example  $(\mathbf{x}, y)$ . The 0-1 loss function is defined as

$$L(y, h(\mathbf{x})) = \begin{cases} 0, & h(\mathbf{x}) = y \\ 1, & h(\mathbf{x}) \neq y \end{cases} \quad (6.4.42)$$

That is, the loss is zero if the classifier gets the prediction right, and one if the classifier gets the prediction wrong. We may define the *risk* of a classifier as the expected loss using  $h(\cdot)$  over the joint distribution of  $(\mathbf{X}, Y)$ , i.e.

$$R(h) = \mathbb{E}[L(Y, h(\mathbf{X}))] \quad (6.4.43)$$

We can show that for the 0-1 loss function, the risk is equal to the probability of misclassification:

$$R(h) = \mathbb{E}[L(Y, h(\mathbf{X}))] \quad (6.4.44)$$

$$= \mathbb{E}[\mathbb{I}_{\{Y \neq h(\mathbf{X})\}}] \quad (6.4.45)$$

$$= \Pr(Y \neq h(\mathbf{X})) \quad (6.4.46)$$

We can attempt to derive the classifier which minimises the risk. Rewriting the risk by conditioning on  $\mathbf{X}$ :

$$R(h) = \mathbb{E}_{\mathbf{X}} [\mathbb{E}[L(Y, h(\mathbf{X}))|\mathbf{X}]] \quad (6.4.47)$$

$$= \mathbb{E}_{\mathbf{X}} \left[ \sum_{k=0}^K L(k, h(\mathbf{x})) \Pr(Y = k | \mathbf{X} = \mathbf{x}) \right] \quad (6.4.48)$$

A property of the optimal classifier  $h^*(\mathbf{x})$  is that it should minimise the inner conditional expectation pointwise:

$$\mathbb{E}[L(Y, h^*(\mathbf{x}))|\mathbf{X} = \mathbf{x}] \leq \mathbb{E}[L(Y, h(\mathbf{x}))|\mathbf{X} = \mathbf{x}] \quad (6.4.49)$$

for all  $\mathbf{x} \in \mathcal{X}$ , which means

$$\sum_{k=0}^K L(k, h^*(\mathbf{x})) \Pr(Y = k | \mathbf{X} = \mathbf{x}) \leq \sum_{k=0}^K L(k, h(\mathbf{x})) \Pr(Y = k | \mathbf{X} = \mathbf{x}) \quad (6.4.50)$$

From this, we deduce that the optimal prediction is found by

$$h^*(\mathbf{x}) = \operatorname{argmin}_g \sum_{k=0}^K L(k, g) \Pr(Y = k | \mathbf{X} = \mathbf{x}) \quad (6.4.51)$$

Due to the 0-1 loss function we have  $L(k, g) = 1$  when  $k \neq g$ , so this simplifies to

$$h^*(\mathbf{x}) = \operatorname{argmin}_{g \in \{0, \dots, K-1\}} \sum_{k \neq g}^K \Pr(Y = k | \mathbf{X} = \mathbf{x}) \quad (6.4.52)$$

$$= \operatorname{argmax}_{g \in \{0, \dots, K-1\}} \{1 - \Pr(Y = g | \mathbf{X} = \mathbf{x})\} \quad (6.4.53)$$

or equivalently,

$$h^*(\mathbf{x}) = \operatorname{argmax}_{g \in \{0, \dots, K-1\}} \Pr(Y = g | \mathbf{X} = \mathbf{x}) \quad (6.4.54)$$

This is known as the Bayes-optimal classifier, and intuitively means that we should classify  $\mathbf{X}$  in its most probable class (in a Bayesian sense, hence the Bayes moniker).

## Bayes Error

As shown above, the probability of misclassification (i.e. error rate) of a classifier is equal to its risk with a 0-1 loss function. The error rate with the Bayes-optimal classifier

$$R(h^*) = \Pr(Y \neq h^*(\mathbf{X})) \quad (6.4.55)$$

is known as the Bayes-optimal risk or the Bayes error rate, because it represents the error rate with the ‘best possible’ classifier. Thus, the Bayes error rate can be interpreted as the irreducible part of error rate, such that even with infinite data (i.e. essentially having knowledge of the true underlying distribution), it is impossible for a classifier to perform better. This is because the data-generating process is inherently stochastic.

## Naïve Bayes

To classify the label  $y$  for  $n$  features  $\mathbf{x} = (x_1, \dots, x_n)$ , we can use Bayes' theorem to write the posterior distribution as:

$$p(y|x_1, \dots, x_n) = \frac{p(y)p(\mathbf{x}|y)}{p(\mathbf{x})} \quad (6.4.56)$$

which is proportional to the denominator:

$$p(y|\mathbf{x}) \propto p(y)p(\mathbf{x}|y) \quad (6.4.57)$$

$$\propto p(\mathbf{x}, y) \quad (6.4.58)$$

where the joint distribution  $p(\mathbf{x}, y)$  factorises to

$$p(\mathbf{x}, y) = p(x_1|x_2, \dots, x_n, y) \dots p(x_n|y) p(y) \quad (6.4.59)$$

The naivety here is to assume that the features of  $\mathbf{x}$  are independent, so that we can instead factorise  $p(\mathbf{x}, y)$  as

$$p(\mathbf{x}, y) = p(y) \prod_{i=1}^n p(x_i|y) \quad (6.4.60)$$

so that the posterior is proportional to

$$p(y|\mathbf{x}) \propto p(y) \prod_{i=1}^n p(x_i|y) \quad (6.4.61)$$

Applying the Bayes-optimal classification rule (i.e. the maximum a posteriori class), the naïve Bayes classifier is

$$h(\mathbf{x}) = \operatorname{argmax}_{y \in \{0, \dots, K-1\}} p(y|\mathbf{x}) \quad (6.4.62)$$

$$= \operatorname{argmax}_{y \in \{0, \dots, K-1\}} p(y) \prod_{i=1}^n p(x_i|y) \quad (6.4.63)$$

### 6.4.8 Bayesian Linear Regression [160]

Consider a linear model of the form with weights  $\mathbf{w}$ :

$$f(x) = x^\top \mathbf{w} \quad (6.4.64)$$

with the data generating process

$$y = f(x) + \varepsilon \quad (6.4.65)$$

where the noise distribution is given by  $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ . Denote the data by  $\mathcal{D} = (X, \mathbf{y})$  where  $X = [x_1 \ x_2 \ \dots \ x_n]$  consists of the feature vectors  $x_1, \dots, x_n$  aggregated column-wise and  $\mathbf{y}$  is a vector of responses. Constructing the likelihood under an i.i.d. assumption:

$$\mathcal{L}(\mathcal{D}|X, \mathbf{w}) = p(\mathbf{y}|X, \mathbf{w}) \quad (6.4.66)$$

$$= \prod_{i=1}^n p(y_i|x_i, \mathbf{w}) \quad (6.4.67)$$

Since  $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ , then each

$$y_i = f(x_i) + \varepsilon_i \quad (6.4.68)$$

$$\sim \mathcal{N}(f(x_i), \sigma_n^2) \quad (6.4.69)$$

$$\sim \mathcal{N} \left( x_i^\top \mathbf{w}, \sigma_n^2 \right) \quad (6.4.70)$$

The form of the Gaussian density is given by

$$p(y_i|x_i, \mathbf{w}) = \frac{1}{\sigma_n \sqrt{2\pi}} \exp \left[ -\frac{(y_i - x_i^\top \mathbf{w})^2}{2\sigma_n^2} \right] \quad (6.4.71)$$

hence the likelihood becomes

$$\prod_{i=1}^n p(y_i|x_i, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sigma_n \sqrt{2\pi}} \exp \left[ -\frac{(y_i - x_i^\top \mathbf{w})^2}{2\sigma_n^2} \right] \quad (6.4.72)$$

$$= \frac{1}{\sigma_n \sqrt{2\pi}} \exp \left[ -\sum_{i=1}^n \frac{(y_i - x_i^\top \mathbf{w})^2}{2\sigma_n^2} \right] \quad (6.4.73)$$

$$= \frac{1}{\sigma_n \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma_n^2} \sum_{i=1}^n (y_i - x_i^\top \mathbf{w})^2 \right] \quad (6.4.74)$$

$$= \frac{1}{\sigma_n \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma_n^2} \sum_{i=1}^n \| \mathbf{y} - X^\top \mathbf{w} \|^2 \right] \quad (6.4.75)$$

Therefore the likelihood can be written compactly as

$$p(\mathbf{y}|X, \mathbf{w}) = \mathcal{N} \left( X^\top \mathbf{w}, \sigma_n^2 I \right) \quad (6.4.76)$$

Suppose the prior for the weights is a Gaussian:

$$\mathbf{w} \sim \mathcal{N} (\mathbf{0}, \Sigma_p) \quad (6.4.77)$$

$$p(\mathbf{w}) \propto \exp \left( -\frac{1}{2} \mathbf{w}^\top \Sigma_p^{-1} \mathbf{w} \right) \quad (6.4.78)$$

By Bayes' theorem, then the posterior distribution for the weights is

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w}|X)}{p(\mathbf{y}|X)} \quad (6.4.79)$$

where  $\mathbf{w}$  and  $\mathbf{y}$  are the variables being exchanged, and all conditioned on  $X$ . Since we can assume that  $X$  and  $\mathbf{w}$  are independent, then

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w})}{p(\mathbf{y}|X)} \quad (6.4.80)$$

This can be rewritten as

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w})}{\int p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}} \quad (6.4.81)$$

to express that the marginal likelihood  $p(\mathbf{y}|X)$  is obtained via marginalisation of the likelihood multiplied by prior. Hence  $p(\mathbf{y}|X)$  is a constant with respect to  $\mathbf{w}$ , and the posterior is proportional to the numerator:

$$p(\mathbf{w}|\mathbf{y}, X) \propto p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w}) \quad (6.4.82)$$

$$\propto \exp \left[ -\frac{1}{2\sigma_n^2} (\mathbf{y} - X^\top \mathbf{w})^\top (\mathbf{y} - X^\top \mathbf{w}) \right] \exp \left( -\frac{1}{2} \mathbf{w}^\top \Sigma_p^{-1} \mathbf{w} \right) \quad (6.4.83)$$

$$= \exp \left[ -\frac{1}{2} \left[ \frac{(\mathbf{y} - X^\top \mathbf{w})^\top (\mathbf{y} - X^\top \mathbf{w})}{\sigma_n^2} + \mathbf{w}^\top \Sigma_p^{-1} \mathbf{w} \right] \right] \quad (6.4.84)$$

$$= \exp \left[ -\frac{1}{2} \left[ \frac{1}{\sigma_n^2} \left( \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top X^\top \mathbf{w} - \mathbf{w}^\top X \mathbf{y} + \mathbf{w}^\top X X^\top \mathbf{w} + \mathbf{w}^\top \sigma_n^2 \Sigma_p^{-1} \mathbf{w} \right) \right] \right] \quad (6.4.85)$$

$$= \exp \left[ -\frac{1}{2} \left[ \mathbf{w}^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) \mathbf{w} + \frac{\mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top X^\top \mathbf{w} - \mathbf{w}^\top X \mathbf{y}}{\sigma_n^2} \right] \right] \quad (6.4.86)$$

We now complete the square by first writing

$$\begin{aligned} p(\mathbf{w} | \mathbf{y}, X) \propto & \exp \left[ -\frac{1}{2} \left[ \mathbf{w}^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) \mathbf{w} \right. \right. \\ & - \underbrace{\frac{1}{\sigma_n^2} \mathbf{y}^\top X^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1} \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)}_{I} \mathbf{w} \\ & \left. \left. - \underbrace{\frac{1}{\sigma_n^2} \mathbf{w}^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1} X \mathbf{y} + \frac{\mathbf{y}^\top \mathbf{y}}{\sigma_n^2}}_I \right] \right] \quad (6.4.87) \end{aligned}$$

Define  $\bar{\mathbf{w}} := \frac{1}{\sigma_n^2} \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1} X \mathbf{y}$  and by noting that  $X X^\top + \sigma_n^2 \Sigma_p^{-1}$  is symmetric, we have  $\bar{\mathbf{w}}^\top = \frac{1}{\sigma_n^2} \mathbf{y}^\top X^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1}$ . Then we can see that

$$\begin{aligned} p(\mathbf{w} | \mathbf{y}, X) \propto & \exp \left[ -\frac{1}{2} \left[ \mathbf{w}^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) \mathbf{w} \right. \right. \\ & - \underbrace{\frac{1}{\sigma_n^2} \mathbf{y}^\top X^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1} \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)}_{\bar{\mathbf{w}}^\top} \mathbf{w} \\ & \left. \left. - \bar{\mathbf{w}}^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) \underbrace{\frac{1}{\sigma_n^2} \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1} X \mathbf{y} + \frac{\mathbf{y}^\top \mathbf{y}}{\sigma_n^2}}_{\bar{\mathbf{w}}} \right] \right] \quad (6.4.88) \end{aligned}$$

and we can finish completing the square by writing

$$\begin{aligned} p(\mathbf{w} | \mathbf{y}, X) \propto & \exp \left[ -\frac{1}{2} \left[ (\mathbf{w} - \bar{\mathbf{w}})^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) (\mathbf{w} - \bar{\mathbf{w}}) \right. \right. \\ & \left. \left. - \bar{\mathbf{w}}^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) \bar{\mathbf{w}} + \frac{\mathbf{y}^\top \mathbf{y}}{\sigma_n^2} \right] \right] \quad (6.4.89) \end{aligned}$$

$$\begin{aligned} = & \exp \left[ -\frac{1}{2} \left[ (\mathbf{w} - \bar{\mathbf{w}})^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) (\mathbf{w} - \bar{\mathbf{w}}) \right. \right. \\ & \left. \left. - \frac{1}{\sigma_n^4} \mathbf{y}^\top X^\top \left( \frac{X X^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1} X \mathbf{y} + \frac{\mathbf{y}^\top \mathbf{y}}{\sigma_n^2} \right] \right] \quad (6.4.90) \end{aligned}$$

$$\propto \exp \left[ -\frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^\top \left( \frac{XX^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right) (\mathbf{w} - \bar{\mathbf{w}}) \right] \quad (6.4.91)$$

where the terms not involving  $\mathbf{w}$  have been ignored due to proportionality. Hence it can be deduced

$$p(\mathbf{w}|\mathbf{y}, X) = \mathcal{N} \left( \bar{\mathbf{w}}, \left( \frac{XX^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)^{-1} \right) \quad (6.4.92)$$

For ease of notation, let  $A = \left( \frac{XX^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2} \right)$  so that  $\bar{\mathbf{w}} = \frac{1}{\sigma_n^2} A^{-1} X \mathbf{y}$  and

$$p(\mathbf{w}|\mathbf{y}, X) = \mathcal{N} \left( \frac{1}{\sigma_n^2} A^{-1} X \mathbf{y}, A^{-1} \right) \quad (6.4.93)$$

This posterior distribution represents a distribution over the possible linear models. To obtain a predictive distribution  $f_* := f(x_*)$  at a test point  $x_*$ , we can integrate over the posterior:

$$p(f_*|x_*, X, \mathbf{y}) = \int p(f_*|x_*, \mathbf{w}, X, \mathbf{y}) p(\mathbf{w}|x_*, X, \mathbf{y}) d\mathbf{w} \quad (6.4.94)$$

By the i.i.d. assumption,  $p(f_*|x_*, \mathbf{w}, X, \mathbf{y}) = p(f_*|x_*, \mathbf{w})$  and also by independence of the weights and the test point,  $p(\mathbf{w}|x_*, X, \mathbf{y}) = p(\mathbf{w}|X, \mathbf{y})$  which gives

$$p(f_*|x_*, X, \mathbf{y}) = \int p(f_*|x_*, \mathbf{w}) p(\mathbf{w}|X, \mathbf{y}) d\mathbf{w} \quad (6.4.95)$$

where  $p(\mathbf{w}|X, \mathbf{y})$  as above and  $p(f_*|x_*, \mathbf{w}) = \mathcal{N}(x_*^\top \mathbf{w}, \sigma_n^2)$ . But rather than evaluating the integral, note that  $f_*$  is just a linear transformation of the random variable  $\mathbf{w}$ , given by  $f_* = x_*^\top \mathbf{w}$ . By using the property of affine transformations of Gaussian random variables, this gives

$$p(f_*|x_*, X, \mathbf{y}) = \mathcal{N} \left( \frac{1}{\sigma_n^2} x_*^\top A^{-1} X \mathbf{y}, x_*^\top A^{-1} x_* \right) \quad (6.4.96)$$

### Basis Function Bayesian Linear Regression

A regression model can be linear in a basis of  $x$ , denoted  $\phi(x)$ , given by

$$f(x) = \phi(x)^\top \mathbf{w} \quad (6.4.97)$$

Defining the data matrix aggregated column-wise  $\Phi = [\phi(x_1) \ \phi(x_2) \ \dots \ \phi(x_n)]$ , the predictive distribution for a test point  $x_*$  is now given by

$$p(f_*|x_*, X, \mathbf{y}) = \mathcal{N} \left( \frac{1}{\sigma_n^2} \phi(x_*)^\top A^{-1} \Phi \mathbf{y}, \phi(x_*)^\top A^{-1} \phi(x_*) \right) \quad (6.4.98)$$

where  $A = \frac{\Phi \Phi^\top + \sigma_n^2 \Sigma_p^{-1}}{\sigma_n^2}$ . Denote  $\phi_* := \phi(x_*)$  for convenience. We can show that the mean can be written as

$$\frac{1}{\sigma_n^2} \phi_*^\top A^{-1} \Phi \mathbf{y} = \phi_*^\top \Sigma_p \Phi (K + \sigma_n^2 I)^{-1} \mathbf{y} \quad (6.4.99)$$

where  $K = \Phi^\top \Sigma_p \Phi$ . To do this, it suffices to show that

$$\frac{A^{-1} \Phi}{\sigma_n^2} = \Sigma_p \Phi (K + \sigma_n^2 I)^{-1} \quad (6.4.100)$$

Using the definition of  $A$ :

$$\left( \Phi \Phi^\top + \sigma_n^2 \Sigma_p^{-1} \right)^{-1} \Phi = \Sigma_p \Phi \left( K + \sigma_n^2 I \right)^{-1} \quad (6.4.101)$$

$$\Phi \left( K + \sigma_n^2 I \right) = \left( \Phi \Phi^\top + \sigma_n^2 \Sigma_p^{-1} \right) \Sigma_p \Phi \quad (6.4.102)$$

$$\Phi \left( \Phi^\top \Sigma_p \Phi + \sigma_n^2 I \right) = \left( \Phi \Phi^\top + \sigma_n^2 \Sigma_p^{-1} \right) \Sigma_p \Phi \quad (6.4.103)$$

$$\Phi \Phi^\top \Sigma_p \Phi + \sigma_n^2 \Phi = \Phi \Phi^\top \Sigma_p \Phi + \sigma_n^2 \Phi \quad (6.4.104)$$

as required. We also find an alternative form for the predictive covariance. Again using the definition of  $A$ :

$$\phi_*^\top A^{-1} \phi_* = \phi_*^\top \left( \Sigma_p^{-1} + \Phi \sigma_n^{-2} I \Phi^\top \right)^{-1} \Phi \quad (6.4.105)$$

Now apply the matrix inversion lemma:

$$\left( Z + UWV^\top \right)^{-1} = Z^{-1} - Z^{-1}U \left( W^{-1} + V^\top Z^{-1}U \right)^{-1} V^\top Z^{-1} \quad (6.4.106)$$

using  $Z = \Sigma_p^{-1}$ ,  $U = \Phi$ ,  $W = \sigma_n^{-2}I$  and  $V = \Phi$  to give

$$A^{-1} = \Sigma_p - \Sigma_p \Phi \left( \sigma_n I + \Phi^\top \Sigma_p \Phi \right)^{-1} \Phi^\top \Sigma_p \quad (6.4.107)$$

Then use the definition of  $K$ :

$$A^{-1} = \Sigma_p - \Sigma_p \Phi \left( K + \sigma_n^2 I \right)^{-1} \Phi^\top \Sigma_p \quad (6.4.108)$$

Hence an alternative form for the predictive covariance is

$$\phi_*^\top A^{-1} \phi_* = \phi_*^\top \Sigma_p \phi_* - \phi_*^\top \Sigma_p \Phi \left( K + \sigma_n^2 I \right)^{-1} \Phi^\top \Sigma_p \phi_* \quad (6.4.109)$$

The predictive distribution has the new form

$$p(f_*|x_*, X, \mathbf{y}) = \mathcal{N} \left( \phi_*^\top \Sigma_p \Phi \left( K + \sigma_n^2 I \right)^{-1} \mathbf{y}, \phi_*^\top \Sigma_p \phi_* - \phi_*^\top \Sigma_p \Phi \left( K + \sigma_n^2 I \right)^{-1} \Phi^\top \Sigma_p \phi_* \right) \quad (6.4.110)$$

Note that the terms in blue will be of the form  $\phi(x)^\top \Sigma_p \phi(x')$  where  $x$  and  $x'$  can be either from the test set or training set.

#### 6.4.9 Type II Maximum Likelihood [143]

#### 6.4.10 Hierarchical Bayes Modelling [160]

Consider a model over parameters  $\mathbf{w}$  (e.g. weights in a neural network), hyperparameters  $\boldsymbol{\theta}$  (e.g. regularisation in a cost function) and a discrete set of structures  $\mathcal{H}_i$  (e.g. number of layers and nodes). We can conduct inference on these using a hierarchical approach. On the lowest level (level 1) is inference over the weights, given inputs  $X$ , outputs  $\mathbf{y}$  and  $\boldsymbol{\theta}$ ,  $\mathcal{H}_i$ .

$$p(\mathbf{w}|\mathbf{y}, X, \boldsymbol{\theta}, \mathcal{H}_i) = \frac{p(\mathbf{y}|\mathbf{w}, X, \boldsymbol{\theta}, \mathcal{H}_i)p(\mathbf{w}|X, \boldsymbol{\theta}, \mathcal{H}_i)}{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)} \quad (6.4.111)$$

As a common assumption is that  $\mathbf{w}$  and  $X$  are independent

$$p(\mathbf{w}|\mathbf{y}, X, \boldsymbol{\theta}, \mathcal{H}_i) = \frac{p(\mathbf{y}|\mathbf{w}, X, \boldsymbol{\theta}, \mathcal{H}_i)p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i)}{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)} \quad (6.4.112)$$

A property of hyperparameters is that they can be selected before we gather training data and before we begin training. It follows that  $\boldsymbol{\theta}$  should be independent of  $\mathbf{y}$ , giving

$$p(\mathbf{w}|\mathbf{y}, X, \boldsymbol{\theta}, \mathcal{H}_i) = \frac{p(\mathbf{y}|\mathbf{w}, X, \mathcal{H}_i)p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i)}{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)} \quad (6.4.113)$$

In level 2, we conduct inference over  $\boldsymbol{\theta}$ .

$$p(\boldsymbol{\theta}|\mathbf{y}, X, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)p(\boldsymbol{\theta}|X, \mathcal{H}_i)}{p(\mathbf{y}|X, \mathcal{H}_i)} \quad (6.4.114)$$

Again, we argue that  $\boldsymbol{\theta}$  should be independent of the data  $X$ , so

$$p(\boldsymbol{\theta}|\mathbf{y}, X, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)p(\boldsymbol{\theta}|\mathcal{H}_i)}{p(\mathbf{y}|X, \mathcal{H}_i)} \quad (6.4.115)$$

Note that the marginal likelihood in level 1 can be expressed as in integral over  $\mathbf{w}$ :

$$p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i) = \int p(\mathbf{y}|\mathbf{w}, X, \mathcal{H}_i)p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i)d\mathbf{w} \quad (6.4.116)$$

Notice that this level 1 marginal likelihood is the same as the level 2 likelihood. In practical terms, we will conduct inference on  $\boldsymbol{\theta}$  by integrating over  $\mathbf{w}$  using some prior for  $\mathbf{w}$  to find the likelihood. We also require a hyperprior  $p(\boldsymbol{\theta}|\mathcal{H}_i)$  for  $\boldsymbol{\theta}$ . Once this is done, then we may perform inference on  $\mathbf{w}$ .

In level 3, inference is performed over the model structures. As  $\mathcal{H}_i$  was defined to be discrete, we use  $P(\cdot)$  to denote the probability mass function (in contrast to  $p(\cdot)$  for the probability density function). Bayes' theorem for the posterior of  $\mathcal{H}_i$  gives

$$P(\mathcal{H}_i|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathcal{H}_i)P(\mathcal{H}_i)}{p(\mathbf{y}|X)} \quad (6.4.117)$$

The likelihood function  $p(\mathbf{y}|X, \mathcal{H}_i)$  is the same as the marginal likelihood for level 2, and can be computed using the integral

$$p(\mathbf{y}|X, \mathcal{H}_i) = \int p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)p(\boldsymbol{\theta}|\mathcal{H}_i)d\boldsymbol{\theta} \quad (6.4.118)$$

As  $P(\mathcal{H}_i)$  is a probability mass function, the level 3 marginal likelihood is computed using the sum

$$p(\mathbf{y}|X) = \sum_i p(\mathbf{y}|X, \mathcal{H}_i)P(\mathcal{H}_i) \quad (6.4.119)$$

ie. it is a weighted average of some probability density functions, and acts as a normalising constant for the posterior of  $\mathcal{H}_i$ .

## 6.5 Posterior Approximations

### 6.5.1 Laplace's Approximation

Laplace's approximation may be used to approximate definite integrals of the form  $\int_a^b \exp(Mf(x))dx$  where  $f(x)$  is a twice-differentiable function which has a unique global minimum  $a < x_0 < b$  and  $M$  is a large number. The terminals  $a$  and  $b$  are allowed to be infinite. The approximation first begins with a second-order Taylor approximation of  $f(x)$  about its global minimum  $x_0$ :

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 \quad (6.5.1)$$

Since  $x_0$  is a global maximum,  $f'(x_0) = 0$  (also note  $f''(x_0) \leq 0$ ) so the first order term vanishes:

$$f(x) \approx f(x_0) + \frac{1}{2} f''(x_0) (x - x_0)^2 \quad (6.5.2)$$

$$= f(x_0) - \frac{1}{2} |f''(x_0)| (x - x_0)^2 \quad (6.5.3)$$

Putting this approximation in the integral:

$$\int_a^b \exp(Mf(x)) dx \approx \int_a^b \exp\left[M\left(f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2\right)\right] dx \quad (6.5.4)$$

$$= e^{Mf(x_0)} \int_a^b \exp\left[-\frac{M}{2}|f''(x_0)|(x - x_0)^2\right] dx \quad (6.5.5)$$

As  $M$  is assumed large, the exponential inside the integrand decays quickly, such that even for finite  $a, b$  we can approximate the integral with the Gaussian integral:

$$\int_a^b \exp\left[\frac{M}{2}f''(x_0)(x - x_0)^2\right] dx \approx \int_{-\infty}^{\infty} \exp\left[\frac{M}{2}|f''(x_0)|(x - x_0)^2\right] dx \quad (6.5.6)$$

$$= \sqrt{\frac{2\pi}{M|f''(x_0)|}} \quad (6.5.7)$$

Hence this gives Laplace's approximation:

$$\int_a^b \exp(Mf(x)) dx \approx e^{Mf(x_0)} \sqrt{\frac{2\pi}{M|f''(x_0)|}} \quad (6.5.8)$$

### Multivariate Laplace's Approximation

Laplace's approximation can be extended to the multivariate case. We approximate the integral  $\int_{\mathcal{X}} \exp(Mf(\mathbf{x})) d\mathbf{x}$  where twice-differentiable  $f(\mathbf{x})$  is a scalar function of  $n$ -dimensional  $\mathbf{x}$  which has a unique global maximum  $\mathbf{x}_0 \in \mathcal{X}$ , and  $M$  is large. Deriving the approximation follows the analogous arguments as the univariate case:

$$\int_{\mathcal{X}} \exp(Mf(\mathbf{x})) d\mathbf{x} \approx \int_{\mathcal{X}} \exp\left[M\left(f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^\top \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\right)\right] d\mathbf{x} \quad (6.5.9)$$

$$= \int_{\mathcal{X}} \exp\left[M\left(f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\right)\right] d\mathbf{x} \quad (6.5.10)$$

$$= e^{Mf(\mathbf{x}_0)} \int_{\mathcal{X}} \exp\left[-\frac{M}{2}(\mathbf{x} - \mathbf{x}_0)^\top (-\nabla^2 f(\mathbf{x}_0))(\mathbf{x} - \mathbf{x}_0)\right] d\mathbf{x} \quad (6.5.11)$$

$$\approx e^{Mf(\mathbf{x}_0)} \int_{-\infty}^{\infty} \exp\left[-\frac{M}{2}(\mathbf{x} - \mathbf{x}_0)^\top (-\nabla^2 f(\mathbf{x}_0))(\mathbf{x} - \mathbf{x}_0)\right] d\mathbf{x} \quad (6.5.12)$$

$$= e^{Mf(\mathbf{x}_0)} \sqrt{\left(\frac{2\pi}{M}\right)^n \cdot \frac{1}{\det(-\nabla^2 f(\mathbf{x}_0))}} \quad (6.5.13)$$

### 6.5.2 Expectation Propagation [65, 196]

Using the expectation propagation algorithm, we construct an estimate of the posterior  $p(\theta|X)$  for some parameters  $\theta$ , which we denote the approximation by  $q(\theta)$ . We also suppress the

dependence on the data  $X$  in  $q(\theta)$ , because we take it as given and fixed. Assume that the joint distribution  $p(X, \theta)$  can be conveniently factorised as

$$p(X, \theta) = \prod_{i=0}^n f_i(\theta) \quad (6.5.14)$$

For instance, we could write

$$p(X, \theta) = p(\theta) \prod_{i=1}^n p(x_i | \theta) \quad (6.5.15)$$

which is in terms of the likelihood for an i.i.d. sample. From Bayes' theorem, we then obtain

$$p(\theta | X) = \frac{p(X, \theta)}{p(X)} \quad (6.5.16)$$

$$= \frac{1}{p(X)} \prod_{i=0}^n f_i(\theta) \quad (6.5.17)$$

To approximate  $p(\theta | X)$ , we also put  $q(\theta)$  into a factorised form

$$q(\theta) = \frac{1}{Z} \prod_{i=0}^n \hat{f}_i(\theta) \quad (6.5.18)$$

where  $Z$  is treated as the normalising constant. Moreover, we assume that the posterior can be well-approximated if we constrain  $q(\theta)$  to lie in the exponential family of distributions. That is, we can write

$$q(\theta) = h(\theta) g(\eta) \exp\left(\eta^\top T(\theta)\right) \quad (6.5.19)$$

where  $\eta$  are parameters of the exponential family and  $T(\theta)$  are sufficient statistics for  $\eta$ . Next, we describe the approach to measure ‘closeness’ of the approximation, which will be measured in terms of the Kullback-Leibler divergence of  $p(\theta | X)$  from  $q(\theta)$ , defined by

$$\text{KL}(p \| q) = \int p(\theta | X) \log\left(\frac{p(\theta | X)}{q(\theta)}\right) d\theta \quad (6.5.20)$$

Under the exponential family assumption, this can be expressed as

$$\text{KL}(p \| q) = \int p(\theta | X) \log\left(\frac{p(\theta | X)}{h(\theta) g(\eta) \exp(\eta^\top T(\theta))}\right) d\theta \quad (6.5.21)$$

$$= - \int p(\theta | X) \log g(\eta) d\theta - \int p(\theta | X) \eta^\top T(\theta) d\theta + \underbrace{\int p(\theta | X) \log\left(\frac{p(\theta | X)}{h(\theta)}\right) d\theta}_{C} \quad (6.5.22)$$

$$= -\log g(\eta) - \int p(\theta | X) \eta^\top T(\theta) d\theta + C \quad (6.5.23)$$

where the last term does not depend on  $\eta$ . To minimise the KL divergence, we differentiate it with respect to  $\eta$  and obtain

$$\nabla_\eta \text{KL}(p \| q) = -\frac{1}{g(\eta)} \nabla_\eta g(\eta) - \int p(\theta | X) T(\theta) d\theta \quad (6.5.24)$$

$$= -\frac{1}{g(\eta)} \nabla_\eta g(\eta) - \mathbb{E}_p[T(\theta)] \quad (6.5.25)$$

where we have used  $\mathbb{E}_p [\cdot]$  to denote expectation with respect to the distribution  $p(\theta|X)$ . Setting the gradient to zero, this yields

$$\mathbb{E}_p [T(\theta)] = -\frac{1}{g(\eta)} \nabla_\eta g(\eta) \quad (6.5.26)$$

Now return to the form of  $q(\theta)$ , from which:

$$1 = \int q(\theta) d\theta \quad (6.5.27)$$

$$= \int h(\theta) g(\eta) \exp(\eta^\top T(\theta)) d\theta \quad (6.5.28)$$

$$= g(\eta) \int h(\theta) \exp(\eta^\top T(\theta)) d\theta \quad (6.5.29)$$

Differentiating both sides with respect to  $\eta$  using the product rule, we have

$$0 = \nabla_\eta g(\eta) \cdot \int h(\theta) \exp(\eta^\top T(\theta)) d\theta + g(\eta) \int h(\theta) T(\theta) \exp(\eta^\top T(\theta)) d\theta \quad (6.5.30)$$

$$= \frac{1}{g(\eta)} \nabla_\eta g(\eta) + \int q(\theta) T(\theta) d\theta \quad (6.5.31)$$

where we have used  $\int h(\theta) \exp(\eta^\top T(\theta)) d\theta = \frac{1}{g(\eta)}$  from the fact  $1 = \int q(\theta) d\theta$  in order to simplify the first term. Recognise that this then becomes

$$\mathbb{E}_q [T(\theta)] = -\frac{1}{g(\eta)} \nabla_\eta g(\eta) \quad (6.5.32)$$

and thus

$$\mathbb{E}_p [T(\theta)] = \mathbb{E}_q [T(\theta)] \quad (6.5.33)$$

This shows that in order to minimise the KL divergence when  $q(\theta)$  is part of the exponential family, we can just match their moments (via the expectations of the sufficient statistics). This result is very similar to that seen in maximum likelihood for exponential families, and can even be explicitly related by invoking the equivalence between minimum KL divergence and maximum likelihood estimation.

### Expectation Propagation Algorithm

Minimising the KL divergence between the posterior  $p(\theta|X)$  from the approximation  $q(\theta)$  is generally intractable because it essentially requires knowing  $p(\theta|X)$  in the first place. Instead, we can instead minimise in terms of the factors  $f_i(\theta)$  and  $\widehat{f}_i(\theta)$ . The expectation propagation algorithm operates in an iterative fashion. Beginning from an initial approximation  $q(\theta)$  of the posterior and and the  $i^{\text{th}}$  factor  $\widehat{f}_i(\theta)$ , we perform the following steps.

1. Obtain the *cavity distribution*  $q_{\setminus i}(\theta)$ , by ‘pulling out’ the approximating factor, so that

$$q_{\setminus i}(\theta) \propto \frac{q(\theta)}{\widehat{f}_i(\theta)} \quad (6.5.34)$$

2. Form the *titled distribution*, given by

$$\widetilde{q}(\theta) \propto f_i(\theta) q_{\setminus i}(\theta) \quad (6.5.35)$$

which replaces with removed approximating factor by its corresponding factor from the target distribution  $p(\theta|X)$ .

3. Having both  $q(\theta)$  and  $\tilde{q}(\theta)$ , we minimise the KL divergence with

$$\eta^+ = \underset{\eta}{\operatorname{argmin}} \text{KL}(\tilde{q}(\theta) \| q(\theta)) \quad (6.5.36)$$

with respect to the  $\eta$  in  $q(\theta)$  (we do not modify the  $\eta$  in  $\tilde{q}(\theta)$ ). As shown above, this amounts to setting the moments of  $q(\theta)$  equal to those of  $\tilde{q}(\theta)$ . Denote the updated approximation  $q(\theta)$  by  $q^+(\theta)$ .

4. Next, recover the updated approximating factor via

$$\hat{f}_i^+(\theta) \propto \frac{q^+(\theta)}{q_{\setminus i}(\theta)} \quad (6.5.37)$$

We can update all of the factors consecutively (where each factor is being optimised in the context of all the factors), and then re-iterate for convergence. Afterwards, if we are interested in the marginal likelihood  $p(X)$ , this can be approximated with

$$p(X) \approx \int q(\theta) d\theta \quad (6.5.38)$$

### 6.5.3 Variational Inference [65]

Variational inference is another iterative approach similar to expectation propagation for approximating the posterior  $p(\theta|y)$  of some parameters  $\theta$ , given the data  $y$ . Suppose the parameters  $\theta$  are  $d$ -dimensional, and we aim to approximate the posterior with  $q(\theta)$ . The standard approach is to impose a class of distributions on the approximation, and constrain the components to be independent so that

$$q(\theta) = \prod_{j=1}^d q_j(\theta_j) \quad (6.5.39)$$

Let  $\theta_{\setminus j}$  denote all parameters except the  $j^{\text{th}}$ :

$$\theta_{\setminus j} := (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_d) \quad (6.5.40)$$

and let  $q_{\setminus j}(\theta_{\setminus j})$  denote the posterior approximation but with the  $j^{\text{th}}$  component factored out:

$$q_{\setminus j}(\theta_{\setminus j}) := \frac{q(\theta)}{q_j(\theta_j)} \quad (6.5.41)$$

$$= \prod_{\iota \neq j}^d q_\iota(\theta_\iota) \quad (6.5.42)$$

To describe one iteration of the algorithm, we perform for each component  $j = 1, \dots, d$ , using the current posterior approximation  $q(\theta)$ :

1. Obtain the expectation

$$\mathbb{E}_{q_{\setminus j}} [\log p(\theta, y)] = \int q_{\setminus j}(\theta_{\setminus j}) \log(p(y|\theta)p(\theta)) d\theta_{\setminus j} \quad (6.5.43)$$

Since we have integrated out  $\theta_{\setminus j}$ , this will only be a function of  $\theta_j$ .

2. Set the update of the  $j^{\text{th}}$  component such that

$$q_j^+(\theta_j) \propto \exp \left( \mathbb{E}_{q_{\setminus j}} [\log p(\theta, y)] \right) \quad (6.5.44)$$

where for now, we assume that such an update still belongs within the class considered for  $q(\theta)$ .

Once we have updated each of  $q_1(\theta_1), \dots, q_d(\theta_d)$ , we can repeat the procedure again for convergence. It can be shown that each update on any individual  $j$  will decrease the Kullback-Leibler divergence  $\text{KL}(q(\theta) \| p(\theta|y))$ . Note that this is different from expectation propagation, where we minimised the KL divergence of  $p(\theta|y)$  from  $q(\theta)$ . In variational inference, we instead try to minimise the KL divergence of  $q(\theta)$  from  $p(\theta|y)$ . We write this KL divergence as

$$\text{KL}(q(\theta) \| p(\theta|y)) = \int q(\theta) \log \left( \frac{q(\theta)}{p(\theta|y)} \right) d\theta \quad (6.5.45)$$

$$= - \int q(\theta) \log \left( \frac{p(\theta, y)}{q(\theta)p(y)} \right) d\theta \quad (6.5.46)$$

$$= - \int q(\theta) \log \left( \frac{p(\theta, y)}{q(\theta)} \right) d\theta + \int q(\theta) \log p(y) d\theta \quad (6.5.47)$$

$$= - \int q(\theta) \log \left( \frac{p(\theta, y)}{q(\theta)} \right) d\theta + \log p(y) \int q(\theta) d\theta \quad (6.5.48)$$

$$= -\mathbb{E}_q \left[ \log \left( \frac{p(\theta, y)}{q(\theta)} \right) \right] + \log p(y) \quad (6.5.49)$$

Note that the term  $\log p(y)$  does not depend on  $\theta$ . We call the term  $\mathbb{E}_q \left[ \log \left( \frac{p(\theta, y)}{q(\theta)} \right) \right]$  the *variational lower bound*, which we aim to minimise the negative of. Using the factorisation  $q(\theta) = q_j(\theta_j) q_{\setminus j}(\theta_{\setminus j})$ , the variational lower bound becomes

$$-\mathbb{E}_q \left[ \log \left( \frac{p(\theta, y)}{q(\theta)} \right) \right] = - \int \int q_j(\theta_j) q_{\setminus j}(\theta_{\setminus j}) \log \left( \frac{p(\theta, y)}{q_j(\theta_j) q_{\setminus j}(\theta_{\setminus j})} \right) d\theta_j d\theta_{\setminus j} \quad (6.5.50)$$

$$= \int \int q_j(\theta_j) q_{\setminus j}(\theta_{\setminus j}) (-\log p(\theta, y) + \log q_j(\theta_j) + \log q_{\setminus j}(\theta_{\setminus j})) d\theta_j d\theta_{\setminus j} \quad (6.5.51)$$

$$= - \int q_j(\theta_j) \int q_{\setminus j}(\theta_{\setminus j}) \log p(\theta, y) d\theta_{\setminus j} d\theta_j \quad (6.5.52)$$

$$+ \int q_j(\theta_j) \log q_j(\theta_j) \underbrace{\int q_{\setminus j}(\theta_{\setminus j}) d\theta_{\setminus j}}_1 d\theta_j$$

$$+ \int q_{\setminus j}(\theta_{\setminus j}) \log q_{\setminus j}(\theta_{\setminus j}) \underbrace{\int q_j(\theta_j) d\theta_j}_1 d\theta_{\setminus j}$$

$$= - \int q_j(\theta_j) \int q_{\setminus j}(\theta_{\setminus j}) \log p(\theta, y) d\theta_{\setminus j} d\theta_j + \int q_j(\theta_j) \log q_j(\theta_j) d\theta_j \\ + \int q_{\setminus j}(\theta_{\setminus j}) \log q_{\setminus j}(\theta_{\setminus j}) d\theta_{\setminus j} \quad (6.5.53)$$

$$= - \int q_j(\theta_j) \mathbb{E}_{q_{\setminus j}} [\log p(\theta, y)] d\theta_j + \int q_j(\theta_j) \log q_j(\theta_j) d\theta_j \\ + \int q_{\setminus j}(\theta_{\setminus j}) \log q_{\setminus j}(\theta_{\setminus j}) d\theta_{\setminus j} \quad (6.5.54)$$

where the last term also does not depend on  $q_j(\theta_j)$ , and so can be ignored for the purpose of minimisation. As mentioned above, integrating out  $\theta_{\setminus j}$  causes  $\mathbb{E}_{q_{\setminus j}} [\log p(\theta, y)]$  to only be a function of  $\theta_j$ , so we denote the normalised density from this function by

$$g(\theta_j) \propto \exp \left( \mathbb{E}_{q_{\setminus j}} [\log p(\theta, y)] \right) \quad (6.5.55)$$

The log density will be given as

$$\log g(\theta_j) = \mathbb{E}_{q_{\setminus j}} [\log p(\theta, y)] + c \quad (6.5.56)$$

where  $c$  is some constant. Thus, the variational lower bound becomes

$$-\mathbb{E}_q \left[ \log \left( \frac{p(\theta, y)}{q(\theta)} \right) \right] = - \int q_j(\theta_j) \log g(\theta_j) d\theta_j + \int q_j(\theta_j) \log q_j(\theta_j) d\theta_j + c' \quad (6.5.57)$$

$$= - \int q_j(\theta_j) \log \left( \frac{g(\theta_j)}{q_j(\theta_j)} \right) d\theta_j + c' \quad (6.5.58)$$

where we have absorbed  $c$  and  $\int q_{\setminus j}(\theta_{\setminus j}) \log q_{\setminus j}(\theta_{\setminus j}) d\theta_{\setminus j}$  into the constant  $c'$ . This allows us to see that the KL divergence will be minimised if  $q_j(\theta_j)$  is updated to be equal to  $g(\theta_j)$ , thus our update should be such that

$$q_j^+(\theta_j) = g(\theta_j) \quad (6.5.59)$$

$$\propto \exp \left( \mathbb{E}_{q_{\setminus j}} [\log p(\theta, y)] \right) \quad (6.5.60)$$

$$= \exp \left( \mathbb{E}_{q_{\setminus j}} [\log p(y|\theta) p(\theta)] \right) \quad (6.5.61)$$

### Variational Inference for Exponential Families

Suppose the likelihood and prior are chosen such that they form an exponential family conjugate pair. That is, the likelihood takes the form

$$p(y|\theta) \propto \varphi(\theta)^n \exp \left( \eta(\theta)^\top T(y) \right) \quad (6.5.62)$$

while the prior takes the form

$$p(\theta) \propto \varphi(\theta)^m \exp \left( \eta(\theta)^\top \mathbf{s} \right) \quad (6.5.63)$$

where  $p(\theta)$  is implicitly parameterised by some hyperparameters contained in  $\mathbf{s}$ . Thus

$$p(y|\theta) p(\theta) \propto \varphi(\theta)^{m+n} \exp \left( \eta(\theta)^\top (\mathbf{s} + T(y)) \right) \quad (6.5.64)$$

and taking logs, we have

$$\log p(y|\theta) p(\theta) = \eta(\theta)^\top (\mathbf{s} + T(y)) + (m+n) \log \varphi(\theta) + \gamma \quad (6.5.65)$$

where  $\gamma$  is some constant. Taking expectations with respect to  $q_{\setminus j}(\theta_{\setminus j})$ ,

$$\mathbb{E}_{q_{\setminus j}} [\log p(y|\theta) p(\theta)] = \mathbb{E}_{q_{\setminus j}} \left[ \eta(\theta)^\top (\mathbf{s} + T(y)) + (m+n) \log \varphi(\theta) + \gamma \right] \quad (6.5.66)$$

$$= (\mathbf{s} + T(y))^\top \mathbb{E}_{q_{\setminus j}} [\eta(\theta)] + \mathbb{E}_{q_{\setminus j}} [(m+n) \log \varphi(\theta) + \gamma] \quad (6.5.67)$$

So when variational inference is performed, the updated  $j^{\text{th}}$  approximating factor  $q_j^+(\theta_j)$  will be

$$q_j^+(\theta_j) \propto \exp \left( \mathbb{E}_{q_{\setminus j}} [\log p(y|\theta) p(\theta)] \right) \quad (6.5.68)$$

$$= \exp \left( (\mathbf{s} + T(y))^\top \mathbb{E}_{q_{\setminus j}} [\eta(\theta)] + \mathbb{E}_{q_{\setminus j}} [(m+n) \log \varphi(\theta) + \gamma] \right) \quad (6.5.69)$$

$$= h_j(\theta_j) \exp \left( (\mathbf{s} + T(y))^\top \bar{\eta}_j(\theta_j) \right) \quad (6.5.70)$$

where

$$h_j(\theta_j) = \exp \left( \mathbb{E}_{q_{\setminus j}} [(m+n) \log \varphi(\theta) + \gamma] \right) \quad (6.5.71)$$

$$\bar{\eta}_j(\theta_j) = \mathbb{E}_{q_{\setminus j}} [\eta(\theta)] \quad (6.5.72)$$

Hence the approximating factor  $q_j^+(\theta_j)$  will also be a member of the exponential family, with sufficient statistic  $\bar{\eta}_j(\theta_j)$ . This suggests that the steps in variational inference may be analytically tractable if exponential families are used in the likelihood and prior.

### Expectation Maximisation as Variational Inference [65]

The expectation maximisation algorithm can be treated as a special case of variational inference. Recall in the EM algorithm that we perform inference over the latent variables  $z$  and the parameters  $\theta$ . In the framework of variational inference, we consider these both as parameters, partitioned into  $(z, \theta)$ . Thus the posterior is factorised as

$$p(z, \theta|y) = p(z|y, \theta)p(\theta|y) \quad (6.5.73)$$

where we approximate each of the factors with

$$q_z(z) \approx p(z|y, \theta) \quad (6.5.74)$$

$$q_\theta(\theta) \approx p(\theta|y) \quad (6.5.75)$$

In each iteration of the EM algorithm, we do not constrain the functional form of  $q_z(z)$ , and instead let it be

$$q_z(z) = p(z|y, \hat{\theta}) \quad (6.5.76)$$

where  $\hat{\theta}$  is the current point estimate of  $\theta$ . To update  $q_\theta(\theta)$ , the variational inference approach is to take the expectation

$$\mathbb{E}_{q_z} [\log p(z, \theta, y)] = \int q_z(z) \log p(z, \theta, y) dz \quad (6.5.77)$$

which is the same approach taken as in the expectation step of EM. If we now constrain  $q_\theta(\theta)$  to be a point mass, it is sensible to set the point mass to be at the mode of  $\mathbb{E}_{q_z} [\log p(z, \theta, y)]$ , which is exactly the approach taken in the maximisation step of EM.

#### 6.5.4 Posterior Sampling

If we are able to sample from the posterior  $p(\theta|x)$ , we can generate  $N$  samples  $\theta_1, \dots, \theta_N$  and approximate the posterior using the empirical distribution of the samples:

$$p(\theta|x) \approx \frac{1}{N} \sum_{i=1}^N \delta(\theta - \theta_i) \quad (6.5.78)$$

This is useful if we want to construct a histogram to visualise the shape of the posterior distribution, or if we are particularly interested in the moments of the posterior, from which we can approximate using the sample moments:

$$\widehat{\mathbb{E}}[\theta|x] = \frac{1}{N} \sum_{i=1}^N \theta_i \quad (6.5.79)$$

$$\widehat{\text{Var}}(\theta|x) = \frac{1}{N} \sum_{i=1}^N (\theta_i - \widehat{\mathbb{E}}[\theta|x]) (\theta_i - \widehat{\mathbb{E}}[\theta|x])^\top \quad (6.5.80)$$

As the posterior  $p(\theta|x)$  is proportional to the likelihood  $p(x|\theta)$  multiplied by the prior  $p(\theta)$ :

$$p(\theta|x) \propto p(x|\theta)p(\theta) \quad (6.5.81)$$

this means that if all else fails, we can still use methods such as acceptance-rejection sampling or Markov chain Monte-Carlo to generate samples. This is because we do not necessarily need to know the posterior exactly; we just need to be able to evaluate it up to a normalising constant, which can be done as long as we can evaluate the likelihood and prior.

## 6.6 Bayesian Networks

6.6.1 Factor Graphs

6.6.2 Probabilistic Graphical Models

6.6.3 Structure Learning in Graphical Models

6.6.4 Causality in Graphical Models

## 6.7 Bernstein-von Mises Theorem [199]

## 6.8 Cox's Theorem [13, 100]

## 6.9 Subjective Probability [87, 176]

6.9.1 Dempster-Shafer Theory

# Chapter 7

# Markov Processes

## 7.1 Finite-State Discrete-Time Markov Chains

### 7.1.1 Markov Models

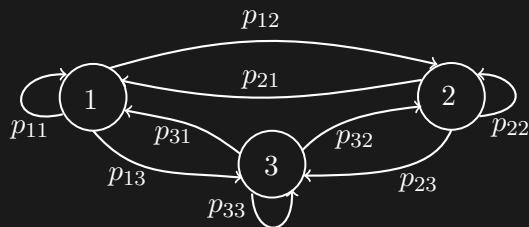
A Markov model can be used to model a specific class of systems or processes that randomly changes over time. We let the state variable  $X_t$  be a variable which captures all possible configurations of the system at any given time. If the *state-space* consists of finitely many  $N$  states, we can enumerate each of the states by

$$X_t \in \{1, \dots, N\} \quad (7.1.1)$$

In a discrete-time Markov model, the state is only allowed to change at discrete time instants. If the state is  $X_t = i$  at time  $t$ , then the probability that the state will transition to  $X_{t+1} = j$  at the next time instant is called the *transition probability*, and is denoted

$$p_{ij} = \Pr(X_{t+1} = j | X_t = i) \quad (7.1.2)$$

Markov models can be depicted using state diagrams.



### Markov Property

The evolution over time of a Markov model is known as a Markov process. All Markov process  $X_t$  satisfy the Markov property, which says that the random variable  $X_{t+1}$  is conditionally independent with  $X_{t-1}, \dots, X_0$ , given  $X_t$ . Formally,

$$\Pr(X_{t+1} = j | X_t = i_t, X_{t-1} = i_{t-1}, \dots, X_0 = i_0) = \Pr(X_{t+1} = j | X_t = i_t) \quad (7.1.3)$$

$$= p_{i_t j} \quad (7.1.4)$$

In other words, the transition probability is only determined by the current state; the previous history of states is irrelevant.

### 7.1.2 Stochastic Matrices

Not to be confused with a *random matrix*, a stochastic matrix is a class of matrix which contains a probability distribution in its rows or columns.

## Right Stochastic Matrices

A right stochastic (or *row-stochastic*) matrix is a square matrix of non-negative elements, such that each row sums to 1. This means that each row is a probability mass function. Hence a right stochastic matrix  $P$  also satisfies the property

$$P\mathbf{1} = \mathbf{1} \quad (7.1.5)$$

where  $\mathbf{1}$  is a column vector of ones. The equation  $P\mathbf{1} = \mathbf{1}$  resembles the equation which defines eigenvalues and eigenvectors, so we can see that  $P$  has an eigenvalue of  $\lambda = 1$ , with corresponding right-eigenvector that can be made to be proportional to  $\mathbf{1}$  (recalling that eigenvectors are arbitrary up to a scaling).

**Lemma 7.1.** *For a right stochastic matrix  $P$ , all the eigenvalues satisfy*

$$|\lambda| \leq 1 \quad (7.1.6)$$

*Proof.* Let the dimension of  $P$  be  $N \times N$ . The equation defining eigenvalues and eigenvectors is  $P\mathbf{v} = \lambda\mathbf{v}$ . The  $j^{\text{th}}$  row of this system of equations can be written as

$$\sum_{i=1}^N p_{ji} v_i = \lambda v_j \quad (7.1.7)$$

Denote the largest entry of  $\mathbf{v}$  by

$$|v_{j^*}| = \max_j |v_j| \quad (7.1.8)$$

We have for any eigenvalue  $\lambda$ :

$$|\lambda| \cdot |v_{j^*}| = |\lambda v_{j^*}| \quad (7.1.9)$$

$$= \left| \sum_{i=1}^N p_{j^*i} v_i \right| \quad (7.1.10)$$

$$\leq \sum_{i=1}^N p_{j^*i} \cdot |v_i| \quad (7.1.11)$$

by the Cauchy-Schwarz inequality, noting that each  $p_{j^*i}$  is non-negative by the property of stochastic matrices. Then using the characterisation of  $|v_{j^*}|$ , this gives:

$$\sum_{i=1}^N p_{j^*i} \cdot |v_i| \leq \sum_{i=1}^N p_{j^*i} \cdot |v_{j^*}| \quad (7.1.12)$$

$$= |v_{j^*}| \quad (7.1.13)$$

since  $\sum_{i=1}^N p_{j^*i} = 1$  for right stochastic matrices. Therefore  $|\lambda| \leq 1$ .  $\square$

Apart from these restrictions, the eigenvalues of stochastic matrices are generally allowed to be complex, zero, or there may even be multiple eigenvalues that are one. Also, since the  $\ell_\infty$  operator norm of  $P$  can be computed by the maximum  $\ell_1$  norm of a row, and we have that all rows sum to one, then

$$\|P\|_\infty = 1 \quad (7.1.14)$$

## Left Stochastic Matrices

A left stochastic (or *column-stochastic*) matrix is a square matrix of non-negative elements, such that each column sums to 1. Hence each column is a probability mass function, and a left stochastic matrix  $P'$  satisfies the property

$$\mathbf{1}^\top P' = \mathbf{1}^\top \quad (7.1.15)$$

Also, if  $P'$  is a left stochastic matrix, then  $(P')^\top$  is a right stochastic matrix. Since the eigenvalues of  $P'$  are the same as the eigenvalues of  $(P')^\top$ , we know that  $\lambda = 1$  will be an eigenvalue of  $P'$ , with corresponding left-eigenvalue (i.e. a row vector  $\mathbf{v}$  that satisfies  $\mathbf{v} = \mathbf{v}P'$ ) that can be made proportional to  $\mathbf{1}^\top$ . Moreover, all the eigenvalues will satisfy  $|\lambda| \leq 1$ . Also, since the  $\ell_1$  operator norm of  $P'$  can be computed by the maximum  $\ell_1$  norm of a column, and we have that all columns sum to one, then

$$\|P'\|_1 = 1 \quad (7.1.16)$$

## Doubly Stochastic Matrices

A doubly stochastic matrix is a square matrix of non-negative elements, such that every row and column sums to 1. Thus a doubly stochastic matrix is a right stochastic matrix as well as a left stochastic matrix, and will inherit the properties of both.

### 7.1.3 Markov Chain Probabilities

#### State Probability Vectors

Let  $X_t$  be a Markov chain on finite state-space  $\mathcal{X} = \{1, \dots, N\}$ . The state probability vector  $\mathbf{p}_t$  is by convention usually a row vector

$$\mathbf{p}_t = [\Pr(X_t = 1) \ \dots \ \Pr(X_t = N)] \quad (7.1.17)$$

containing the probabilities that  $X_t$  will belong to each state at time  $t$ . An alternative (but for all purposes equivalent) convention is to represent  $\mathbf{p}_t$  using a column vector.

#### Transition Matrices

In an  $N$ -state Markov model, there are  $N^2$  transition probabilities, which we can represent using an  $N \times N$  matrix called a transition matrix. If the convention is to use state probability row vectors, then this should be paired with the convention

$$T = \begin{bmatrix} p_{11} & \dots & p_{1N} \\ \vdots & \ddots & \vdots \\ p_{N1} & \dots & p_{NN} \end{bmatrix} \quad (7.1.18)$$

$$= \begin{bmatrix} \Pr(X_{t+1} = 1 | X_t = 1) & \dots & \Pr(X_{t+1} = N | X_t = 1) \\ \vdots & \ddots & \vdots \\ \Pr(X_{t+1} = 1 | X_t = N) & \dots & \Pr(X_{t+1} = N | X_t = N) \end{bmatrix} \quad (7.1.19)$$

Note that each row, as a conditional distribution, sums to one. Hence this transition matrix is a **right stochastic matrix**. The alternative convention, if using state probability column vectors, is to write transition matrices as the transpose of above:

$$T' = \begin{bmatrix} p_{11} & \dots & p_{N1} \\ \vdots & \ddots & \vdots \\ p_{1N} & \dots & p_{NN} \end{bmatrix} \quad (7.1.20)$$

$$= \begin{bmatrix} \Pr(X_{t+1} = 1|X_t = 1) & \dots & \Pr(X_{t+1} = 1|X_t = N) \\ \vdots & \ddots & \vdots \\ \Pr(X_{t+1} = N|X_t = 1) & \dots & \Pr(X_{t+1} = N|X_t = N) \end{bmatrix} \quad (7.1.21)$$

which will be a left stochastic matrix, since each column sums to one.

### Chapman-Kolmogorov Equations

Using the law of total probability, we can write

$$\Pr(X_{t+1} = j) = \sum_{i=1}^N \Pr(X_{t+1} = j|X_t = i) \Pr(X_t = i) \quad (7.1.22)$$

We can implement this for each element of the state probability vector, using matrix multiplication. This yields the Chapman-Kolmogorov equations, which relates successive state probability vectors using the transition matrix:

$$\begin{aligned} & [\Pr(X_{t+1} = 1) \dots \Pr(X_{t+1} = N)] \\ &= [\Pr(X_t = 1) \dots \Pr(X_t = N)] \begin{bmatrix} \Pr(X_{t+1} = 1|X_t = 1) & \dots & \Pr(X_{t+1} = N|X_t = 1) \\ \vdots & \ddots & \vdots \\ \Pr(X_{t+1} = 1|X_t = N) & \dots & \Pr(X_{t+1} = N|X_t = N) \end{bmatrix} \end{aligned} \quad (7.1.23)$$

or compactly,

$$\mathbf{p}_{t+1} = \mathbf{p}_t T \quad (7.1.24)$$

The equivalent convention is to write

$$\mathbf{p}_t^\top = T^\top \mathbf{p}_{t+1}^\top \quad (7.1.25)$$

where  $\mathbf{p}_t^\top$  and  $\mathbf{p}_{t+1}^\top$  are state probability column vectors, and  $T^\top$  is a right stochastic transition matrix. We can use the Chapman-Kolmogorov equations to recursively compute the state probability vector at any time  $n$ , beginning from an initial distribution  $\mathbf{p}_0$ . Immediately,  $\mathbf{p}_1 = \mathbf{p}_0 T$  and generally,

$$\mathbf{p}_n = \mathbf{p}_0 T^n \quad (7.1.26)$$

### Multi-Step Transition Probabilities

The matrix  $T^n$  is called the  $n$ -step transition matrix, because the  $ij^{\text{th}}$  element of  $T^n$  is the  $n$ -step transition probability

$$[T^n]_{ij} = \Pr(X_{t+n} = j|X_t = i) \quad (7.1.27)$$

This can be seen by viewing an  $n$ -step transition as just taking a single ‘large’ step, so the elements of  $T^n$  must contain the respective conditional probabilities for an  $n$ -step transition, in the same way that  $T$  contains the probabilities for a single step.

### Markov Chain Path Probabilities

Consider the probability of a Markov chain taking the path  $i_0, i_1, \dots, i_n$ . Using the Markov property, the probability of this is

$$\Pr(X_0 = i_0, \dots, X_n = i_n) = \Pr(X_n = i_n|X_{n-1} = i_{n-1}) \times \dots \times \Pr(X_1 = i_1|X_0 = i_0) \Pr(X_0 = i_0) \quad (7.1.28)$$

$$= p_{i_{n-1}i_n} \times \dots \times p_{i_0i_1} \mathbf{p}_{0,i_0} \quad (7.1.29)$$

### 7.1.4 Markov Chain Properties

#### Accessible States

State  $j$  is said to be accessible from state  $i$  if it is possible after some number of time periods to reach state  $j$  from state  $i$ . Formally, there must exist some  $n \geq 0$  such that

$$\Pr(X_{t+n} = j | X_t = i) > 0 \quad (7.1.30)$$

This means that every state is accessible from itself, since  $\Pr(X_t = i | X_t = i) = 1$ .

#### Communicating States

States  $i$  and  $j$  are said to communicate if  $j$  is accessible from  $i$ , and  $i$  is also accessible from  $j$ .

#### Communicating Classes

A communicating class is a subset of the states where every state in the class communicates with each other. It is possible to partition a Markov chain into its communicating classes.

#### Irreducible Markov Chains

An irreducible Markov chain is one where all the states form a communicating class. Intuitively, it is possible to reach any given state from any other state within some number of time periods.

#### Recurrent States

State  $i$  is recurrent if for all  $j \neq i$  accessible from  $i$ , then  $i$  is accessible from  $j$ . That is, if we are at state  $i$  and then leave it, it is possible to return back to state  $i$  at some point in the future. Hence a recurrent state will be visited infinitely often, and we can say for a recurrent state  $i$ :

$$\sum_{n=1}^{\infty} \Pr(X_n = i | X_0 = i) = \infty \quad (7.1.31)$$

#### Transient States

A transient state is not recurrent. This means that if  $i$  is transient, then for all  $j \neq i$  accessible from  $i$ , there is at least one  $j$  in which  $i$  is not accessible. Thus a transient state will be visited a finite number of times. Also state  $i$  is transient if and only if

$$\sum_{n=1}^{\infty} \Pr(X_n = i | X_0 = i) < \infty \quad (7.1.32)$$

#### Recurrent Classes

A recurrent class is a set of states in which every state is accessible from one another, and there are no states outside the class that are accessible from within the class. This means every recurrent class is also a communicating class, but not every communicating class will be a recurrent class because once a Markov process is within a recurrent class, it will stay in that class forever. Any Markov chain can be decomposed into:

- At least one recurrent class (e.g. for an irreducible Markov chain, the whole state-space would be a recurrent class).
- Possibly some transient states.

## Mean Recurrence Time

Consider an irreducible Markov chain, which has a single recurrent class. For a particular recurrent state  $s$ , let  $\tau_i$  denote the *mean first passage time*, which is the expected number of steps to reach state  $s$ , beginning from state  $i$ . Trivially,  $\tau_s = 0$ , and for all  $i \neq s$ , we have the system of equations

$$\tau_i = 1 + \sum_{j=1}^N p_{ij} \tau_j \quad (7.1.33)$$

These equations are intuitive because from state  $i$ , we will incur one additional step no matter what, and then the average over the mean first passage times from state  $j$ , weighted by the probability of transitioning to  $j$ . By solving these equations, we can explicitly obtain the mean first passage times  $\tau_1, \dots, \tau_N$ . From the mean first passage times, we can calculate the mean recurrence time of  $s$ , which is the expected number of steps until the state  $s$  is reached again, given the current state is  $s$ . Denoted  $\tilde{\tau}_s$ , and using similar reasoning as for the mean first passage times, this is given by

$$\tilde{\tau}_s = 1 + \sum_{j=1}^N p_{sj} \tau_j \quad (7.1.34)$$

## Periodic States [127]

A state  $i$  has period  $d$  if it can only reoccur at multiples of  $d$ , i.e.

$$\Pr(X_{t+n} = i | X_t = i) = 0 \quad (7.1.35)$$

whenever  $n$  is not a multiple of  $d$ . If there are many  $d$  with this property, then we take the period to be largest. A state is said to be periodic if it has period  $d > 1$ .

## Aperiodic States

A state is said to be aperiodic if it has period  $d = 1$ .

## Periodic Classes [21]

A recurrent class is said to be periodic if we can group the states into  $d > 1$  disjoint subsets  $S_1, \dots, S_d$ , so that all transitions from  $S_k$  will be to  $S_{k+1}$  (or if we are at  $S_d$ , then it will wrap back around to  $S_1$ ). That is, if the current state is from  $S_k$ , then the next state will definitely be from  $S_{k+1}$ , and we will visit  $S_k$  in multiples of  $d$ . Since we require  $d > 1$ , then given any  $n$  and current state  $i$ , there must exist at least one or more states  $j$  such that

$$\Pr(X_{t+n} = j | X_t = i) = 0 \quad (7.1.36)$$

## Aperiodic Classes

A recurrent class which is not periodic is said to be aperiodic. Hence by taking the negation of the requirement for periodic classes, this means there exists some  $n$  such that

$$\Pr(X_{t+n} = j | X_t = i) > 0 \quad (7.1.37)$$

for all  $i, j$  in the class.

## Periodic Markov Chains

An irreducible Markov chain is said to be periodic if all its states have periods greater than one.

## Aperiodic Markov Chains

An irreducible Markov chain is said to be aperiodic if all its states have period  $d = 1$ .

## Regular Markov Chains [74]

A Markov chain with transition matrix  $T$  is said to be regular (or *primitive*) if there is some integer  $n$  where all the elements of  $T^n$  are positive. For instance, any Markov chain where the state diagram is ‘fully connected’ (i.e. it is possible to go from any state to any other state in a single jump) will be regular.

**Theorem 7.1.** *A finite-state Markov chain is regular if and only if it is irreducible and aperiodic.*

*Proof.* By the definition of irreducibility, then for any pair of states  $i, j$  there exists some  $n$  such that

$$\Pr(X_{t+n} = j | X_t = i) > 0 \quad (7.1.38)$$

If the Markov chain were periodic, it might be that there is no  $n$  such that all transition probabilities are simultaneously positive. Being aperiodic however, this guarantees there exists some  $n$  such that the  $n$ -step transition probabilities are positive for all  $i, j$ . Recalling that the  $n$ -step transition probabilities are the elements of  $T^n$ , this matches the definition of a regular Markov chain.  $\square$

### 7.1.5 Absorbing Markov Chains

#### Absorbing States

A recurrent state  $i$  is absorbing if

$$\Pr(X_{t+1} = j | X_t = i) = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \quad (7.1.39)$$

Hence if a Markov process reaches an absorbing state, it will stay there forever.

#### Canonical Form of Absorbing Markov Chains

A Markov chain is absorbing if every state is either an absorbing state or a transient state. As the ordering of the states in the state-space can be arbitrary, we can group all transient states together at the beginning and represent the transition matrix of an absorbing Markov chain by the block matrix

$$T = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix} \quad (7.1.40)$$

where  $Q$  are the probabilities of transitioning from transient states to other transient states, and  $R$  contains the probabilities of transitioning to absorbing states (where the state will remain forever, indicated by the identity matrix).

#### Fundamental Matrix of Absorbing Markov Chains

Consider an absorbing Markov chain with  $m$  transient states, and suppose the initial state is one of the transient states. Let  $\mathbf{M}$  be an  $m \times m$  matrix whereby the  $ij^{\text{th}}$  element is the expected number of times state  $j$  gets visited (because being absorbed eventually), given the initial state

was  $i$ . This matrix is known as the fundamental matrix. To determine this, we can express by definition

$$\mathbf{M} = \begin{bmatrix} \mathbb{E} [\sum_{t=0}^{\infty} \mathbb{I}_{\{X_t=1\}} | X_0 = 1] & \dots & \mathbb{E} [\sum_{t=0}^{\infty} \mathbb{I}_{\{X_t=m\}} | X_0 = 1] \\ \vdots & \ddots & \vdots \\ \mathbb{E} [\sum_{t=0}^{\infty} \mathbb{I}_{\{X_t=1\}} | X_0 = m] & \dots & \mathbb{E} [\sum_{t=0}^{\infty} \mathbb{I}_{\{X_t=m\}} | X_0 = m] \end{bmatrix} \quad (7.1.41)$$

where we have used indicators to count the number of visits. Then

$$\mathbf{M} = \sum_{t=0}^{\infty} \begin{bmatrix} \mathbb{E} [\mathbb{I}_{\{X_t=1\}} | X_0 = 1] & \dots & \mathbb{E} [\mathbb{I}_{\{X_t=m\}} | X_0 = 1] \\ \vdots & \ddots & \vdots \\ \mathbb{E} [\mathbb{I}_{\{X_t=1\}} | X_0 = m] & \dots & \mathbb{E} [\mathbb{I}_{\{X_t=m\}} | X_0 = m] \end{bmatrix} \quad (7.1.42)$$

$$= \sum_{t=0}^{\infty} \begin{bmatrix} \Pr(X_t = 1 | X_0 = 1) & \dots & \Pr(X_t = m | X_0 = 1) \\ \vdots & \ddots & \vdots \\ \Pr(X_t = 1 | X_0 = m) & \dots & \Pr(X_t = m | X_0 = m) \end{bmatrix} \quad (7.1.43)$$

$$= \sum_{t=0}^{\infty} Q^t \quad (7.1.44)$$

from the definition of the multi-step transition matrix. By the characterisation of transient states, this infinite sum converges, and we can perform (analogously to a geometric series):

$$\mathbf{M} = I + Q + Q^2 + \dots \quad (7.1.45)$$

$$= I + Q(I + Q + Q^2 + \dots) \quad (7.1.46)$$

$$= I + Q\mathbf{M} \quad (7.1.47)$$

Thus

$$\mathbf{M} = (I - Q)^{-1} \quad (7.1.48)$$

### Expected Absorption Time

We can also count the total number of visits to all transient states from initial state  $i$ . The expectation of this is called the expected absorption time. A vector of the expected absorption times  $\mu_1, \dots, \mu_m$  can be compactly written as

$$\boldsymbol{\mu} = \mathbf{M}\mathbf{1} \quad (7.1.49)$$

where  $\mathbf{1}$  is a vector of ones and the  $i^{\text{th}}$  element of  $\boldsymbol{\mu}$  is the expected time to absorption given the initial state was  $i$ . Via the form above, we can show

$$\mathbf{M}\mathbf{1} = (I + Q\mathbf{M})\mathbf{1} \quad (7.1.50)$$

$$\boldsymbol{\mu} = \mathbf{1} + Q\boldsymbol{\mu} \quad (7.1.51)$$

which is a system of equations with each row taking the form

$$\mu_i = 1 + \sum_{j=1}^m p_{ij}\mu_j \quad (7.1.52)$$

This can be generalised over the whole state-space as

$$\mu_i = 1 + \sum_{j=1}^N p_{ij}\mu_j \quad (7.1.53)$$

where we take  $\mu_j = 0$  for  $j > m$ , since these are already absorbing states. Note the resemblance of this system of equations to those in computing the mean first passage time and mean recurrence time.

## Absorption Probabilities

For a given absorbing state  $s$ , let  $a_i$  denote the probability of eventually being absorbed by state  $s$  from current state  $i$ . We trivially know that  $a_s = 1$  and  $a_i = 0$  for all absorbing states  $i \neq s$ . For transient states, the absorption probabilities satisfy

$$a_i = \sum_{j=1}^N p_{ij} a_j \quad (7.1.54)$$

which follows from application of the law of total probability and the Markov property.

### 7.1.6 Stationary Distributions

For a finite-state Markov chain with right stochastic transition matrix  $P$ , the stationary distribution is defined as the probability mass function whose row vector representation  $\pi$  satisfies

$$\pi = \pi P \quad (7.1.55)$$

Via the Chapman-Kolmogorov equations, the interpretation of this is that if the state probability vector is  $\pi$  at time  $t$ , then the state probability vector will also be  $\pi$  at time  $t+1$ . In this way, the stationary distribution is sometimes also referred to as the *steady-state*, *invariant*, or *equilibrium* distribution. The ‘stationary’ qualifier is also fitting because if the initial state probability vector is made to be  $\pi$ , then the path probability is

$$\Pr(X_0 = i_0, \dots, X_n = i_n) = p_{i_{n-1}i_n} \times \dots \times p_{i_0i_1} \pi_{i_0} \quad (7.1.56)$$

$$= \Pr(X_t = i_0, \dots, X_{t+n} = i_n) \quad (7.1.57)$$

for any  $t$ , hence the process is **strictly stationary**. From a dynamical systems perspective, we can view the Chapman-Kolmogorov equations as the evolution of the state probability vector through the discrete-time linear difference equation

$$\mathbf{v}_{t+1} = P^\top \mathbf{v}_t \quad (7.1.58)$$

where  $\mathbf{v}$  is a column vector, which we treat as the ‘state’ of the system. Thus a stationary distribution is an equilibrium point of this system.

### Existence of Stationary Distribution for Finite Markov Chains

Any finite-state discrete-time Markov chain has a stationary distribution. Let  $P$  be the right stochastic transition matrix, which we know has a largest eigenvalue of  $\lambda = 1$ . The stationary distribution defined as is the solution to

$$\pi = \pi P \quad (7.1.59)$$

$$\pi \mathbf{1} = 1 \quad (7.1.60)$$

Looking at  $\pi = \pi P$ , we can see that this resembles the definition of left-eigenvalues, when the eigenvalue is  $\lambda = 1$ . As we know there exists an eigenvalue of one, we can be assured that there exists a solution to the above system. Thus to compute the stationary distribution(s), we can:

- Compute the left-eigenvalue(s) of  $P$  corresponding to the eigenvalue(s) one, and then normalise the entries such that  $\pi$  is a valid distribution.
- Alternatively, we can compute the usual right-eigenvalue(s) of the left stochastic  $P^\top$ , and the resulting normalised solution will be  $\pi^\top$ .

Note in contrast that although  $P$  is guaranteed to have a right-eigenvalue of  $\mathbf{1}$  (normalised to  $N^{-1}\mathbf{1}$ ) by definition, this will in general not be the stationary distribution.

Although an eigenvector  $\pi$  is assured, it is not immediately clear that  $\pi$  will actually be a valid distribution (i.e. all its elements are real and non-negative). We provide some intuition for why  $\pi$  will be valid. For simplicity, first consider a Markov chain that consists of a single **recurrent class**, i.e. it is irreducible. By this characterisation, every state will be visited infinitely many times. Consider the long-run proportion of time spent in state  $i$ . Analogous to waiting times in the geometric distribution, we can express this long-run proportion as

$$\pi_i = \frac{1}{\tilde{\tau}_i} \quad (7.1.61)$$

where  $\tilde{\tau}_i$  is the mean recurrence time of state  $i$ . As  $\pi_i$  is characterised as a long-run proportion, this suggests that it is not changing with time. So the vector of  $\pi_i$  should be the stationary distribution, which is definitely valid since the  $\tilde{\tau}_i$  are real and positive. Now consider an arbitrary Markov chain. We know that it can be decomposed into at least one recurrent class, and possibly some transient states. For a transient state  $i$ , the long-run proportion will be

$$\pi_i = 0 \quad (7.1.62)$$

since it will be visited a finite number of times. We can then apply our arguments for a single recurrent class in a modular fashion. Within each class, the long-run proportions will be positive, but with some appropriate normalisation when considered as a mix between all the recurrent classes. We have thus heuristically argued that all the elements of  $\pi$  will be real and non-negative.

Note that a Markov chain may have more than one stationary distribution. For instance, an absorbing Markov chain with multiple absorbing states will have a stationary distribution corresponding to each absorbing state.

## Limiting Distributions

Limiting distributions express the probability of finding a Markov chain in a particular state after a long time. With right stochastic transition matrix  $T$  and initial state probability row vector  $\mathbf{p}_0$ , the limiting distribution is defined as

$$\mathbf{p}_\infty := \lim_{n \rightarrow \infty} \mathbf{p}_0 T^n \quad (7.1.63)$$

The limiting distribution is not identical to the stationary distribution. Rather, limiting distributions are subsets of stationary distributions. That is to say, any limiting distribution will be a stationary distribution, but the stationary distribution may not be the limiting distribution. Moreover, a stationary distribution is only a property of the transition matrix  $T$ , while the limiting distribution is a property of both  $T$  and the initial distribution  $\mathbf{p}_0$ . There are several possibilities for the existence and characterisation of a limiting distribution.

- Given  $\mathbf{p}_0$  and  $T$ , the Markov chain may not have a limiting distribution. We can typically use periodic Markov chains to construct such examples. For instance, consider the period 2 Markov chain with transition matrix

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (7.1.64)$$

which will not have a limiting distribution if the initial state were  $\mathbf{p}_0 = (1, 0)$  or  $\mathbf{p}_0 = (0, 1)$ , because then the state will flip deterministically back and forth between states 1 and 2. However, we can verify that  $\pi = (1/2, 1/2)$  is a stationary distribution of  $T$ . Thus the Markov chain will have a limiting distribution in the special case  $\mathbf{p}_0 = \pi$ .

- Given  $T$ , the Markov chain may have a limiting distribution, but the limiting distribution will generally depend on  $\mathbf{p}_0$ . For example, in an absorbing Markov chain, the initial state may affect the probabilities of being absorbed into the various absorbing states.
- Given  $T$ , the Markov chain has a limiting distribution, and the limiting distribution does not depend on  $\mathbf{p}_0$ . For these sorts of chains, we can imagine that the transitions are sufficiently ‘chaotic’ so that it does not matter what happened initially, after a long amount of time has passed. As the limiting distribution is also a stationary distribution in this case, we can say that the process is asymptotically stationary.

Given  $\mathbf{p}_0$  and  $T$  where the limiting distribution exists, we can compute  $\mathbf{p}_\infty$  as follows, provided that  $T$  is diagonalisable. For diagonalisable  $T$ , we can perform eigendecomposition by writing

$$T = VDV^{-1} \quad (7.1.65)$$

where  $D$  is a diagonal matrix of the eigenvalues. Then note that for any power  $n$ :

$$T^n = (VDV^{-1})^n \quad (7.1.66)$$

$$= \left( VDV^{-1} \right) \left( VDV^{-1} \right) \dots \left( VDV^{-1} \right) \quad (7.1.67)$$

$$= VD^nV^{-1} \quad (7.1.68)$$

Hence the limiting distribution is

$$\mathbf{p}_\infty = \lim_{n \rightarrow \infty} \mathbf{p}_0 T^n \quad (7.1.69)$$

$$= \mathbf{p}_0 V \left( \lim_{n \rightarrow \infty} D^n \right) V^{-1} \quad (7.1.70)$$

Since all eigenvalues  $|\lambda| \leq 1$  for  $T$  and with  $D$  being diagonal, then  $D_\infty := (\lim_{n \rightarrow \infty} D^n)$  should be simple to compute, as any diagonal element not corresponding to  $\lambda = 1$  should (i.e. with the exception of complex  $\lambda$  with  $|\lambda| = 1$ ) vanish in the limit. Assuming that all eigenvalues with modulus one are real, then using floor notation, we can take

$$D_\infty = \text{diag} \{ \lfloor |\lambda_1| \rfloor, \dots, \lfloor |\lambda_N| \rfloor \} \quad (7.1.71)$$

and get the limiting distribution from

$$\mathbf{p}_\infty = \mathbf{p}_0 V D_\infty V^{-1} \quad (7.1.72)$$

If the above approach fails (e.g. when  $T$  is not diagonalisable), we can always approximate  $\mathbf{p}_\infty$  by

$$\mathbf{p}_\infty \approx \mathbf{p}_0 T^n \quad (7.1.73)$$

where  $n$  is taken to be a large integer.

### Perron-Frobenius Theory

When applied to finite-state Markov chains, the Perron-Frobenius theorem formalises the intuition that a single recurrent class has a valid stationary distribution. The implication of the theorem comprises several statements pertaining to regular (i.e. aperiodic and irreducible) Markov chains. Among them, it ensures that a regular Markov chains will have ‘well-behaved’ properties in that:

- The stationary distribution is unique.
- The limiting distribution always coincides with the stationary distribution.

To show the existence of a valid stationary distribution, we only need irreducibility.

**Theorem 7.2.** *Let  $P$  be the right stochastic matrix of an irreducible Markov chain. Then it has a valid stationary distribution  $\pi$  with all positive elements.*

*Proof.* The logically equivalent contrapositive (which we aim to show) is that if  $\pi$  is a left-eigenvector corresponding to the 1-eigenvalue with at least one non-positive element, then the chain is reducible. By reordering the states and writing

$$\pi = [\pi_+ \ \pi_-] \quad (7.1.74)$$

where  $\pi_+ > \mathbf{0}$  and  $\pi_- \leq \mathbf{0}$ , the goal is to show that the transition matrix can be partitioned into

$$P = \begin{bmatrix} P_+ & \mathbf{0} \\ \mathbf{0} & P_- \end{bmatrix} \quad (7.1.75)$$

This would mean there is at least one recurrent class, which implies the chain is reducible. We can formulate this approach as a linear program. Suppose  $\pi_+$  is of length  $m$  and  $\pi_-$  is of length  $N - m$ , where  $1 \leq m \leq N - 1$  (i.e. we allow for at least one non-positive component). Denote the sets  $\mathcal{I}_+ = \{1, \dots, m\}$  and  $\mathcal{I}_- = \{m + 1, \dots, N\}$ , which we will use to index the elements of  $P_+$  and  $P_-$  respectively. Then consider the linear program

$$\begin{aligned} \min_P \quad & \sum_{i,j \in \mathcal{I}_+} p_{ij} + \sum_{i,j \in \mathcal{I}_-} p_{ij} \\ \text{s.t.} \quad & p_{ij} \geq 0, \quad i, j = 1, \dots, N \\ & \sum_{j=1}^N p_{ij} = 1, \quad i = 1, \dots, N \\ & \sum_{i=1}^N \pi_i p_{ij} = \pi_j, \quad j = 1, \dots, N \end{aligned} \quad (7.1.76)$$

where the elements of right stochastic  $P$  must be subject to the constraints  $p_{ij} \geq 0$  and  $\sum_{j=1}^N p_{ij} = 1$ , and the constraint  $\sum_{i=1}^N \pi_i p_{ij} = \pi_j$  represents each entry of the left-eigenvector relationship  $\pi = \pi P$ . If  $P$  can indeed be partitioned as desired, then it suffices to show that the minimum objective of this linear program is  $N$  (because the sum of elements in any stochastic matrix is  $N$ , so it means all the elements in the off-diagonal blocks must be zero). Letting  $\mathbf{z} = \text{vec}(P^\top)$  denote our decision variable, this linear program has the more compact form

$$\begin{aligned} \min_{\mathbf{z}} \quad & \mathbf{c}^\top \mathbf{z} \\ \text{s.t.} \quad & A\mathbf{z} = \mathbf{b} \\ & \mathbf{z} \geq \mathbf{0} \end{aligned} \quad (7.1.77)$$

where  $\mathbf{c}$  is an appropriately structured matrix of zeros and ones, while

$$\underbrace{\left[ \begin{array}{ccc|c|ccc} 1 & \dots & 1 & \cdots & 1 & \dots & 1 \\ \hline \pi_1 & \dots & 0 & | & \pi_N & \dots & 0 \\ \vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots \\ 0 & \dots & \pi_1 & | & 0 & \dots & \pi_N \end{array} \right]}_A = \underbrace{\begin{bmatrix} p_{11} \\ \vdots \\ p_{1N} \\ \vdots \\ p_{N1} \\ \vdots \\ p_{NN} \end{bmatrix}}_{\mathbf{z}} = \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \\ \pi_1 \\ \vdots \\ \pi_N \end{bmatrix}}_b \quad (7.1.78)$$

Our linear program has a convenient Lagrangian dual form which can be written as

$$\begin{aligned} \max_{\mathbf{y}} \quad & \mathbf{b}^\top \mathbf{y} \\ \text{s.t.} \quad & A^\top \mathbf{y} \leq \mathbf{c} \end{aligned} \tag{7.1.79}$$

where  $\mathbf{y}$  is a vector of Lagrange multipliers. The structure of  $A^\top \mathbf{y} \leq \mathbf{c}$  takes on the form

$$\underbrace{\left[ \begin{array}{c|ccc} 1 & \pi_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & \pi_1 \\ \hline \ddots & & \vdots & \\ \hline 1 & \pi_N & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & \pi_N \end{array} \right]}_{\mathbf{A}^\top} \underbrace{\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}}_{\mathbf{y}} \leq \mathbf{c} \tag{7.1.80}$$

so we summarise that for all  $i \in \{1, \dots, N\}$  and  $j \in \{1, \dots, N\}$  that

$$\lambda_i + \pi_i \mu_j \leq \begin{cases} 1, & i, j \in \mathcal{I}_+ \\ 1, & i, j \in \mathcal{I}_- \\ 0, & \text{otherwise} \end{cases} \tag{7.1.81}$$

From this, we conclude in the dual objective that

$$\mathbf{b}^\top \mathbf{y} = [1 \ \dots \ 1 \ \pi_1 \ \dots \ \pi_N] \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \mu_1 \\ \vdots \\ \mu_N \end{bmatrix} \tag{7.1.82}$$

$$= \sum_{i=1}^N (\lambda_i + \pi_i \mu_i) \tag{7.1.83}$$

$$\leq N \tag{7.1.84}$$

From strong duality of linear programs (which says that the primal solution is equal to the dual solution), this shows that

$$\mathbf{c}^\top \mathbf{z} = \sum_{i,j \in \mathcal{I}_+} p_{ij} + \sum_{i,j \in \mathcal{I}_-} p_{ij} \tag{7.1.85}$$

$$\geq N \tag{7.1.86}$$

as desired.  $\square$

Moreover, if the Markov chain is aperiodic in addition to irreducible (meaning the chain is regular), then this stationary distribution is unique [177]. To develop this, we use the following preliminary results.

**Lemma 7.2.** *If  $P$  is a right stochastic matrix, then*

$$\sup_{\|\mathbf{v}\|_1=1} \left\{ \min_j \left\{ \frac{\sum_{i=1}^N p_{ij} \cdot v_i}{v_j} \right\} \right\} = 1 \tag{7.1.87}$$

*Proof.* We first find a lower bound for the supremum by exhibiting the stationary distribution  $\pi$ , which satisfies  $\sum_{i=1}^N p_{ij} \cdot \pi_i = \pi_j$  for all  $j$ . Thus

$$\frac{\sum_{i=1}^N p_{ij} \cdot \pi_i}{\pi_j} = 1 \quad (7.1.88)$$

for all  $j$ , and we have

$$\min_j \left\{ \frac{\sum_{i=1}^N p_{ij} \cdot \pi_i}{\pi_j} \right\} = 1 \quad (7.1.89)$$

hence

$$\sup_{\|\mathbf{v}\|_1=1} \left\{ \min_j \left\{ \frac{\sum_{i=1}^N p_{ij} \cdot v_i}{v_j} \right\} \right\} \geq 1 \quad (7.1.90)$$

To find an upper bound, we use the characterisation of the  $\ell_\infty$  norm of a right stochastic matrix:

$$\|P\|_\infty = \sup_{\mathbf{v} \neq \mathbf{0}} \left\{ \frac{\|\mathbf{v}P\|_\infty}{\|\mathbf{v}\|_\infty} \right\} \quad (7.1.91)$$

$$= \sup_{\mathbf{v} \neq \mathbf{0}} \left\{ \frac{\max_j \sum_{i=1}^N p_{ij} \cdot v_i}{\max_j v_j} \right\} \quad (7.1.92)$$

$$= 1 \quad (7.1.93)$$

From this, we derive the upper bound

$$1 = \sup_{\mathbf{v} \neq \mathbf{0}} \left\{ \frac{\max_j \sum_{i=1}^N p_{ij} \cdot v_i}{\max_j v_j} \right\} \quad (7.1.94)$$

$$\geq \sup_{\mathbf{v} \neq \mathbf{0}} \left\{ \min_j \left\{ \frac{\sum_{i=1}^N p_{ij} \cdot v_i}{v_j} \right\} \right\} \quad (7.1.95)$$

$$\geq \sup_{\|\mathbf{v}\|_1=1} \left\{ \min_j \left\{ \frac{\sum_{i=1}^N p_{ij} \cdot v_i}{v_j} \right\} \right\} \quad (7.1.96)$$

The proof is completed by noticing that the lower bound and upper bound are both equal to one.  $\square$

**Lemma 7.3.** *Let  $\mathbf{v}$  be a (possibly complex) left-eigenvector corresponding to  $|\lambda| = 1$  of the right stochastic  $P$  for a regular Markov chain. Then the element-wise modulus of  $\mathbf{v}$ , denoted  $\mathbf{v}_+$ , is also a left-eigenvector, i.e.*

$$\sum_{i=1}^N p_{ij} \cdot |v_i| = |v_j| \quad (7.1.97)$$

for all  $j$ .

*Proof.* Note that the result is trivial for the stationary distribution  $\pi$  corresponding to  $\lambda = 1$ . For possibly complex  $\mathbf{v}$  and  $\lambda$ , begin from

$$|\lambda| |v_j| = |\lambda v_j| \quad (7.1.98)$$

$$= \left| \sum_{i=1}^N p_{ij} \cdot v_i \right| \quad (7.1.99)$$

$$\leq \sum_{i=1}^N p_{ij} \cdot |v_i| \quad (7.1.100)$$

for all  $j$ , using the Cauchy-Schwarz inequality. Putting  $|\lambda| = 1$ , this yields

$$|v_j| \leq \sum_{i=1}^N p_{ij} \cdot |v_i| \quad (7.1.101)$$

We show via contradiction that this inequality is always satisfied with equality. Suppose there is at least one  $j$  such that  $|v_j| < \sum_{i=1}^N p_{ij} \cdot |v_i|$ . In vector form, write this as

$$\mathbf{z} := \mathbf{v}_+ P - \mathbf{v}_+ \quad (7.1.102)$$

$$\geq \mathbf{0} \quad (7.1.103)$$

In the case that at least one  $|v_j| < \sum_{i=1}^N p_{ij} \cdot |v_i|$ , this means  $\mathbf{z}$  has some strictly positive elements. With  $P$  being regular, then  $P^n > \mathbf{0}$  for some  $n$ , so it follows that  $\mathbf{z}P^n > \mathbf{0}$ . We then have

$$\mathbf{v}_+ P \cdot P^n - \mathbf{v}_+ P^n = (\mathbf{v}_+ P^n) P - \mathbf{v}_+ P^n \quad (7.1.104)$$

$$> \mathbf{0} \quad (7.1.105)$$

so

$$\frac{\sum_{i=1}^N p_{ij} \cdot [\mathbf{v}_+ P^n]_i}{[\mathbf{v}_+ P^n]_j} > 1 \quad (7.1.106)$$

for each  $j$ . But this leads to a contradiction, as we have found a vector  $\mathbf{v}_+ P^n$  which contradicts the supremum

$$\sup_{\|\mathbf{v}\|_1=1} \left\{ \min_j \left\{ \frac{\sum_{i=1}^N p_{ij} \cdot v_i}{v_j} \right\} \right\} = 1 \quad (7.1.107)$$

found earlier.  $\square$

**Theorem 7.3.** *For a regular Markov chain with right stochastic matrix  $P$ , the eigenvalues satisfy  $|\lambda| < 1$  for all  $\lambda \neq 1$ .*

*Proof.* We can show the property that if  $|\lambda| = 1$ , then it must be real. Let  $\mathbf{v}$  be a left-eigenvector corresponding to  $|\lambda| = 1$ . From above, we know that if  $|\lambda| = 1$ , then

$$|v_j| = \left| \sum_{i=1}^N p_{ij} \cdot v_i \right| \quad (7.1.108)$$

$$= \sum_{i=1}^N p_{ij} \cdot |v_i| \quad (7.1.109)$$

As  $P$  is regular, then for some  $n$  such that  $P^n > \mathbf{0}$ , we can start from  $\mathbf{v}P^n = \lambda^n \mathbf{v}$ , put  $|\lambda| = 1$  to get

$$|\lambda|^n |v_j| = |v_j| \quad (7.1.110)$$

$$= \left| \sum_{i=1}^N [P^n]_{ij} \cdot v_i \right| \quad (7.1.111)$$

$$\leq \sum_{i=1}^N [P^n]_{ij} \cdot |v_i| \quad (7.1.112)$$

and then apply essentially the same reasoning to show

$$\left| \sum_{i=1}^N [P^n]_{ij} \cdot v_i \right| = \sum_{i=1}^N [P^n]_{ij} \cdot |v_i| \quad (7.1.113)$$

Now note that the modulus of a sum of non-zero complex numbers is equal to the sum of the modulus if and only if their angle is the same. We know that each element  $[P^n]_{ij} > 0$  and  $\mathbf{v}_+ > \mathbf{0}$  as shown earlier, hence  $v_1, \dots, v_N$  will all have the same angle in the complex plane. For each row in  $\lambda \mathbf{v} = \mathbf{v}P$ , if we express

$$\lambda v_j = \sum_{i=1}^N p_{ij} \cdot v_i \quad (7.1.114)$$

in complex exponential form, we can cancel out exponentials until we are left with

$$|\lambda| \cdot \angle(\lambda) \cdot |v_j| = \sum_{i=1}^N p_{ij} \cdot |v_i| \quad (7.1.115)$$

where  $\angle(\lambda)$  denotes the angle of  $\lambda$ , as a complex exponential on the unit disc in the complex plane. Comparing this relation to  $|v_j| = \sum_{i=1}^N p_{ij} \cdot |v_i|$ , it must be that  $\lambda$  is real, i.e.  $\lambda = 1$ .  $\square$

**Theorem 7.4.** *For a regular Markov chain with right stochastic  $P$ , the stationary distribution  $\pi$  is unique.*

*Proof.* We can show that the eigenvector corresponding to  $\lambda = 1$  is unique, up to a scaling, which also implies that there is only one eigenvalue  $\lambda = 1$  (we can also say that this eigenvalue is *simple*, as it has *arithmetic multiplicity* of one). We do this via a contradiction argument. Let  $\pi$  be the left-eigenvector that can be chosen with strictly positive elements, corresponding to  $\lambda = 1$ . Suppose  $\mathbf{v}$  is another (possibly complex) left-eigenvector also corresponding to  $\lambda = 1$  that is linearly independent with  $\pi$ . As we have shown above, then the vector obtained from taking the element-wise modulus

$$\mathbf{v}_+ := [|v_1| \ \dots \ |v_N|] \quad (7.1.116)$$

also satisfies the left-eigenvector equation. We already see that  $\mathbf{v}_+ \geq \mathbf{0}$ , but because  $P$  is also regular, this implies that

$$\mathbf{v}_+ P^n = \mathbf{v}_+ \quad (7.1.117)$$

$$> \mathbf{0} \quad (7.1.118)$$

for some  $n$ , since  $P^n$  has all elements strictly positive. Now consider the family of vectors defined by

$$\eta = \pi - c\mathbf{v} \quad (7.1.119)$$

for some possibly complex scalar  $c$ . We are assured that  $\eta \neq \mathbf{0}$  since the vectors are linearly independent, so if such a family were to exist, then every member would also be a left-eigenvector corresponding to  $\lambda = 1$  because

$$\eta P = (\pi - c\mathbf{v}) P \quad (7.1.120)$$

$$= \pi - c\mathbf{v} \quad (7.1.121)$$

$$= \eta \quad (7.1.122)$$

Consider a choice of  $c$  that makes at least one entry of  $\eta$  equal to zero. We have the property that  $\eta_+ \geq \mathbf{0}$ , but we also know that  $\eta_+ \geq \mathbf{0}$  for a regular  $P$  implies that  $\eta_+ > \mathbf{0}$ . This leads to a contradiction that at least one entry of  $\eta$  is zero, which means such a family of eigenvectors cannot exist. We conclude that there cannot be two linearly independent left-eigenvectors associated with  $\lambda = 1$ , and that any two left-eigenvectors must be linearly dependent (i.e. same up to a scaling). Now consider the scaling for  $\pi$  which makes all the elements sum to one. This is the unique stationary distribution for the Markov chain.  $\square$

**Theorem 7.5.** For a regular Markov chain, the stationary distribution is identical to the limiting distribution.

*Proof.* Let  $P$  be the right stochastic matrix. For simplicity, consider the case where  $P$  is diagonalisable (but generally,  $P$  may not be diagonalisable). Then  $P$  may be written as

$$P = VDV^{-1} \quad (7.1.123)$$

where  $V$  contains the right-eigenvectors of  $P$  as its columns, and  $V^{-1}$  contains the left-eigenvectors of  $P$  as its rows. As  $P$  is right-stochastic, we know that  $\mathbf{1}$  is a right-eigenvector corresponding to the unique eigenvalue  $\lambda = 1$ , and as  $P$  is regular, we know that the unique stationary distribution  $\pi$  is the left-eigenvector corresponding to  $\lambda = 1$ . Also, all other eigenvalues  $\lambda \neq 1$  have  $|\lambda| < 1$ , so following the approach to compute limiting distributions, we may write

$$\bar{P} := \lim_{n \rightarrow \infty} P^n \quad (7.1.124)$$

$$= V \left( \lim_{n \rightarrow \infty} D^n \right) V^{-1} \quad (7.1.125)$$

$$= \begin{bmatrix} 1 & \dots \\ \vdots & \dots \\ 1 & \dots \end{bmatrix} \text{diag}\{1, 0, \dots, 0\} \begin{bmatrix} \pi_1 & \dots & \pi_N \\ \vdots & \vdots & \vdots \end{bmatrix} \quad (7.1.126)$$

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \pi_1 & \dots & \pi_N \\ \vdots & \vdots & \vdots \end{bmatrix} \quad (7.1.127)$$

$$= \begin{bmatrix} \pi_1 & \dots & \pi_N \\ \vdots & \vdots & \vdots \\ \pi_1 & \dots & \pi_N \end{bmatrix} \quad (7.1.128)$$

In more compact notation, we can rewrite this as

$$\bar{P} = [\mathbf{1} \ \mathbf{0}] \begin{bmatrix} \pi \\ \vdots \end{bmatrix} \quad (7.1.129)$$

$$= \mathbf{1}\pi \quad (7.1.130)$$

Let  $\mathbf{p}_0$  be any arbitrary initial distribution, then

$$\mathbf{p}_\infty := \mathbf{p}_0 \bar{P} \quad (7.1.131)$$

$$= \mathbf{p}_0 \mathbf{1}\pi \quad (7.1.132)$$

$$= \pi \quad (7.1.133)$$

since  $\mathbf{p}_0 \mathbf{1} = 1$  always. Thus, the limiting distribution coincides with the stationary distribution. For the case where  $P$  is not diagonalisable, we can still validate that  $\lim_{n \rightarrow \infty} P^n = \mathbf{1}\pi$  as  $\bar{P}$  must satisfy idempotence, i.e.  $\bar{P}^2 = \bar{P}$ . This is demonstrated by

$$\bar{P}^2 = \mathbf{1}\pi \mathbf{1}\pi \quad (7.1.134)$$

$$= \mathbf{1}(\pi \mathbf{1})\pi \quad (7.1.135)$$

$$= \mathbf{1}\pi \quad (7.1.136)$$

$$= \bar{P} \quad (7.1.137)$$

□

## Mixing Times

### 7.1.7 Reversible Markov Chains [170]

#### Reverse Markov Property

Note that the Markov property also implies a ‘reverse’ Markov property, i.e.

$$\Pr(X_{t-1} = j | X_t = i, X_{t+1} = i_1, \dots) = \Pr(X_{t-1} = j | X_t = i) \quad (7.1.138)$$

To see why, recognise that independence is a two-way relation, i.e. if event  $E_1$  is independent with  $E_2$ , then  $E_2$  is independent with  $E_1$ . This also extends to conditional independence. Thus the following two statements are equivalent:

- $(X_{t+1}, X_{t+2}, \dots)$  is conditionally independent with  $X_{t-1}$  given  $X_t$ .
- $X_{t-1}$  is conditionally independent with  $(X_{t+1}, X_{t+2}, \dots)$  given  $X_t$ .

If we were to represent the first statement with

$$\Pr(X_{t+1} = i_1, X_{t+2} = i_2 | X_t = i, X_{t-1} = j) = \Pr(X_{t+1} = i_1, X_{t+2} = i_2 | X_t = i) \quad (7.1.139)$$

which occurs as a consequence of the Markov property, then the second statement would be represented with

$$\Pr(X_{t-1} = j | X_t = i, X_{t+1} = i_1, X_{t+2} = i_2, \dots) = \Pr(X_{t-1} = j | X_t = i) \quad (7.1.140)$$

which is the reverse Markov property. Hence a Markov process played ‘in reverse’ is another Markov process.

#### Time-Reversible Markov Chains

Consider a stationary regular Markov chain. By stationary, we mean that the process has been active for a ‘long time’, so that the initial distribution is the unique stationary distribution  $\pi$ , and moreover the distribution at any time is also  $\pi$ . Denote the transition probabilities of the reversed process by

$$q_{ij} = \Pr(X_{t-1} = j | X_t = i) \quad (7.1.141)$$

In terms of the forward transition probabilities, this is shown to be

$$q_{ij} = \frac{\Pr(X_{t-1} = j, X_t = i)}{\Pr(X_t = i)} \quad (7.1.142)$$

$$= \frac{\Pr(X_t = i | X_{t-1} = j) \Pr(X_{t-1} = j)}{\Pr(X_t = i)} \quad (7.1.143)$$

$$= \frac{\pi_j p_{ji}}{\pi_i} \quad (7.1.144)$$

If the reversed Markov chain has the same transition probabilities as the forward chain, i.e.  $q_{ij} = p_{ij}$  for all  $i, j$ , then the process is time-reversible, and we call the chain a time-reversible Markov chain. Substituting the relation above, we see that a time-reversible Markov chain must satisfy

$$p_{ij} = \frac{\pi_j p_{ji}}{\pi_i} \quad (7.1.145)$$

$$\pi_i p_{ij} = \pi_j p_{ji} \quad (7.1.146)$$

for all  $i, j$ . These are known as the *detailed balance equations*. Reinterpreting these equations, we have

$$\Pr(X_{t+1} = j | X_t = i) \Pr(X_t = i) = \Pr(X_{t+1} = i | X_t = j) \Pr(X_t = j) \quad (7.1.147)$$

$$\Pr(X_t = i, X_{t+1} = j) = \Pr(X_t = j, X_{t+1} = i) \quad (7.1.148)$$

That is, observing the consecutive pair  $(i, j)$  is as likely as observing the consecutive pair  $(j, i)$ , for all pairs of states.

### 7.1.8 Strong Markov Property [149]

### 7.1.9 Maximum Likelihood Estimation of Markov Chains

## 7.2 Infinite-State Discrete-Time Markov Chains

### 7.2.1 Countable-State Discrete-Time Markov Chains

We can extend the state-space of Markov models to be countably infinite, such as  $\mathcal{X} = \{\dots, -1, 0, 1, \dots\}$  or  $\mathcal{X} = \{1, 2, \dots\}^d$ . In this case, the transition matrix would require a stochastic matrix of infinite size to represent, however taking  $\mathcal{X} = \{1, 2, \dots\}$  for simplicity, the transition probabilities still satisfy

$$\sum_{j=1}^{\infty} \Pr(X_{t+1} = j | X_t = i) = 1 \quad (7.2.1)$$

for all  $i$ . The Chapman-Kolmogorov equations for a countable state Markov chain are

$$\Pr(X_{t+1} = j) = \sum_{i=1}^{\infty} \Pr(X_{t+1} = j | X_t = i) \Pr(X_t = i) \quad (7.2.2)$$

### Null Recurrent States

If the mean recurrence time of a state  $s$  satisfies  $\tilde{\tau}_s < \infty$ , then state  $s$  is said to be *positive recurrent*. Otherwise if  $\tilde{\tau}_s = \infty$ , then state  $s$  is said to be null recurrent. In finite-state Markov chains, every recurrent state is positive recurrent, but in infinite-state chains, a recurrent state can be null recurrent.

### Stationary Distributions with Countable-State

Suppose the state-space is  $\mathcal{X} = \{1, 2, \dots\}$ . Then the stationary distribution represented by the infinite dimensional vector  $\pi$  is any vector which satisfies

$$\pi_j = \sum_{i=1}^{\infty} \Pr(X_{t+1} = j | X_t = i) \pi_i \quad (7.2.3)$$

and  $\pi_j \geq 0$  for each  $j = 1, 2, \dots$ , and of course  $\sum_{i=1}^{\infty} \pi_i = 1$ . A Markov chain with countably infinite state-space need not have a stationary distribution, unlike the case of finite state-space. For instance, consider a chain on state-space  $\{1, 2, \dots\}$  with transition probabilities

$$\Pr(X_{t+1} = i + 1 | X_t = i) = 1 \quad (7.2.4)$$

for all  $i$ . Then a stationary distribution does not exist, because the state always increments each transition.

### 7.2.2 Uncountable-State Discrete-Time Markov Chains [114]

For a discrete-time Markov model, we can consider the state  $x_t$  as belonging to an uncountable state-space  $\mathcal{X}$ , such as  $\mathbb{R}^d$ . To do this, we now consider the probability density of  $x_t$ . This generalises countable-state Markov chains, because we can represent probability mass functions using Dirac delta mixtures. Denote the initial density by  $p(x_0)$ , which satisfies

$$\Pr(x_0 \in S) = \int_S p(x_0) dx_0 \quad (7.2.5)$$

for every  $S \subseteq \mathcal{X}$ . The Markov property for uncountable state-space can now be expressed as

$$p(x_{t+1}|x_t, \dots, x_0) = p(x_{t+1}|x_t) \quad (7.2.6)$$

and

$$\Pr(x_{t+1} \in S|x_t, \dots, x_0) = \Pr(x_{t+1} \in S|x_t) \quad (7.2.7)$$

#### Transition Densities

The probabilistic evolution of the state is governed by the transition density  $p(x_{t+1}|x_t)$  which satisfies

$$\Pr(x_t \in S|x_t) = \int_S p(x_{t+1}|x_t) dx_{t+1} \quad (7.2.8)$$

We also call the function

$$\kappa(S, x) = \Pr(x_t \in S|x_t = x) \quad (7.2.9)$$

the *transition kernel*. The analogous Chapman-Kolmogorov equation using transition densities is performed using the chain rule of probability and marginalisation:

$$p(x_{t+1}) = \int_{\mathcal{X}} p(x_{t+1}, x_t) dx_t \quad (7.2.10)$$

$$= \int_{\mathcal{X}} p(x_{t+n}|x_t) p(x_t) dx_t \quad (7.2.11)$$

The  $n$ -step transition density  $p(x_{t+n}|x_t)$  can be obtained via applying the chain rule of probability, the Markov property, and marginalisation over  $x_{t+n-1}, \dots, x_{t+1}$ .

$$p(x_{t+n}|x_t) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} p(x_{t+n}, \dots, x_{t+1}|x_t) dx_{t+n-1} \dots dx_{t+1} \quad (7.2.12)$$

$$= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} p(x_{t+n}|x_{t+n-1}, \dots, x_t) \dots p(x_{t+1}|x_t) dx_{t+n-1} \dots dx_{t+1} \quad (7.2.13)$$

$$= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} p(x_{t+n}|x_{t+n-1}) \dots p(x_{t+1}|x_t) dx_{t+n-1} \dots dx_{t+1} \quad (7.2.14)$$

From this, the  $n$ -step transition kernel can be expressed as

$$\kappa^n(S, x) = \Pr(x_{t+n} \in S|x_t = x) \quad (7.2.15)$$

$$= \int_S p(x_{t+n}|x_t) dx_{t+n} \quad (7.2.16)$$

#### Gaussian Markov Processes

A Gaussian Markov process (not be confused with the Gauss-Markov theorem) is a stochastic process which satisfies the Markov property as well as the requirements of a Gaussian process.

## Harris Recurrence [153]

### 7.3 Continuous-Time Markov Processes [3]

#### 7.3.1 Countable-State Continuous-Time Markov Processes

#### 7.3.2 Uncountable-State Continuous-Time Markov Processes

### 7.4 Time-Inhomogeneous Markov Chains

### 7.5 Hidden Markov Models

#### 7.5.1 Discrete-Time Hidden Markov Models

##### Finite-State Discrete-Time Hidden Markov Models [190]

A hidden Markov model (HMM) consists of two processes,  $(X_t, Y_t)$ . The process  $X_t$  is the state process, follows the Markov property, and is treated as the ‘hidden’ (or *latent*) variable. The process  $Y_t$  is the observation process which has been ‘corrupted’ from the state process, and is observed in lieu of the state. Hence a hidden Markov model is suitable for modelling processes where the true state is not directly known/observed. Suppose there are a finite number  $N$  states and also a finite number  $M$  observation symbols

$$Y_t \in \{1, \dots, M\} \quad (7.5.1)$$

Then the HMM is specified by the usual  $N \times N$  transition matrix  $P$  such that

$$\mathbf{p}_{t+1} = \mathbf{p}_t P \quad (7.5.2)$$

where  $\mathbf{p}_k$  is the state probability vector for  $X_k$ . In addition, we introduce the  $N \times M$  matrix  $C$  consisting of the elements

$$C_{ij} = \Pr(Y_t = j | X_t = i) \quad (7.5.3)$$

These are known as the observation (or *emission*) probabilities, and a property is that the observation  $Y_k$  is conditionally independent to all other observations and hidden variables, given  $X_t$ . Let  $\mathbf{q}_t$  denote the observation probability vector:

$$\mathbf{q}_t = [\Pr(Y_t = 1) \ \dots \ \Pr(Y_t = M)] \quad (7.5.4)$$

Then  $\mathbf{q}_t$  is given by

$$\mathbf{q}_t = \left[ \sum_{i=1}^N \Pr(X_t = i) \Pr(Y_t = 1 | X_t = i) \ \dots \ \sum_{i=1}^N \Pr(X_t = i) \Pr(Y_t = M | X_t = i) \right] \quad (7.5.5)$$

$$= \mathbf{p}_t C \quad (7.5.6)$$

Overall, an HMM with finite state-space and finite observation symbols is specified by the pair  $(P, C)$  and the initial state probability vector  $\mathbf{p}_0$ . Note that  $Y_t$  does not necessarily need to be a Markov process, so HMMs can still be used to model processes which are not observed to follow the Markov property.

#### Continuous-Observation Discrete-Time Hidden Markov Models

We can also allow for the observation  $Y_t$  to take on a continuum of values. Instead of a matrix  $C$ , we instead specify an observation likelihood density  $p(y|X_t = i)$ . An example of an observation process which takes on a continuum of values is the additive noise:

$$Y_t = h(X_t) + V_t \quad (7.5.7)$$

where  $h : \{1, \dots, N\} \rightarrow \mathbb{R}$  is some function of the state, and the noise  $V_t$  is an i.i.d. sequence of a continuous random variable.

### Difference Equation Representation of Hidden Markov Models [114]

A difference equation representation of hidden Markov models is attainable if we define the state to take on vectors:

$$X_t \in \{\mathbf{e}_1, \dots, \mathbf{e}_N\} \quad (7.5.8)$$

where  $\mathbf{e}_i$  is the  $i^{\text{th}}$  unit basis column vector in  $\mathbb{R}^N$ . Define

$$W_t := X_{t+1} - \mathbb{E}[X_{t+1}|X_0, \dots, X_t] \quad (7.5.9)$$

$$= X_{t+1} - \mathbb{E}[X_{t+1}|X_t] \quad (7.5.10)$$

by the Markov property. Suppose  $X_t = \mathbf{e}_i$ . Then

$$\mathbb{E}[X_{t+1}|X_t = \mathbf{e}_i] = \sum_{j=1}^N \Pr(X_{t+1} = \mathbf{e}_j | X_t = \mathbf{e}_i) \mathbf{e}_j \quad (7.5.11)$$

$$= \sum_{j=1}^N P_{ij} \mathbf{e}_j \quad (7.5.12)$$

Note that  $\sum_{j=1}^N P_{ij} \mathbf{e}_j$  consists of a vector made up of the elements in the  $i^{\text{th}}$  row of the transition matrix  $P$ , so it can be expressed as  $P^\top \mathbf{e}_i = P^\top X_t$ . Therefore

$$W_t = X_{t+1} - P^\top X_t \quad (7.5.13)$$

or in difference equation form:

$$X_{t+1} = P^\top X_t + W_t \quad (7.5.14)$$

Taking conditional expectations of  $W_t$ , we also see

$$\mathbb{E}[W_{t+1}|X_0, \dots, X_t] = \mathbb{E}[X_{t+1}|X_0, \dots, X_t] - \mathbb{E}[X_{t+1}|X_0, \dots, X_t] \quad (7.5.15)$$

$$= 0 \quad (7.5.16)$$

so  $W_t$  is a martingale difference sequence. Furthermore, if we assume additive observation noise, we have the equations

$$X_{t+1} = P^\top X_t + W_t \quad (7.5.17)$$

$$Y_t = CX_t + V_t \quad (7.5.18)$$

where we can write  $h(X_t) = CX_t$  without loss of generality since  $X_t$  is a unit basis vector.

### Continuous-State Discrete-Time Hidden Markov Models

A hidden Markov model with continuous-valued states and measurements can be specified using the same setup as the Bayes filter. We require an initial state density  $p(x_0)$ , state transition density  $p(x_{t+1}|x_t)$  and an observation likelihood  $p(y_t|x_t)$ .

#### 7.5.2 Forward Algorithm

The forward algorithm in hidden Markov models is a method for computing the joint distribution  $p(x_t, \mathbf{y}_{0:t})$ , where  $\mathbf{y}_{0:t}$  denotes all the observations up until time  $t$ , i.e.  $\mathbf{y}_{0:t} = (y_0, \dots, y_t)$ . We may generally assume  $p(x_t, \mathbf{y}_{0:t})$  to be a probability density, as probability masses can be represented with densities using Dirac delta distributions. A naive way to compute  $p(x_t, \mathbf{y}_{0:t})$  is to first compute  $p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t})$  using the transition densities and observation likelihoods by

$$p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) = p(\mathbf{y}_{0:t}|\mathbf{x}_{0:t}) p(\mathbf{x}_{0:t}) \quad (7.5.19)$$

where because of the Markov property

$$p(\mathbf{x}_{0:t}) = p(x_t|x_{t-1}) \dots p(x_1|x_0) p(x_0) \quad (7.5.20)$$

$$= p(x_0) \prod_{k=1}^t p(x_k|x_{k-1}) \quad (7.5.21)$$

and because of conditional independence of the observation

$$p(\mathbf{y}_{0:t}|\mathbf{x}_{0:t}) = p(y_t|\mathbf{x}_{0:t}, \mathbf{y}_{0:(t-1)}) p(\mathbf{y}_{0:(t-1)}|\mathbf{x}_{0:t}) \quad (7.5.22)$$

$$= p(y_t|x_t) p(\mathbf{y}_{0:(t-1)}|\mathbf{x}_{0:t}) \quad (7.5.23)$$

$$\vdots \quad (7.5.24)$$

$$= p(y_t|x_t) \dots p(y_0|x_0) \quad (7.5.25)$$

$$= \prod_{k=0}^t p(y_k|x_k) \quad (7.5.26)$$

Hence

$$p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) = p(x_0) \prod_{k=1}^t p(x_k|x_{k-1}) \cdot \prod_{j=0}^t p(y_j|x_j) \quad (7.5.27)$$

Then we may marginalise out  $\mathbf{x}_{0:(t-1)}$  to obtain  $p(x_t, \mathbf{y}_{0:t})$ :

$$p(x_t, \mathbf{y}_{0:t}) = \int \dots \int p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) dx_{t-1} \dots dx_0 \quad (7.5.28)$$

However as this approach involves a multidimensional integral, it may not be very efficient. The forward algorithm instead uses a recursive approach to compute  $p(x_t, \mathbf{y}_{0:t})$  from  $p(x_{t-1}, \mathbf{y}_{0:(t-1)})$ . The procedure is derived as follows. By marginalising only over  $x_{t-1}$ , we get

$$p(x_t, \mathbf{y}_{0:t}) = \int p(x_t, x_{t-1}, \mathbf{y}_{0:t}) dx_{t-1} \quad (7.5.29)$$

$$= \int p(y_t|x_t, x_{t-1}, \mathbf{y}_{0:(t-1)}) p(x_t, x_{t-1}, \mathbf{y}_{0:(t-1)}) dx_{t-1} \quad (7.5.30)$$

$$= \int p(y_t|x_t, x_{t-1}, \mathbf{y}_{0:(t-1)}) p(x_t|x_{t-1}, \mathbf{y}_{0:(t-1)}) p(x_{t-1}, \mathbf{y}_{0:(t-1)}) dx_{t-1} \quad (7.5.31)$$

using the chain rule of probability. By applying the Markov property and conditional independence of the observation, this simplifies to

$$p(x_t, \mathbf{y}_{0:t}) = \int p(y_t|x_t) p(x_t|x_{t-1}) p(x_{t-1}, \mathbf{y}_{0:(t-1)}) dx_{t-1} \quad (7.5.32)$$

$$= p(y_t|x_t) \int p(x_t|x_{t-1}) p(x_{t-1}, \mathbf{y}_{0:(t-1)}) dx_{t-1} \quad (7.5.33)$$

where we notice that it is recursive in terms of  $p(x_{t-1}, \mathbf{y}_{0:(t-1)})$ . Therefore computing  $p(x_t, \mathbf{y}_{0:t})$  requires successively computing  $p(x_0, y_0)$ ,  $p(x_1, \mathbf{y}_{0:1})$ , etc. The first term  $p(x_0, y_0)$  itself may be computed by

$$p(x_0, y_0) = p(y_0|x_0) p(x_0) \quad (7.5.34)$$

where we presume  $p(x_0)$  to be known. In the case where the state-space  $\mathcal{X}$  and observation symbols  $\mathcal{Y}$  are finite, then the recursion can be alternatively expressed using summation notation

$$p(x_t, \mathbf{y}_{0:t}) = p(y_t|x_t) \sum_{x_{t-1} \in \mathcal{X}} p(x_t|x_{t-1}) p(x_{t-1}, \mathbf{y}_{0:(t-1)}) \quad (7.5.35)$$

where the distributions now represent probability mass functions.

## Hidden Markov Model Filtering by Forward Algorithm

The Bayes filter outlines an approach to obtain the distribution  $p(x_t | \mathbf{y}_{0:t})$  in an HMM. We can also use the forward algorithm to explicitly compute  $p(x_t | \mathbf{y}_{0:t})$ . This is done by normalisation:

$$p(x_t | \mathbf{y}_{0:t}) = \frac{p(x_t, \mathbf{y}_{0:t})}{p(\mathbf{y}_{0:t})} \quad (7.5.36)$$

$$= \frac{p(x_t, \mathbf{y}_{0:t})}{\int p(x_t, \mathbf{y}_{0:t}) dx_t} \quad (7.5.37)$$

or if considering finite states:

$$p(x_t | \mathbf{y}_{0:t}) = \frac{p(x_t, \mathbf{y}_{0:t})}{\sum_{x_t \in \mathcal{X}} p(x_t, \mathbf{y}_{0:t})} \quad (7.5.38)$$

where each  $p(x_t, \mathbf{y}_{0:t})$  for  $x_t \in \mathcal{X}$  is computed using the forward algorithm.

### 7.5.3 Forward-Backward Algorithm [15]

The forward-backward algorithm computes the distribution  $p(x_t | \mathbf{y}_{0:T})$ , i.e. the distribution of the hidden state at time  $t$ , given all observations up to time  $T \geq t$ . Using Bayes' Theorem (all conditioned on  $\mathbf{y}_{0:t}$ ), we write

$$p(x_t | \mathbf{y}_{0:T}) = p(x_t | \mathbf{y}_{0:t}, \mathbf{y}_{(t+1):T}) \quad (7.5.39)$$

$$= \frac{p(\mathbf{y}_{(t+1):T} | \mathbf{y}_{0:t}, x_t) p(x_t | \mathbf{y}_{0:t})}{p(\mathbf{y}_{(t+1):T} | \mathbf{y}_{0:t})} \quad (7.5.40)$$

$$= \frac{p(\mathbf{y}_{(t+1):T} | x_t) p(x_t | \mathbf{y}_{0:t})}{p(\mathbf{y}_{(t+1):T} | \mathbf{y}_{0:t})} \quad (7.5.41)$$

by conditional independence of the observation. We then focus on computing the numerator, because the distribution can be normalised eventually:

$$p(x_t | \mathbf{y}_{0:T}) = \frac{p(\mathbf{y}_{(t+1):T} | x_t) p(x_t | \mathbf{y}_{0:t})}{\int p(\mathbf{y}_{(t+1):T} | x_t) p(x_t | \mathbf{y}_{0:t}) dx_t} \quad (7.5.42)$$

$$\propto p(\mathbf{y}_{(t+1):T} | x_t) p(x_t | \mathbf{y}_{0:t}) \quad (7.5.43)$$

Observe that  $p(x_t | \mathbf{y}_{0:t})$  can be obtained via filtering (e.g. using the forward algorithm) while we need to use a backwards recursion to compute  $p(\mathbf{y}_{(t+1):T} | x_t)$ . This backwards recursion can be derived similarly to the forwards algorithm (by applying the chain rule of probability and conditional independence properties):

$$p(\mathbf{y}_{(t+1):T} | x_t) = \int p(y_{t+1}, \mathbf{y}_{(t+2):T}, x_{t+1} | x_t) dx_{t+1} \quad (7.5.44)$$

$$= \int p(y_{t+1} | \mathbf{y}_{(t+2):T}, x_{t+1}, x_t) p(\mathbf{y}_{(t+2):T}, x_{t+1} | x_t) dx_{t+1} \quad (7.5.45)$$

$$= \int p(y_{t+1} | x_{t+1}) p(\mathbf{y}_{(t+2):T} | x_{t+1}, x_t) p(x_{t+1} | x_t) dx_{t+1} \quad (7.5.46)$$

$$= \int p(y_{t+1} | x_{t+1}) p(\mathbf{y}_{(t+2):T} | x_{t+1}) p(x_{t+1} | x_t) dx_{t+1} \quad (7.5.47)$$

where we have  $p(\mathbf{y}_{(t+2):T} | x_{t+1}, x_t) = p(\mathbf{y}_{(t+2):T} | x_{t+1})$  because  $x_t$  and the observations  $\mathbf{y}_{(t+2):T}$  are conditionally independent given  $x_{t+1}$ . Thus we see that we can successively compute  $p(y_T | x_{T-1})$ ,  $p(\mathbf{y}_{(T-1):T} | x_{T-2})$ , etc. until  $p(\mathbf{y}_{(t+1):T} | x_t)$ , beginning from

$$p(y_T | x_{T-1}) = \int p(y_T | x_T) p(x_T | x_{T-1}) dx_T \quad (7.5.48)$$

### Hidden Markov Model Smoothing by Forward-Backward Algorithm

The forward-backward algorithm can also be used to obtain the posterior joint distribution of the hidden states (i.e. smoothed estimates). First consider the posterior for the pair  $(x_t, x_{t+1})$ . This is given by

$$p(x_t, x_{t+1} | \mathbf{y}_{0:T}) = p(x_{t+1} | x_t, \mathbf{y}_{0:T}) p(x_t | \mathbf{y}_{0:T}) \quad (7.5.49)$$

$$= p(x_{t+1} | x_t) p(x_t | \mathbf{y}_{0:T}) \quad (7.5.50)$$

where  $p(x_t | \mathbf{y}_{0:T})$  is obtained via the usual forward-backward algorithm. Alternatively,

$$p(x_t, x_{t+1} | \mathbf{y}_{0:T}) = \frac{p(x_t, x_{t+1}, \mathbf{y}_{0:T})}{p(\mathbf{y}_{0:T})} \quad (7.5.51)$$

$$\propto p(x_t, x_{t+1}, \mathbf{y}_{0:T}) \quad (7.5.52)$$

$$= p(x_t, x_{t+1}, \mathbf{y}_{0:t}, y_{t+1}, \mathbf{y}_{(t+2):T}) \quad (7.5.53)$$

$$= p(\mathbf{y}_{(t+2):T} | x_t, x_{t+1}, \mathbf{y}_{0:t}, y_{t+1}) p(y_{t+1} | x_t, x_{t+1}, \mathbf{y}_{0:t}) p(x_{t+1} | x_t, \mathbf{y}_{0:t}) p(x_t, \mathbf{y}_{0:t}) \quad (7.5.54)$$

$$= p(\mathbf{y}_{(t+2):T} | x_{t+1}) p(y_{t+1} | x_{t+1}) p(x_{t+1} | x_t) p(x_t, \mathbf{y}_{0:t}) \quad (7.5.55)$$

where  $p(x_t, \mathbf{y}_{0:t})$  is directly computed from the forward algorithm and  $p(\mathbf{y}_{(t+2):T} | x_{t+1})$  from the backward recursion. Iterating either procedure, we can compute the full smoothed posterior  $p(\mathbf{x}_{0:T} | \mathbf{y}_{0:T})$ , e.g.

$$p(\mathbf{x}_{0:T} | \mathbf{y}_{0:T}) = p(x_T | x_{T-1}) \dots p(x_1 | x_0) p(x_0 | \mathbf{y}_{0:T}) \quad (7.5.56)$$

$$= p(x_0 | \mathbf{y}_{0:T}) \prod_{k=1}^T p(x_k | x_{k-1}) \quad (7.5.57)$$

#### 7.5.4 Hidden Markov Model Prediction

To obtain the predictive hidden state distribution  $p(x_{T+1} | \mathbf{y}_{0:T})$ , we can begin from the filtering distribution  $p(x_T | \mathbf{y}_{0:T})$  (which can be computed using the forward algorithm or Bayes filter), obtain the joint distribution  $p(x_{T+1}, x_T | \mathbf{y}_{0:T})$ , and then marginalise out  $x_T$  as follows:

$$p(x_{T+1} | \mathbf{y}_{0:T}) = \int p(x_{T+1}, x_T | \mathbf{y}_{0:T}) dx_T \quad (7.5.58)$$

$$= \int p(x_{T+1} | x_T) p(x_T | \mathbf{y}_{0:T}) dx_T \quad (7.5.59)$$

This can be generalised in the same fashion to obtain the  $h$ -step ahead predictive distribution  $p(x_{T+h} | \mathbf{y}_{0:T})$  by marginalising out the preceding variables:

$$p(x_{T+h} | \mathbf{y}_{0:T}) = \int \dots \int p(x_{T+h}, \dots, x_T | \mathbf{y}_{0:T}) dx_T \dots dx_{T+h-1} \quad (7.5.60)$$

$$= \int \dots \int \prod_{k=1}^h p(x_{T+k} | x_{T+k-1}) p(x_T | \mathbf{y}_{0:T}) dx_T \dots dx_{T+h-1} \quad (7.5.61)$$

We can also obtain the  $h$ -step ahead predictive distribution for the observation,  $p(y_{T+h} | \mathbf{y}_{0:T})$ . This involves chaining the observation likelihood and marginalising out  $x_{T+h}$ :

$$p(y_{T+h} | \mathbf{y}_{0:T}) = \int p(y_{T+h} | x_{T+h}) p(x_{T+h} | \mathbf{y}_{0:T}) dx_{T+h} \quad (7.5.62)$$

### 7.5.5 Viterbi Algorithm

The Viterbi algorithm aims to find the most likely sequence of states from a sequence of observations, that is

$$\hat{\mathbf{x}}_{0:T} = \underset{\mathbf{x}_{0:T}}{\operatorname{argmax}} p(\mathbf{x}_{0:T} | \mathbf{y}_{0:T}) \quad (7.5.63)$$

We can proceed by instead maximising over the joint distribution:

$$\underset{\mathbf{x}_{0:T}}{\operatorname{argmax}} p(\mathbf{x}_{0:T} | \mathbf{y}_{0:T}) = \underset{\mathbf{x}_{0:T}}{\operatorname{argmax}} p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T}) \quad (7.5.64)$$

since  $p(\mathbf{x}_{0:T} | \mathbf{y}_{0:T}) \propto p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T})$  where the constant of proportionality only depends on the observations  $\mathbf{y}_{0:T}$ , and thus do not affect the maximisation. Using the conditional independence properties of hidden Markov models, the problem can be written as

$$\max_{\mathbf{x}_{0:T}} p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T}) = \max_{\mathbf{x}_{0:T}} \left\{ \prod_{k=0}^T p(y_k | x_k) p(x_k | x_{k-1}) \right\} \quad (7.5.65)$$

where we take  $p(x_0 | x_{-1}) = p(x_0)$ . In log form, this becomes

$$\underset{\mathbf{x}_{0:T}}{\operatorname{argmax}} p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T}) = \underset{\mathbf{x}_{0:T}}{\operatorname{argmax}} \left\{ \sum_{k=0}^T \log(p(x_k | x_{k-1})) \right\} \quad (7.5.66)$$

The summations and maximisations in this representation can be separated as follows:

$$\begin{aligned} & \max_{\mathbf{x}_{0:T}} \{ \log p(y_0 | x_0) p(x_0) + \log p(y_1 | x_1) p(x_1 | x_0) + \dots + \log p(y_T | x_T) p(x_T | x_{T-1}) \} \\ &= \max_{x_0} \left\{ \log p(y_0 | x_0) p(x_0) + \max_{x_1} \left\{ \log p(y_1 | x_1) p(x_1 | x_0) + \dots + \max_{x_T} \{ \log p(y_T | x_T) p(x_T | x_{T-1}) \} \dots \right\} \right\} \end{aligned} \quad (7.5.67)$$

which shows how the Viterbi algorithm is connected to dynamic programming. This connection also suggests that we may be able to compute the estimates in a backwards order, i.e.  $\hat{x}_T, \hat{x}_{T-1}, \dots, \hat{x}_0$ . However, this cannot be done directly because the innermost maximisation of  $p(y_T | x_T) p(x_T | x_{T-1})$  still depends on  $x_{T-1}$  (and implicitly all of  $\mathbf{x}_{0:(T-1)}$ ). Hence we need to somehow propagate this dependence forward; the Viterbi algorithm can be thought of as doing this in the first stage. A recursive way for performing this is as follows. First consider

$$p(\mathbf{x}_{0:(t+1)}, \mathbf{y}_{0:(t+1)}) = p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}, x_{t+1}, y_{t+1}) \quad (7.5.68)$$

$$= p(x_{t+1}, y_{t+1} | \mathbf{x}_{0:t}, \mathbf{y}_{0:t}) p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \quad (7.5.69)$$

$$= p(y_{t+1} | x_{t+1}) p(x_{t+1} | x_t) p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \quad (7.5.70)$$

Hence

$$\max_{\mathbf{x}_{0:t}} p(\mathbf{x}_{0:(t+1)}, \mathbf{y}_{0:(t+1)}) = p(y_{t+1} | x_{t+1}) \max_{\mathbf{x}_{0:t}} \{ p(x_{t+1} | x_t) p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \} \quad (7.5.71)$$

$$= p(y_{t+1} | x_{t+1}) \max_{x_t} \left\{ p(x_{t+1} | x_t) \max_{\mathbf{x}_{0:(t-1)}} p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \right\} \quad (7.5.72)$$

as the term  $p(x_{t+1} | x_t)$  does not contain any of the variables from  $\mathbf{x}_{0:(t-1)}$ . Define

$$V_t(x_t) = \max_{\mathbf{x}_{0:(t-1)}} p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \quad (7.5.73)$$

which is written purely as a function of  $x_t$  since we have maximised with respect to  $\mathbf{x}_{0:(t-1)}$ , and given all the observations  $\mathbf{y}_{0:t}$  are known (hence fixed). From above, we can then see the recursion

$$V_{t+1}(x_{t+1}) = p(y_{t+1} | x_{t+1}) \max_{x_t} \{ p(x_{t+1} | x_t) V_t(x_t) \} \quad (7.5.74)$$

The steps in the Viterbi algorithm are summarised as follows [201].

1. Initialise  $V_0(x_0) = p(y_0|x_0)p(x_0)$ . If there are finitely many states, we can instead represent this function with a vector  $\mathbf{V}_0$  where the  $i^{\text{th}}$  element is  $\Pr(x_0 = i)p(y_0|x_0 = i)$ .
2. For each  $t = 1, \dots, T$ , perform the recursion

$$V_t(x_t) = p(y_t|x_t) \max_{x_{t-1}} \{p(x_t|x_{t-1}) V_{t-1}(x_{t-1})\} \quad (7.5.75)$$

and define

$$U_t(x_t) = \operatorname{argmax}_{x_{t-1}} \{p(x_t|x_{t-1}) V_{t-1}(x_{t-1})\} \quad (7.5.76)$$

Likewise, if there are finitely many states, we can alternatively represent these functions as vectors  $\mathbf{V}_t$  and  $\mathbf{U}_t$  respectively.

3. Let

$$v^* = \max_{x_T} V_T(x_T) \quad (7.5.77)$$

$$\hat{x}_T = \operatorname{argmax}_{x_T} V_T(x_T) \quad (7.5.78)$$

4. Perform backtracking to obtain the remainder of the estimates:

$$\hat{x}_{T-1} = U_T(\hat{x}_T) \quad (7.5.79)$$

$$\vdots \quad (7.5.80)$$

$$\hat{x}_t = U_{t+1}(\hat{x}_{t+1}) \quad (7.5.81)$$

$$\vdots \quad (7.5.82)$$

$$\hat{x}_0 = U_1(\hat{x}_1) \quad (7.5.83)$$

thus we have  $\hat{\mathbf{x}}_{0:T} = (\hat{x}_0, \dots, \hat{x}_T)$ .

To explain why this works, first consider from the definition of  $V_t(x_t)$  that:

$$v^* = \max_{x_T} \left\{ \max_{\mathbf{x}_{0:(T-1)}} \{p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T})\} \right\} \quad (7.5.84)$$

Hence  $v^*$  is the likelihood of the most likely sequence. Next, we have

$$\hat{x}_T = \operatorname{argmax}_{x_T} \left\{ \max_{\mathbf{x}_{0:(T-1)}} \{p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T})\} \right\} \quad (7.5.85)$$

Therefore  $\hat{x}_T$  belongs on the most likely sequence, as everything else has been maximised with respect to  $\mathbf{x}_{0:(T-1)}$ . In the first backtracking step, we have

$$\hat{x}_{T-1} = \operatorname{argmax}_{x_{T-1}} \{p(\hat{x}_T|x_{T-1}) V_{T-1}(x_{T-1})\} \quad (7.5.86)$$

$$= \operatorname{argmax}_{x_{T-1}} \left\{ p(\hat{x}_T|x_{T-1}) \max_{\mathbf{x}_{0:(T-2)}} p(\mathbf{x}_{0:(T-1)}, \mathbf{y}_{0:(T-1)}) \right\} \quad (7.5.87)$$

$$= \operatorname{argmax}_{x_{T-1}} \left\{ p(y_T|\hat{x}_T) p(\hat{x}_T|x_{T-1}) \max_{\mathbf{x}_{0:(T-2)}} p(\mathbf{x}_{0:(T-1)}, \mathbf{y}_{0:(T-1)}) \right\} \quad (7.5.88)$$

$$= \operatorname{argmax}_{x_{T-1}} \left\{ \max_{\mathbf{x}_{0:(T-2)}, x_T} p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T}) \right\} \quad (7.5.89)$$

Because  $\hat{x}_T$  belongs on the most likely trajectory  $\hat{\mathbf{x}}_{0:T}$ , and we are simultaneously maximising with respect to  $\mathbf{x}_{0:(T-2)}$ , then it follows that  $\hat{x}_{T-1}$  must also belong to  $\hat{\mathbf{x}}_{0:T}$ . Generally, this applies to all backtracked computations:

$$\hat{x}_t = U_{t+1}(\hat{x}_{t+1}) \quad (7.5.90)$$

$$= \operatorname{argmax}_{x_t} \{ p(\hat{x}_{t+1}|x_t) V_t(x_t) \} \quad (7.5.91)$$

$$= \operatorname{argmax}_{x_t} \left\{ p(\hat{x}_{t+1}|x_t) \max_{\mathbf{x}_{0:(t-1)}} p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \right\} \quad (7.5.92)$$

$$= \operatorname{argmax}_{x_t} \left\{ \left( \prod_{k=t+1}^{T-1} p(y_{k+1}|\hat{x}_k) p(\hat{x}_{k+1}|\hat{x}_k) \right) p(y_{t+1}|\hat{x}_{t+1}) p(\hat{x}_{t+1}|x_t) \max_{\mathbf{x}_{0:(t-1)}} p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \right\} \quad (7.5.93)$$

$$= \operatorname{argmax}_{x_t} \left\{ \max_{\mathbf{x}_{0:(t-1)}, \mathbf{x}_{(t+1):T}} p(\mathbf{x}_{0:t}, \mathbf{y}_{0:t}) \right\} \quad (7.5.94)$$

so  $\hat{x}_t$  belongs on the trajectory  $\hat{\mathbf{x}}_{0:T}$ . We can view the role of  $U_t(x_t)$  as the most likely estimate of  $x_{t-1}$  (from  $\hat{\mathbf{x}}_{0:T}$ ) as a function of the most likely estimate of  $x_t$ .

### 7.5.6 Baum-Welch Algorithm [76]

The Baum-Welch algorithm is the application of the EM algorithm for the maximum likelihood estimation of the parameters of a finite-state and finite-observation symbol HMM, given the observation sequence  $\mathbf{y}_{0:T}$ . Let there be a nominal  $N$  states, from the set  $\{1, \dots, N\}$ , and suppose there are  $M$  observation symbols, from the set  $\{1, \dots, M\}$ . The parameters to be estimated are

$$\theta = (A, C, \pi) \quad (7.5.95)$$

where  $A$  is the  $N \times N$  transition probability matrix,  $C$  is the  $N \times M$  emission probability matrix, and  $\pi$  denotes the initial state probability vector. Using the framework of the EM algorithm, closed-form solutions to the maximisation step can be found. For ease of notation, denote  $y := \mathbf{y}_{0:T}$  and  $x = \mathbf{x}_{0:T}$ . We use  $y$  as the observed data, while  $x$  plays the role of the latent variables. Suppose  $\hat{\theta}_k$  is the current parameter estimate at iteration  $k$  of the EM algorithm. We can then formulate

$$Q(\theta; \hat{\theta}_k) = \mathbb{E}_{X|y; \hat{\theta}_k} [\log p(X, y; \theta)] \quad (7.5.96)$$

$$= \sum_x p(x|y; \hat{\theta}_k) \cdot \log p(x, y; A, C, \pi) \quad (7.5.97)$$

Note that the sum is taken over all possible length  $T + 1$  sequences for  $\mathbf{x}_{0:T}$ . Then in the maximisation step, suppressing the constraints on  $\theta$  in the meantime, we may write:

$$\max_{\theta} Q(\theta; \hat{\theta}_k) = \max_{\theta} \left\{ \sum_x p(x|y; \hat{\theta}_k) \cdot \log p(x, y; A, C, \pi) \right\} \quad (7.5.98)$$

$$= \max_{\theta} \left\{ \sum_x p(x|y; \hat{\theta}_k) \cdot \log \left( p(x_0; \pi) \prod_{t=1}^T p(x_t|x_{t-1}; A) \cdot \prod_{t=0}^T p(y_t|x_t; C) \right) \right\} \quad (7.5.99)$$

$$= \max_{\theta} \left\{ \sum_x p(x|y; \hat{\theta}_k) \cdot \left( \log p(x_0; \pi) + \sum_{t=1}^T \log p(x_t|x_{t-1}; A) + \sum_{t=0}^T \log p(y_t|x_t; C) \right) \right\} \quad (7.5.100)$$

Notice this maximisation can be split up, in that

$$\begin{aligned} \max_{\theta} Q(\theta; \hat{\theta}_k) &= \max_A \left\{ \sum_x p(x|y; \hat{\theta}_k) \left( \sum_{t=1}^T \log p(x_t|x_{t-1}; A) \right) \right\} \\ &+ \max_{\pi} \left\{ \sum_x p(x|y; \hat{\theta}_k) \cdot \log p(x_0; \pi) \right\} + \max_C \left\{ \sum_x p(x|y; \hat{\theta}_k) \left( \sum_{t=0}^T \log p(y_t|x_t; C) \right) \right\} \end{aligned} \quad (7.5.101)$$

This shows that we can maximise separately with respect to  $A$ ,  $C$  and  $\pi$ . Beginning with  $A$ , we have:

$$\hat{A}_{k+1} = \operatorname{argmax}_A \left\{ \sum_x p(x|y; \hat{\theta}_k) \left( \sum_{t=1}^T \log p(x_t|x_{t-1}; A) \right) \right\} \quad (7.5.102)$$

$$= \operatorname{argmax}_A \left\{ \sum_x p(x|y; \hat{\theta}_k) \left( \sum_{i=1}^N \sum_{j=1}^N n_{i,j}(x) \cdot \log a_{i,j} \right) \right\} \quad (7.5.103)$$

$$= \operatorname{argmax}_A \left\{ \sum_{i=1}^N \sum_{j=1}^N \left( \sum_x p(x|y; \hat{\theta}_k) \cdot n_{i,j}(x) \right) \cdot \log a_{i,j} \right\} \quad (7.5.104)$$

where the  $a_{i,j}$  are elements of  $A$ , and  $n_{i,j}(x)$  denotes the number of transitions from state  $i$  to  $j$  in the sequence  $x = \mathbf{x}_{0:T}$ . As we take  $A$  to be right-stochastic (i.e. each row sums to one), we can further separate the problem by solving for each row  $\mathbf{a}_1, \dots, \mathbf{a}_N$ . Explicitly, and with constraints, the problem is now

$$\begin{aligned} \max_{\mathbf{a}_i} \quad & \sum_{j=1}^N \left( \sum_x p(x|y; \hat{\theta}_k) \cdot n_{i,j}(x) \right) \cdot \log a_{i,j} \\ \text{s.t.} \quad & \sum_{j=1}^N a_{i,j} = 1 \\ & a_{i,j} \geq 0, \quad j = 1, \dots, N \end{aligned} \quad (7.5.105)$$

Let  $q_{i,j} := \sum_x p(x|y; \hat{\theta}_k) \cdot n_{i,j}(x)$ , so after normalisation,  $\frac{q_{i,j}}{\sum_{j=1}^N q_{i,j}}$  are the masses of a distribution. Dividing the objective by  $\sum_{j=1}^N q_{i,j}$  (which does not affect the maximiser), we can see that the problem now becomes similar to the derivation of the maximum entropy distribution over a finite support. Using the same Gibb's inequality argument, the objective is then maximised when we set

$$\hat{a}_{i,j} = \frac{q_{i,j}}{\sum_{j=1}^N q_{i,j}} \quad (7.5.106)$$

To compute  $q_{i,j}$  itself more easily (as an alternative to literally summing over all the sequences), we use from definition

$$q_{i,j} := \sum_x p(x|y; \hat{\theta}_k) \cdot n_{i,j}(x) \quad (7.5.107)$$

$$= \mathbb{E}_{X|y; \hat{\theta}_k} [n_{i,j}(X)] \quad (7.5.108)$$

$$= \mathbb{E}_{X|y; \hat{\theta}_k} \left[ \sum_{t=1}^T \mathbb{I}_{\{X_{t-1}=i, X_t=j\}} \right] \quad (7.5.109)$$

$$= \sum_{t=1}^T \mathbb{E}_{X|y;\hat{\theta}_k} [\mathbb{I}_{\{X_{t-1}=i, X_t=j\}}] \quad (7.5.110)$$

$$= \sum_{t=1}^T \Pr(X_{t-1} = i, X_t = j | y; \hat{\theta}_k) \quad (7.5.111)$$

where the probability  $\Pr(X_{t-1} = i, X_t = j | y; \hat{\theta}_k)$  can be computed using smoothing via the forward-backward algorithm. Now to find the update for  $\pi$ ,

$$\hat{\pi}_{k+1} = \operatorname{argmax}_{\pi} \left\{ \sum_x p(x | y; \hat{\theta}_k) \cdot \log p(x_0; \pi) \right\} \quad (7.5.112)$$

$$= \operatorname{argmax}_{\pi} \left\{ \sum_x p(x | y; \hat{\theta}_k) \cdot \log \pi_x \right\} \quad (7.5.113)$$

Analogously, this is solved via

$$\hat{\pi}_{i,k+1} = \Pr(X_0 = i | y; \hat{\theta}_k) \quad (7.5.114)$$

which can again be computed via the forward-backward algorithm. Lastly for  $C$ , we obtain

$$\hat{C}_{k+1} = \operatorname{argmax}_C \left\{ \sum_x p(x | y; \hat{\theta}_k) \left( \sum_{t=0}^T \log p(y_t | x_t; C) \right) \right\} \quad (7.5.115)$$

$$= \operatorname{argmax}_C \left\{ \sum_x p(x | y; \hat{\theta}_k) \left( \sum_{i=1}^N \sum_{\ell=1}^M \eta_{i,\ell}(x, y) \log c_{i,\ell} \right) \right\} \quad (7.5.116)$$

$$= \operatorname{argmax}_C \left\{ \sum_{i=1}^N \sum_{\ell=1}^M \left( \sum_x p(x | y; \hat{\theta}_k) \cdot \eta_{i,\ell}(x, y) \right) \cdot \log c_{i,\ell} \right\} \quad (7.5.117)$$

where  $\eta_{i,\ell}(x, y) := \sum_{t=0}^T \mathbb{I}_{\{x_t=i, y_t=\ell\}}$  denotes the number of times the pair  $(i, \ell)$  appears in  $(\mathbf{x}_{0:T}, \mathbf{y}_{0:T})$ . As with above, we can separate this problem by solving for each of the rows of  $C$ , denoted  $\mathbf{c}_1, \dots, \mathbf{c}_N$ . This gives

$$\begin{aligned} \max_{\mathbf{c}_i} & \sum_{\ell=1}^M \left( \sum_x p(x | y; \hat{\theta}_k) \cdot \eta_{i,\ell}(x, y) \right) \cdot \log c_{i,\ell} \\ \text{s.t. } & \sum_{\ell=1}^M c_{i,\ell} = 1 \\ & c_{i,\ell} \geq 0, \quad \ell = 1, \dots, M \end{aligned} \quad (7.5.118)$$

Defining  $g_{i,\ell} := \sum_x p(x | y; \hat{\theta}_k) \cdot \eta_{i,\ell}(x, y)$  and applying the same reasoning as above, the objective is maximised when

$$\hat{c}_{i,\ell} = \frac{g_{i,\ell}}{\sum_{\ell=1}^M g_{i,\ell}} \quad (7.5.119)$$

Additionally, we can compute each  $g_{i,\ell}$  by

$$g_{i,\ell} = \mathbb{E}_{X|y;\hat{\theta}_k} [\eta_{i,\ell}(X, y)] \quad (7.5.120)$$

$$= \mathbb{E}_{X|y;\hat{\theta}_k} \left[ \sum_{t=0}^T \mathbb{I}_{\{X_t=i, y_t=\ell\}} \right] \quad (7.5.121)$$

$$= \sum_{t=0}^T \mathbb{E}_{X|y; \hat{\theta}_k} [\mathbb{I}_{\{X_t=i, y_t=\ell\}}] \quad (7.5.122)$$

$$= \sum_{t=0}^T \mathbb{E}_{X|y; \hat{\theta}_k} [\mathbb{I}_{\{X_t=i\}}] \cdot \mathbb{I}_{\{y_t=\ell\}} \quad (7.5.123)$$

since  $y$  has been already conditioned on. Hence

$$g_{i,\ell} = \sum_{t=0}^T \Pr(X_t = i | y; \hat{\theta}_k) \cdot \mathbb{I}_{\{y_t=\ell\}} \quad (7.5.124)$$

where the probabilities may be computed from the forward-backward algorithm once again.

### 7.5.7 Hidden Markov Model Estimation by Method of Moments [114]

The stationary distribution and transition matrix of a finite state and observation symbol hidden Markov model can be estimated using a method of moments approach. We assume that:

- The Markov chain transition matrix  $A \in \mathbb{R}^{N \times N}$  is regular, so there exists a unique stationary distribution  $\pi$  with all positive elements by the Perron-Frobenius theorem.
- The initial distribution is equal to the stationary distribution  $\pi$  (i.e. the process has been ‘active’ for a long time before we started collecting data).
- The observation matrix  $C \in \mathbb{R}^{N \times M}$  is known, and so does not need to be estimated.

To proceed with the estimation, we first construct the  $M \times M$  matrix  $S$ , such that each of its elements is

$$s_{ij} = \Pr(y_t = i, y_{t+1} = j) \quad (7.5.125)$$

Each of these elements can be computed by

$$s_{ij} = \sum_{k=1}^N \sum_{\ell=1}^N \Pr(y_t = i, y_{t+1} = j, x_t = k, x_{t+1} = \ell) \quad (7.5.126)$$

$$= \sum_{\ell=1}^N \Pr(y_{t+1} = j | x_{t+1} = \ell) \sum_{k=1}^N \Pr(x_{t+1} = \ell | x_t = k) \Pr(y_t = i | x_t = k) \Pr(x_t = k) \quad (7.5.127)$$

However, a more compact representation is given by

$$S = C^\top A^\top \text{diag}\{\pi\} C \quad (7.5.128)$$

To show this, we first write out  $\text{diag}\{\pi\} C$  as

$$\text{diag}\{\pi\} C = \begin{bmatrix} \pi_1 & & \\ & \ddots & \\ & & \pi_N \end{bmatrix} \begin{bmatrix} \Pr(y_t = 1 | x_t = 1) & \dots & \Pr(y_t = M | x_t = 1) \\ \vdots & \ddots & \vdots \\ \Pr(y_t = 1 | x_t = N) & \dots & \Pr(y_t = M | x_t = N) \end{bmatrix} \quad (7.5.129)$$

$$= \begin{bmatrix} \Pr(y_t = 1, x_t = 1) & \dots & \Pr(y_t = M, x_t = 1) \\ \vdots & \ddots & \vdots \\ \Pr(y_t = 1, x_t = N) & \dots & \Pr(y_t = M, x_t = N) \end{bmatrix} \quad (7.5.130)$$

Next,  $C^\top A^\top$  can be written as

$$C^\top A^\top = \begin{bmatrix} \Pr(y_{t+1} = 1|x_{t+1} = 1) & \dots & \Pr(y_{t+1} = 1|x_{t+1} = N) \\ \vdots & \ddots & \vdots \\ \Pr(y_{t+1} = M|x_{t+1} = 1) & \dots & \Pr(y_{t+1} = M|x_{t+1} = N) \end{bmatrix} \quad (7.5.131)$$

$$\times \begin{bmatrix} \Pr(x_{t+1} = 1|x_t = 1) & \dots & \Pr(x_{t+1} = 1|x_t = N) \\ \vdots & \ddots & \vdots \\ \Pr(x_{t+1} = N|x_t = 1) & \dots & \Pr(x_{t+1} = N|x_t = N) \end{bmatrix}$$

$$= \begin{bmatrix} \Pr(y_{t+1} = 1|x_t = 1) & \dots & \Pr(y_{t+1} = 1|x_t = N) \\ \vdots & \ddots & \vdots \\ \Pr(y_{t+1} = M|x_t = 1) & \dots & \Pr(y_{t+1} = M|x_t = N) \end{bmatrix} \quad (7.5.132)$$

Thus

$$C^\top A^\top \text{diag}\{\pi\} C = \begin{bmatrix} \Pr(y_{t+1} = 1|x_t = 1) & \dots & \Pr(y_{t+1} = 1|x_t = N) \\ \vdots & \ddots & \vdots \\ \Pr(y_{t+1} = M|x_t = 1) & \dots & \Pr(y_{t+1} = M|x_t = N) \end{bmatrix} \quad (7.5.133)$$

$$\times \begin{bmatrix} \Pr(y_t = 1, x_t = 1) & \dots & \Pr(y_t = M, x_t = 1) \\ \vdots & \ddots & \vdots \\ \Pr(y_t = 1, x_t = N) & \dots & \Pr(y_t = M, x_t = N) \end{bmatrix}$$

$$= \begin{bmatrix} \Pr(y_t = 1, y_{t+1} = 1) & \dots & \Pr(y_t = 1, y_{t+1} = M) \\ \vdots & \ddots & \vdots \\ \Pr(y_t = M, y_{t+1} = 1) & \dots & \Pr(y_t = M, y_{t+1} = M) \end{bmatrix} \quad (7.5.134)$$

$$= S \quad (7.5.135)$$

Suppose we have gathered observations  $\mathbf{y}_{0:T}$ . A method of moments estimation for  $S$  would involve counting consecutive pairs and normalising:

$$\widehat{S} = \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} \mathbb{I}_{\{y_t=1, y_{t-1}=1\}} & \dots & \mathbb{I}_{\{y_t=1, y_{t-1}=M\}} \\ \vdots & \ddots & \vdots \\ \mathbb{I}_{\{y_t=M, y_{t-1}=1\}} & \dots & \mathbb{I}_{\{y_t=M, y_{t-1}=M\}} \end{bmatrix} \quad (7.5.136)$$

Let  $B := A^\top \text{diag}\{\pi\}$  contain the variables to be estimated. We aim to find the matrix  $B$  such that  $\widehat{S} \approx C^\top BC$ . This can be formally done by solving the optimisation problem

$$\begin{aligned} \widehat{B} &= \underset{B}{\operatorname{argmin}} \quad \left\| \widehat{S} - C^\top BC \right\|_F^2 \\ \text{s.t.} \quad B &\geq 0 \\ \mathbf{1}^\top B \mathbf{1} &= 1 \end{aligned} \quad (7.5.137)$$

where  $\|\cdot\|_F$  is the Frobenius norm. Note that the matrix  $B$  contains the joint distribution over  $(x_t, x_{t+1})$ , which explains the constraints. This problem is convex, which we can verify by rearranging it into a constrained linear least squares problem. Let  $\mathbf{b} := \text{vec}(B)$ , and then by the ‘vec trick’, we have  $\text{vec}(C^\top BC) = (C^\top \otimes C^\top) \mathbf{b}$ . Therefore the problem is equivalently

$$\begin{aligned} \widehat{\mathbf{b}} &= \underset{\mathbf{b}}{\operatorname{argmin}} \quad \left\| \text{vec}(\widehat{S}) - (C^\top \otimes C^\top) \mathbf{b} \right\|_2^2 \\ \text{s.t.} \quad \mathbf{b} &\geq \mathbf{0} \\ \mathbf{1}^\top \mathbf{b} &= 1 \end{aligned} \quad (7.5.138)$$

Once  $\widehat{B}$  is obtained, we estimate  $\pi$  by

$$\widehat{\pi} = \mathbf{1}^\top \widehat{B} \quad (7.5.139)$$

(this amounts to marginalising out one of the  $(x_t, x_{t+1})$ ), and lastly we estimate  $A$  using the relation

$$\widehat{A} = \widehat{B} \operatorname{diag}\{\widehat{\pi}\}^{-1} \quad (7.5.140)$$

Note that  $\operatorname{diag}\{\pi\}^{-1}$  is guaranteed to exist under the assumption that the Markov chain is regular.

## 7.6 Markov Decision Processes

### 7.6.1 Discrete-Time Markov Decision Processes

A Markov decision process (also known as a controlled Markov process) is a stochastic process which can be minimally described as a process involving two different variables. There is a process for the states:  $X_0, X_1, X_2, \dots$ , and a process for actions:  $A_0, A_1, A_2, \dots$  in discrete-time. A Markov decision process then consists of the tuple  $(\mathcal{X}, \mathcal{A}_x, p(x'|x, a))$ , where

- $\mathcal{X}$  is the state-space (which we consider to generally be countable).
- $\mathcal{A}_x$  is the action space (which may also generally be countable), containing all the possible actions which are possible from state  $x$ .
- $p(x'|x, a) = \Pr(X_{t+1} = x' | X_t = x, A_t = a)$  is the transition probability to new state  $x'$  from previous state  $x$  after taking action  $a$ .

Implicitly, there is also assumed to be some policy

$$\pi(a|x) = \Pr(A_t = a | X_t = x) \quad (7.6.1)$$

which gives the probability distribution of choosing action  $a$  at state  $x$ . The support of this distribution will then be  $\mathcal{A}_x$ . The combination of the Markov decision process  $(\mathcal{X}, \mathcal{A}_x, p(x'|x, a))$  with a policy  $\pi(a|x)$  then induces the stochastic process  $X_0, X_1, X_2, \dots$  to behave like a Markov chain, with the transition probability  $p_\pi(x'|x) = \Pr(X_{t+1} = x' | X_t = x)$ , which is computed via marginalisation:

$$p_\pi(x'|x) = \sum_{a \in \mathcal{A}_x} p(x'|x, a) \pi(a|x) \quad (7.6.2)$$

### Episodic Markov Decision Processes [121]

An analogous absorbing state for a Markov decision process can be defined as a state  $x$  such that  $p(x|x, a) = 1$  for any action  $a \in \mathcal{A}_x$ . Then a Markov decision process is said to be episodic if there exists an absorbing state that will be reached almost surely with any policy.

### Finite-Horizon Markov Decision Processes

A Markov decision process is said to be finite-horizon if it is episodic, and in addition the absorbing state in question is always reached after a fixed finite  $T$  number of transitions.

### Communicating Markov Decision Processes

As the analogous concept of communicability, we say that a Markov decision process is communicating if for any pair of states  $x_1, x_2 \in \mathcal{A}$ , there exists some policy such that  $x_2$  can be reached from  $x_1$  eventually in the future while following the policy. This concept is opposite to that of episodic Markov decision processes.

7.6.2 Partially Observable Markov Decision Processes

7.6.3 Continuous-Time Markov Decision Processes

## 7.7 Markov Networks

7.7.1 Belief Propagation

7.8 Semi-Markov Chains [153, 203]

7.9 Quasistationary Distributions [45]

## Chapter 8

# Measure Theoretic Probability

## 8.1 Probability Spaces

### 8.1.1 Concepts in Probability Spaces

#### Elementary Events

An elementary event is an event which contains only a single outcome in the sample space. For example, in the experiment that a coin is flipped twice, the outcomes are  $\{H, T\}$ ,  $\{T, H\}$ ,  $\{H, H\}$  and  $\{T, T\}$ . The event ‘heads followed by tails’ is an elementary event because there is only a single outcome associated to it:  $\{H, T\}$ . However, the event ‘at least one heads’ is not an elementary event because there are three outcomes associated to it:  $\{H, T\}$ ,  $\{T, H\}$  and  $\{H, H\}$ .

#### Power Sets

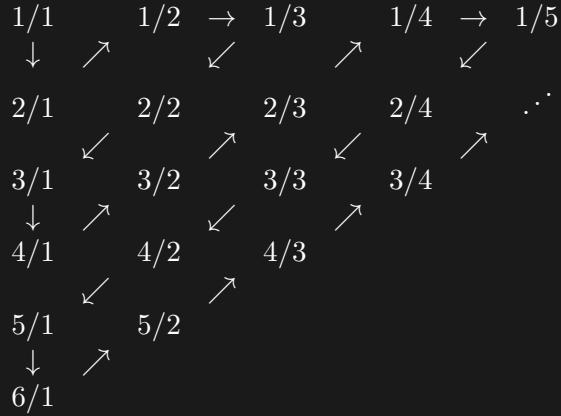
The power set of a set  $S$  is the set of all subsets of  $S$ , including the empty set and  $S$  itself. If  $S$  is a finite set with cardinality  $|S| = n$ , then the number of subsets of  $S$  is  $2^n$  (related to the binomial theorem). This motivates the notation for the power set of  $S$  as  $2^S$ .

The power set of all functions from  $Y$  to  $X$  can be denoted  $X^Y$ .

#### Countable Sets

A countable set  $S$  has the same cardinality  $|S|$  as some subset of the natural numbers. Intuitively speaking, a set is countable we can in some way assign numbers  $1, 2, 3, \dots$  to each element in  $S$  (i.e. we can ‘enumerate’ the elements of  $S$ ). A finite set will always be countable, but some infinite sets can still be countably infinite. For example, the set of natural numbers itself is trivially countably infinite. We can also enumerate through the set of positive rational numbers

in the following way:



Similarly, the set of all rational numbers is also countable. One possible enumeration is that we can imagine starting at 0, and then enumerating back and forth between a second ‘layer’ of a grid of negative rational numbers behind the grid of positive rational numbers.

### Uncountable Sets

An uncountable set is an infinite set that has ‘too many’ elements to be countable. An example of an uncountable set is the set of real numbers.

### Borel Sets

A Borel set is a set in topological space  $\Omega$  (e.g. sample space) that can be formed from the operations of countable union, countable intersection, and relative complement (i.e. the relative complement of  $A$  in  $B$  is denoted  $A \setminus B$ ) of open (or equivalently, of closed sets) in  $\Omega$ .

### $\sigma$ -algebras

The  $\sigma$ -algebra of a set  $\Omega$  is a collection of subsets of  $\Omega$  which:

- includes the empty subset.
- is closed under complement (i.e. the complement of a member of the  $\sigma$ -algebra is also a member of the  $\sigma$ -algebra).
- is closed under countable unions (i.e. the countable union of members of the  $\sigma$ -algebra is also a member of the  $\sigma$ -algebra).
- is closed under countable intersections (i.e. the countable intersection of members of the  $\sigma$ -algebra is also a member of the  $\sigma$ -algebra).

In general, the  $\sigma$ -algebra of  $\Omega$  is a subset of the power set of  $\Omega$ .

### Sub- $\sigma$ -algebras

Suppose  $\mathcal{F}$  is a  $\sigma$ -algebra on  $\Omega$ . Then another  $\sigma$ -algebra  $\mathcal{F}_1$  on  $\Omega$  is said to be a sub- $\sigma$ -algebra of  $\mathcal{F}$  if  $\mathcal{F}_1 \subseteq \mathcal{F}$ .

## Borel $\sigma$ -algebras

The Borel  $\sigma$ -algebra of a set  $\Omega$  is the collection of all Borel sets of  $\Omega$ . The Borel  $\sigma$ -algebra gives the smallest  $\sigma$ -algebra containing all open sets (or equivalently, all closed sets) of  $\Omega$ .

- In the case where  $\Omega$  is a countable set, then the power set is identical to the Borel  $\sigma$ -algebra.
- If  $\Omega$  is the real line  $\mathbb{R}$ , then the Borel  $\sigma$ -algebra includes all ‘reasonable’ (i.e. measurable) open and closed intervals, as well as their countable union/intersection and relative complement.

## Boundary Sets

The boundary set of  $\mathcal{B}$  is the set of points in the closure of  $\mathcal{B}$  but not in the interior of  $\mathcal{B}$ . The boundary set is denoted  $\partial\mathcal{B}$ .

## Continuity Sets

For a random variable (or vector)  $X$ , a Borel set  $\mathcal{B}$  is called a continuity set if

$$\mathbb{P}(X \in \partial\mathcal{B}) = 0 \quad (8.1.1)$$

For example, if  $X$  is a Bernoulli random variable, then  $[0, 1]$  nor  $(0, 1)$  are considered continuity sets, however  $(-1, -0.5)$ ,  $[0.1, 0.9]$  and  $[1.1, \infty)$  would be considered continuity sets. A continuity set can be made of any points at which the cumulative distribution function of  $X$  is continuous.

### 8.1.2 Probability Triple

In measure theoretic probability, a probability space is a measure space denoted with the triple  $(\Omega, \mathcal{F}, \mathbb{P})$  where  $\Omega$  is the sample space,  $\mathcal{F}$  is the event space and  $\mathbb{P}$  is a probability measure.

The set  $\Omega$  contains each elementary event. The event space  $\mathcal{F}$  is formed by taking a  $\sigma$ -algebra of  $\Omega$ . A typical choice of  $\sigma$ -algebra is the Borel  $\sigma$ -algebra. The function  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$  is called a probability measure, which maps events to probabilities.

## Probability Axioms

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space with sample space  $\Omega$ , event space  $\mathcal{F}$  and probability measure  $\mathbb{P}$ . The axioms of probability (sometimes known as Kolmogorov’s axioms) are:

1. The probability of an event  $E \in \mathcal{F}$  is a nonnegative real number

$$0 \leq \mathbb{P}(E) \in \mathbb{R} \quad (8.1.2)$$

for all  $E \in \mathcal{F}$ .

2. The probability that at least one of the elementary events in the entire sample space will occur is 1.

$$\mathbb{P}(\Omega) = 1 \quad (8.1.3)$$

3. Any countable sequence of disjoint (mutually exclusive) events  $E_1, E_2, \dots$  satisfies

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) \quad (8.1.4)$$

From the axioms, we can deduce further properties

- If  $A \subseteq B$ , then we have the monotonicity property

$$\mathbb{P}(A) \leq \mathbb{P}(B) \quad (8.1.5)$$

This can be seen by defining events  $E_1 = A$ ,  $E_2 = B \setminus A$  where  $A \subseteq B$ , and  $E_i = \emptyset$  for all  $i \geq 3$ . These sets are all disjoint (we may alternatively define disjoint sets to be sets whose intersection is the empty set). Additionally, we have  $\bigcup_{i=1}^{\infty} E_i = B$ . By the third axiom

$$\sum_{i=1}^{\infty} \mathbb{P}(E_i) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) + \sum_{i=3}^{\infty} \mathbb{P}(\emptyset) \quad (8.1.6)$$

$$= \mathbb{P}(B) \quad (8.1.7)$$

The terms  $\mathbb{P}(B \setminus A)$  and  $\sum_{i=3}^{\infty} \mathbb{P}(\emptyset)$  are non-negative, hence  $\mathbb{P}(A) \leq \mathbb{P}(B)$ .

- The probability of the empty set is zero.

$$\mathbb{P}(\emptyset) = 0 \quad (8.1.8)$$

In deriving the above, we saw that the sum in  $\mathbb{P}(A) + \mathbb{P}(B \setminus A) + \sum_{i=3}^{\infty} \mathbb{P}(\emptyset)$  was convergent. Therefore it must be that  $\mathbb{P}(\emptyset) = 0$ , otherwise the sum would be infinite.

- The numeric bound applies to any event  $E \in \mathcal{F}$ :

$$0 \leq \mathbb{P}(E) \leq 1 \quad (8.1.9)$$

This can be shown by applying the non-negativity axiom and the monotonicity property to any subset of  $\Omega$  since  $\mathbb{P}(\Omega) = 1$ .

## Almost Sure Events

An event  $E \in \mathcal{F}$  is said to happen *almost surely* if the event of  $E$  not happening has probability measure zero, that is  $\mathbb{P}(\bar{E}) = 0$ . Alternatively, we can say that  $E$  has probability measure  $\mathbb{P}(E) = 1$ . In contrast, *almost never* describes events which have probability measure zero.

## Atomic Probability Spaces

A measurable set  $A \in \mathcal{F}$  of a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is said to be *atomic* if  $\mathbb{P}(A) > 0$  for each measurable subset  $B \subseteq A$ , either  $\mathbb{P}(B) = 0$  or  $\mathbb{P}(B) = \mathbb{P}(A)$ . This intuitively means that we cannot ‘break down’ the event  $A$  any further in terms of probability. Elementary events are atomic, provided they have positive measure. Any probability space with countable sample space can be considered to be *purely atomic* (as any outcome of the sample space with zero measure can simply be omitted). A probability space without any atoms is referred to as *non-atomic*.

### 8.1.3 Measurability

#### Measurable Sets

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a valid probability space. Then all the members of the collection  $\mathcal{F}$  are said to be measurable sets, because we can assign a number  $\mathbb{P}(E)$  to each  $E \in \mathcal{F}$ .

## Non-Measurable Sets

An example of sets which are not measurable are Vitali sets. A Vitali set  $V$  is a subset of the interval  $[0, 1]$  such that for each real number  $r \in \mathbb{R}$ , there is exactly one number  $v \in V$  such that  $v - r$  is a rational number. A numerical example of this is with the real number  $10 + 1/\sqrt{2}$ , we can choose  $1/\sqrt{2} \in [0, 1]$  such that  $1/\sqrt{2} - (10 - 1/\sqrt{2})$  is rational. There are uncountably many Vitali sets, but all Vitali sets will satisfy this property.

We can show that Vitali sets are unmeasurable in the following way. Define the enumeration of all rational numbers in  $[-1, 1]$  to be  $q_1, q_2, \dots$ . Define the translated Vitali sets  $V_k := V + q_k = \{v + q_k : v \in V\}$  for  $k = 1, 2, \dots$  so that  $v_k = v + q_k \in V_k$ . These sets must be disjoint in order to satisfy the definition that there is exactly one  $v \in V$  for each real number. If shifting  $V$  by  $q_k$  causes  $V$  and  $V_k$  to have elements in common, then it implies  $V$  is not a Vitali set. Another way to state this is that there is no gap between any two elements in  $V$  equal to a rational number. If this were the case, then there could be more than one  $v \in V$  that could make  $v - r$  rational. Hence shifting  $V$  by a rational number  $q_k$  will ensure  $V_k$  and  $V$  will be disjoint. We can write the following inclusion

$$\bigcup_k V_k \subseteq [-1, 2] \quad (8.1.10)$$

since  $v \in [0, 1]$  and  $q_k \in [-1, 1]$ , then  $-1 \leq v + q_k \leq 2$ . Furthermore, consider any real number  $v_i \in [0, 1]$ . Then by definition there will be exactly one  $v \in [0, 1]$  such that  $v - v_i = -q_i$ , where  $q_i \in [-1, 1]$  since  $-1 \leq v - v_i \leq 1 \Rightarrow -1 \leq q_i \leq 1$ . Therefore  $v_i \in V_i$  and we can write

$$[0, 1] \subseteq \bigcup_k V_k \subseteq [-1, 2] \quad (8.1.11)$$

Suppose we can take the Lebesgue measure  $\lambda(\cdot)$  of these inclusions.

$$1 \leq \sum_{k=1}^{\infty} \lambda(V_k) \leq 3 \quad (8.1.12)$$

Since the Lebesgue measure is translation invariant, we have  $\lambda(V_k) = \lambda(V)$  (a constant) and then

$$1 \leq \sum_{k=1}^{\infty} \lambda(V) \leq 3 \quad (8.1.13)$$

This results in a contradiction. The infinite sum must either be zero or infinity, but neither is between 1 and 3. So a Vitali set is not Lebesgue measurable.

## Measurable Functions

A real valued function  $f : \Omega \rightarrow \mathbb{R}$  is said to be measurable if for every  $B \in \mathcal{B}$  (where  $\mathcal{B}$  is a  $\sigma$ -algebra of  $\mathbb{R}$ ), the preimage of  $B$  under  $f$  is an element of the  $\sigma$ -algebra of  $\Omega$ , denoted  $\mathcal{F}$ . Explicitly, we require

$$f^{-1}(B) := \{x \in \Omega | f(x) \in B\} \in \mathcal{F} \quad (8.1.14)$$

for all  $B \in \mathcal{B}$ . Although measurability is formally subject to the choice of  $\sigma$ -algebras  $\mathcal{B}$  and  $\mathcal{F}$ , we normally take these to be the Borel  $\sigma$ -algebras.

## Non-Measurable Functions

An example of a non-measurable function is an indicator function for a non-measurable set, such as a Vitali set. In that case, the preimage  $f^{-1}(1)$  becomes an unmeasurable set, which cannot exist in the  $\sigma$ -algebra  $\mathcal{F}$ .

**Essential Supremum****Essential Infimum****8.1.4 Borel-Cantelli Lemma**

Let  $E_1, E_2, \dots$  be a sequence of events in some probability space. The *limit supremum* of the sequence of events is the set of all outcomes where the event  $E_n$  occurs infinitely many times in the infinite sequence of events. That is,

$$\limsup_{n \rightarrow \infty} E_n = \bigcup_{k=1}^{\infty} E_k \cap \bigcup_{k=2}^{\infty} E_k \cap \dots \quad (8.1.15)$$

$$= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_k \quad (8.1.16)$$

The Borel-Cantelli lemma gives a characterisation of almost sure convergence.

**Lemma 8.1.** *For a sequence of events  $E_1, E_2, \dots$  in a probability space, if the sum of probabilities is finite*

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \quad (8.1.17)$$

*then the probability the event occurs infinitely often is zero:*

$$\mathbb{P}\left(\limsup_{n \rightarrow \infty} E_n\right) = 0 \quad (8.1.18)$$

*Proof.* Since  $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$ , then the series of probabilities converges, meaning  $\sum_{n=N}^{\infty} \mathbb{P}(E_n) \rightarrow 0$  as  $N \rightarrow \infty$ . Hence taking the greatest lower bound of the series gives

$$\inf_{N \geq 1} \sum_{n=N}^{\infty} \mathbb{P}(E_n) = 0 \quad (8.1.19)$$

Evaluating  $\mathbb{P}(\limsup_{n \rightarrow \infty} E_n)$  yields

$$\mathbb{P}\left(\limsup_{n \rightarrow \infty} E_n\right) = \mathbb{P}\left(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} E_n\right) \quad (8.1.20)$$

Note that by the chain rule of probability,  $\Pr(A \cap B) = \Pr(B|A)\Pr(A) \leq \Pr(A)$  so we can use the generalisation of this argument to state

$$\mathbb{P}\left(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} E_n\right) \leq \inf_{N \geq 1} \mathbb{P}\left(\bigcup_{n=N}^{\infty} E_n\right) \quad (8.1.21)$$

$$\leq \inf_{N \geq 1} \sum_{n=N}^{\infty} \mathbb{P}(E_n) = 0 \quad (8.1.22)$$

where the latter inequality comes from the generalisation of the concept  $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$  (this is called Boole's inequality). Hence  $\mathbb{P}(\limsup_{n \rightarrow \infty} E_n) = 0$   $\square$

An alternative proof is provided:

*Proof.* Let  $\mathbb{I}_n$  denote the indicator function for the event  $E_n$ . Then by the linearity of expectation

$$\mathbb{E} \left[ \sum_{n=1}^{\infty} \mathbb{I}_n \right] = \sum_{n=1}^{\infty} \mathbb{E} [\mathbb{I}_n] \quad (8.1.23)$$

$$= \sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \quad (8.1.24)$$

Suppose there is a non-zero probability that  $E_n$  occurs infinitely often, meaning  $\mathbb{P}(\sum_{n=1}^{\infty} \mathbb{I}_n = \infty) > 0$ . However if we attempt to find  $\mathbb{E}[\sum_{n=1}^{\infty} \mathbb{I}_n]$  by Lebesgue integration this gives

$$\mathbb{E} \left[ \sum_{n=1}^{\infty} \mathbb{I}_n \right] = \int_{\{\sum_{n=1}^{\infty} \mathbb{I}_n = \infty\}} \left( \sum_{n=1}^{\infty} \mathbb{I}_n \right) d\mathbb{P} + \int_{\{\sum_{n=1}^{\infty} \mathbb{I}_n \neq \infty\}} \left( \sum_{n=1}^{\infty} \mathbb{I}_n \right) d\mathbb{P} \quad (8.1.25)$$

$$\geq \int_{\{\sum_{n=1}^{\infty} \mathbb{I}_n = \infty\}} \left( \sum_{n=1}^{\infty} \mathbb{I}_n \right) d\mathbb{P} \quad (8.1.26)$$

Since there is a non-zero probability that  $\sum_{n=1}^{\infty} \mathbb{I}_n = \infty$ , then the integral is infinite, i.e.

$$\int_{\{\sum_{n=1}^{\infty} \mathbb{I}_n = \infty\}} \left( \sum_{n=1}^{\infty} \mathbb{I}_n \right) d\mathbb{P} = \infty \quad (8.1.27)$$

This results in  $\mathbb{E}[\sum_{n=1}^{\infty} \mathbb{I}_n] \geq \infty$ , which is a contradiction.  $\square$

### Converse Borel-Cantelli Lemma

We have the converse result that if the events  $E_1, E_2, \dots$  are independent and the sum of the probabilities diverges to infinity, i.e.

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty \quad (8.1.28)$$

then the probability that  $E_n$  occurs infinitely many times is 1, i.e.

$$\mathbb{P} \left( \limsup_{n \rightarrow \infty} E_n \right) = 1 \quad (8.1.29)$$

*Proof.* We can equivalently show that there is zero probability that  $E_n$  will occur a finite amount of times, i.e.

$$1 - \mathbb{P} \left( \limsup_{n \rightarrow \infty} E_n \right) = 0 \quad (8.1.30)$$

This probability can be rewritten as

$$1 - \mathbb{P} \left( \limsup_{n \rightarrow \infty} E_n \right) = \lim_{N \rightarrow \infty} \mathbb{P} \left( \bigcap_{n=N}^{\infty} \overline{E}_n \right) \quad (8.1.31)$$

where  $\overline{E}_n$  is the complement of  $E_n$ . Intuitively, this expresses the probability that there is some integer large enough such that any subsequent  $E_n$  can no longer occur. Since the  $E_n$  are independent, we can show

$$\mathbb{P} \left( \bigcap_{n=N}^{\infty} \overline{E}_n \right) = \prod_{n=N}^{\infty} \mathbb{P}(\overline{E}_n) \quad (8.1.32)$$

$$= \prod_{n=N}^{\infty} (1 - \mathbb{P}(E_n)) \quad (8.1.33)$$

$$\leq \prod_{n=N}^{\infty} \exp(-\mathbb{P}(E_n)) \quad (8.1.34)$$

where we have used the fact that  $1 - x \leq e^{-x}$  for  $x \geq 0$ . Hence

$$\mathbb{P}\left(\bigcap_{n=N}^{\infty} \bar{E}_n\right) \leq \prod_{n=N}^{\infty} \exp(-\mathbb{P}(E_n)) \quad (8.1.35)$$

$$= \exp\left(-\sum_{n=N}^{\infty} \mathbb{P}(E_n)\right) \quad (8.1.36)$$

$$= 0 \quad (8.1.37)$$

since  $-\sum_{n=N}^{\infty} \mathbb{P}(E_n) = -\infty$  by hypothesis. Therefore

$$1 - \mathbb{P}\left(\limsup_{n \rightarrow \infty} E_n\right) = \lim_{N \rightarrow \infty} \mathbb{P}\left(\bigcap_{n=N}^{\infty} \bar{E}_n\right) \quad (8.1.38)$$

$$= 0 \quad (8.1.39)$$

which completes the proof.  $\square$

## 8.2 Lebesgue Integration

### 8.2.1 Riemann Integrability

Lebesgue integration can be used to answer valid questions in probability involving integrals, where Riemann integration cannot.

#### Darboux Integrals

Darboux integrals are an equivalent formulation of integration compared to Riemann integration. That is, a function is Riemann integrable if and only if it is Darboux integrable, and the two integrals are equal, whenever they exist. Consider a function  $f(x)$  that is to be integrated on the domain  $[a, b]$ . Moreover, suppose  $f(x)$  is bounded over this interval. Denote a partition of  $[a, b]$  by

$$\mathcal{P} = \{x_0, \dots, x_n\} \quad (8.2.1)$$

which consists on  $n$  disjoint non-empty subintervals, denoted  $\mathcal{I}_1, \dots, \mathcal{I}_n$ , with endpoints

$$a = x_0 < \dots < x_n = b \quad (8.2.2)$$

Define the *upper Darboux sum* by

$$U = \sum_{k=1}^n \sup_{x \in \mathcal{I}_k} f(x) \cdot (x_k - x_{k-1}) \quad (8.2.3)$$

which can be thought of as the sum of the area of upper rectangles in a graph of  $f(x)$  (which over-approximates the area beneath the curve). Similarly, define the *lower Darboux sum* as

$$L = \sum_{k=1}^n \inf_{x \in \mathcal{I}_k} f(x) \cdot (x_k - x_{k-1}) \quad (8.2.4)$$

which can be thought of as the sum of the area of lower rectangles in a graph of  $f(x)$  (which under-approximates the area beneath the curve). From this, we let the *upper Darboux integral* be defined as

$$\underline{U} = \inf_{\mathcal{P} \in \Pi} U \quad (8.2.5)$$

where  $\Pi$  is the set of all partitions of  $[a, b]$ . We also let the *lower Darboux integral* be defined as

$$\bar{L} = \sup_{\mathcal{P} \in \Pi} L \quad (8.2.6)$$

We can imagine that these infima and suprema are reached in the limit, if we are allowed infinitely many infinitesimally thin subintervals. If  $\underline{U} = \bar{L}$ , we say that the function is Darboux integrable, and call the common value

$$\int_a^b f(x) dx = \underline{U} = \bar{L} \quad (8.2.7)$$

the Darboux integral. By equivalence, we say that a function is Riemann integrable if and only if  $\underline{U} = \bar{L}$ .

### Non-Riemann Integrable Functions

We exhibit a function that is not Riemann integrable, but whose integral otherwise yields a valid probability computation. Consider the *Dirichlet function*  $f : [0, 1] \rightarrow \mathbb{R}$ , defined by

$$f(x) = \begin{cases} 1, & x \in [0, 1] \cap \mathbb{Q} \\ 0, & x \in [0, 1] \cap (\mathbb{R} \setminus \mathbb{Q}) \end{cases} \quad (8.2.8)$$

where  $\mathbb{Q}$  is the set of rational numbers. That is, the Dirichlet function is the indicator function for the rational numbers in  $[0, 1]$ , and for compactness we may denote it by  $f(x) = \mathbb{I}_{\{x \in \mathbb{Q}\}}$ . Since the set of rational and irrational numbers are both dense in the reals (i.e. any rational number is arbitrarily close to an irrational number, and vice-versa), then any non-empty interval  $\mathcal{I}_k$  has

$$\sup_{x \in \mathcal{I}_k} f(x) = 1 \quad (8.2.9)$$

$$\inf_{x \in \mathcal{I}_k} f(x) = 0 \quad (8.2.10)$$

Thus the upper Darboux sum is given by

$$U = \sum_{k=1}^n (x_k - x_{k-1}) \quad (8.2.11)$$

$$= x_n - x_0 \quad (8.2.12)$$

$$= 1 \quad (8.2.13)$$

while its lower Darboux sum is

$$L = \sum_{k=1}^n 0 \cdot (x_k - x_{k-1}) \quad (8.2.14)$$

$$= 0 \quad (8.2.15)$$

This shows that  $\underline{U} = 1 \neq \bar{L} = 0$ , meaning the Dirichlet function is not Riemann integrable. The intuition is that the function is too discontinuous to be integrated in a Riemann sense. However, the integral of the Dirichlet function:

$$\int_0^1 \mathbb{I}_{\{x \in \mathbb{Q}\}} dx = \mathbb{E}_X [\mathbb{I}_{\{X \in \mathbb{Q}\}}] \quad (8.2.16)$$

$$= \mathbb{P}(X \in \mathbb{Q}) \quad (8.2.17)$$

is characterised as the probability that  $X$  uniformly distributed on  $[0, 1]$  is a rational number. Applying the logic that there are uncountably many irrational numbers compared to only countably infinitely many rational numbers, this probability should be zero. This rationale can also be used to demonstrate that the Dirichlet function is measurable, as  $f^{-1}(1)$  is the set of all rational numbers in  $[0, 1]$  (with measure zero), while  $f^{-1}(0)$  is the set of all irrational numbers in  $[0, 1]$ , which has measure one. Using Lebesgue integration, it can indeed be shown that

$$\int_0^1 \mathbb{I}_{\{x \in \mathbb{Q}\}} dx = 0 \quad (8.2.18)$$

### 8.2.2 Lebesgue Integral over Probability Spaces

In a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , the Lebesgue integral of a measurable function  $f(x)$  over a measurable subset  $E$  of  $\Omega$  with respect to the measure  $\mathbb{P}$  may be denoted as

$$\int_E f(x) d\mathbb{P}(x) \quad (8.2.19)$$

#### Simple Functions

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , suppose there is an indicator function  $\mathbb{I}_E(\omega) : \Omega \rightarrow \mathbb{R}$  where  $E$  is a measurable subset of  $\Omega$ . That is,

$$\mathbb{I}_E(\omega) = \begin{cases} 1, & \omega \in E \\ 0, & \text{otherwise} \end{cases} \quad (8.2.20)$$

To decide what value the Lebesgue integral of this function should take, the only reasonable choice is to set the integral as

$$\int_{\Omega} \mathbb{I}_E(\omega) d\mathbb{P}(\omega) = \mathbb{P}(E) \quad (8.2.21)$$

A simple function is a finite linear combination of indicator functions for sets  $E_1, \dots, E_n$  that are a partition of  $\Omega$  is denoted

$$f_n(\omega) = \sum_{i=1}^n a_i \mathbb{I}_{E_i}(\omega) \quad (8.2.22)$$

where  $a_i$  are non-negative coefficients. The integrals for such functions can be defined as

$$\int_{\Omega} f_n(\omega) d\mathbb{P}(\omega) = \sum_i a_i \mathbb{P}(E_i) \quad (8.2.23)$$

#### Monotone Convergence Theorem

For non-negative functions which cannot be simply written as a finite linear combination of indicator functions, one can take the limit of a sequence of simple functions as  $n \rightarrow \infty$ . The Lebesgue integral of such functions can then be defined as

$$\int_{\Omega} f(\omega) d\mathbb{P}(\omega) = \sup_{f_n(\omega) = \sum_{i=1}^n a_i \mathbb{I}_{E_i}(\omega)} \left\{ \int_{\Omega} f_n(\omega) d\mathbb{P}(\omega) \middle| 0 \leq f_n(\omega) \leq f(\omega) \forall \omega \in \Omega \right\} \quad (8.2.24)$$

i.e. take the Lebesgue integral of the ‘largest’ simple function bounded by  $f(\omega)$ . The monotone convergence theorem asserts that if there is a non-decreasing sequence of non-negative functions

(i.e.  $0 \leq f_n(\omega) \leq f_{n+1}(\omega)$  for all  $\omega \in \Omega$  and for all  $n \geq 0$ ), and if  $f_n(\omega)$  converges to  $f(\omega)$  pointwise:

$$\lim_{n \rightarrow \infty} f_n(\omega) = f(\omega) \quad (8.2.25)$$

for all  $\omega \in \Omega$ , then the limit of the Lebesgue integral of  $f_n(\omega)$  as  $n \rightarrow \infty$  is the same as the definition from above:

$$\lim_{n \rightarrow \infty} \int_{\Omega} f_n(\omega) d\mathbb{P}(\omega) = \int_{\Omega} f(\omega) d\mathbb{P}(\omega) \quad (8.2.26)$$

Note that the sequence of simple functions as  $n \rightarrow \infty$  is non-decreasing by virtue of the coefficients  $a_i$  being defined to be non-negative.

### Lebesgue Integral of Signed Functions

The Lebesgue integral can be generalised to functions which can take on negative values. Define  $f^+(\omega) = \sup\{f(\omega), 0\}$  (i.e. the positive part) and  $f^-(\omega) = \sup\{-f(\omega), 0\}$  (i.e. the negative part). Note that both  $f^+(\omega) \geq 0$  and  $f^-(\omega) \geq 0$ . Then the Lebesgue integral can be defined as

$$\int_{\Omega} f(\omega) d\mathbb{P}(\omega) = \int_{\Omega} f^+(\omega) d\mathbb{P}(\omega) - \int_{\Omega} f^-(\omega) d\mathbb{P}(\omega) \quad (8.2.27)$$

whenever at least one of the integrals is finite. Note that the Lebesgue integral generalises the Riemann integral, so the Lebesgue integral of a function is the same as the Riemann integral, if the Riemann integral is defined/exists.

### 8.2.3 Monte-Carlo Characterisation of Lebesgue Integral [112]

An intuitive characterisation of the Lebesgue integral  $\int_{[0,1]} f(x) dx$  is to set it to be equal to the probability limit of  $\frac{1}{n} \sum_{i=1}^n f(X_i)$  where  $X_i$  are generated uniformly on  $[0, 1]$ . This characterisation is justified by the Monte-Carlo estimate of an integral.

### 8.2.4 Measure-Theoretic Random Variables

In measure theoretic probability, a real valued random variable  $X$  is defined as a measurable function  $X : \Omega \rightarrow \mathbb{R}$ . For each  $\omega \in \Omega$ , the function  $X(\omega)$  assigns a number to each outcome in the sample space. For example, in an experiment involving a sequence of coin tosses, let the sample space be  $\Omega = \{H, T, HH, HT, TH, TT\}$ . We can define the random variable  $X(\omega)$  to be the number of heads tossed. In that case, we have  $X(H) = 1$ ,  $X(T) = 0$ ,  $X(HH) = 2$ ,  $X(HT) = 1$ ,  $X(TH) = 1$ ,  $X(TT) = 0$ .

The function  $X^{-1} : \mathcal{B} \rightarrow \mathcal{F}$  is a mapping from the Borel  $\sigma$ -algebra of  $\mathbb{R}$  to a  $\sigma$ -algebra of  $\Omega$ . It is defined to be

$$X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\} \quad (8.2.28)$$

That is, given a Borel set  $B$ , the function finds all the elementary events  $\omega \in \Omega$  which lead to the random variable  $X$  being in  $B$ . For the example sample space  $\Omega$  above, we can say that  $X^{-1}(\{2\}) = \{HH\}$  and  $X^{-1}(\{1\}) = \{H, HT, TH\}$ .

#### $\sigma$ -algebra Generated by a Random Variable

Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable. Denote by  $\mathcal{B}$  the Borel  $\sigma$ -algebra of  $\mathbb{R}$ . Then the  $\sigma$ -algebra generated by  $X$ , denoted  $\sigma(X)$ , is defined as

$$\sigma(X) = \{X^{-1}(B) | B \in \mathcal{B}\} \quad (8.2.29)$$

That is, for every Borel set of  $\mathbb{R}$  we find the subset of  $\Omega$  which leads to  $X$  being in the Borel set. The collection of all these subsets is known as the  $\sigma$ -algebra generated by  $X$ . This will also be known as the smallest  $\sigma$ -algebra for which  $X$  is measurable.

### Absolutely Continuous Random Variables

Consider a probability space with the real line as the sample space, and a random variable  $X$  which is just the identity mapping (i.e. a random variable  $X$  on the real line). Then  $X$  is said to be an absolutely continuous random variable if there exists a bounded density function  $f(x)$  such that

$$\mathbb{P}(A) = \int_A f(x) dx \quad (8.2.30)$$

for all Borel sets  $A$ . This is a stronger condition than for continuous random variables, which is just that  $\mathbb{P}(X = x) = 0$  for all  $x \in \mathbb{R}$ . Many common families of continuous random variables are absolutely continuous. However, an example of a continuous random variable which is not absolutely continuous comes from the Cantor distribution.

### Singular Distributions

A singular distribution is a distribution that is concentrated on a set of Lebesgue measure zero, however the probability of each point on that set is zero. Since the probability of any point is zero, then this necessitates that the cumulative distribution function is continuous. An example of a singular distribution is the bivariate distribution of comonotonic random variables  $X$  and  $Y$ . The probability will be entirely concentrated on the line  $x = y$ , which has Lebesgue measure zero in  $\mathbb{R}^2$ . Another example of a singular distribution is the Cantor distribution.

Singular distributions will not have probability density functions. To see why, suppose the distribution is concentrated on the set  $\mathcal{X}$  with Lebesgue measure zero. Then there is no function  $f(x)$  such that

$$\int_{\mathcal{X}} f(x) dx = 1 \quad (8.2.31)$$

without using point masses, as it is required by definition of singular distributions that every point  $x \in \mathcal{X}$  has  $\mathbb{P}(X = x) = 0$ .

Singular distributions are related to, but not the exact same as, degenerate distributions. Comonotonic random variables will have both singular and degenerate distributions. However in the univariate case, a Dirac delta distribution is degenerate but not singular, while the Cantor distribution is singular but not degenerate.

The decomposition of a CDF can be extended to include continuous singular components. We can now write the CDF of any random variable (or random vector, for that matter) as

$$F(x) = p_{\text{a.c.}} F_{\text{a.c.}}(x) + p_{\text{disc.}} F_{\text{disc.}}(x) + p_{\text{sing.}} F_{\text{sing.}}(x) \quad (8.2.32)$$

with  $p_{\text{a.c.}}, p_{\text{disc.}}, p_{\text{sing.}} \geq 0$  and  $p_{\text{a.c.}} + p_{\text{disc.}} + p_{\text{sing.}} = 1$ , where:

- $p_{\text{a.c.}} F_{\text{a.c.}}(x)$  corresponds to the absolutely continuous component,
- $p_{\text{disc.}} F_{\text{disc.}}(x)$  corresponds to the discrete component,
- $p_{\text{sing.}} F_{\text{sing.}}(x)$  corresponds to the singular component.

The discrete component can be obtained by taking the summation over the probability mass function, while the continuous component can be obtained by integrating over the probability density function. As the singular component does not have a valid density however, it can instead be obtained by taking the difference:

$$p_{\text{sing.}} F_{\text{sing.}}(x) = F(x) - p_{\text{a.c.}} F_{\text{a.c.}}(x) - p_{\text{disc.}} F_{\text{disc.}}(x) \quad (8.2.33)$$

### 8.2.5 Measure-Theoretic Expectation

The expectation of a random variable  $X(\omega)$  is defined using the Lebesgue integral

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega) \quad (8.2.34)$$

### 8.2.6 Fatou's Lemma

### 8.2.7 Dominated Convergence Theorem

Suppose there is a sequence of random variables  $\{X_n\}$  converging to  $X$  almost surely as  $n \rightarrow \infty$ . If there exists a random variable  $Y$  such that  $\mathbb{E}[Y] < \infty$  and  $|X_n| < Y$  for all  $n$ , then this implies expectation and limits can be exchanged, i.e.

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \mathbb{E}\left[\lim_{n \rightarrow \infty} X_n\right] = \mathbb{E}[X] \quad (8.2.35)$$

Although it might seem intuitive that the expectation and limit be always exchangeable, a simple counterexample can be given. Let a sequence of random variables be defined by

$$X_n = \begin{cases} 2^n, & \text{w.p. } \frac{1}{2^n} \\ 0, & \text{w.p. } 1 - \frac{1}{2^n} \end{cases} \quad (8.2.36)$$

The expectation can be calculated to be  $\mathbb{E}[X_n] = 1$  for all  $n$ . However, notice that  $X_n \xrightarrow{\text{a.s.}} 0$ . This can be shown using the Borel-Cantelli lemma, by noting the convergent series

$$\sum_{n=1}^{\infty} \Pr(X_n = 1) = \sum_{n=1}^{\infty} \frac{1}{2^n} = 1 \quad (8.2.37)$$

so the event  $\{X_n = 1\}$  occurs finitely many times, meaning  $\Pr(\lim_{n \rightarrow \infty} X_n = 0) = 1$ . Therefore in this case,

$$\mathbb{E}[X_n] = 1 \neq \mathbb{E}\left[\lim_{n \rightarrow \infty} X_n\right] = \mathbb{E}[0] = 0 \quad (8.2.38)$$

because there does not exist a random variable  $Y$  with finite expectation such that  $|X_n| < Y$  for all  $n$ .

## 8.3 Radon-Nikodym Derivatives

### 8.3.1 Radon-Nikodym Theorem

### 8.3.2 Smoothing Law

## 8.4 Product Measures

### 8.4.1 Fubini's Theorem

## 8.5 Convergence of Probability Measures

## 8.6 Measure Theoretic Stochastic Processes

### 8.6.1 Concept of a Stochastic Process [9]

In the language of measure-theoretic probability, a stochastic process  $X(t; \omega)$  with  $t \in T$  and  $\omega \in \Omega$  on probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  may be treated as a function of two arguments, where  $T$  is called the index set and  $\Omega$  is the sample space. When  $T \subseteq \mathbb{Z}$ , the process is a discrete-time process while if  $T \subseteq \mathbb{R}$ , the process is a continuous-time process. For fixed  $t \in T$ ,  $X(t, \cdot)$  is a random variable while for fixed  $\omega \in \Omega$ ,  $X(\cdot, \omega)$  is a realisation of the process, sometimes referred to as a sample path.

### Finite-Dimensional Distributions

For time indices  $t_1, \dots, t_k$ , we can assign a cumulative distribution function for the random vector  $(X(t_1; \omega), \dots, X(t_k; \omega))$  using the measure  $\mathbb{P}$ , denoted by

$$F_{t_1, \dots, t_k}(x_1, \dots, x_k) = \mathbb{P}(\{X(t_1; \omega) \leq x_1, \dots, X(t_k; \omega) \leq x_k\}) \quad (8.6.1)$$

called the finite-dimensional distribution (also known as the  $k^{\text{th}}$  order distribution). A stochastic process may be uniquely determined by all its finite-dimensional distributions (the conditions which guarantee this are covered in the Kolmogorov extension theorem).

### Stochastic Process Ensembles

The space of sample paths  $\mathcal{X}$  may be called the sample function space, or also the ensemble. So a stochastic process can be described as a function which maps a point  $\omega \in \Omega$  to the ensemble  $\mathcal{X}$ . The measure  $\mathbb{P}$  defined on  $\mathcal{F}$  induces a measure on the space  $\mathcal{X}$  in the following way. Consider the sets  $A \in \mathcal{X}$  such that the corresponding sets  $\{\omega : X(\cdot; \omega) \in A\} \in \Omega$  (i.e. the elementary events attached to the mapping which produce  $A$ ) are measurable. Then we can define this induced measure as

$$\mathbb{P}'(\{X(\cdot, \omega) \in A\}) = \mathbb{P}(\{\omega : X(\cdot; \omega) \in A\}) \quad (8.6.2)$$

### Non-Borel Sets of Stochastic Processes

In continuous-time processes, the set of bounded sample paths  $\{\omega : X(t; \omega) \leq c, \forall t \in (a, b)\}$  is not a Borel set, because the construction of this set would require the intersection of uncountably many sets (i.e. the set  $(a, b)$  is uncountable). Hence in general it is not possible to give the probability that sample paths are bounded. However, this restriction is lifted for the analogous set in discrete-time processes, because the set of time indices would be countable.

**8.6.2 Filtrations****Adapted Processes****8.6.3 Martingales****Supermartingales****Submartingales****Martingale Difference Sequences****Martingale Central Limit Theorem****8.6.4 Stopping Times****8.6.5 Law of the Iterated Logarithm****8.6.6 Doob Decomposition Theorem****8.7 Ergodic Theory****8.7.1 Birkhoff's Ergodic Theorem [112]**

# Chapter 9

# Advanced Statistics

## 9.1 Asymptotic Statistics

### 9.1.1 Law of Large Numbers for Correlated Sequences

The traditional law of large numbers only applies to independent (or uncorrelated) sequences. However, we can additionally show a law of large numbers that holds when the sequences are ‘weakly’ correlated in a particular way, using notion of **ergodicity** as the strongest possible characterisation of a process which satisfies the law of large numbers. Let  $X_t$  be a second order process with constant mean function  $\mathbb{E}[X_t] = 0$  and autocovariance  $C(t, s) = \mathbb{E}[(X_t - \mu)(X_s - \mu)]$  which is ‘weakly’ correlated in the sense that for any  $s$ ,

$$\sum_{t=-\infty}^{\infty} C(t, s) \leq c \quad (9.1.1)$$

where  $c < \infty$  is a constant which does not depend on  $s$ . In this case we call the autocovariance *summable*. Then the time-average of the process converges in probability to  $\mu$ , i.e.

$$\frac{1}{T} \sum_{t=1}^T X_t \xrightarrow{\text{P}} \mu \quad (9.1.2)$$

*Proof.* Let  $S_T = \sum_{t=1}^T X_t$  and  $\bar{X}_t = S_T/T$ . So  $\mathbb{E}[S_T] = T\mu$  and we can show for the variance of  $S_T$ :

$$\text{Var}(S_T) = \mathbb{E}[(S_T - T\mu)^2] \quad (9.1.3)$$

$$= \mathbb{E}\left[\left(\sum_{t=1}^T (X_t - \mu)\right)^2\right] \quad (9.1.4)$$

$$= \mathbb{E}\left[\left(\sum_{t=1}^T (X_t - \mu)\right) \left(\sum_{s=1}^T (X_s - \mu)\right)\right] \quad (9.1.5)$$

$$= \sum_{s=1}^T \sum_{t=1}^T \mathbb{E}[(X_t - \mu)(X_s - \mu)] \quad (9.1.6)$$

$$\leq \sum_{s=1}^T \sum_{t=-\infty}^{\infty} C(t, s) \quad (9.1.7)$$

$$\leq Tc \quad (9.1.8)$$

Then using Chebychev's inequality, for any  $\varepsilon > 0$ :

$$\Pr(|\bar{X}_t - \mu| > \varepsilon) \leq \frac{\text{Var}(\bar{X}_t - \mu)}{\varepsilon^2} \quad (9.1.9)$$

$$= \frac{\text{Var}(S_t - T\mu)}{T^2\varepsilon^2} \quad (9.1.10)$$

$$= \frac{\mathbb{E}[(S_T - T\mu)^2]}{T^2\varepsilon^2} \quad (9.1.11)$$

$$\leq \frac{c}{T\varepsilon^2} \quad (9.1.12)$$

which satisfies convergence in probability:

$$\lim_{T \rightarrow \infty} \Pr(|\bar{X}_t - \mu| > \varepsilon) = 0 \quad (9.1.13)$$

□

We can naturally specialise the autocovariance condition to weakly stationary processes, which can be written as  $C(\tau)$  in terms of the time difference  $\tau$ . Then since  $C(\tau) = C(-\tau)$ , we require

$$\sum_{\tau=0}^{\infty} C(\tau) \leq c \quad (9.1.14)$$

for some constant  $c < \infty$ . This characterises that the autocovariance should decay 'fast enough' (e.g. exponentially decaying autocovariance) in order for the law of large numbers to hold. This is analogous to the mean-ergodicity condition for weakly stationary continuous-time processes.

### 9.1.2 Pointwise Convergence in Probability

Pointwise convergence in probability extends the notion of convergence in probability to sequences which are parametrised by some parameter  $\theta$  from a parameter space  $\Theta$ . For example, the sequence  $\{X_n\}$  can be thought of as estimators and  $\theta$  is a parameter of the data generating process. A formal definition is given as follows.

Denote by  $X$  a random variable obtained from an experiment parametrised by  $\theta \in \Theta$ , and denote by  $X_\theta$  the random variable obtained by fixing  $\theta$  during the experiment. Denote by  $\{X_{1,\theta}, X_{2,\theta}, \dots, X_{n,\theta}\}$  a sequence of random variables obtained by fixing  $\theta$  during the experiment. Then the sequence is said to be pointwise convergent in probability to  $X$  if for each  $\theta \in \Theta$

$$\lim_{n \rightarrow \infty} \Pr(|X_{n,\theta} - X_\theta| > \varepsilon) = 0 \quad (9.1.15)$$

for all  $\varepsilon > 0$ . That is, convergence in probability holds for any value of  $\theta \in \Theta$ . An alternative equivalent definition is that given any  $\varepsilon > 0$ ,  $\delta > 0$  and  $\theta \in \Theta$ , one can find an integer  $n_0$  (possibly dependent on  $\varepsilon, \delta, \theta$ ) such that

$$\Pr(|X_{n,\theta} - X_\theta| > \varepsilon) < \delta \quad (9.1.16)$$

if  $n > n_0$ . Pointwise convergence in probability can be analogously defined for sequences of random vectors (using  $\|\cdot\|$  rather than  $|\cdot|$ ).

### 9.1.3 Uniform Convergence in Probability

Uniform convergence in probability is a stronger condition of pointwise convergence in probability. Denote by  $X$  a random variable obtained from an experiment parametrised by  $\theta \in \Theta$ , and denote by  $X_\theta$  the random variable obtained by fixing  $\theta$  during the experiment. Denote by  $\{X_{1,\theta}, X_{2,\theta}, \dots, X_{n,\theta}\}$  a sequence of random variables obtained by fixing  $\theta$  during the experiment. Then the sequence is said to be uniformly convergent in probability to  $X$  if

$$\lim_{n \rightarrow \infty} \Pr \left( \sup_{\theta \in \Theta} |X_{n,\theta} - X_\theta| > \varepsilon \right) = 0 \quad (9.1.17)$$

That is,  $\sup_{\theta \in \Theta} |X_{n,\theta} - X_\theta|$  converges in probability to zero. By examining an alternative equivalent definition, we have that given any  $\varepsilon > 0$ ,  $\delta > 0$ , one can find an integer  $n_0$  (possibly dependent on  $\varepsilon, \delta$ ) such that

$$\Pr(|X_{n,\theta} - X_\theta| > \varepsilon) < \delta \quad (9.1.18)$$

for all  $\theta \in \Theta$  if  $n > n_0$ . The difference from pointwise convergence in probability is that one must now relax dependence on  $\theta$ . So we need to find a  $n_0$  for a given  $\varepsilon, \delta$  that works for all  $\theta \in \Theta$ . Uniform convergence in probability can be analogously defined for sequences of random vectors (using  $\|\cdot\|$  rather than  $|\cdot|$ ).

### 9.1.4 Delta Method [204]

The delta method allows us to derive limiting distributions of functions of statistics with known limiting distributions, which allows us to perform inference on them. Suppose  $\hat{\theta}_n$  be a sequence of estimators for  $\theta \in \mathbb{R}^k$  such that

$$r_n (\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathbf{Z} \quad (9.1.19)$$

where  $\mathbf{Z}$  has some arbitrary distribution and  $r_n > 0$  represents an increasing sequence where  $r_n \rightarrow \infty$  as  $n \rightarrow \infty$ . We assume  $\hat{\theta}_n \xrightarrow{P} \theta$  so that  $\mathbb{E}[\mathbf{Z}] = 0$ . Let  $h : \mathbb{R}^k \rightarrow \mathbb{R}^m$  be a mapping that is differentiable at  $\theta^*$ . Then a first-order Taylor approximation of  $h(\hat{\theta}_n)$  about  $\theta^*$  gives

$$h(\hat{\theta}_n) \approx h(\theta^*) + \frac{\partial h(\theta)}{\partial \theta} \Big|_{\theta=\theta^*}^\top (\hat{\theta}_n - \theta^*) \quad (9.1.20)$$

Rearranging:

$$h(\hat{\theta}_n) - h(\theta^*) \approx \frac{\partial h(\theta)}{\partial \theta} \Big|_{\theta=\theta^*}^\top (\hat{\theta}_n - \theta^*) \quad (9.1.21)$$

$$r_n (h(\hat{\theta}_n) - h(\theta^*)) \approx \frac{\partial h(\theta)}{\partial \theta} \Big|_{\theta=\theta^*}^\top r_n (\hat{\theta}_n - \theta^*) \quad (9.1.22)$$

Since  $\frac{\partial h(\theta)}{\partial \theta} \Big|_{\theta=\theta^*}^\top$  is just a matrix that performs left-multiplication, then by Slutsky's theorem

$$r_n (h(\hat{\theta}_n) - h(\theta^*)) \xrightarrow{d} \frac{\partial h(\theta)}{\partial \theta} \Big|_{\theta=\theta^*}^\top \mathbf{Z} \quad (9.1.23)$$

where the approximation vanishes as  $\hat{\theta}_n \xrightarrow{P} \theta$ . Specialising this to the case where  $h : \mathbb{R}^k \rightarrow \mathbb{R}$ , we can write

$$r_n (h(\hat{\theta}_n) - h(\theta^*)) \xrightarrow{d} \nabla_{\theta} h(\theta^*)^\top \mathbf{Z} \quad (9.1.24)$$

The frequently-encountered scenario is when  $r_n = \sqrt{n}$  and  $\mathbf{Z}$  is a  $\mathcal{N}(0, \Sigma)$  distribution. In that case, we have

$$\sqrt{n} (h(\hat{\theta}_n) - h(\theta^*)) \xrightarrow{d} \mathcal{N}(0, \nabla_{\theta} h(\theta^*)^\top \Sigma \nabla_{\theta} h(\theta^*)) \quad (9.1.25)$$

### 9.1.5 Weierstrass Approximation Theorem [161]

### 9.1.6 Edgeworth Series Expansions

The (probabilists') Hermite polynomials are a polynomial sequence defined by

$$\text{He}_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2} \quad (9.1.26)$$

The first few Hermite polynomials obtained via evaluation are as follows.

$$\text{He}_0(x) = e^{x^2/2} e^{-x^2/2} \quad (9.1.27)$$

$$= 1 \quad (9.1.28)$$

$$\text{He}_1(x) = -e^{x^2/2} \times -xe^{-x^2/2} \quad (9.1.29)$$

$$= x \quad (9.1.30)$$

$$\text{He}_2(x) = e^{x^2/2} \times -\frac{d}{dx} xe^{-x^2/2} \quad (9.1.31)$$

$$= e^{x^2/2} \left( x^2 e^{-x^2/2} - e^{-x^2/2} \right) \quad (9.1.32)$$

$$= x^2 - 1 \quad (9.1.33)$$

A recurrence relation can also be obtained for the Hermite polynomials. Note that

$$\text{He}'_{n-1}(x) = (-1)^{n-1} \frac{d}{dx} \left[ e^{x^2/2} \frac{d^{n-1}}{dx^{n-1}} e^{-x^2/2} \right] \quad (9.1.34)$$

$$= (-1)^{n-1} xe^{x^2/2} \frac{d^{n-1}}{dx^{n-1}} e^{-x^2/2} + (-1)^{n-1} e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2} \quad (9.1.35)$$

$$= x \text{He}_{n-1}(x) - (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2} \quad (9.1.36)$$

$$= x \text{He}_{n-1}(x) - \text{He}_n(x) \quad (9.1.37)$$

Hence

$$\text{He}_{n+1}(x) = x \text{He}_n(x) - \text{He}_n(x) \quad (9.1.38)$$

## 9.2 Empirical Measures

### 9.2.1 Empirical Distribution Function

For a sample  $X_1, \dots, X_n$ , the empirical distribution function  $\hat{F}_n(x)$  gives the proportion of observations in the sample less than or equal to  $x$ :

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x\}} \quad (9.2.1)$$

If the sample is i.i.d. with common cumulative distribution function  $F(x)$ , then  $\mathbb{I}_{\{X_i \leq x\}}$  is a Bernoulli random variable with mean  $F(x)$ , and  $\hat{F}_n(x)$  is an unbiased estimator for  $F(x)$ .

### Empirical Density Function

The density function for a sample  $X_1, \dots, X_n$  may be expressed using a mixture of Dirac delta distributions:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - X_i) \quad (9.2.2)$$

### 9.2.2 Glivenko-Cantelli Theorem

By the strong law of large numbers,  $\widehat{F}_n(x) \xrightarrow{\text{a.s.}} F(x)$ . Hence the empirical distribution function converges pointwise almost surely to the population cumulative distribution function. The Glivenko-Cantelli theorem gives a stronger result, which is that the empirical distribution function converges uniformly to the population cumulative distribution function. That is, for any  $\varepsilon > 0$  we can find a  $n^*$  such that for all  $n \geq n^*$ , we have

$$\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \leq \varepsilon \quad (9.2.3)$$

almost surely.

*Proof.* For simplicity, consider a continuous random variable (hence  $F(x)$  is continuous). For any integer  $m \geq 1$ , we can fix  $m+1$  points on the extended real number line by

$$-\infty = x_0 < x_1 < \cdots < x_m = \infty \quad (9.2.4)$$

such that the difference in CDF between successive points satisfies

$$F(x_j) - F(x_{j-1}) = \frac{1}{m} \quad (9.2.5)$$

for  $j = 1, \dots, m$ . This is because  $F(x)$  is continuous non-decreasing. Then for each  $x \in \mathbb{R}$ , there exists an integer  $j^* \in \{1, \dots, m\}$  such that  $x_{j^*-1} \leq x \leq x_{j^*}$ . So for each  $x \in \mathbb{R}$ , we can upper bound

$$\widehat{F}_n(x) - F(x) \leq \widehat{F}_n(x_{j^*}) - F(x_{j^*-1}) \quad (9.2.6)$$

$$= \widehat{F}_n(x_{j^*}) - \left( F(x_{j^*}) - \frac{1}{m} \right) \quad (9.2.7)$$

$$= \widehat{F}_n(x_{j^*}) - F(x_{j^*}) + \frac{1}{m} \quad (9.2.8)$$

Similarly, for each  $x \in \mathbb{R}$ , we can lower bound

$$\widehat{F}_n(x) - F(x) \geq \widehat{F}_n(x_{j^*-1}) - F(x_{j^*}) \quad (9.2.9)$$

$$= \widehat{F}_n(x_{j^*-1}) - \left( F(x_{j^*-1}) + \frac{1}{m} \right) \quad (9.2.10)$$

$$= \widehat{F}_n(x_{j^*-1}) - F(x_{j^*-1}) - \frac{1}{m} \quad (9.2.11)$$

Combining both bounds,

$$\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \leq \max_{j \in \{1, \dots, m\}} |\widehat{F}_n(x_j) - F(x_j)| + \frac{1}{m} \quad (9.2.12)$$

We see that  $\max_{j \in \{1, \dots, m\}} |\widehat{F}_n(x_j) - F(x_j)| \xrightarrow{\text{a.s.}} 0$  due to pointwise convergence. That is for each  $m$ , there exists an  $n^*$  such that  $\Pr(\max_{j \in \{1, \dots, m\}} |\widehat{F}_n(x_j) - F(x_j)| = 0) = 1$  for  $n \geq n^*$ . Then, also for  $n \geq n^*$ :

$$\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \leq \frac{1}{m} \quad (9.2.13)$$

almost surely. □

### 9.2.3 Dvoretzky-Kiefer-Wolfowitz Inequality [113]

Let  $X_1, \dots, X_n$  denote an i.i.d. sample from a distribution with CDF  $F(x)$  and let  $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{X_i \leq x}$  denote the usual empirical distribution function indexed in  $n$ . The Dvoretzky-Kiefer-Wolfowitz inequality states that

$$\Pr \left( \sup_{x \in \mathbb{R}} \left\{ \sqrt{n} \left| \widehat{F}_n(x) - F(x) \right| \right\} > \epsilon \right) \leq 2e^{-2\epsilon^2} \quad (9.2.14)$$

for all  $\epsilon > 0$ . To provide some intuition behind this result, first recognise that for any  $x$  such that  $F(x) \in (0, 1)$ , the Central Limit Theorem gives that

$$\sqrt{n} (\widehat{F}_n(x) - F(x)) \xrightarrow{\text{d}} \mathcal{N}(0, F(x)(1 - F(x))) \quad (9.2.15)$$

since  $\widehat{F}_n(x)$  is essentially the sample proportion of observations not greater than  $x$ , which we can compute to have variance  $\frac{F(x)(1 - F(x))}{n}$  (e.g. from the binomial distribution). Hence  $\widehat{F}_n(x)$  will be approximately normal for large  $n$ . However, recall that a two-sided concentration inequality for a normal random variable  $Y \sim \mathcal{N}(\mu, \sigma^2)$  is

$$\Pr(|Y - \mu| > \epsilon) \leq 2 \exp \left( -\frac{\epsilon^2}{2\sigma^2} \right) \quad (9.2.16)$$

for all  $\epsilon > 0$ . It turns out that if we apply the approximately normal  $\widehat{F}_n(x)$  to this concentration inequality, we will recover the Dvoretzky-Kiefer-Wolfowitz inequality. Note that  $F(x)(1 - F(x))$  is maximised at  $F(x) = 1/2$ , with a maximum of  $1/4$  (essentially Popoviciu's inequality). Hence

$$\text{Var}(\widehat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n} \quad (9.2.17)$$

$$\leq \frac{1}{4n} \quad (9.2.18)$$

Treating  $\widehat{F}_n(x)$  as normal and plugging it into the concentration inequality gives for any  $x \in \mathbb{R}$ :

$$\Pr \left( \left| \widehat{F}_n(x) - F(x) \right| > \epsilon \right) \leq 2 \exp \left( -\frac{\epsilon^2}{2 \text{Var}(\widehat{F}_n(x))} \right) \quad (9.2.19)$$

Taking the worst-case supremum over  $x$  on either side separately:

$$\Pr \left( \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right| > \epsilon \right) \leq 2 \exp \left( -\frac{\epsilon^2}{2 \sup_{x' \in \mathbb{R}} \text{Var}(\widehat{F}_n(x'))} \right) \quad (9.2.20)$$

$$= 2 \exp \left( -\frac{\epsilon^2}{2/(4n)} \right) \quad (9.2.21)$$

$$= 2e^{-2n\epsilon^2} \quad (9.2.22)$$

Applying the substitution  $\epsilon = \sqrt{n}\varepsilon$  then yields the Dvoretzky-Kiefer-Wolfowitz inequality as claimed:

$$\Pr \left( \sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right| > \epsilon \right) \leq 2e^{-2\epsilon^2} \quad (9.2.23)$$

### 9.2.4 Kolmogorov-Smirnov Distance

The Kolmogorov-Smirnov distance between two distributions with CDFs  $F(x)$  and  $G(x)$  is defined as

$$d = \sup_{x \in \mathbb{R}} |G(x) - F(x)| \quad (9.2.24)$$

That is, the size of the largest ‘gap’ between a graph of the two CDFs.

### Kolmogorov-Smirnov Statistic

Let  $F_n(x)$  be the empirical CDF of a sample from a distribution with CDF  $F(x)$ . The Kolmogorov-Smirnov statistic is defined as the Kolmogorov-Smirnov distance between  $F_n(x)$  and  $F(x)$ :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \quad (9.2.25)$$

### Kolmogorov-Smirnov Goodness-of-Fit Test [46, 66]

## 9.3 Order Statistics

### 9.3.1 Distribution of a Single Order Statistic

Let  $X_1, \dots, X_n$  be i.i.d. random variables with cumulative distribution function  $F(x)$ . If a sample of these random variables are arranged in ascending order denoted as follows:

$$X_{(1)} \leq \dots \leq X_{(k)} \leq \dots \leq X_{(n)} \quad (9.3.1)$$

then we call  $X_{(k)}$  the  $k^{\text{th}}$  order statistic. The cumulative distribution of  $X_{(k)}$ , denoted  $F_{(k)}$ , can be expressed in terms of  $F(x)$ . By definition,

$$F_{(k)}(x) = \Pr(X_{(r)} \leq x) \quad (9.3.2)$$

which can be rewritten as

$$F_{(k)}(x) = \Pr(X_{(1)} \leq \dots \leq X_{(k)} \leq x) \quad (9.3.3)$$

This is the probability that at least  $k$  out of the  $X_1, \dots, X_n$  are less than or equal to  $x$ , which can be formulated using the upper cumulative binomial probability with ‘success’ probability  $F(x)$ .

$$F_{(k)}(x) = \sum_{i=k}^n \binom{n}{i} F(x)^i [1 - F(x)]^{n-i} \quad (9.3.4)$$

or a lower cumulative binomial probability with success probability  $1 - F(x)$ , by making the change of variables  $j = n - i$ :

$$F_{(k)}(x) = \sum_{j=0}^{n-k} \binom{n}{j} [1 - F(x)]^j F(x)^{n-j} \quad (9.3.5)$$

Using the alternative form of the cumulative binomial distribution in terms of the regularised incomplete Beta function, this gives

$$F_{(k)}(x) = \frac{1}{B(k, n - k + 1)} \int_0^{F(x)} t^{k-1} (1 - t)^{n-k} dt \quad (9.3.6)$$

If  $X$  is continuous with probability density function  $f(x)$ , this form can be used to derive the probability density function of the  $k^{\text{th}}$  order statistic using the Fundamental Theorem of Calculus:

$$f_{(k)}(x) = \frac{dF_{(k)}(x)}{dx} \quad (9.3.7)$$

$$= \frac{1}{B(k, n-k+1)} \frac{d}{dx} \int_0^{F(x)} t^{k-1} (1-t)^{n-k} dt \quad (9.3.8)$$

$$= \frac{1}{B(k, n-k+1)} \frac{dF_{(k)}(x)}{dx} \frac{d}{dF_{(k)}(x)} \int_0^{F(x)} t^{k-1} (1-t)^{n-k} dt \quad (9.3.9)$$

$$= \frac{1}{B(k, n-k+1)} f(x) F(x)^{k-1} [1 - F(x)]^{n-k} \quad (9.3.10)$$

$$= \frac{n!}{(k-1)! (n-k)!} F(x)^{k-1} [1 - F(x)]^{n-k} f(x) \quad (9.3.11)$$

$$= k \binom{n}{k} F(x)^{k-1} [1 - F(x)]^{n-k} f(x) \quad (9.3.12)$$

As special cases, we have for the maximum order statistic:

$$f_{(n)}(x) = nF(x)^{n-1} f(x) \quad (9.3.13)$$

$$F_{(n)}(x) = F(x)^n \quad (9.3.14)$$

and for the minimum order statistic:

$$f_{(1)}(x) = n [1 - F(x)]^{n-1} f(x) \quad (9.3.15)$$

$$F_{(1)}(x) = 1 - [1 - F(x)]^n \quad (9.3.16)$$

where the latter cumulative distribution can be derived from the upper tail of the binomial distribution.

### Order Statistics of Uniform Distribution

If a sample  $X_1, \dots, X_n$  are i.i.d. uniform distributed on  $(0, 1)$ , then the cumulative distribution  $F(x) = x$  for  $0 < x < 1$  and the density function  $f(x) = 1$  for  $0 < x < 1$ . Applying the expression for the  $k^{\text{th}}$  order statistic:

$$f_{(k)}(x) = \frac{n!}{(k-1)! (n-k)!} x^{k-1} (1-x)^{n-k} \quad (9.3.17)$$

$$= \frac{x^{k-1} (1-x)^{n-k}}{B(k, n-k+1)} \quad (9.3.18)$$

on support  $(0, 1)$ . This is the same form as the Beta distribution. Hence the  $k^{\text{th}}$  order statistic of uniform random variables is Beta-distributed with parameters  $k$  and  $n - k + 1$ . Hence the mean of the  $k^{\text{th}}$  order statistic is

$$\mathbb{E}[X_{(k)}] = \frac{k}{n+1} \quad (9.3.19)$$

### Extreme Order Statistics of the Gaussian Distribution

For a standard Gaussian sample, the density of the maximum order statistic in terms of the standard Gaussian CDF and PDF is

$$f_{(n)}(x) = n\Phi(x)^{n-1} \phi(x) \quad (9.3.20)$$

Note that this resembles a generalised skew normal distribution with natural number parameter  $n - 1$  and shape parameter  $\lambda = 1$ .

### 9.3.2 Joint Distribution of Order Statistics

For an i.i.d. random sample  $X_1, \dots, X_n$  with cumulative distribution  $F(x)$  and probability density function  $f(x)$ , denote the order statistics

$$X_{(1)} \leq \cdots \leq X_{(n)} \quad (9.3.21)$$

let some  $r, s$  satisfy  $1 \leq r < s \leq n$  such that

$$X_{(1)} \leq \cdots \leq X_{(r)} \leq X_{(s)} \leq \cdots \leq X_{(n)} \quad (9.3.22)$$

The joint distribution of  $X_{(r)}, X_{(s)}$  can be derived as follows. We begin with the cumulative distribution:

$$F_{(r)(s)}(x, y) = \Pr(X_{(r)} \leq x, X_{(s)} \leq y) \quad (9.3.23)$$

For  $x < y$ , this can be stated as the probability that there are at least  $r$  order statistics  $X_{(i)}$  less than or equal to  $x$ , and at least  $s$  order statistics  $X_{(j)}$  less than or equal to  $y$ . This probability can be computed with the summation:

$$F_{(r)(s)}(x, y) = \sum_{j=s}^n \sum_{i=r}^j \Pr(|\{X : X_{(k)} \leq x\}| = i, |\{X : X_{(k)} \leq y\}| = j) \quad (9.3.24)$$

where each summand is the probability that there are exactly  $i$  order statistics less than or equal to  $x$  and exactly  $j$  order statistics less than or equal to  $y$ . Then this probability can be written as

$$F_{(r)(s)}(x, y) = \sum_{j=s}^n \sum_{i=r}^j \frac{n!}{i!(j-i)!(n-j)!} [F(x)]^i [F(y) - F(x)]^{j-i} [1 - F(y)]^{n-j} \quad (9.3.25)$$

because in the summand,  $[F(x)]^i$  is the probability that  $i$  particular values in the sample fall less than or equal to  $x$ ,  $[F(y) - F(x)]^{j-i}$  is the probability that  $j-i$  particular samples fall in between  $x$  and  $y$ , and  $[1 - F(y)]^{n-j}$  is the probability that  $n-j$  particular samples fall greater than  $y$ . The coefficient  $\frac{n!}{i!(j-i)!(n-j)!}$  is the corresponding multinomial coefficient for such a combination. For the case  $x \geq y$ , if  $X_{(s)} \leq y$  then we have  $X_{(r)} \leq x$  automatically satisfied since  $X_{(r)} \leq X_{(s)} \leq y \leq x$ . Therefore we only need to consider the probability of  $X_{(s)} \leq y$ . This becomes the cumulative distribution of a single order statistic:

$$F_{(r)(s)}(x, y) = F_{(s)}(y) \quad (9.3.26)$$

$$= \sum_{i=s}^n \binom{n}{i} F(y)^i [1 - F(y)]^{n-i} \quad (9.3.27)$$

The form of the cumulative joint distribution generalises to  $k$  order statistics. Suppose we have  $1 \leq n_1 < \cdots < n_k \leq n$ . Then the cumulative distribution of the joint distribution of  $X_{(n_1)}, \dots, X_{(n_k)}$  is

$$\begin{aligned} & F_{(n_1)\dots(n_k)}(x_{(n_1)}, \dots, x_{(n_k)}) \\ &= \sum_{i_k=n_k}^n \sum_{i_{k-1}=n_{k-1}}^{i_k} \cdots \sum_{i_1=n_1}^{i_2} \frac{n!}{i_1!(i_2-i_1)!\times\cdots\times(n-i_k)!} [F(x_{(n_1)})]^{i_1} [F(x_{(n_2)}) - F(x_{(n_1)})]^{i_2-i_1} \\ & \quad \times \cdots \times [1 - F(x_{(n_k)})]^{n-i_k} \end{aligned} \quad (9.3.28)$$

for the case  $x_{(n_1)} < \dots < x_{(n_k)}$ . Note that if we do not have  $x_{(1)} \leq \dots \leq x_{(k)}$ , then the joint cumulative distribution can be evaluated by:

$$F_{(n_1)\dots(n_k)}(x_{(n_1)}, \dots, x_{(n_k)}) = F_{(n_1)\dots(n_k)}(x_{(n_1)}^*, \dots, x_{(n_k)}^*) \quad (9.3.29)$$

where

$$x_{(n_k)}^* = x_{(n_k)} \quad (9.3.30)$$

$$x_{(n_{k-1})}^* = \min \{x_{(n_{k-1})}, x_{(n_k)}^*\} \quad (9.3.31)$$

$$\vdots \quad (9.3.32)$$

$$x_{(n_1)}^* = \min \{x_{(n_1)}, x_{(n_2)}^*\} \quad (9.3.33)$$

since  $(x_{(n_1)}^*, \dots, x_{(n_k)}^*)$  satisfies  $x_{(1)}^* \leq \dots \leq x_{(k)}^*$  and the joint probability density for the region between  $(x_{(n_1)}^*, \dots, x_{(n_k)}^*)$  and  $(x_{(n_1)}, \dots, x_{(n_k)})$  is zero.

### Joint Density of Order Statistics [50]

We can also derive the joint probability density, starting with the joint probability density of two order statistics. This can be obtained from differentiating the cumulative distribution, however instead here we derive it starting from the equation

$$f_{(r)(s)}(x, y) dx dy = \Pr(x < X_{(r)} \leq x + dx, y < X_{(s)} \leq y + dy) \quad (9.3.34)$$

Note that this requires exactly  $r - 1$  order statistics to be below  $x$ , exactly  $s - r - 1$  order statistics to be between  $x$  and  $y$ , and exactly  $n - s$  order statistics to be above  $y$ . This can be computed using a multinomial probability as with deriving the cumulative distribution:

$$\begin{aligned} f_{(r)(s)}(x, y) dx dy \\ = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} [F(x)]^{r-1} f(x) dx [F(y) - F(x)]^{s-r-1} f(y) dy [1 - F(y)]^{n-s} \end{aligned} \quad (9.3.35)$$

Hence

$$f_{(r)(s)}(x, y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} [F(x)]^{r-1} f(x) [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s} \quad (9.3.36)$$

which is valid for  $x \leq y$ , otherwise the density  $f_{(r)(s)}(x, y) = 0$  when  $x > y$ . This can be generalised to  $k$  order statistics:

$$f_{(n_1)\dots(n_k)}(x_{(n_1)}, \dots, x_{(n_k)}) = \begin{cases} \frac{n!}{(n_1-1)!(n_2-n_1-1)!\times\cdots\times(n-n_k)!} \\ \quad \times [F(x_{(n_1)})]^{n_1-1} f(x_{(n_1)}) & x_{(n_1)} \leq \dots \leq x_{(n_k)} \\ \quad \times [F(x_{(n_2)}) - F(x_{(n_1)})]^{n_2-n_1-1} f(x_{(n_2)}) \\ \quad \times \cdots \times f(x_{(n_k)}) [1 - F(x_{(n_k)})]^{n-n_k}, & \text{otherwise} \\ 0, & \text{otherwise} \end{cases} \quad (9.3.37)$$

This can be more compactly written if we define  $x_{(n_0)} := -\infty$ ,  $x_{(n_{k+1})} := \infty$ ,  $n_0 := 0$ ,  $n_{k+1} = n + 1$ , and use indicator functions for  $x_{(n_1)} \leq \dots \leq x_{(n_k)}$ . We have

$$f_{(n_1)\dots(n_k)}(x_{(n_1)}, \dots, x_{(n_k)}) = n! \left( \prod_{j=1}^k f(x_{(n_j)}) \right) \prod_{j=0}^k \left[ \frac{\left( F(x_{(n_{j+1})}) - F(x_{(n_j)}) \right)^{n_{j+1}-n_j-1}}{(n_{j+1}-n_j-1)!} \right] \mathbb{I}_{\{x_{(n_1)} \leq \dots \leq x_{(n_k)}\}} \quad (9.3.38)$$

Note that we have a special case when  $n = k$ , which becomes

$$f_{(1)\dots(n)}(x_{(1)}, \dots, x_{(n)}) = n! f(x_{(1)}) \times \dots \times f(x_{(n)}) \mathbb{I}_{\{x_{(1)} \leq \dots \leq x_{(k)}\}} \quad (9.3.39)$$

This can be derived intuitively, because the joint distribution of the sample is

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = f(x_1) \times \dots \times f(x_n) \quad (9.3.40)$$

and for every possible  $x_{(1)} \leq \dots \leq x_{(k)}$ , there are  $n!$  equally likely events which could have led to it.

### 9.3.3 Conditional Distribution of Order Statistics

For some  $r, s$  satisfying  $1 \leq r < s \leq n$ , we derive the joint conditional PDF of the order statistics  $X_{(r+1)}, \dots, X_{(s-1)}$  given  $X_{(1)} = x_{(1)}, \dots, X_{(r)} = x_{(r)}, X_{(s)} = x_{(s)}, \dots, X_{(n)} = x_{(n)}$  (i.e. conditional on the order statistics outside  $r+1, \dots, s-1$ ). From above, the expression for the joint distribution for the entire order statistics (ignoring support) is:

$$f_{(1)\dots(n)}(x_{(1)}, \dots, x_{(n)}) = n! f(x_{(1)}) \times \dots \times f(x_{(n)}) \quad (9.3.41)$$

By applying the formula for the general form, the joint distribution of order statistics  $1, \dots, r, s, \dots, n$  is:

$$f_{(1)\dots(r)(s)\dots(n)}(x_{(1)}, \dots, x_{(r)}, x_{(s)}, \dots, x_{(n)}) = \frac{n!}{(s-r-1)!} f(x_{(1)}) \dots f(x_{(r)}) \times [F(x_{(s)}) - F(x_{(r)})]^{s-r-1} f(x_{(s)}) \dots f(x_{(n)}) \quad (9.3.42)$$

Taking their quotient gives the conditional distribution, thus doing so we get some cancellations:

$$f_{(r+1)\dots(s-1)|X_{(i)}=x_{(i)}, i \leq r, i \geq s}(x_{(r+1)}, \dots, x_{(s-1)}) = \frac{f_{(1)\dots(n)}(x_{(1)}, \dots, x_{(n)})}{f_{(1)\dots(r)(s)\dots(n)}(x_{(1)}, \dots, x_{(r)}, x_{(s)}, \dots, x_{(n)})} \quad (9.3.43)$$

$$= \frac{(s-r-1)!}{[F(x_{(s)}) - F(x_{(r)})]^{s-r-1}} f(x_{(r+1)}) \dots f(x_{(s-1)}) \quad (9.3.44)$$

$$= (s-r-1)! \prod_{i=r+1}^{s-1} \frac{f(x_{(i)})}{F(x_{(s)}) - F(x_{(r)})} \quad (9.3.45)$$

when  $x_{(1)} \leq \dots \leq x_{(n)}$ . We see from this form that the conditional distribution is the same as the joint distribution of order statistics for a sample of size  $s-r-1$ , if the parent distribution were the original distribution truncated between  $x_{(r)}$  and  $x_{(s)}$ . Since  $x_{(1)}, \dots, x_{(r-1)}, \dots, x_{(s+1)}, x_{(n)}$  also do not appear in the density, we deduce that  $X_{(r+1)}, \dots, X_{(s-1)}$  are conditionally independent with  $X_{(1)}, \dots, X_{(r-1)}, X_{(s+1)}, X_{(n)}$  given  $X_{(r)}, X_{(s)}$ . Therefore

$$f_{(r+1)\dots(s-1)|X_{(i)}=x_{(i)}, i \leq r, i \geq s}(x_{(r+1)}, \dots, x_{(s-1)}) = f_{(r+1)\dots(s-1)|X_{(r)}=x_{(r)}, X_{(r)}=x_{(r)}}(x_{(r+1)}, \dots, x_{(s-1)}) \quad (9.3.46)$$

This conforms well with intuition, since we only need to know  $x_{(r)}$  and  $x_{(s)}$  to determine the truncated distribution. In a similar way, we can also write the conditional distribution of the lower and upper order statistics given  $X_{(s)}$  and  $X_{(r)}$  respectively:

$$f_{(1)\dots(s-1)|X_{(s)}=x_{(s)}}(x_{(1)}, \dots, x_{(s-1)}) = (s-1)! \prod_{i=1}^{s-1} \frac{f(x_{(i)})}{F(x_{(s)})} \quad (9.3.47)$$

$$f_{(r+1)\dots(n)|X_{(r)}=x_{(r)}}(x_{(r+1)}, \dots, x_{(n)}) = (n-r)! \prod_{i=r+1}^n \frac{f(x_{(i)})}{1 - F(x_{(r)})} \quad (9.3.48)$$

on their respective supports, where the truncation interpretation similarly applies (with one-sided truncation rather than two). We can also obtain the marginal conditional distributions of the proceeding/succeeding order statistics. The distribution of  $f_{(r+1)|X_{(r)}=x}(y)$  is the same distribution as the minimum order statistic of a sample of size  $n-r$  from the truncated parent distribution on the right of  $x$  (which has PDF and CDF  $\frac{f(y)}{1 - F(x)}$  and  $\frac{1}{1 - F(x)} \int_x^y f(t) dt = \frac{F(y) - F(x)}{1 - F(x)}$  respectively):

$$f_{(r+1)|X_{(r)}=x}(y) = (n-r) \left[ 1 - \frac{F(y) - F(x)}{1 - F(x)} \right]^{n-r-1} \frac{f(y)}{1 - F(x)} \quad (9.3.49)$$

$$= (n-r) \left[ \frac{1 - F(y)}{1 - F(x)} \right]^{n-r-1} \frac{f(y)}{1 - F(x)} \quad (9.3.50)$$

on support  $y > x$ . Analogously, the distribution of  $f_{(s-1)|X_{(s)}=x}(y)$  is the same distribution as the maximum order statistic of a sample of size  $s-1$  from the truncated parent distribution on the left of  $x$  (which has PDF and CDF  $\frac{f(y)}{F(x)}$  and  $\frac{1}{F(x)} \int_{-\infty}^y f(t) dt = \frac{F(y)}{F(x)}$  respectively):

$$f_{(s-1)|X_{(s)}=x}(y) = (s-1) \left[ \frac{F(y)}{F(x)} \right]^{s-2} \frac{f(y)}{F(x)} \quad (9.3.51)$$

on support  $y < x$ . The conditional independence also shows that the sequences  $\{X_{(1)}, X_{(2)}, \dots\}$  and  $\{X_{(n)}, X_{(n-1)}, \dots\}$  are Markov processes, with transition densities given above.

### 9.3.4 Spacings of Order Statistics

For an i.i.d. sample  $X_1, \dots, X_n$  denote the  $i^{\text{th}}$  order statistic by  $X_{i:n}$ . The spacing between the  $i^{\text{th}}$  and  $(i+1)^{\text{th}}$  order statistics is defined by  $S_{i:n} = X_{(i+1):n} - X_{i:n}$ .

#### Expected Spacings of Subsequent Order Statistics

To find the expected spacing  $\mathbb{E}[S_{i:n}]$ , note that using the cumulative distribution functions of the order statistics, their expectations can be expressed as

$$\mathbb{E}[X_{(i+1):n}] = \int_0^\infty [1 - F_{(i+1):n}(x)] dx - \int_{-\infty}^0 F_{(i+1):n}(x) dx \quad (9.3.52)$$

$$\mathbb{E}[X_{i:n}] = \int_0^\infty [1 - F_{i:n}(x)] dx - \int_{-\infty}^0 F_{i:n}(x) dx \quad (9.3.53)$$

where  $F_{(i+1):n}(x)$  and  $F_{i:n}(x)$  denote the cumulative distribution functions of  $X_{(i+1):n}$  and  $X_{i:n}$  respectively. Then

$$\mathbb{E}[S_{i:n}] = \mathbb{E}[X_{(i+1):n} - X_{i:n}] \quad (9.3.54)$$

$$= \int_0^\infty [1 - F_{(i+1):n}(x) - 1 + F_{i:n}(x)] dx - \int_{-\infty}^0 (F_{(i+1):n}(x) - F_{i:n}(x)) dx \quad (9.3.55)$$

$$= \int_0^\infty [F_{i:n}(x) - F_{(i+1):n}(x)] dx + \int_{-\infty}^0 [F_{i:n}(x) - F_{(i+1):n}(x)] dx \quad (9.3.56)$$

$$= \int_{-\infty}^\infty [F_{i:n}(x) - F_{(i+1):n}(x)] dx \quad (9.3.57)$$

Using the binomial form of the CDFs where  $F(x)$  is the CDF of  $X_i$ :

$$F_{i:n}(x) - F_{(i+1):n}(x) = \sum_{r=i}^n \binom{n}{r} F(x)^r [1 - F(x)]^{n-r} - \sum_{r=i+1}^n \binom{n}{r} F(x)^r [1 - F(x)]^{n-r} \quad (9.3.58)$$

$$= \binom{n}{i} F(x)^i [1 - F(x)]^{n-i} \quad (9.3.59)$$

Hence

$$\mathbb{E}[S_{i:n}] = \binom{n}{i} \int_{-\infty}^\infty F(x)^i [1 - F(x)]^{n-i} dx \quad (9.3.60)$$

### 9.3.5 Fisher–Tippett–Gnedenko Theorem [50]

Consider the maximum order statistic, denoted  $X_{n:n}$ , in an i.i.d. sample of size  $n$  from a population with CDF  $F(x)$ . The distribution of said statistic is  $[F(x)]^n$ . The Fisher-Tippett-Gnedenko theorem asserts that if the distribution  $[F(x)]^n$  after appropriate scaling and shifting (i.e. we are able to find some constants  $a_n > 0$ ,  $b_n$  and take  $[F(a_n x + b_n)]^n$ ) converges in distribution to some distribution  $G(x)$ , then  $G(x)$  can only be a distribution from one of the following three families:

- Gumbel distribution with form  $G_1(x) = \exp(-e^{-x})$
- Fréchet distribution with form  $G_2(x) = \exp(-x^{-\alpha})$  where  $\alpha > 0$
- Reversed Weibull distribution with form

$$G_3(x) = \begin{cases} \exp[-(-x)^k], & x \leq 0 \\ 1, & x > 0 \end{cases} \quad (9.3.61)$$

where  $k > 0$ .

These distributions are sometimes referred to as the extreme value distributions (Type I, Type II and Type III respectively). This result can be reasoned by considering the maximum across  $m$  samples of size  $n$  and comparing against the maximum in a single sample of size  $mn$ . If a limiting distribution  $G(x)$  exists, then the standardised distributions of both should converge to  $G(x)$  as  $n \rightarrow \infty$ . The distribution of the former can also be written as  $[G(x)]^m$ . It follows that there should exist constants  $a_m > 0$ ,  $b_m$  such that

$$[G(a_m x + b_m)]^m = G(x) \quad (9.3.62)$$

We verify this by showing for each of the three extreme value distributions above that the CDF to an exponent  $m > 0$  gives another CDF of the same family up to location and scale parameters. For the Gumbel family:

$$[G_1(x)]^m = [\exp(-e^{-x})]^m \quad (9.3.63)$$

$$= \exp(-me^{-x}) \quad (9.3.64)$$

$$= \exp(-e^{-(x-\log m)}) \quad (9.3.65)$$

$$= G_1(x - \log m) \quad (9.3.66)$$

For the Fréchet family:

$$[G_2(x)]^m = [\exp(-x^{-\alpha})]^m \quad (9.3.67)$$

$$= \exp(-mx^{-\alpha}) \quad (9.3.68)$$

$$= \exp\left[-\left(m^{-1/\alpha}x\right)^{-\alpha}\right] \quad (9.3.69)$$

$$= G_2\left(m^{-1/\alpha}x\right) \quad (9.3.70)$$

For the reversed Weibull family, considering the form when  $x \leq 0$ :

$$[G_3(x)]^m = \left[\exp\left[-(-x)^k\right]\right]^m \quad (9.3.71)$$

$$= \exp\left[-m(-x)^k\right] \quad (9.3.72)$$

$$= \exp\left[-\left(-m^{1/k}x\right)^k\right] \quad (9.3.73)$$

$$= G_3\left(m^{1/k}x\right) \quad (9.3.74)$$

### 9.3.6 Central Limit Theorem for Order Statistics

#### Asymptotic Distribution of a Single Central Order Statistic

**Theorem 9.1** ([4]). *For some  $0 < p < 1$ , let index  $i = \lfloor np \rfloor + 1$ . Then the order statistic  $X_{i:n}$  (which we call a central order statistic) is representative for the  $p^{\text{th}}$  sample quantile of  $X$ , the latter which has PDF  $f(x)$  and CDF  $F(x)$ . If  $f(F^{-1}(p))$  is finite, positive and continuous, then*

$$\sqrt{n}(X_{i:n} - F^{-1}(p)) \xrightarrow{\text{d}} \mathcal{N}\left(0, \frac{p(1-p)}{[f(F^{-1}(p))]^2}\right) \quad (9.3.75)$$

as  $n \rightarrow \infty$ .

*Proof.* We first show the result for a uniform random variable  $U$  on  $(0, 1)$ . Recall that  $U_{i:n}$  is a Beta( $i, n-i+1$ ) random variable. Combining this with the fact that the sum of independent exponential random variables (with same rate parameter  $\lambda$ ) is Erlang distributed (and hence Gamma distributed), and the distribution relationship between the Gamma and Beta distribution, we can write  $U_{i:n}$  as

$$U_{i:n} = \frac{A_n}{A_n + B_n} \quad (9.3.76)$$

where

$$A_n = \sum_{j=1}^i Z_j \quad (9.3.77)$$

$$B_n = \sum_{j=i+1}^{n+1} Z_j \quad (9.3.78)$$

where  $Z_j$  are independent exponential random variables with rate parameter 1. By the central limit theorem, with  $\mathbb{E}[A_n] = i$ ,

$$\frac{A_n - i}{\sqrt{i}} \xrightarrow{\text{d}} \mathcal{N}(0, 1) \quad (9.3.79)$$

Since  $i/n \rightarrow p$ , when multiplying out by  $\sqrt{i/n}$  gives

$$\frac{A_n - i}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, p) \quad (9.3.80)$$

Similarly, with  $\mathbb{E}[B_n] = n - i + 1$ ,

$$\frac{B_n - (n - i + 1)}{\sqrt{n - i + 1}} \xrightarrow{d} \mathcal{N}(0, 1) \quad (9.3.81)$$

and since  $(n - i + 1)/n \rightarrow 1 - p$ , multiplying out by  $\sqrt{(n - i + 1)/n}$  gives

$$\frac{B_n - (n - i + 1)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1 - p) \quad (9.3.82)$$

Consider the mixture

$$C_n = (1 - p) \frac{A_n - i}{\sqrt{n}} - p \frac{B_n - (n - i + 1)}{\sqrt{n - i + 1}} \quad (9.3.83)$$

Since by independence

$$\text{Var}(C_n) = (1 - p)^2 \text{Var}\left(\frac{A_n - i}{\sqrt{n}}\right) + p^2 \text{Var}\left(\frac{B_n - (n - i + 1)}{\sqrt{n - i + 1}}\right) \quad (9.3.84)$$

and asymptotically, for  $C_n \xrightarrow{d} C$ :

$$\text{Var}(C) = (1 - p)^2 p + p^2 (1 - p) \quad (9.3.85)$$

$$= p \left[ (1 - p)^2 + p(1 - p) \right] \quad (9.3.86)$$

$$= p(1 - 2p + p^2 + p - p^2) \quad (9.3.87)$$

$$= p(1 - p) \quad (9.3.88)$$

then

$$C_n \xrightarrow{d} \mathcal{N}(0, p(1 - p)) \quad (9.3.89)$$

Now express  $\sqrt{n}(U_{i:n} - p)$  as

$$\sqrt{n}(U_{i:n} - p) = \sqrt{n} \left( \frac{A_n}{A_n + B_n} - p \right) \quad (9.3.90)$$

$$= \sqrt{n} \left( \frac{A_n - pA_n - pB_n}{A_n + B_n} \right) \quad (9.3.91)$$

$$= \sqrt{n} \left( \frac{(1 - p)A_n - pB_n}{A_n + B_n} \right) \quad (9.3.92)$$

$$= \sqrt{n} \left[ \frac{(1 - p)A_n - i(1 - p) - pB_n + p(n - i + 1) + i(1 - p) - p(n - i + 1)}{A_n + B_n} \right] \quad (9.3.93)$$

$$= \sqrt{n} \left[ \frac{(1 - p)(A_n - i) - p[B_n - (n - i + 1)] + i - pi - np + pi - p}{A_n + B_n} \right] \quad (9.3.94)$$

$$= \sqrt{n} \left( \frac{\sqrt{n}C_n + i - np - p}{A_n + B_n} \right) \quad (9.3.95)$$

$$= \frac{nC_n + \sqrt{n}(i - np - p)}{A_n + B_n} \quad (9.3.96)$$

$$= \frac{C_n + (i - np - p) / \sqrt{n}}{(A_n + B_n) / n} \quad (9.3.97)$$

We have  $(i - np - p) / \sqrt{n} \rightarrow 0$ , and  $(A_n + B_n) / n = \frac{1}{n} \sum_{j=1}^n Z_j + \frac{Z}{n} \xrightarrow{\text{P}} 1$  due to the weak law of large numbers. Then by Slutsky's theorem:

$$\sqrt{n}(U_{i:n} - p) \xrightarrow{\text{d}} \mathcal{N}(0, p(1-p)) \quad (9.3.98)$$

which satisfies the theorem since  $F^{-1}(p) = p$  and  $f(x) = 1$  for the uniform distribution. To extend this to arbitrary distributions, we use the inverse transform  $X_{i:n} = F^{-1}(U_{i:n})$ . Applying Taylor's theorem about  $p$  (and specifically, using the Lagrange form of the remainder), we have

$$F^{-1}(U_{i:n}) = F^{-1}(p) + \frac{U_{i:n} - p}{f(F^{-1}(D_n))} \quad (9.3.99)$$

where  $D_n$  is a random variable that is between  $U_{i:n}$  and  $p$ , and we can show that  $\frac{dF^{-1}(p)}{dp} = \frac{1}{f(F^{-1}(p))}$  using the inverse function theorem. After rearranging, we can write this relationship expressed in terms of  $X_{i:n}$  as

$$\sqrt{n}(X_{i:n} - F^{-1}(p)) = \sqrt{n} \frac{U_{i:n} - p}{f(F^{-1}(D_n))} \quad (9.3.100)$$

It is clear that for the uniform distribution,  $D_n \xrightarrow{\text{P}} p$  so using the continuous mapping theorem (under the continuity assumption),  $f(F^{-1}(D_n)) \xrightarrow{\text{P}} f(F^{-1}(p))$ . By again applying Slutsky's theorem along with asymptotic normality of  $\sqrt{n}(U_{i:n} - p)$ , we therefore show that the central order statistic  $X_{i:n}$  is asymptotically normal with:

$$\sqrt{n}(X_{i:n} - F^{-1}(p)) \xrightarrow{\text{d}} \mathcal{N}\left(0, \frac{p(1-p)}{[f(F^{-1}(p))]^2}\right) \quad (9.3.101)$$

□

### Asymptotic Joint Distribution of Central Order Statistics

**Theorem 9.2** ([50]). *With some  $0 < p_1 < \dots < p_m < 1$ , let indices  $i_r = \lfloor np_r \rfloor + 1$  for  $1 \leq r \leq m$ . Then  $X_{i_1:n}, \dots, X_{i_m:n}$  are the central order statistics of  $X$ , the latter which has PDF  $f(x)$  and CDF  $F(x)$ . If  $f(F^{-1}(p))$  is finite, positive and continuous, then*

$$\begin{aligned} & \sqrt{n} \left( \begin{bmatrix} X_{i_1:n} \\ X_{i_2:n} \\ \vdots \\ X_{i_m:n} \end{bmatrix} - \begin{bmatrix} F^{-1}(p_1) \\ F^{-1}(p_2) \\ \vdots \\ F^{-1}(p_m) \end{bmatrix} \right) \\ & \xrightarrow{\text{d}} \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} \frac{p_1(1-p_1)}{[f(F^{-1}(p_1))]^2} & \frac{p_1(1-p_2)}{f(F^{-1}(p_1))f(F^{-1}(p_2))} & \cdots & \frac{p_1(1-p_m)}{f(F^{-1}(p_1))f(F^{-1}(p_m))} \\ \frac{p_1(1-p_2)}{f(F^{-1}(p_1))f(F^{-1}(p_2))} & \frac{p_2(1-p_2)}{[f(F^{-1}(p_2))]^2} & \cdots & \frac{p_2(1-p_m)}{f(F^{-1}(p_2))f(F^{-1}(p_m))} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_1(1-p_m)}{f(F^{-1}(p_1))f(F^{-1}(p_m))} & \frac{p_2(1-p_m)}{f(F^{-1}(p_2))f(F^{-1}(p_m))} & \cdots & \frac{p_m(1-p_m)}{[f(F^{-1}(p_m))]^2} \end{bmatrix} \right) \end{aligned} \quad (9.3.102)$$

*Proof.* In the univariate case, it was shown that

$$X_{i:n} - F^{-1}(p) = \frac{U_{i:n} - p}{f(F^{-1}(D_n))} \quad (9.3.103)$$

where  $f(F^{-1}(D_n)) \xrightarrow{P} f(F^{-1}(p))$ . Let  $\tilde{F}_n(x)$  denote the empirical CDF of the sample  $X_1, \dots, X_n$  (i.e. the proportion of observations in the sample less than or equal to  $x$ ). By the Glivenko-Cantelli theorem,  $\tilde{F}_n(x)$  converges almost surely to  $F(x)$ . Note also that  $F(X_{i:n}) = U_{i:n}$  and  $X_{i:n} \xrightarrow{P} F^{-1}(p)$ . Putting these arguments together, we analyse the asymptotic behaviour of  $\tilde{F}_n(F^{-1}(p))$  since it will have the same asymptotic distribution as  $U_{i:n}$ . Consider two  $j, k$  with  $j < k$  (hence  $F^{-1}(p_j) \leq F^{-1}(p_k)$ ). By writing the empirical CDF as a sum of indicators

$$\tilde{F}_n(x) = \frac{1}{n} \sum_{\iota=1}^n \mathbb{I}_{X_\iota \leq x} \quad (9.3.104)$$

We can compute the covariance between  $\tilde{F}_n(F^{-1}(p_j))$  and  $\tilde{F}_n(F^{-1}(p_k))$ :

$$\begin{aligned} \text{Cov}(\tilde{F}_n(F^{-1}(p_j)), \tilde{F}_n(F^{-1}(p_k))) &= \mathbb{E} \left[ \frac{1}{n^2} \sum_{\iota=1}^n \sum_{\kappa=1}^n \mathbb{I}_{X_\iota \leq F^{-1}(p_j)} \mathbb{I}_{X_\kappa \leq F^{-1}(p_k)} \right] \\ &\quad - \mathbb{E} \left[ \frac{1}{n} \sum_{\iota=1}^n \mathbb{I}_{X_\iota \leq F^{-1}(p_j)} \right] \mathbb{E} \left[ \frac{1}{n} \sum_{\kappa=1}^n \mathbb{I}_{X_\kappa \leq F^{-1}(p_k)} \right] \end{aligned} \quad (9.3.105)$$

$$= \mathbb{E} \left[ \frac{1}{n^2} \sum_{\iota=1}^n \sum_{\kappa=1}^n \mathbb{I}_{X_\iota \leq F^{-1}(p_j)} \mathbb{I}_{X_\kappa \leq F^{-1}(p_k)} \right] - p_j p_k \quad (9.3.106)$$

$$= p_j - p_j p_k \quad (9.3.107)$$

$$= p_j(1 - p_k) \quad (9.3.108)$$

where  $\frac{1}{n^2} \sum_{\iota=1}^n \sum_{\kappa=1}^n \mathbb{I}_{X_\iota \leq F^{-1}(p_j)} \mathbb{I}_{X_\kappa \leq F^{-1}(p_k)}$  gives the proportion of pairs of observations where both observations are less than or equal to  $F^{-1}(p_j)$ , and thus is same as the proportion of observations less than or equal to  $F^{-1}(p_j)$ . Moreover, recognise that  $\tilde{F}_n(F^{-1}(p))$  is the sample mean of i.i.d. Bernoulli random variables with mean  $p$ . Applying the multivariate central limit theorem,

$$\sqrt{n} \begin{bmatrix} \tilde{F}_n(F^{-1}(p_j)) - p_j \\ \tilde{F}_n(F^{-1}(p_k)) - p_k \end{bmatrix} \xrightarrow{d} \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} p_j(1-p_j) & p_j(1-p_k) \\ p_j(1-p_k) & p_k(1-p_k) \end{bmatrix} \right) \quad (9.3.109)$$

By a scaling,

$$\sqrt{n} \begin{bmatrix} \frac{\tilde{F}_n(F^{-1}(p_j)) - p_j}{f(F^{-1}(p_j))} \\ \frac{\tilde{F}_n(F^{-1}(p_k)) - p_k}{f(F^{-1}(p_k))} \end{bmatrix} \xrightarrow{d} \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} \frac{p_j(1-p_j)}{[f(F^{-1}(p_j))]^2} & \frac{p_j(1-p_k)}{f(F^{-1}(p_j))f(F^{-1}(p_k))} \\ \frac{p_j(1-p_k)}{f(F^{-1}(p_j))f(F^{-1}(p_k))} & \frac{p_k(1-p_k)}{[f(F^{-1}(p_k))]^2} \end{bmatrix} \right) \quad (9.3.110)$$

Then it follows that for any pair  $j < k$

$$\sqrt{n} \begin{bmatrix} X_{i_j:n} - F^{-1}(p_j) \\ X_{i_k:n} - F^{-1}(p_k) \end{bmatrix} \xrightarrow{d} \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} \frac{p_j(1-p_j)}{[f(F^{-1}(p_j))]^2} & \frac{p_j(1-p_k)}{f(F^{-1}(p_j))f(F^{-1}(p_k))} \\ \frac{p_j(1-p_k)}{f(F^{-1}(p_j))f(F^{-1}(p_k))} & \frac{p_k(1-p_k)}{[f(F^{-1}(p_k))]^2} \end{bmatrix} \right) \quad (9.3.111)$$

□

### 9.3.7 L-Statistics

An L-statistic is a statistic which is a linear combination of order statistics.

## 9.4 Computational Statistics

### 9.4.1 Monte-Carlo Estimation

Suppose we wish to evaluate the multivariate integral  $\int_D g(\mathbf{x}) d\mathbf{x}$ , which may arise as some form of expectation or probability (recalling that probabilities can be expressed as the expectation of an indicator). A solution to estimate the integral is to i.i.d. sample  $\mathbf{X}_i$  uniformly on  $D$ , which by the Strong Law of Large Numbers, satisfies

$$\frac{1}{N} \sum_{i=1}^N g(\mathbf{X}_i) d\mathbf{x} \xrightarrow{\text{a.s.}} \mathbb{E}[g(\mathbf{X}_i)] \quad (9.4.1)$$

$$= \int_D \frac{1}{V} g(\mathbf{x}) d\mathbf{x} \quad (9.4.2)$$

where  $V = \int_D d\mathbf{x}$  is the integrated volume of  $D$ , thus the density of  $\mathbf{X}_i$ . Then we can approximate our desired integral by

$$\int_D g(\mathbf{x}) d\mathbf{x} \approx \frac{V}{N} \sum_{i=1}^N g(\mathbf{X}_i) d\mathbf{x} \quad (9.4.3)$$

### Monte-Carlo Probability Estimation

Suppose we want to estimate the probability of an event  $A$  via a Monte-Carlo approach. Using the fact

$$\Pr(A) = \mathbb{E}[\mathbb{I}_A] \quad (9.4.4)$$

where  $\mathbb{I}_A$  is a Bernoulli distributed indicator random variable for event  $A$ . If we are able to sample  $\mathbb{I}_{A,1}, \dots, \mathbb{I}_{A,N}$  (e.g. via a simulation), then the probability can be estimated by

$$\Pr(A) \approx \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{A,i} \quad (9.4.5)$$

### 9.4.2 Acceptance-Rejection Sampling [165, 166]

Acceptance-rejection sampling is a method for sampling from a *target density*  $f(\mathbf{x})$  (which can generally be a multivariate density), which is known (i.e. can be evaluated) only up to a multiplicative constant. This can be useful for example when the inverse CDF is not known or difficult to evaluate (so inverse transform sampling cannot be used). To perform acceptance-rejection sampling, we require a *proposal density*  $g(\mathbf{x})$  that satisfies the following conditions.

- The supports are compatible, i.e.  $g(\mathbf{x}) > 0$  whenever  $f(\mathbf{x}) > 0$ . This means that anything that can be sampled from  $f(\mathbf{x})$  must also have the ability to be sampled from  $g(\mathbf{x})$ .
- There exists a constant  $M$  such that

$$f(\mathbf{x}) \leq M g(\mathbf{x}) \quad (9.4.6)$$

for all  $\mathbf{x}$ . Note that if  $f(\mathbf{x})$  and  $g(\mathbf{x})$  share the same support, then this inequality will be trivially satisfied for all  $\mathbf{x}$  outside their supports. Also note that this inequality can be expressed in several ways such as

$$\frac{f(\mathbf{x})}{g(\mathbf{x})} \leq M \quad (9.4.7)$$

$$\frac{f(\mathbf{x})}{M g(\mathbf{x})} \leq 1 \quad (9.4.8)$$

To perform acceptance-rejection sampling, we use the following steps.

1. Generate a sample  $\mathbf{X}$  from  $g(\mathbf{x})$ , and independently sample  $U$  from the uniform distribution on  $[0, 1]$ .
2. Accept sample  $\mathbf{Y} = \mathbf{X}$  if

$$U \leq \frac{f(\mathbf{x})}{Mg(\mathbf{x})} \quad (9.4.9)$$

otherwise return to the first step and start again.

Then  $\mathbf{Y}$  will have the same density of  $f(\mathbf{x})$ . To see why this is the case, we start with the characterisation of the lower-tail probability of  $\mathbf{Y}$  in terms of the acceptance rule. We assume for simplicity that  $f(\mathbf{x})$  is univariate, however the steps are analogous (and extends naturally) for the multivariate case. The probability that  $Y \leq y$  is given by

$$\Pr(Y \leq y) = \Pr\left(X \leq y \middle| U \leq \frac{f(X)}{Mg(X)}\right) \quad (9.4.10)$$

$$= \frac{\Pr\left(X \leq y, U \leq \frac{f(X)}{Mg(X)}\right)}{\Pr\left(U \leq \frac{f(X)}{Mg(X)}\right)} \quad (9.4.11)$$

using the definition of conditional probability. Writing this in terms of integrals (and marginalising over  $f(x)$  in the denominator:

$$\Pr(Y \leq y) = \frac{\int_{-\infty}^y \int_0^{f(x)/(Mg(x))} du \cdot g(x) dx}{\int_{-\infty}^{\infty} \int_0^{f(x)/(Mg(x))} du \cdot g(x) dx} \quad (9.4.12)$$

$$= \frac{\int_{-\infty}^y \frac{f(x)}{Mg(x)} \cdot g(x) dx}{\int_{-\infty}^{\infty} \frac{f(x)}{Mg(x)} \cdot g(x) dx} \quad (9.4.13)$$

where we recall that the uniform distribution over  $[0, 1]$  has unit density on  $[0, 1]$ . Hence this is why the condition  $\frac{f(\mathbf{x})}{Mg(\mathbf{x})} \leq 1$  is required, which means that the inner integral evaluates to the upper terminal. Otherwise, we would be integrating over the region where the uniform density is zero, so we cannot perform the above step. After cancellation of the  $g(x)$  densities, we get

$$\Pr(Y \leq y) = \frac{\frac{1}{M} \int_{-\infty}^y f(x) dx}{\frac{1}{M} \int_{-\infty}^{\infty} f(x) dx} \quad (9.4.14)$$

$$= \frac{\int_{-\infty}^y f(x) dx}{1} \quad (9.4.15)$$

$$= \int_{-\infty}^y f(x) dx \quad (9.4.16)$$

which gives the definition of the CDF for density  $f(x)$ . Hence  $Y$  will also be distributed with density  $f(x)$ . We can also see why we do not require  $f(\mathbf{x})$  to integrate to one, because any normalising constant can be absorbed into  $M$  (which does not need to be a tight constant), and then gets cancelled out.

### 9.4.3 Importance Sampling [115]

Suppose Monte-Carlo sampling is used to estimate some quantity  $\mu = \mathbb{E}[H(\mathbf{X})]$  where  $\mathbf{X}$  is a random vector with density  $f(\mathbf{x})$  and  $H(\cdot)$  is a real-valued function. Then a straightforward estimator using  $N$  i.i.d. samples  $\mathbf{X}$  drawn from the density  $f(\mathbf{x})$  is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N H(\mathbf{X}_i) \quad (9.4.17)$$

However, there may be some distributions which are ‘difficult’ to sample from, such that the variance of  $\hat{\mu}$  will be impractically large. This may arise for example in rare event simulation, where  $H(\cdot)$  is an indicator function for a rare event  $A$  and  $\hat{\mu}$  is an estimator for  $\Pr(A)$ . Since  $A$  is rare, a plain Monte-Carlo sample as above may not even yield a single occurrence of  $A$ . The idea behind importance sampling is to re-weight the sampling density in hope of reducing the variance of  $\hat{\mu}$ . This is done as follows. We call  $f(\mathbf{x})$  the *target* or *nominal* density, and introduce a density  $g(\mathbf{x})$  called the *proposal* or *importance* density. The density  $g(\mathbf{x})$  must satisfy the property that  $H(\mathbf{x})f(\mathbf{x}) = 0$  wherever  $g(\mathbf{x}) = 0$ . This is so that

$$\mu = \int H(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad (9.4.18)$$

$$= \int_{\{H(\mathbf{x})f(\mathbf{x}) \neq 0\}} H(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad (9.4.19)$$

$$= \int_{\{g(\mathbf{x}) \neq 0\}} H(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} \quad (9.4.20)$$

$$= \mathbb{E}_g \left[ H(\mathbf{X}) \frac{f(\mathbf{X})}{g(\mathbf{X})} \right] \quad (9.4.21)$$

Hence we can sample from the density  $g(\mathbf{x})$  and estimate the average of the random variable  $H(\mathbf{X}) \frac{f(\mathbf{X})}{g(\mathbf{X})}$ . This gives the unbiased importance sampling estimator from  $N$  i.i.d. samples of  $\mathbf{X}$  drawn from  $g(\mathbf{x})$ :

$$\hat{\mu}' = \frac{1}{N} \sum_{i=1}^N H(\mathbf{X}_i) \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)} \quad (9.4.22)$$

By slight abuse of nomenclature, the weighting ratio  $w(\mathbf{x}) = \frac{f(\mathbf{x})}{g(\mathbf{x})}$  is called the *likelihood ratio*.

There is an optimal choice of importance density  $g^*(\mathbf{x})$  which minimises  $\text{Var}_g(\hat{\mu}')$ . It can be shown that the optimal importance density is

$$g^*(\mathbf{x}) = \frac{H(\mathbf{x}) f(\mathbf{x})}{\int H(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}} \quad (9.4.23)$$

$$= \frac{H(\mathbf{x}) f(\mathbf{x})}{\mu} \quad (9.4.24)$$

Note that this implicitly requires either  $H(\mathbf{x}) \geq 0$  or  $H(\mathbf{x}) \leq 0$  (i.e. it does not ever switch signs) to guarantee that the density  $g^*(\mathbf{x}) \geq 0$ . With this, we can show that when samples of  $\mathbf{X}$  are drawn i.i.d. from  $g^*(\mathbf{x})$ :

$$\text{Var}_{g^*}(\hat{\mu}') = \text{Var}_{g^*} \left( \frac{1}{N} \sum_{i=1}^N H(\mathbf{X}_i) \frac{f(\mathbf{X}_i)}{g^*(\mathbf{X}_i)} \right) \quad (9.4.25)$$

$$= \text{Var}_{g^*} \left( \frac{1}{N} \sum_{i=1}^N \mu \right) \quad (9.4.26)$$

$$= 0 \quad (9.4.27)$$

This means that just a single sample will yield the desired  $\mu$ . However it is usually not possible to find  $g^*(\mathbf{x})$  because it requires  $\mu$  itself. In practice, good importance sampling densities are chosen to be ‘close’ to the minimum variance density  $g^*(\mathbf{x})$  [115].

### Importance Sampling for Unnormalised Target Densities

Suppose that the target density  $f(\mathbf{x})$  is no longer normalised. That is, we write

$$p(\mathbf{x}) = \frac{f(\mathbf{x})}{Z} \quad (9.4.28)$$

where  $p(\mathbf{x})$  is the normalised density and  $Z = \int f(\mathbf{x}) d\mathbf{x}$  is the normalising constant. Regardless, importance sampling can still be used to estimate the expectation  $\mathbb{E}_p[H(\mathbf{X})]$ . This is done by noticing that if we put  $H(\mathbf{x}) = 1$ , we end up estimating

$$\int 1 \cdot f(\mathbf{x}) = Z \quad (9.4.29)$$

Hence by applying the same importance sampling technique, an estimate of  $Z$  is given by

$$\widehat{Z} = \frac{1}{N} \sum_{i=1}^N w(\mathbf{X}_i) \quad (9.4.30)$$

where each  $\mathbf{X}_i \sim g(\mathbf{x})$  from the importance density  $g(\mathbf{x})$ , while  $w(\mathbf{x}) = \frac{f(\mathbf{x})}{g(\mathbf{x})}$  are the unnormalised importance weights. Thus dividing out by this normalising constant, our estimate for the expectation of interest is

$$\widehat{\mu}' = \frac{1}{N} \sum_{i=1}^N H(\mathbf{X}_i) \frac{1}{\widehat{Z}} \cdot \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)} \quad (9.4.31)$$

$$= \frac{1}{N} \sum_{i=1}^N H(\mathbf{X}_i) \frac{w(\mathbf{X}_i)}{\widehat{Z}} \quad (9.4.32)$$

$$= \frac{1}{N} \sum_{i=1}^N H(\mathbf{X}_i) \frac{w(\mathbf{X}_i)}{\frac{1}{N} \sum_{j=1}^N w(\mathbf{X}_j)} \quad (9.4.33)$$

$$= \sum_{i=1}^N H(\mathbf{X}_i) \cdot \frac{w(\mathbf{X}_i)}{\sum_{j=1}^N w(\mathbf{X}_j)} \quad (9.4.34)$$

$$= \sum_{i=1}^N H(\mathbf{X}_i) \cdot W(\mathbf{X}_i) \quad (9.4.35)$$

where  $W(\mathbf{X}_i) = \frac{w(\mathbf{X}_i)}{\sum_{i=1}^N w(\mathbf{X}_i)}$  are the normalised importance weights.

### Importance Sampling for Distribution Approximation

By estimating the expectation  $\mathbb{E}_p[H(\mathbf{X})]$  with the quantity  $\sum_{i=1}^N H(\mathbf{X}_i) \cdot W(\mathbf{X}_i)$ , we have implicitly approximated the density  $p(\mathbf{x})$  with the weighted point masses

$$\widehat{p}(\mathbf{x}) = \sum_{i=1}^N \delta(\|\mathbf{x} - \mathbf{X}_i\|) \quad (9.4.36)$$

Thus, we can use importance sampling to approximate a distribution that we cannot directly sample from, but can evaluate the density up to a normalising constant (which is useful in situations such as approximating the posterior).

#### 9.4.4 Markov Chain Monte-Carlo

Markov chain Monte-Carlo (MCMC) methods can be used to generate a sample that is approximately distributed from a target distribution that is known up a normalising constant. This requirement is similar to acceptance-rejection sampling, except we do not need to know an explicit constant  $M$  that the target distribution satisfies. The idea behind MCMC is to construct a time-reversible Markov chain that has the target distribution as its stationary and limiting distribution, so if we run the chain for a long time, the distribution of the next sample should approximately the same as the target distribution.

##### Metropolis-Hastings Algorithm for Discrete Distributions [170]

Let  $\mathcal{X}$  be a countably infinite state-space that we wish to sample from. We first demonstrate how to construct a time-reversible Markov chain  $X_t$  that has a desired distribution over  $\mathcal{X}$ . As  $\mathcal{X}$  is countably infinite, we work with

$$\mathcal{X} = \{1, 2, \dots\} \quad (9.4.37)$$

Consider an algorithm which as the following steps in a single iteration:

1. Suppose the current state is  $X_t = i$ . Generate the candidate (or *proposal*)  $Y_{t+1}$  according to the proposal distribution denoted by

$$q_{ij} = \Pr(Y_{t+1} = j | X_t = i) \quad (9.4.38)$$

2. Let the generated value of  $Y_{t+1}$  be  $j$ . Set  $X_{t+1}$  according to

$$\Pr(X_{t+1} = j | Y_{t+1} = j, X_t = i) = \alpha_{ij} \quad (9.4.39)$$

$$\Pr(X_{t+1} = i | Y_{t+1} = j, X_t = i) = 1 - \alpha_{ij} \quad (9.4.40)$$

where  $\alpha_{ij}$  is known as the acceptance probability. That is, we accept the proposal with probability  $\alpha_{ij}$ , otherwise the state does not change.

We can see that  $X_t$  is indeed a Markov chain satisfying the Markov property, with transition probabilities

$$p_{ij} = \Pr(X_{t+1} = j | X_t = i) \quad (9.4.41)$$

$$= \Pr(X_{t+1} = j | Y_{t+1} = j, X_t = i) \Pr(Y_{t+1} = j | X_t = i) \quad (9.4.42)$$

$$= q_{ij} \alpha_{ij} \quad (9.4.43)$$

when  $j \neq i$ , and

$$p_{ii} = \Pr(X_{t+1} = i | X_t = i) \quad (9.4.44)$$

$$= q_{ii} + \sum_{k \neq i} q_{ik} (1 - \alpha_{ik}) \quad (9.4.45)$$

whenever  $j = i$ , since the state does not change whenever the proposal is generated to be  $Y_{t+1} = i$ , or else whenever the proposal is not accepted. Recall that a time-reversible Markov chain satisfies the detailed balance equations

$$\pi_i p_{ij} = \pi_j p_{ji} \quad (9.4.46)$$

for all  $i, j$ , where  $\pi$  is the stationary distribution. Since this is trivially satisfied for  $j = i$ , we require that

$$\pi_i q_{ij} \alpha_{ij} = \pi_j q_{ji} \alpha_{ji} \quad (9.4.47)$$

for all  $j \neq i$ . This can be shown if we let the acceptance probability be

$$\alpha_{ij} = \min \left\{ \frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1 \right\} \quad (9.4.48)$$

The key idea is that if  $\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \geq 1$ , then its reciprocal  $\frac{\pi_i q_{ij}}{\pi_j q_{ji}} \leq 1$ , or vice-versa. Thus if we ever have  $\alpha_{ij} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$ , this implies  $\alpha_{ji} = 1$ . Then

$$\pi_i q_{ij} \alpha_{ij} = \pi_i q_{ij} \cdot \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \quad (9.4.49)$$

$$= \pi_j q_{ji} \quad (9.4.50)$$

$$= \pi_j q_{ji} \alpha_{ji} \quad (9.4.51)$$

which satisfies the detailed balance equations. Now suppose we wish to make the stationary distribution the distribution which has un-normalised masses  $b_1, b_2, \dots$ , etc. Putting  $\pi_j \propto b_j$ , this can be done with the acceptance probability

$$\alpha_{ij} = \min \left\{ \frac{b_j q_{ji}}{b_i q_{ij}}, 1 \right\} \quad (9.4.52)$$

and we can see that the normalising constant is not needed since it cancels out in the ratio. This is the Metropolis-Hastings algorithm.

### Metropolis-Hastings Algorithm for Continuous Distributions [115]

The Metropolis-Hastings algorithm can be applied in much the same way for sampling from continuous distributions. Let  $f(x)$  denote a possibly multivariate target density on support  $\mathcal{X}$ , and known up to a normalising constant. We need a proposal distribution, from which we can sample from the same support as  $f(x)$ , according to the conditional density  $q(y|x)$ . The Metropolis-Hastings is analogous to the discrete case, whereby:

1. Suppose the current state is  $x_t$ . Generate a proposal  $y_{t+1}$  according to  $q(y_{t+1}|x_t)$ .
2. Accept the proposal  $x_{t+1} = y_{t+1}$  with probability

$$\alpha(x_t, y_{t+1}) = \min \left\{ \frac{f(y_{t+1}) q(x_t|y_{t+1})}{f(x_t) q(y_{t+1}|x_t)}, 1 \right\} \quad (9.4.53)$$

otherwise the state is unchanged,  $x_{t+1} = x_t$ .

To demonstrate that this also results in a time-reversible Markov chain with stationary distribution equal to  $f(x)$ , we use the analogous detailed balance equations for uncountable state-space:

$$\pi(x) p(y|x) = \pi(y) p(x|y) \quad (9.4.54)$$

Let  $\kappa(x'|x)$  denote the transition density under the Metropolis-Hastings algorithm. Since there is a positive probability that  $x' = x$ , this density will be given by the hybrid distribution

$$\kappa(y|x) = \alpha(x, y) q(y|x) + \delta(y - x) \int_{\mathcal{X}} q(y|x) (1 - \alpha(x, y)) dy \quad (9.4.55)$$

where  $\delta(\cdot)$  is the Dirac delta function. This is analogous to the discrete case (after combining both cases  $y \neq x$  and  $y = x$  into a single equation). We can show that this transition density satisfies the detailed balance equation

$$f(x) \kappa(y|x) = f(y) \kappa(x|y) \quad (9.4.56)$$

by considering the cases  $y \neq x$  and  $y = x$  separately. If  $y \neq x$ , then  $\delta(y - x)$  and with similar arguments to the discrete case,  $\alpha(x, y) = \frac{f(y) q(x|y)}{f(x) q(y|x)}$  implies  $\alpha(y, x) = 1$ , so

$$f(x) \kappa(y|x) = f(x) \alpha(x, y) q(y|x) \quad (9.4.57)$$

$$= f(x) q(y|x) \cdot \frac{f(y) q(x|y)}{f(x) q(y|x)} \quad (9.4.58)$$

$$= f(y) q(x|y) \quad (9.4.59)$$

$$= f(y) \alpha(y, x) q(x|y) \quad (9.4.60)$$

$$= f(y) \kappa(x|y) \quad (9.4.61)$$

For the case  $y = x$ , note that

$$\delta(y - x) \int_{\mathcal{X}} q(y|x) (1 - \alpha(x, y)) dy = \delta(x - y) \int_{\mathcal{X}} q(x|y) (1 - \alpha(y, x)) dx \quad (9.4.62)$$

holds trivially under  $x = y$ , and both sides are equal to zero otherwise. The other parts of the detailed balance equation are also trivially satisfied.

If the proposal density can be written as  $q(y|x) = q(y)$ , i.e. the proposal  $y$  is sampled independently of  $x$ , then this is known as an *independence sampler*. In this way, the algorithm also becomes quite similar to the acceptance rejection algorithm.

### Metropolis Algorithm [65]

If the Metropolis-Hastings proposal distribution is chosen to be a symmetric distribution, i.e.  $q(y|x) = q(x|y)$  for all  $x, y \in \mathcal{X}$ , then the acceptance probability will just involve the ratio of the target densities:

$$\alpha(x, y) = \min \left\{ \frac{f(y)}{f(x)}, 1 \right\} \quad (9.4.63)$$

This is known as just the Metropolis Algorithm, or also a *random walk sampler* [115].

### Gibbs Sampling [170]

Gibbs sampling is suitable as a MCMC method to sample from  $d$ -dimensional distributions, with  $d > 1$ . The requirement to use Gibbs sampling is that we can generate samples of an individual component, conditioned on all other components (but perhaps it is difficult to generate all the components simultaneously, which necessitates Gibbs sampling). Suppose  $f(\mathbf{x})$  is the target density we wish to sample from. For simplicity of notation in the algorithm, let  $\mathbf{X} = (X_1, \dots, X_d)$  denote the current state, and  $\mathbf{Y} = (Y_1, \dots, Y_d)$  denotes the proposal, while  $\mathbf{X}' = (X'_1, \dots, X'_d)$  denotes the next state. A single iteration of the basic Gibbs sampler is as follows:

1. Pick an index  $j \in \{1, \dots, d\}$  uniformly at random.
2. Sample a value  $Y_j$  from the conditional distribution of the  $j^{\text{th}}$  component given all other components, i.e. from the distribution  $f(x_j|x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)$ .
3. Set the proposal as

$$\mathbf{Y} = (X_1, \dots, Y_j, \dots, X_d) \quad (9.4.64)$$

so that all components except the  $j^{\text{th}}$  are equal to the current state  $\mathbf{X}$ .

4. The next state is automatically equal to the proposal,  $\mathbf{X}' = \mathbf{Y}$ .

The algorithm steps are performed analogously when sampling from a discrete distribution. This basic version of the Gibbs sampler can also be considered a special case of the Metropolis-Hastings algorithm. In the discrete case, the proposal density specified in terms of the conditional distributions is given by

$$\Pr(\mathbf{Y} = (y_1, \dots, y_d) | \mathbf{X} = (x_1, \dots, x_d)) = \begin{cases} \frac{1}{d} \Pr(X_j = y_j | \mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j}), & \mathbf{y}_{\setminus j} = \mathbf{x}_{\setminus j} \\ 0, & \text{otherwise} \end{cases} \quad (9.4.65)$$

for each  $j \in \{1, \dots, d\}$ , where we have used the shorthand notation of subscript  $\setminus j$  to denote all components except the  $j^{\text{th}}$ , e.g.  $\mathbf{x}_{\setminus j} := (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)$ , etc. Also notice that as a Metropolis-Hastings algorithm, the acceptance probability of Gibbs sampling is always one. This can be shown explicitly, firstly through noting the conditional probability:

$$\Pr(X_j = y_j | \mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j}) = \frac{\Pr(X_j = y_j, \mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j})}{\Pr(\mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j})} \quad (9.4.66)$$

$$= \frac{\Pr(\mathbf{X} = \mathbf{y})}{\Pr(\mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j})} \quad (9.4.67)$$

$$= \frac{p(\mathbf{y})}{\Pr(\mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j})} \quad (9.4.68)$$

whenever  $\mathbf{y}_{\setminus j} = \mathbf{x}_{\setminus j}$ , with  $p(\mathbf{y}) := \Pr(\mathbf{X} = \mathbf{y})$ . Hence

$$q(\mathbf{y} | \mathbf{x}) := \Pr(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) \quad (9.4.69)$$

$$= \begin{cases} \frac{1}{d} \cdot \frac{\Pr(\mathbf{X} = \mathbf{y})}{\Pr(\mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j})}, & \mathbf{y}_{\setminus j} = \mathbf{x}_{\setminus j} \\ 0, & \text{otherwise} \end{cases} \quad (9.4.70)$$

$$= \mathbb{I}_{\{\mathbf{y}_{\setminus j} = \mathbf{x}_{\setminus j}\}} \frac{1}{d} \cdot \frac{p(\mathbf{y})}{\Pr(\mathbf{X}_{\setminus j} = \mathbf{x}_{\setminus j})} \quad (9.4.71)$$

Swapping variables, we also have

$$q(\mathbf{x} | \mathbf{y}) = \mathbb{I}_{\{\mathbf{y}_{\setminus j} = \mathbf{x}_{\setminus j}\}} \frac{1}{d} \cdot \frac{p(\mathbf{x})}{\Pr(\mathbf{X}_{\setminus j} = \mathbf{y}_{\setminus j})} \quad (9.4.72)$$

Thus the acceptance probability for any  $\mathbf{x}, \mathbf{y}$  can be computed as

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ \frac{p(\mathbf{y}) q(\mathbf{x} | \mathbf{y})}{p(\mathbf{x}) q(\mathbf{y} | \mathbf{x})}, 1 \right\} \quad (9.4.73)$$

$$= \min \left\{ \frac{p(\mathbf{y}) p(\mathbf{x})}{p(\mathbf{x}) p(\mathbf{y})}, 1 \right\} \quad (9.4.74)$$

$$= 1 \quad (9.4.75)$$

since the denominators in  $q(\mathbf{y} | \mathbf{x})$  and  $q(\mathbf{x} | \mathbf{y})$  will cancel out in the ratio, as the acceptance probability will only ever be evaluated where  $\mathbf{y}_{\setminus j} = \mathbf{x}_{\setminus j}$  for some  $j \in \{1, \dots, d\}$ .

### Systematic Gibbs Sampling [115]

A variant of Gibbs sampling ‘cycles’ through each of the components deterministically to generate a proposal within a single iteration. Let  $f(\mathbf{x})$  denote the target density, and  $f(x_j | \mathbf{x}_{-j})$  the respective conditional densities of the  $j^{\text{th}}$  components given all other components. To perform a single iteration at current state  $\mathbf{X} = \mathbf{x}$ :

1. Generate  $Y_1 = y_1$  from  $f(x_1 | \mathbf{x}_{\setminus 1})$ .
2. Generate  $Y_j = y_j$  from  $f(x_j | y_1, \dots, y_{j-1}, x_{j+1}, \dots, x_d)$  for each of  $j = 2, \dots, d-1$ .
3. Generate  $Y_d = y_d$  from  $f(x_d | y_1, \dots, y_{d-1})$ .
4. Set the next state as  $\mathbf{X}' = \mathbf{y} := (y_1, \dots, y_d)$ .

Using notation  $\mathbf{x}_{1:j} := (x_1, \dots, x_j)$ , this algorithm induces a transition density of the Markov chain given by

$$\vec{\kappa}(\mathbf{y} | \mathbf{x}) = \prod_{j=1}^d f(y_j | \mathbf{y}_{1:(j-1)}, \mathbf{x}_{(j+1):d}) \quad (9.4.76)$$

To show that this Markov chain indeed has the target density  $f(\mathbf{x})$  as a stationary distribution, introduce the ‘reverse’ transition density  $\overleftarrow{\kappa}(\mathbf{x} | \mathbf{y})$  that updates the vector  $\mathbf{y}$  in the reverse order from component  $d$  to component 1.

$$\overleftarrow{\kappa}(\mathbf{x} | \mathbf{y}) = f(x_d | \mathbf{y}_{1:(d-1)}) f(x_{d-1} | \mathbf{y}_{1:(d-2)}, x_d) \dots f(x_1 | \mathbf{x}_{2:d}) \quad (9.4.77)$$

$$= \prod_{j=1}^d f(x_j | \mathbf{y}_{1:(j-1)}, \mathbf{x}_{(j+1):d}) \quad (9.4.78)$$

The ratio of transition densities is

$$\frac{\vec{\kappa}(\mathbf{y} | \mathbf{x})}{\overleftarrow{\kappa}(\mathbf{x} | \mathbf{y})} = \frac{\prod_{i=1}^d f(y_i | \mathbf{y}_{1:(i-1)}, \mathbf{x}_{(i+1):d})}{\prod_{j=1}^d f(x_j | \mathbf{y}_{1:(j-1)}, \mathbf{x}_{(j+1):d})} \quad (9.4.79)$$

$$= \frac{\prod_{i=1}^d f(\mathbf{y}_{1:i}, \mathbf{x}_{(i+1):d}) / f(\mathbf{y}_{1:(i-1)}, \mathbf{x}_{(i+1):d})}{\prod_{j=1}^d f(\mathbf{y}_{1:(j-1)}, \mathbf{x}_{j:d}) / f(\mathbf{y}_{1:(j-1)}, \mathbf{x}_{(j+1):d})} \quad (9.4.80)$$

$$= \frac{\prod_{i=1}^d f(\mathbf{y}_{1:i}, \mathbf{x}_{(i+1):d})}{\prod_{j=1}^d f(\mathbf{y}_{1:(j-1)}, \mathbf{x}_{j:d})} \quad (9.4.81)$$

$$= \frac{f(\mathbf{y}) \prod_{i=1}^{d-1} f(\mathbf{y}_{1:i}, \mathbf{x}_{(i+1):d})}{f(\mathbf{x}) \prod_{j=2}^d f(\mathbf{y}_{1:(j-1)}, \mathbf{x}_{j:d})} \quad (9.4.82)$$

by factoring out the last and first multiplicands in the numerator and denominator respectively. Then by the substitution  $k = j - 1$ , we have

$$\frac{\vec{\kappa}(\mathbf{y} | \mathbf{x})}{\overleftarrow{\kappa}(\mathbf{x} | \mathbf{y})} = \frac{f(\mathbf{y}) \prod_{i=1}^{d-1} f(\mathbf{y}_{1:i}, \mathbf{x}_{(i+1):d})}{f(\mathbf{x}) \prod_{i=1}^{d-1} f(\mathbf{y}_{1:i}, \mathbf{x}_{(i+1):d})} \quad (9.4.83)$$

$$= \frac{f(\mathbf{y})}{f(\mathbf{x})} \quad (9.4.84)$$

Note that the one regularity condition we need for the above to hold is that for every  $\mathbf{x}$  where all the marginal densities  $f(x_i) > 0$ , the joint density  $f(\mathbf{x}) > 0$ , i.e. the support of the joint distribution  $\mathcal{X}$  is equal to the Cartesian product of the marginal distributions. Hence we get

$$f(\mathbf{x}) \vec{\kappa}(\mathbf{y} | \mathbf{x}) = f(\mathbf{y}) \overleftarrow{\kappa}(\mathbf{x} | \mathbf{y}) \quad (9.4.85)$$

which is similar to the detailed balance equations. The difference however, is that we integrate both sides with respect to  $\mathbf{x} \in \mathcal{X}$  and obtain

$$\int_{\mathcal{X}} f(\mathbf{x}) \vec{\kappa}(\mathbf{y} | \mathbf{x}) d\mathbf{x} = \int_{\mathcal{X}} f(\mathbf{y}) \overleftarrow{\kappa}(\mathbf{x} | \mathbf{y}) d\mathbf{x} \quad (9.4.86)$$

$$= f(\mathbf{y}) \int_{\mathcal{X}} \overleftarrow{\kappa}(\mathbf{x}|\mathbf{y}) d\mathbf{x} \quad (9.4.87)$$

$$= f(\mathbf{y}) \quad (9.4.88)$$

These are the Chapman-Kolmogorov equations for the transition density  $\overrightarrow{\kappa}(\mathbf{y}|\mathbf{x})$ , showing that  $f(\cdot)$  is the stationary distribution.

### 9.4.5 Latin Hypercube Sampling

### 9.4.6 Quasi Monte-Carlo Estimation [148]

In traditional Monte-Carlo estimation, we aim to approximate integrals over some domain  $D$  of the form  $\int_D f(\mathbf{x}) d\mathbf{x}$ , which can, for example, represent some probability or expectation of interest. This approximation can be conducted by

$$\int_D f(\mathbf{x}) d\mathbf{x} \approx \frac{V}{N} \sum_{i=1}^N f(\mathbf{X}_i) \quad (9.4.89)$$

using  $N$  nodes, where each node  $\mathbf{X}_i$  is sampled uniformly from the domain  $D$  and  $V = \int_D d\mathbf{x}$ . In a quasi Monte-Carlo approach, we instead perform

$$\int_D f(\mathbf{x}) d\mathbf{x} \approx \frac{V}{N} \sum_{i=1}^N f(\mathbf{x}_i) \quad (9.4.90)$$

where the nodes  $\mathbf{x}_i$  are drawn from a *high-discrepancy* sequence, which is essentially allowed to be a deterministic sequence but ‘resembles’ a uniform distribution over  $D$ . That is, points will appear to be spread uniformly over  $D$ , even though there may be some underlying pattern in the way they are spread. Quasi Monte-Carlo methods can be used to improve the accuracy of the approximation compared to traditional methods (for some fixed  $N$ ).

### 9.4.7 Monte-Carlo Confidence Intervals [162]

## 9.5 Resampling Methods

### 9.5.1 Jackknife

Given a sample of size  $n$ , the Jackknife method for an estimator involves aggregating the estimates for each size  $n - 1$  subsample. Suppose the parameter to be estimated is the population mean. Formally, the Jackknife estimate involves first taking  $n$  sample means with each observation removed:

$$\bar{x}_i = \frac{1}{n-1} \sum_{j=1, j \neq i}^n x_j \quad (9.5.1)$$

for  $i = 1, \dots, n$ . The Jackknife estimate of the population mean is then the mean of all the subsample means:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i \quad (9.5.2)$$

The Jackknife estimate of the variance of the estimator can also be calculated using the distribution of  $\bar{x}_i$  as follows:

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n (\bar{x}_i - \hat{\theta})^2 \quad (9.5.3)$$

This estimator is an unbiased estimator of the variance of the sample mean.

*Proof.* Notice that  $\bar{x}_i = \frac{n\bar{x} - x_i}{n-1}$  implies 1-21-2  $(n-1)\bar{x}_i = n\bar{x} - x_i$  where  $\bar{x}$  is the sample mean. The term  $\bar{x}_i - \hat{\theta}$  can be manipulated to become

$$\bar{x}_i - \hat{\theta} = \frac{n\bar{x} - x_i}{n-1} - \frac{1}{n} \sum_{i=1}^n \bar{x}_i \quad (9.5.4)$$

$$= \frac{1}{n-1} \left( n\bar{x} - x_i - \frac{1}{n} \sum_{i=1}^n (n-1)\bar{x}_i \right) \quad (9.5.5)$$

$$= \frac{1}{n-1} \left( n\bar{x} - x_i - \frac{1}{n} \sum_{i=1}^n (n\bar{x} - x_i) \right) \quad (9.5.6)$$

$$= \frac{1}{n-1} \left( n\bar{x} - x_i - n\bar{x} + \frac{1}{n} \sum_{i=1}^n x_i \right) \quad (9.5.7)$$

$$= \frac{1}{n-1} (\bar{x} - x_i) \quad (9.5.8)$$

Hence

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n (\bar{x}_i - \hat{\theta})^2 \quad (9.5.9)$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{x} - x_i)^2 \quad (9.5.10)$$

Recall that  $\frac{1}{(n-1)} \sum_{i=1}^n (\bar{x} - x_i)^2$  is an unbiased estimator of the population variance, and so  $\frac{1}{n(n-1)} \sum_{i=1}^n (\bar{x} - x_i)^2$  is an unbiased estimator of the variance of the sample mean.  $\square$

## 9.5.2 Bootstrapping

### Empirical Bootstrap [212]

The bootstrap method can be used to compute standard errors of very general estimators. Let  $T_n = g(X_1, \dots, X_n)$  be a statistic. In order to estimate  $\text{Var}(T_n)$  with respect to the data generating process (which is not presumed known), we compute

$$\widehat{\text{Var}}(T_n) = \frac{1}{N} \sum_{i=1}^N \left( T_{n,i}^* - \bar{T}_n^* \right)^2 \quad (9.5.11)$$

where  $T_{n,i}^* = g(X_{1,i}^*, \dots, X_{n,i}^*)$ , is the statistic calculated from the  $i^{\text{th}}$  resample of the data from the empirical distribution function defined by the original data  $X_1, \dots, X_n$ .  $\bar{T}_n^*$  is the sample mean of the statistic across all simulations. Hence the bootstrap estimate contains two levels of approximation. The first approximation is made by using the empirical distribution to replace the data generating process. The second level of approximation is by estimating the variance of  $T_n$  with respect to the empirical distribution by using a Monte Carlo average.

### Parametric Bootstrap

The parametric bootstrap differs from the empirical bootstrap in the way that the resampled data is generated. In the parametric bootstrap, suppose we have a sample  $X_1, \dots, X_n$  from a parametric distribution  $F(x; \theta)$  parametrised in  $\theta$ . Let the estimator for  $\theta$  from the sample be

$$\hat{\theta} = T_n(X_1, \dots, X_n) \quad (9.5.12)$$

To generate each resampled dataset, we randomly generate from the distribution  $F(x; \hat{\theta})$ . Performing the estimate of  $\theta$  on the resampled data gives us each bootstrap replication, from which we can estimate the variance of the sampling distribution.

### Block Bootstrap

The block bootstrap is suitable for resampling dependent data (such as time-series data). Denote the data by  $(X_1, \dots, X_n)$ , and we ideally assume that the process generating the data is stationary or weakly stationary. Decide on an integer block length  $k$ . Then the data can be split into

$$(X_1, \dots, X_n) = (\mathbf{X}_{1:k}, \mathbf{X}_{(k+1):2k}, \dots, \mathbf{X}_{(mk+1):n}) \quad (9.5.13)$$

where  $m = \lfloor n/k \rfloor$ . First for simplicity, assume  $n$  is divisible by  $k$ . To generate a bootstrapped dataset, resample  $m$  blocks with replacement and combine them in an arbitrary order, giving  $(\mathbf{X}_{1:k}^*, \dots, \mathbf{X}_{(mk-k+1):mk}^*)$ . Although each resampled block is independent of every other block, the idea is that we have preserved the dependence between observations within each block. If we let  $k = 1$ , this reduces to the empirical bootstrap, which yields a sample without the desired dependence. Hence the approximation is improved by letting  $k$  grow as  $n$  grows.

To address the case where  $n$  is not divisible by  $k$ , one approach is to resample  $m$  blocks of length  $k$ , and append a resampled block (from the original dataset  $(X_1, \dots, X_n)$ ) with length equal to remainder of  $n$  divided by  $k$ .

### 9.5.3 Bootstrap Confidence Intervals

#### Normal Bootstrap Interval

By assuming  $T_n$  is normally distributed, a simple confidence interval can be formed using  $T_n \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(T_n)}$ .

#### Pivotal Bootstrap Interval

Let  $\theta$  denote the population parameter for  $T$  and let  $\hat{\theta}_n$  denote the sample statistic. Define the pivot  $R_n = \hat{\theta}_n - \theta$  which can be thought of as a ‘zeroed’ random variable. Let  $\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,N}^*$  denote resampled bootstrap replications of  $\hat{\theta}_n$ . There exists some CDF of the pivot

$$F_{R_n}(r) = \Pr(R_n \leq r) \quad (9.5.14)$$

Then  $F_{R_n}^{-1}(\cdot)$  is the quantile function and there exists some  $a, b$  such that

$$F_{R_n}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - a \quad (9.5.15)$$

$$F_{R_n}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - b \quad (9.5.16)$$

We can then show

$$\Pr(a \leq \theta \leq b) = \Pr\left(\hat{\theta}_n - b \leq R_n \leq \hat{\theta}_n - a\right) \quad (9.5.17)$$

$$= F_{R_n}\left(\hat{\theta}_n - a\right) - F_{R_n}\left(\hat{\theta}_n - b\right) \quad (9.5.18)$$

$$= F_{R_n}\left(F_{R_n}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - F_{R_n}\left(F_{R_n}^{-1}\left(\frac{\alpha}{2}\right)\right) \quad (9.5.19)$$

$$= 1 - \alpha \quad (9.5.20)$$

Hence  $(a, b)$  is an exact confidence interval. We estimate  $a$  and  $b$  by first estimating  $F_{R_n}(r)$  from resampling:

$$\hat{F}_{R_n}(r) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(R_{n,i}^* \leq r) \quad (9.5.21)$$

where  $\mathbb{I}(\cdot)$  is the indicator and  $R_{n,i}^* := \hat{\theta}_{n,i}^* - \hat{\theta}_n$ . Let  $r_\alpha^*$  denote the  $\alpha$  sample quantile from resampled bootstrap replications of  $R_{n,1}^*, \dots, R_{n,N}^*$  and similarly let  $\theta_\alpha^*$  denote the  $\alpha$  sample quantile from resampled bootstrap replications of  $\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,N}^*$ . Then  $b$  may be estimated by

$$\hat{b} = \hat{\theta}_n - \hat{F}_{R_n}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r_{\alpha/2}^* \quad (9.5.22)$$

Likewise,  $a$  can be estimated by

$$\hat{a} = \hat{\theta}_n - \hat{F}_{R_n}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r_{1-\alpha/2}^* \quad (9.5.23)$$

Note that  $r_\alpha^* = \theta_\alpha^* - \hat{\theta}_n$  because  $R_{n,i}^* := \hat{\theta}_{n,i}^* - \hat{\theta}_n$ . Hence a  $1 - \alpha$  pivotal bootstrap confidence interval is given by  $(2\hat{\theta}_n - \hat{\theta}_{1-\alpha/2}^*, 2\hat{\theta}_n - \hat{\theta}_{\alpha/2}^*)$ .

### Percentile Bootstrap Interval

The percentile bootstrap interval is an ‘intuitive’ interval using percentiles from the bootstrapped distribution. This is justified as follows. Assume there exists a monotonic transformation  $m(\cdot)$  such that the random variable formed by  $U := m(T)$  is normally distributed with  $U \sim \mathcal{N}(\phi, c^2)$  where  $\phi := m(\theta)$  is the monotonically transformed population parameter. Note that we do not suppose that we know  $m(\cdot)$ , only that it exists. A confidence interval for  $\phi$  would be given by  $U \pm cz_{\alpha/2}$  since

$$\Pr(U - cz_{\alpha/2} \leq \phi \leq U + cz_{\alpha/2}) = \Pr\left(-z_{\alpha/2} \leq \frac{U - \phi}{c} \leq z_{\alpha/2}\right) \quad (9.5.24)$$

$$= 1 - \alpha \quad (9.5.25)$$

The estimates of  $U \pm cz_{\alpha/2}$  would come from bootstrapping. Denote  $U_i^* = m(\theta_{n,i}^*)$  during resampling and let  $u_\alpha^*$  be the  $\alpha$  sample quantile of the resampled bootstrap replications of  $U_1^*, \dots, U_N^*$ . So then

$$1 - \alpha = \Pr(U - cz_{\alpha/2} \leq \phi \leq U + cz_{\alpha/2}) \quad (9.5.26)$$

$$\approx \Pr(u_{\alpha/2}^* \leq \phi \leq u_{1-\alpha/2}^*) \quad (9.5.27)$$

$$= \Pr(m(u_{\alpha/2}^*) \leq m(\phi) \leq m(u_{1-\alpha/2}^*)) \quad (9.5.28)$$

Since the monotonic transformation preserves quantiles, then

$$\Pr(m(u_{\alpha/2}^*) \leq m(\theta) \leq m(u_{1-\alpha/2}^*)) = \Pr(\theta_{\alpha/2}^* \leq \phi \leq \theta_{1-\alpha/2}^*) \quad (9.5.29)$$

So the percentile bootstrap interval is given by  $(\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*)$ .

### 9.5.4 Bootstrap Hypothesis Tests

#### Bootstrap Test for Location Parameter [56]

Let  $X_1, \dots, X_n$  be an i.i.d. sample from distribution  $F(x)$  and let  $Y_1, \dots, Y_m$  be another i.i.d. sample from distribution  $G(x)$ . We wish to test the null hypothesis that  $F(x)$  and  $G(x)$  have

the same mean, against the alternative that the mean of  $F(x)$  is greater than the mean of  $G(x)$ . The test statistic is given by

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/n + s_Y^2/m}} \quad (9.5.30)$$

where  $\bar{X}$ ,  $\bar{Y}$  are the respective sample means and  $s_X^2$ ,  $s_Y^2$  are the respective sample variances. This is the same as the Welch's  $t$ -test statistic. Rather than using the  $t$ -distribution however, we compute the  $p$ -value with a bootstrap approach. Define augmented datasets

$$\tilde{X}_i = X_i - \bar{X} + \bar{Z} \quad (9.5.31)$$

$$\tilde{Y}_i = Y_i - \bar{Y} + \bar{Z} \quad (9.5.32)$$

where

$$\bar{Z} = \frac{n\bar{X} + m\bar{Y}}{n + m} \quad (9.5.33)$$

is the sample mean of the combined dataset. We can see that this transforms the empirical distributions of the datasets to have the same mean, while preserving the original shape. Also if the null is true, then  $\bar{X}$ ,  $\bar{Y}$ ,  $\bar{Z}$  should all be roughly the same, and the augmented datasets will be ‘close’ to the originals. In each bootstrap replication, we resample with replacement a sample of size  $n$  from the  $\tilde{X}_i$ , and a sample of size  $m$  from the  $\tilde{Y}_i$ . From this, we compute the bootstrapped replication of  $T$ , denoted  $T_j^*$ . For  $N$  replications, the  $p$ -value is then calculated by

$$p = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{T_i^* \geq T\}} \quad (9.5.34)$$

which is effectively  $T$  evaluated at the complementary empirical distribution function of the bootstrapped replications. An analogous calculation using both tails follows for a two-tailed test.

### Parametric Bootstrap Hypothesis Test [111]

Let  $T_n(X_1, \dots, X_n)$  be a test statistic constructed such that we should reject the null hypothesis in favour of an appropriate alternative if  $T_n$  is sufficiently large (such as a **goodness-of-fit test statistic**). Suppose we do not know the distribution of  $T_n$  under the null. Instead, we can approximate the null distribution using Monte-Carlo, with  $N$  replications of the test statistic under the null hypothesis (and any **nuisance parameters** required to generate resamples can be substituted with their estimates from the original sample). Denote these by  $T_{n,i}^*$ . Then the  $p$ -value can be calculated as

$$p = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{T_{n,i}^* \geq T_n\}} \quad (9.5.35)$$

which is effectively  $T_n$  evaluated at the complementary empirical distribution function of the bootstrapped replications.

### 9.5.5 Permutation Tests

#### Permutation Test for Location Parameter [68]

Let  $\mathbf{X} = (X_1, \dots, X_n)$  be an i.i.d. sample from a symmetric distribution, with location parameter  $\theta$ . We would like to test the null hypothesis

$$H_0 : \theta = \theta_0 \quad (9.5.36)$$

against the alternative

$$H_A : \theta \neq \theta_0 \quad (9.5.37)$$

for a two-sided test, or either  $\theta > \theta_0$  or  $\theta < \theta_0$  for a one-sided test. Consider a test statistic which is the sum of deviations from the null  $\theta_0$ :

$$T(\mathbf{X}, \theta_0) = \sum_{i=1}^n (X_i - \theta_0) \quad (9.5.38)$$

A permutation test involves estimating the distribution of this test statistic under the null  $\theta_0$  by perturbing the signs of the deviations. Since under the null we have assumed that the underlying distribution is symmetric about  $\theta_0$ , then any sample  $\mathbf{X}$  but with the signs of its deviations arbitrarily perturbed (to either  $-1$  or  $1$ ) is equally likely. We could perform this for all  $2^n$  possible perturbations, and the resulting distribution of the sign-perturbed test statistics will be our approximation of the test statistic under the null. That is, for each  $j$  in  $\{1, \dots, 2^n\}$ , we compute

$$T_j = \sum_{i=1}^n \varsigma_{j,i} (X_i - \theta_0) \quad (9.5.39)$$

where we could enumerate each of the  $2^n$  possibilities by

$$\varsigma_1 = (-1, \dots, -1) \quad (9.5.40)$$

$$\vdots \quad (9.5.41)$$

$$\varsigma_{2^n} = (1, \dots, 1) \quad (9.5.42)$$

using a base-2 convention. The resulting distribution of the  $T_j$ s is sometimes called the permutation distribution, and note that it will be symmetric about zero. We then use the permutation distribution to compute the  $p$ -value of our test statistic  $T(\mathbf{X}, \theta_0)$  in the test procedure. Intuitively, if the null is true, then  $T(\mathbf{X}, \theta_0)$  will be ‘close’ to zero, resulting in a large  $p$ -value. However if the null is false, then  $T(\mathbf{X}, \theta_0)$  should be:

- significantly far from zero (if testing against  $\theta \neq \theta_0$ ),
- significantly higher than zero (if testing against  $\theta > \theta_0$ ),
- or significantly lower than zero (if testing against  $\theta < \theta_0$ ),

resulting in a small  $p$ -value. Note that the smallest increment in the  $p$ -value will be  $1/2^n$  in the one-tailed case, or  $1/2^{n-1}$  in the two-tailed case. Also if  $n$  is large, computing all of the  $2^n$  sign-perturbed test statistics will take prohibitively long. In that case, we can form a Monte-Carlo approximation of the permutation distribution by randomly sampling the  $\varsigma_j$ . It is this way that the permutation test is sometimes also referred to as a randomised test.

### Confidence Intervals from Permutation Tests [68]

The results of permutation tests can be used to obtain confidence intervals, by ‘inverting the test’. In fact, we can use this as a general procedure to construct confidence intervals from a hypothesis test. Consider a random sample  $\mathbf{X}$  supported on  $\mathcal{X}$ , and parameter  $\theta$  from parameter space  $\Theta$ . Let  $S(\mathbf{X})$  denote a confidence interval for  $\theta$  (i.e.  $S(\mathbf{X}) \subset \Theta$ ), and let  $\mathcal{A}(\theta_0)$  denote the non-reject region for the test statistic  $T(\mathbf{X}, \theta_0)$  under the null hypothesis  $\theta = \theta_0$  (i.e.  $\mathcal{A}(\theta_0) \subset \mathcal{T}$  where  $\mathcal{T}$  is the set of all possible values  $T(\mathbf{X}, \theta_0)$  can take on). Consider the following method to construct  $S(\mathbf{X})$ . For all  $\mathbf{x} \in \mathcal{X}$ ,  $\theta_0 \in \Theta$ , we put

$$\theta_0 \in S(\mathbf{x}) \quad (9.5.43)$$

if and only if

$$T(\mathbf{x}, \theta_0) \in \mathcal{A}(\theta_0) \quad (9.5.44)$$

That is,  $S(\mathbf{x})$  contains all the values of  $\theta_0$  across  $\theta_0 \in \Theta$  such that we do not reject the null in a test of  $\theta = \theta_0$ . To see why this would form an appropriate confidence interval for the same level of significance, we reverse our statement for the method used to construct the confidence interval, and say that for all  $\mathbf{x} \in \mathcal{X}$ ,  $\theta_0 \in \Theta$ , we have  $T(\mathbf{x}, \theta_0) \in \mathcal{A}(\theta_0)$  if and only if  $\theta_0 \in S(\mathbf{x})$ . This is equivalent to writing  $\mathcal{A}(\theta_0)$  as

$$\mathcal{A}(\theta_0) = \{T(\mathbf{x}, \theta_0) : \theta_0 \in S(\mathbf{x}), \mathbf{x} \in \mathcal{X}\} \quad (9.5.45)$$

since in words,  $\mathcal{A}(\theta_0)$  contains all the values of  $T(\mathbf{x}, \theta_0)$  across  $\mathbf{x} \in \mathcal{X}$  such that  $\theta_0 \in S(\mathbf{x})$ . Whereas our initial statement is equivalent to writing  $S(\mathbf{x})$ .

$$S(\mathbf{x}) = \{\theta_0 : T(\mathbf{x}, \theta_0) \in \mathcal{A}(\theta_0), \theta_0 \in \Theta\} \quad (9.5.46)$$

Therefore, this shows that  $S(\mathbf{x}) = \{\theta_0 : T(\mathbf{x}, \theta_0) \in \mathcal{A}(\theta_0), \theta_0 \in \Theta\}$  if and only if  $\mathcal{A}(\theta_0) = \{T(\mathbf{x}, \theta_0) : \theta_0 \in S(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ . Now suppose we have a test with level of significance  $\alpha$ . This means the test satisfies

$$\Pr(T(\mathbf{X}, \theta_0) \in \mathcal{A}(\theta_0) | \theta = \theta_0) \geq 1 - \alpha \quad (9.5.47)$$

which is to say we do not make a rejection with more than  $\alpha$  probability, assuming the null. Then from the equivalence we have shown above, it follows that

$$\Pr(\theta_0 \in S(\mathbf{X}) | \theta = \theta_0) = \Pr(T(\mathbf{X}, \theta_0) \in \mathcal{A}(\theta_0) | \theta = \theta_0) \quad (9.5.48)$$

$$\geq 1 - \alpha \quad (9.5.49)$$

Thus,  $S(\mathbf{X})$  is a  $100(1 - \alpha)\%$  confidence interval for  $\theta$ . Confidence intervals constructed this way from inverting tests are sometimes called *consonance intervals* [108]. Specialising to constructing confidence intervals for a location parameter from a permutation test, we would find a range for the null hypothesis  $\theta = \theta_0$  such that the null is not rejected at a level of significance. Two-sided tests yield two-sided confidence intervals, while one-sided tests yield one-sided confidence intervals.

Note that the reverse procedure can in principle be used to obtain rejection regions from confidence intervals (such as to obtain bootstrap hypothesis tests from bootstrap confidence intervals).

### Permutation Test for Difference in Location Parameters [58]

Let  $(\mathbf{X}, \mathbf{Y}) = (X_1, \dots, X_n, Y_1, \dots, Y_m)$  be two samples and assume that  $\mathbf{X}$  and  $\mathbf{Y}$  are drawn from a population with the same shape, but with the only difference being a shift in location parameter. We may think of  $\mathbf{X}$  as being a control group, and  $\mathbf{Y}$  as being a treatment group, and we would like to test whether the treatment has any effect. Naturally, the null hypothesis is that the location parameters are the same, and the alternative hypothesis is the appropriate specification depending on whether the test is one-tailed or two-tailed. Take the test statistic as the difference in means:

$$T(\mathbf{X}, \mathbf{Y}) = \frac{1}{m} \sum_{j=1}^m Y_j - \frac{1}{n} \sum_{i=1}^n X_i \quad (9.5.50)$$

Intuitively, if the null is true, then  $T(\mathbf{X}, \mathbf{Y})$  should be close to zero. We can approximate the sampling distribution of  $T(\mathbf{X}, \mathbf{Y})$  under the null by the following permutation test procedure. If the null is true, the sequence  $(X_1, \dots, X_n, Y_1, \dots, Y_m)$  is an exchangeable sequence. Hence we can calculate the the test statistic under all  $\binom{n+m}{n}$  combinations of dividing the pooled sample into two groups in order to form the permutation distribution. The  $p$ -value can then be calculated by comparing  $T(\mathbf{X}, \mathbf{Y})$  to the permutation distribution. If we were to condition on the observed values, and let only the treatment be randomly assigned across the overall

sample, then the test becomes an *exact test* (in that we can compute the  $p$ -value exactly; there is no approximation). Conditioning on the observed values under the null is akin to saying that we were going to observe those values anyway, regardless of whether or not the treatment was applied to any individual observation. This lets us compute exact  $p$ -values under the null hypothesis because we can obtain the exact distribution of the test statistic.

## 9.6 Survival Analysis

### 9.6.1 Survival Function

### 9.6.2 Censored Data

#### Tobit Regression

#### Cox Regression

### 9.6.3 Kaplan-Meier Estimator [89]

### 9.6.4 Greenwood's Formula

## 9.7 Nonparametric Statistics

### 9.7.1 Nonparametric Mass Estimation

Suppose  $X$  is a random variable supported on a finite set, taking values in  $\{a_1, \dots, a_K\}$ . Based on some measurements of  $X$ , we would like to estimate the probability mass function of  $X$ , without any parametric assumptions. The way to do this is to essentially ‘parametrise’ the distribution of  $X$  with a  $K$ -way categorical distribution. Suppose we observe a total of  $n$  independent samples of  $X$ , and let  $n_1, \dots, n_K$  denote the number of observations in  $a_1, \dots, a_K$  respectively, such that  $n_1 + \dots + n_K = n$ . Then the likelihood of the respective probabilities  $p_1, \dots, p_K$  given the data is

$$\mathcal{L}(p_1, \dots, p_K | X_1, \dots, X_n) = \prod_{k=1}^K p_k^{n_k} \quad (9.7.1)$$

So the log-likelihood is

$$\log \mathcal{L}(p_1, \dots, p_K | X_1, \dots, X_n) = \sum_{k=1}^K n_k \log p_k \quad (9.7.2)$$

The maximum likelihood estimate is obtained by solving the following convex optimisation problem:

$$\begin{aligned} \min_{p_1, \dots, p_K} \quad & - \sum_{k=1}^K n_k \log p_k \\ \text{s.t.} \quad & p_1 + \dots + p_K = 1 \\ & p_1 \geq 0, \dots, p_K \geq 0 \end{aligned} \quad (9.7.3)$$

where we can impose additional constraints based on known structure of the distribution (e.g. maximum or minimum probabilities).

### 9.7.2 Nonparametric Density Estimation [91]

#### Histogram Density Estimation

We consider a histogram-based approach to formally specify the estimate of the probability distribution function of a given univariate sample  $x_1, \dots, x_n$ . In this method, the required

hyperparameters are an origin  $x_{(0)}$  and bin width  $h$ , producing the bins

$$B_1 = [x_{(0)}, x_{(0)} + h] \quad (9.7.4)$$

$$B_2 = [x_{(0)} + h, x_{(0)} + 2h] \quad (9.7.5)$$

$$\vdots \quad (9.7.6)$$

$$B_m = [x_{(0)} + (m - 1)h, x_{(0)} + mh] \quad (9.7.7)$$

$$\vdots \quad (9.7.8)$$

$$B_M = [x_{(0)} + (M - 1)h, x_{(0)} + Mh] \quad (9.7.9)$$

where the total number of bins  $M$  is induced by the choice of  $x_{(0)}$ ,  $h$  and the range of the data, such that  $x_{(0)} + Mh \geq \max_i x_i$ . Moreover,  $x_{(0)}$  should be chosen such that  $x_{(0)} \leq \min_i x_i$  (i.e. the bins cover all the data). For a point  $x$ , let  $\iota(x)$  denote the index of the bin that  $x$  falls in, i.e.

$$\iota(x) = \mathbb{I}_{\{x \in B_1\}} + 2\mathbb{I}_{\{x \in B_2\}} + \cdots + M\mathbb{I}_{\{x \in B_M\}} \quad (9.7.10)$$

and let  $n_m := \sum_{i=1}^n \mathbb{I}_{\{x_i \in B_m\}}$  denote the number of samples falling in bin  $m$ . Then we can express the histogram density estimator as

$$\hat{f}(x) = \frac{1}{nh} n_{\iota(x)} \quad (9.7.11)$$

and take  $\hat{f}(x) = 0$  whenever  $x$  does not fall into any of the bins. This specification is essentially a mixture of uniform distributions. We can verify that this estimate is a valid density function, because

$$\int_{-\infty}^{\infty} \hat{f}(x) dx = \frac{n_1}{nh} \int_{B_1} dx + \cdots + \frac{n_M}{nh} \int_{B_M} dx \quad (9.7.12)$$

And as  $\int_{B_m} dx = h$  due to the chosen bin length, then

$$\int_{-\infty}^{\infty} \hat{f}(x) dx = \frac{n_1 + \cdots + n_M}{n} \quad (9.7.13)$$

$$= 1 \quad (9.7.14)$$

## Kernel Density Estimation

The empirical density function estimator from a given univariate sample  $x_1, \dots, x_n$  yields

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - x_i) \quad (9.7.15)$$

However this density as well as the histogram density estimator will not be smooth, which can be problematic when trying to model continuous distributions, or if we explicitly assume the density to be smooth. To introduce smoothness, we can replace the Dirac delta density  $\delta(\cdot)$  with a *kernel*  $K(\cdot)$ , which can be any smooth probability density function with a mean of zero. A common choice is the standard Gaussian density. Moreover, we can scale the density by a factor  $h > 0$ , which we call the *bandwidth*. This gives the kernel density estimator (also called the Rosenblatt-Parzen estimator):

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (9.7.16)$$

The bandwidth  $h$  controls the amount of smoothing. As  $h$  becomes small, this is like letting the variance go to zero, so  $\frac{1}{h} K\left(\frac{x - x_i}{h}\right)$  approaches the Dirac delta density  $\delta(x - x_i)$  as  $h \rightarrow 0$ .

## Multivariate Kernel Density Estimation

Kernel density estimation can be extended to multivariate densities. Consider a multivariate sample  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . We now define a multivariate kernel  $K(\mathbf{x})$  (i.e. smooth multivariate probability density with mean zero) and introduce the *bandwidth matrix*  $\mathbf{H}$ , which is symmetric positive-definite. We now scale by a factor of  $\mathbf{H}^{1/2}$ , and the appropriate transformation of the kernel is given by  $\det(\mathbf{H}^{-1/2}) K(\mathbf{H}^{-1/2}\mathbf{x})$ . This yields the density estimator

$$\hat{f}(\mathbf{x}) = \frac{1}{n \det(\mathbf{H}^{1/2})} \sum_{i=1}^n K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right) \quad (9.7.17)$$

A common choice of  $K(\mathbf{x})$  is the standard multivariate Gaussian density, in which case  $\mathbf{H}$  plays the role of the covariance matrix. There are also several ways to construct a multivariate kernel from a symmetric univariate kernel  $k(x)$ :

- We can take a product. With  $d$  dimensions so that  $\mathbf{x} = (x_1, \dots, x_d)$ , we have

$$K(\mathbf{x}) = \prod_{j=1}^d k(x_j) \quad (9.7.18)$$

- We can produce a spherically symmetric kernel by

$$K(\mathbf{x}) = \frac{1}{c} k\left(\sqrt{\mathbf{x}^\top \mathbf{x}}\right) \quad (9.7.19)$$

where  $c = \int k\left(\sqrt{\mathbf{x}^\top \mathbf{x}}\right) d\mathbf{x}$  is the normalising constant.

## Characteristic Function Density Estimation

A natural estimator for the characteristic function  $\varphi(t) = \mathbb{E}[e^{-itX}]$  from a sample  $X_1, \dots, X_n$  is

$$\hat{\varphi}(t) = \frac{1}{n} \sum_{j=1}^n e^{-itX_j} \quad (9.7.20)$$

Since the characteristic function is the Fourier transform of the density function, we can obtain an estimate of the density function via the inverse Fourier transform of  $\hat{\varphi}(t)$ :

$$\hat{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\varphi}(t) e^{ixt} dt \quad (9.7.21)$$

Note however that the Fourier transform of the Dirac delta function  $\delta(x - x_j)$  is  $e^{-itx_j}$ , so we find that this estimate is equivalent to the empirical density function.

## Damped Characteristic Function Density Estimation

In characteristic function density estimation, consider multiplying  $\hat{\varphi}(t)$  by a *damping function*  $\psi(t)$ , which has the property  $\psi(0) = 1$  and falls to zero as  $|t| \rightarrow \infty$ . More generally, consider the scaled damping function  $\psi(ht)$  where  $h > 0$  controls the rate of ‘dropoff’ to zero. This way, as  $h \rightarrow 0$ , we will have  $\psi(t) = 1$  everywhere and  $\hat{\varphi}(t)$  will be unchanged. Then the damped characteristic function estimator for the density is

$$\hat{f}(x) = \int_{-\infty}^{\infty} \hat{\varphi}(t) \psi(ht) e^{ixt} dt \quad (9.7.22)$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{n} \sum_{j=1}^n e^{-itX_j} \psi(ht) e^{ixt} dt \quad (9.7.23)$$

Using a change of variables  $s = ht$ ,

$$\widehat{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{n} \sum_{j=1}^n e^{-iX_j s/h} \psi(s) e^{ixs/h} \frac{1}{h} ds \quad (9.7.24)$$

$$= \frac{1}{nh} \sum_{j=1}^n \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi(s) \exp\left(is \frac{x - X_j}{h}\right) ds \quad (9.7.25)$$

$$= \frac{1}{nh} \sum_{j=1}^n K\left(\frac{x - X_j}{h}\right) \quad (9.7.26)$$

where we recognise  $K(\cdot)$  is the inverse Fourier transform of  $\psi(\cdot)$ . Hence damped characteristic function density estimation is equivalent to kernel density estimation with bandwidth  $h$ .

### 9.7.3 Nonparametric Regression

#### Nadaraya-Watson Estimator [32]

Suppose we have observations  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ . A nonparametric model of the density  $p(\mathbf{x}, y)$  can be constructed as follows. Let  $K(\mathbf{x})$  be a multivariate kernel for  $\mathbf{x}$ . For compactness of notation, for bandwidth matrix  $\mathbf{H}$ , denote

$$K_{\mathbf{H}}(\mathbf{x}) := \frac{1}{\det(\mathbf{H}^{1/2})} K\left(\mathbf{H}^{-1/2}\mathbf{x}\right) \quad (9.7.27)$$

Let  $\kappa(y)$  be a univariate kernel for  $y$ , and for compactness of notation denote

$$\kappa_h(y) := \frac{1}{h} \kappa\left(\frac{y}{h}\right) \quad (9.7.28)$$

By forming a joint kernel for  $\mathbf{x}, y$  via their product, we have a multivariate kernel density estimate for the joint density  $p(\mathbf{x}, y)$  given by

$$\widehat{p}(\mathbf{x}, y) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) \kappa_h(y - y_i) \quad (9.7.29)$$

Also, we have a multivariate kernel density estimator for the marginal density  $p(\mathbf{x})$  given by

$$\widehat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) \quad (9.7.30)$$

Thus the natural estimate of the conditional density  $p(y|\mathbf{x})$  is

$$\widehat{p}(y|\mathbf{x}) = \frac{\widehat{p}(\mathbf{x}, y)}{\widehat{p}(\mathbf{x})} \quad (9.7.31)$$

$$= \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) \kappa_h(y - y_i)}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)} \quad (9.7.32)$$

Now consider a regression function given by the estimated conditional expectation:

$$f(\mathbf{x}) = \widehat{\mathbb{E}}[Y|\mathbf{X} = \mathbf{x}] \quad (9.7.33)$$

This can be computed from our estimate of the conditional density by

$$\widehat{\mathbb{E}}[Y|\mathbf{X} = \mathbf{x}] = \int y \widehat{p}(y|\mathbf{x}) dy \quad (9.7.34)$$

$$= \int y \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) \kappa_h(y - y_i)}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)} dy \quad (9.7.35)$$

$$= \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) \int y \kappa_h(y - y_i) dy}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)} \quad (9.7.36)$$

Observe that  $\int y \kappa_h(y - y_i) dy = y_i$  because a kernel must be zero-mean, so  $\kappa_h(y - y_i)$  is centered at  $y = y_i$ . Hence

$$\hat{\mathbb{E}}[Y|\mathbf{X} = \mathbf{x}] = \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) y_i}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)} \quad (9.7.37)$$

$$= \sum_{i=1}^n \frac{K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{j=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_j)} y_i \quad (9.7.38)$$

Therefore the regression function takes the form of a weighted average

$$f(\mathbf{x}) = \sum_{i=1}^n w_i(\mathbf{x}) y_i \quad (9.7.39)$$

with weightings

$$w_i(\mathbf{x}) = \frac{K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{j=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_j)} \quad (9.7.40)$$

This is known as the Nadaraya-Watson estimator. Note that this estimator does not depend at all on the choice of kernel  $\kappa(y)$  for  $y$ . This form is also quite intuitive, because the weightings are higher for observations that are close to  $\mathbf{x}$ . So the estimator can be thought of as providing a rough average of the  $y_i$  values corresponding to the nearest  $\mathbf{x}_i$  values to  $\mathbf{x}$ .

### Kernel Ridge Regression [25]

Recall that the ridge regression estimator is given by

$$\hat{\beta} = (\lambda I + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (9.7.41)$$

Now consider a generalisation where we perform ridge regression in a feature space under the basis transformation  $\phi(\mathbf{x}) : \mathbb{R}^p \rightarrow \mathbb{R}^d$ . This basis transform can potentially ‘lift’ the features into a higher dimensional space. For the data  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ , denote

$$\Phi = \begin{bmatrix} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \phi(\mathbf{x}_n)^\top \end{bmatrix} \quad (9.7.42)$$

Hence the ridge regression estimator is

$$\hat{\beta} = (\lambda I + \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y} \quad (9.7.43)$$

Restrict the regularisation parameter  $\lambda > 0$  so it is strictly positive. We can derive a ‘dual form’ for this estimator. Using the matrix inversion lemma

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \quad (9.7.44)$$

with  $A = \lambda I$ ,  $U = \Phi^\top$ ,  $V = \Phi$  and  $C = I$ , we have

$$(\lambda I + \Phi^\top \Phi)^{-1} = \frac{1}{\lambda} I - \frac{1}{\lambda} \Phi^\top \left( I + \frac{1}{\lambda} \Phi \Phi^\top \right)^{-1} \frac{1}{\lambda} \Phi \quad (9.7.45)$$

Post-multiplying both sides by  $\Phi^\top$ :

$$\left(\lambda I + \Phi^\top \Phi\right)^{-1} \Phi^\top = \frac{1}{\lambda} \Phi^\top - \frac{1}{\lambda} \Phi^\top \left(I + \frac{1}{\lambda} \Phi \Phi^\top\right)^{-1} \frac{1}{\lambda} \Phi \Phi^\top \quad (9.7.46)$$

$$= \frac{1}{\lambda} \Phi^\top \left(I + \frac{1}{\lambda} \Phi \Phi^\top\right)^{-1} \left[\left(I + \frac{1}{\lambda} \Phi \Phi^\top\right) - \frac{1}{\lambda} \Phi \Phi^\top\right] \quad (9.7.47)$$

$$= \frac{1}{\lambda} \Phi^\top \left(I + \frac{1}{\lambda} \Phi \Phi^\top\right)^{-1} \quad (9.7.48)$$

$$= \frac{1}{\lambda} \Phi^\top \lambda \left(\lambda I + \Phi \Phi^\top\right)^{-1} \quad (9.7.49)$$

$$= \Phi^\top \left(\lambda I + \Phi \Phi^\top\right)^{-1} \quad (9.7.50)$$

So the dual form is given by

$$\hat{\beta} = \Phi^\top \left(\lambda I + \Phi \Phi^\top\right)^{-1} \mathbf{y} \quad (9.7.51)$$

and moreover, the prediction  $y_*$  at test point  $\mathbf{x}_*$  is

$$y_* = \phi(\mathbf{x}_*)^\top \hat{\beta} \quad (9.7.52)$$

$$= \phi(\mathbf{x}_*)^\top \Phi^\top \left(\lambda I + \Phi \Phi^\top\right)^{-1} \mathbf{y} \quad (9.7.53)$$

See that

$$\phi(\mathbf{x}_*)^\top \Phi^\top = \begin{bmatrix} \phi(\mathbf{x}_*)^\top \phi(\mathbf{x}_1) & \dots & \phi(\mathbf{x}_*)^\top \phi(\mathbf{x}_n) \end{bmatrix} \quad (9.7.54)$$

$$\Phi \Phi^\top = \begin{bmatrix} \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_1) & \dots & \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_1) & \dots & \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_n) \end{bmatrix} \quad (9.7.55)$$

Hence the prediction only depends on the dot products between members of  $(\phi(\mathbf{x}_*), \phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n))$ . For a feature mapping  $\phi : \mathcal{X} \rightarrow \mathcal{K}$ , we define the *kernel* [86]  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  as the dot product

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}') \quad (9.7.56)$$

Note that the kernel is symmetric, i.e.  $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$ . Thus we can write the prediction entirely in terms of kernels. This motivates the so-called ‘kernel trick’, whereby instead of directly choosing a basis, we can instead choose a kernel. However this kernel needs to satisfy some properties in order to be valid. For one, the kernel must necessarily be a symmetric function. Note that the ‘Gram matrix’  $\Phi \Phi^\top$  is positive semidefinite. Hence if we generally denote the Gram matrix in terms of kernels as

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \quad (9.7.57)$$

then another property of a valid kernel is that  $\mathbf{K}$  should be positive semidefinite. The kernel trick is powerful enough that for some choices of kernels, it could be the kernel of an infinite-dimensional basis. That is, we could interpret the basis as coming from a Hilbert space - an infinite-dimensional function where a notion of inner product  $\langle \cdot, \cdot \rangle$  is defined. In this case, we would refine the definition of the kernel to being an inner product:

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \quad (9.7.58)$$

rather than a dot product (which would be valid if we were considering only Euclidean space). Writing the prediction in terms of kernels, we have

$$y_* = [k(\mathbf{x}_*, \mathbf{x}_1) \dots k(\mathbf{x}_*, \mathbf{x}_n)] \left( \lambda I + \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right)^{-1} \mathbf{y} \quad (9.7.59)$$

$$= \mathbf{k}^\top (\mathbf{x}_*)^\top (\lambda I + \mathbf{K})^{-1} \mathbf{y} \quad (9.7.60)$$

Recognise that this prediction is now inherently nonparametric; we have done away with  $\hat{\beta}$ , and instead each prediction must be made by evaluating the kernels between points in the data and the test point.

### Isotonic Regression

Isotonic regression (which can also be referred to as monotonic regression) is a nonparametric shape-constrained regression method which aims to estimate the labels which agree with monotonicity constraints on the regressors. To illustrate, consider the dataset of size  $n$  consists of features

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \quad (9.7.61)$$

with each  $\mathbf{x}_i \in \mathbb{R}^d$  and corresponding labels

$$\mathbf{y} = (y_1, \dots, y_n) \quad (9.7.62)$$

Suppose it is known that a ‘true’ regression function  $f(\mathbf{x})$  should satisfy  $f(\mathbf{x}) \leq f(\mathbf{x}')$  if and only if  $\mathbf{x} \leq \mathbf{x}'$  (component-wise inequality). Then the aim is to find estimates  $\hat{y}_1, \dots, \hat{y}_n$  which are ‘close’ to the original labels, but however respect the monotonicity order implied by the dataset. Isotonic regression can be performed as follows. We can construct a directed acyclic graph  $\mathcal{G}(V, E)$  with vertices  $V$  and edges  $E$ , where each vertex represents a point in  $\mathbf{X}$ . The edges define a partial ordering between all the points, such that there is an edge from vertex  $a$  to vertex  $b$  if  $\mathbf{x}_a \leq \mathbf{x}_b$ . These are the constraints that the regression function  $f(\mathbf{x})$  has to satisfy. Then a least-squares approach is to solve the quadratic program:

$$\begin{aligned} \min_{\hat{\mathbf{y}}} \quad & \sum_{i=1}^n (\hat{y}_i - y_i)^2 \\ \text{s.t.} \quad & \hat{y}_a \leq \hat{y}_b, \quad \forall (a, b) \in E \end{aligned} \quad (9.7.63)$$

A possible extension/generalisation is to perform weighted least-squares with a weighting vector  $\mathbf{w} = (w_1, \dots, w_n)$ , giving the problem

$$\begin{aligned} \min_{\hat{\mathbf{y}}} \quad & \sum_{i=1}^n w_i (\hat{y}_i - y_i)^2 \\ \text{s.t.} \quad & \hat{y}_a \leq \hat{y}_b, \quad \forall (a, b) \in E \end{aligned} \quad (9.7.64)$$

### 9.7.4 Splines

#### Regression Splines

#### Smoothing Splines

### 9.7.5 Nonparametric Hypothesis Testing

#### Wilcoxon Rank-Sum Test [90]

The Wilcoxon rank-sum test is for testing hypotheses that one random variable is ‘stochastically different’ from another. Let  $\{X_1, \dots, X_m\}$  and  $\{Y_1, \dots, Y_n\}$  be independent samples (both

within sample and between samples) from continuous distributions with cumulative distribution function  $F(x)$  and  $G(y)$  respectively. The null hypothesis is that these random variables are stochastically the same, that is

$$H_0 : F(t) = G(t) \quad (9.7.65)$$

for all  $t$ . A *two-tailed test* specifies the alternative that

$$H_A : F(t) \neq G(t) \quad (9.7.66)$$

for at least one  $t$ . It is also sometimes appropriate to specify the null and alternative as

$$H_0 : \Pr(X > Y) = \Pr(Y > X) \quad (9.7.67)$$

$$H_A : \Pr(X > Y) \neq \Pr(Y > X) \quad (9.7.68)$$

For an *upper-tailed test*, we can specify the hypotheses

$$H_0 : F(t) = G(t) \quad (9.7.69)$$

for all  $t$ , and the alternative that  $Y$  is ‘stochastically larger’ than  $X$  by

$$H_A : F(t) \geq G(t) \quad (9.7.70)$$

for all  $t$ , with at least some inequality strict. For a *lower-tailed test*, we can specify the hypotheses

$$H_0 : F(t) = G(t) \quad (9.7.71)$$

for all  $t$ , and the alternative that  $Y$  is ‘stochastically less’ than  $X$  by

$$H_A : F(t) \leq G(t) \quad (9.7.72)$$

for all  $t$ , with at least some inequality strict. Analogous specifications involving the probabilities  $\Pr(X > Y)$  and  $\Pr(Y > X)$  are also possible.

To compute the test statistic  $W$ , combine the two samples into a single sample of size  $N = n + m$  and order from least to greatest, assigning the ranks from 1 to  $N$  for each element (i.e. the lowest is assigned rank 1). Since  $X$  and  $Y$  were assumed continuous, we need not consider ties. Let  $R_i$  denote the rank of  $Y_i$ . The Wilcoxon rank-sum statistic is given by the sum of ranks:

$$W = \sum_{i=1}^n R_i \quad (9.7.73)$$

Under the null hypothesis, the ranks  $R_1, \dots, R_n$  may be treated as a random sample of size  $n$  from a finite population  $\{1, \dots, N\}$ . Hence the statistic  $\frac{W}{n}$  under the null hypothesis has the same distribution of the sample mean of sample size  $n$  from the finite population  $\{1, \dots, N\}$ . We can compute the mean of this population as

$$\mu = \frac{1}{N} \sum_{j=1}^N j \quad (9.7.74)$$

$$= \frac{N(N+1)}{2N} \quad (9.7.75)$$

$$= \frac{N+1}{2} \quad (9.7.76)$$

and the population variance as

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^N j^2 - \mu^2 \quad (9.7.77)$$

$$= \frac{1}{N} \cdot \frac{N(N+1)(2N+1)}{6} - \left( \frac{N+1}{2} \right)^2 \quad (9.7.78)$$

$$= \frac{2N^2 + 3N + 1}{6} - \frac{N^2 + 2N + 1}{4} \quad (9.7.79)$$

$$= \frac{N^2 - 1}{12} \quad (9.7.80)$$

$$= \frac{(N+1)(N-1)}{12} \quad (9.7.81)$$

Then from the theory of the sampling distribution of the sample mean from a finite population:

$$\mathbb{E} \left[ \frac{W}{n} \right] = \frac{N+1}{2} \quad (9.7.82)$$

$$\begin{aligned} \text{Var} \left( \frac{W}{n} \right) &= \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \\ &= \frac{m}{N-1} \cdot \frac{(N+1)(N-1)}{12n} \\ &= \frac{m(N+1)}{12n} \end{aligned} \quad (9.7.83)$$

Hence in the null distribution:

$$\mathbb{E}[W] = \frac{n(N+1)}{2} \quad (9.7.84)$$

$$\text{Var}(W) = \frac{mn(N+1)}{12} \quad (9.7.85)$$

Also note that the null distribution of  $W$  is symmetric about its mean and has support  $\left\{ \frac{n(n+1)}{2}, \dots, \frac{n(N+m+1)}{2} \right\}$ , so that

$$\Pr(W \leq \mathbb{E}[W] - x) = \Pr(W \geq \mathbb{E}[W] + x) \quad (9.7.86)$$

for all  $x$  in  $\left\{ -\frac{n(m+1)}{2}, \dots, \frac{n(m+1)}{2} \right\}$ . Making the substitution  $w = \mathbb{E}[W] - x$ , we get

$$\Pr(W \leq w) = \Pr(W \geq 2\mathbb{E}[W] - w) \quad (9.7.87)$$

for all  $w$  in the support of  $W$ . This property may be convenient for computing  $p$ -values and deriving equivalent rejection rules between upper and lower-tailed tests. From the asymptotic normality of the sample mean from a finite population, the standardised statistic is approximately normally distributed from the standard normal, for large  $n$  and  $N$ :

$$\frac{W - n(N+1)/2}{\sqrt{mn(N+1)/12}} \xrightarrow{\text{approx.}} \mathcal{N}(0, 1) \quad (9.7.88)$$

so in this instance, critical values can just be taken from the standard normal distribution.

### Mann-Whitney $U$ -Test

The Mann-Whitney  $U$ -test gives an equivalent test to the Wilcoxon rank-sum test (so hypotheses may be specified in the same way). Given samples  $\{X_1, \dots, X_m\}$  and  $\{Y_1, \dots, Y_n\}$  under the same assumptions as in the Wilcoxon rank-sum test, the test statistic is

$$U = \sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{X_i < Y_j} \quad (9.7.89)$$

We can show that the Wilcoxon rank-sum test statistic  $W = \sum_{i=1}^n R_i$  can be written in terms of  $U$ . Observe that  $R_i$  will equal the number of  $X$ -values lower than  $Y_i$ , plus the number of  $Y$ -values lower than  $Y_i$ , plus 1. So

$$R_i = \sum_{j=1}^m \mathbb{I}_{X_j < Y_i} + \sum_{k=1}^n \mathbb{I}_{Y_k < Y_i} + 1 \quad (9.7.90)$$

Hence

$$W = \sum_{i=1}^n \left( \sum_{j=1}^m \mathbb{I}_{X_j < Y_i} + \sum_{k=1}^n \mathbb{I}_{Y_k < Y_i} + 1 \right) \quad (9.7.91)$$

$$= \sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{X_j < Y_i} + \sum_{i=1}^n \sum_{k=1}^n \mathbb{I}_{Y_k < Y_i} + n \quad (9.7.92)$$

$$= U + 0 + 1 + \dots + (n-1) + n \quad (9.7.93)$$

$$= U + \frac{n(n+1)}{2} \quad (9.7.94)$$

#### 9.7.6 Nonparametric Confidence Intervals

##### Cumulative Distribution Function Confidence Intervals

A pointwise confidence interval can be constructed for any individual point on the empirical CDF. Using the fact that the empirical CDF  $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{X_i \leq x}$  is a sample proportion, then any standard techniques for constructing intervals for sample proportions can be applied (such as a normal approximation and  $z$ -scores for larger samples).

A ‘simultaneous’ confidence interval on the empirical CDF applies to the entirety of CDF, so that there is a coverage probability of  $1 - \alpha$  that the bound is violated anywhere along the empirical CDF. Such an interval can be constructed using the Dvoretzsky-Kiefer-Wolfowitz inequality. Beginning from

$$\Pr \left( \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2} \quad (9.7.95)$$

we apply DeMorgan’s laws to obtain

$$\Pr \left( \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \leq \varepsilon \right) \geq 1 - 2e^{-2n\varepsilon^2} \quad (9.7.96)$$

Thus for at least a  $1 - \alpha$  coverage probability, we set

$$\alpha = 2e^{-2n\varepsilon^2} \quad (9.7.97)$$

which rearranges to

$$\varepsilon = \sqrt{\frac{\log(2/\alpha)}{2n}} \quad (9.7.98)$$

Note that is only a function of  $n$  and  $\alpha$ , hence the confidence interval will be nonparametric (i.e. not dependent on the distribution  $F(x)$ ). As we have a bound on the worst-case deviation, then for any point  $x$  along the CDF:

$$\Pr\left(\left|\widehat{F}_n(x) - F(x)\right| \leq \sqrt{\frac{\log(2/\alpha)}{2n}}\right) \geq 1 - \alpha \quad (9.7.99)$$

or equivalently:

$$\Pr\left(\widehat{F}_n(x) - \sqrt{\frac{\log(2/\alpha)}{2n}} \leq F(x) \leq \widehat{F}_n(x) + \sqrt{\frac{\log(2/\alpha)}{2n}}\right) \geq 1 - \alpha \quad (9.7.100)$$

Therefore a  $1 - \alpha$  confidence interval (which actually has coverage probability of at least  $1 - \alpha$ ) running along every point on the empirical CDF is given by

$$\left[\widehat{F}_n(x) - \sqrt{\frac{\log(2/\alpha)}{2n}}, \widehat{F}_n(x) + \sqrt{\frac{\log(2/\alpha)}{2n}}\right] \quad (9.7.101)$$

which since  $0 \leq F(x) \leq 1$ , can be further restricted to

$$\left[\max\left\{0, \widehat{F}_n(x) - \sqrt{\frac{\log(2/\alpha)}{2n}}\right\}, \min\left\{\widehat{F}_n(x) + \sqrt{\frac{\log(2/\alpha)}{2n}}, 1\right\}\right] \quad (9.7.102)$$

### Quantile Function Confidence Intervals

Bounds from the Dvoretzsky-Kiefer-Wolfowitz inequality can also be inverted to give a non-parametric and finite-sample confidence interval on the quantile function (inverse cumulative distribution function), defined as  $F^{-1}(p) = \inf_{F(x) \geq p} x$ . We invert the event inside

$$\Pr\left(\widehat{F}_n(x) - \sqrt{\frac{\log(2/\alpha)}{2n}} \leq F(x) \leq \widehat{F}_n(x) + \sqrt{\frac{\log(2/\alpha)}{2n}}\right) \geq 1 - \alpha \quad (9.7.103)$$

to give

$$\Pr\left(\inf_{\widehat{F}_n(x) + \sqrt{n^{-1} \log(\sqrt{2/\alpha})} \geq p} \{x\} \leq \inf_{F(x) \geq p} x \leq \inf_{\widehat{F}_n(x) - \sqrt{n^{-1} \log(\sqrt{2/\alpha})} \geq p} \{x\}\right) \geq 1 - \alpha \quad (9.7.104)$$

where the inequality signs are switched around because quantiles of the lower confidence bound will be greater, and vice-versa. Hence a  $1 - \alpha$  confidence interval for the quantile function is

$$\Pr\left(\inf_{\widehat{F}_n(x) + \sqrt{n^{-1} \log(\sqrt{2/\alpha})} \geq p} \{x\} \leq F^{-1}(p) \leq \inf_{\widehat{F}_n(x) - \sqrt{n^{-1} \log(\sqrt{2/\alpha})} \geq p} \{x\}\right) \geq 1 - \alpha \quad (9.7.105)$$

Note that this bound is only informative when  $p$  is not too extreme, under consideration of the width of the confidence band on the CDF. That is, it is only informative when neither  $p \leq \sqrt{n^{-1} \log(\sqrt{2/\alpha})}$  nor  $p \geq 1 - \sqrt{n^{-1} \log(\sqrt{2/\alpha})}$ .

## 9.8 Robust Statistics

### 9.8.1 Robust Point Estimation

Trimmed Mean

### 9.8.2 Robust Regression

Least Absolute Deviations

Random Sample Consensus Algorithm

Theil-Sen Estimator

### 9.8.3 Sandwich Estimators

### 9.8.4 Robust Design

Taguchi Method [10]

# Chapter 10

## Stochastic Calculus

### 10.1 Continuity of Stochastic Processes

#### 10.1.1 Continuity in Mean-Square

A second-order stochastic process  $X(t)$  is said to be continuous in mean-square at time  $\tau$  if

$$\lim_{h \rightarrow 0} \mathbb{E} [|X(\tau + h) - X(\tau)|^2] = 0 \quad (10.1.1)$$

**Theorem 10.1** ([9]). *A second order stochastic process  $X(t)$  is mean-square continuous at time  $\tau$  if and only if the mean function  $m_X(t) = \mathbb{E}[X(t)]$  is continuous at  $\tau$  and the autocovariance function  $C_X(t, s) = \mathbb{E}[X(t)X(s)] - m_X(t)m_X(s)$  is continuous at  $(\tau, \tau)$ .*

*Proof.* Expanding the term  $[X(t+h) - m(t+h) - X(t) + m(t)]^2$ , we have

$$\begin{aligned} & [X(t+h) - m(t+h) - X(t) + m(t)]^2 \\ &= [X(t+h) - X(t)]^2 - 2[X(t+h) - X(t)][m(t+h) - m(t)] + [m(t+h) - m(t)]^2 \end{aligned} \quad (10.1.2)$$

Hence

$$|X(t+h) - X(t)|^2 = [X(t+h) - X(t)]^2 \quad (10.1.3)$$

$$\begin{aligned} & [X(t+h) - m(t+h) - X(t) + m(t)]^2 \\ &+ 2[X(t+h) - X(t)][m(t+h) - m(t)] - [m(t+h) - m(t)]^2 \end{aligned} \quad (10.1.4)$$

Taking the expectation of both sides,

$$\begin{aligned} \mathbb{E} [|X(t+h) - X(t)|^2] &= \mathbb{E} [|X(t+h) - X(t) - (m(t+h) - m(t))|^2] \\ &+ 2\mathbb{E} [X(t+h) - X(t)][m(t+h) - m(t)] - [m(t+h) - m(t)]^2 \end{aligned} \quad (10.1.5)$$

$$\begin{aligned} &= \text{Var}(X(t+h) - X(t)) \\ &+ 2[m(t+h) - m(t)]^2 - [m(t+h) - m(t)]^2 \end{aligned} \quad (10.1.6)$$

$$\begin{aligned} &= \text{Var}(X(t+h)) - 2\text{Cov}(X(t+h), X(t)) + \text{Var}(X(t)) + [m(t+h) - m(t)]^2 \\ &\quad (10.1.7) \end{aligned}$$

$$= C_X(t+h, t+h) - 2C_X(t+h, t) + C_X(t, t) + [m(t+h) - m(t)]^2 \quad (10.1.8)$$

Taking the limit of both sides, we find that  $\lim_{h \rightarrow 0} \mathbb{E} [|X(t+h) - X(t)|^2] = 0$  is equivalent to  $m(\cdot)$  and  $C_X(\cdot, \cdot)$  being continuous, since the terms  $[m(t+h) - m(t)]^2$  and  $C_X(t+h, t+h) - 2C_X(t+h, t) + C_X(t, t) = \text{Var}(X(t+h) - X(t))$  are non-negative, meaning that they both must converge to zero.  $\square$

### 10.1.2 Continuity in Probability

Consider a stochastic process  $X : T \times \Omega \rightarrow \mathbb{R}$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The process  $X$  is said to be continuous in probability at time  $\tau \in T$  if for all  $\varepsilon > 0$ :

$$\lim_{h \rightarrow 0} \mathbb{P}(\{\omega \in \Omega : |X(\tau + h, \omega) - X(\tau, \omega)| \geq \varepsilon\}) = 0 \quad (10.1.9)$$

### 10.1.3 Continuity in Distribution

A stochastic process  $X : T \times \Omega \rightarrow \mathbb{R}$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is said to be continuous in distribution at time  $\tau \in T$  if

$$\lim_{h \rightarrow 0} F_{\tau+h}(x) = F_\tau(x) \quad (10.1.10)$$

at all points  $x$  which  $F_\tau(x)$  is continuous, where  $F_\tau(x)$  denotes the cumulative distribution of  $X(\tau, \cdot)$ .

### 10.1.4 Continuity with Probability One

A stochastic process  $X : T \times \Omega \rightarrow \mathbb{R}$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is said to be continuous with probability one at time  $\tau \in T$  if

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{h \rightarrow 0} |X(\tau + h, \omega) - X(\tau, \omega)| = 0\right\}\right) = 1 \quad (10.1.11)$$

### 10.1.5 Sample Continuity

A stochastic process  $X : T \times \Omega \rightarrow \mathbb{R}$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is said to be sample continuous if sample paths  $X(\cdot, \omega)$  are continuous for  $\mathbb{P}$ -almost all  $\omega \in \Omega$  (i.e. almost surely with respect to the measure  $\mathbb{P}$ ).

### 10.1.6 càdlàg Stochastic Processes

#### 10.1.7 Kolmogorov-Chentsov Continuity Theorem [12, 16]

The Kolmogorov-Chentsov continuity theorem provides sufficient conditions on a stochastic process such that the process is sample continuous, or has a ‘continuous version’ that is sample continuous. Formally, consider a stochastic process  $X : T \times \Omega \rightarrow \mathbb{R}$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Another stochastic process  $Y : T \times \Omega \rightarrow \mathbb{R}$  on the same probability space is said to be a modification of  $X$  if for each  $\tau \in T$ :

$$\mathbb{P}(\{\omega \in \Omega : X(\tau, \omega) = Y(\tau, \omega)\}) = 1 \quad (10.1.12)$$

This implies that  $X$  and  $Y$  will have the same finite dimensional distributions. If there are constants  $\alpha, \beta, c > 0$  such that

$$\mathbb{E}[|X(\tau, \cdot) - X(s, \cdot)|^\alpha] \leq c|\tau - s|^{1+\beta} \quad (10.1.13)$$

for all  $0 \leq s < \tau$ , then there exists a modification  $\tilde{X}$  of  $X$  that is sample continuous.

## 10.2 Mean-Square Stochastic Calculus

### 10.2.1 Differentiability in Mean-Square

A second-order stochastic process  $X(t)$  is said to be differentiable in mean-square at time  $\tau$  with derivative  $X'(\tau)$  if the limit

$$\lim_{k \rightarrow 0} \left[ \frac{X(\tau + k) - X(\tau)}{k} \right] = X'(\tau) \quad (10.2.1)$$

exists in the sense that

$$\lim_{h \rightarrow 0} \mathbb{E} \left[ \left| \frac{X(\tau+h) - X(\tau)}{h} - X'(\tau) \right|^2 \right] = 0 \quad (10.2.2)$$

**Theorem 10.2** ([9]). A second-order stochastic process  $X(t)$  is differentiable in mean-square at time  $\tau$  if and only if the mean function  $m_X(t) = \mathbb{E}[X(t)]$  is differentiable at  $\tau$  and the mixed second partial derivative of the covariance function  $\frac{\partial^2 C_X(t, s)}{\partial t \partial s}$  exists at  $(\tau, \tau)$ .

*Proof.* Firstly note that the mixed second partial derivative may be defined as

$$\frac{\partial^2 C_X(t, s)}{\partial t \partial s} = \frac{\partial}{\partial s} \lim_{h \rightarrow 0} \left[ \frac{C_X(t+h, s) - C_X(t, s)}{h} \right] \quad (10.2.3)$$

$$= \lim_{k \rightarrow 0} \lim_{h \rightarrow 0} \left[ \frac{1}{k} \left( \frac{C_X(t+h, s+k) - C_X(t, s+k)}{h} - \frac{C_X(t+h, s) - C_X(t, s)}{h} \right) \right] \quad (10.2.4)$$

$$= \lim_{h, k \rightarrow 0} \left[ \frac{C_X(t+h, s+k) - C_X(t, s+k) - C_X(t+h, s) + C_X(t, s)}{hk} \right] \quad (10.2.5)$$

Now, via quadratic expansion, we can write

$$\begin{aligned} & \left[ \frac{X(\tau+h) - X(\tau)}{h} - \frac{X(\tau+k) - X(\tau)}{k} \right]^2 \\ &= \left( \frac{X(\tau+h) - X(\tau)}{h} \right)^2 - 2 \left( \frac{X(\tau+h) - X(\tau)}{h} \right) \left( \frac{X(\tau+k) - X(\tau)}{k} \right) + \left( \frac{X(\tau+k) - X(\tau)}{k} \right)^2 \end{aligned} \quad (10.2.6)$$

Take the expectation of the central term on the right-hand side and using the definition of covariance, write

$$\begin{aligned} & \mathbb{E} \left[ \left( \frac{X(\tau+h) - X(\tau)}{h} \right) \left( \frac{X(\tau+k) - X(\tau)}{k} \right) \right] \\ &= \text{Cov} \left( \frac{X(\tau+h) - X(\tau)}{h}, \frac{X(\tau+k) - X(\tau)}{k} \right) + \mathbb{E} \left[ \frac{X(\tau+h) - X(\tau)}{h} \right] \mathbb{E} \left[ \frac{X(\tau+k) - X(\tau)}{k} \right] \end{aligned} \quad (10.2.7)$$

which becomes

$$\begin{aligned} & \mathbb{E} \left[ \left( \frac{X(\tau+h) - X(\tau)}{h} \right) \left( \frac{X(\tau+k) - X(\tau)}{k} \right) \right] \\ &= \frac{C_X(\tau+h, \tau+k) - C_X(\tau, \tau+k) - C_X(\tau+h, \tau) + C_X(\tau, \tau)}{hk} \\ & \quad + \left( \frac{m(\tau+h) - m(\tau)}{h} \right) \left( \frac{m(\tau+k) - m(\tau)}{k} \right) \end{aligned} \quad (10.2.8)$$

Taking the limit, we get

$$\lim_{h, k \rightarrow 0} \mathbb{E} \left[ \left( \frac{X(\tau+h) - X(\tau)}{h} \right) \left( \frac{X(\tau+k) - X(\tau)}{k} \right) \right] = \frac{\partial^2 C_X(t, s)}{\partial t \partial s} \Big|_{s, t=\tau} + \frac{dm(t)}{dt} \Big|_{t=\tau} \quad (10.2.9)$$

We can do the same for the other terms in the quadratic expansion, but since we have already done the more general case (the other terms are just a special case of  $h = k$ ), we will find that

$$\lim_{h \rightarrow 0} \mathbb{E} \left[ \left( \frac{X(\tau+h) - X(\tau)}{h} \right)^2 \right] = \frac{\partial^2 C_X(t, s)}{\partial t \partial s} \Big|_{s, t=\tau} + \frac{dm(t)}{dt} \Big|_{t=\tau} \quad (10.2.10)$$

$$\lim_{k \rightarrow 0} \mathbb{E} \left[ \left( \frac{X(\tau+k) - X(\tau)}{k} \right)^2 \right] = \frac{\partial^2 C_X(t, s)}{\partial t \partial s} \Big|_{s, t=\tau} + \frac{dm(t)}{dt} \Big|_{t=\tau} \quad (10.2.11)$$

As long as  $\frac{\partial^2 C_X(t, s)}{\partial t \partial s} \Big|_{s, t=\tau}$  and  $\frac{dm(t)}{dt} \Big|_{t=\tau}$  exist, then via cancellations this means

$$\lim_{k \rightarrow 0} \mathbb{E} \left[ \left( \frac{X(\tau+h) - X(\tau)}{h} - \frac{X(\tau+k) - X(\tau)}{k} \right)^2 \right] = 0 \quad (10.2.12)$$

To show the reverse direction, we argue that the derivatives must exist in order for them to cancel out and make the limit zero.  $\square$

### 10.2.2 Integrability in Mean-Square

Let  $X(t)$  be a second order stochastic process, and consider a time interval  $[a, b] \in T$ . Let the time indices denoted by  $a = t_0 < t_1 < \dots < t_n = b$  be a subdivision of  $[a, b]$ . We consider the sum

$$\mathcal{I}_n = \sum_{k=1}^n X(\tau_k)(t_k - t_{k-1}) \quad (10.2.13)$$

such that  $t_{k-1} \leq \tau_k \leq t_k$ . The stochastic process is said to be mean-square Riemann integrable if there exists a limit in mean-square as  $n \rightarrow \infty$  in such a way that the subdivisions get arbitrarily close together, i.e.

$$\lim_{n \rightarrow \infty} \max_{1 \leq k \leq n} |t_k - t_{k-1}| = 0 \quad (10.2.14)$$

This limit is the mean-square Riemann integral over  $[a, b]$  and is denoted by

$$\mathcal{I} = \int_a^b X(\tau) d\tau \quad (10.2.15)$$

**Theorem 10.3** ([9, 29]). *A second-order stochastic process  $X(t)$  is mean-square Riemann integrable on  $[a, b]$  if and only if the mean function  $\mu_X(t) = \mathbb{E}[X(t)]$  is Riemann integrable on  $[a, b]$  and the autocovariance function  $C_X(t, s) = \text{Cov}(X(t), X(s))$  is Riemann integrable on  $[a, b] \times [a, b]$ .*

*Proof.* Let  $\mathcal{I}_n$  and  $\mathcal{I}_m$  be the Riemann sums associated with two subdivisions (or ‘partition sequences’) of the time interval  $[a, b]$ , denoted  $\{t_k\}_n$  and  $\{t_j\}_m$  respectively. We can write

$$\mathbb{E}[(\mathcal{I}_n - \mathcal{I}_m)^2] = \mathbb{E}[\mathcal{I}_n^2] - 2\mathbb{E}[\mathcal{I}_n \mathcal{I}_m] + \mathbb{E}[\mathcal{I}_m] \quad (10.2.16)$$

Mean-square integrability of  $X(t)$  over  $[a, b]$  requires that  $\lim_{n, m \rightarrow \infty} \mathbb{E}[(\mathcal{I}_n - \mathcal{I}_m)^2] = 0$  for any pair or partition sequences  $\{t_k\}_n, \{t_j\}_m$ . This condition is equivalent to the right-hand side also having a limit of zero, which just requires that  $\lim_{n, m \rightarrow \infty} \mathbb{E}[\mathcal{I}_n \mathcal{I}_m]$  exists for any  $\{t_k\}_n, \{t_j\}_m$  (the other terms can be treated as special cases where  $n = m$ ). Since we can write

$$\mathbb{E}[\mathcal{I}_n \mathcal{I}_m] = \text{Cov}(\mathcal{I}_n, \mathcal{I}_m) + \mathbb{E}[\mathcal{I}_n] \mathbb{E}[\mathcal{I}_m] \quad (10.2.17)$$

$$\begin{aligned} &= \sum_{k=1}^n \sum_{j=1}^m C_X(\tau_k, \tau_j)(t_k - t_{k-1})(t_j - t_{j-1}) \\ &\quad + \left( \sum_{k=1}^n \mu_X(\tau_k)(t_k - t_{k-1}) \right) \left( \sum_{j=1}^m \mu_X(\tau_j)(t_j - t_{j-1}) \right) \end{aligned} \quad (10.2.18)$$

Then this becomes equivalent to the mean function being Riemann integrable on  $[a, b]$  and the autocovariance function being Riemann integrable on  $[a, b] \times [a, b]$ .  $\square$

---

### 10.2.3 Mean Square Stochastic Differential Equations [190]

## 10.3 Continuous-Time Martingales

### 10.3.1 Semimartingales

### 10.3.2 Doob-Meyer Decomposition Theorem

## 10.4 Itô Calculus

### 10.4.1 Itô Integral [38]

Suppose  $f(W(t), t)$ , which is a functional (i.e. function of a function) of the Wiener process  $W(t)$ , is an adapted process. Roughly speaking, this means that the process  $f(W(t), t)$  does not depend on information from the future. Consider the sum

$$S_n = \sum_{i=1}^n f(W(t_{i-1}), t_{i-1}) \cdot (W(t_i) - W(t_{i-1})) \quad (10.4.1)$$

where the  $t_i$  are in the interval  $[a, b]$ , with  $a \geq 0$  such that

$$a = t_0 < t_1 < \dots < t_n = b \quad (10.4.2)$$

For simplicity, it suffices to treat these times as being equidistant:

$$t_i = (b - a) \frac{i}{n} \quad (10.4.3)$$

The sum resembles a Riemann sum, and in the same way that we can take the limit of a Riemann sum to define the Riemann integral, we can take the limit of  $S_n$  to define a stochastic integral. The Itô integral of  $f(W(t), t)$  with respect to the Wiener process is defined as

$$\int_a^b f(W(t), t) dW(t) = \lim_{n \rightarrow \infty} S_n \quad (10.4.4)$$

Note that the integral is itself a random variable, which is what  $S_n$  converges in probability to.

### 10.4.2 Itô Processes

### 10.4.3 Itô's Lemma

## 10.5 Stratonovich Integral

## 10.6 Stochastic Differential Equations

### 10.6.1 Diffusions

### 10.6.2 Stochastic Partial Differential Equations

### 10.6.3 Backward Stochastic Differential Equations

## 10.7 Malliavin Calculus

## 10.8 Numerical Stochastic Differential Equations

### 10.8.1 Euler–Maruyama Method

### 10.8.2 Milstein Method

### 10.8.3 Runge-Kutta Method

## Chapter 11

# Probabilistic Combinatorics

### 11.1 Stirling's Approximation

### 11.2 Inclusion-Exclusion Principle

The inclusion-exclusion principle generalises the method of counting the elements in the union of two finite sets, namely

$$|A \cup B| = |A| + |B| - |A \cap B| \quad (11.2.1)$$

and thus can be used to generalise the addition rule of probability. Let  $A_1, \dots, A_n$  each be a finite set. Then the inclusion-exclusion principle expresses the cardinality of the union of all the sets, in terms of the cardinalities of intersections between the sets:

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{k=i \leq n} |A_i| - \sum_{i \leq i < j \leq n} |A_i \cap A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cap A_j \cap A_k| - \dots + (-1)^{n-1} |A_1 \cap \dots \cap A_n| \quad (11.2.2)$$

$$= \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} |A_{i_1} \cap \dots \cap A_{i_k}| \right) \quad (11.2.3)$$

This is a sum involving the intersection of every combination  $k$  from  $n$  sets, for all  $k = 1, \dots, n$ . Thus the  $k^{\text{th}}$  term in the outer sum will itself have  $\binom{n}{k}$  summands.

*Proof.* Consider a single element contained in  $\bigcup_{i=1}^n A_i$ , and since the ordering of sets is arbitrary, assume without loss of generality that this element is contained in all the sets  $A_1, \dots, A_t$  for some  $t \in \{1, \dots, n\}$ . If the inclusion-exclusion principle indeed holds, then we need to show that this element is counted precisely once in the sum  $\sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} |A_{i_1} \cap \dots \cap A_{i_k}| \right)$ . Since  $|A_{i_1} \cap \dots \cap A_{i_k}|$  is at most one when all of  $i_1, \dots, i_k \leq t$  and zero otherwise, we can count using indicators:

$$\sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbb{I}_{\{i_1, \dots, i_k \leq t\}} \right) = \binom{t}{1} - \binom{t}{2} + \dots + (-1)^{t+1} \binom{t}{t} \quad (11.2.4)$$

since there are  $\binom{t}{k}$  such indices from  $1, \dots, n$  such that  $k$  of them are less than or equal to  $t$ . Using the binomial theorem for expansion, we can show for the right-hand side that

$$0 = (1 - 1)^t \quad (11.2.5)$$

$$= 1^t \binom{t}{0} + 1^{t-1} (-1) \binom{t}{1} + \dots + (-1)^t \binom{t}{t} \quad (11.2.6)$$

$$= 1 + (-1) \left[ \binom{t}{1} + \cdots + (-1)^{t+1} \binom{t}{t} \right] \quad (11.2.7)$$

Rearranging,

$$\binom{t}{1} + \cdots + (-1)^{t+1} \binom{t}{t} = 1 \quad (11.2.8)$$

as required.  $\square$

### 11.2.1 Probabilistic Inclusion-Exclusion Principle

Letting  $A_1, \dots, A_n$  be events in some probability space, we can divide out the inclusion-exclusion principle by the cardinality of the sample space to obtain the probabilistic version of the inclusion-exclusion principle:

$$\begin{aligned} \Pr \left( \bigcup_{i=1}^n A_i \right) &= \sum_{k=i \leq n} \Pr(A_i) - \sum_{i \leq i < j \leq n} \Pr(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} \Pr(A_i \cap A_j \cap A_k) - \dots \\ &\quad + (-1)^{n-1} \Pr(A_1 \cap \dots \cap A_n) \end{aligned} \quad (11.2.9)$$

$$= \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} \Pr(A_{i_1} \cap \dots \cap A_{i_k}) \right) \quad (11.2.10)$$

With  $n = 3$ , we obtain the addition rule of probability for three events:

$$\begin{aligned} \Pr(A \cup B \cup C) &= \Pr(A) + \Pr(B) + \Pr(C) \\ &\quad - \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) + \Pr(A \cap B \cap C) \end{aligned} \quad (11.2.11)$$

Another special case is when the probability  $\Pr(A_{i_1} \cap \dots \cap A_{i_k})$  only depends on  $k$  and not on the particular events being intersected, say

$$\Pr(A_{i_1} \cap \dots \cap A_{i_k}) = p_k \quad (11.2.12)$$

for each  $k$ . Then we can write

$$\Pr \left( \bigcup_{i=1}^n A_i \right) = \sum_{k=1}^n (-1)^{k+1} \binom{n}{k} p_k \quad (11.2.13)$$

### 11.2.2 Complementary Inclusion-Exclusion Principle

Combining with DeMorgan's laws, we obtain the complementary form of the inclusion-exclusion principle, where we can count the intersection of complements by

$$\left| \bigcap_{i=1}^n \overline{A}_i \right| = \left| S \setminus \bigcup_{i=1}^n A_i \right| \quad (11.2.14)$$

$$= |S| - \left| \bigcup_{i=1}^n A_i \right| \quad (11.2.15)$$

$$= |S| - \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} |A_{i_1} \cap \dots \cap A_{i_k}| \right) \quad (11.2.16)$$

where  $S$  is the universal set. Alternatively, we count the intersection of the  $A_i$  by

$$\left| \bigcap_{i=1}^n A_i \right| = |S| - \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} |\bar{A}_{i_1} \cap \dots \cap \bar{A}_{i_k}| \right) \quad (11.2.17)$$

Translated to probability, this implies

$$\Pr \left( \bigcap_{i=1}^n \bar{A}_i \right) = 1 - \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} \Pr(A_{i_1} \cap \dots \cap A_{i_k}) \right) \quad (11.2.18)$$

or

$$\Pr \left( \bigcap_{i=1}^n A_i \right) = 1 - \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} \Pr(\bar{A}_{i_1} \cap \dots \cap \bar{A}_{i_k}) \right) \quad (11.2.19)$$

In the special case that  $\Pr(A_{i_1} \cap \dots \cap A_{i_k}) = p_k$ , we have

$$\Pr \left( \bigcap_{i=1}^n \bar{A}_i \right) = 1 - \sum_{k=1}^n (-1)^{k+1} \binom{n}{k} p_k \quad (11.2.20)$$

$$= \sum_{k=0}^n (-1)^k \binom{n}{k} p_k \quad (11.2.21)$$

where we take  $p_0 = 1$ , which is natural, as we can view  $p_k$  equivalently as

$$p_k = \Pr(S \cap A_{i_1} \cap \dots \cap A_{i_k}) \quad (11.2.22)$$

so  $p_0 = \Pr(S) = 1$ .

### 11.2.3 Bonferroni Inequalities [119]

## 11.3 Pigeonhole Principle

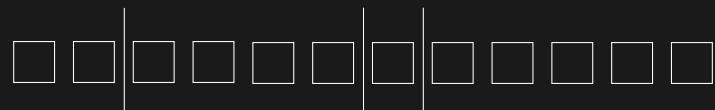
## 11.4 Partitions

### 11.4.1 Integer Compositions

Consider the positive integer  $n$ , composed of the sum of  $k \leq n$  positive integers by

$$n = n_1 + n_2 + \dots + n_k \quad (11.4.1)$$

where  $n_1, \dots, n_k \geq 1$  (i.e. there are no ‘empty’ partitions). We ask how many different distinct integer sequences  $(n_1, \dots, n_k)$  there are. Picture a ‘block’ of  $n$  objects, with spaces in between each object. There are  $n-1$  such spaces (since the sides are excluded). We then imagine placing  $k-1$  ‘separators’ among these spaces, in order to divide the block into  $k$  partitions.



Then the number of ways to place these separators is the number of possible partitions, given by

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(k-1)! (n-k)!} \quad (11.4.2)$$

Thus for an integer  $n$ , the total number of compositions (where anywhere between the sum of 1 to  $n$  integers is allowed) is given by

$$\sum_{k=1}^n \binom{n-1}{k-1} = \sum_{k=0}^{n-1} \binom{n-1}{k} \quad (11.4.3)$$

$$= 2^{n-1} \quad (11.4.4)$$

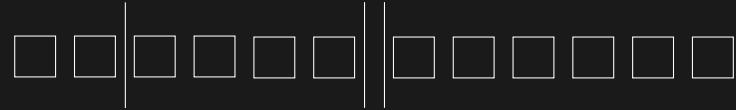
using the identity for the sum of binomial coefficients.

### Weak Integer Compositions

If instead weak compositions are allowed (i.e.  $n_1, \dots, n_k \geq 0$ ), then we can extend the number of compositions above by enumerating through possibilities where some of the blocks being separated are empty. Note that if we allow for  $n_1, \dots, n_k \geq 0$ , there can be as few as 1 and as many as  $k - 1$  empty blocks being separated (since the sum must still equal  $n$ ). For an arbitrary  $i$  empty blocks being separated, there are  $\binom{k}{i}$  different possibilities, and for each of these possibilities we can count the integer compositions of  $n$  for a sequence of  $k - i$  integers. Hence the number of ways is given by

$$\binom{n-1}{k-1} + \sum_{i=1}^{k-1} \binom{k}{i} \times \binom{n-1}{k-1-i} = \sum_{i=0}^{k-1} \binom{k}{i} \times \binom{n-1}{k-1-i} \quad (11.4.5)$$

Another way to count weak integer compositions is to now imagine arranging  $k - 1$  among  $n$ , where we are allowed to have no blocks between separators (corresponding to a zero term). The total number of ways to arrange  $k - 1$  separators among  $n + k - 1$  total objects is  $\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$ .



Comparing with the other method of counting above, this reveals the identity

$$\binom{n+k-1}{k-1} = \sum_{i=0}^{k-1} \binom{k}{i} \times \binom{n-1}{k-1-i} \quad (11.4.6)$$

Also, the formula  $\binom{n+k-1}{k-1} = \binom{k+n-1}{n}$  is the same as the number of unordered ways to sample  $n$  objects from a total of  $k$  (noting that  $n \leq k$ ), with replacement.

### Integer Partitions

If integer compositions cannot be distinguished by the ordering of the sequence  $(n_1, \dots, n_k)$  (i.e. order of summation does not matter), then these are known as the integer partitions of  $n$ . We are effectively counting the number of sets of positive integers  $\{n_1, \dots, n_k\}$  such that  $n_1 + \dots + n_k = n$ .

#### 11.4.2 Stirling Numbers of the Second Kind

The number of ways to partition a set of  $n$  labelled objects into  $k$  non-empty unlabelled subsets is known as a Stirling number of the second kind, and denoted  $\left\{ \begin{smallmatrix} n \\ k \end{smallmatrix} \right\}$ . In other words, this counts number of ways to deposit  $n$  objects labelled  $1, \dots, n$  into  $k$  unlabelled bins. Note that this

number is generally different from the number of ways to partition an integer. A formula for  $\{n\}_k$  is

$$\{n\}_k = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^n \quad (11.4.7)$$

To derive this formula, we can use the inclusion-exclusion principle, in particular the special case of the complementary form. Consider the universal set  $S$  as all the partitions of the  $n$  objects into  $k$  distinguishable (i.e. labelled) and possibly empty boxes, for which there will be  $|S| = k^n$  elements. Let  $A_j$  denote the set of all partitions where the  $j^{\text{th}}$  box is empty. Since the Stirling number counts partitions into indistinguishable non-empty boxes (i.e. partitions are equivalent up to a re-ordering of the boxes), we can count the number of partitions in the former quantity by

$$k! \{n\}_k = \left| \bigcap_{j=1}^k \overline{A}_j \right| \quad (11.4.8)$$

i.e. the number of elements in the set where all the boxes are non-empty. Using the complementary form of the inclusion-exclusion principle, we can write

$$\left| \bigcap_{j=1}^k \overline{A}_j \right| = |S| - \sum_{i=1}^k (-1)^{i+1} \left( \sum_{1 \leq j_1 < \dots < j_i \leq n} |A_{j_1} \cap \dots \cap A_{j_i}| \right) \quad (11.4.9)$$

$$= \sum_{k=0}^k (-1)^i \left( \sum_{1 \leq j_1 < \dots < j_i \leq n} |S \cap A_{j_1} \cap \dots \cap A_{j_i}| \right) \quad (11.4.10)$$

The term  $|S \cap A_{j_1} \cap \dots \cap A_{j_i}|$  is the number of ways that  $i$  of the  $k$  boxes are empty, and so given by

$$|S \cap A_{j_1} \cap \dots \cap A_{j_i}| = (k-i)^n \quad (11.4.11)$$

As the cardinality depends on  $i$  and not on the indices  $j_1, \dots, j_i$ , this simplifies to

$$\left| \bigcap_{i=1}^k \overline{A}_i \right| = \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^n \quad (11.4.12)$$

whereby we can rearrange for  $\{n\}_k$  to obtain the formula.

### Coupon Collection Probabilities

Suppose there are  $N$  labelled objects, from which we sample  $n$  times with replacement. Then there will be  $N^n$  possible sequences of samples. An alternative way to count this is by using Stirling numbers of the second kind. Consider a subset of  $k$  of the objects; if all the samples fall within this subset, then there are  $\{n\}_k$  ways in which they can fall without considering the order of the subsets, or  $k! \{n\}_k$  ways when considering the order of the subsets. Also, there are  $\binom{N}{k}$  ways to pick a subset of  $k$  from the  $N$  objects. Since order matters in the  $N^n$  possible sequences, then via this counting logic, we have the identity

$$\sum_{k=0}^n k! \{n\}_k \binom{N}{k} = N^n \quad (11.4.13)$$

or alternatively,

$$\sum_{k=0}^n \{n\}_k {}^N P_k = N^n \quad (11.4.14)$$

This identity works for both cases  $n \geq N$  and  $N \geq n$ , since if  $k > N$  we have  $\binom{N}{k} = 0$ , or if  $N \geq n$ , then there can be at most  $n$  different objects in the sample. Similar to Vandermonde's identity, we can derive a discrete distribution directly from this identity, since

$$\sum_{k=0}^n \frac{k!}{N^n} \left\{ \begin{matrix} n \\ k \end{matrix} \right\} \binom{N}{k} = 1 \quad (11.4.15)$$

In this case, consider there to be  $N$  distinct coupons and we sample  $n$  coupons uniformly with replacement. Let  $X$  be a random variable for the number of coupons that we have collected at least once. Then the distribution of  $X$  is given by

$$\Pr(X = k) = \frac{k!}{N^n} \left\{ \begin{matrix} n \\ k \end{matrix} \right\} \binom{N}{k} \quad (11.4.16)$$

### 11.4.3 Bell Numbers

The  $n^{\text{th}}$  Bell number, denoted  $B_n$ , counts all the possible ways to partition  $n$  labelled objects into unlabelled non-empty subsets (without respect to the number of subsets). Hence the Bell numbers are related to Stirling numbers of the second kind by

$$B_n = \sum_{k=0}^n \left\{ \begin{matrix} n \\ k \end{matrix} \right\} \quad (11.4.17)$$

### Dobiński's Formula

## 11.5 Catalan Numbers [119]

### 11.6 Derangements

A rearrangement of a set such that no element appears in its original position is called a derangement. The number of derangements of a set of size  $n$  is denoted  $!n$ . The number of derangements can be counted by considering the following problem: each element in  $\{1, \dots, n\}$  is to be assigned a label from  $1, \dots, n$ . We are interested in the number of ways to assign labels in which no element is assigned its own label. That is, each element has exactly one 'forbidden' label which it cannot be assigned. Suppose element 1 is assigned an arbitrary label  $i \neq 1$ . There are  $n - 1$  such possibilities. Then we can group the remaining possibilities pertaining to element  $i$  as follows:

- Element  $i$  is reciprocally assigned label 1. Then the remaining possibilities becomes the same as with  $n - 2$  elements and  $n - 2$  labels.
- Element  $i$  is not assigned label 1. Then remaining possibilities becomes the same as with  $n - 1$  elements and  $n - 1$  labels, because each element still has exactly one forbidden label (label 1 acts as element  $i$ 's 'forbidden' label, because it can no longer be assigned it).

Hence the recurrence relation is obtained for the number of derangements  $!n$ :

$$!n = (n - 1) [! (n - 2) + ! (n - 1)] \quad (11.6.1)$$

Letting  $!0 = 1$  and  $!1 = 0$ , then this defines the entirety of the sequence in  $n$  for the number of derangements.

An explicit formula for the number of derangements can also be obtained via the inclusion-exclusion principle. In a rearrangement of a set, we call an element which remains in its original

position ‘fixed’. There are  $n!$  possibilities in which there are at least zero fixed elements (same as the number of arrangements). There are  $\binom{n}{1}(n-1)!$  possibilities in which there are at least one fixed element (because there are  $\binom{n}{1}$  ways to choose one fixed element and  $(n-1)!$  arrangements of the remaining  $n-1$  elements). Hence we can develop the general formula that there are  $\binom{n}{i}(n-i)!$  arrangements in which there are at least  $i$  fixed elements. Applying the inclusion exclusion principle:

$$!n = n! - \binom{n}{1}(n-1)! + \binom{n}{2}(n-2)! - \dots \quad (11.6.2)$$

$$= \sum_{i=0}^n (-1)^i \binom{n}{i} (n-i)! \quad (11.6.3)$$

$$= \sum_{i=0}^n (-1)^i \frac{n!}{i!(n-1)!} (n-i)! \quad (11.6.4)$$

$$= n! \sum_{i=0}^n \frac{(-1)}{i!} \quad (11.6.5)$$

## 11.7 Twelvefold Way

## 11.8 Probabilisitic Method

## 11.9 Random Graphs

### 11.9.1 Erdős-Rényi Graphs

## Part II

# Applications

## Chapter 12

# Information Theory

### 12.1 Entropy

#### 12.1.1 Shannon Entropy

For a discrete random variable  $X$  with probability mass function  $P(x)$ , the *self-information* of the event  $\{X = x\}$  is defined as

$$I(x) = \log \frac{1}{P(x)} \quad (12.1.1)$$

The Shannon entropy or *information entropy* of a discrete random variable  $X$  with probability mass function  $P(x)$  can be defined as the expected self-information:

$$H[X] = \mathbb{E}[-\log P(X)] \quad (12.1.2)$$

$$= - \sum_{i=1}^n P(x_i) \log P(x_i) \quad (12.1.3)$$

$$= \sum_{i=1}^n P(x_i) \log \frac{1}{P(x_i)} \quad (12.1.4)$$

Note that since  $0 \leq P(x_i) \leq 1$ , then entropy is non-negative (and when  $P(x_i) \rightarrow 0$ , the term in the summation approaches zero). If the logarithm used is base 2, then the units of entropy is measured in bits. If the logarithm is the natural logarithm, then the units are in nats. There are various interpretations of entropy.

- Entropy can be thought of as a measure of uncertainty (in a different way to variance). The larger the entropy, the more uncertainty in the random variable. A deterministic variable is the most certain kind of random variable, and has an entropy of zero (minimum uncertainty) since  $\log 1 = 0$ .
- Entropy measured in bits can be thought of as the lower bound on the number of bits it requires to transmit/store the outcome of an experiment. For example, in the deterministic case we require no bits because we already know the outcome of all experiments. Consider the outcome of a fair coin toss, with probability mass function

$$\Pr(X = x) = \begin{cases} 1/2, & x = 0 \\ 1/2, & x = 1 \\ 0, & \text{elsewhere} \end{cases} \quad (12.1.5)$$

The calculation of entropy yields  $H[X] = 1$  bit. This agrees with the intuition that it requires a minimum of 1 bit to convey the outcome of a fair coin toss (i.e. 0 for tails, 1

for heads). In the case of an unfair coin, for example with the probability mass function

$$\Pr(X = x) = \begin{cases} 1/4, & x = 0 \\ 3/4, & x = 1 \\ 0, & \text{elsewhere} \end{cases} \quad (12.1.6)$$

The calculation of entropy for this distribution yields  $H[X] \approx 0.8113$  bits. This is a lower bound, so rounded up gives 1 bit. We still require at least 1 bit to convey the outcome of an unfair coin toss because the outcome is still binary (either we get heads or we do not).

Now imagine an experiment with four possible outcomes distributed uniformly (e.g. two fair coin tosses). The calculation of entropy does not depend on the events, only the associated probabilities. So in the case the probabilities are each  $1/4$ , the entropy works out to be 2 bits. Once again this agrees with the intuition that it would take a minimum of 2 bits to store the outcome of two fair coin tosses.

- Lastly, entropy can also be thought of as the expected value of a random variable that expresses the amount of information contained in each event (the random variable in question being the self-information  $-\log P(X)$ ). To illustrate, consider the random variable with distribution

$$\Pr(X = x) = \begin{cases} \frac{1}{n}, & x = 0 \\ \frac{n-1}{n}, & x = 1 \\ 0, & \text{elsewhere} \end{cases} \quad (12.1.7)$$

for some large integer  $n$ . We regard the occurrence of the event  $X = 0$  as containing more information than the occurrence of the event  $X = 1$ , since it is rarer and hence more of a ‘surprise’ (the occurrence of  $X = 0$  says more about the experiment than the occurrence  $X = 1$ ). This is reflected in  $-\log \frac{1}{n} = \log n > -\log \frac{n-1}{n} = \log \frac{n}{n-1}$ .

### 12.1.2 Differential Entropy

The continuous analog of information entropy is differential entropy, where the discrete sum is converted to an integral over the probability density function  $f(x)$ .

$$h(X) = - \int f(x) \log f(x) dx \quad (12.1.8)$$

Note however that this measure of entropy does not share all the same properties as information entropy defined above. One such property is non-negativity. Since  $f(x)$  may be greater than 1, then it is possible for  $-\log f(x) < 0$ . Despite this, we can still use differential entropy as a measure of uncertainty.

### 12.1.3 Joint Entropy

For a pair of discrete random variables  $X$  and  $Y$  with joint probability mass function  $p(x, y)$ , the joint entropy is defined as

$$H[X, Y] = -\mathbb{E}[\log p(X, Y)] \quad (12.1.9)$$

$$= \sum_x \sum_y p(x, y) \log \frac{1}{p(x, y)} \quad (12.1.10)$$

If  $X$  and  $Y$  are independent, then  $p(x, y) = p(x)p(y)$  so

$$H[X, Y] = \sum_x \sum_y p(x)p(y) \log \frac{1}{p(x)p(y)} \quad (12.1.11)$$

$$= \sum_x \sum_y p(x)p(y) \left( \log \frac{1}{p(x)} + \log \frac{1}{p(y)} \right) \quad (12.1.12)$$

$$= \sum_x p(x) \log \frac{1}{p(x)} \underbrace{\sum_y p(y)}_{=1} + \sum_y p(y) \log \frac{1}{p(y)} \underbrace{\sum_x p(x)}_{=1} \quad (12.1.13)$$

$$= H[X] + H[Y] \quad (12.1.14)$$

Joint entropy generalises to discrete random vectors  $\mathbf{X}$ , so that

$$H[\mathbf{X}] = -\mathbb{E}[\log p(\mathbf{X})] \quad (12.1.15)$$

Furthermore, if the components of  $\mathbf{X} = (X_1, \dots, X_n)$  are i.i.d. copies of  $X$ , we can similarly show

$$H[\mathbf{X}] = n H[X] \quad (12.1.16)$$

#### 12.1.4 Conditional Entropy

For a pair of discrete random variables  $X$  and  $Y$  with joint probability mass function  $p(x, y)$  and conditional distribution  $p(y|x)$ , the conditional entropy is defined as

$$H[Y|X] = -\mathbb{E}[\log p(Y|X)] \quad (12.1.17)$$

$$= \sum_x \sum_y p(x, y) \log \frac{1}{p(y|x)} \quad (12.1.18)$$

$$= \sum_x p(x) \sum_y p(y|x) \log \frac{1}{p(y|x)} \quad (12.1.19)$$

We can interpret  $H[Y|X]$  as like the information content of  $Y$  contained in  $X$ . If  $X$  and  $Y$  are independent, then  $p(y|x) = p(y)$  and

$$H[Y|X] = H[Y] \quad (12.1.20)$$

which aligns with the notion that the conditional entropy is unchanged if  $X$  contains no information about  $Y$ . The conditional entropy of  $X$  on itself is

$$H[X|X] = 0 \quad (12.1.21)$$

since  $p(x|x) = \Pr(X=x|X=x) = 1$ . This is in agreement with the intuition that knowing  $X$  reveals perfect information about  $X$ .

#### 12.1.5 Chain Rule of Entropy

The chain rule of entropy says that

$$H[X, Y] = H[X] + H[Y|X] \quad (12.1.22)$$

*Proof.*

$$H[X, Y] = \sum_x \sum_y p(x, y) \log \frac{1}{p(x, y)} \quad (12.1.23)$$

$$= \sum_x \sum_y p(x, y) \log \frac{1}{p(y|x)p(x)} \quad (12.1.24)$$

$$= - \sum_x \sum_y p(x, y) \log p(x) - \sum_x \sum_y p(x, y) \log p(y|x) \quad (12.1.25)$$

$$= - \sum_x p(x) \log p(x) - \sum_x \sum_y p(x, y) \log p(y|x) \quad (12.1.26)$$

$$= H[X] + H[Y|X] \quad (12.1.27)$$

□

A corollary is that if  $X$  and  $Y$  are independent,

$$H[X, Y] = H[X] + H[Y] \quad (12.1.28)$$

and if  $X$  and  $Y$  are independent and identically distributed,

$$H[X, Y] = 2H[X] \quad (12.1.29)$$

This generalises the the entropy of a random vector  $\mathbf{X} = (X_1, \dots, X_n)$  so that

$$H[\mathbf{X}] = H[X_1, \dots, X_n] \quad (12.1.30)$$

$$= H[X_1] + H[X_2|X_1] + H[X_3|X_2, X_1] + \dots + H[X_n|X_{n-1}, \dots, X_1] \quad (12.1.31)$$

$$= \sum_{i=1}^n H[X_i|X_{i-1}, \dots, X_1] \quad (12.1.32)$$

This also generalises to joint entropy between discrete random vectors  $\mathbf{X}$  and  $\mathbf{Y}$  so that

$$H[\mathbf{X}, \mathbf{Y}] = H[\mathbf{X}] + H[\mathbf{Y}|\mathbf{X}] \quad (12.1.33)$$

### Chain Rule of Conditional Entropy

It is valid to condition the chain rule of entropy on another random variable  $Z$ , so that we have

$$H[X, Y|Z] = H[X|Z] + H[Y|X, Z] \quad (12.1.34)$$

#### 12.1.6 Entropy of Functions

Let  $X$  be a discrete random variable, and let  $g(X)$  be a function of  $X$ . By the chain rule of entropy,

$$H[X, g(X)] = H[X] + H[g(X)|X] \quad (12.1.35)$$

$$= H[X] \quad (12.1.36)$$

since  $H[g(X)|X] = 0$  (i.e. all information about  $g(X)$  is revealed through  $X$ ). Also by the chain rule of entropy,

$$H[X, g(X)] = H[g(X)] + H[X|g(X)] \quad (12.1.37)$$

$$\geq H[g(X)] \quad (12.1.38)$$

since  $H[X|g(X)] \geq 0$ . Hence this gives the inequality for entropy of functions of random variables:

$$H[g(X)] \leq H[X] \quad (12.1.39)$$

This is intuitive because the distribution of  $g(X)$  cannot be any less informative than  $X$  itself. In the extreme case where  $g(X)$  equals a constant, then  $H[g(X)] = 0$ . This generalises to functions of random vectors so that

$$H[g(\mathbf{X})] \leq H[\mathbf{X}] \quad (12.1.40)$$

## Entropy of Sums

Suppose there are two independent discrete random variables  $X$  and  $Y$ , and their sum is  $Z = X + Y$ . We can obtain upper and lower bounds on  $H[Z]$ . Firstly, we can show

$$H[Z|X] = \sum_x \Pr(X=x) \sum_z \Pr(Z=z|X=x) \log \frac{1}{\Pr(Z=z|X=x)} \quad (12.1.41)$$

$$= \sum_x \Pr(X=x) \sum_z \Pr(Y=z-x|X=x) \log \frac{1}{\Pr(Y=z-x|X=x)} \quad (12.1.42)$$

$$= \sum_x \Pr(X=x) \sum_y \Pr(Y=y|X=x) \log \frac{1}{\Pr(Y=y|X=x)} \quad (12.1.43)$$

$$= H[Y|X] \quad (12.1.44)$$

This shows that after knowing  $X$ , the uncertainty in information about  $Z$  is due to  $Y$ . Then since  $X$  and  $Y$  are independent,  $H[Y|X] = H[Y]$ . Then by the inequality  $H[Z] \geq H[Z|X]$  that conditioning never increases entropy (shown using mutual information), we can write

$$H[Y] \leq H[Z] \quad (12.1.45)$$

and similarly

$$H[X] \leq H[Z] \quad (12.1.46)$$

Combining both cases gives  $\max\{H[X], H[Y]\} \leq H[Z]$ . The interpretation of this is the intuitive idea that adding a random variable to another never reduces the uncertainty. This can also be visualised using the convolution, which smoothes out the density and makes it less concentrated, increasing the entropy. An upper bound can be obtained by first using the chain rule of entropy to write

$$H[X, Z] = H[Z] + H[X|Z] \quad (12.1.47)$$

$$= H[X] + H[Z|X] \quad (12.1.48)$$

Hence

$$H[Z] = H[X] + H[Z|X] - H[X|Z] \quad (12.1.49)$$

$$= H[X] + H[Y] - H[X|Z] \quad (12.1.50)$$

since  $H[Z|X] = H[Y]$  as established above. Then since  $H[X|Z] \geq 0$ , then

$$H[Z] \leq H[X] + H[Y] \quad (12.1.51)$$

We can even establish conditions on  $X$  and  $Y$  for equality to hold. Since  $Z = X + Y$  is a function of  $(X, Y)$ , then  $H[Z] \leq H[X, Y]$  and by the chain rule of entropy  $H[X, Y] = H[X] + H[X|Y] = H[X] + H[Y]$  due to independence. Therefore

$$H[Z] \leq H[X, Y] = H[X] + H[Y] \quad (12.1.52)$$

and equality holds if we can show  $H[Z] = H[X, Y]$ . This requires that  $(X, Y)$  be a function of  $Z$ , which implies that  $X$  or  $Y$  be constant.

## Entropy of Scalings [47]

Note that if a discrete random variable is scaled by a non-zero factor, then its entropy does not change, because the probability masses themselves do not change. However, this property does

not extend to continuous random variables. We have for a continuous random variable  $X$  and its scaling  $Y = aX$  that

$$h(aX) = - \int_{-\infty}^{\infty} f_Y(y) \log f_Y(y) dy \quad (12.1.53)$$

$$= - \int_{-\infty}^{\infty} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \quad (12.1.54)$$

since  $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$ . Then apply the change of variables  $x = \frac{y}{|a|}$  to give

$$h(aX) = - \int_{-\infty}^{\infty} \frac{1}{|a|} f_X\left(\frac{x|a|}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{x|a|}{a}\right)\right) |a| dx \quad (12.1.55)$$

$$= - \int_{-\infty}^{\infty} f_X(x \cdot \text{sgn}(a)) \log\left(\frac{1}{|a|} f_X(x \cdot \text{sgn}(a))\right) dx \quad (12.1.56)$$

The  $\text{sgn}(a)$  term does not make a difference in the computation since the integral is taken over the real line, so

$$h(aX) = - \int_{-\infty}^{\infty} f_X(x) \log\left(\frac{1}{|a|} f_X(x)\right) dx \quad (12.1.57)$$

$$= - \int_{-\infty}^{\infty} f_X(x) (\log f_X(x) - |a|) dx \quad (12.1.58)$$

$$= - \int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx + \int_{-\infty}^{\infty} f_X(x) \log |a| dx \quad (12.1.59)$$

$$= h(X) + \log |a| \quad (12.1.60)$$

This expresses that increasing (decreasing) the variance of a random variable will increase (decrease) the entropy, provided the random variance remains in the same family of distribution.

### Entropy of Linear Transformations

The differential entropy of a continuous random vector  $\mathbf{X}$  under an invertible linear transformation  $\mathbf{Y} = A\mathbf{X}$  is found by

$$h(A\mathbf{X}) = - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}) \log f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \quad (12.1.61)$$

Using the relation for densities of linear transformations of random vectors,

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det(A)|} f_{\mathbf{X}}(A^{-1}\mathbf{y}) \quad (12.1.62)$$

Hence

$$h(A\mathbf{X}) = - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{|\det(A)|} f_{\mathbf{X}}(A^{-1}\mathbf{y}) \log\left(\frac{1}{|\det(A)|} f_{\mathbf{X}}(A^{-1}\mathbf{y})\right) d\mathbf{y} \quad (12.1.63)$$

Applying the change of variables  $\mathbf{x} = A^{-1}\mathbf{y}$  and using the Jacobian determinant as the scale factor of the volume element  $d\mathbf{y} = |\det(A)| d\mathbf{x}$ , this becomes

$$h(A\mathbf{X}) = - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log\left(\frac{1}{|\det(A)|} f_{\mathbf{X}}(\mathbf{x})\right) d\mathbf{x} \quad (12.1.64)$$

$$= - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} + \log |\det(A)| \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (12.1.65)$$

$$= h(\mathbf{X}) + \log |\det(A)| \quad (12.1.66)$$

$$(12.1.67)$$

### 12.1.7 Cross Entropy

For two probability mass functions  $p(x_i)$  and  $q(x_i)$ , where  $q(x_i)$  is the approximating distribution of  $p(x_i)$ , the cross entropy between  $p$  and  $q$  is defined as

$$H_{p,q} = -\mathbb{E}_p [\log q(X)] \quad (12.1.68)$$

$$= \sum_i p(x_i) \log \frac{1}{q(x_i)} \quad (12.1.69)$$

Similar properties of joint entropies apply to cross entropies of multivariate distributions. Suppose we have multivariate distributions  $p(x_1, \dots, x_n)$  and the approximating distribution  $q(x_1, \dots, x_n)$  (note the slight abuse in notation from before). Then the cross entropy is

$$H_{p_{1:n},q_{1:n}} = \sum_{x_1} \cdots \sum_{x_n} p(x_1, \dots, x_n) \log \frac{1}{q(x_1, \dots, x_n)} \quad (12.1.70)$$

If  $X_1, \dots, X_n$  are independent (subsequently this also implies  $q(x_1, \dots, x_n) = q(x_1) \dots q(x_n)$  if the knowledge of independence is also incorporated into the approximating distribution), then

$$H_{p_{1:n},q_{1:n}} = \sum_{x_1} \cdots \sum_{x_n} p(x_1) \dots p(x_n) \log \frac{1}{q(x_1) \dots q(x_n)} \quad (12.1.71)$$

$$= - \sum_{x_1} \cdots \sum_{x_n} p(x_1) \dots p(x_n) \log q(x_1) \dots q(x_n) \quad (12.1.72)$$

$$= - \sum_{x_1} \cdots \sum_{x_n} p(x_1) \dots p(x_n) \log q(x_1) - \cdots - \sum_{x_1} \cdots \sum_{x_n} p(x_1) \dots p(x_n) \log q(x_n) \quad (12.1.73)$$

$$= - \sum_{x_1} p(x_1) \log q(x_1) - \cdots - \sum_{x_n} p(x_n) \log q(x_n) \quad (12.1.74)$$

$$= \sum_{i=1}^n \sum_{x_i} p(x_i) \log \frac{1}{q(x_i)} \quad (12.1.75)$$

$$= \sum_{i=1}^n H_{p_i, q_i} \quad (12.1.76)$$

### 12.1.8 Entropy Rate

Let  $\mathbf{X}_n = \{X_1, X_2, \dots, X_n\}$  be a random sequence. Then the entropy rate is defined as

$$H[\mathbf{X}_n] = \lim_{n \rightarrow \infty} \frac{1}{n} H[X_1, X_2, \dots, X_n] \quad (12.1.77)$$

if the limit exists. Moreover if the sequence is i.i.d., then the entropy rate is

$$H[\mathbf{X}_n] = \lim_{n \rightarrow \infty} \frac{1}{n} H[X_1, X_2, \dots, X_n] \quad (12.1.78)$$

$$= \lim_{n \rightarrow \infty} \frac{n H[X]}{n} \quad (12.1.79)$$

$$= H[X] \quad (12.1.80)$$

### 12.1.9 Asymptotic Equipartition Property

The asymptotic equipartition property is the analogue of the weak law of large numbers for entropy. If  $X_1, \dots, X_n$  are i.i.d. with joint probability mass function  $p(X_1, \dots, X_n)$ , then

$$\frac{1}{n} \log \frac{1}{p(X_1, \dots, X_n)} \xrightarrow{P} H[X] \quad (12.1.81)$$

as  $n \rightarrow \infty$ .

*Proof.* Because of i.i.d., we can write

$$\frac{1}{n} \log \frac{1}{p(X_1, \dots, X_n)} = -\frac{1}{n} \sum_{i=1}^n \log p(X_i) \quad (12.1.82)$$

which converges in probability to  $-\mathbb{E}[\log p(X)] = H[X]$  due to the weak law of large numbers.  $\square$

The asymptotic equipartition property also holds for continuous random variables and differential entropy. If  $X_1, \dots, X_n$  are i.i.d. with joint probability density function  $f(X_1, \dots, X_n)$ , then

$$\frac{1}{n} \log \frac{1}{f(X_1, \dots, X_n)} \xrightarrow{P} H[X] \quad (12.1.83)$$

as  $n \rightarrow \infty$ .

### 12.1.10 Typicality

#### Typical Sets

### 12.1.11 Rényi Entropy

The Rényi entropy generalises the Shannon entropy. Let  $X$  be a discrete random variable on finite support with  $N$  elements, and respective probability masses  $p_1, \dots, p_N$ . Then the Rényi entropy of order  $\alpha$  with  $\alpha > 0$  is defined as

$$H_\alpha[X] = \frac{1}{1-\alpha} \ln \left( \sum_{i=1}^N p_i^\alpha \right) \quad (12.1.84)$$

The Shannon entropy can then be viewed as a limiting special case as  $\alpha \rightarrow 1$ . To show this, we write

$$\lim_{\alpha \rightarrow 1} H_\alpha[X] = \lim_{\alpha \rightarrow 1} \left( \frac{1}{1-\alpha} \ln \left( \sum_{i=1}^N p_i^\alpha \right) \right) \quad (12.1.85)$$

$$= \lim_{\alpha \rightarrow 1} \left( \frac{1}{-1} \cdot \frac{1}{\sum_{i=1}^N p_i^\alpha} \sum_{i=1}^N \frac{d}{d\alpha} p_i^\alpha \right) \quad (12.1.86)$$

using L'Hôpital's rule. Note the derivative  $\frac{d}{d\alpha} p_i^\alpha$  can be evaluated by

$$\frac{d}{d\alpha} p_i^\alpha = \frac{d}{d\alpha} e^{\alpha \ln p_i} \quad (12.1.87)$$

$$= (\ln p_i) e^{\alpha \ln p_i} \quad (12.1.88)$$

$$= p_i^\alpha \ln p_i \quad (12.1.89)$$

Hence evaluating the limit as  $\alpha \rightarrow 1$  (with  $\sum_{i=1}^N p_i$ ) yields

$$\lim_{\alpha \rightarrow 1} H_\alpha[X] = - \sum_{i=1}^N p_i \ln p_i \quad (12.1.90)$$

which is the Shannon entropy.

## 12.2 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence is a measure of *relative entropy* between two distributions, which roughly speaking gives a measure of the amount of information lost when approximating one distribution with the other distribution. For discrete probability distributions  $P(x)$  and  $Q(x)$ , the Kullback-Leibler divergence from  $Q$  to  $P$  is defined as

$$\text{KL}(P\|Q) = - \sum_i P(x_i) \log \frac{Q(x_i)}{P(x_i)} \quad (12.2.1)$$

$$= \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)} \quad (12.2.2)$$

$$= \sum_i P(x_i) [\log P(x_i) - \log Q(x_i)] \quad (12.2.3)$$

Here,  $Q$  is treated as the approximating distribution and  $P$  is the ‘true’ distribution. The KL divergence is finite only for when  $Q(x_i) = 0$  implies  $P(x_i) = 0$  for all  $i$ . If there exists an  $i$  for which  $Q(x_i) = 0$  and  $P(x_i) > 0$ , we take  $\text{KL}(P\|Q) = \infty$  [47].

The KL divergence is also easily generalised to multivariate distributions; consider the joint mass function  $p(x, y)$  and the approximating joint mass function  $q(x, y)$ .

$$\text{KL}(p\|q) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} \quad (12.2.4)$$

The continuous analogue of KL divergence for probability density functions  $p$  and  $q$  is

$$\text{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx \quad (12.2.5)$$

### 12.2.1 Gibbs’ Inequality

A property of the KL divergence between probability mass functions is that it is always non-negative, i.e.  $\text{KL}(P\|Q) \geq 0$ . This is known as Gibbs’ inequality.

*Proof.*

$$\text{KL}(P\|Q) = \sum_{P(x_i)>0} P(x_i) \log \frac{P(x_i)}{Q(x_i)} \quad (12.2.6)$$

$$= - \sum_{P(x_i)>0} P(x_i) \log \frac{Q(x_i)}{P(x_i)} \quad (12.2.7)$$

As  $-\log$  is a convex function, then using Jensen’s inequality

$$\text{KL}(P\|Q) \geq - \log \sum_{P(x_i)>0} P(x_i) \frac{Q(x_i)}{P(x_i)} = - \log \sum_{P(x_i)>0} Q(x_i) \quad (12.2.8)$$

Or

$$-\text{KL}(P\|Q) \leq \log \sum_{P(x_i)>0} Q(x_i) \quad (12.2.9)$$

Since  $\sum Q(x_i) \leq 1$ , then

$$-\text{KL}(P\|Q) \leq \log 1 = 0 \quad (12.2.10)$$

Hence

$$\text{KL}(P\|Q) \geq 0 \quad (12.2.11)$$

□

Intuitively, there is always information loss when approximating one distribution with another distribution, with no information loss ( $\text{KL}(P\|Q) = 0$ ) occurring only when the distributions of  $P$  and  $Q$  are identical. Also note that in general,  $\text{KL}(P\|Q) \neq \text{KL}(Q\|P)$ . An alternative way to write the KL divergence is as

$$\text{KL}(P\|Q) = \sum_i P(x_i) \log P(x_i) - \sum_i P(x_i) \log Q(x_i) \quad (12.2.12)$$

$$= \left[ - \sum_i P(x_i) \log Q(x_i) \right] - \left[ - \sum_i P(x_i) \log P(x_i) \right] \quad (12.2.13)$$

$$= H_{P,Q} - H_P \quad (12.2.14)$$

where  $H_{P,Q}$  is the cross entropy of  $P$  and  $Q$ , and  $H_P$  is the entropy of  $P$ . Note however that the KL divergence between probability density functions does not necessarily follow Gibbs' inequality (yet it can still be interpreted as the amount of information lost).

### Log Sum Inequality

The log sum inequality may also be used to show non-negativity of the KL divergence.

**Theorem 12.1.** *For non-negative numbers  $p_1, \dots, p_n$  and  $q_1, \dots, q_n$  such that  $\sum_{i=1}^n p_i = p$  and  $\sum_{i=1}^n q_i = q$ , we have*

$$\sum_{i=1}^n p_i \log \frac{p_i}{q_i} \geq p \log \frac{p}{q} \quad (12.2.15)$$

*Proof.* Rearranging, we may equivalently show

$$\sum_{i=1}^n p_i \log \frac{p_i}{q_i} - p \log \frac{p}{q} \geq 0 \quad (12.2.16)$$

$$\sum_{i=1}^n p_i \log \frac{p_i}{q_i} - \sum_{i=1}^n p_i \log \frac{p}{q} \geq 0 \quad (12.2.17)$$

$$\sum_{i=1}^n p_i \log \frac{qp_i}{pq_i} \geq 0 \quad (12.2.18)$$

Note that for all positive  $x$ , we have  $\log \frac{1}{x} \geq 1 - x$ . This can be shown via calculus:

$$\frac{d}{dx} \left( \log \frac{1}{x} - 1 + x \right) = -\frac{1}{x^2} \cdot \frac{1}{1/x} + 1 \quad (12.2.19)$$

$$= -\frac{1}{x} + 1 \quad (12.2.20)$$

Thus  $\log \frac{1}{x} - 1 + x$  attains a minimum at  $x = 0$ , and consequently the minimum attained is zero. Applying  $\log \frac{1}{x} \geq 1 - x$ , we get

$$\sum_{i=1}^n p_i \log \frac{qp_i}{pq_i} \geq \sum_{i=1}^n p_i \left( 1 - \frac{pq_i}{qp_i} \right) \quad (12.2.21)$$

$$= \sum_{i=1}^n p_i - \sum_{i=1}^n \frac{pq_i}{q} \quad (12.2.22)$$

$$= p - p \frac{\sum_{i=1}^n q_i}{q} \quad (12.2.23)$$

$$= 0 \quad (12.2.24)$$

□

Applying the log sum inequality with  $p = 1$  and  $q = 1$  (i.e. to discrete probability distributions) gives

$$\text{KL}(P\|Q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i} \quad (12.2.25)$$

$$\geq p \log \frac{p}{q} \quad (12.2.26)$$

$$= 0 \log 1 \quad (12.2.27)$$

$$= 0 \quad (12.2.28)$$

An analogous inequality and arguments applies for continuous distributions, e.g. for densities  $p(x)$  and  $q(x)$  over support  $\mathcal{X}$ , we have

$$\text{KL}(p(x)\|q(x)) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \quad (12.2.29)$$

$$\geq \int_{\mathcal{X}} p(x) \left( 1 - \frac{q(x)}{p(x)} \right) dx \quad (12.2.30)$$

$$= \int_{\mathcal{X}} p(x) dx - \int_{\mathcal{X}} q(x) dx \quad (12.2.31)$$

$$= 1 - 1 \quad (12.2.32)$$

$$= 0 \quad (12.2.33)$$

or more generally,

$$\int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \geq \int_{\mathcal{X}} p(x) dx \log \frac{\int_{\mathcal{X}} p(x) dx}{\int_{\mathcal{X}} q(x) dx} \quad (12.2.34)$$

## 12.2.2 Chain Rule of KL Divergence

Let  $p(x, y)$  and  $q(x, y)$  be a pair of bivariate probability mass functions. The KL divergence of  $p$  from  $q$  can be expressed as

$$\text{KL}(p(x, y)\|q(x, y)) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} \quad (12.2.35)$$

$$= \sum_x p(x) \sum_y p(y|x) \log \left( \frac{p(y|x)p(x)}{q(y|x)q(x)} \right) \quad (12.2.36)$$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \sum_y p(y|x) \log \left( \frac{p(y|x)}{q(y|x)} \right) \quad (12.2.37)$$

$$= \text{KL}(p(x)\|q(x)) + \mathbb{E}_X \left[ \mathbb{E}_Y \left[ \log \left( \frac{p(Y|X)}{q(Y|X)} \right) \middle| X \right] \right] \quad (12.2.38)$$

The second term can be viewed as like a *conditional KL divergence*, which we denote by

$$\text{KL}(p(x|y)\|q(x|y)|p(x)) = \sum_x p(x) \sum_y p(y|x) \log \left( \frac{p(y|x)}{q(y|x)} \right) \quad (12.2.39)$$

Hence the chain rule of KL divergence is

$$\text{KL}(p(x, y) \| q(x, y)) = \text{KL}(p(x) \| q(x)) + \text{KL}(p(x|y) \| q(x|y)|p(x)) \quad (12.2.40)$$

Through induction, a more general rule over multivariate distributions can be derived as

$$\text{KL}(p(x_1, \dots, x_n) \| q(x_1, \dots, x_n)) = \sum_{i=1}^n \text{KL}(p(x_i|x_{i-1}, \dots, x_1) \| q(x_i|x_{i-1}, \dots, x_1)|p(x_{i-1}, \dots, x_1)) \quad (12.2.41)$$

### 12.2.3 Mutual Information

Consider two discrete random variables  $X, Y$  with joint probability mass function  $p(x, y)$  and marginal probability mass functions  $p(x)$  and  $p(y)$  respectively. The mutual information  $\text{MI}(X; Y)$  is the KL divergence between the joint distribution and the approximating distribution  $q(x, y) = p(x)p(y)$ .

$$\text{MI}(X; Y) = \text{KL}(p(x, y) \| p(x)p(y)) \quad (12.2.42)$$

$$= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (12.2.43)$$

We can see that if  $X$  and  $Y$  are independent, i.e.  $p(x, y) = p(x)p(y)$ , then the mutual information will be zero. In that sense, we can think of mutual information as the amount of information loss by assuming independence of random variables. Also, it also reinforces the idea that independent random variables convey no information about each other. Mutual information can be rewritten as

$$\text{MI}(X; Y) = \sum_x \sum_y p(x, y) \log \frac{p(x|y)}{p(x)} \quad (12.2.44)$$

$$= \sum_x \sum_y p(x, y) \log p(x|y) - \sum_x \sum_y p(x, y) \log p(x) \quad (12.2.45)$$

$$= \sum_x \sum_y p(x, y) p(x) \log p(x|y) - \sum_x p(x) \log p(x) \quad (12.2.46)$$

$$= H[X] - H[X|Y] \quad (12.2.47)$$

By symmetry we can also show

$$\text{MI}(X; Y) = H[Y] - H[Y|X] \quad (12.2.48)$$

Hence in this way,  $\text{MI}(X; Y)$  can be thought of as the information gain (i.e. loss in entropy) about  $X$  from knowing  $Y$ , and likewise about  $Y$  from knowing  $X$ . By using the chain rule of entropy, we have

$$\text{MI}(X; Y) = H[X] + H[Y] - H[X, Y] \quad (12.2.49)$$

Also note that

$$\text{MI}(X; X) = H[X] - H[X|X] = H[X] - 0 = H[X] \quad (12.2.50)$$

Given that  $\text{MI}(X; Y) \geq 0$  by Gibb's inequality, we can also write

$$H[X] \geq H[X|Y] \quad (12.2.51)$$

$$H[Y] \geq H[Y|X] \quad (12.2.52)$$

which expresses that conditioning never increases entropy.

## Conditional Mutual Information

We can define the conditional mutual information  $\text{MI}(X; Y|Z)$  in the same conceptual way as the mutual information, except using the conditional distributions  $p(x, y|z)$ ,  $p(x|z)$  and  $p(y|z)$ :

$$\text{MI}(X; Y|Z) = \sum_x \sum_y p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \quad (12.2.53)$$

From this, the analogous properties from mutual information can be derived:

$$\text{MI}(X; Y|Z) = H[X|Z] + H[Y|Z] - H[X, Y|Z] \quad (12.2.54)$$

$$H[X|Z] \geq H[X|Y, Z] \quad (12.2.55)$$

$$H[Y|Z] \geq H[Y|X, Z] \quad (12.2.56)$$

## Chain Rule of Mutual Information

The chain rule for mutual information says that

$$\text{MI}(X_1, X_2; Y) = \text{MI}(X_1; Y) + \text{MI}(X_2; Y|X_1) \quad (12.2.57)$$

*Proof.* Write mutual information in terms of entropy as

$$\text{MI}(X_1, X_2; Y) = H[X_1, X_2] - H[X_1, X_2|Y] \quad (12.2.58)$$

Expanding both these terms using the chain rule of entropy, this gives

$$\text{MI}(X_1, X_2; Y) = H[X_1] + H[X_2|X_1] - (H[X_1|Y] + H[X_2|X_1, Y]) \quad (12.2.59)$$

Reorganising the terms, we see

$$\text{MI}(X_1, X_2; Y) = (H[X_1] - H[X_1|Y]) + (H[X_2|X_1] + H[X_2|X_1, Y]) \quad (12.2.60)$$

$$= \text{MI}(X_1; Y) + \text{MI}(X_2; Y|X_1) \quad (12.2.61)$$

□

A generalisation to more variables is possible, giving

$$\text{MI}(X_1, \dots, X_n; Y) = \text{MI}(X_1; Y) + \text{MI}(X_2; Y|X_1) + \dots + \text{MI}(X_n; Y|X_{n-1}, \dots, X_1) \quad (12.2.62)$$

$$= \sum_{i=1}^n \text{MI}(X_i; Y|X_{i-1}, \dots, X_1) \quad (12.2.63)$$

## Data Processing Inequality

The data processing inequality formalises the notion that no processing of data can further improve the information that can be gleaned from that data. Let  $X, Y, Z$  be a **Markov chain**, in that order (i.e.  $X$  and  $Z$  are conditionally independent given  $Y$ ). Then the inequality states for their mutual information:

$$\text{MI}(X; Y) \geq \text{MI}(X; Z) \quad (12.2.64)$$

*Proof.* Using the chain rule for mutual information, write the mutual information  $\text{MI}(Y, Z; X)$  as

$$\text{MI}(Y, Z; X) = \text{MI}(Y; X) + \text{MI}(Z; X|Y) \quad (12.2.65)$$

which can be alternatively expanded as

$$\text{MI}(Y, Z; X) = \text{MI}(Z; X) + \text{MI}(Y; X|Z) \quad (12.2.66)$$

Equating the two yields

$$\text{MI}(Y; X) + \text{MI}(Z; X|Y) = \text{MI}(Z; X) + \text{MI}(Y; X|Z) \quad (12.2.67)$$

Now since  $X$  and  $Z$  are conditionally independent given  $Y$ , and the mutual information of independent random variables is zero, it follows that  $\text{MI}(Z; X|Y) = 0$ . Our equality reduces to

$$\text{MI}(Y; X) = \text{MI}(Z; X) + \text{MI}(Y; X|Z) \quad (12.2.68)$$

Then use the property that mutual information is non-negative (i.e.  $\text{MI}(Y; X|Z) \geq 0$ ) to give our desired inequality (up to a symmetry):

$$\text{MI}(Y; X) \geq \text{MI}(Z; X) \quad (12.2.69)$$

□

Intuitively,  $X$  can be thought of as the variable we would like to infer, and  $Y$  is some data observed that is dependent on  $Y$ . Then  $Z$  can be an arbitrary processing of the data. Then the inequality roughly says that the information gained about  $X$  from knowing  $Z$  is no more than the information gained from knowing  $Y$ .

### Lautum Information

If the ordering of the distributions in the definition in the mutual information is reversed, this is known as the lautum information (so-called because “lautum” is the reverse spelling of “mutual”) [150].

$$\text{LI}(X; Y) = \text{KL}(p(x)p(y)\|p(x,y)) \quad (12.2.70)$$

$$= \sum_x \sum_y p(x)p(y) \log \frac{p(x)p(y)}{p(x,y)} \quad (12.2.71)$$

A heuristic characterisation of the lautum information is the amount of information lost by approximating  $p(x)p(y)$  with  $p(x,y)$ , i.e. the price paid for assuming dependence, when two things are actually independent.

#### 12.2.4 Information Processing Inequality

Let  $X$  and  $Y$  be random variables with distributions  $p$  and  $q$  respectively. Consider any function  $f(\cdot)$  so that  $W = f(X)$  and  $Z = f(Y)$ , with distributions  $\tilde{p}$  and  $\tilde{q}$  respectively. Then

$$\text{KL}(\tilde{p}\|\tilde{q}) \leq \text{KL}(p\|q) \quad (12.2.72)$$

That is, if the random variables undergo a common transformation, then there is reduced information loss between the distributions.

*Proof.* Suppose  $X$  and  $Y$  are discrete random variables on support  $\mathcal{X}$ . Denote the image of  $f(\cdot)$  by  $f(\mathcal{X})$ , and the preimage by  $f^{-1}(x)$ . Then splitting the summation over  $\mathcal{X}$  into the image and then preimage, we have

$$\text{KL}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad (12.2.73)$$

$$= \sum_{j \in f(\mathcal{X})} \sum_{x \in f^{-1}(j)} p(x) \log \frac{p(x)}{q(x)} \quad (12.2.74)$$

$$\geq \sum_{j \in f(\mathcal{X})} \Pr(W = j) \log \frac{\Pr(W = j)}{\Pr(Z = j)} \quad (12.2.75)$$

$$= \text{KL}(\tilde{p} \parallel \tilde{q}) \quad (12.2.76)$$

where we have used the log sum inequality, since

$$\sum_{x \in f^{-1}(j)} p(x) = \Pr(W = j) \quad (12.2.77)$$

$$\sum_{x \in f^{-1}(j)} q(x) = \Pr(Z = j) \quad (12.2.78)$$

The steps in the case where  $X$  and  $Y$  are continuous random variables are analogous.  $\square$

### 12.2.5 Asymptotic Equipartition Property for the KL Divergence

A version of the asymptotic equipartition property exists for the KL divergence, which can be proved in a similar fashion. Let  $X_1, \dots, X_n$  be i.i.d. with joint probability mass function  $p(X_1, \dots, X_n)$ . Let  $q(X_1, \dots, X_n)$  be any other probability mass function on the support of  $X$  (which plays the part of the approximating distribution). Then

$$\frac{1}{n} \log \frac{p(X_1, \dots, X_n)}{q(X_1, \dots, X_n)} \xrightarrow{p} \text{KL}(p \parallel q) \quad (12.2.79)$$

as  $n \rightarrow \infty$ .

### 12.2.6 Equivalence Between Minimum KL Divergence and MLE

We show that finding a parameter which minimises the KL divergence between the parametrised distribution and a empirical distribution is equivalent to finding the maximum likelihood estimate. This is shown for the case where the family of distributions is continuous, by the result analogously holds for families of discrete distributions. Suppose some i.i.d. data  $x_1, \dots, x_n$  is collected, resulting in an empirical distribution function with density function  $\tilde{p}(x)$ .

$$\tilde{p}(x) = \sum_{i=1}^n \frac{1}{n} \delta(x - x_i) \quad (12.2.80)$$

where  $\delta(\cdot)$  is the Dirac distribution (i.e. the empirical density is just a continuous representation using impulses of the empirical mass function). Suppose there is a family of models with probability density  $q(x; \theta)$  on support  $\mathcal{X}$  (which is implicitly assumed to be a superset of all the data), parametrised by  $\theta$ . We seek to minimise the KL divergence between the empirical density function  $\tilde{p}(x)$  and the approximating distribution  $q(x; \theta)$ . By definition of the KL divergence,

$$\text{KL}(\tilde{p}(x) \parallel q(x; \theta)) = \mathbb{E} \left[ \log \frac{\tilde{p}(x)}{q(x; \theta)} \right] \quad (12.2.81)$$

where it is understood that the expectation is taken over the empirical density. Rewriting the KL divergence in terms of cross entropy and entropy,

$$\text{KL}(\tilde{p}(x) \parallel q(x; \theta)) = -\mathbb{E}[\log q(x; \theta)] + \mathbb{E}[\log \tilde{p}(x)] \quad (12.2.82)$$

Note that the entropy of  $\tilde{p}(x)$  is unaffected by  $\theta$ . Evaluating the (negative) cross-entropy term  $\mathbb{E}[\log q(x; \theta)]$ :

$$\mathbb{E}[\log q(x; \theta)] = \int_{\mathcal{X}} \tilde{p}(x) \log q(x; \theta) dx \quad (12.2.83)$$

$$= \int_{\mathcal{X}} \sum_{i=1}^n \frac{1}{n} \delta(x - x_i) \log q(x; \theta) dx \quad (12.2.84)$$

$$= \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{X}} \delta(x - x_i) \log q(x; \theta) dx \quad (12.2.85)$$

$$= \frac{1}{n} \sum_{i=1}^n \log q(x_i; \theta) \quad (12.2.86)$$

Hence

$$\operatorname{argmin}_{\theta} \text{KL}(\tilde{p}(x) \| q(x; \theta)) = \operatorname{argmin}_{\theta} \left\{ - \sum_{i=1}^n \log q(x_i; \theta) \right\} \quad (12.2.87)$$

which is the minimiser of the negative log likelihood when the data is i.i.d.

### 12.2.7 Symmetrised KL Divergence

The symmetrised KL divergence is defined as

$$\text{KL}_{\text{sym}}(P \| Q) = \text{KL}(P \| Q) + \text{KL}(Q \| P) \quad (12.2.88)$$

where it attains the property that  $\text{KL}_{\text{sym}}(P \| Q) = \text{KL}_{\text{sym}}(Q \| P)$ .

### 12.2.8 Method of Types

Consider a distribution  $q$  on finite support  $\mathcal{X}$ , and let  $(X_1, \dots, X_n)$  be an i.i.d. sequence from  $q$ . Use  $\mathbf{x}$  to denote a realisation of the sequence. Denote  $P_{\mathbf{x}}$  as the *type* of the sequence  $\mathbf{x}$ , which is essentially the empirical distribution. That is, for each  $x \in \mathcal{X}$ , we have

$$P_{\mathbf{x}}(x) = \frac{n_{\mathbf{x}}(x)}{n} \quad (12.2.89)$$

where  $n_{\mathbf{x}}(x)$  is the number of times the symbol  $x$  appears in  $\mathbf{x}$ . Let  $\mathcal{P}_n$  denote the set of all realisable empirical distributions (i.e. types) which may result of drawing i.i.d. sequences  $\mathbf{x}$  from  $q$ . We have a simple upper bound on the size of  $\mathcal{P}_n$  as

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|} \quad (12.2.90)$$

since each element in  $P_{\mathbf{x}}$  always has the same denominator of  $n$ , while the numerator can take on  $n+1$  possible values from the set  $\{0, \dots, n\}$ . Then there are  $|\mathcal{X}|$  different elements in  $P_{\mathbf{x}}$ . For a given  $P \in \mathcal{P}_n$ , we define the *type class* of  $P$  as

$$T(P) = \{\mathbf{x} \in \mathcal{X}^n : P_{\mathbf{x}} = P\} \quad (12.2.91)$$

which is the set of sequences  $\mathbf{x}$  which lead to a type of  $P$ . The multinomial coefficient immediately gives the size of the type class as

$$|T(P_{\mathbf{x}})| = \binom{n}{n_{\mathbf{x}}(\alpha_1), \dots, n_{\mathbf{x}}(\alpha_{|\mathcal{X}|})} \quad (12.2.92)$$

for symbols  $\alpha_1, \dots, \alpha_{|\mathcal{X}|} \in \mathcal{X}$ . Looser but more manipulable bounds for this can be obtained. We first claim the following lemma for the distribution over sequences.

**Lemma 12.1.** *Let  $\mathbf{x}$  be the realisation of an i.i.d. sequence from the distribution  $q(x)$ . The joint distribution denoted  $q^n(\mathbf{x}) := \prod_{i=1}^n q(x_i)$  is given by:*

$$q^n(\mathbf{x}) = \exp[-n(H[P_{\mathbf{x}}] + \text{KL}(P_{\mathbf{x}} \| q))] \quad (12.2.93)$$

where the logarithm used is the natural logarithm, or

$$q^n(\mathbf{x}) = 2^{-n(H[P_{\mathbf{x}}] + \text{KL}(P_{\mathbf{x}}||q))} \quad (12.2.94)$$

where the logarithm used is base 2.

*Proof.* To show this, we begin from

$$q^n(\mathbf{x}) = \prod_{i=1}^n q(x_i) \quad (12.2.95)$$

$$= \prod_{\alpha \in \mathcal{X}} q(\alpha)^{n_{\mathbf{x}}(\alpha)} \quad (12.2.96)$$

$$= \prod_{\alpha \in \mathcal{X}} q(\alpha)^{nP_{\mathbf{x}}(\alpha)} \quad (12.2.97)$$

$$= \prod_{\alpha \in \mathcal{X}} \exp \left( \ln q(\alpha)^{nP_{\mathbf{x}}(\alpha)} \right) \quad (12.2.98)$$

$$= \prod_{\alpha \in \mathcal{X}} \exp(nP_{\mathbf{x}}(\alpha) \ln q(\alpha)) \quad (12.2.99)$$

Adding and subtracting  $P_{\mathbf{x}}(\alpha) \ln P_{\mathbf{x}}(\alpha)$ :

$$q^n(\mathbf{x}) = \prod_{\alpha \in \mathcal{X}} \exp[n(P_{\mathbf{x}}(\alpha) \ln q(\alpha) - P_{\mathbf{x}}(\alpha) \ln P_{\mathbf{x}}(\alpha) + P_{\mathbf{x}}(\alpha) \ln P_{\mathbf{x}}(\alpha))] \quad (12.2.100)$$

$$= \exp \left[ \sum_{\alpha \in \mathcal{X}} n(P_{\mathbf{x}}(\alpha) \ln q(\alpha) - P_{\mathbf{x}}(\alpha) \ln P_{\mathbf{x}}(\alpha) + P_{\mathbf{x}}(\alpha) \ln P_{\mathbf{x}}(\alpha)) \right] \quad (12.2.101)$$

$$= \exp \left[ \sum_{\alpha \in \mathcal{X}} n \left( -P_{\mathbf{x}}(\alpha) \ln \frac{P_{\mathbf{x}}(\alpha)}{q(\alpha)} + P_{\mathbf{x}}(\alpha) \ln P_{\mathbf{x}}(\alpha) \right) \right] \quad (12.2.102)$$

Then using the definition of the KL divergence, we have

$$q^n(\mathbf{x}) = \exp[n(-\text{KL}(P_{\mathbf{x}}||q) - H[P_{\mathbf{x}}])] \quad (12.2.103)$$

$$= \exp[-n(H[P_{\mathbf{x}}] + \text{KL}(P_{\mathbf{x}}||q))] \quad (12.2.104)$$

as claimed. In the same way, we can show the analogous holds when the log is taken in base 2.  $\square$

This shows that the probability of an i.i.d. sequence only depends on the type (intuitively, permuting the sequence does not affect the probability).

**Theorem 12.2.** *The size of the type class is upper bounded by:*

$$|T(P_{\mathbf{x}})| \leq 2^{nH[P_{\mathbf{x}}]} \quad (12.2.105)$$

*Proof.* Since the distribution  $q(x)$  is arbitrary above, setting it to be the same as the empirical distribution  $P_{\mathbf{x}}$  yields  $q^n(\mathbf{x}) = 2^{-nH[P_{\mathbf{x}}]}$  as long as  $\mathbf{x} \in T(P_{\mathbf{x}})$ . Applying this fact gives

$$1 \geq \sum_{\mathbf{x} \in T(P_{\mathbf{x}})} q^n(\mathbf{x}) \quad (12.2.106)$$

$$= \sum_{\mathbf{x} \in T(P_{\mathbf{x}})} 2^{-nH[P_{\mathbf{x}}]} \quad (12.2.107)$$

$$= |T(P_{\mathbf{x}})| 2^{-nH[P_{\mathbf{x}}]} \quad (12.2.108)$$

Therefore

$$|T(P_{\mathbf{x}})| \leq 2^{nH[P_{\mathbf{x}}]} \quad (12.2.109)$$

$\square$

**Theorem 12.3.** *The size of the type class is lower bounded by:*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n H[P_{\mathbf{x}}]} \leq |T(P_{\mathbf{x}})| \quad (12.2.110)$$

*Proof.* For notational simplicity, we suppress the dependence on  $\mathbf{x}$  and work with  $P$ . Let  $P^n$  denote the joint distribution of an i.i.d. sequence sampled from the empirical distribution  $P$  itself. We can write

$$\sum_{P' \in \mathcal{P}_n} \Pr_{P^n} (\mathbf{X} \in T(P')) = 1 \quad (12.2.111)$$

because the support of  $P^n$  is a subset of  $\mathcal{X}^n$ , and we have

$$\bigcup_{P' \in \mathcal{P}_n} T(P') = \mathcal{X}^n \quad (12.2.112)$$

where the  $T(P')$  are disjoint, thus we follow Kolmogorov's axioms. Then take the bound

$$1 = \sum_{P' \in \mathcal{P}_n} \Pr_{P^n} (\mathbf{X} \in T(P')) \quad (12.2.113)$$

$$\leq |\mathcal{P}_n| \max_{P' \in \mathcal{P}_n} \Pr_{P^n} (\mathbf{X} \in T(P')) \quad (12.2.114)$$

$$\leq (n+1)^{|\mathcal{X}|} \max_{P' \in \mathcal{P}_n} \Pr_{P^n} (\mathbf{X} \in T(P')) \quad (12.2.115)$$

using the upper bound on the size of  $\mathcal{P}_n$ . Next, we claim that

$$\max_{P' \in \mathcal{P}_n} \Pr_{P^n} (\mathbf{X} \in T(P')) = \Pr_{P^n} (\mathbf{X} \in T(P)) \quad (12.2.116)$$

which says that the most likely sequence to be generated by resampling from  $P$  is a sequence of type  $P$  itself. This is intuitive and not very surprising, however it can be formally shown by (for instance) considering the maximum likelihood estimator for the multinomial distribution, whereby the estimates of the categorical probabilities correspond to the empirical proportions themselves. Thus

$$1 \leq (n+1)^{|\mathcal{X}|} \Pr_{P^n} (\mathbf{X} \in T(P)) \quad (12.2.117)$$

$$= (n+1)^{|\mathcal{X}|} \sum_{\mathbf{x} \in T(P)} P^n(\mathbf{x}) \quad (12.2.118)$$

$$= (n+1)^{|\mathcal{X}|} \sum_{\mathbf{x} \in T(P)} 2^{-n H[P]} \quad (12.2.119)$$

$$= (n+1)^{|\mathcal{X}|} |T(P)| 2^{-n H[P]} \quad (12.2.120)$$

where  $P^n(\mathbf{x}) = 2^{-n H[P]}$  in the same way as proving the upper bound. Lastly rearranging, this yields the required bound

$$|T(P)| \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n H[P]} \quad (12.2.121)$$

□

**Corollary 12.1.** *The probability of a type class  $T(P)$  with respect to the distribution  $q$  is bounded by*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n \text{KL}(P||q)} \leq \Pr_{q^n} (\mathbf{X} \in T(P)) \leq 2^{-n \text{KL}(P||q)} \quad (12.2.122)$$

*Proof.* For the upper bound, since we already know that the probability of any sequence only depends on its type class, we have

$$\Pr_{q^n}(\mathbf{X} \in T(P)) = 2^{-n(H[P] + KL(P||q))} \quad (12.2.123)$$

$$= 2^{-nH[P]} \cdot 2^{-nKL(P||q)} \quad (12.2.124)$$

$$\leq 2^{-nKL(P||q)} \quad (12.2.125)$$

as  $-nH[P] \leq 0$  by non-negativity of Shannon entropy, so  $2^{-nH[P]} \leq 1$ . For the lower bound, again applying the fact that the probability only depends on the type class,

$$\Pr_{q^n}(\mathbf{X} \in T(P)) = \sum_{\mathbf{x} \in T(P)} q^n(\mathbf{x}) \quad (12.2.126)$$

$$= |T(P)| 2^{-n(H[P] + KL(P||q))} \quad (12.2.127)$$

$$\geq \frac{2^{-n(H[P] + KL(P||q))}}{(n+1)^{|\mathcal{X}|}} 2^{nH[P]} \quad (12.2.128)$$

$$= \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nKL(P||q)} \quad (12.2.129)$$

where we used the lower bound for the size of the type class.  $\square$

### Sanov's Theorem

Sanov's theorem provides a large deviations bound in terms of the KL divergence. Consider the distribution  $q(x)$  on finite support  $\mathcal{X}$ , and an i.i.d. sample  $\mathbf{X}$  of size  $n$  from the joint distribution  $q^n(\mathbf{x})$ . Let  $E$  be a set of probability distributions over the same support, which we can think of as a ‘rare’ event for the empirical distribution of  $\mathcal{X}$ . From the method of types, and with minor abuse of notation using  $q^n(\cdot)$ , we express the probability of this rare event as

$$q^n(E) = \sum_{P_x \in E} \Pr_{q^n}(\mathbf{X} \in T(P_x)) \quad (12.2.130)$$

**Theorem 12.4.** *The probability  $q^n(E)$  is upper bounded by*

$$q^n(E) \leq (n+1)^{|\mathcal{X}|} 2^{-nKL(P^*||q)} \quad (12.2.131)$$

where

$$P^* = \operatorname{argmin}_{P \in E} KL(P||q) \quad (12.2.132)$$

*Proof.* Let  $\mathcal{P}_n$  be the set of all realisable types from a sample of size  $n$  so that

$$q^n(E) = \sum_{P_x \in E \cap \mathcal{P}_n} \Pr_{q^n}(\mathbf{X} \in T(P_x)) \quad (12.2.133)$$

$$\leq \sum_{P_x \in E \cap \mathcal{P}_n} 2^{-nKL(P||q)} \quad (12.2.134)$$

$$= |E \cap \mathcal{P}_n| 2^{-nKL(P||q)} \quad (12.2.135)$$

$$\leq |\mathcal{P}_n| 2^{-nKL(P||q)} \quad (12.2.136)$$

$$\leq (n+1)^{|\mathcal{X}|} 2^{-nKL(P||q)} \quad (12.2.137)$$

where the first inequality comes from the upper bound on the probability  $\Pr_{q^n}(\mathbf{X} \in T(P_x))$ , and the last inequality comes from the upper bound on  $\mathcal{P}_n$ . Taking the ‘worst case’ over  $E \cap \mathcal{P}_n$ , we have

$$q^n(E) \leq (n+1)^{|\mathcal{X}|} \max_{P \in E \cap \mathcal{P}_n} 2^{-nKL(P||q)} \quad (12.2.138)$$

$$\leq (n+1)^{|\mathcal{X}|} \max_{P \in E} 2^{-n \text{KL}(P||q)} \quad (12.2.139)$$

$$= (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in E} \text{KL}(P||q)} \quad (12.2.140)$$

$$= (n+1)^{|\mathcal{X}|} 2^{-n \text{KL}(P^*||q)} \quad (12.2.141)$$

□

Note that  $(n+1)^{|\mathcal{X}|}$  in the numerator is polynomial with  $n$  while  $2^{n \text{KL}(P^*||q)}$  in the denominator is exponential with  $n$ . So intuitively, this result says the probability of a rare event gets smaller as the sample size increases. The rate at which it gets small is controlled by the size  $|\mathcal{X}|$ , and the ‘rarity’ of  $E$ , represented by  $P^*$  (which is distribution in  $E$  that is the ‘closest’ from  $q$ ).

We can also characterise the probability in the limit if  $E$  is a suitable ‘nice’ set. More precisely, consider the case where  $E$  is closure of its interior. Roughly speaking, the set  $E$  has at least some ‘filling’. So for example, a point set  $E$  would be excluded, but  $E$  as a non-empty region of the probability simplex would be acceptable.

**Corollary 12.2.** *If  $E$  is the closure of its interior, then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log q^n(E) = -\text{KL}(P^*||q) \quad (12.2.142)$$

*Proof.* For arbitrary  $n$ , note that  $P^*$  need not be a member of  $\mathcal{P}_n$ . But if  $E$  is the closure of its interior, then for sufficiently large  $n \geq \bar{n}$ , we can find a distribution in  $E \cap \mathcal{P}_n$  that is ‘close’ to  $P^*$  (as  $\mathcal{P}_n$  becomes more and more ‘refined’ over the probability simplex). This means we are able to construct a sequence of distributions  $P[n]$  such that  $P[n] \in E \cap \mathcal{P}_n$  for all  $n \geq \bar{n}$  and  $\lim_{n \rightarrow \infty} \text{KL}(P[n]||q) = \text{KL}(P^*||q)$ . Using the lower bound the probability of a type class, then for every  $n \geq \bar{n}$  we have

$$q^n(E) = \sum_{P_x \in E} \Pr_{q^n}(\mathbf{X} \in T(P_x)) \quad (12.2.143)$$

$$\geq \Pr_{q^n}(\mathbf{X} \in T(P[n])) \quad (12.2.144)$$

$$\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n \text{KL}(P[n]||q)} \quad (12.2.145)$$

Hence

$$\frac{1}{n} \log_2 q^n(E) \geq -\frac{|\mathcal{X}| \log_2(n+1)}{n} - \text{KL}(P[n]||q) \quad (12.2.146)$$

and a lower bound on the limit of  $\frac{1}{n} q^n(E)$  is

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log q^n(E) \geq \liminf_{n \rightarrow \infty} \left( -\frac{|\mathcal{X}| \log_2(n+1)}{n} - \text{KL}(P[n]||q) \right) \quad (12.2.147)$$

$$= -\text{KL}(P^*||q) \quad (12.2.148)$$

In the same way for the upper bound,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log q^n(E) \leq \limsup_{n \rightarrow \infty} \left( \frac{|\mathcal{X}| \log(n+1)}{n} - \text{KL}(P^*||q) \right) \quad (12.2.149)$$

$$= -\text{KL}(P^*||q) \quad (12.2.150)$$

which establishes the limit, because the bounds for the limit superior and limit inferior are equal. □

This limit indicates that the asymptotic rate of the probability for the rare event becoming small only depends on  $\text{KL}(P^*||q)$ , i.e. for large  $n$

$$q^n(E) \approx 2^{-n\text{KL}(P^*||q)} \quad (12.2.151)$$

Sanov's theorem can also be applied to continuous distributions, by appropriately quantising the support using a similar concept as with the chi-squared goodness-of-fit test.

## 12.3 Maximum Entropy Distributions [75]

### 12.3.1 Principle of Maximum Entropy

The principle of maximum entropy roughly states that, given limited information about a prior, then the best choice of prior distribution (out of all possible priors which satisfy the limited information given) is the distribution with maximum entropy. Intuitively, this distribution makes the least number of assumptions about the variable in question.

### 12.3.2 Maximum Entropy Distributions on Finite Support

#### Discrete Uniform Maximum Entropy Distribution

Consider all probability mass functions supported on a finite set  $\mathcal{X}$ , with cardinality  $|\mathcal{X}| = N$ . If there are no other constraints (apart from being a valid probability mass function), it can be shown that the discrete uniform distribution (i.e. with probability mass function  $Q(x) = \frac{1}{N}$ ) is the distribution which maximises entropy over all other possible distributions. To show this, first calculate the entropy of the uniform distribution to be

$$H_Q = - \sum_{x \in \mathcal{X}} \frac{1}{N} \log \frac{1}{N} \quad (12.3.1)$$

$$= \log N \sum_{x \in \mathcal{X}} \frac{1}{N} \quad (12.3.2)$$

$$= \log |\mathcal{X}| \quad (12.3.3)$$

Now consider an arbitrary distribution  $P(x)$ , also supported on  $\mathcal{X}$ . The Kullback-Leibler divergence of  $P$  from  $Q$  is

$$\text{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{1/N} \quad (12.3.4)$$

$$= \sum_{x \in \mathcal{X}} P(x) \log P(x) - \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{N} \quad (12.3.5)$$

$$= -H_P + H_Q \quad (12.3.6)$$

By Gibb's inequality,  $\text{KL}(P||Q) \geq 0$  hence

$$H_P \leq H_Q \quad (12.3.7)$$

which shows that any arbitrary distribution on  $\mathcal{X}$  cannot have greater entropy than the discrete uniform distribution. This also proves that the entropy of any random variable  $X$  supported on  $\mathcal{X}$  is upper bounded by

$$H[X] \leq \log |\mathcal{X}| \quad (12.3.8)$$

An alternative derivation would involve formulating and solving the constrained optimisation problem

$$\begin{aligned} \min_{p_1, \dots, p_N} \quad & -\sum_{i=1}^N p_i \log p_i \\ \text{s.t.} \quad & p_1 + \dots + p_N = 1 \\ & p_i \geq 0, \quad i = 1, \dots, n \end{aligned} \tag{12.3.9}$$

using the method of Lagrange multipliers.

This result agrees with the intuition that the uniform distributions contain the most amount of ‘surprise’ (i.e. everything is equally likely). This characterisation can also be used to demonstrate that more entropy does not necessarily equate to more variance. Suppose a distribution were modified from the uniform distribution by shifting some of the mass more towards the bounds of the support. Then the variance would clearly increase, but we know the resulting distribution will have less entropy than the uniform distribution.

### 12.3.3 Maximum Entropy Distributions on Bounded Support

For any continuous distribution on support  $[a, b]$ , the continuous uniform distribution has the maximum entropy.

### 12.3.4 Maximum Entropy Distributions on Unbounded Support

For a prescribed mean, the exponential distribution has the maximum entropy among all continuous distributions supported on  $[0, \infty)$ .

For a prescribed mean and variance, the Gaussian distribution has the maximum entropy among all continuous distributions supported on  $(-\infty, \infty)$ .

For a prescribed mean, variance, skewness and kurtosis, the maximum entropy distribution among continuous distributions supported on  $(-\infty, \infty)$  takes the form

$$f_X(x) \propto \exp(ax + bx^2 + cx^3 + dx^4) \tag{12.3.10}$$

However there may be no solution (if the skewness and kurtosis lie in certain regions) and the solution (if it exists) can be a bimodal distribution [167].

### 12.3.5 Maximum Entropy of Exponential Families [75, 143]

## 12.4 Coding Theory [47]

### 12.4.1 Source Coding

Suppose there is a random variable  $X$  with discrete support  $\mathcal{X}$ . A source code for  $X$  is a mapping from  $\mathcal{X}$  to a set  $\mathcal{D}$  of finite length strings using symbols from an alphabet of length  $D$  (called a  $D$ -ary alphabet). Each of these strings is called a *codeword*. A binary alphabet with symbols 0 and 1 is a case of  $D = 2$ .

#### Prefix Codes

A (source) code is said to be *nonsingular* if every element in  $\mathcal{X}$  maps uniquely to  $\mathcal{D}$ . The *extension* of a code is the mapping from sequences of elements from  $\mathcal{X}$  to concatenated strings from

$\mathcal{D}$  corresponding to the sequence. A code is *uniquely decodable* if its extension is non-singular (that is, it is enough to determine the sequence of events just by looking at the concatenated string). An example of uniquely decodable codes are prefix codes, whereby no codeword is a prefix of any other codeword. Prefix codes can be decoded instantaneously, i.e. without needing to look at the future sequence of codewords.

### Kraft Inequality

The Kraft inequality, (also known as the Kraft-McMillan inequality) gives the necessary and sufficient conditions for the existence of a prefix code. Consider a source code and let  $\ell_i$  denote the length of the codeword for source symbol  $x_i$ . The Kraft inequality itself is

$$\sum_i D^{-\ell_i} \leq 1 \quad (12.4.1)$$

The assertions made are:

- Any prefix code over an alphabet of size  $D$  must satisfy the Kraft inequality (necessary condition).
- Conversely, given some codeword lengths satisfying the Kraft inequality, there exists a prefix code with those codeword lengths.

To show the necessary condition, consider a  $D$ -ary tree (i.e. each node has  $D$  descendants); the special case of  $D = 2$  is a binary tree. We can imagine any codeword as a node within the tree, and spelling out the codeword amounts to taking a path along the tree to the node. We make the following two observations:

1. For an arbitrary codeword of length  $\ell$ , the number of descendants in the tree of length  $\bar{\ell}$  will be  $D^{\bar{\ell}-\ell}$ .
2. Fix  $\bar{\ell}$  to be the length of the longest codeword. The sets of descendants at depth  $\bar{\ell}$  for each codeword will be disjoint. This is due to a contradiction argument, i.e. if they were not disjoint, it would imply that one codeword is a descendant of another, and the code would no longer be prefix.

There are also cannot be more than  $D^{\bar{\ell}}$  codewords of length  $\bar{\ell}$ , thus summing over all the codewords gives

$$\sum_i D^{\bar{\ell}-\ell_i} \leq D^{\bar{\ell}} \quad (12.4.2)$$

Dividing out by  $D^{\bar{\ell}}$ , this yields the Kraft inequality  $\sum_i D^{-\ell_i} \leq 1$ .

To show the converse, consider the following constructive description of assigning codewords. Let  $\bar{\ell}$  denote the length of the longest codeword. Given codeword of length  $\ell_i$ , assign it to an available node at depth  $\ell_i$ , and ‘delete’ (i.e. make unavailable to be assigned) all the descendants of this node, otherwise the prefix condition may be violated. We now count the total number of leaf nodes (i.e. nodes at depth  $\bar{\ell}$ , with no descendants) made unavailable this way; the number is  $\sum_i D^{\bar{\ell}-\ell_i}$ . There are also a total of  $D^{\bar{\ell}}$  leaf nodes that can be deleted. So be assured our scheme will work, we require

$$\sum_i D^{\bar{\ell}-\ell_i} \leq D^{\bar{\ell}} \quad (12.4.3)$$

But after rearranging, this is the Kraft inequality  $\sum_i D^{-\ell_i} \leq 1$ , which is guaranteed to hold by hypothesis.

Note that having smaller codeword lengths  $\ell_i$  brings the sum  $\sum_i D^{-\ell_i}$  closer to one. So in a sense, satisfying the Kraft inequality with equality hints at the code being more ‘efficient’, whereas there will be some redundancy in a code which satisfies the Kraft inequality with strict inequality.

### **D-adic Distributions**

A discrete probability distribution is said to be  $D$ -adic if each probability in the mass function satisfies

$$\Pr(X = x) = \frac{1}{D^n} \quad (12.4.4)$$

for some natural number  $n$  (which can be different depending on  $x$ ). In the case  $D = 2$ , this is called a *dyadic* distribution, e.g. a distribution with masses  $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$ .

### **Optimal Prefix Codes**

Consider the design of a prefix code which minimises expected codeword length

$$L = \sum_{i=1}^m p_i \ell_i \quad (12.4.5)$$

where  $p_i$  is the probability  $x_i \in \mathcal{X}$ , with  $|\mathcal{X}| = m$ . We can formulate this as an optimisation problem

$$\begin{aligned} & \min_{\ell_1, \dots, \ell_m} && \sum_{i=1}^m p_i \ell_i \\ & \text{s.t.} && \sum_{i=1}^m D^{-\ell_i} \leq 1 \\ & && \ell_i \in \mathbb{N}, \quad \ell = 1, \dots, m \end{aligned} \quad (12.4.6)$$

where the Kraft inequality is a constraint in order to satisfy the prefix code requirement. If we instead ignore the integer constraints (only requiring the lengths to be positive) and assume that the Kraft inequality is satisfied with equality at the solution (corresponding to the intuition that smaller  $\ell_i$  makes  $\sum_{i=1}^m D^{-\ell_i}$  larger), we have the relaxed problem

$$\begin{aligned} & \min_{\ell_1, \dots, \ell_m} && \sum_{i=1}^m p_i \ell_i \\ & \text{s.t.} && \sum_{i=1}^m D^{-\ell_i} = 1 \\ & && \ell_i > 0, \quad \ell = 1, \dots, m \end{aligned} \quad (12.4.7)$$

Introduce  $d_i := D^{-\ell_i}$  so  $\ell_i = -\log_D d_i$ ,  $d_i \in (0, 1)$  and attempting to solve the problem in the  $d_i$ , we notice that is now resembles the **maximum entropy problem**, which we know to have the solution

$$d_i^* = p_i \quad (12.4.8)$$

for each  $i$ , i.e. choose  $d_i$  as the same distribution as  $p_i$ , making the KL divergence zero (attaining the lower bound in **Gibb's inequality**). Thus

$$\ell_i^* = -\log_D p_i \quad (12.4.9)$$

and evaluating the optimum yields

$$L^* = \sum_{i=1}^m p_i \ell_i^* \quad (12.4.10)$$

$$= - \sum_{i=1}^m p_i \log_D p_i \quad (12.4.11)$$

$$= H[X] \quad (12.4.12)$$

which is the entropy of random variable  $X$  from the distribution of the  $p_i$ . We can show this quantity lower bounds the actual expected codeword length  $L$  from a ‘valid’ prefix code with integer lengths, i.e.  $L \geq L^* = H[X]$  for any valid  $\ell_1, \dots, \ell_m$ . This is done as follows:

$$L - L^* = \sum_{i=1}^m p_i \ell_i + \sum_{i=1}^m p_i \log_D p_i \quad (12.4.13)$$

$$= - \sum_{i=1}^m p_i \log_D D^{-\ell_i} + \sum_{i=1}^m p_i \log_D p_i \quad (12.4.14)$$

Let  $c := \sum_{i=1}^m D^{-\ell_i} \leq 1$ , then

$$L - L^* = \sum_{i=1}^m p_i \log_D p_i - \sum_{i=1}^m p_i \log_D D^{-\ell_i} + \sum_{i=1}^m p_i \log_D c - \log_D c \quad (12.4.15)$$

$$= \sum_{i=1}^m p_i \log_D \left( \frac{p_i c}{D^{-\ell_i}} \right) + \log_D \frac{1}{c} \quad (12.4.16)$$

Now let  $r_i := \frac{D^{-\ell_i}}{\sum_{i=1}^m D^{-\ell_i}} = \frac{D^{-\ell_i}}{c}$  be the masses of a distribution, so

$$L - L^* = \sum_{i=1}^m p_i \log_D \left( \frac{p_i}{r_i} \right) + \log_D \frac{1}{c} \quad (12.4.17)$$

$$= \text{KL}(p||r) + \log_D \frac{1}{c} \quad (12.4.18)$$

We know  $\text{KL}(p||r) \geq 0$  by Gibb’s inequality. Also, the Kraft inequality implies  $\frac{1}{c} \geq 1$  so  $\log_D \frac{1}{c} \geq 0$ , therefore

$$L - L^* \geq 0 \quad (12.4.19)$$

Notice in fact that  $L = L^*$  if and only if the optimal code lengths  $\ell_i^* = -\log_D p_i$  are all natural numbers. This means  $\log_D \frac{1}{p_i} \in \mathbb{N}$  for each  $i$ , which is only possible if each  $p_i = \frac{1}{D^n}$  for some  $n \in \mathbb{N}$  (i.e. we require the distribution  $p_i$  to be  $D$ -adic). This property also gives an instructive way to find a code with optimal integer codeword lengths. We look at

$$L - L^* = \text{KL}(p||r) + \log_D \frac{1}{c} \quad (12.4.20)$$

and seek to minimise  $\text{KL}(p||r)$ , noting that the term  $\log_D \frac{1}{c}$  is minimised ‘automatically’, as shorter codeword lengths bring the Kraft inequality closer to equality. Therefore we should find the  $D$ -adic distribution  $r_i$  which minimises the KL divergence  $\text{KL}(p||r)$ , and then assign codeword lengths according to

$$\ell_i = \log_D \frac{1}{r_i} \quad (12.4.21)$$

## Shannon Coding

Shannon coding is an intuitive coding scheme for prefix codes which assigns codeword lengths by

$$\ell_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil \quad (12.4.22)$$

These lengths satisfy the Kraft inequality, since

$$\sum_i D^{-\ell_i} = \sum_i D^{-\lceil \log_D (1/p_i) \rceil} \quad (12.4.23)$$

$$\leq \sum_i D^{-\log_D (1/p_i)} \quad (12.4.24)$$

$$= \sum_i p_i = 1 \quad (12.4.25)$$

which guarantees the existence of a prefix code satisfying these lengths. We have already established a lower bound on the expected codeword length as  $H[X] \leq L = \sum_i p_i \ell_i$ . An upper bound for  $L$  can be shown to be

$$L = \sum_i p_i \ell_i \quad (12.4.26)$$

$$= \sum_i p_i \left\lceil \log_D \frac{1}{p_i} \right\rceil \quad (12.4.27)$$

$$< \sum_i p_i \left( \log_D \frac{1}{p_i} + 1 \right) \quad (12.4.28)$$

$$= \sum_i p_i \log_D \frac{1}{p_i} + \sum_i p_i \quad (12.4.29)$$

$$= H[X] + 1 \quad (12.4.30)$$

Hence

$$H[X] \leq L < H[X] + 1 \quad (12.4.31)$$

In the binary  $D = 2$  case, we can interpret this by saying that the expected codeword length by Shannon coding will differ by less 1 bit worse on average, compared to the theoretical lower bound.

## Asymptotic Optimality of Shannon Coding

Suppose we are able to group random variables together before coding it. That is, we consider a source code for the random vector  $\mathbf{X}_n = (X_1, \dots, X_n)$  which will be supported on  $\mathcal{X}^n = \mathcal{X} \times \dots \times \mathcal{X}$ , but it just another discrete random variable. Applying the entropy bounds for Shannon coding, we have

$$H[\mathbf{X}_n] \leq L_n < H[\mathbf{X}_n] + 1 \quad (12.4.32)$$

where  $L_n$  is the expected codeword length for block length  $n$ . As a direct way to compare against coding one input  $X$  at a time, consider  $\frac{1}{n} L_n$ , which is the average codeword length per input. Assuming that the inputs  $\mathbf{X}_n$  are an i.i.d. sequence, then using properties of the joint entropy for i.i.d. sequences, we have

$$n H[X] \leq L_n < n H[X] + 1 \quad (12.4.33)$$

and dividing out by  $n$ , we obtain

$$H[X] \leq \frac{1}{n} L_n < H[X] + \frac{1}{n} \quad (12.4.34)$$

Thus asymptotically as  $n \rightarrow \infty$ , we have  $\frac{1}{n}L_n \leq H[X]$  and we can attain the optimal expected codeword length for the average codeword length per input.

### Coding Theory Characterisation of Cross Entropy

Based the distribution of  $X$ , denoted  $p(x_i)$ , an optimal code can be designed which minimises the expected codeword length. Denote these optimal lengths  $\ell_i^*$ , ignoring the integer constraints. The entropy of  $X$  (using log base  $D$ ) gives a lower bound on the expected codeword length of the optimal prefix code using a  $D$ -ary alphabet:

$$H[X] \leq \sum_i p(x_i) \ell_i^* \quad (12.4.35)$$

Suppose however an optimal code is designed using an assumed/approximating distribution  $q(x_i)$ , yielding lengths  $l_i^* \geq \log_D \frac{1}{q(x_i)}$ . The actual expected code length would then be lower bounded by

$$H_{p,q} = \sum_i p_i(x_i) \log_D \frac{1}{q(x_i)} \quad (12.4.36)$$

which is the **cross entropy**. Thus an operational interpretation of cross entropy between  $p(x_i)$  and  $q(x_i)$  is the lower bound on actual expected codeword length of the optimal prefix code with an approximating distribution  $q(x_i)$  for  $p(x_i)$ .

$$H_{p,q} \leq \sum_i p(x_i) l_i^* \quad (12.4.37)$$

Since the KL divergence can also be written as  $KL(p\|r) = H_{p,q} - H_p$ , then an operational interpretation of the KL divergence would be the difference in lower bounds of expected codeword length from when the distribution  $p(x_i)$  is known perfectly.

#### 12.4.2 Channel Coding

A communication channel can be described with an input set  $\mathcal{X}$ , an output set  $\mathcal{Y}$ , and a conditional distribution  $p(y|x)$ . The symbol  $x \in \mathcal{X}$  can be interpreted as the symbol we wish to send, and  $y \in \mathcal{Y}$  is the symbol received, which ideally should be the same as  $x$ , but may get corrupted by some noise according to  $p(y|x)$ . We say that a channel is *memoryless* if the current output is conditionally independent of previous inputs or outputs, given the current input.

#### Discrete Channels

A discrete channel is a channel in which  $\mathcal{X}$  and  $\mathcal{Y}$  are finite sets, hence  $p(y|x)$  can be represented using a finite size **transition matrix**. An example of a discrete channel is a *binary channel*, in which  $\mathcal{X} = \{0, 1\}$  and  $\mathcal{Y} = \{0, 1\}$ .

#### Channel Capacity

Note that if a distribution  $p(x)$  over  $\mathcal{X}$  is specified over the input, then this defines the joint distribution  $p(x, y) = p(y|x)p(x)$ . The channel capacity of a discrete channel is defined as

$$C = \max_{p(x)} MI(X; Y) \quad (12.4.38)$$

That is, the maximum mutual information between  $X$  and  $Y$ , with respect to the distribution  $p(x)$  (which induces the joint distribution). A property of channel capacity is that

$$C \geq 0 \quad (12.4.39)$$

because of the property  $\text{MI}(X; Y) \geq 0$  for mutual information. Another property is the simple upper bound

$$C \leq \log |\mathcal{X}| \quad (12.4.40)$$

since

$$C = \max_{p(x)} \text{MI}(X; Y) \quad (12.4.41)$$

$$= \max_{p(x)} \{\text{H}[X] - \text{H}[X|Y]\} \quad (12.4.42)$$

$$\leq \max_{p(x)} \{\text{H}[X]\} \quad (12.4.43)$$

$$= \log |\mathcal{X}| \quad (12.4.44)$$

where  $\max_{p(x)} \{\text{H}[X]\} = \log |\mathcal{X}|$  due to the maximum entropy distribution. By symmetry of the mutual information, we can also show that

$$C \leq \log |\mathcal{Y}| \quad (12.4.45)$$

### Fano's Inequality

Suppose  $X \in \mathcal{X}$  is a random variable that we wish to estimate the value of, based on the observation of another random variable  $Y$ , which is somehow correlated with  $X$ . Let the estimate of  $X$  be given as  $\hat{X} = g(Y)$  for some function  $g(\cdot)$ . Define the error probability as

$$P_e = \Pr(\hat{X} \neq X) \quad (12.4.46)$$

Assume that  $X, Y$  take on values from a finite set, and  $(X, Y, \hat{X})$  forms a Markov chain. Then in terms of logarithm base 2:

$$P_e \geq \frac{\text{H}[X|Y] - 1}{\log |\mathcal{X}|} \quad (12.4.47)$$

which intuitively says that if  $\text{H}[X|Y]$  is large (meaning that  $Y$  contains little information about  $X$ ), then the error probability  $P_e$  cannot be very small.

*Proof.* Write the error probability as

$$\Pr(\hat{X} \neq X) = \sum_{x \in \mathcal{X}} p(x) \Pr(\hat{X} \neq X | X = x) \quad (12.4.48)$$

and define and indicator variable for the error  $E = \mathbb{I}_{\{\hat{X} \neq X\}}$ . From the chain rule of conditional entropy, we can write  $\text{H}[E, X | \hat{X}]$  in two ways:

$$\text{H}[E, X | \hat{X}] = \text{H}[X | \hat{X}] + \text{H}[E | X, \hat{X}] \quad (12.4.49)$$

$$\text{H}[E, X | \hat{X}] = \text{H}[E | \hat{X}] + \text{H}[X | E, \hat{X}] \quad (12.4.50)$$

We know that  $H[E|\widehat{X}] \leq H[E]$ , since conditioning never increases entropy, via the properties of mutual information. Also,  $H[X|E, \widehat{X}] = 0$  since  $E$  is entirely determined if  $X$  and  $\widehat{X}$  are known. Furthermore, as  $H[X|E, \widehat{X}]$  is an expectation, it can be written and bounded as

$$H[X|E, \widehat{X}] = \Pr(E=0)H[X|\widehat{X}, E=0] + \Pr(E=1)H[X|\widehat{X}, E=1] \quad (12.4.51)$$

$$= (1 - P_e) \cdot 0 + P_e H[X|\widehat{X}, E=1] \quad (12.4.52)$$

$$\leq P_e H[X] \quad (12.4.53)$$

$$\leq P_e \log |\mathcal{X}| \quad (12.4.54)$$

where  $H[X|\widehat{X}, E=0] = 0$  because we know  $X = \widehat{X}$  with certainty if  $\widehat{X}$  and  $E=0$  are given. Also, the bound  $\log |\mathcal{X}|$  comes from the bound on entropy of a random variable. Putting all these arguments together, we obtain

$$H[X|\widehat{X}] \leq H[E] + P_e \log |\mathcal{X}| \quad (12.4.55)$$

Now since  $(X, Y, \widehat{X})$  forms a Markov chain, use the data processing inequality to say that

$$MI(X; Y) \geq MI(X; \widehat{X}) \quad (12.4.56)$$

Since  $MI(X; Y) = H[X] - H[X|Y]$ , this means that

$$H[X] - H[X|Y] \geq H[X] - H[X|\widehat{X}] \quad (12.4.57)$$

$$H[X|Y] \leq H[X|\widehat{X}] \quad (12.4.58)$$

and therefore

$$H[X|Y] \leq H[E] + P_e \log |\mathcal{X}| \quad (12.4.59)$$

Rearranging, this gives

$$P_e \geq \frac{H[X|Y] - H[E]}{\log |\mathcal{X}|} \quad (12.4.60)$$

$$\geq \frac{H[X|Y] - 1}{\log |\mathcal{X}|} \quad (12.4.61)$$

where we used  $H[E] \leq 1$  since  $E$  is a binary random variable, i.e. it takes one bit to transmit the result of a binary random variable, by the source coding characterisation of entropy.  $\square$

### Channel Coding Theorem [92]

#### Gaussian Channels

##### 12.4.3 Differential Privacy

##### 12.4.4 Perplexity

The perplexity of a probability distribution  $p(x)$  is defined as 2 to the exponent of the entropy of  $p(x)$ , measured in bits. That is,

$$2^{H_p} = 2^{-\sum_x p(x) \log p(x)} \quad (12.4.62)$$

Perplexity may be roughly interpreted as a measure of difficulty of guessing an outcome from  $p(x)$ . Let  $L$  be a random variable for the number of bits it takes to transmit an event from  $p(x)$  using an optimal prefix code. An event that can be transmitted in  $L$  bits of information takes  $2^L$  guesses on average to guess correctly, by guessing uniformly at random. By taking the interpretation that the entropy  $H_p$  is a lower bound on expected bits to transmit an event from  $p(x)$  using this optimal prefix code, then we have

$$2^{H_p} \leq 2^{\mathbb{E}[L]} \quad (12.4.63)$$

$$\leq \mathbb{E}[2^L] \quad (12.4.64)$$

where the last inequality comes from Jensen's inequality. Therefore perplexity is a lower bound on the expected 'difficulty' of guessing events from  $p(x)$ .

### Perplexity of a Probability Model

Suppose we have a probability model (i.e. distribution)  $q(x)$  for a sample with empirical distribution  $p(x)$ . The perplexity of the model may be defined as  $2$  to the exponent of the cross entropy  $H_{p,q}$ , measured in bits, i.e.

$$2^{H_{p,q}} = 2^{-\sum_x p(x) \log q(x)} \quad (12.4.65)$$

This perplexity may be roughly interpreted as the average difficulty in guessing events from the sample using the model, and can be used as a measure of model fit.

## 12.5 Information Criteria

### 12.5.1 Akaike Information Criterion [130]

The AIC of a model parametrised by  $\theta$  can be written in terms of its log likelihood  $\log \mathcal{L}(\theta)$ .

$$\text{AIC}(\theta) = -2 \max_{\theta} \{\log \mathcal{L}(\theta)\} + 2 \dim(\theta) \quad (12.5.1)$$

where  $\dim(\theta)$  is the length of the parameter vector  $\theta$ . To derive the AIC [130], first consider a prediction error term  $\varepsilon_i(\theta)$  parameterised by  $\theta$ , which is the prediction error of the  $i^{\text{th}}$  observation in the dataset  $\mathcal{D}_N$ , where  $N$  is the sample size. Let  $\ell(\varepsilon)$  be a cost on the prediction error. The estimator  $\hat{\theta}_N$  is an extremum estimator for the mean cost  $V_N(\theta, \mathcal{D}_N)$  given by

$$\hat{\theta}_N = \operatorname{argmin}_{\theta} V_N(\theta, \mathcal{D}_N) \quad (12.5.2)$$

$$= \operatorname{argmin}_{\theta} \left\{ \frac{1}{N} \sum_{i=1}^N \ell(\varepsilon_i(\theta)) \right\} \quad (12.5.3)$$

Define  $\bar{V}(\theta)$  as the limit of the mean cost as  $N \rightarrow \infty$ . That is,

$$\bar{V}(\theta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \ell(\varepsilon_i(\theta)) \quad (12.5.4)$$

If the errors  $\varepsilon_i$  are i.i.d. with the distribution of some 'true error'  $\varepsilon$ , then by the Law of Large Numbers we can write this as:

$$\bar{V}(\theta) = \mathbb{E}[\ell(\varepsilon(\theta)) | \theta] \quad (12.5.5)$$

where the conditional expectation is taken over the data generating process. For an estimate  $\hat{\theta}_N$ , we seek to relate the terms  $\mathbb{E}[V_N(\hat{\theta}_N, \mathcal{D}_N)]$  and  $\mathbb{E}[\bar{V}(\hat{\theta}_N)]$ , where the expectations are

taken over the estimator  $\hat{\theta}_N$ . There is an important distinction between these two terms, which is that the term  $\mathbb{E} [V_N(\hat{\theta}_N, \mathcal{D}_N)]$  expresses the expected mean cost as evaluated on the training data  $\mathcal{D}_N$ , while the term  $\mathbb{E} [\bar{V}(\hat{\theta}_N)]$  represents the expected cost of prediction on unseen data (i.e. validation data). Note that we assume that the training data and validation data follow the same data generating process. To obtain these expectations, first perform a Taylor series expansion of  $\bar{V}(\theta)$  about the ‘true’ parameter  $\theta^*$ .

$$\bar{V}(\hat{\theta}_N) \approx \bar{V}(\theta^*) + \frac{1}{2} (\hat{\theta}_N - \theta^*)^\top \nabla_\theta^2 \bar{V}(\theta^*) (\hat{\theta}_N - \theta^*) \quad (12.5.6)$$

Note that  $\theta^*$  minimises  $\bar{V}(\theta)$ , so the gradient vanishes. Perform a similar expansion for  $V_N(\theta, \mathcal{D}_N)$  about  $\hat{\theta}_N$ , noting that this is minimised at  $\hat{\theta}_N$ .

$$V_N(\theta^*, \mathcal{D}_N) \approx V_N(\hat{\theta}_N, \mathcal{D}_N) + \frac{1}{2} (\theta^* - \hat{\theta}_N)^\top \nabla_\theta^2 \bar{V}(\hat{\theta}_N) (\theta^* - \hat{\theta}_N) \quad (12.5.7)$$

Rearranging gives

$$V_N(\hat{\theta}_N, \mathcal{D}_N) \approx V_N(\theta^*, \mathcal{D}_N) - \frac{1}{2} (\hat{\theta}_N - \theta^*)^\top \nabla_\theta^2 \bar{V}(\hat{\theta}_N) (\hat{\theta}_N - \theta^*) \quad (12.5.8)$$

Take the expectation of the expression for  $\bar{V}(\hat{\theta}_N)$ :

$$\mathbb{E} [\bar{V}(\hat{\theta}_N)] \approx \bar{V}(\theta^*) + \frac{1}{2} \mathbb{E} \left[ \text{trace} \left( (\hat{\theta}_N - \theta^*)^\top \nabla_\theta^2 \bar{V}(\theta^*) (\hat{\theta}_N - \theta^*) \right) \right] \quad (12.5.9)$$

since  $\bar{V}(\theta^*)$  is not a random quantity and the quadratic form is a scalar. Then since the trace is invariant to cyclic permutations,

$$\mathbb{E} [\bar{V}(\hat{\theta}_N)] \approx \bar{V}(\theta^*) + \frac{1}{2} \text{trace} \left( \nabla_\theta^2 \bar{V}(\theta^*) \mathbb{E} \left[ (\hat{\theta}_N - \theta^*)^\top (\hat{\theta}_N - \theta^*) \right] \right) \quad (12.5.10)$$

where we have used the facts that the trace is a linear operator (and so commutes with sums hence expectations), and  $\nabla_\theta^2 \bar{V}(\theta^*)$  is also not a random quantity. Then suppose that  $\hat{\theta}_N$  has an asymptotic covariance of

$$\mathbb{E} \left[ (\hat{\theta}_N - \theta^*)^\top (\hat{\theta}_N - \theta^*) \right] \rightarrow \frac{1}{N} C \quad (12.5.11)$$

as  $N \rightarrow \infty$ . Then for large  $N$ ,

$$\mathbb{E} [\bar{V}(\hat{\theta}_N)] \approx \bar{V}(\theta^*) + \frac{1}{2N} \text{trace} (\nabla_\theta^2 \bar{V}(\theta^*) C) \quad (12.5.12)$$

Now take the expectation of the expression for  $V_N(\hat{\theta}_N, \mathcal{D}_N)$ :

$$\mathbb{E} [V_N(\hat{\theta}_N, \mathcal{D}_N)] \approx V_N(\theta^*, \mathcal{D}_N) - \frac{1}{2} \mathbb{E} \left[ \text{trace} \left( (\hat{\theta}_N - \theta^*)^\top \nabla_\theta^2 \bar{V}(\hat{\theta}_N) (\hat{\theta}_N - \theta^*) \right) \right] \quad (12.5.13)$$

as similar to before. For large  $N$ ,  $V_N(\theta^*, \mathcal{D}_N) \approx \bar{V}(\theta^*)$  and  $\nabla_\theta^2 \bar{V}(\hat{\theta}_N) \approx \nabla_\theta^2 \bar{V}(\theta^*)$  so

$$\mathbb{E} [V_N(\hat{\theta}_N, \mathcal{D}_N)] \approx \bar{V}(\theta^*) - \frac{1}{2N} \text{trace} (\nabla_\theta^2 \bar{V}(\theta^*) C) \quad (12.5.14)$$

Combining both these expectations, we see that

$$\mathbb{E} [\bar{V}(\hat{\theta}_N)] \approx \mathbb{E} [V_N(\hat{\theta}_N, \mathcal{D}_N)] + \frac{1}{N} \text{trace} (\nabla_\theta^2 \bar{V}(\theta^*) C) \quad (12.5.15)$$

If our criterion is to minimise the expected prediction error on validation data  $\mathbb{E} \left[ \bar{V} \left( \hat{\theta}_N \right) \right]$ , then we seek to minimise the right hand side. Suppose that the prediction error is chosen to be the log-likelihood, i.e.

$$V_N(\theta, \mathcal{D}_N) = -\frac{1}{N} \log \mathcal{L}(\theta; \mathcal{D}_N) \quad (12.5.16)$$

The best estimate of  $\mathbb{E} \left[ V_N \left( \hat{\theta}_N, \mathcal{D}_N \right) \right]$  is  $-\frac{1}{N} \log \mathcal{L} \left( \hat{\theta}_N; \mathcal{D}_N \right)$  itself (which becomes increasingly better for large  $N$ ), while the asymptotic covariance  $C$  is the inverse of the Fisher information (i.e. in this case  $C = \nabla_{\theta}^2 \bar{V}(\theta^*)^{-1}$ ). Therefore an estimate of the expected prediction error is

$$\widehat{\mathbb{E}} \left[ \bar{V} \left( \hat{\theta}_N \right) \right] = -\frac{1}{N} \max_{\theta} \log \mathcal{L}(\theta; \mathcal{D}_N) + \frac{\text{trace} \left( \nabla_{\theta}^2 \bar{V}(\theta^*) \nabla_{\theta}^2 \bar{V}(\theta^*)^{-1} \right)}{N} \quad (12.5.17)$$

$$= -\frac{1}{N} \log \mathcal{L} \left( \hat{\theta}_N; \mathcal{D}_N \right) + \frac{\dim(\hat{\theta}_N)}{N} \quad (12.5.18)$$

this is the same as the expression above for AIC (up to multiplication by a constant). Intuitively, the AIC aims to trade off between a high likelihood and low model complexity (measured by the number of parameters) by introducing a penalty on the number of parameters, since too high model complexity may lead to overfitting or loss in interpretability of the model. The AIC can be used for model selection, where we select the model with the lowest AIC. There also exist formulae for corrected AIC based on sample size  $N$  [44].

### 12.5.2 Bayesian Information Criterion

The Bayesian information criterion (BIC), also known as the Schwarz information criterion, is a property of a model parametrised by  $\theta$  and can be defined in terms of its maximum log likelihood  $\log \mathcal{L}(\theta)$  from  $n$  observations.

$$\text{BIC}(\theta) = \log n \cdot \dim(\theta) - 2 \max_{\theta} \{ \log \mathcal{L}(\theta) \} \quad (12.5.19)$$

Like the AIC, the BIC is also used as a model selection technique, where we aim to select the model with lowest BIC (as defined above). The BIC can be derived as follows. Let  $M_1, M_2, \dots$  be multiple competing models. Given data  $\mathbf{y} = (y_1, y_2, \dots, y_n)$ , we address the goal of finding the model with the highest posterior probability:

$$p(M_i | \mathbf{y}) = \frac{p(\mathbf{y} | M_i) p(M_i)}{p(\mathbf{y})} \quad (12.5.20)$$

If we assume that each model is equally likely with a flat prior (i.e.  $p(M_i)$  is a constant), then the task becomes the same as finding the model with the highest marginal likelihood:  $M^* = \text{argmax}_i p(\mathbf{y} | M_i)$ . We call  $p(\mathbf{y} | M_i)$  the marginal likelihood because it is obtained by marginalising over the prior distribution for the parameters  $\boldsymbol{\theta}_i$  associated with model  $M_i$ :

$$p(\mathbf{y} | M_i) = \int p(\mathbf{y} | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \quad (12.5.21)$$

Denote  $k_i$  to be the number of parameters in model  $M_i$  so that  $\boldsymbol{\theta}_i$  has dimension  $k_i$ . Also note that  $p(\mathbf{y} | \boldsymbol{\theta}_i)$  is the likelihood of  $\boldsymbol{\theta}_i$  given the data. Now suppose that  $p(\mathbf{y} | \boldsymbol{\theta}_i) p(\boldsymbol{\theta})$  is twice differentiable in  $\boldsymbol{\theta}_i$  and has a unique global maximum  $\boldsymbol{\theta}_i^*$ , which implies  $\log p(\mathbf{y} | \boldsymbol{\theta}_i) p(\boldsymbol{\theta})$  has the same global maximum. It is then valid to apply a Laplace approximation for  $p(\mathbf{y} | M_i)$ :

$$p(\mathbf{y} | M_i) = \int \exp [\log p(\mathbf{y} | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i)] d\boldsymbol{\theta}_i \quad (12.5.22)$$

$$\approx \exp [\log p(\mathbf{y}|\boldsymbol{\theta}_i^*) p(\boldsymbol{\theta}_i^*)] \sqrt{\frac{(2\pi)^{k_i}}{\det(-\nabla_{\boldsymbol{\theta}_i}^2 p(\mathbf{y}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i))}} \quad (12.5.23)$$

$$= p(\mathbf{y}|\boldsymbol{\theta}_i^*) p(\boldsymbol{\theta}_i^*) \sqrt{\frac{(2\pi)^{k_i}}{\det(-\nabla_{\boldsymbol{\theta}_i}^2 \log p(\mathbf{y}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i))}} \quad (12.5.24)$$

Taking the log of both sides:

$$\log p(\mathbf{y}|M_i) \approx \log p(\mathbf{y}|\boldsymbol{\theta}_i^*) + \log p(\boldsymbol{\theta}_i^*) + \frac{k_i}{2} \log 2\pi - \frac{1}{2} \log \det(-\nabla_{\boldsymbol{\theta}_i}^2 p(\mathbf{y}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i)) \quad (12.5.25)$$

Note that under independence of each observation,

$$\log p(\mathbf{y}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i) = \sum_{j=1}^n \log p(y_j|\boldsymbol{\theta}_i) + \log p(\boldsymbol{\theta}_i) \quad (12.5.26)$$

Hence

$$-\nabla_{\boldsymbol{\theta}_i}^2 \log p(\mathbf{y}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i) = -\sum_{j=1}^n \nabla_{\boldsymbol{\theta}_i}^2 \log p(y_j|\boldsymbol{\theta}_i) - \nabla_{\boldsymbol{\theta}_i}^2 \log p(\boldsymbol{\theta}_i) \quad (12.5.27)$$

where each  $-\nabla_{\boldsymbol{\theta}_i}^2 \log p(y_j|\boldsymbol{\theta}_i)$  is the observed information for a single observation. For large  $n$ , then by application of the Law of Large Numbers we have

$$-\nabla_{\boldsymbol{\theta}_i}^2 \log p(\mathbf{y}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i) \approx nI_{\boldsymbol{\theta}_i} - \nabla_{\boldsymbol{\theta}_i}^2 \log p(\boldsymbol{\theta}_i) \quad (12.5.28)$$

where  $I_{\boldsymbol{\theta}_i}$  is the Fisher information for a single observation. Moreover, for large  $n$ , the overall Hessian term is dominated by the term involving  $n$  so we are able to approximate:

$$\det(-\nabla_{\boldsymbol{\theta}_i}^2 \log p(\mathbf{y}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i)) \approx \det(nI_{\boldsymbol{\theta}_i}) \quad (12.5.29)$$

$$= n^{k_i} \det(I_{\boldsymbol{\theta}_i}) \quad (12.5.30)$$

Thus the log marginal likelihood of model  $M_i$  gets approximated by

$$\log p(\mathbf{y}|M_i) \approx \log p(\mathbf{y}|\boldsymbol{\theta}_i^*) + \log p(\boldsymbol{\theta}_i^*) + \frac{k_i}{2} \log 2\pi - \frac{k_i}{2} \log n - \frac{1}{2} \log \det(I_{\boldsymbol{\theta}_i}) \quad (12.5.31)$$

Finally, for large  $n$  we again ignore terms which do not contain  $n$ , leaving

$$\log p(\mathbf{y}|M_i) \approx \log p(\mathbf{y}|\boldsymbol{\theta}_i^*) - \frac{k_i}{2} \log n \quad (12.5.32)$$

The right-hand side of this is the BIC of model  $M_i$ , where we select the model with the highest BIC.

### 12.5.3 Deviance Information Criterion

## 12.6 Optimal Experimental Design

### 12.6.1 Optimal Experimental Design for Least Squares

Consider a data generating process

$$y_i = \beta^\top x_i + \varepsilon_i \quad (12.6.1)$$

where  $\varepsilon_i$  are i.i.d. and  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ . For an estimation problem comprising the ordinary least squares estimator, the associated covariance matrix on the estimate  $\hat{\beta}$  with sample size  $n$  is

$$\text{Cov}(\hat{\beta}) = \sigma^2 \left( \sum_{i=1}^n x_i x_i^\top \right)^{-1} \quad (12.6.2)$$

where we have assumed the regressor vectors  $x_1, \dots, x_n$  are non-random. The goal of optimal experimental design as the experiment designer is to choose the vectors  $x_1, \dots, x_n$  which in some sense reduces the covariance of the estimates  $\hat{\beta}$ . Note that the Fisher information in the maximum likelihood estimate is the inverse of this, given by

$$\mathcal{I}(\beta) = \frac{1}{\sigma^2} \sum_{i=1}^n x_i x_i^\top \quad (12.6.3)$$

so by minimising the covariance we are in a sense designing the most ‘informative’ experiments by maximising the Fisher information. One tractable formulation involves introducing a ‘menu’ of  $m$  fixed vectors  $v_1, \dots, v_m$  which we can choose each sample point  $x_i$  from. Then for each of these vectors we have a weighting  $\lambda_j$  which can be interpreted in one of two ways:

1. The weighting  $\lambda_j$  between 0 and 1 indicates the (approximate) proportion of sample points at  $v_j$ . That is, we can take the rounding of  $\lambda_j m$  and choose that many sample points at  $v_j$ .
2. The weighting  $\lambda_j$  between 0 and 1 can represent the probability of picking a sample point at  $v_j$ , in a randomised experiment. Note that in a randomised experiment the Fisher information is

$$\mathcal{I}(\beta) = \frac{1}{\sigma^2} \mathbb{E} \left[ \sum_{i=1}^n x_i x_i^\top \right] \quad (12.6.4)$$

$$= \frac{1}{\sigma^2} \sum_{j=1}^m \lambda_j v_j v_j^\top \quad (12.6.5)$$

so this is still a valid interpretation.

The optimal experimental design problem then becomes a problem of finding a probability distribution with masses  $\lambda_j$  which minimise the covariance. Since the covariance is matrix-valued, we require introducing several alternative scalarisations, some which have geometric interpretations in terms of constructed confidence ellipsoids.

### A-Optimal Design

In A-optimal design, we solve the optimisation problem to minimise the trace of the covariance matrix:

$$\begin{aligned} \min_{\lambda_1, \dots, \lambda_m} \quad & \text{trace} \left( \left( \sum_{j=1}^m \lambda_j v_j v_j^\top \right)^{-1} \right) \\ \text{s.t.} \quad & \lambda_1 + \dots + \lambda_m = 1 \\ & \lambda_j \geq 0, \quad j = 1, \dots, m \end{aligned} \quad (12.6.6)$$

Note that we have left out the variance  $\sigma^2$  as it leaves the problem unchanged up to a positive scaling. This problem can be formulated as a semi-definite program, which is a convex optimisation problem.

## D-Optimal Design

In D-optimal design, we minimise the determinant of the covariance matrix (or equivalently, the log determinant):

$$\begin{aligned} \min_{\lambda_1, \dots, \lambda_m} \quad & \det \left( \left( \sum_{j=1}^m \lambda_j v_j v_j^\top \right)^{-1} \right) \\ \text{s.t.} \quad & \lambda_1 + \dots + \lambda_m = 1 \\ & \lambda_j \geq 0, \quad j = 1, \dots, m \end{aligned} \tag{12.6.7}$$

which is also a convex optimisation problem. This corresponds to minimising the volume of the confidence ellipsoid for a fixed confidence level.

## E-Optimal Design

In E-optimal design, we minimise the 2-norm of the covariance matrix:

$$\begin{aligned} \min_{\lambda_1, \dots, \lambda_m} \quad & \left\| \left( \sum_{j=1}^m \lambda_j v_j v_j^\top \right)^{-1} \right\|_2 \\ \text{s.t.} \quad & \lambda_1 + \dots + \lambda_m = 1 \\ & \lambda_j \geq 0, \quad j = 1, \dots, m \end{aligned} \tag{12.6.8}$$

which is also a convex optimisation problem since it can be cast as a semi-definite program. Since this is equivalent to minimising the maximum eigenvalue of the covariance matrix, this corresponds to minimising the maximum diameter of the confidence ellipsoid for a fixed confidence level.

## Kiefer-Wolfowitz Theorem [121]

## 12.7 Statistical Distances

### 12.7.1 Total Variation Distance

Let  $\mathbb{P}$  and  $\mathbb{Q}$  be two probability measures defined on the same sample space  $\Omega$  with  $\sigma$ -algebra  $\mathcal{F}$ . The total variation distance (sometimes referred to as the total variational distance) between  $\mathbb{P}$  and  $\mathbb{Q}$  is defined as

$$\delta(\mathbb{P}, \mathbb{Q}) = \sup_{E \in \mathcal{F}} |\mathbb{P}(E) - \mathbb{Q}(E)| \tag{12.7.1}$$

Roughly speaking, this conveys the distance between probability distributions in terms of the largest possible difference in probabilities that the distributions can assign to the same event. If  $\Omega = \mathbb{R}$ , then we have the following relationship with the Kolmogorov-Smirnov distance (denoted  $d(\mathbb{P}, \mathbb{Q})$ ):

$$d(\mathbb{P}, \mathbb{Q}) \leq \delta(\mathbb{P}, \mathbb{Q}) \tag{12.7.2}$$

because the total variation distance considers more than just the cumulative distribution function.

## Variation Distance [71]

A related distance to the total variation distance is known as the variation distance. It may be defined as

$$\|\mathbb{P} - \mathbb{Q}\|_1 = \sup_{\mathcal{E}} \sum_{E \in \mathcal{E}} |\mathbb{P}(E) - \mathbb{Q}(E)| \tag{12.7.3}$$

where the supremum is over all possible partitions  $\mathcal{E}$  of the sample space  $\Omega$  into events, i.e.  $\bigcup_{E \in \mathcal{E}} = \Omega$ . If  $\Omega$  is countable, e.g. corresponding to discrete distributions, then the variation distance coincides with the  $\ell_1$  norm between the two distributions:

$$\|\mathbb{P} - \mathbb{Q}\|_1 = \sum_{\omega \in \Omega} |\mathbb{P}(\omega) - \mathbb{Q}(\omega)| \quad (12.7.4)$$

To show this equivalence, first consider that  $\Omega$  is a valid partition, and by definition that

$$\|\mathbb{P} - \mathbb{Q}\|_1 = \sup_{\mathcal{E}} \sum_{E \in \mathcal{E}} |\mathbb{P}(E) - \mathbb{Q}(E)| \quad (12.7.5)$$

$$\geq \sum_{\omega \in \Omega} |\mathbb{P}(\omega) - \mathbb{Q}(\omega)| \quad (12.7.6)$$

Conversely however, for any event  $E \subset \Omega$  we have

$$|\mathbb{P}(E) - \mathbb{Q}(E)| = \left| \sum_{\omega \in E} \mathbb{P}(\omega) - \sum_{\omega \in E} \mathbb{Q}(\omega) \right| \quad (12.7.7)$$

$$= \left| \sum_{\omega \in E} (\mathbb{P}(\omega) - \mathbb{Q}(\omega)) \right| \quad (12.7.8)$$

$$\leq \sum_{\omega \in E} |\mathbb{P}(\omega) - \mathbb{Q}(\omega)| \quad (12.7.9)$$

by the triangle inequality. Since this is shown to hold for any event  $E$ , then

$$\sup_{\mathcal{E}} \sum_{E \in \mathcal{E}} |\mathbb{P}(E) - \mathbb{Q}(E)| \leq \sum_{\omega \in \Omega} |\mathbb{P}(\omega) - \mathbb{Q}(\omega)| \quad (12.7.10)$$

The variation distance is also explicitly related to the total variation distance by a factor of 2:

$$\|\mathbb{P} - \mathbb{Q}\|_1 = 2\delta(\mathbb{P}, \mathbb{Q}) \quad (12.7.11)$$

*Proof.* Let  $\mathcal{E} = \{E, \bar{E}\}$  be any partition into two complementary events. By definition of the variation distance,

$$\|\mathbb{P} - \mathbb{Q}\|_1 \geq |\mathbb{P}(E) - \mathbb{Q}(E)| + |\mathbb{P}(\bar{E}) - \mathbb{Q}(\bar{E})| \quad (12.7.12)$$

$$= |\mathbb{P}(E) - \mathbb{Q}(E)| + |(1 - \mathbb{P}(E)) - (1 - \mathbb{Q}(E))| \quad (12.7.13)$$

$$= 2|\mathbb{P}(E) - \mathbb{Q}(E)| \quad (12.7.14)$$

Since  $E$  can be anything from the  $\sigma$ -algebra  $\mathcal{F}$ , then

$$\|\mathbb{P} - \mathbb{Q}\|_1 \geq 2 \sup_{E \in \mathcal{F}} |\mathbb{P}(E) - \mathbb{Q}(E)| \quad (12.7.15)$$

$$= 2\delta(\mathbb{P}, \mathbb{Q}) \quad (12.7.16)$$

where we have recognised the definition of the total variation distance. Conversely, let  $\mathcal{E}^*$  be the partition which achieves the supremum in the definition of the variation distance, provided it exists. If it does not exist, we can still find a partition  $\mathcal{E}^*$  that approximately achieves the supremum, given by

$$\sum_{E \in \mathcal{E}^*} |\mathbb{P}(E) - \mathbb{Q}(E)| \geq \|\mathbb{P} - \mathbb{Q}\|_1 - \varepsilon \quad (12.7.17)$$

for some  $\varepsilon > 0$  (or otherwise this holds with equality and  $\varepsilon = 0$  in the case that  $\mathcal{E}^*$  achieves the supremum). Now define the set  $\mathcal{A} \subset \Omega$  as the union of all the  $E \in \mathcal{E}^*$  for which  $\mathbb{P}(E) \geq \mathbb{Q}(E)$ . Then we can express

$$\sum_{E \in \mathcal{E}^*} |\mathbb{P}(E) - \mathbb{Q}(E)| = \sum_{E \in \mathcal{A}} (\mathbb{P}(E) - \mathbb{Q}(E)) + \sum_{E \in \mathcal{E}^* \setminus \mathcal{A}} (\mathbb{Q}(E) - \mathbb{P}(E)) \quad (12.7.18)$$

$$= \mathbb{P}(\mathcal{A}) - \mathbb{Q}(\mathcal{A}) + \mathbb{Q}(\mathcal{E}^* \setminus \mathcal{A}) - \mathbb{P}(\mathcal{E}^* \setminus \mathcal{A}) \quad (12.7.19)$$

$$= \mathbb{P}(\mathcal{A}) - \mathbb{Q}(\mathcal{A}) + (1 - \mathbb{Q}(\mathcal{A})) - (1 - \mathbb{P}(\mathcal{A})) \quad (12.7.20)$$

$$= 2(\mathbb{P}(\mathcal{A}) - \mathbb{Q}(\mathcal{A})) \quad (12.7.21)$$

$$\leq 2 \sup_{E \in \mathcal{F}} |\mathbb{P}(E) - \mathbb{Q}(E)| \quad (12.7.22)$$

Thus

$$\|\mathbb{P} - \mathbb{Q}\|_1 - \varepsilon \leq 2 \sup_{E \in \mathcal{F}} |\mathbb{P}(E) - \mathbb{Q}(E)| \quad (12.7.23)$$

$$= 2\delta(\mathbb{P}, \mathbb{Q}) \quad (12.7.24)$$

but since  $\varepsilon$  is either zero or can be made arbitrarily small, this completes the proof.  $\square$

Another characterisation of the total variation distance and variation distance is in terms of an event  $B$  which achieves the supremum.

**Theorem 12.5.** *The total variation distance can be expressed as*

$$\delta(\mathbb{P}, \mathbb{Q}) = \sup_{E \in \mathcal{F}} |\mathbb{P}(E) - \mathbb{Q}(E)| \quad (12.7.25)$$

$$= \mathbb{P}(B) - \mathbb{Q}(B) \quad (12.7.26)$$

where in the case that the sample space  $\Omega$  is countable, the event  $B$  is given by

$$B := \{\omega \in \Omega : \mathbb{P}(\omega) \geq \mathbb{Q}(\omega)\} \quad (12.7.27)$$

Otherwise,  $B$  takes on an analogous event except using the partition which achieves the supremum in the definition of the variation distance.

*Proof.* For any event  $E$ , then by definition of  $B$  we have  $\mathbb{P}(\omega) - \mathbb{Q}(\omega) \geq 0$  for any  $\omega \in B$  and moreover any  $\omega \in E \cap B$ . Likewise, we have  $\mathbb{P}(\omega) - \mathbb{Q}(\omega) < 0$  for any  $\omega \in \overline{B}$  and moreover any  $\omega \in E \cap \overline{B}$ . Using these properties, we show

$$\mathbb{P}(E) - \mathbb{Q}(E) = \mathbb{P}(E \cap B) + \mathbb{P}(E \cap \overline{B}) - \mathbb{Q}(E \cap B) - \mathbb{Q}(E \cap \overline{B}) \quad (12.7.28)$$

$$= \mathbb{P}(E \cap B) - \mathbb{Q}(E \cap B) + \underbrace{\mathbb{P}(E \cap \overline{B}) - \mathbb{Q}(E \cap \overline{B})}_{<0} \quad (12.7.29)$$

$$\leq \mathbb{P}(E \cap B) - \mathbb{Q}(E \cap B) \quad (12.7.30)$$

and

$$\mathbb{P}(B) - \mathbb{Q}(B) = \mathbb{P}(E \cap B) + \mathbb{P}(\overline{E} \cap B) - \mathbb{Q}(E \cap B) - \mathbb{Q}(\overline{E} \cap B) \quad (12.7.31)$$

$$= \mathbb{P}(E \cap B) - \mathbb{Q}(E \cap B) + \underbrace{\mathbb{P}(\overline{E} \cap B) - \mathbb{Q}(\overline{E} \cap B)}_{\geq 0} \quad (12.7.32)$$

$$\geq \mathbb{P}(E \cap B) - \mathbb{Q}(E \cap B) \quad (12.7.33)$$

Putting these together, this implies

$$\mathbb{P}(E) - \mathbb{Q}(E) \leq \mathbb{P}(B) - \mathbb{Q}(B) \quad (12.7.34)$$

On the other hand, we similarly have

$$\mathbb{Q}(E) - \mathbb{P}(E) = \mathbb{Q}(E \cap B) + \mathbb{Q}(E \cap \overline{B}) - \mathbb{P}(E \cap B) - \mathbb{P}(E \cap \overline{B}) \quad (12.7.35)$$

$$\leq \mathbb{Q}(E \cap \overline{B}) - \mathbb{P}(E \cap \overline{B}) \quad (12.7.36)$$

and

$$\mathbb{Q}(\overline{B}) - \mathbb{P}(\overline{B}) = \mathbb{Q}(E \cap \overline{B}) + \mathbb{Q}(\overline{E} \cap \overline{B}) - \mathbb{P}(E \cap \overline{B}) - \mathbb{P}(\overline{E} \cap \overline{B}) \quad (12.7.37)$$

$$\geq \mathbb{Q}(E \cap \overline{B}) - \mathbb{P}(E \cap \overline{B}) \quad (12.7.38)$$

so

$$\mathbb{Q}(E) - \mathbb{P}(E) \leq \mathbb{Q}(\overline{B}) - \mathbb{P}(\overline{B}) \quad (12.7.39)$$

$$= (1 - \mathbb{Q}(B)) - (1 - \mathbb{P}(B)) \quad (12.7.40)$$

$$= \mathbb{P}(B) - \mathbb{Q}(B) \quad (12.7.41)$$

Therefore

$$|\mathbb{P}(E) - \mathbb{Q}(E)| \leq \mathbb{P}(B) - \mathbb{Q}(B) \quad (12.7.42)$$

for any event  $E \in \mathcal{F}$  but since the event  $B$  achieves this with equality, then

$$\sup_{E \in \mathcal{F}} |\mathbb{P}(E) - \mathbb{Q}(E)| = \mathbb{P}(B) - \mathbb{Q}(B) \quad (12.7.43)$$

□

### Pinsker's Inequality [30, 55]

The total variation distance is related to the KL divergence by

$$\delta(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} \text{KL}(\mathbb{P} \parallel \mathbb{Q})} \quad (12.7.44)$$

or equivalently,

$$\text{KL}(\mathbb{P} \parallel \mathbb{Q}) \geq 2\delta(\mathbb{P}, \mathbb{Q})^2 \quad (12.7.45)$$

*Proof.* We first prove the result in the simpler case with binary sample space  $\Omega = \{0, 1\}$ . Let  $q = \mathbb{Q}(1)$  and  $p = \mathbb{P}(1)$ . Then

$$\text{KL}(\mathbb{P} \parallel \mathbb{Q}) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \quad (12.7.46)$$

and

$$\delta(\mathbb{P}, \mathbb{Q}) = \max \{|p - q|, |(1-p) - (1-q)|\} \quad (12.7.47)$$

$$= |p - q| \quad (12.7.48)$$

Thus we are required to show

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq 2(p-q)^2 \quad (12.7.49)$$

Equivalently, define the function in  $q$ :

$$\varphi(q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - 2(p-q)^2 \quad (12.7.50)$$

and show that  $\varphi(q)$  is non-negative. Differentiating, we find

$$\varphi'(q) = p \left( -\frac{p}{q^2} \right) \cdot \frac{1}{p/q} + (1-p) \left[ \frac{1-p}{(1-q)^2} \right] \cdot \frac{1-q}{1-p} + 4(p-q) \quad (12.7.51)$$

$$= -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) \quad (12.7.52)$$

$$= \frac{-p(1-q) + q(1-p)}{q(1-q)} + 4(p-q) \quad (12.7.53)$$

$$= \frac{q-p}{q(1-q)} + 4(p-q) \quad (12.7.54)$$

$$= (q-p) \left[ \frac{1}{q(1-q)} - 4 \right] \quad (12.7.55)$$

Because  $q(1-q)$  attains a maximum of  $1/4$ , then the second factor is non-negative. Thus the behaviour of  $\varphi(q)$  depends on the sign of  $q-p$ . If  $q < p$ , then  $\varphi(q)$  is non-increasing, and if  $q > p$ , then  $\varphi(q)$  is non-decreasing. Then at  $q = p$ , the function  $\varphi(q)$  must attain a minimum, and this minimum is  $\varphi(p) = 0$ , therefore  $\varphi(q)$  is non-negative. Considering the general case, let  $E^*$  be the event which achieves the supremum as defined in the total variation distance. If such an event does not exist, we can still find an event which achieves arbitrarily close to the supremum. Then we only worry about the former case and write

$$\delta(\mathbb{P}, \mathbb{Q}) = |\mathbb{P}(E^*) - \mathbb{Q}(E^*)| \quad (12.7.56)$$

Now define two measures  $\mathbb{P}^*$  and  $\mathbb{Q}^*$  over the sample space  $\{0, 1\}$  which act as indicators for the event  $E^*$ , so that

$$\mathbb{P}(E^*) = \mathbb{P}^*(1) \quad (12.7.57)$$

$$\mathbb{Q}(E^*) = \mathbb{Q}^*(1) \quad (12.7.58)$$

Then because we have already shown Pinsker's inequality over a binary sample space:

$$\delta(\mathbb{P}, \mathbb{Q})^2 = (\mathbb{P}(E^*) - \mathbb{Q}(E^*))^2 \quad (12.7.59)$$

$$= (\mathbb{P}^*(1) - \mathbb{Q}^*(1))^2 \quad (12.7.60)$$

$$\leq \frac{1}{2} \text{KL}(\mathbb{P}^* \parallel \mathbb{Q}^*) \quad (12.7.61)$$

$$\leq \frac{1}{2} \text{KL}(\mathbb{P} \parallel \mathbb{Q}) \quad (12.7.62)$$

where the last inequality comes from the information processing inequality, since the indicator  $\mathbb{I}_{E^*}$  can be determined by (i.e. is a function of) realisations from  $\mathbb{P}$  and  $\mathbb{Q}$ .  $\square$

## 12.7.2 Hellinger Distance

### 12.7.3 $f$ -Divergence [49]

The  $f$ -divergence is a concept that can be used to generalise other measures of statistical distance such as the Kullback-Leibler divergence and the Hellinger distance. Suppose (for simplicity) that distributions  $P$  and  $Q$  are discrete on support  $\mathcal{X}$ . Let  $f(t)$  be a convex function that is defined for  $t > 0$ , and  $f(1) = 0$ . The  $f$ -divergence of a distribution  $P$  from  $Q$  is defined by

$$D_f(P \parallel Q) = \sum_{x \in \mathcal{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right) \quad (12.7.63)$$

To be precise, we need to take

$$0f(0) = 0 \quad (12.7.64)$$

$$f(0) = \lim_{t \rightarrow 0} f(t) \quad (12.7.65)$$

$$0f\left(\frac{a}{0}\right) = \lim_{t \rightarrow 0} tf\left(\frac{a}{t}\right) \quad (12.7.66)$$

Rewriting the  $f$ -divergence as

$$D_f(P\|Q) = \mathbb{E}_Q \left[ f \left( \frac{P(X)}{Q(X)} \right) \right] \quad (12.7.67)$$

we can see that the  $f$ -divergence may be intuitively thought of as the average of the function  $f(\cdot)$  of the odds ratio between  $P$  and  $Q$ . Note that if we set  $f(t) = t \log t$ , then we recover the Kullback-Leibler divergence:

$$D_{t \log t}(P\|Q) = \sum_{x \in \mathcal{X}} Q(x) \frac{P(x)}{Q(x)} \log \left( \frac{P(x)}{Q(x)} \right) \quad (12.7.68)$$

$$= \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right) \quad (12.7.69)$$

#### 12.7.4 Wasserstein Distance

Let  $(\Omega, \delta)$  be a metric space (i.e.  $\Omega$  is some set and the metric  $\delta$  defines some notion of distance between two points on that set). For two probability measures  $\mathbb{P}$  and  $\mathbb{Q}$ , the  $p$ -Wasserstein distance between them is denoted

$$W_p(\mathbb{P}, \mathbb{Q}) = \left( \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\Omega \times \Omega} \delta(x, y)^p d\gamma \right)^{1/p} \quad (12.7.70)$$

where  $\Gamma(\mathbb{P}, \mathbb{Q})$  denotes the set of all joint probability measures with marginals  $\mathbb{P}$  and  $\mathbb{Q}$  respectively. To provide an alternative definition, suppose  $X$  and  $Y$  are random objects on support as a subset of  $\Omega$ , with marginal distributions  $\mathbb{P}$  and  $\mathbb{Q}$  respectively. Then the  $p$ -Wasserstein distance is equivalently

$$W_p(\mathbb{P}, \mathbb{Q}) = \left( \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{\gamma} [\delta(X, Y)^p] \right)^{1/p} \quad (12.7.71)$$

#### Optimal Transportation

The Wasserstein distance can be used in the context of the optimal transportation problem. Suppose there is a normalised mass (i.e. of mass one) with distribution  $\mathbb{P}$  on  $\Omega$ . We wish to move (i.e. transport) the mass around in  $\Omega$  in such a way that the resulting mass distribution looks like  $\mathbb{Q}$ . A *transport plan* is a function  $\gamma(x, y)$  that tells us that we should transport  $\gamma(x, y)$  ‘amount’ of mass from point  $x$  to point  $y$ . In fact,  $\gamma(x, y)$  will be a valid joint distribution of  $\mathbb{P}$  and  $\mathbb{Q}$ . To see this, we setup some ‘mass conservation’ equations. Firstly, the total amount of mass transported from point  $x$  should be equal to the amount of mass that was there to begin with:

$$\int_{\Omega} \gamma(x, y) dy = \mathbb{P}(x) \quad (12.7.72)$$

Secondly, the total amount of mass transported to point  $y$  should be the total summed or integrated over the transport plan:

$$\int_{\Omega} \gamma(x, y) dx = \mathbb{Q}(y) \quad (12.7.73)$$

Hence we see that  $\gamma(x, y)$  satisfies the equations for obtaining the marginal distributions from the joint distribution (marginalisation). Suppose there is a cost  $c(x, y) \geq 0$  from transporting a unit of mass from point  $x$  to point  $y$ . For a given transport plan  $\gamma(x, y)$ , the total transportation cost  $C$  incurred is then

$$C = \int_{\Omega} \int_{\Omega} c(x, y) \gamma(x, y) dx dy \quad (12.7.74)$$

where  $\gamma(x, y) dxdy$  is the infinitesimal amount of mass transported from  $x$  to  $y$ , or  $d\gamma(x, y)$ . Hence we write

$$C = \int_{\Omega \times \Omega} c(x, y) d\gamma(x, y) \quad (12.7.75)$$

The optimal transportation problem is to then find an optimal transport plan which minimises the total transportation cost:

$$C^* = \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\Omega \times \Omega} c(x, y) d\gamma(x, y) \quad (12.7.76)$$

where this search occurs over all possible transport plans, which is the same as all possible joint distributions between  $\mathbb{P}$  and  $\mathbb{Q}$ . Suppose  $c(x, y)$  is chosen to reflect the ‘distance’  $\delta(x, y)$  between  $x$  and  $y$ . Then the optimal transportation cost  $C^*$  is identical to the 1-Wasserstein distance  $W_1(\mathbb{P}, \mathbb{Q})$  between  $\mathbb{P}$  and  $\mathbb{Q}$ . Defining alternative appropriate specifications of the total transportation cost will recover the other Wasserstein distances.

### Kantorovich-Rubinstein Duality

The 1-Wasserstein distance with metric  $\delta(x, y) = \|x - y\|$  has a dual form given by

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f: |f(x) + f(y)| \leq \|x - y\|} \left\{ \left| \int_{\Omega} f(x) d\mathbb{P}(x) - \int_{\Omega} f(y) d\mathbb{Q}(y) \right| \right\} \quad (12.7.77)$$

That is, the supremum over all functions  $f$  such that  $f$  is 1-Lipschitz continuous. To show this duality, we begin with

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\Omega \times \Omega} \|x - y\| d\gamma(x, y) \quad (12.7.78)$$

$$\begin{aligned} &= \inf_{\gamma} \left\{ \int_{\Omega \times \Omega} \|x - y\| d\gamma(x, y) \right. \\ &\quad \left. + \sup_{f, g} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) - \int_{\Omega \times \Omega} (f(x) + g(y)) d\gamma(x, y) \right\} \right\} \end{aligned} \quad (12.7.79)$$

where we have relaxed the constraint  $\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})$  in the outer infimum and instead encoded it in the inner supremum (which are over functions  $f$  and  $g$ ). This is because if we want  $\gamma$  to have  $\mathbb{P}$  and  $\mathbb{Q}$  as marginals, then we should be able to write

$$\int_{\Omega \times \Omega} f(x) d\gamma(x, y) = \mathbb{E}_{\gamma}[f(X)] \quad (12.7.80)$$

$$= \int_{\Omega} f(x) d\mathbb{P}(x) \quad (12.7.81)$$

$$= \mathbb{E}_{\mathbb{P}}[f(X)] \quad (12.7.82)$$

and

$$\int_{\Omega \times \Omega} g(y) d\gamma(x, y) = \mathbb{E}_{\gamma}[g(Y)] \quad (12.7.83)$$

$$= \int_{\Omega} g(y) d\mathbb{Q}(x) \quad (12.7.84)$$

$$= \mathbb{E}_{\mathbb{Q}}[g(Y)] \quad (12.7.85)$$

so that if  $\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})$ , then

$$\int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) - \int_{\Omega \times \Omega} (f(x) + g(y)) d\gamma(x, y) = 0 \quad (12.7.86)$$

holds for any  $f, g$ , hence the supremum is zero. Otherwise, if  $\gamma \notin \Gamma(\mathbb{P}, \mathbb{Q})$ , we could make the supremum large via some appropriate choice of  $f, g$ . Thus, satisfying the infimum will require  $\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})$ . Re-organising the terms, we have

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma} \left\{ \sup_{f,g} \left\{ \int_{\Omega \times \Omega} (\|x - y\| - f(x) - g(y)) d\gamma(x, y) + \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) \right\} \right\} \quad (12.7.87)$$

$$= \sup_{f,g} \left\{ \inf_{\gamma} \left\{ \int_{\Omega \times \Omega} (\|x - y\| - f(x) - g(y)) d\gamma(x, y) \right\} + \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) \right\} \quad (12.7.88)$$

where we can exchange the order of the infimum and supremum because of the **minimax principle**. Another way to formulate this is via constrained optimisation and the Lagrangian dual. Suppose  $X \sim \mathbb{P}$  and  $Y \sim \mathbb{Q}$ , with densities  $p(x)$  and  $q(y)$  respectively. Then the Wasserstein distance is the solution to the optimisation problem

$$\begin{aligned} & \inf_{\pi \geq 0} \int_{\Omega \times \Omega} \|x - y\| \pi(x, y) dx dy \\ & \text{s.t.} \quad \int_{\Omega} \pi(x, y) dy = p(x), \quad x \in \Omega \\ & \quad \int_{\Omega} \pi(x, y) dx = q(y), \quad y \in \Omega \end{aligned} \quad (12.7.89)$$

which is with respect to the joint density  $\pi(x, y)$ . If the sample space  $\Omega$  is infinite, then this becomes an infinite-dimensional optimisation problem with an infinite number of constraints. In that case however, we can still use infinite-dimensional Lagrange multipliers (i.e. functions), which we will introduce  $f, g$  for, and write the Lagrangian as

$$\begin{aligned} \mathcal{L}(\pi, f, g) &= \int_{\Omega \times \Omega} \|x - y\| \pi(x, y) dx dy \\ &- \int_{\Omega} \left( \int_{\Omega} \pi(x, y) dy - p(x) \right) f(x) dx - \int_{\Omega} \left( \int_{\Omega} \pi(x, y) dx - q(y) \right) g(y) dy \end{aligned} \quad (12.7.90)$$

The Lagrangian dual function is then

$$\mathcal{L}^*(f, g) = \inf_{\pi \geq 0} \mathcal{L}(\pi, f, g) \quad (12.7.91)$$

Recognise that the objective to optimise is actually a linear functional (i.e. analogous to a linear function) in  $\pi$ , and moreover the equality constraints are linear in  $\pi$ . Hence we can treat this problem as a linear program, and invoke strong duality of linear programs to write the solution as

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f,g} \mathcal{L}^*(f, g) \quad (12.7.92)$$

$$= \sup_{f,g} \left\{ \inf_{\pi \geq 0} \left\{ \int_{\Omega \times \Omega} (\|x - y\| - f(x) - g(y)) \pi(x, y) dx dy \right\} + \int_{\Omega} f(x) p(x) dx + \int_{\Omega} g(y) q(y) dy \right\} \quad (12.7.93)$$

This arrives at the same formulation as via the minimax principle above. Now consider  $f, g$  such that  $\|x - y\| \geq f(x) + g(y)$  for all  $x, y$ . Then the infimum would be

$$\inf_{\pi \geq 0} \left\{ \int_{\Omega \times \Omega} (\|x - y\| - f(x) - g(y)) \pi(x, y) dx dy \right\} = 0 \quad (12.7.94)$$

which is achieved by  $\pi(x, y) = 0$ . Conversely, if there exists some  $x, y$  such that  $\|x - y\| < f(x) + g(y)$ , then we would have

$$\inf_{\pi \geq 0} \left\{ \int_{\Omega \times \Omega} (\|x - y\| - f(x) - g(y)) \pi(x, y) dx dy \right\} < 0 \quad (12.7.95)$$

which is achieved by concentrating  $\pi$  at this particular  $x, y$ . So in order to satisfy the supremum, we must have  $\|x - y\| \geq f(x) + g(y)$ , and we can write the Wasserstein distance in the simplified form

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f, g: f(x) + g(y) \leq \|x - y\|} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) \right\} \quad (12.7.96)$$

Now consider optimising with respect to  $g$ , given  $f$ . It is clear that  $g$  must satisfy

$$g(y) \leq \inf_x \{\|x - y\| - f(x)\} \quad (12.7.97)$$

otherwise we would violate  $\|x - y\| \geq f(x) + g(y)$ . But we also see that increasing  $g$  also increases the term inside the supremum, so we should take

$$g^*(y) := \inf_x \{\|x - y\| - f(x)\} \quad (12.7.98)$$

Therefore this dual form of the Wasserstein distance becomes an optimisation with respect to a single function  $f$ :

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_f \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g^*(y) d\mathbb{Q}(y) \right\} \quad (12.7.99)$$

$$= \sup_f \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} \inf_x \{\|x - y\| - f(x)\} d\mathbb{Q}(y) \right\} \quad (12.7.100)$$

We show that this is equivalent to a problem where we only consider  $f$  over a space of 1-Lipschitz continuous functions. Naturally, we have

$$W_1(\mathbb{P}, \mathbb{Q}) \geq \sup_{f: |f(x) + f(y)| \leq \|x - y\|} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g^*(y) d\mathbb{Q}(y) \right\} \quad (12.7.101)$$

because we are restricting the space of functions which the supremum is with respect to. If  $f$  is 1-Lipschitz, we can further simplify the right-hand side. Note that

$$g^*(y) = \inf_x \{\|x - y\| - f(x)\} \quad (12.7.102)$$

$$= -f(y) \quad (12.7.103)$$

because intuitively, 1-Lipschitz continuity condition makes the term  $\|x - y\|$  ‘more important’ to minimise than the  $-f(x)$ , thus the infimum is attained at  $x = y$ . Substituting  $g^*$  for  $-f$ :

$$W_1(\mathbb{P}, \mathbb{Q}) \geq \sup_{f: |f(x) + f(y)| \leq \|x - y\|} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) - \int_{\Omega} f(y) d\mathbb{Q}(y) \right\} \quad (12.7.104)$$

On the other hand, we show the inequality also holds in the other direction. First introduce the function

$$h(x) := \inf_y \{\|x - y\| - g(y)\} \quad (12.7.105)$$

We claim that  $h$  is 1-Lipschitz. Through the definition of  $h$ , we have for any  $u \in \Omega$ :

$$h(x) \leq \|x - u\| - g(u) \quad (12.7.106)$$

$$\leq \|x - y\| + \|y - u\| - g(u) \quad (12.7.107)$$

but since this is valid for any  $u$ , we then have

$$h(x) \leq \inf_u \{\|x - y\| + \|y - u\| - g(u)\} \quad (12.7.108)$$

$$= \|x - y\| + \inf_u \{\|y - u\| - g(u)\} \quad (12.7.109)$$

$$= \|x - y\| + h(y) \quad (12.7.110)$$

which means  $h(x) - h(y) \leq \|x - y\|$ . Simply by swapping the roles of  $x$  and  $y$ , we similarly obtain  $h(y) - h(x) \leq \|y - x\|$  and thus

$$|h(x) - h(y)| \leq \|x - y\| \quad (12.7.111)$$

which confirms that  $h$  is 1-Lipschitz. Returning to the requirement  $f(x) \leq \|x - y\| - g(y)$ , this means that for any  $y$ ,

$$f(x) \leq \inf_y \{\|x - y\| - g(y)\} \quad (12.7.112)$$

$$= h(x) \quad (12.7.113)$$

Also by letting  $u = x$  in the inequality  $h(x) \leq \|x - u\| - g(u)$ , we have

$$h(x) \leq \|x - x\| - g(x) \quad (12.7.114)$$

$$= -g(x) \quad (12.7.115)$$

Putting these two inequalities together, we get

$$\int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) \leq \int_{\Omega} h(x) d\mathbb{P}(x) - \int_{\Omega} h(y) d\mathbb{Q}(y) \quad (12.7.116)$$

Hence

$$\sup_{f,g: f(x)+g(y) \leq \|x-y\|} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) \right\} \leq \int_{\Omega} h(x) d\mathbb{P}(x) - \int_{\Omega} h(y) d\mathbb{Q}(y) \quad (12.7.117)$$

Since  $h$  is 1-Lipshitz, if we instead consider all functions that are 1-Lipschitz, then

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f,g: f(x)+g(y) \leq \|x-y\|} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) + \int_{\Omega} g(y) d\mathbb{Q}(y) \right\} \quad (12.7.118)$$

$$\leq \sup_{f: |f(x)+f(y)| \leq \|x-y\|} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) - \int_{\Omega} f(y) d\mathbb{Q}(y) \right\} \quad (12.7.119)$$

As we have established this inequality in both directions, it follows that

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f: |f(x)+f(y)| \leq \|x-y\|} \left\{ \int_{\Omega} f(x) d\mathbb{P}(x) - \int_{\Omega} f(y) d\mathbb{Q}(y) \right\} \quad (12.7.120)$$

Then lastly, as  $f$  being 1-Lipschitz implies that  $-f$  is also 1-Lipschitz, we can equivalently write

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f: |f(x)+f(y)| \leq \|x-y\|} \left\{ \left| \int_{\Omega} f(x) d\mathbb{P}(x) - \int_{\Omega} f(y) d\mathbb{Q}(y) \right| \right\} \quad (12.7.121)$$

### Wasserstein Distance and Kolmogorov-Smirnov Distance

A relationship can be established between the **Kolmogorov-Smirnov distance** and the 1-Wasserstein distance with respect to the metric  $\delta(x, y) = |x - y|$ .

**Theorem 12.6.** *Let  $X$  and  $Y$  be univariate random variables with probability measures  $\mathbb{P}$  and  $\mathbb{Q}$  respectively on  $\mathbb{R}$ . Suppose the density of  $Y$ , denoted  $p_Y(y)$ , is upper-bounded by a constant  $C$  for all  $y \in \mathbb{R}$ . Then*

$$\text{KS}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{2C W_1(\mathbb{P}, \mathbb{Q})} \quad (12.7.122)$$

where

$$\text{KS}(\mathbb{P}, \mathbb{Q}) = \sup_{t \in \mathbb{R}} |\Pr(X \leq t) - \Pr(Y \leq t)| \quad (12.7.123)$$

is the *Kolmogorov-Smirnov distance*.

*Proof.* Introduce the function

$$g_1(x) = \begin{cases} 1, & x \leq t \\ 1 - \frac{x-t}{\varepsilon}, & t < x \leq t + \varepsilon \\ 0, & x > t + \varepsilon \end{cases} \quad (12.7.124)$$

That is, the function which is one for  $x \leq t$ , zero for  $x \geq t + \varepsilon$ , and a linear interpolation in-between. Also introduce the function

$$g_2(x) = \begin{cases} 1, & x \leq t - \varepsilon \\ 1 - \frac{x-t+\varepsilon}{\varepsilon}, & t - \varepsilon < x \leq t \\ 0, & x > t \end{cases} \quad (12.7.125)$$

which is one for  $x \leq t - \varepsilon$ , zero for  $x \geq t$ , and a linear interpolation in-between. Thus, these two functions ‘sandwich’ the indicator  $\mathbb{I}_{\{x \leq t\}}$  by:

$$g_2(x) \leq \mathbb{I}_{\{x \leq t\}} \leq g_1(x) \quad (12.7.126)$$

Using this property, we have

$$\Pr(X \leq t) - \Pr(Y \leq t) = \mathbb{E}[\mathbb{I}_{\{X \leq t\}}] - \Pr(Y \leq t) \quad (12.7.127)$$

$$\leq \mathbb{E}[g_1(X)] - \Pr(Y \leq t) \quad (12.7.128)$$

$$= \mathbb{E}[g_1(X)] - \mathbb{E}[g_1(Y)] + \mathbb{E}[g_1(Y)] - \Pr(Y \leq t) \quad (12.7.129)$$

We seek to bound each pair of terms. Recall the **Kantorovich-Rubinstein dual representation** of the Wasserstein distance is

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f: |f(x) - f(y)| \leq |x - y|} \{|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]|\} \quad (12.7.130)$$

Note that  $g_1(x)$  is  $\frac{1}{\varepsilon}$ -Lipschitz, so  $\varepsilon g_1(x)$  is 1-Lipschitz hence

$$\mathbb{E}[\varepsilon g_1(X)] - \mathbb{E}[\varepsilon g_1(Y)] \leq W_1(\mathbb{P}, \mathbb{Q}) \quad (12.7.131)$$

or

$$\mathbb{E}[g_1(X)] - \mathbb{E}[g_1(Y)] \leq \frac{1}{\varepsilon} W_1(\mathbb{P}, \mathbb{Q}) \quad (12.7.132)$$

For the other pair of terms, we have

$$\mathbb{E}[g_1(Y)] - \Pr(Y \leq t) = \mathbb{E}[g_1(Y) - \mathbb{I}_{\{Y \leq t\}}] \quad (12.7.133)$$

$$= \int_{-\infty}^{\infty} p_Y(y) (g_1(y) - \mathbb{I}_{\{y \leq t\}}) dy \quad (12.7.134)$$

$$= \int_t^{t+\varepsilon} p_Y(y) \left(1 - \frac{x-t}{\varepsilon}\right) dy \quad (12.7.135)$$

$$\leq \int_t^{t+\varepsilon} C \left(1 - \frac{x-t}{\varepsilon}\right) dy \quad (12.7.136)$$

$$= \frac{C\varepsilon}{2} \quad (12.7.137)$$

where the last integral is effectively the area of a triangle with base  $\varepsilon$  and height  $C$ . Applying these upper bounds, we have for any  $t \in \mathbb{R}$ :

$$\Pr(X \leq t) - \Pr(Y \leq t) \leq \frac{1}{\varepsilon} W_1(\mathbb{P}, \mathbb{Q}) + \frac{C\varepsilon}{2} \quad (12.7.138)$$

Now we can similarly bound

$$\Pr(Y \leq t) - \Pr(X \leq t) = \Pr(Y \leq t) - \mathbb{E}[\mathbb{I}_{\{X \leq t\}}] \quad (12.7.139)$$

$$\leq \Pr(Y \leq t) - \mathbb{E}[g_2(X)] \quad (12.7.140)$$

$$= \Pr(Y \leq t) - \mathbb{E}[g_2(Y)] + \mathbb{E}[g_2(Y)] - \mathbb{E}[g_2(X)] \quad (12.7.141)$$

and following analogous arguments, we obtain

$$\mathbb{E}[g_2(Y)] - \mathbb{E}[g_2(X)] \leq \frac{1}{\varepsilon} W_1(\mathbb{P}, \mathbb{Q}) \quad (12.7.142)$$

$$\Pr(Y \leq t) - \mathbb{E}[g_2(Y)] \leq \frac{C\varepsilon}{2} \quad (12.7.143)$$

which gives

$$\Pr(Y \leq t) - \Pr(X \leq t) \leq \frac{1}{\varepsilon} W_1(\mathbb{P}, \mathbb{Q}) + \frac{C\varepsilon}{2} \quad (12.7.144)$$

for any  $t \in \mathbb{R}$ . Therefore the Kolmogorov-Smirnov distance is also bounded by

$$\sup_{t \in \mathbb{R}} |\Pr(X \leq t) - \Pr(Y \leq t)| \leq \frac{1}{\varepsilon} W_1(\mathbb{P}, \mathbb{Q}) + \frac{C\varepsilon}{2} \quad (12.7.145)$$

This bound can be optimised with respect to  $\varepsilon$ . Taking the derivative:

$$\frac{d}{d\varepsilon} \left( \frac{1}{\varepsilon} W_1(\mathbb{P}, \mathbb{Q}) + \frac{C\varepsilon}{2} \right) = -\frac{W_1(\mathbb{P}, \mathbb{Q})}{\varepsilon^2} + \frac{C}{2} \quad (12.7.146)$$

and equating to zero yields

$$\frac{W_1(\mathbb{P}, \mathbb{Q})}{\varepsilon^2} = \frac{C}{2} \quad (12.7.147)$$

$$\varepsilon = \sqrt{\frac{2W_1(\mathbb{P}, \mathbb{Q})}{C}} \quad (12.7.148)$$

Substituting this value of  $\varepsilon$  leads to the claimed bound:

$$\sup_{t \in \mathbb{R}} |\Pr(X \leq t) - \Pr(Y \leq t)| \leq \sqrt{\frac{C}{2W_1(\mathbb{P}, \mathbb{Q})}} W_1(\mathbb{P}, \mathbb{Q}) + \frac{C}{2} \sqrt{\frac{2W_1(\mathbb{P}, \mathbb{Q})}{C}} \quad (12.7.149)$$

$$= \sqrt{\frac{CW_1(\mathbb{P}, \mathbb{Q})}{2}} + \sqrt{\frac{CW_1(\mathbb{P}, \mathbb{Q})}{2}} \quad (12.7.150)$$

$$= \sqrt{2CW_1(\mathbb{P}, \mathbb{Q})} \quad (12.7.151)$$

□

## Wasserstein Distance and Total Variation Distance

A connection between the Wasserstein distance and total variation distance  $\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}}$  can be established. Firstly, we require the following characterisation of the total variation distance.

**Theorem 12.7.** *Let  $X$  and  $Y$  be random variables on the same sample space  $\Omega$ , corresponding to probability measures  $\mathbb{P}$  and  $\mathbb{Q}$  respectively. Then*

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \Pr(X \neq Y) \quad (12.7.152)$$

where the infimum is taken over all joint distributions for  $(X, Y)$  with marginals given by  $\mathbb{P}$  and  $\mathbb{Q}$  respectively.

*Proof.* For any event  $E \in \mathcal{F}$  and any joint distribution  $\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})$ ,

$$\mathbb{P}(E) - \mathbb{Q}(E) = \Pr(X \in E) - \Pr(Y \in E) \quad (12.7.153)$$

$$= \Pr(X \in E) - (1 - \Pr(Y \notin E)) \quad (12.7.154)$$

$$= \Pr(X \in E) + \Pr(Y \notin E) - 1 \quad (12.7.155)$$

and since

$$1 \geq \Pr(X \in E \cup Y \notin E) \quad (12.7.156)$$

$$= \Pr(X \in E) + \Pr(Y \in E) - \Pr(X \in E, Y \notin E) \quad (12.7.157)$$

then

$$\mathbb{P}(E) - \mathbb{Q}(E) = \Pr(X \in E) + \Pr(Y \notin E) - 1 \quad (12.7.158)$$

$$\leq \Pr(X \in E) + \Pr(Y \notin E) - \Pr(X \in E) - \Pr(Y \in E) + \Pr(X \in E, Y \notin E) \quad (12.7.159)$$

$$= \Pr(X \in E, Y \notin E) \quad (12.7.160)$$

$$\leq \Pr(X \neq Y) \quad (12.7.161)$$

where the last inequality follows because

$$\Pr(X \neq Y) = \bigcup_{E \in \mathcal{F}} \Pr(X \in E, Y \notin E) \quad (12.7.162)$$

Now since  $E$  can be any event and  $\gamma$  can be any joint distribution, it follows that

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = \sup_{E \in \mathcal{F}} |\mathbb{P}(E) - \mathbb{Q}(E)| \quad (12.7.163)$$

$$\leq \Pr(X \neq Y) \quad (12.7.164)$$

for all joint distributions  $\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})$ . All that remains is to show that there exist a joint distribution that attains this inequality with equality, which we do so by construction. Introduce the event  $B$  defined by

$$B := \{\omega \in \Omega : \mathbb{P}(\omega) \geq \mathbb{Q}(\omega)\} \quad (12.7.165)$$

in the case where  $\Omega$  is countable. In the case where  $\Omega$  is uncountable, we can introduce an analogously defined event pertaining to the partition that achieves the supremum in the variation distance. Here, we proceed assuming that  $\Omega$  is countable; the steps are analogous otherwise. Let

$$q := \sum_{\omega \in \Omega} \min \{\mathbb{P}(\omega), \mathbb{Q}(\omega)\} \quad (12.7.166)$$

$$= \mathbb{P}(\overline{B}) + \mathbb{Q}(B) \quad (12.7.167)$$

$$= 1 - \mathbb{P}(B) + \mathbb{Q}(B) \quad (12.7.168)$$

$$= 1 - (\mathbb{P}(B) - \mathbb{Q}(B)) \quad (12.7.169)$$

$$= 1 - \|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \quad (12.7.170)$$

where we recall the alternative characterisation of the total variation distance in the last equality. Hence

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = 1 - q \quad (12.7.171)$$

Define the following functions:

$$c_1(\omega) := \frac{\mathbb{P}(\omega) - \mathbb{Q}(\omega)}{\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}}} \mathbb{I}_{\{\omega \in B\}} \quad (12.7.172)$$

$$c_2(\omega) := \frac{\mathbb{Q}(\omega) - \mathbb{P}(\omega)}{\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}}} \mathbb{I}_{\{\omega \notin B\}} \quad (12.7.173)$$

$$c_3(\omega) := \frac{\min\{\mathbb{P}(\omega), \mathbb{Q}(\omega)\}}{q} \quad (12.7.174)$$

Note that  $c_1(\omega) \geq 0$ ,  $c_2(\omega) \geq 0$  and  $c_3(\omega) \geq 0$ . We also can show that

$$\sum_{\omega \in B} c_1(\omega) = \frac{\mathbb{P}(B) - \mathbb{Q}(B)}{\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}}} \quad (12.7.175)$$

$$= 1 \quad (12.7.176)$$

and  $\sum_{\omega \in \Omega} c_1(\omega) = 1$ . Furthermore,

$$\sum_{\omega \notin \Omega} c_2(\omega) = \frac{\mathbb{Q}(\overline{B}) - \mathbb{P}(\overline{B})}{\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}}} \quad (12.7.177)$$

$$= \frac{\mathbb{P}(B) - \mathbb{Q}(B)}{\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}}} \quad (12.7.178)$$

$$= 1 \quad (12.7.179)$$

and  $\sum_{\omega \in \Omega} c_2(\omega) = 1$ . As for  $c_3(\cdot)$ ,

$$\sum_{\omega \in \Omega} c_3(\omega) = \frac{\sum_{\omega \in \Omega} \min\{\mathbb{P}(\omega), \mathbb{Q}(\omega)\}}{q} \quad (12.7.180)$$

$$= \frac{q}{q} \quad (12.7.181)$$

$$= 1 \quad (12.7.182)$$

Hence  $c_1(\cdot)$ ,  $c_2(\cdot)$  and  $c_3(\cdot)$  are all valid probability distributions. Introduce the binary random variable  $\xi$  such that  $\Pr(\xi = 1) = p$ , and consider the conditional distribution

$$\Pr(X = x, Y = y | \xi) = c_3(x) \mathbb{I}_{\{x=y\}} \xi + (1 - \xi) c_1(x) c_2(y) \quad (12.7.183)$$

Since  $\Pr(\xi = 1) = q$ , the joint distribution over  $(X, Y)$  is given by

$$\Pr(X = x, Y = y) = qc_3(x) \mathbb{I}_{\{x=y\}} + (1 - q) c_1(x) c_2(y) \quad (12.7.184)$$

We compute the marginal distributions. For  $X$ :

$$\Pr(X = x) = \sum_{y \in \Omega} \Pr(X = x, Y = y) \quad (12.7.185)$$

$$= q \sum_{y \in \Omega} c_3(y) \mathbb{I}_{\{x=y\}} + (1-q) c_1(x) \sum_{y \in \Omega} c_2(y) \quad (12.7.186)$$

$$= qc_3(x) + (1-q) c_1(x) \quad (12.7.187)$$

$$= \frac{q \min \{\mathbb{P}(x), \mathbb{Q}(x)\}}{q} + (1-q) \frac{\mathbb{P}(x) - \mathbb{Q}(x)}{1-q} \mathbb{I}_{\{x \in B\}} \quad (12.7.188)$$

$$= \min \{\mathbb{P}(x), \mathbb{Q}(x)\} + (\mathbb{P}(x) - \mathbb{Q}(x)) \mathbb{I}_{\{x \in B\}} \quad (12.7.189)$$

$$= \mathbb{P}(x) \quad (12.7.190)$$

Similarly for  $Y$ :

$$\Pr(Y = y) = \sum_{x \in \Omega} \Pr(X = x, Y = y) \quad (12.7.191)$$

$$= q \sum_{x \in \Omega} c_3(x) \mathbb{I}_{\{x=y\}} + (1-q) c_2(y) \sum_{x \in \Omega} c_1(x) \quad (12.7.192)$$

$$= qc_3(y) + (1-q) c_2(y) \quad (12.7.193)$$

$$= \frac{q \min \{\mathbb{P}(y), \mathbb{Q}(y)\}}{q} + (1-q) \frac{\mathbb{Q}(y) - \mathbb{P}(y)}{1-q} \mathbb{I}_{\{y \notin B\}} \quad (12.7.194)$$

$$= \min \{\mathbb{P}(y), \mathbb{Q}(y)\} + (\mathbb{Q}(y) - \mathbb{P}(y)) \mathbb{I}_{\{y \notin B\}} \quad (12.7.195)$$

$$= \mathbb{Q}(y) \quad (12.7.196)$$

This confirms that our construction is a valid joint distribution with the required marginal distributions. Lastly, we demonstrate that using this distribution achieves the total variation distance:

$$\Pr(X \neq Y) = \sum_{x \in \Omega} \sum_{y \in \Omega} \mathbb{I}_{\{x \neq y\}} \Pr(X = x, Y = y) \quad (12.7.197)$$

$$= \sum_{x \in \Omega} \sum_{y \in \Omega} qc_3(x) \underbrace{\mathbb{I}_{\{x=y\}} \mathbb{I}_{\{x \neq y\}}}_{0} + \sum_{x \in \Omega} \sum_{y \in \Omega} (1-q) c_1(x) c_2(y) \mathbb{I}_{\{x \neq y\}} \quad (12.7.198)$$

$$= (1-q) \sum_{x \in \Omega} c_1(x) \sum_{y \in \Omega} c_2(y) \mathbb{I}_{\{x \neq y\}} \quad (12.7.199)$$

$$= (1-q) \sum_{x \in \Omega} c_1(x) (1 - c_2(x)) \quad (12.7.200)$$

$$= (1-q) \left( \sum_{x \in \Omega} c_1(x) - \sum_{x \in \Omega} \frac{(\mathbb{P}(x) - \mathbb{Q}(x)) (\mathbb{Q}(x) - \mathbb{P}(x))}{(1-q)^2} \mathbb{I}_{\{x \in B\}} \mathbb{I}_{\{x \notin B\}} \right) \underbrace{0}_{0} \quad (12.7.201)$$

$$= (1-q) (1 - 0) \quad (12.7.202)$$

$$= 1 - q \quad (12.7.203)$$

$$= \|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \quad (12.7.204)$$

□

If  $X$  and  $Y$  are identically distributed, then clearly  $\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = 0$  and the joint distribution achieving the infimum is **comonotonic**. If  $X$  and  $Y$  are not identically distributed, then this definition of the total variation distance characterises how close a *coupling* between  $X$  and  $Y$  can get to being comonotonic.

If we were to write  $\Pr_\gamma(X \neq Y)$  as an expectation of an indicator, then

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\gamma [\mathbb{I}_{\{X \neq Y\}}] \quad (12.7.205)$$

If we now consider  $\delta(x, y) = \mathbb{I}_{\{x \neq y\}}$  as a metric (transportation cost), we see that

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{\gamma} [\delta(x, y)] \quad (12.7.206)$$

$$= W_1(\mathbb{P}, \mathbb{Q}) \quad (12.7.207)$$

That is, the total variation distance is equivalent to the 1-Wasserstein distance with transportation cost  $\mathbb{I}_{\{x \neq y\}}$ .

### Empirical Wasserstein Distance [28]

Let  $X_1, \dots, X_n$  be a univariate sample from one distribution, and let  $Y_1, \dots, Y_n$  be a univariate sample from another (possibly same) distribution. We use  $\widehat{\mathbb{P}}$  and  $\widehat{\mathbb{Q}}$  respectively to denote the empirical distributions, and we use the notation  $X_{(1)} \leq \dots \leq X_{(n)}$  to denote the order statistics. Then for  $p \geq 1$ , the  $p$ -Wasserstein distance between the empirical distributions for the metric  $\delta(x, y) = |x - y|$  is

$$W_p(\widehat{\mathbb{P}}, \widehat{\mathbb{Q}}) = \left( \frac{1}{n} \sum_{i=1}^n |X_{(i)} - Y_{(i)}|^p \right)^{1/p} \quad (12.7.208)$$

*Proof.* For simplicity, consider  $p = 1$ ; the reasoning is the same for  $p > 1$ . We first claim that for any permutation  $\pi$  of the integers  $\{1, \dots, n\}$ :

$$\inf_{\pi} \sum_{i=1}^n |X_{(i)} - Y_{(\pi(i))}| = \sum_{i=1}^n |X_{(i)} - Y_{(i)}| \quad (12.7.209)$$

That is, pairing the order statistics minimises the sum of absolute distances between pairings, compared to any other pairing. To show this, it suffices to show that if we begin with the optimal permutation and ‘swap’ the pairings of any two indices (say  $j$  and  $k$ ), then this will never reduce the sum of absolute distances. Upon making this swap, the sum increases by  $|X_{(j)} - Y_{(k)}| + |X_{(k)} - Y_{(j)}|$  and reduces by  $|X_{(j)} - Y_{(j)}| + |X_{(k)} - Y_{(k)}|$ , so we aim to show that this net change is non-negative:

$$|X_{(j)} - Y_{(k)}| + |X_{(k)} - Y_{(j)}| - |X_{(j)} - Y_{(j)}| - |X_{(k)} - Y_{(k)}| \geq 0 \quad (12.7.210)$$

To this end, we assume without loss of generality that  $j = 1$  and  $k = 2$  (because only relative distances matter) and that  $X_{(1)} \leq Y_{(1)}$  (because the identity of each sample is arbitrary). Then consider the possible orderings between the variables  $X_{(1)}, X_{(2)}, Y_{(1)}, Y_{(2)}$ :

- In the case that  $X_{(1)} \leq X_{(2)} \leq Y_{(1)} \leq Y_{(2)}$ , we have

$$|X_{(1)} - Y_{(1)}| + |X_{(2)} - Y_{(2)}| = (Y_{(1)} - X_{(1)}) + (Y_{(2)} - X_{(2)}) \quad (12.7.211)$$

$$= (Y_{(1)} - X_{(2)}) + (Y_{(2)} - X_{(1)}) \quad (12.7.212)$$

$$= |X_{(1)} - Y_{(2)}| + |X_{(2)} - Y_{(1)}| \quad (12.7.213)$$

and so the inequality is satisfied with equality.

- In the case that  $X_{(1)} \leq Y_{(1)} \leq X_{(2)} \leq Y_{(2)}$ , we have

$$|X_{(1)} - Y_{(1)}| + |X_{(2)} - Y_{(2)}| = (Y_{(1)} - X_{(1)}) + (Y_{(2)} - X_{(2)}) \quad (12.7.214)$$

$$= (Y_{(1)} - X_{(2)}) + (Y_{(2)} - X_{(1)}) \quad (12.7.215)$$

$$\leq (X_{(2)} - Y_{(1)}) + (Y_{(2)} - X_{(1)}) \quad (12.7.216)$$

$$= |X_{(1)} - Y_{(2)}| + |X_{(2)} - Y_{(1)}| \quad (12.7.217)$$

since  $Y_{(1)} - X_{(2)}$  is non-positive.

- In the case that  $X_{(1)} \leq Y_{(1)} \leq Y_{(2)} \leq X_{(2)}$ , we have

$$|X_{(1)} - Y_{(1)}| + |X_{(2)} - Y_{(2)}| = (Y_{(1)} - X_{(1)}) + (X_{(2)} - Y_{(2)}) \quad (12.7.218)$$

$$= (Y_{(2)} - X_{(1)}) + (X_{(2)} - Y_{(1)}) - 2(Y_{(2)} - Y_{(1)}) \quad (12.7.219)$$

$$\leq (Y_{(2)} - X_{(1)}) + (X_{(2)} - Y_{(1)}) \quad (12.7.220)$$

$$= |X_{(1)} - Y_{(2)}| + |X_{(2)} - Y_{(1)}| \quad (12.7.221)$$

since  $Y_{(2)} - Y_{(1)} \geq 0$ .

The same logic applies for  $p > 1$ . Through the definition of the Wasserstein distance, now consider finding the joint distribution with marginals  $\widehat{\mathbb{P}}$  and  $\widehat{\mathbb{Q}}$  which minimises the total transportation cost. Note that any such distribution with the required marginals can be constructed just by choosing a permutation of the integers  $\{1, \dots, n\}$  (e.g. ‘filling in’  $n$  spots of an  $n \times n$  grid so that every row and column has exactly one spot filled in). Thus if we choose the permutation that matches up the order statistics, this has already been shown to be the minimiser:

$$W_p(\widehat{\mathbb{P}}, \widehat{\mathbb{Q}}) = \left( \inf_{\gamma \in \Gamma(\widehat{\mathbb{P}}, \widehat{\mathbb{Q}})} \mathbb{E}_\gamma [|X - Y|^p] \right)^{1/p} \quad (12.7.222)$$

$$= \left( \inf_{\pi} \frac{1}{n} \sum_{i=1}^n |X_i - Y_{\pi(i)}|^p \right)^{1/p} \quad (12.7.223)$$

$$= \left( \frac{1}{n} \sum_{i=1}^n |X_{(i)} - Y_{(i)}|^p \right)^{1/p} \quad (12.7.224)$$

□

## 12.8 Algorithmic Information Theory

### 12.8.1 Kolmogorov Complexity [47]

Abstracting away from the finer details, we can view a *Turing machine* as a mapping from a set of finite-length strings of a finite-length alphabet as input (although, the binary alphabet  $\{0, 1\}$  will be considered for simplicity), to a set of finite-length or possibly infinite-length binary strings. We denote this mapping by the function  $f : \{0, 1\}^* \rightarrow \{0, 1\}^* \cup \{0, 1\}^\infty$  where  $*$  represents strings of arbitrarily finite length. If the output is infinite-length, the computation of the Turing machine is said to ‘not halt’, and furthermore the function is considered to be undefined.

A *universal computer* or *universal Turing machine*  $\mathcal{U}$  is a Turing machine that can ‘mimic’ the behaviour of any other Turing machine  $\mathcal{T}$  [129]. In essence,  $\mathcal{U}$  needs to be able to simulate  $\mathcal{T}$ ’s actions for any given input.

Kolmogorov complexity can be thought of as a precursor to entropy. While entropy is defined for probability distributions, Kolmogorov complexity may be defined for any object which has a finite-length representation. Let  $x$  be a finite-length binary string. The *Kolmogorov complexity* of  $x$  with respect to a universal computer  $\mathcal{U}$  is the length of the shortest input (or program)  $s$  that outputs  $x$ . That is,

$$K_{\mathcal{U}}(x) = \min_{s: \mathcal{U}(s)=x} l(s) \quad (12.8.1)$$

where  $l(s)$  denotes the length of  $s$  and  $\mathcal{U}(s)$  denotes the output of  $\mathcal{U}$  with  $s$  as an input. Kolmogorov complexity of an object relates to its ‘data compressibility’, since intuitively an object which is more compressible should be able to be described with a shorter length program.

### Conditional Kolmogorov Complexity

The conditional Kolmogorov Complexity given the length of  $x$ , denoted  $K_{\mathcal{U}}(x|l(x))$ , is defined as the Kolmogorov complexity when  $\mathcal{U}$  ‘knows’  $l(x)$ . This is different from the standard Kolmogorov complexity. To illustrate, if we consider programs of the form which instruct the computer to simply ‘print’  $x$ , then a computer which does not know the length of  $x$  would need to include the length ‘coded in’ to instruct the computer when to halt, whereas the same type of program on a computer which did know  $l(x)$  would implicitly know when to halt without the length being coded in. In this way, we can state a simple upper bound for the conditional Kolmogorov Complexity:

$$K_{\mathcal{U}}(x|l(x)) \leq l(x) + c_{\mathcal{U}} \quad (12.8.2)$$

where  $c_{\mathcal{U}}$  is a constant independent of  $x$ , but may for example be associated with the ‘overhead’ of running instructions on  $\mathcal{U}$ . Since universal computers can simulate one another, we can drop the  $\mathcal{U}$  subscript for convenience and instead write

$$K(x|l(x)) \leq l(x) + c \quad (12.8.3)$$

for some fixed but unspecified universal computer  $\mathcal{U}$ . An upper bound on the Kolmogorov complexity in terms of the conditional Kolmogorov complexity is

$$K(x) \leq K(x|l(x)) + 2 \log l(x) + c \quad (12.8.4)$$

which characterises the additional length a program might require to code in the length of  $x$ . To derive this bound, we describe a naïve way to code in the length of the binary string  $x$ . We require at most  $\log_2 l(x)$  bits to describe the length, however we also need a ‘termination’ code to let the computer know when we have finished describing the length. One way is to repeat each bit twice (either 00 or 11), and then terminate with the code 01 or 10 to distinguish it from a repeated bit. This way, we require an additional  $2 \log_2 l(x) + 2$  bits to describe the length, and the extra 2 bit termination code can be absorbed into the constant  $c$ .

### Kolmogorov Complexity and Entropy

The entropy of an i.i.d. sequence can be related to its Kolmogorov complexity by the following lower and upper bounds.

**Theorem 12.8.** *Let  $\mathbf{X} = (X_1, \dots, X_n)$  be an i.i.d. sequence of random variable  $X$  from distribution  $p(x)$  with finite support  $\mathcal{X}$ . Denote the joint distribution of  $\mathbf{X}$  by  $p_n(\mathbf{x}) = \prod_{i=1}^n p(x_i)$ . Then for all  $n$ , there exists a constant  $c$  such that*

$$H[X] \leq \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{x}|n) \leq H[X] + \frac{(|\mathcal{X}| - 1) \log n}{n} + \frac{c}{n} \quad (12.8.5)$$

*Proof.* We first derive the lower bound. For each  $\mathbf{x} = (x_1, \dots, x_n)$  of length  $n$ , assign it the shortest-length program  $s$  such that a universal computer  $\mathcal{U}$  outputs  $\mathcal{U}(s) = \mathbf{x}$  given  $n$ . By the definition of the conditional Kolmogorov complexity, the length of this program satisfies

$$l(s) = K(\mathbf{x}|n) \quad (12.8.6)$$

Now, we argue that for the set  $\mathcal{S}$  of all programs that halt, we have

$$\sum_{s \in \mathcal{S}} 2^{-l(s)} \leq 1 \quad (12.8.7)$$

This holds because of the Kraft inequality encountered in coding theory. As  $s$  halts, then there cannot be another halting program that has  $s$  as a prefix. Thus, the set  $\mathcal{S}$  forms a prefix

code, and their lengths are already shown to satisfy the Kraft inequality above. Then from the characterisation of entropy in relation to prefix codes, we know that the entropy lower bounds the expected codeword length, i.e.

$$H[\mathbf{X}] \leq \mathbb{E}_{\mathbf{X}}[l(s)] \quad (12.8.8)$$

Hence

$$n H[X] \leq \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) l(s) \quad (12.8.9)$$

$$H[X] \leq \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{x}|n) \quad (12.8.10)$$

To derive the upper bound, we describe  $\mathbf{x}$  using the [method of types](#), which will form an upper bound on  $K(\mathbf{x}|n)$ . Denote  $P_{\mathbf{x}}$  as the type of the sequence  $\mathbf{x}$  (i.e. essentially the empirical distribution). To represent the type  $P_{\mathbf{x}}$ , we use  $(|\mathcal{X}| - 1) \log_2 n$  bits. This is because we use up to  $\log_2 n$  bits to count the frequency for  $|\mathcal{X}| - 1$  of the symbols, and frequency of the remaining symbol can be uniquely determined since  $n$  is given. Then for a particular  $P_{\mathbf{x}}$ , we still need to index which of the sequences  $\mathbf{x}$  led to type  $P_{\mathbf{x}}$ . Denote  $T(P_{\mathbf{x}})$  as the type class of  $P_{\mathbf{x}}$ . Using the upper bound of the size of the type class:

$$|T(P_{\mathbf{x}})| \leq 2^{n H[P_{\mathbf{x}}]} \quad (12.8.11)$$

This establishes that it takes at most  $\log_2 |T(P_{\mathbf{x}})| \leq n H[P_{\mathbf{x}}]$  bits to index the sequence  $\mathbf{x}$  from the class of type  $P_{\mathbf{x}}$ . Hence the type representation takes no more than  $n H[P_{\mathbf{x}}] + (|\mathcal{X}| - 1) \log_2 n$  bits, and we can write

$$K(\mathbf{x}|n) \leq n H[P_{\mathbf{x}}] + (|\mathcal{X}| - 1) \log_2 n + c \quad (12.8.12)$$

Taking the expectation of both sides for a random sequence  $\mathbf{X}$ :

$$\mathbb{E}[K(\mathbf{X}|n)] \leq n \mathbb{E}[H[P_{\mathbf{X}}]] + (|\mathcal{X}| - 1) \log_2 n + c \quad (12.8.13)$$

where we can evaluate

$$\mathbb{E}[H[P_{\mathbf{X}}]] = \mathbb{E}\left[-\sum_{\alpha \in \mathcal{X}} P_{\mathbf{X}}(\alpha) \log P_{\mathbf{X}}(\alpha)\right] \quad (12.8.14)$$

$$\leq -\sum_{\alpha \in \mathcal{X}} \mathbb{E}[P_{\mathbf{X}}(\alpha)] \log \mathbb{E}[P_{\mathbf{X}}(\alpha)] \quad (12.8.15)$$

$$= -\sum_{\alpha \in \mathcal{X}} p(\alpha) \log p(\alpha) \quad (12.8.16)$$

$$= H[X] \quad (12.8.17)$$

due Jensen's inequality, since  $-y \log y$  is a concave function. Also, we used the property  $\mathbb{E}[P_{\mathbf{X}}(\alpha)] = p(\alpha)$  (which says the expected empirical frequency of symbol  $\alpha$  is equal to the probability of symbol  $\alpha$ ). Hence

$$\sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{X}|n) \leq n H[X] + (|\mathcal{X}| - 1) \log_2 n + c \quad (12.8.18)$$

$$\frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{X}|n) \leq H[X] + \frac{(|\mathcal{X}| - 1) \log_2 n}{n} + \frac{c}{n} \quad (12.8.19)$$

□

**Corollary 12.3.** *An upper and lower bound of the unconditional Kolmogorov complexity satisfy*

$$H[X] \leq \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{x}) \leq H[X] + \frac{(|\mathcal{X}|+1) \log n}{n} + \frac{c}{n} \quad (12.8.20)$$

*Proof.* Obtaining the lower bound follows the same arguments, since the set of all halting programs is still a prefix code and the program lengths  $K(\mathbf{x})$  will satisfy the Kraft inequality (it was not mandatory to condition on  $n$ ). To prove the upper bound, we use the upper bound for the Kolmogorov complexity  $K(\mathbf{x}) \leq K(\mathbf{x}|l(x)) + 2 \log l(\mathbf{x}) + c$  so that

$$\sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{x}) \leq \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) (K(\mathbf{x}|n) + 2 \log n + c) \quad (12.8.21)$$

$$= \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{x}|n) + 2 \log n + c \quad (12.8.22)$$

Combining this with our upper bound for  $\frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{x}|n)$  yields

$$\frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} p_n(\mathbf{x}) K(\mathbf{x}) \leq H[X] + \frac{(|\mathcal{X}|+1) \log n}{n} + \frac{c}{n} \quad (12.8.23)$$

where we have abused notation and absorbed all constants into a single term  $c$ .  $\square$

Taking limits as  $n \rightarrow \infty$ , we see the the upper and lower bounds converge and we get

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[K(\mathbf{X})] = H[X] \quad (12.8.24)$$

or alternatively,

$$\lim_{n \rightarrow \infty} \mathbb{E}[K(\mathbf{X})] = H[\mathbf{X}] \quad (12.8.25)$$

and so the entropy  $H[X]$  can be characterised as the limiting value of  $\frac{1}{n} \mathbb{E}[K(\mathbf{X})]$ , the latter which may be thought of as the average per unit-length ‘compressibility’ achieved by a universal computer on random i.i.d. strings made out of  $X$ .

### 12.8.2 Universal Probability [47]

The universal probability of a binary string  $x$  with respect to a universal computer  $\mathcal{U}$  is defined as

$$P_{\mathcal{U}}(x) = \Pr(\mathcal{U}(s) = x) \quad (12.8.26)$$

over the distribution of programs  $s$  which are an i.i.d. binary sequence of Bernoulli (0.5) random variables, which are drawn until  $x$  is eventually (or fails to be) an output of the computer. For a program  $s$  with length  $l(s)$  such that  $\mathcal{U}(s) = x$ , the probability that the first  $l(s)$  random bits are equal to  $s$  is  $\left(\frac{1}{2}\right)^{l(s)} = 2^{-l(s)}$ . Thus,

$$P_{\mathcal{U}}(x) = \sum_{s: \mathcal{U}(s)=x} 2^{-l(s)} \quad (12.8.27)$$

The universal probability is related to the Kolmogorov complexity by the following bounds.

**Theorem 12.9.** *For any universal computer  $\mathcal{U}$ , there exists a constant  $c$  not dependent of binary string  $x$  such that*

$$K_{\mathcal{U}}(x) - c \leq \log_2 \frac{1}{P_{\mathcal{U}}(x)} \leq K_{\mathcal{U}}(x) \quad (12.8.28)$$

for all  $x$ .

*Proof.* For ease of notation, we suppress dependence on  $\mathcal{U}$ . The upper bound is proven first. Let  $s^*$  be the shortest program for  $x$ . Then

$$P(x) = \sum_{s: \mathcal{U}(s)=x} 2^{-l(s)} \quad (12.8.29)$$

$$\geq 2^{-l(s^*)} \quad (12.8.30)$$

$$= 2^{-K(x)} \quad (12.8.31)$$

by definition of the Kolmogorov complexity. The bound  $P(x) \geq 2^{-K(x)}$  rearranges to  $\log_2 \frac{1}{P(x)} \leq K(x)$ . For the lower bound, we rearrange it to

$$K(x) \leq \log_2 \frac{1}{P(x)} + c \quad (12.8.32)$$

and aim to show the existence of programs for  $x$  with length of approximately  $\log_2 \frac{1}{P(x)}$ . However,  $P(x)$  is noncomputable (related to the halting problem, since we cannot try all programs to see whether they will output  $x$ , as we do not know whether or not they will halt). Instead, we develop a construction which approximates  $P(x)$  and assigns a program of length approximately  $\log_2 \frac{1}{P(x)}$  for each  $x$ . Consider a list of all programs in the order in which they halt, paired with its associated output:

$$\{(s_1, x_1), (s_2, x_2), \dots\} \quad (12.8.33)$$

For the  $k^{\text{th}}$  program, we estimate  $P(x_k)$  by

$$\widehat{P}(x_k) = \sum_{\{i : x_i = x_k, i \leq k\}} 2^{-l(s_i)} \quad (12.8.34)$$

That is, we sum using the lengths of all programs which output  $x_k$ , that have come before  $s_k$ . Note that this is an underapproximation, i.e.  $\widehat{P}(x_k) \leq P(x_k)$ . Then for each  $k$ , define the integer

$$n_k = \left\lceil \log_2 \frac{1}{\widehat{P}(x_k)} \right\rceil \quad (12.8.35)$$

which will later act as the levels of a binary tree. Our procedure for the construction is detailed in several steps. In the first step, we determine the list of triplets

$$\{(s_1, x_1, n_1), (s_2, x_2, n_2), \dots\} \quad (12.8.36)$$

In the second step, we trim away all the triplets in the list that share the same  $x_k$ ,  $n_k$  as a previous triplet. That is, we go through each  $(x_k, n_k)$  and if we find that it is the same as a previous  $(x_{k'}, n_{k'})$ , we remove the  $k^{\text{th}}$  triplet from the list. This ensures that for any given  $x_k$ , all the  $n_i$  for  $\{i : x_i = x_k\}$  are distinct. In the third step, denote the trimmed down list by

$$\{(s'_1, x'_1, n'_1), (s'_2, x'_2, n'_2), \dots\} \quad (12.8.37)$$

and construct a binary tree as follows. For each  $k$ , put  $(s'_k, x'_k, n'_k)$  at the first node available at level  $n'_k + 1$  in the binary tree. Then, all the descendants of that node cannot be assigned to another tripe (this is to ensure that we maintain a prefix code). To show that this construction will be successful, we show that the codewords given by the binary tree with lengths  $n_k + 1$  satisfy the Kraft inequality:

$$\sum_{k=1}^{\infty} 2^{-(n'_k + 1)} \leq 1 \quad (12.8.38)$$

Begin from the sum and write it out as

$$\sum_{k=1}^{\infty} 2^{-(n'_k+1)} = \sum_{x_j \in \{0,1\}^*} \sum_{\{i:x_i=x_j\}} 2^{-(n'_i+1)} \quad (12.8.39)$$

In the outer sum, we sum over the set  $\{0,1\}^*$  of all binary strings of arbitrary length, while in the inner sum we sum over the indices  $i$  of programs which lead to output equal to  $x_j$ . Working on the inner sum, we have

$$\sum_{\{i:x_i=x_j\}} 2^{-(n_i+1)} = 2^{-1} \sum_{\{i:x_i=x_j\}} 2^{-n'_i} \quad (12.8.40)$$

Now for each  $i$ , we have that

$$-n'_i = - \left\lceil \log_2 \frac{1}{\widehat{P}(x_i)} \right\rceil \quad (12.8.41)$$

$$= \left\lfloor -\log_2 \frac{1}{\widehat{P}(x_i)} \right\rfloor \quad (12.8.42)$$

$$= \left\lfloor \log_2 \widehat{P}(x_i) \right\rfloor \quad (12.8.43)$$

$$\leq \lfloor \log_2 P(x_i) \rfloor \quad (12.8.44)$$

Thus for any program  $s'_i$  that outputs  $x_j$ , its level is bounded by  $-n'_i \leq \lfloor \log_2 P(x_j) \rfloor$ . But by construction, the levels for all triplets with the same output are necessarily distinct. So for the inner sum:

$$2^{-1} \sum_{\{i:x_i=x_j\}} 2^{-n'_i} \leq 2^{-1} \left( 2^{\lfloor \log_2 P(x_j) \rfloor} + 2^{\lfloor \log_2 P(x_j) \rfloor - 1} + 2^{\lfloor \log_2 P(x_j) \rfloor - 2} + \dots \right) \quad (12.8.45)$$

$$= 2^{-1} 2^{\lfloor \log_2 P(x_j) \rfloor} \left( 1 + \frac{1}{2} + \frac{1}{4} + \dots \right) \quad (12.8.46)$$

$$\leq 2^{-1} 2^{\lfloor \log_2 P(x_j) \rfloor} 2 \quad (12.8.47)$$

$$= 2^{\lfloor \log_2 P(x_j) \rfloor} \quad (12.8.48)$$

$$\leq 2^{\log_2 P(x_j)} \quad (12.8.49)$$

$$= P(x_j) \quad (12.8.50)$$

Including the outer sum again,

$$\sum_{x_j \in \{0,1\}^*} \sum_{\{i:x_i=x_j\}} 2^{-(n'_i+1)} = \sum_{x_j \in \{0,1\}^*} P(x_j) \quad (12.8.51)$$

$$\leq 1 \quad (12.8.52)$$

as we have summed over probabilities of mutually exclusive events. Hence the Kraft inequality holds, which shows our construction successfully yields a prefix code for all the triplets. So for any output  $x'_k$ , we could instruct the computer to run the program  $s'_k$  located at level  $n'_k + 1$ . We also know there exists another program for  $x'_i = x'_k$  where the approximation  $\widehat{P}(x'_i)$  is going to be ‘close enough’ to  $P(x'_k)$ . The latter has a codeword of length

$$n'_i + 1 = \left\lceil \log_2 \frac{1}{\widehat{P}(x'_i)} \right\rceil + 1 \quad (12.8.53)$$

$$= \left\lceil \log_2 \frac{1}{P(x'_k)} \right\rceil + 1 \quad (12.8.54)$$

$$\leq \log_2 \frac{1}{P(x'_k)} + 2 \quad (12.8.55)$$

This proves the existence of a program for  $x$  with length increasing as  $\log_2 \frac{1}{P(x)}$ , and therefore the Kolmogorov complexity of  $x$  satisfies

$$K(x) \leq \log_2 \frac{1}{P(x)} + c \quad (12.8.56)$$

with overhead constant  $c$ . □

This result establishes an equivalence between the Kolmogorov complexity and universal probability as a measure of complexity. The difference in Kolmogorov complexity between two universal computers is bounded by a constant, as is  $K(x)$  and  $\log_2 \frac{1}{P(x)}$ . Thus  $\log_2 \frac{1}{P(x)}$  can be treated as a measure of complexity. The less ‘complex’  $x$  is, the more likely it is for a program of random bits to output  $x$ , and so the smaller  $\log_2 \frac{1}{P(x)}$  is.

### 12.8.3 Minimum Description Length [47]

Consider the problem of choosing a probability model  $p$  from some model class  $\mathcal{P}$  that ‘best’ fits some data  $\mathcal{D}$ . The minimum description length principle suggests to pick the  $p \in \mathcal{P}$  such that the quantity

$$\ell(\mathcal{D}; p) = K(p) + \log \frac{1}{p(\mathcal{D})} \quad (12.8.57)$$

is minimised. This quantity can be interpreted as the length required to describe the data (hence the namesake). This description is given by a two-stage procedure. If  $\mathcal{P}$  consists of purely probability mass functions, then the Kolmogorov complexity  $K(p)$  is the length required to describe  $p$ , and  $\log \frac{1}{p(\mathcal{D})}$  is the length required (in bits, if logarithm base 2 is used) to describe the data  $\mathcal{D}$  using the model  $p$ . Choosing the minimum description length formalises the principle of *Occam’s razor*, which asserts that the simplest explanation is usually the right one.

#### Crude Minimum Description Length [75]

A problem with performing minimum description length in practice using the Kolmogorov complexity is that the Kolmogorov complexity is non-computable. Thus we can consider a ‘crude’ approximation where  $K(p)$  is instead replaced by an ad-hoc description length  $\hat{K}(p)$ . One way to construct a description is to suppose that each  $p \in \mathcal{P}$  can be represented by a parameter

$$\theta \in \Theta = \bigcup_{k=1}^{\infty} \Theta_k \quad (12.8.58)$$

where each  $\Theta_k$  is  $k$ -dimensional, and allowed to be an infinite compact set. Thus,  $k$  captures the ‘complexity’ of the model because a higher complexity requires more parameters to describe. Once given  $k$ , then to describe any  $\theta \in \Theta_k$  (which can potentially take on an infinite number of values), we need to introduce a numerical precision  $d$ . The number  $d$  (in bits) allows us to discretise each dimension in  $\Theta_k$  by gridding it uniformly into  $2^d$  values. This is because if we allow up to  $d$  bits to describe each dimension, then the parameter value in each dimension can take up to  $2^d$  values this way. Overall, any  $\theta \in \Theta$  can be encoded by:

- Using  $\log_2 k$  bits to describe dimension  $k$ , using a prefix code for the integers.
- Using  $\log_2 d$  bits to describe the numerical precision  $d$ , also using a prefix code for the integers.
- Given  $k$  and  $d$ , using  $kd$  to describe every component of  $k$ -dimensional  $\theta$  (since each component can be described in  $d$  bits).

Thus the description length of  $p_\theta$  becomes

$$\widehat{K}(p_\theta) = \log_2 k + \log_2 d + kd \quad (12.8.59)$$

where  $k$  and  $d$  implicitly depend on  $\theta$ . Note that if  $\Theta_k$  is uncountably infinite (e.g. it is a compact region of  $\mathbb{R}^k$ ), then  $d$  must be infinite to represent an irrational  $\theta$ , however  $d$  will be finite if we only need to represent rational  $\theta$ .

The crude minimum description length principle for data  $\mathcal{D}$  is given by

$$\theta^* = \operatorname{argmin}_{\theta \in \Theta} \left\{ \log_2 k + \log_2 d + kd + \log_2 \frac{1}{p_\theta(\mathcal{D})} \right\} \quad (12.8.60)$$

If we ignore or fix the complexity terms (i.e. those containing  $k$  and  $d$ ), then the problem reduces to  $\theta^* = \operatorname{argmin}_{\theta \in \Theta} \{-\log p_\theta(\mathcal{D})\}$ , which is equivalent to maximum likelihood estimation. A practically implementable method to perform minimum description length is then to:

1. For each  $k = 1, 2, \dots$  up to some maximum dimension  $K$ , compute the maximum likelihood estimators:

$$\theta_k^* = \operatorname{argmin}_{\theta \in \Theta_k} \{-\log p_\theta(\mathcal{D})\} \quad (12.8.61)$$

2. Using the maximum likelihood estimates as a starting basis, then for each  $d = 1, 2, \dots$  up to some maximum precision  $D$ , compute the discretised maximum likelihood estimators:

$$\theta_{k,d}^* = \operatorname{argmin}_{\theta \in \Theta_{k,d}} \{-\log p_\theta(\mathcal{D})\} \quad (12.8.62)$$

for each  $k = 1, \dots, K$ , where  $\Theta_{k,d}$  denotes  $\Theta_k$  discretised to precision  $d$ .

3. Compute the complexity terms involving  $k$  and  $d$  for each of the preceding estimates, and return the estimate with the shortest description length.

Minimum description length can be thought of as providing a similar functionality as information criteria, because like information criteria, the objective is to minimise the negative log likelihood but with an added penalty term for the model complexity (which in this case, the length required to describe the model).

## 12.9 Information Geometry

### 12.9.1 Statistical Manifolds [40]

Consider a family of parametrised probability density functions:

$$\mathcal{S} = \{p(x; \xi) : \xi \in \Xi\} \quad (12.9.1)$$

where we treat each density function as a function on the support  $\mathcal{X}$ , i.e.

$$p(x; \xi) : \mathcal{X} \rightarrow \mathbb{R} \quad (12.9.2)$$

which is parametrised by  $\xi$  over the parameter space  $\Xi \subseteq \mathbb{R}^n$ . Suppose we have the following regularity conditions:

- The mapping from  $\xi$  to  $p(x; \xi)$  is one-to-one, i.e. for any  $\xi$  we can identify a unique density function, and vice-versa.
- The partial derivatives  $\left( \frac{\partial p}{\partial \xi_1}, \dots, \frac{\partial p}{\partial \xi_n} \right)$  when treated as functions of  $x$  (and hence infinite-dimensional vectors), are linearly independent. So one function cannot be a multiple of another, and roughly speaking, it means that there are no ‘redundant’ parameters.

If  $\mathcal{S}$  satisfies these regularity conditions, then we can call  $\mathcal{S}$  a statistical manifold, which is a geometric structure whereby each point on the manifold maps to a probability density function. We may also use parameters  $\xi = (\xi_1, \dots, \xi_n)$  as a coordinate system to move along the manifold. Note that the choice of coordinate system is not unique, as we can reparametrise the model. For example, the family of univariate Gaussian distributions may be treated as a two-dimensional manifold on the coordinate system  $(\mu, \sigma)$ , where  $\mu \in \mathbb{R}$  parametrises the mean and  $\sigma > 0$  parametrises the standard deviation. Alternatively, we could use the coordinate system  $(\mu, v)$ , where  $\mu \in \mathbb{R}$  parametrises the mean and  $v > 0$  parametrises the variance.

### 12.9.2 Fisher Information Metric

A statistical manifold can also be equipped with a metric, which causes the manifold to become a Riemannian manifold. Informally, the metric can be used to define the notion of lengths, distances and angles between different points on the manifold (i.e. different distributions). We formulate a heuristic construction of the Fisher information metric. Rather than considering each point on the statistical manifold  $\mathcal{S}$  as mapping to the density  $p(x; \xi)$ , instead consider each point as mapping to the log-likelihood function  $\log p(x; \xi)$ . This is an equivalent representation since the log is a monotone transformation, and it just suffices to show that the linear independence regularity condition is also equivalent. We have

$$\frac{\partial \log p(x; \xi)}{\partial \xi_j} = \frac{1}{p(x; \xi)} \cdot \frac{\partial p(x; \xi)}{\partial \xi_j} \quad (12.9.3)$$

Thus we have proportionality:

$$\begin{bmatrix} \frac{\partial \log p(x; \xi)}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log p(x; \xi)}{\partial \xi_n} \end{bmatrix} = \frac{1}{p(x; \xi)} \begin{bmatrix} \frac{\partial p(x; \xi)}{\partial \xi_1} \\ \vdots \\ \frac{\partial p(x; \xi)}{\partial \xi_n} \end{bmatrix} \quad (12.9.4)$$

and if any two functions of  $x$  on the right are proportional, then the corresponding two functions on the left are also proportional, and vice versa. Taking a log is also justified via the characterisation of the information on an event.

Now the *tangent space* to a point on a manifold can be intuitively thought of as a linearisation to the manifold at that point. It is the space spanned by the  $n$  tangent vectors at that point, denoted by  $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ . As established above, the tangent vectors to the log-likelihood are linearly independent, and given by

$$(\mathbf{e}_1, \dots, \mathbf{e}_n) = \left( \frac{\partial \log p(x; \xi)}{\partial \xi_1}, \dots, \frac{\partial \log p(x; \xi)}{\partial \xi_n} \right) \quad (12.9.5)$$

So for a small change  $\Delta \xi_j$  in  $\xi_j$ , the function  $\log p(x; \xi)$  changes by approximately  $\frac{\partial \log p(x; \xi)}{\partial \xi_j} \Delta \xi_j$ . We can also interpret this as the change in information for a small change in the parameters.

Recognise that these functions are also score functions.

Having the concept of a tangent space, we can now introduce an inner product defined over the tangent space. As the inner product generalises the dot product, this will allow us to work with generalised notions of length, distance and angle. Recall that any real-valued function over the support  $\mathcal{X}$  can be treated as a random variable, because it maps elements of the sample space to real values. Hence the infinite-dimensional vectors  $\log p(X; \xi)$  and moreover  $\frac{\partial \log p(X; \xi)}{\partial \xi_j}$  can be considered as random variables. Thus a valid inner product is the expectation of the product of random variables:

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \left\langle \frac{\partial \log p(x; \xi)}{\partial \xi_i}, \frac{\partial \log p(x; \xi)}{\partial \xi_j} \right\rangle \quad (12.9.6)$$

$$= \mathbb{E} \left[ \frac{\partial \log p(X; \xi)}{\partial \xi_i} \frac{\partial \log p(X; \xi)}{\partial \xi_j} \right] \quad (12.9.7)$$

Note the similarity of this with the  $ij^{\text{th}}$  element of the Fisher information matrix. Explicitly, we define the Fisher information metric at a point  $\xi$  as

$$g_{ij}(\xi) = \mathbb{E}_{X \sim p(x; \xi)} \left[ \frac{\partial \log p(X; \xi)}{\partial \xi_i} \cdot \frac{\partial \log p(X; \xi)}{\partial \xi_j} \right] \quad (12.9.8)$$

$$= \int_{\mathcal{X}} \frac{\partial \log p(x; \xi)}{\partial \xi_i} \cdot \frac{\partial \log p(x; \xi)}{\partial \xi_j} p(x; \xi) dx \quad (12.9.9)$$

Since the Fisher information matrix is the covariance of the score, an intuitive interpretation of the Fisher information is how information in one direction changes with change in information in another direction.

### Invariance of Fisher Information

Let  $X$  be a random vector from sample space  $\mathcal{X} \subseteq \mathbb{R}^n$ , whose density is from a statistical manifold  $p_X(x; \xi)$  with Fisher information metric  $g_{X,ij}(\xi)$ . Consider an invertible transform  $h : \mathcal{X} \rightarrow \mathcal{Y}$  so that  $Y = h(X)$  is another random vector on support  $\mathcal{Y} \subseteq \mathbb{R}^n$ . Denote the density of  $Y$  and the associated Fisher information metric computed from this density as  $p_Y(y; \xi)$  and  $g_{Y,ij}(\xi)$  respectively, but under the same parametrisation  $\xi$ , i.e.

$$g_{Y,ij}(\xi) = \mathbb{E}_{Y \sim p_Y(y; \xi)} \left[ \frac{\partial \log p_Y(Y; \xi)}{\partial \xi_i} \cdot \frac{\partial \log p_Y(Y; \xi)}{\partial \xi_j} \right] \quad (12.9.10)$$

Then

$$g_{X,ij}(\xi) = g_{Y,ij}(\xi) \quad (12.9.11)$$

*Proof.* Using the formula for invertible transformations of random vectors, it follows that

$$p_X(x; \xi) = p_Y(h(x); \xi) \left| \det \left( \frac{\partial h(x)}{\partial x} \right) \right| \quad (12.9.12)$$

$$= p_Y(y; \xi) \left| \det \left( \frac{\partial h(x)}{\partial x} \right) \right| \quad (12.9.13)$$

Taking the log-likelihood,

$$\log p_X(x; \xi) = \log p_Y(y; \xi) + \log \left| \det \left( \frac{\partial h(x)}{\partial x} \right) \right| \quad (12.9.14)$$

Since  $h(\cdot)$  does not depend on  $\xi$ , then

$$\frac{\partial \log p_X(x; \xi)}{\partial \xi_j} = \frac{\partial \log p_Y(y; \xi)}{\partial \xi_j} \quad (12.9.15)$$

Hence

$$g_{X,ij}(\xi) = \int_{\mathcal{X}} \frac{\partial \log p_X(x; \xi)}{\partial \xi_i} \cdot \frac{\partial_X \log p(x; \xi)}{\partial \xi_j} p_X(x; \xi) dx \quad (12.9.16)$$

$$= \int_{\mathcal{X}} \frac{\partial \log p_Y(y; \xi)}{\partial \xi_j} \cdot \frac{\partial \log p_Y(y; \xi)}{\partial \xi_j} p_Y(y; \xi) \left| \det \left( \frac{\partial h(x)}{\partial x} \right) \right| dx \quad (12.9.17)$$

$$= \int_{\mathcal{Y}} \frac{\partial \log p_Y(y; \xi)}{\partial \xi_j} \cdot \frac{\partial \log p_Y(y; \xi)}{\partial \xi_j} p_Y(y; \xi) dy \quad (12.9.18)$$

$$= g_{Y,ij}(\xi) \quad (12.9.19)$$

□

This property is known as the invariance of the Fisher information to reparametrisations of the sample space. The ideas are similar to those for the Jeffrey's prior.

Additionally, we can show another property known as the Fisher information metric being *covariant* to reparametrisations of the parameter space. Let  $\theta \in \mathbb{R}^n$  be a reparametrisation of  $\xi \in \mathbb{R}^n$ , so that  $\theta$  and  $\xi$  are related by the invertible transformation  $\xi = \varphi(\theta)$ . Let the density under parametrisation  $\theta$  be denoted by

$$p_\theta(x; \theta) = p_\xi(x; \varphi(\theta)) \quad (12.9.20)$$

$$= p_\xi(x; \xi) \quad (12.9.21)$$

Then the Fisher information matrix under parametrisation by  $\theta$ , denoted  $\mathcal{I}_\theta(\theta)$ , in terms of the Fisher information matrix under parametrisation  $\xi$ , denoted  $\mathcal{I}_\xi(\xi)$ , is given by

$$\mathcal{I}_\theta(\theta) = \frac{\partial \varphi(\theta)}{\partial \theta}^\top \mathcal{I}_\xi(\xi) \frac{\partial \varphi(\theta)}{\partial \theta} \quad (12.9.22)$$

$$= \frac{\partial \xi}{\partial \theta}^\top \mathcal{I}_\xi(\xi) \frac{\partial \xi}{\partial \theta} \quad (12.9.23)$$

where

$$\frac{\partial \xi}{\partial \theta} = \frac{\partial \varphi(\theta)}{\partial \theta} \quad (12.9.24)$$

is the  $n \times n$  Jacobian matrix.

*Proof.* We first write  $\mathcal{I}_\theta(\theta)$  as

$$\mathcal{I}_\theta(\theta) = \mathbb{E} \left[ (\nabla_\theta \log p_\theta(X; \theta)) (\nabla_\theta \log p_\theta(X; \theta))^\top \right] \quad (12.9.25)$$

$$= \int_{\mathcal{X}} (\nabla_\theta \log p_\theta(x; \theta)) (\nabla_\theta \log p_\theta(x; \theta))^\top p_\theta(x; \theta) dx \quad (12.9.26)$$

$$= \int_{\mathcal{X}} \frac{1}{p_\theta(x; \theta)} (\nabla_\theta p_\theta(x; \theta)) (\nabla_\theta p_\theta(x; \theta))^\top \frac{1}{p_\theta(x; \theta)} p_\theta(x; \theta) dx \quad (12.9.27)$$

$$= \int_{\mathcal{X}} \frac{1}{p_\theta(x; \theta)} (\nabla_\theta p_\theta(x; \theta)) (\nabla_\theta p_\theta(x; \theta))^\top dx \quad (12.9.28)$$

Via the chain rule, we have

$$\nabla_\theta p_\theta(x; \theta) = \nabla_\theta p_\xi(x; \xi) \quad (12.9.29)$$

$$= \frac{\partial \xi^\top}{\partial \theta} \nabla_\xi p_\xi(x; \xi) \quad (12.9.30)$$

Hence

$$\mathcal{I}_\theta(\theta) = \int_{\mathcal{X}} \frac{1}{p_\theta(x; \theta)} \frac{\partial \xi^\top}{\partial \theta} \nabla_\xi p_\xi(x; \xi) \nabla_\xi p_\xi(x; \xi)^\top \frac{\partial \xi}{\partial \theta} dx \quad (12.9.31)$$

$$= \frac{\partial \xi^\top}{\partial \theta} \int_{\mathcal{X}} \frac{1}{p_\xi(x; \xi)} \nabla_\xi p_\xi(x; \xi) \nabla_\xi p_\xi(x; \xi)^\top dx \frac{\partial \xi}{\partial \theta} \quad (12.9.32)$$

$$= \frac{\partial \xi^\top}{\partial \theta} \int_{\mathcal{X}} (\nabla_\xi \log p_\xi(x; \xi)) (\nabla_\xi \log p_\xi(x; \xi))^\top p_\xi(x; \xi) dx \frac{\partial \xi}{\partial \theta} \quad (12.9.33)$$

$$= \frac{\partial \xi^\top}{\partial \theta} \mathbb{E} [(\nabla_\xi \log p_\xi(X; \xi)) (\nabla_\xi \log p_\xi(X; \xi))^\top] \frac{\partial \xi}{\partial \theta} \quad (12.9.34)$$

$$= \frac{\partial \xi^\top}{\partial \theta} \mathcal{I}_\xi(\xi) \frac{\partial \xi}{\partial \theta} \quad (12.9.35)$$

□

### Fisher Information as Curvature of Kullback-Leibler Divergence

Consider a family of distributions  $p(x; \xi)$  parametrised by the parameter vector  $\xi$ . For some particular  $\xi'$ , the Fisher information metric can be shown to be the second partial derivative with respect to  $\xi'$  of the KL divergence of  $p(x; \xi)$  from  $p(x; \xi')$  evaluated at  $\xi$ . That is,

$$\left. \frac{\partial^2}{\partial \xi'_i \partial \xi'_j} \text{KL}(p(x; \xi) \| p(x; \xi')) \right|_{\xi'=\xi} = \left. \frac{\partial^2}{\partial \xi'_i \partial \xi'_j} (-\mathbb{E}[\log p(X; \xi')] + \mathbb{E}[\log p(X; \xi)]) \right|_{\xi'=\xi} \quad (12.9.36)$$

$$= - \left. \frac{\partial^2}{\partial \xi'_i \partial \xi'_j} \mathbb{E}[\log p(X; \xi')] \right|_{\xi'=\xi} \quad (12.9.37)$$

$$= - \int_{\mathcal{X}} p(x; \xi) \left. \frac{\partial^2}{\partial \xi'_i \partial \xi'_j} \log p(x; \xi') \right|_{\xi'=\xi} dx \quad (12.9.38)$$

$$= - \int_{\mathcal{X}} p(x; \xi) \frac{\partial^2}{\partial \xi_i \partial \xi_j} \log p(x; \xi) dx \quad (12.9.39)$$

$$= -\mathbb{E} \left[ \frac{\partial^2}{\partial \xi_i \partial \xi_j} \log p(X; \xi) \right] \quad (12.9.40)$$

Then the Hessian  $\nabla_\xi^2 \text{KL}(p(x; \xi) \| p(x; \xi')) \Big|_{\xi'=\xi}$  can be formed from the matrix of all second partial derivatives. Next if we treat  $\log p(x; \xi)$  as the log likelihood, then this gives equivalence to the **Fisher information matrix**. Performing a second order Taylor expansion for the KL divergence in  $\xi'$  about  $\xi$  gives

$$\begin{aligned} \text{KL}(p(x; \xi) \| p(x; \xi')) &\approx \text{KL}(p(x; \xi) \| p(x; \xi)) + (\xi' - \xi)^\top \nabla_{\xi'} \text{KL}(p(x; \xi) \| p(x; \xi')) \Big|_{\xi'=\xi} \\ &\quad + \frac{1}{2} (\xi' - \xi)^\top \nabla_{\xi'}^2 \text{KL}(p(x; \xi) \| p(x; \xi')) \Big|_{\xi'=\xi} (\xi' - \xi) \end{aligned} \quad (12.9.41)$$

$$= \frac{1}{2} (\xi' - \xi)^\top \nabla_{\xi'}^2 \text{KL}(p(x; \xi) \| p(x; \xi')) \Big|_{\xi'=\xi} (\xi' - \xi) \quad (12.9.42)$$

where we have used the properties of the KL divergence that

$$\text{KL}(p(x; \xi) \| p(x; \xi)) = 0 \quad (12.9.43)$$

and takes on a minimum value of zero (by Gibb's inequality), so the gradient also vanishes, i.e.

$$\nabla_{\xi'} \text{KL}(p(x; \xi) \| p(x; \xi'))|_{\xi'=\xi} = 0 \quad (12.9.44)$$

Then the Fisher information metric may be interpreted as the curvature of the KL divergence with respect to  $\xi'$  at  $\xi$ . The intuition this way is that the larger the Fisher information (i.e. the larger the curvature), the more easily the parameter can be distinguished, hence the more ‘information’ implied by  $\xi$ .

### Fisher Information Distance

The Fisher information metric induces a notion of Fisher information distance between points on a statistical manifold. Consider a statistical manifold  $p(x; \xi)$ ; the aim is to compute the information distance between points  $\xi$  and  $\xi'$ . To do this, we use Fisher information metric on the tangent space local to  $\xi$ , so it helps to think of  $\xi$  and  $\xi'$  being ‘close together’ for the computation to be valid. A notion of squared distance is then given by the inner product of their difference

$$d(\xi, \xi')^2 = \langle \xi' - \xi, \xi' - \xi \rangle \quad (12.9.45)$$

which we have define the inner product as the Fisher information metric. This computation is analogous to that which would have been done in Euclidean space. We can split  $\xi' - \xi$  into constituent components in terms of the basis  $(\mathbf{e}_1, \dots, \mathbf{e}_n)$  from the tangent space:

$$\xi' - \xi = (\xi'_1 - \xi_1) \mathbf{e}_1 + \dots + (\xi'_n - \xi_n) \mathbf{e}_n \quad (12.9.46)$$

Hence

$$\begin{aligned} \langle \xi' - \xi, \xi' - \xi \rangle &= \langle (\xi'_1 - \xi_1) \mathbf{e}_1 + \dots + (\xi'_n - \xi_n) \mathbf{e}_n, (\xi'_1 - \xi_1) \mathbf{e}_1 + \dots + (\xi'_n - \xi_n) \mathbf{e}_n \rangle \\ &\quad (12.9.47) \end{aligned}$$

$$\begin{aligned} &= (\xi'_1 - \xi_1) \langle \mathbf{e}_1, \mathbf{e}_1 \rangle (\xi'_1 - \xi_1) + \dots + (\xi'_n - \xi_n) \langle \mathbf{e}_n, \mathbf{e}_n \rangle (\xi'_n - \xi_n) \\ &\quad + \dots + (\xi'_n - \xi_n) \langle \mathbf{e}_n, \mathbf{e}_1 \rangle (\xi'_1 - \xi_1) + \dots + (\xi'_n - \xi_n) \langle \mathbf{e}_n, \mathbf{e}_n \rangle (\xi'_n - \xi_n) \\ &\quad (12.9.48) \end{aligned}$$

$$= [\xi'_1 - \xi_1 \quad \dots \quad \xi'_n - \xi_n] \begin{bmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle & \dots & \langle \mathbf{e}_1, \mathbf{e}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{e}_n, \mathbf{e}_1 \rangle & \dots & \langle \mathbf{e}_n, \mathbf{e}_n \rangle \end{bmatrix} \begin{bmatrix} \xi'_1 - \xi_1 \\ \vdots \\ \xi'_n - \xi_n \end{bmatrix} \quad (12.9.49)$$

$$= (\xi' - \xi)^T \begin{bmatrix} g_{11}(\xi) & \dots & g_{1n}(\xi) \\ \vdots & \ddots & \vdots \\ g_{n1}(\xi) & \dots & g_{nn}(\xi) \end{bmatrix} (\xi' - \xi) \quad (12.9.50)$$

$$= \Delta\xi^T \mathbf{G}(\xi) \Delta\xi \quad (12.9.51)$$

where  $\Delta\xi := \xi' - \xi$  and  $\mathbf{G}(\xi)$  is the Fisher information matrix. Recognise that this quadratic form is the second order approximation of the KL divergence, so we have approximately

$$d(\xi, \xi')^2 = \Delta\xi^T \mathbf{G}(\xi) \Delta\xi \quad (12.9.52)$$

$$\approx 2 \text{KL}(p(x; \xi) \| p(x; \xi')) \quad (12.9.53)$$

Thus the Fisher information metric is a way to formally characterise the KL divergence as a notion of information ‘distance’.

The representation of Fisher information distance above is only valid for  $\xi$  ‘close’ to  $\xi'$ , but for points further apart, it is not a proper distance measure because the distance is not symmetric

(as we have used the tangent space local to  $\xi$ ). Thus, we need to compute the Fisher information *geodesic* distance, by first writing the differential local squared distance as

$$ds^2(\xi) = d\xi^\top \mathbf{G}(\xi) d\xi \quad (12.9.54)$$

Conceptually, we then need to integrate  $ds$  along the geodesic (a notion of the shortest path) between points  $\xi_1$  and  $\xi_2$ :

$$d(\xi_1, \xi_2) = \int_{\xi_1}^{\xi_2} ds(\xi) \quad (12.9.55)$$

which results in symmetry  $d(\xi_1, \xi_2) = d(\xi_2, \xi_1)$ .

### 12.9.3 Natural Gradients [97]

For a statistical manifold  $p(x; \xi)$  with parameter space  $\Xi$ , let  $J(\xi) : \Xi \rightarrow \mathbb{R}$  be a function on the manifold, that we wish to minimise with respect to  $\xi$ . It is known that negative gradient direction  $-\nabla_\xi J(\xi)$  is the steepest descent direction, but only in the case where we consider the metric between two points  $\xi, \xi'$  to be the Euclidean distance  $d(\xi, \xi') = \sqrt{(\xi' - \xi)^\top (\xi' - \xi)}$ . However as we have defined the Fisher information metric on the statistical manifold, we look to find the direction in  $\Xi$  that is the steepest descent direction with respect to the Fisher information metric. Recall that the square of local information distance between two nearby points  $\xi$  and  $\xi + \Delta\xi$  is given by

$$d(\xi, \xi + \Delta\xi)^2 = \Delta\xi^\top \mathbf{G}(\xi) \Delta\xi \quad (12.9.56)$$

where  $\mathbf{G}(\xi)$  is the Fisher information matrix. For a fixed step size in information distance (which recall can be related to the KL divergence), we want to find the direction  $\mathbf{a}$  from the current point  $\xi$  that reduces the objective function the most. Let  $\Delta\xi = \varepsilon\mathbf{a}$ , where  $\varepsilon$  is a small but fixed positive constant. Since  $\varepsilon$  is small, the squared local distance approximation above is valid, and we can also write

$$J(\xi + \varepsilon\mathbf{a}) = J(\xi) + \varepsilon\nabla_\xi J(\xi)^\top \mathbf{a} \quad (12.9.57)$$

To fix the step size, we impose the constraint

$$\mathbf{a}^\top \mathbf{G}(\xi) \mathbf{a} = 1 \quad (12.9.58)$$

Thus finding the steepest descent direction can be formulated as the optimisation problem

$$\max_{\mathbf{a}: \mathbf{a}^\top \mathbf{G}(\xi) \mathbf{a} = 1} \{J(\xi) - J(\xi + \varepsilon\mathbf{a})\} = \max_{\mathbf{a}: \mathbf{a}^\top \mathbf{G}(\xi) \mathbf{a} = 1} \left\{ -\varepsilon \nabla_\xi J(\xi)^\top \mathbf{a} \right\} \quad (12.9.59)$$

$$= \min_{\mathbf{a}: \mathbf{a}^\top \mathbf{G}(\xi) \mathbf{a} = 1} \left\{ \varepsilon \nabla_\xi J(\xi)^\top \mathbf{a} \right\} \quad (12.9.60)$$

This is a constrained problem with equality constraint  $\mathbf{a}^\top \mathbf{G}(\xi) \mathbf{a} - 1 = 0$ , which can be solved with the method of Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\mathbf{a}, \lambda) = \varepsilon \nabla_\xi J(\xi)^\top \mathbf{a} + \lambda \left( \mathbf{a}^\top \mathbf{G}(\xi) \mathbf{a} - 1 \right) \quad (12.9.61)$$

with Lagrange multiplier  $\lambda > 0$ , as the constraint is always active. Differentiating the Lagrangian, we find

$$\nabla_{\mathbf{a}} \mathcal{L}(\mathbf{a}, \lambda) = \varepsilon \nabla_\xi J(\xi) + 2\lambda \mathbf{G}(\xi) \mathbf{a} \quad (12.9.62)$$

We can then set  $\nabla_{\mathbf{a}} \mathcal{L}(\mathbf{a}, \lambda) = 0$  to find the steepest descent direction  $\mathbf{a}^*$ :

$$2\lambda \mathbf{G}(\xi) \mathbf{a}^* = -\varepsilon \nabla_\xi J(\xi) \quad (12.9.63)$$

$$\mathbf{a}^* = -\frac{\varepsilon}{2\lambda} \mathbf{G}(\xi)^{-1} \nabla_\xi J(\xi) \quad (12.9.64)$$

Thus

$$\mathbf{a}^* \propto -\mathbf{G}(\xi)^{-1} \nabla_\xi J(\xi) \quad (12.9.65)$$

as  $\varepsilon$  and  $\lambda$  are both positive. With this, we call

$$\tilde{\nabla} J(\xi) := \mathbf{G}(\xi)^{-1} \nabla_\xi J(\xi) \quad (12.9.66)$$

the natural gradient of  $J(\cdot)$ .

### Natural Gradient Descent

An iterative optimisation method analogous to steepest gradient descent can be derived based on the natural gradient. Letting  $\xi \in \Xi$  be the decision variable, the update for  $\xi_t$  at each iteration is given by

$$\xi_{t+1} = \xi_t - \eta_t \mathbf{G}(\xi_t)^{-1} \nabla_\xi J(\xi_t) \quad (12.9.67)$$

where  $\eta_t$  is the step size. We can compare this update to that in a Newton or Gauss-Newton algorithm, where rather than computing or approximating the Hessian, we use the Fisher information instead.

An example of this in application is for minimisation of a function  $f(x) : \mathcal{X} \rightarrow \mathbb{R}$ , which may be a black box or have discrete search space  $\mathcal{X}$ , so we cannot use gradients. Consider the *stochastic relaxation* of the problem, which is to introduce a statistical manifold  $p(x; \xi)$ ,  $\xi \in \Xi$  where  $\xi$  could for instance contain a location parameter, and minimise

$$J(\xi) = \mathbb{E}_{X \sim p(x; \xi)} [f(X)] \quad (12.9.68)$$

Thus we have transformed the problem from search over  $\mathcal{X}$  (which may be discrete) to search over  $\Xi$ , for which we can use natural gradient descent. We can also imagine the stochastic relaxation as having the effect of ‘smoothing’ the objective function surface. Note that an implementation of a natural gradient descent algorithm may require the natural gradient to be estimated. A Monte-Carlo estimation of the natural gradient can be performed in the following way:

$$\tilde{\nabla} J(\xi) = \mathbf{G}(\xi)^{-1} \nabla_\xi J(\xi) \quad (12.9.69)$$

$$= \mathbf{G}(\xi)^{-1} \nabla_\xi \mathbb{E}_{X \sim p(x; \xi)} [f(X)] \quad (12.9.70)$$

$$= \mathbf{G}(\xi)^{-1} \nabla_\xi \int_{\mathcal{X}} p(x; \xi) f(x) dx \quad (12.9.71)$$

$$= \int_{\mathcal{X}} f(x) \mathbf{G}(\xi)^{-1} \nabla_\xi p(x; \xi) dx \quad (12.9.72)$$

$$= \int_{\mathcal{X}} f(x) \mathbf{G}(\xi)^{-1} p(x; \xi) \nabla_\xi \log p(x; \xi) dx \quad (12.9.73)$$

$$= \mathbb{E}_{X \sim p(x; \xi)} [f(X) \mathbf{G}(\xi)^{-1} \nabla_\xi \log p(X; \xi)] \quad (12.9.74)$$

$$\approx \frac{1}{N} \sum_{i=1}^N f(X_i) \mathbf{G}(\xi)^{-1} \nabla_\xi \log p(X_i; \xi) \quad (12.9.75)$$

Thus we have gone from the gradient of an expectation to the expectation of a gradient, which can be estimated from samples  $X_i$  drawn from the distribution  $p(x; \xi)$ . This form assumes that the Fisher information matrix is analytically known. If this is not the case, it can still be estimated (as is done in maximum likelihood inference, with the observed Fisher information).

### 12.9.4 Bregman Divergence [97]

Suppose  $\psi(\xi)$  is a differentiable convex function in  $\xi$ . At a point  $\xi_0$ , a local linear (i.e. first order Taylor) approximation to  $\psi(\xi)$  is given by

$$\widehat{\psi}(\xi) = \psi(\xi_0) + \nabla_{\xi}\psi(\xi)^{\top}(\xi - \xi_0) \quad (12.9.76)$$

Since  $\psi(\xi)$  is convex, then the graph of  $\psi(\xi)$  will upper bound  $\widehat{\psi}(\xi)$ , which is the supporting hyperplane at point  $\xi_0$ . Thus the term

$$D_{\psi}(\xi\|\xi_0) := \psi(\xi) - \widehat{\psi}(\xi) \quad (12.9.77)$$

$$= \psi(\xi) - \psi(\xi_0) - \nabla_{\xi}\psi(\xi)^{\top}(\xi - \xi_0) \quad (12.9.78)$$

will be non-negative, and we call  $D_{\psi}(\xi\|\xi_0)$  the Bregman divergence of  $\xi$  from  $\xi_0$  derived from convex function  $\psi(\cdot)$ .

#### Kullback-Leibler Divergence as Bregman Divergence

The KL divergence between discrete distributions on finite support (i.e. categorical distributions) can be considered as a particular case of a Bregman divergence. Let  $\xi = (p_1, \dots, p_n)$  be a probability vector, so that  $\sum_{i=1}^n p_i = 1$ . Consider the function

$$\psi(\xi) = \sum_{i=1}^n p_i \log p_i \quad (12.9.79)$$

which can be shown to be convex, i.e we can compute  $\frac{\partial^2 \psi(\xi)}{p_i^2} = \frac{1}{p_i} > 0$ . This function also happens to be the negative entropy. Let  $\mathbf{p}$  and  $\mathbf{q}$  be two probability vectors. Computing the Bregman divergence of  $\mathbf{p}$  from  $\mathbf{q}$ , we have

$$D_{\psi}(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n q_i \log q_i - \mathbf{1}^{\top} \mathbf{p} + \mathbf{1}^{\top} \mathbf{q} - \sum_{i=1}^n p_i \log q_i + \sum_{i=1}^n q_i \log q_i \quad (12.9.80)$$

$$= \sum_{i=1}^n p_i \log p_i - \cancel{\mathbf{1}^{\top} \mathbf{p}} + \cancel{\mathbf{1}^{\top} \mathbf{q}} - \sum_{i=1}^n p_i \log q_i \quad (12.9.81)$$

$$= \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n p_i \log q_i \quad (12.9.82)$$

$$= \sum_{i=1}^n p_i \log \frac{p_i}{q_i} \quad (12.9.83)$$

which is the KL divergence of  $\mathbf{p}$  from  $\mathbf{q}$ .

### 12.9.5 Information Projection

# Chapter 13

# Econometrics

## 13.1 Economic Data

### 13.1.1 Generation of Economic Data

#### Experimental Data

In experimental data, we are able to generate data by controlling the values of the regressors. Hence the design matrix  $\mathbf{X}$  may be treated as a non-random quantity.

#### Observational Data

We treat observational data (also known as *non-experimental data*) ‘as observed’, where there is no ability to control for the regressor values. Hence the data matrix  $\mathbf{X}$  is treated as a random quantity, which is drawn from some underlying population distribution. We call this data as having *stochastic regressors* or *random design*, as opposed to having *non-stochastic regressors* or *non-random design*.

### 13.1.2 Types of Economic Data

#### Cross-Sectional Data

Cross-sectional data records observations within a particular time period. Each observation may be indexed by  $i$ , for each ‘individual’.

#### Time-Series Data

Time-series data records observations across multiple time periods. Each observation may be indexed by  $t$ , for each time period.

#### Panel Data

Also known as longitudinal data, panel data records an observation for each individual  $i$  across multiple time periods  $t$ . This means each observation is indexed by  $i$  and  $t$ .

## 13.2 Model Specification

### 13.2.1 Causal Interpretation of Models

A model for some relationship between  $X$  and  $Y$  may assert that  $X$  causes a change in  $Y$ . The equation which describes exactly how  $X$  causes a change in  $Y$  is known as a causal equation.

## Linear Causal Equations

A linear causal equation between response  $Y$  and explanatory variables  $X_1, X_2, \dots, X_k$  may be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + U \quad (13.2.1)$$

where  $U$  is known as a disturbance/error term which contains all other factors which affect  $Y$ , and may be considered as a random variable. The interpretation of the coefficients  $\beta_1, \beta_2, \dots$  are the causal effects. That is, a 1 unit increase in the explanatory variable causes a respective increase in  $Y$  equal to the value of the coefficient, holding all else constant. Note that  $\beta_0$  need not necessarily have an interpretation since it can be absorbed into  $U$ .

### 13.2.2 Statistical Interpretation of Models

Suppose that the population means of some random variables  $Y_i$  exhibit variation conditional on the values of some explanatory variables  $X_i$ . This relationship for the conditional expectation may be modelled by a general function  $f(\cdot)$ , called the *population regression function* (PRF):

$$\mathbb{E}[Y_i|X_i] = f(X_i) \quad (13.2.2)$$

### Linear Population Regression Functions

A linear population regression function may be obtained by taking the expectation of a linear causal equation conditional on all the explanatory variables:

$$\mathbb{E}[Y_i|X_{1,i}, \dots, X_{k,i}] = \mathbb{E}[\beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i} + U_i|X_{1,i}, \dots, X_{k,i}] \quad (13.2.3)$$

$$= \beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i} + \mathbb{E}[U_i|X_{1,i}, \dots, X_{k,i}] \quad (13.2.4)$$

If we assume the conditional expectation of the error term  $\mathbb{E}[U_i|X_{1,i}, \dots, X_{k,i}] = 0$ , which will be the case if  $U_i$  is independent with  $X_{1,i}, \dots, X_{k,i}$  and  $\mathbb{E}[U_i] = 0$  (which the latter condition can be assumed without loss of generality because any non-zero mean can be absorbed into  $\beta_0$ ), then the population regression function becomes:

$$\mathbb{E}[Y_i|X_{1,i}, \dots, X_{k,i}] = \beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i} \quad (13.2.5)$$

This gives rise to the conditional mean interpretation of the model. We would say that a 1 unit increase in  $X_{1,i}$  is associated with a  $\beta_1$  unit increase in  $Y_i$  on average, holding all else constant. The slope coefficients  $\beta_1, \dots, \beta_k$  are sometimes called the marginal effects of their respective explanatory variables. The coefficient  $\beta_0$  may not necessarily have a valid interpretation (depending on context), but it is the conditional mean of  $Y_i$  when all the explanatory variables are zero. By comparing the population regression function to the causal equation, we can see that

$$Y_i = \mathbb{E}[Y_i|X_{1,i}, \dots, X_{k,i}] + U_i \quad (13.2.6)$$

or equivalently,

$$U_i = Y_i - \mathbb{E}[Y_i|X_{1,i}, \dots, X_{k,i}] \quad (13.2.7)$$

### Linear Sample Regression Functions

The coefficients  $\beta_0, \dots, \beta_k$  in a linear population regression function are population parameters, so they must be estimated. Linear regression via ordinary least squares is a standard way to estimate these parameters. The estimated population regression function is known as a sample regression function and may be written in the form:

$$\widehat{\mathbb{E}}[Y_i|X_{1,i}, \dots, X_{k,i}] = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1,i} + \dots + \widehat{\beta}_k X_{k,i} \quad (13.2.8)$$

where  $\widehat{\beta}_1, \dots, \widehat{\beta}_k$  become the estimates of the marginal effects. Hence the interpretations of the coefficients becomes an interpretation on the estimated conditional mean.

### 13.2.3 Log-Level Models

Suppose we have a simple causal equation of the form

$$\log Y = \beta_0 + \beta_1 X + V \quad (13.2.9)$$

where  $V$  is the error term. This is known as a log-level (or ‘semilog’) specification, and implicitly we require that  $Y$  is a positive variable. Then  $\beta_1 \times 100\%$  may be roughly interpreted as the percentage increase in  $Y$  for a 1 unit increase in  $X$ , holding all else constant. To see this, first note

$$\frac{\partial \log Y}{\partial Y} = \frac{1}{Y} \quad (13.2.10)$$

$$\frac{\partial \log Y}{\partial X} = \beta_1 \quad (13.2.11)$$

Taking the ratio gives

$$\beta_1 = \frac{\partial Y}{\partial X} \cdot \frac{1}{Y} \quad (13.2.12)$$

or in ‘difference’ form:

$$\beta_1 \Delta X \approx \frac{\Delta Y}{Y} \quad (13.2.13)$$

So for a 1 unit increase in  $X$  we have  $\beta_1 \approx \frac{\Delta Y}{Y}$  which is the proportional change in  $Y$ , which can be interpreted using percentages.

### Geometric Mean Interpretation of Log-Level Models

Suppose we have an intercept-only population regression function (i.e. unconditional mean) of the form

$$\mathbb{E} [\log Y_i] = \beta_0 \quad (13.2.14)$$

The estimate of the unconditional mean from a sample is given by

$$\widehat{\mathbb{E}} [\log Y_i] = \frac{1}{n} \sum_{i=1}^n \log Y_i \quad (13.2.15)$$

If we take the exponential of this, we see that

$$\exp \left( \widehat{\mathbb{E}} [\log Y_i] \right) = \exp \left( \frac{1}{n} \sum_{i=1}^n \log Y_i \right) \quad (13.2.16)$$

$$= \exp \left( \sum_{i=1}^n \log Y_i \right)^{1/n} \quad (13.2.17)$$

$$= \left[ \prod_{i=1}^n \exp (\log Y_i) \right]^{1/n} \quad (13.2.18)$$

$$= \left( \prod_{i=1}^n Y_i \right)^{1/n} \quad (13.2.19)$$

which is the geometric mean of  $Y_i$ . Hence the exponential of the unconditional mean of logged variables may be interpreted as the geometric mean.

## Log-Level Models with Multiple Explanatory Variables

These interpretations can be extended to population regression functions with multiple explanatory variables. Suppose we have

$$\mathbb{E}[\log Y_i | X_{1,i}, \dots, X_{k,i}] = \beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i} \quad (13.2.20)$$

with error term  $V_i = \log Y_i - \mathbb{E}[\log Y_i | X_{1,i}, \dots, X_{k,i}]$ . Note that if we make the assumption that  $V_i$  hence  $\log Y_i$  (conditioned on the explanatory variables) is normally distributed, then this implicitly assumes  $Y_i$  conditioned on the explanatory variables is lognormally distributed. Taking the exponential of the causal equation, we get

$$Y_i = \exp(\beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i}) \exp(V_i) \quad (13.2.21)$$

and taking expectations conditional on the explanatory variables gives

$$\mathbb{E}[Y_i | X_{1,i}, \dots, X_{k,i}] = \mathbb{E}[\exp(\beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i}) \exp(V_i) | X_{1,i}, \dots, X_{k,i}] \quad (13.2.22)$$

$$= \exp(\beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i}) \mathbb{E}[\exp(V_i) | X_{1,i}, \dots, X_{k,i}] \quad (13.2.23)$$

Assume that  $\mathbb{E}[\exp(V_i) | X_{1,i}, \dots, X_{k,i}]$  equals some constant  $\kappa_0$  (which will be the case if  $V_i$  were independent with  $X_{1,i}, \dots, X_{k,i}$ ). Then

$$\mathbb{E}[Y_i | X_{1,i}, \dots, X_{k,i}] = \kappa_0 \exp(\beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i}) \quad (13.2.24)$$

To interpret  $\beta_1$  by determining the marginal effect of  $X_{1,i}$  (which could be any arbitrary explanatory variable) in proportional terms, we write

$$\begin{aligned} & \frac{\mathbb{E}[Y_i | X_{1,i} = x+1, X_{2,i} = x_2, \dots, X_{k,i} = x_k] - \mathbb{E}[Y_i | X_{1,i} = x, X_{2,i} = x_2, \dots, X_{k,i} = x_k]}{\mathbb{E}[Y_i | X_{1,i} = x, X_{2,i} = x_2, \dots, X_{k,i} = x_k]} \\ &= \frac{\mathbb{E}[Y_i | X_{1,i} = x+1, X_{2,i} = x_2, \dots, X_{k,i} = x_k]}{\mathbb{E}[Y_i | X_{1,i} = x, X_{2,i} = x_2, \dots, X_{k,i} = x_k]} - 1 \end{aligned} \quad (13.2.25)$$

which yields

$$\frac{\mathbb{E}[Y_i | X_{1,i} = x+1, X_{2,i} = x_2, \dots, X_{k,i} = x_k]}{\mathbb{E}[Y_i | X_{1,i} = x, X_{2,i} = x_2, \dots, X_{k,i} = x_k]} - 1 = \frac{\kappa_0 \exp(\beta_0 + \beta_1(x+1) + \dots + \beta_k x_k)}{\kappa_0 \exp(\beta_0 + \beta_1 x + \dots + \beta_k x_k)} - 1 \quad (13.2.26)$$

$$= \exp(\beta_1) - 1 \quad (13.2.27)$$

A Taylor approximation of  $\exp(\beta_1)$  about zero gives  $\exp(\beta_1) \approx 1 + \beta_1$ , so

$$\exp(\beta_1) - 1 \approx \beta_1 \quad (13.2.28)$$

and we can interpret this by saying that for an increase in  $X_{1,i}$  by 1 unit, the mean of  $Y_i$  increases by approximately  $\beta_1 \times 100\%$ , holding all else constant. We can also arrive at a geometric mean interpretation in a different way. Taking the exponential of the population regression function, we see

$$\exp(\mathbb{E}[\log Y_i | X_{1,i}, \dots, X_{k,i}]) = \exp(\beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i}) \quad (13.2.29)$$

and conditioning this on  $X_{1,i} = x+1, X_{2,i} = x_2, \dots, X_{k,i} = x_k$  yields

$$\exp(\mathbb{E}[\log Y_i | X_{1,i} = x+1, \dots, X_{k,i} = x_k]) = \exp(\beta_0 + \beta_1(x+1) + \dots + \beta_k x_k) \quad (13.2.30)$$

$$= \exp(\beta_1) \exp(\beta_0 + \beta_1 x + \dots + \beta_k x_k) \quad (13.2.31)$$

$$\approx (1 + \beta_1) \exp(\mathbb{E}[\log Y_i | X_{1,i} = x, \dots, X_{k,i} = x_k]) \quad (13.2.32)$$

Analogously to the geometric mean interpretation above, the interpretation here is that for a 1 unit increase in  $X_{1,i}$ , the geometric mean of  $Y_i$  increases by approximately  $\beta_1 \times 100\%$ , holding all else constant.

### 13.2.4 Level-Log Models

Consider a causal model where some (but not necessarily all) explanatory variables may be logged:

$$Y = \beta_0 + \beta_1 \log X_1 + \beta_2 X_2 + \cdots + U \quad (13.2.33)$$

where the logged variables are implicitly positive. To analyse the marginal effect of the logged explanatory variable, we use a simplified population regression function with a single logged explanatory variable (which is of no considerable loss because we are interested in the marginal effect all else held constant):

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 \log X_i \quad (13.2.34)$$

We investigate the marginal effect of a proportional change in  $X_i$  by 1%:

$$\mathbb{E}[Y_i|X_i = 1.01x] - \mathbb{E}[Y_i|X_i = x] = \beta_1 (\log 1.01x - \log x) \quad (13.2.35)$$

$$= \beta_1 \log 1.01 \quad (13.2.36)$$

Hence the exact marginal effect on  $\mathbb{E}[Y_i|X_i]$  will be  $\beta_1 \log 1.01$ , however a Taylor approximation of  $\log(1+x)$  about  $x=0$  gives  $\log(1+x) \approx x$  for small  $x$ . Therefore

$$\mathbb{E}[Y_i|X_i = 1.01x] - \mathbb{E}[Y_i|X_i = x] \approx \frac{\beta_1}{100} \quad (13.2.37)$$

and we can interpret that an 1% increase in  $X_i$ , is associated with an  $\frac{\beta_1}{100}$  increase in  $Y_i$  on average, holding all else constant.

### 13.2.5 Log-Log Models

Given a log-log specification between two variables, e.g.

$$\log Y = \beta_0 + \beta_1 \log X + U \quad (13.2.38)$$

where  $U$  is the error term, we can interpret  $\beta_1$  as ratio of percentage changes. This is also known as the  $Y$ -elasticity of  $X$ . That is, for a 1% increase in  $X$ , there is a  $\beta_1$  percent increase in  $Y$ , holding all else constant. To see this, first write  $\beta_1$  as a ratio of proportional changes:

$$\frac{\Delta Y}{Y} \div \frac{\Delta X}{X} = \beta_1 \quad (13.2.39)$$

Then for small changes in  $X$  and  $Y$  such that we can approximate them with differentials:

$$\frac{dY}{Y} \approx \beta_1 \frac{dX}{X} \quad (13.2.40)$$

$$\int \frac{1}{Y} dY \approx \beta_1 \int \frac{1}{X} dX \quad (13.2.41)$$

$$\log Y \approx \beta_1 \log X + U \quad (13.2.42)$$

which gives back the log-log specification. Note that this implicitly requires  $X$  and  $Y$  to be positive variables. We can also derive this relationship using the marginal effect. Consider the population regression function

$$\mathbb{E}[\log Y_i|X_i] = \beta_0 + \beta_1 \log X_i \quad (13.2.43)$$

where any other explanatory variables may be left out for simplicity. Following the same steps as in the log-level model, we take the exponential of the causal equation:

$$Y = \exp(\beta_0 + \beta_1 \log X + U) \quad (13.2.44)$$

and then assume that  $\mathbb{E}[U_i|X_i] = \kappa_0$  is a constant, then the marginal effect of a 1% increase in  $X_i$  is

$$\frac{\mathbb{E}[Y_i|X_i = 1.01x] - \mathbb{E}[Y_i|X_i = x]}{\mathbb{E}[Y_i|X_i = x]} = \frac{\mathbb{E}[Y_i|X_i = 1.01x]}{\mathbb{E}[Y_i|X_i = x]} - 1 \quad (13.2.45)$$

$$= \exp[\beta_1(\log 1.01x - \log x)] - 1 \quad (13.2.46)$$

Applying the Taylor approximations  $\exp(\beta_1 \log 1.01) - 1 \approx \beta_1 \log 1.01$  for small  $\beta_1$  and  $\log 1.01 \approx 0.01$ , then

$$\frac{\mathbb{E}[Y_i|X_i = 1.01x] - \mathbb{E}[Y_i|X_i = x]}{\mathbb{E}[Y_i|X_i = x]} \approx 0.01\beta_1 \quad (13.2.47)$$

or

$$\mathbb{E}[Y_i|X_i = 1.01x] \approx (1 + 0.01\beta_1)\mathbb{E}[Y_i|X_i = x] \quad (13.2.48)$$

So the interpretation is that for a 1% increase in  $X_1$ , the mean of  $Y_i$  increases by approximately  $\beta_1\%$ , holding all else constant.

### 13.2.6 Quadratic Models

A causal equation specified as a quadratic in an explanatory variable:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U \quad (13.2.49)$$

and corresponding population regression function

$$\mathbb{E}[Y_i|X] = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \quad (13.2.50)$$

is known as a quadratic regression model (but note that the coefficients can still be estimated with ordinary least squares). Here, the interpretation of  $\beta_2$  and the marginal effect of  $X_i$  is not so simple because the marginal effect will in general depend on the value of  $X_i$ . However, the sign of  $\beta_2$  may still be given an interpretation. If  $\beta_2 > 0$ , then the explanatory variable is said to have a *convex quadratic effect*  $Y_i$ , while if  $\beta_2 < 0$  it is said to have a *concave quadratic effect*.

### 13.2.7 Interaction Terms [192]

If a causal equation has two explanatory variables  $X_{1,i}$  and  $X_{2,i}$  and an additional regressor as the product between  $X_{1,i}$  and  $X_{2,i}$  as follows:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i} + U_i \quad (13.2.51)$$

then this regressor is known as an interaction term (capturing the ‘interaction’ between  $X_{1,i}$  and  $X_{2,i}$ ). In the PRF form:

$$\mathbb{E}[Y_i|X_{1,i}, X_{2,i}] = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i} \quad (13.2.52)$$

we obtain the marginal effect of  $X_{1,i}$  as  $\beta_1 + \beta_3 X_{2,i}$ , which depends on the value of  $X_{2,i}$ . In this case, the statistical interpretation would be that a 1 unit increase in  $X_{1,i}$  is associated with a  $\beta_1 + \beta_3 X_{2,i}$  unit increase in  $Y_i$  on average, holding all else constant.

### 13.2.8 Specification Tests

A simple conceptual approach for testing the specification of the functional form for a model is to propose additional terms the model might contain (such as quadratics or logs):

$$Y_i = \beta_0 X_i + \beta_1 X_i^2 + \beta_2 \log X_i + U_i \quad (13.2.53)$$

In this instance, tests of the null hypothesis  $\beta_1 = 0$  and/or  $\beta_2 = 0$  will provide evidence on whether the equation is well-specified (with a rejection of the null being in favour of the proposed specification being a proper specification).

### Ramsey RESET Test [218]

If there are many explanatory variables in the equation, including higher order polynomial terms and other nonlinear terms to test their coefficients may result in a large loss in degrees of freedom. An alternative is to use the Ramsey regression specification error test (RESET), which is to first obtain the estimates and their fitted values of a linear specification:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_k X_{k,i} \quad (13.2.54)$$

and then include higher orders of these polynomials as regressors in another model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_k X_{k,i} + \gamma_1 \hat{Y}_i^2 + \cdots + \gamma_{p-1} \hat{Y}_i^p + U_i \quad (13.2.55)$$

where the maximum order  $p$  is arbitrary, but  $p = 2$  or  $p = 3$  are acceptable. Note that this means the specification with respect to the original explanatory variables now becomes

$$\begin{aligned} Y_i = & \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_k X_{k,i} + \gamma_1 (\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_k X_{k,i})^2 \\ & + \cdots + \gamma_{p-1} (\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_k X_{k,i})^p + U_i \end{aligned} \quad (13.2.56)$$

Then the null hypothesis being tested can be  $\gamma_1 = \cdots = \gamma_{p-1} = 0$  against the alternative that at least one of those coefficients is not equal to zero. A rejection of the null suggests that higher order terms could be included in the model.

### David-MacKinnon Test [218]

The David-MacKinnon Test can be used to compare models that are *non-nested* (that is, one model is not simply a special case of the other, like in the RESET test). Suppose we are comparing a linear form:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + U_i \quad (13.2.57)$$

against a linear-log form:

$$Y_i = \delta_0 + \delta_1 \log X_{1,i} + \cdots + \delta_k \log X_{k,i} + U_i \quad (13.2.58)$$

Let  $\hat{Y}_i$  be the fitted values from the linear-log form. As in a similar idea to the RESET test, we include  $\hat{Y}_i$  as a regressor in the linear form:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + \gamma_1 \hat{Y}_i + U_i \quad (13.2.59)$$

Then a rejection of the test with null  $\gamma_1 = 0$  counts as a rejection of the linear model in favour of the linear-log model. We can also perform the test in the reverse direction instead, using the fitted values from the linear form as a regressor in the log-linear form.

### 13.2.9 Quasi-Maximum Likelihood Estimates

In traditional maximum likelihood, it is implicitly assumed that the likelihood describes the true data generating process. If we allow for the parametric family to be potentially misspecified, the estimator is then known as the quasi-maximum likelihood estimator.

### Consistency of Quasi-Maximum Likelihood Estimation [213]

Let  $p(x; \vartheta)$  be a family of densities from which an i.i.d. sample is drawn from, with true parameter  $\vartheta_0$ . Suppose we misspecify the parametric density as  $q(x; \theta)$  when we employ a maximum likelihood estimator. Then under suitable regularity conditions, we can say that the quasi-maximum likelihood estimate is still consistent in the sense that

$$\operatorname{argmin}_{\theta} \left\{ - \sum_{i=1}^n \log q(x_i; \theta) \right\} \xrightarrow{\text{P}} \operatorname{argmin}_{\theta} \text{KL}(p(x; \vartheta_0) \| q(x; \theta)) \quad (13.2.60)$$

In words, the quasi-maximum likelihood estimate  $\hat{\theta}$  converges in value which minimises the KL divergence from  $q(x; \theta)$  (the misspecified family) to  $p(x; \vartheta_0)$  (the true data generating distribution). Some heuristic arguments are available for why this might be possible. As the sample size  $n \rightarrow \infty$ , the empirical distribution of the sample converges in distribution to the underlying distribution  $p(x; \vartheta_0)$ . Since we showed that maximum likelihood is the same as minimising the KL divergence between the specified parametric family and the empirical distribution, then in the limit, we should naturally anticipate that the KL divergence between the misspecified family and the underlying distribution is minimised.

## 13.3 Regression Analysis

### 13.3.1 Sample Regression Function Coefficients

The formula for the sample regression function coefficient estimates by ordinary least squares may be derived by analogy to the population regression function coefficients. Consider the population regression function

$$\mathbb{E}[Y_i|X_i] = X_i^\top \beta \quad (13.3.1)$$

where  $X_i$  is a vector of explanatory variables concatenated with 1, and  $\beta$  is a vector of coefficients. Multiplying both sides by  $X_i$  and taking  $X_i$  inside the conditional expectation, we get

$$X_i \mathbb{E}[Y_i|X_i] = X_i X_i^\top \beta \quad (13.3.2)$$

$$\mathbb{E}[X_i Y_i | X_i] = X_i X_i^\top \beta \quad (13.3.3)$$

Taking the expectations of both sides and using the law of iterated expectations gives

$$\mathbb{E}[\mathbb{E}[X_i Y_i | X_i]] = \mathbb{E}\left[X_i X_i^\top \beta\right] \quad (13.3.4)$$

$$\mathbb{E}[X_i Y_i] = \mathbb{E}\left[X_i X_i^\top\right] \beta \quad (13.3.5)$$

Lastly, left-multiplying both sides by  $\mathbb{E}[X_i X_i^\top]^{-1}$  yields

$$\mathbb{E}\left[X_i X_i^\top\right]^{-1} \mathbb{E}[X_i Y_i] = \beta \quad (13.3.6)$$

To obtain the sample regression function coefficient estimates, we replace the expectations with their analogous sums from the sample:

$$\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n X_i Y_i \quad (13.3.7)$$

$$= \left( \sum_{i=1}^n X_i X_i^\top \right)^{-1} \sum_{i=1}^n X_i Y_i \quad (13.3.8)$$

which is equivalent to the ordinary least squares formula, only converted from matrix form to summation form.

### 13.3.2 Unbiasedness of Ordinary Least Squares

Under particular assumptions, the estimate of the coefficients in ordinary least squares is unbiased. We assume the following:

- We have a ‘simple random sample’ in that each  $(X_i, Y_i)$  is an i.i.d. observation.
- The population regression function governing the data generating process has the form  $\mathbb{E}[Y_i|X_i] = X_i^\top \beta$ .
- There is no perfect multicollinearity.

We can then show that  $\mathbb{E}[\hat{\beta}] = \beta$  as follows. Let  $\mathcal{X} = \{X_1, \dots, X_n\}$ . By the law of iterated expectations and the ordinary least squares formula,

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|\mathcal{X}]] \quad (13.3.9)$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \sum_{i=1}^n X_i Y_i \middle| \mathcal{X}\right]\right] \quad (13.3.10)$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \sum_{i=1}^n X_i \mathbb{E}[Y_i|\mathcal{X}]\right] \quad (13.3.11)$$

By the independence of the sample, we have  $\mathbb{E}[Y_i|\mathcal{X}] = \mathbb{E}[Y_i|X_i]$ . This property is known as strict exogeneity. Roughly speaking,  $X_j$  does not help to explain  $Y_i$  for any  $j \neq i$ . Applying strict exogeneity and the definition of the population regression function,

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\left[\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \sum_{i=1}^n X_i \mathbb{E}[Y_i|X_i]\right] \quad (13.3.12)$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \sum_{i=1}^n X_i X_i^\top \beta\right] \quad (13.3.13)$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^n X_i X_i^\top\right) \beta\right] \quad (13.3.14)$$

$$= \beta \quad (13.3.15)$$

### 13.3.3 Gauss-Markov Theorem in Econometrics

The Gauss-Markov theorem can be stated for when the regressor matrix  $\mathbf{X}$  is random. The general idea is to condition on  $\mathbf{X}$ , for which we need the following assumptions about the errors  $\varepsilon$ :

- Strict exogeneity:  $\mathbb{E}[\varepsilon|\mathbf{X}] = \mathbf{0}$ .
- Spherical errors:  $\text{Cov}(\varepsilon) = \mathbb{E}[\varepsilon \varepsilon^\top] = \sigma^2 I$ .

Then the results of the Gauss-Markov theorem follow if the covariance of the arbitrary linear estimator and OLS are conditioned on  $\mathbf{X}$ .

### 13.3.4 Asymptotic Normality of Ordinary Least Squares

Retaining the same assumptions as for unbiasedness of OLS (simple random sample, linear PRF, no perfect multicollinearity), and introducing the additional assumption that  $X_i$  and  $U_i$  have finite fourth moments [193], we can show the asymptotic normality of the OLS coefficient estimates. Define the prediction errors

$$U_i = Y_i - \mathbb{E}[Y_i|X_i] \quad (13.3.16)$$

$$= Y_i - X_i^\top \boldsymbol{\beta} \quad (13.3.17)$$

and note that  $\mathbb{E}[U_i|X_i] = 0$ . Then substituting  $Y_i = X_i^\top \boldsymbol{\beta} + U_i$  into the OLS formula:

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^n X_i X_i^\top \right)^{-1} \sum_{i=1}^n X_i (X_i^\top \boldsymbol{\beta} + U_i) \quad (13.3.18)$$

$$= \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n X_i (X_i^\top \boldsymbol{\beta} + U_i) \quad (13.3.19)$$

$$= \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right) \boldsymbol{\beta} + \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n X_i U_i \quad (13.3.20)$$

$$= \boldsymbol{\beta} + \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n X_i U_i \quad (13.3.21)$$

By the weak law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \xrightarrow{\text{P}} \mathbb{E}[X_i X_i^\top] \quad (13.3.22)$$

since  $X_i$  has finite fourth moments, so  $X_i X_i^\top$  has finite second moments. Also by the multivariate central limit theorem, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i U_i \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}, \Omega_{XU}) \quad (13.3.23)$$

where  $\Omega_{XU} := \text{Cov}(X_i U_i)$ , since by the law of iterated expectations:

$$\mathbb{E}[X_i U_i] = \mathbb{E}[\mathbb{E}[X_i U_i | X_i]] \quad (13.3.24)$$

$$= \mathbb{E}[X_i \mathbb{E}[U_i | X_i]] \quad (13.3.25)$$

$$= \mathbf{0} \quad (13.3.26)$$

by exogeneity  $\mathbb{E}[U_i | X_i] = 0$  as established above. Note that our finite fourth moments assumption allows  $X_i U_i$  to satisfy the properties required for the Central Limit Theorem. Then write out the OLS formula as

$$\sqrt{n} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i U_i \quad (13.3.27)$$

By applying Slutsky's theorem (and the fact we are dealing with random quantities which converge in distribution to constants), we get

$$\sqrt{n} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} \mathbb{E}[X_i X_i^\top]^{-1} \mathcal{N}(\mathbf{0}, \Omega_{XU}) \quad (13.3.28)$$

Denoting  $\mathbb{E} [X_i X_i^\top] = \Sigma_{XX}$ , this becomes

$$\sqrt{n} (\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{XX}^{-1} \Omega_{XU} \Sigma_{XX}^{-1}) \quad (13.3.29)$$

Compare this to the case where  $X_i$  are non-random with design matrix  $\mathbf{X}$  and  $U_i$  are assumed to be  $\mathcal{N}(0, \sigma^2)$ , which gives the non-asymptotic result  $\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$ .

To apply this result to inference when performing hypothesis tests and constructing confidence intervals, it means that we can use  $z$ -scores. These should also yield similar results to using  $t$ -scores, since we are considering the sample size to be large enough for the Central Limit Theorem to take effect in the first place.

### 13.3.5 Consistency for Regression Variance

Consider the causal equation  $Y_i = \beta^\top X_i + U_i$ , with  $\mathbb{E}[U_i] = 0$  and variance of the regression  $\text{Var}(U_i) = \sigma^2$ , which we write in matrix form as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U} \quad (13.3.30)$$

where  $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ ,  $\mathbf{X} \in \mathbb{R}^{n \times k}$ ,  $\beta \in \mathbb{R}^{k \times 1}$  and  $\mathbf{U} \in \mathbb{R}^{n \times 1}$ . Based on OLS estimates  $\widehat{\beta}$ , our ‘guesses’ of the error terms  $\mathbf{U}$  are the residuals  $\widehat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\widehat{\beta}$ . We can show that our estimate of the variance of the regression

$$\widehat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n \widehat{U}_i^2 \quad (13.3.31)$$

is a consistent estimator for  $\sigma^2$ . Firstly, express  $\widehat{\mathbf{U}}$  as

$$\widehat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\widehat{\beta} \quad (13.3.32)$$

$$= \mathbf{X}\beta + \mathbf{U} - \mathbf{X}\widehat{\beta} \quad (13.3.33)$$

$$= \mathbf{U} - \mathbf{X}(\widehat{\beta} - \beta) \quad (13.3.34)$$

Then

$$\widehat{\sigma}^2 = \frac{1}{n-k} \widehat{\mathbf{U}}^\top \widehat{\mathbf{U}} \quad (13.3.35)$$

$$= \frac{1}{n-k} \left[ \mathbf{U}^\top \mathbf{U} - 2\mathbf{U}^\top \mathbf{X}(\widehat{\beta} - \beta) + (\widehat{\beta} - \beta)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\beta} - \beta) \right] \quad (13.3.36)$$

$$= \frac{1}{n-k} \left[ \sum_{i=1}^n U_i^2 - 2 \sum_{i=1}^n U_i X_i^\top (\widehat{\beta} - \beta) + (\widehat{\beta} - \beta)^\top \sum_{i=1}^n X_i X_i^\top (\widehat{\beta} - \beta) \right] \quad (13.3.37)$$

$$(13.3.38)$$

Under consistency of OLS,  $\widehat{\beta} \xrightarrow{P} \beta$ , while by the Weak Law of Large Numbers:

$$\frac{1}{n-k} \sum_{i=1}^n U_i^2 \xrightarrow{P} \mathbb{E}[U_i^2] \quad (13.3.39)$$

$$\frac{1}{n-k} \sum_{i=1}^n U_i X_i^\top \xrightarrow{P} \mathbb{E}[X_i U_i]^\top \quad (13.3.40)$$

$$\frac{1}{n-k} \sum_{i=1}^n X_i X_i^\top \xrightarrow{P} \mathbb{E}[X_i X_i^\top] \quad (13.3.41)$$

Since  $\mathbb{E}[U_i^2] = \sigma^2$ , then by Slutsky’s theorem, we are left with

$$\widehat{\sigma}^2 \xrightarrow{P} \sigma^2 \quad (13.3.42)$$

### 13.3.6 Homoskedasticity-Only Standard Errors

The asymptotic normality property of OLS can be used to compute standard errors. Assume the homoskedastic case, that is  $\text{Var}(Y_i|X_i) = \sigma^2$ . Then this implies by the definition of variance:

$$\sigma^2 = \mathbb{E} \left[ (Y_i - \mathbb{E}[Y_i|X_i])^2 \middle| X_i \right] \quad (13.3.43)$$

$$= \mathbb{E} [U_i^2 | X_i] \quad (13.3.44)$$

Hence

$$\Omega_{XU} = \text{Cov}(X_i U_i) \quad (13.3.45)$$

$$= \mathbb{E} \left[ (X_i U_i - \mathbb{E}[X_i U_i])(X_i U_i - \mathbb{E}[X_i U_i])^\top \right] \quad (13.3.46)$$

$$= \mathbb{E} [U_i^2 X_i X_i^\top] \quad (13.3.47)$$

since  $\mathbb{E}[X_i U_i] = 0$  as previously established, and note that  $U_i$  is scalar. Therefore by the law of iterated expectations and the property of homoskedasticity,

$$\Omega_{XU} = \mathbb{E} \left[ \mathbb{E} [U_i^2 X_i X_i^\top | X_i] \right] \quad (13.3.48)$$

$$= \mathbb{E} \left[ \mathbb{E} [U_i^2 | X_i] X_i X_i^\top \right] \quad (13.3.49)$$

$$= \mathbb{E} [\sigma^2 X_i X_i^\top] \quad (13.3.50)$$

$$= \sigma^2 \mathbb{E} [X_i X_i^\top] \quad (13.3.51)$$

$$= \sigma^2 \Sigma_{XX} \quad (13.3.52)$$

So in the asymptotic covariance of  $\hat{\beta}$ , given by  $\mathcal{V} = \frac{1}{n} \Sigma_{XX}^{-1} \Omega_{XU} \Sigma_{XX}^{-1}$ , we have

$$\mathcal{V} = \frac{1}{n} \Sigma_{XX}^{-1} \sigma^2 \Sigma_{XX} \Sigma_{XX}^{-1} \quad (13.3.53)$$

$$= \frac{1}{n} \sigma^2 \Sigma_{XX}^{-1} \quad (13.3.54)$$

Thus if we estimate  $\Sigma_{XX}$  with the term

$$\hat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \quad (13.3.55)$$

and estimate  $\sigma^2$  with the residuals

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \hat{\beta})^2 \quad (13.3.56)$$

We have that both these estimators are consistent. By applying the continuous mapping theorem for  $\hat{\Sigma}_{XX}^{-1}$  and Slutsky's theorem, it follows that the standard errors computed from

$$\hat{\mathcal{V}}_{\text{OLS}} = \frac{1}{n} \hat{\sigma}^2 \hat{\Sigma}_{XX}^{-1} \quad (13.3.57)$$

are consistent for  $\mathcal{V}$ . Also note the parallel to the case when we have a design matrix  $\mathbf{X}$ , which we can show

$$\frac{1}{n} \hat{\sigma}^2 \hat{\Sigma}_{XX}^{-1} = \frac{1}{n} \hat{\sigma}^2 \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \quad (13.3.58)$$

$$= \hat{\sigma}^2 \left( \sum_{i=1}^n X_i X_i^\top \right)^{-1} \quad (13.3.59)$$

$$= \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} \quad (13.3.60)$$

These standard errors can be used to perform homoskedasticity-only  $t$ -tests and construct confidence intervals that are asymptotically valid (and with approximately correct size and coverage properties in large samples) if the regressors are stochastic and we do not have the standard assumption that error terms are normally distributed.

### 13.3.7 Heteroskedasticity

Heteroskedasticity is defined as the case where generally  $\text{Var}(Y_i|X_i) \neq \sigma^2$ . That is, if the model is written in the form

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (13.3.61)$$

and if the errors are uncorrelated, then  $\text{Cov}(\boldsymbol{\varepsilon})$  will be equal to some diagonal matrix (but not scaled identity matrix, as this would be homoskedasticity) with non-negative entries.

### 13.3.8 Tests for Heteroskedasticity

**Test for Multiplicative Heteroskedasticity [205]**

**Goldfeld-Quandt Test**

**Breusch-Pagan Test [218]**

**White Test [218]**

### 13.3.9 White Standard Errors

The asymptotic covariance  $\mathcal{V} = \frac{1}{n} \Sigma_{XX}^{-1} \Omega_{XU} \Sigma_{XX}^{-1}$  of  $\hat{\boldsymbol{\beta}}$  gives an idea on how to compute consistent standard errors in the case of heteroskedasticity, where we cannot assume  $\text{Var}(Y_i|X_i) = \sigma^2$ . We have a consistent estimator  $\hat{\Sigma}_{XX}$  for  $\Sigma_{XX}$  as before. To estimate  $\Omega_{XU}$ , we can use

$$\hat{\Omega}_{XU} = \frac{1}{n} \sum_{i=1}^n \hat{U}_i^2 X_i X_i^\top \quad (13.3.62)$$

where  $\hat{U}_i = Y_i - X_i^\top \hat{\boldsymbol{\beta}}$ . This is a consistent estimator, so we have the following heteroskedasticity-consistent standard errors (also known as White standard errors):

$$\hat{\mathcal{V}} = \frac{1}{n} \hat{\Sigma}_{XX}^{-1} \hat{\Omega}_{XU} \hat{\Sigma}_{XX}^{-1} \quad (13.3.63)$$

These standard errors can be used to perform heteroskedasticity-consistent  $t$ -tests and construct confidence intervals that are asymptotically valid (and with approximately correct size and coverage properties in large samples) if the regressors are stochastic and we do not have the standard assumption that error terms are normally distributed.

### 13.3.10 Multicollinearity [72]

Multicollinearity is when there is a ‘high degree’ of correlation between explanatory variables. Strong multicollinearity can lead to imprecise (i.e. high variance) estimates of the slope coefficients. To illustrate why, suppose we have  $K$  explanatory variables (intercept included), with the  $n \times K$  data matrix  $\mathbf{X}$ . For simplicity, but without loss of generality, suppose that all the

values in  $\mathbf{X}$  have been centered so that the sample means (taken along each column) equal zero. Now partition  $\mathbf{X}$  as

$$\mathbf{X} = [\mathbf{X}_{(K)} \quad \mathbf{x}_K] \quad (13.3.64)$$

so that the  $K^{\text{th}}$  explanatory variable (which can be any arbitrary explanatory variable by choice) is on its own. Consider the matrix  $\mathbf{X}^\top \mathbf{X}$ , which can be written as

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{X}_{(K)}^\top \\ \mathbf{x}_K^\top \end{bmatrix} [\mathbf{X}_{(K)} \quad \mathbf{x}_K] \quad (13.3.65)$$

$$= \begin{bmatrix} \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} & \mathbf{X}_{(K)}^\top \mathbf{x}_K \\ \mathbf{x}_K^\top \mathbf{X}_{(K)} & \mathbf{x}_K^\top \mathbf{x}_K \end{bmatrix} \quad (13.3.66)$$

Using block matrix inversion formulae, the  $K^{\text{th}}$  diagonal element of  $(\mathbf{X}^\top \mathbf{X})^{-1}$  (which we denote  $\Xi_{KK}$ ) can be expressed as

$$\Xi_{KK} = \left[ \mathbf{x}_K^\top \mathbf{x}_K - \mathbf{x}_K^\top \mathbf{X}_{(K)} \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K \right]^{-1} \quad (13.3.67)$$

$$= \left[ \mathbf{x}_K^\top \mathbf{x}_K \left( 1 - \frac{\mathbf{x}_K^\top \mathbf{X}_{(K)} \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K}{\mathbf{x}_K^\top \mathbf{x}_K} \right) \right]^{-1} \quad (13.3.68)$$

Now note that if we fitted a regression of  $\mathbf{x}_K$  on  $\mathbf{X}_{(K)}$ , the ordinary least squares estimates  $\hat{\gamma}$  would be

$$\hat{\gamma} = \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K \quad (13.3.69)$$

and the fitted values  $\hat{\mathbf{x}}_K = \mathbf{X}_{(K)} \hat{\gamma}$  would be

$$\hat{\mathbf{x}}_K = \mathbf{X}_{(K)} \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K \quad (13.3.70)$$

Since the data are all centered, the term

$$\hat{\mathbf{x}}_K^\top \hat{\mathbf{x}}_K = \mathbf{x}_K^\top \mathbf{X}_{(K)} \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K \mathbf{X}_{(K)} \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K \quad (13.3.71)$$

$$= \mathbf{x}_K^\top \mathbf{X}_{(K)} \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K \quad (13.3.72)$$

effectively becomes the sum of squares of the regression for  $\mathbf{x}_K$  on  $\mathbf{X}_{(K)}$ . In the same way,  $\mathbf{x}_K^\top \mathbf{x}_K$  is the total sum of squares. Using the characterisation of the multiple R-squared as the ratio of sum of squares of the regression to total sum of squares, we see that

$$R_K^2 = \frac{\mathbf{x}_K^\top \mathbf{X}_{(K)} \left( \mathbf{X}_{(K)}^\top \mathbf{X}_{(K)} \right)^{-1} \mathbf{X}_{(K)}^\top \mathbf{x}_K}{\mathbf{x}_K^\top \mathbf{x}_K} \quad (13.3.73)$$

where  $R_K^2$  is the multiple R-squared of  $\mathbf{x}_K$  on  $\mathbf{X}_{(K)}$ , and we can write

$$\Xi_{KK} = \frac{1}{(1 - R_K^2) S_{KK}} \quad (13.3.74)$$

where  $S_{KK} = \mathbf{x}_K^\top \mathbf{x}_K$ . Now we see that if  $\mathbf{x}_K$  is more strongly correlated with the other explanatory variables (corresponding to a higher degree of multicollinearity), the  $R_K^2$  increases, causing  $\Xi_{KK}$  to increase. Then recall that the covariance of the ordinary least squares estimator is  $\sigma^2(\mathbf{X}^\top \mathbf{X})$ , so the variance of the  $K^{\text{th}}$  coefficient estimator increases. Consequently, we would also see larger standard errors being computed for that coefficient.

## Perfect Multicollinearity

Consider a regressor vector with intercept term:

$$X_i = \begin{bmatrix} 1 \\ X_{1,i} \\ \vdots \\ X_{k,i} \end{bmatrix} \quad (13.3.75)$$

Perfect multicollinearity means that we can always write

$$a_0 + a_1 X_{1,i} + \cdots + a_k X_{k,i} = 0 \quad (13.3.76)$$

for some  $(a_0, \dots, a_k) \neq (0, \dots, 0)$ . For example,  $X_{1,i}$  could be always twice  $X_{k,i}$ , in which case we would write  $X_{1,i} - 2X_{k,i} = 0$ . Now consider the matrix  $\sum_{i=1}^n X_i X_i^\top$  which is being inverted in the OLS estimator. Then for every single  $i = 1, \dots, n$ , and for the same  $(a_0, \dots, a_k)$ , note that

$$X_i X_i^\top \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} 1 & X_{1,i} & \dots & X_{k,i} \\ X_{1,i} & X_{1,i} X_{1,i} & \dots & X_{1,i} X_{k,i} \\ \vdots & \vdots & \ddots & \vdots \\ X_{k,i} & X_{k,i} X_{1,i} & \dots & X_{k,i} X_{k,i} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} \quad (13.3.77)$$

$$= \begin{bmatrix} a_0 + a_1 X_{1,i} + \cdots + a_k X_{k,i} \\ X_{1,i} (a_0 + a_1 X_{1,i} + \cdots + a_k X_{k,i}) \\ \vdots \\ X_{k,i} (a_0 + a_1 X_{1,i} + \cdots + a_k X_{k,i}) \end{bmatrix} \quad (13.3.78)$$

$$= \mathbf{0} \quad (13.3.79)$$

Hence we also have

$$\sum_{i=1}^n X_i X_i^\top \begin{bmatrix} a_0 \\ \vdots \\ a_k \end{bmatrix} = \mathbf{0} \quad (13.3.80)$$

We have thus found a non-zero vector  $\mathbf{a} := (a_0, \dots, a_k)$  which causes  $\sum_{i=1}^n X_i X_i^\top \mathbf{a} = \mathbf{0}$ . It follows (by the invertible matrix theorem) that  $\sum_{i=1}^n X_i X_i^\top$  is singular, ergo non-invertible. So perfect multicollinearity makes the OLS estimator not exist (which is why the absence of perfect multicollinearity is usually stated as an assumption). We could have also shown the non-invertibility by saying that the data matrix  $\mathbf{X}$  would not have been full column-rank, and because  $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^\top \mathbf{X})$ , then  $\mathbf{X}^\top \mathbf{X}$  would not be full rank either.

## Dummy Variable Trap

In a regression model with an intercept and a  $K$ -way categorical variable, the population regression function specification should be

$$Y = \beta_0 + \beta_1 \mathbb{I}_1 + \cdots + \beta_{K-1} \mathbb{I}_{K-1} \quad (13.3.81)$$

where  $\mathbb{I}_j$  is a dummy (i.e. indicator) variable for the  $j^{\text{th}}$  category. Notice that the  $K^{\text{th}}$  category (which can be any arbitrary category of choice) is omitted. This is to avoid the ‘dummy variable trap’, which is if we included all the categories in the regression. If that were the case, then we would have perfect multicollinearity because of the relation

$$\mathbb{I}_1 + \cdots + \mathbb{I}_K = 1 \quad (13.3.82)$$

That is, with the intercept, we would be able to write one explanatory variable as a linear combination of the others. Thus the rank of the data matrix will not be full rank, and the ordinary least squares estimator will fail to have a solution due to the perfect multicollinearity. By dropping the  $K^{\text{th}}$  category, the  $K^{\text{th}}$  category then becomes the ‘base case’, whereby inference on the other variables should be interpreted in relation to the base case.

An alternative approach for avoiding the dummy variable trap is to not include the intercept in the specification, since the intercept term (i.e. 1) in the explanatory variables is required for the linear combination between variables to hold.

### 13.3.11 Frisch-Waugh-Lovell Theorem [72]

For a multiple linear regression with model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (13.3.83)$$

suppose that we partition the regressor matrix  $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$  and parameter vector  $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\top \ \boldsymbol{\beta}_2^\top]^\top$  such that

$$\mathbf{Y} = [\mathbf{X}_1 \ \mathbf{X}_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\varepsilon} \quad (13.3.84)$$

$$= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad (13.3.85)$$

Then the Frisch-Waugh-Lovell theorem says that the ordinary least squares estimate of  $\boldsymbol{\beta}_2$  will be the same as that for the model of the form

$$M_1\mathbf{Y} = M_1(\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) \quad (13.3.86)$$

where  $M_1$  is a residual-maker matrix (also a ‘projection’ matrix):

$$M_1 = I - \mathbf{X}_1 \left( \mathbf{X}_1^\top \mathbf{X}_1 \right)^{-1} \mathbf{X}_1^\top \quad (13.3.87)$$

*Proof.* We begin with the normal equations:

$$\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y} \quad (13.3.88)$$

$$\begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{bmatrix} [\mathbf{X}_1 \ \mathbf{X}_2] \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{bmatrix} \mathbf{Y} \quad (13.3.89)$$

$$\begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{Y} \\ \mathbf{X}_2^\top \mathbf{Y} \end{bmatrix} \quad (13.3.90)$$

Multiplying out to obtain the first row-block, we get

$$\mathbf{X}_1^\top \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1^\top \mathbf{Y} \quad (13.3.91)$$

which can be rearranged to yield

$$\hat{\boldsymbol{\beta}}_1 = \left( \mathbf{X}_1^\top \mathbf{X}_1 \right)^{-1} \left( \mathbf{X}_1^\top \mathbf{Y} - \mathbf{X}_1^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 \right) \quad (13.3.92)$$

$$= \left( \mathbf{X}_1^\top \mathbf{X}_1 \right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \left( \mathbf{Y} - \mathbf{X}_1^\top \hat{\boldsymbol{\beta}}_2 \right) \quad (13.3.93)$$

For the second row-block, we have

$$\mathbf{X}_2^\top \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2^\top \mathbf{Y} \quad (13.3.94)$$

from which substituting the solution for  $\hat{\beta}_1$  above, gives covariance of the ordinary least squares estimator

$$\mathbf{X}_2^\top \mathbf{X}_1 \left[ (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 (\mathbf{Y} - \mathbf{X}_1^\top \hat{\beta}_2) \right] + \mathbf{X}_2^\top \mathbf{X}_2 \hat{\beta}_2 = \mathbf{X}_2^\top \mathbf{Y} \quad (13.3.95)$$

$$\mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{Y} - \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \hat{\beta}_2 + \mathbf{X}_2^\top \mathbf{X}_2 \hat{\beta}_2 = \mathbf{X}_2^\top \mathbf{Y} \quad (13.3.96)$$

$$\mathbf{X}_2^\top \mathbf{X}_2 \hat{\beta}_2 - \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \hat{\beta}_2 = \mathbf{X}_2^\top \mathbf{Y} - \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{Y} \quad (13.3.97)$$

$$\mathbf{X}_2^\top \left[ I - \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \right] \mathbf{X}_2 \hat{\beta}_2 = \mathbf{X}_2^\top \left[ I - \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \right] \mathbf{Y} \quad (13.3.98)$$

From the definition of  $M_1$ :

$$\mathbf{X}_2^\top M_1 \mathbf{X}_2 \hat{\beta}_2 = \mathbf{X}_2^\top M_1 \mathbf{Y} \quad (13.3.99)$$

Using the facts that the residual-maker matrix  $M_1$  is idempotent ( $M_1^2 = M_1$ ) and symmetric, this can be rewritten as

$$\mathbf{X}_2^\top M_1^\top M_1 \mathbf{X}_2 \hat{\beta}_2 = \mathbf{X}_2^\top M_1^\top M_1 \mathbf{Y} \quad (13.3.100)$$

$$\hat{\beta}_2 = \left( (M_1 \mathbf{X}_2)^\top M_1 \mathbf{X}_2 \right)^{-1} (M_1 \mathbf{X}_2)^\top M_1 \mathbf{Y} \quad (13.3.101)$$

Let  $\tilde{\mathbf{X}}_2 := M_1 \mathbf{X}_2$  and  $\tilde{\mathbf{Y}} := M_1 \mathbf{Y}$ . Then

$$\hat{\beta}_2 = \left( \tilde{\mathbf{X}}_2^\top \tilde{\mathbf{X}}_2 \right)^{-1} \tilde{\mathbf{X}}_2^\top \tilde{\mathbf{Y}} \quad (13.3.102)$$

which takes the form of the ordinary least squares estimator for the model

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}_2 \hat{\beta}_2 + M_1 \varepsilon \quad (13.3.103)$$

□

Note that  $\tilde{\mathbf{Y}}$  are the residuals obtained when  $\mathbf{Y}$  is regressed on  $\mathbf{X}_1$  alone. Also, each column of  $\tilde{\mathbf{X}}_2$  contains the residuals obtained when the respective column of  $\mathbf{X}_2$  is regressed on  $\mathbf{X}_1$ . Hence we can additionally interpret the Frisch-Waugh-Lovell theorem as saying that  $\hat{\beta}_2$  is obtained by regressing the residuals of  $\mathbf{Y}$  on  $\mathbf{X}_1$  against the residuals of  $\mathbf{X}_2$  on  $\mathbf{X}_1$ .

**Corollary 13.1** ([51]). *The residuals obtained from OLS estimation of  $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}_2 \hat{\beta}_2 + M_1 \varepsilon$  are identical to the residuals obtained from OLS estimation of  $\mathbf{Y} = \mathbf{X} \beta + \varepsilon$ .*

*Proof.* We can write for the full regression

$$\mathbf{Y} = \mathbf{X}_1 \hat{\beta}_1 + \mathbf{X}_2 \hat{\beta}_2 + M \mathbf{Y} \quad (13.3.104)$$

with the residual-maker matrix  $M = I - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ . Left-multiplying both sides by  $M_1$  yields

$$M_1 \mathbf{Y} = M_1 \mathbf{X}_1 \hat{\beta}_1 + M_1 \mathbf{X}_2 \hat{\beta}_2 + M_1 M \mathbf{Y} \quad (13.3.105)$$

$$= M_1 \mathbf{X}_2 \hat{\beta}_2 + M_1 M \mathbf{Y} \quad (13.3.106)$$

where  $M_1 \mathbf{X}_1 = \mathbf{0}$  since  $M_1$  plays the role of the annihilator matrix. We then apply the fact  $M_1 M = M$ , which can be derived as follows. Using the geometric interpretation of ordinary least squares, the hat matrix  $\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  is a projection matrix which projects onto the subspace spanned by the columns of  $\mathbf{X}$ . The residual-maker matrix is then another projection matrix which projects onto the orthogonal complement of the span of the columns of  $\mathbf{X}$ . Hence

the projection  $MM_1$  consists of a projection onto the orthogonal complement of the span of  $\mathbf{X}_1$ , followed by a projection onto the orthogonal complement of the span of  $\mathbf{X}$ . Since the former is a subspace of the latter, this implies that  $MM_1 = M$ . Also since  $M$  and  $M_1$  are symmetric, it follows that  $M_1M = M$ . An alternative, more tedious way to show this is to write out the expressions for  $M$  and  $M_1$ , and then apply block-matrix inversion formulae to compute the inverse of the block matrix inside  $M$  which involves  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . Either way, we are left with

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}_2 \hat{\boldsymbol{\beta}}_2 + M\mathbf{Y} \quad (13.3.107)$$

which shows that  $M\mathbf{Y}$  are the residuals when  $\tilde{\mathbf{Y}}$  is regressed on  $\tilde{\mathbf{X}}_2$ .  $\square$

### 13.3.12 F-Tests for Linear Restrictions

Suppose in a linear regression of the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (13.3.108)$$

we have the standard assumption that  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$  and the regressors  $\mathbf{X}$  are non-stochastic. We wish to test a linear restriction on the parameters of the form

$$R\boldsymbol{\beta} = r \quad (13.3.109)$$

where  $R \in \mathbb{R}^{q \times k}$  hence there are  $k$  coefficients and we are testing  $q$  linear restrictions. We know that the sampling distribution of  $\hat{\boldsymbol{\beta}}$  satisfies

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathcal{N}\left(0, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right) \quad (13.3.110)$$

hence

$$R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}\left(0, \sigma^2 R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top\right) \quad (13.3.111)$$

Now let  $B = R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top$  so that

$$\mathbf{z} := \frac{B^{-1/2} R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma} \quad (13.3.112)$$

$$\sim \mathcal{N}(0, I_{q \times q}) \quad (13.3.113)$$

Hence by characterisation of the chi-squared distribution,  $\mathbf{z}^\top \mathbf{z} \sim \chi_q^2$ . Additionally, recall from the analysis on the standard errors of OLS that an unbiased estimator for  $\sigma^2$  in terms of the residual sum of squares RSS is  $\hat{\sigma}^2 = \frac{\text{RSS}}{n-k}$ , with  $\frac{\text{RSS}}{\sigma^2} \sim \chi_{n-k}^2$ . We introduce  $\mathbf{d}$ , which satisfies the following:

$$\mathbf{d} := \frac{\hat{\sigma}^2 (n-k)}{\sigma^2} \quad (13.3.114)$$

$$= \frac{\text{RSS}}{\sigma^2} \quad (13.3.115)$$

$$\sim \chi_{n-k}^2 \quad (13.3.116)$$

Also recall that  $\hat{\sigma}^2$  and  $\hat{\boldsymbol{\beta}}$  are independent, hence  $\mathbf{z}^\top \mathbf{z}$  and  $\mathbf{d}$  are independent. Therefore by the characterisation of the  $F$ -distribution,

$$\frac{\mathbf{z}^\top \mathbf{z}/q}{\mathbf{d}/(n-k)} \sim F_{q, n-k} \quad (13.3.117)$$

This statistic can be rearranged into

$$\frac{\mathbf{z}^\top \mathbf{z}/q}{\mathbf{d}/(n-k)} = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top R^\top (B^{1/2})^\top B^{1/2} R (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/q}{\sigma^2} \div \frac{\widehat{\sigma}^2 (n-k)}{\sigma^2 (n-k)} \quad (13.3.118)$$

$$= \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top R^\top (R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top)^{-1} R (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/q}{\widehat{\sigma}^2} \quad (13.3.119)$$

$$= \frac{(R\widehat{\boldsymbol{\beta}} - r)^\top (R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top)^{-1} (R\widehat{\boldsymbol{\beta}} - r)/q}{\widehat{\sigma}^2} \quad (13.3.120)$$

Hence this statistic can be used to test general linear restriction null hypotheses of the form

$$H_0 : R\boldsymbol{\beta} = r \quad (13.3.121)$$

against the alternative that at least one of the restrictions does not hold, for which the above statistic will be  $F_{q,n-k}$ -distributed under the null.

### *F*-Tests for Restricted Regression Models

Consider a special case of the *F*-test in which the restriction takes the form

$$R = \left[ \begin{array}{ccc|c} 0 & \dots & 0 & 1 \\ \vdots & \ddots & \vdots & \ddots \\ 0 & \dots & 0 & 1 \end{array} \right] \quad (13.3.122)$$

$$r = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad (13.3.123)$$

This expresses that the null hypothesis is for a restricted regression where a subset of the regressors are omitted (note that the ordering of regressors is arbitrary so  $R$  can be written in the above form without loss of generality). We then partition the regressor matrix  $\mathbf{X}$  into

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2] \quad (13.3.124)$$

so then

$$R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top = [\mathbf{0} \quad I] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ I \end{bmatrix} \quad (13.3.125)$$

$$= \Xi_{22} \quad (13.3.126)$$

where we denote the block matrix inverse:

$$\begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix}^{-1} = \begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & \Xi_{22} \end{bmatrix} \quad (13.3.127)$$

Using block matrix inversion formulae, we find that

$$\Xi_{22} = \left[ \mathbf{X}_2^\top \mathbf{X}_2 - \mathbf{X}_2^\top \mathbf{X}_1 \left( \mathbf{X}_1^\top \mathbf{X}_1 \right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \right]^{-1} \quad (13.3.128)$$

$$= \left( \mathbf{X}_2^\top M_1 \mathbf{X}_2 \right)^{-1} \quad (13.3.129)$$

where we recognise the residual-maker matrix  $M_1 = I - \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top$ . Therefore

$$\left( R (\mathbf{X}^\top \mathbf{X})^{-1} R^\top \right)^{-1} = \mathbf{X}_2^\top M_1 \mathbf{X}_2 \quad (13.3.130)$$

Partitioning the parameter estimates  $\widehat{\boldsymbol{\beta}}^\top = [\widehat{\boldsymbol{\beta}}_1^\top \quad \widehat{\boldsymbol{\beta}}_2^\top]$ , note then that  $R\widehat{\boldsymbol{\beta}} - r = \widehat{\boldsymbol{\beta}}_2$ . Substituting these obtained expressions into the numerator of the  $F$ -statistic (and leaving out the division by  $q$ ) yields

$$(R\widehat{\boldsymbol{\beta}} - r)^\top \left( R (\mathbf{X}^\top \mathbf{X})^{-1} R^\top \right)^{-1} (R\widehat{\boldsymbol{\beta}} - r) = \widehat{\boldsymbol{\beta}}_2^\top \mathbf{X}_2^\top M_1 \mathbf{X}_2 \widehat{\boldsymbol{\beta}}_2 \quad (13.3.131)$$

By application of the Frisch-Waugh-Lovell theorem, the OLS solution for  $\widehat{\boldsymbol{\beta}}_2$  is

$$\widehat{\boldsymbol{\beta}}_2 = \left( (M_1 \mathbf{X}_2)^\top M_1 \mathbf{X}_2 \right)^{-1} (M_1 \mathbf{X}_2)^\top M_1 \mathbf{Y} \quad (13.3.132)$$

$$= (\mathbf{X}_2^\top M_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top M_1 \mathbf{Y} \quad (13.3.133)$$

where we have used the standard properties of  $M_1$  (idempotence and symmetry). Hence the term in the numerator becomes

$$\begin{aligned} \mathbf{Y}^\top M_1 \mathbf{X}_2 (\mathbf{X}_2^\top M_1 \mathbf{X}_2)^{-1} (\mathbf{X}_2^\top M_1 \mathbf{X}_2) (\mathbf{X}_2^\top M_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top M_1 \mathbf{Y} \\ = \mathbf{Y}^\top M_1 \mathbf{X}_2 (\mathbf{X}_2^\top M_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top M_1 \mathbf{Y} \end{aligned} \quad (13.3.134)$$

We show that this term can be written in terms of the sum of squared residuals of the restricted and unrestricted regressions:

- Let USSR denote the sum of squared residuals for the unrestricted model (that is, the sum of squared residuals for a regression of  $\mathbf{Y}$  on  $\mathbf{X}$ ).
- Let RSSR denote the sum of squared residuals for the restricted model under the restriction being tested  $\boldsymbol{\beta}_2 = \mathbf{0}$  (that is, the sum of squared residuals for a regression of  $\mathbf{Y}$  on just  $\mathbf{X}_1$ ).

Firstly, the residuals of regressing  $\mathbf{Y}$  on  $\mathbf{X}_1$  is given using the residual-maker matrix by  $M_1 \mathbf{Y}$ . Hence

$$\text{RSSR} = \mathbf{Y}^\top M_1^\top M_1 \mathbf{Y} \quad (13.3.135)$$

$$= \mathbf{Y}^\top M_1 \mathbf{Y} \quad (13.3.136)$$

Appealing to another fact from the Frisch-Waugh-Lovell theorem, we have that the residuals from  $\mathbf{Y}$  on  $\mathbf{X}$  are identical to the residuals from  $M_1 \mathbf{Y}$  on  $M_1 \mathbf{X}_2$ , for which the residual-maker matrix of the latter can be written as

$$I - M_1 \mathbf{X}_2 \left[ (M_1 \mathbf{X}_2)^\top (M_1 \mathbf{X}_2) \right]^{-1} (M_1 \mathbf{X}_2)^\top = I - M_1 \mathbf{X}_2 (\mathbf{X}_2^\top M_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top M_1 \quad (13.3.137)$$

Thus the unrestricted sum of squared residuals is computed by

$$\text{USSR} = \mathbf{Y}^\top M_1 \left[ I - M_1 \mathbf{X}_2 (\mathbf{X}_2^\top M_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top M_1 \right] M_1 \mathbf{Y} \quad (13.3.138)$$

$$= \mathbf{Y}^\top M_1 \mathbf{Y} - \mathbf{Y}^\top M_1 \mathbf{X}_2 (\mathbf{X}_2^\top M_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top M_1 \mathbf{Y} \quad (13.3.139)$$

Now consider the difference between USSR and RSSR; we obtain

$$\text{RSSR} - \text{USSR} = \mathbf{Y}^\top M_1 \mathbf{X}_2 \left( \mathbf{X}_2^\top M_1 \mathbf{X}_2 \right)^{-1} \mathbf{X}_2^\top M_1 \mathbf{Y} \quad (13.3.140)$$

which is equal to the original expression in the numerator for the  $F$ -statistic. Also using the relation  $\hat{\sigma}^2 = \text{USSR}/(n - k)$ , this shows that the  $F$ -statistic can be written as

$$\frac{(\text{RSSR} - \text{USSR})/q}{\text{USSR}/(n - k)} \sim F_{q,n-k} \quad (13.3.141)$$

A very special case is when we wish to test the joint significance of the regression by restricting all the coefficients, excluding the intercept coefficient. This means the restricted model is  $\mathbf{Y}$  regressed on an intercept only, giving an estimate as the sample mean of  $\mathbf{Y}$ . Then in this case,  $q = k - 1$  and  $R$  takes the form

$$R = \begin{bmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{bmatrix} \quad (13.3.142)$$

The test statistic becomes

$$\frac{(\text{RSSR} - \text{USSR})/q}{\text{USSR}/(n - k)} = \frac{(\text{TSS} - \text{RSS})/(k - 1)}{\text{RSS}/(n - k)} \quad (13.3.143)$$

$$= \frac{\text{ESS}/(k - 1)}{\text{RSS}/(n - k)} \quad (13.3.144)$$

$$\sim F_{k-1,n-k} \quad (13.3.145)$$

where we recall the computation for the total sum of squares (equivalent to the residual sum of squares when regressed on an intercept only):

$$\text{TSS} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (13.3.146)$$

and the estimated sum of squares from the relation  $\text{TSS} = \text{ESS} + \text{RSS}$ :

$$\text{ESS} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (13.3.147)$$

### 13.3.13 Wald Tests

The  $F$ -test for linear restrictions of the form  $R\beta = r$  is only valid in finite samples if it is known/assumed that the error terms are normally distributed and homoskedastic. If this assumption is not given, then in large samples it is possible to apply the Wald test for linear restrictions. Under some standard assumptions, the OLS estimator is asymptotically distributed with  $\hat{\beta} \stackrel{a}{\sim} \mathcal{N}(\beta, \mathcal{V})$  where  $\mathcal{V} = \text{Cov}(\hat{\beta})$ . The Wald test statistic is given by

$$W = (R\hat{\beta} - r)^\top (R\hat{\mathcal{V}}R^\top)^{-1} (R\hat{\beta} - r) \quad (13.3.148)$$

where  $\hat{\mathcal{V}}$  is an asymptotically consistent estimator for the covariance  $\text{Cov}(\hat{\beta})$ . In general, this allows for a heteroskedasticity-consistent  $\hat{\mathcal{V}}$  to be used, making the Wald test heteroskedasticity-consistent. Recognise the similarity of this statistic to the numerator in the  $F$ -statistic, however in this case we have:

$$R(\hat{\beta} - \beta) \stackrel{a}{\sim} \mathcal{N}(0, RVR^\top) \quad (13.3.149)$$

Then following similar steps as for the  $F$ -test (applying Slutsky's theorem as necessary so we can use  $\widehat{\mathcal{V}}$  instead of  $\mathcal{V}$ ), we can show that the Wald statistic is asymptotically chi-squared distributed with  $q$  degrees of freedom:

$$W \stackrel{a}{\sim} \chi_q^2 \quad (13.3.150)$$

with  $q$  being the number of restrictions tested.

## 13.4 Instrumental Variables Regression

### 13.4.1 Endogeneity

For simplicity, we consider a causal equation of the form

$$Y_i = \delta_0 + \delta_1 X_i + U_i \quad (13.4.1)$$

We say that explanatory variable  $X_i$  is endogenous if  $X_i$  and  $U_i$  are able to be written as

$$\mathbb{E}[U_i|X_i] = \gamma_0 + \gamma_1 X_i \quad (13.4.2)$$

with  $\gamma_1 \neq 0$ . If  $\gamma_1 = 0$ , then we can say  $X_i$  is *exogenous*, i.e.  $U_i$  is mean independent of  $X_i$ , since

$$\mathbb{E}[U_i] = \mathbb{E}[\mathbb{E}[U_i|X_i]] \quad (13.4.3)$$

$$= \mathbb{E}[\gamma_0] \quad (13.4.4)$$

$$= \gamma_0 \quad (13.4.5)$$

$$= \mathbb{E}[U_i|X_i] \quad (13.4.6)$$

Thus conversely, endogeneity is when  $U_i$  is mean dependent on  $X_i$ . Endogeneity causes problems if we are interested in identifying the true causal coefficient  $\delta_1$  (i.e. the marginal effect with all else held constant). This is because if we write out the population regression function as

$$\mathbb{E}[Y_i|X_i] = \delta_0 + \delta_1 X_i + \mathbb{E}[U_i|X_i] \quad (13.4.7)$$

$$= (\delta_0 + \gamma_0) + (\delta_1 + \gamma_1) X_i \quad (13.4.8)$$

then we can see that we will end up estimating  $\delta_1 + \gamma_1$  rather than  $\delta_1$ .

### Randomised Controlled Trials

If we are given freedom to design the values of the explanatory variable  $X_i$ , then a way to decorrelate  $X_i$  from  $U_i$  is to randomly assign the values of  $X_i$  in each observation, independent of anything else. This is known as the method of randomised controlled trials. However, this approach is not possible if in the data we collect we simply 'observe' the value of  $X_i$  (i.e. the values of the explanatory variables are drawn from some distribution in the data generating process).

### 13.4.2 Omitted Variable Bias

Suppose a causal equation of the form

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + U_i \quad (13.4.9)$$

where the explanatory variables are endogenous and  $\mathbb{E}[U_i] = 0$ . Omitting  $X_{2,i}$  from the causal equation, we can write

$$Y_i = \beta_0 + \beta_1 X_{1,i} + V_i \quad (13.4.10)$$

where the effect of  $X_{2,i}$  has been absorbed into the error term. Now if  $X_{1,i}$  and  $X_{2,i}$  are uncorrelated, then  $X_{1,i}$  and  $V_i$  will be uncorrelated, and we are able to identify the coefficient  $\beta_1$ . However assume that  $X_{1,i}$  and  $X_{2,i}$  are correlated with

$$\mathbb{E}[X_{2,i}|X_{1,i}] = \gamma_0 + \gamma_1 X_{1,i} \quad (13.4.11)$$

Then  $X_{1,i}$  will no longer be endogenous in the causal equation with the omitted variable. In the population regression function we can see that

$$\mathbb{E}[Y_i|X_{1,i}] = \beta_0 + \beta_1 X_{1,i} + \mathbb{E}[V_i|X_{1,i}] \quad (13.4.12)$$

$$= \beta_0 + \beta_1 X_{1,i} + \beta_2 \mathbb{E}[X_{2,i}|X_{1,i}] \quad (13.4.13)$$

$$= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) X_{1,i} \quad (13.4.14)$$

Suppose the population regression function has been specified as

$$\mathbb{E}[Y_i|X_{1,i}] = \delta_0 + \delta_1 X_{1,i} \quad (13.4.15)$$

Equating above, we have

$$\delta_1 = \beta_1 + \beta_2 \gamma_1 \quad (13.4.16)$$

Thus ordinary least squares will end up estimating the value of  $\delta_1 = \beta_1 + \beta_2 \gamma_1$ , whereas we may be interested in identifying the value of  $\beta_1$ . The interpretation of each is as follows:

- $\beta_1$  is the marginal effect of  $X_{1,i}$  controlling for  $X_{2,i}$ .
- $\delta_1$  is the marginal effect of  $X_{1,i}$  without controlling for  $X_{2,i}$ .

The bias  $\beta_2 \gamma_1$  is known as the omitted variable bias. The signs of  $\beta_2$  and  $\gamma_1$  determine the direction of bias. Note that this bias is in the sense of estimating the true causal coefficient  $\beta_1$ . Ordinary least squares with an omitted variable under the standard assumptions will be statistically unbiased for  $\delta_1$  (i.e.  $\mathbb{E}[\hat{\delta}_1] = \delta_1$ ), however it will be biased for the true causal coefficient  $\beta_1$  (i.e.  $\mathbb{E}[\hat{\delta}_1] \neq \beta_1$  when  $\beta_2 \neq 0$  and  $\gamma_1 \neq 0$ ).

### 13.4.3 Measurement Errors [193]

Consider the causal equation

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad (13.4.17)$$

where instead of being able to observe the explanatory variable  $X_i$  perfectly, we are instead able to observe the measurement  $\tilde{X}_i$ , with the measurement error  $\tilde{X}_i - X_i$ . We can write the causal equation in terms of the measurement measurement error as

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i - \underbrace{\beta_1 (\tilde{X}_i - X_i)}_{V_i} + U_i \quad (13.4.18)$$

with new error term  $V_i$ . How the measurement error affects the estimation of the causal coefficient  $\beta_1$  depends on whether  $\tilde{X}_i$  is exogenous with  $V_i$ . If  $\tilde{X}_i$  is correlated with the measurement error  $\tilde{X}_i - X_i$  (and by extension  $V_i$ ) then the regressor  $\tilde{X}_i$  is endogenous with the error term  $V_i$  and we will not be able to consistently estimate  $\beta_1$  with ordinary least squares. Note that any measurement in the dependent variable  $Y_i$  can be captured by the error term  $U_i$ , so the conditions of that measurement error would just need to satisfy any conditions we impose on  $U_i$ .

## Classical Measurement Error

In classical measurement error, we assume the measurement is determined by

$$\tilde{X}_i = X_i + W_i \quad (13.4.19)$$

where  $X_i$  is originally exogenous with  $U_i$  and measurement error  $W_i$  satisfies

$$\text{Cov}(W_i, X_i) \quad (13.4.20)$$

$$\text{Cov}(W_i, U_i) \quad (13.4.21)$$

and we denote  $\text{Var}(W_i) = \sigma_W^2$ ,  $\text{Var}(X_i) = \sigma_X^2$ . Thus the error term  $V_i$  becomes

$$V_i = -\beta_1 (\tilde{X}_i - X_i) + U_i \quad (13.4.22)$$

$$= -\beta_1 W_i + U_i \quad (13.4.23)$$

Now the covariance between  $\tilde{X}_i$  and  $V_i$  can be shown to be

$$\text{Cov}(\tilde{X}_i, V_i) = \text{Cov}(X_i + W_i, -\beta_1 W_i + U_i) \quad (13.4.24)$$

$$= -\beta_1 \underbrace{\text{Cov}(X_i, W_i)}_0 + \underbrace{\text{Cov}(X_i, U_i)}_0 - \beta_1 \text{Cov}(W_i, W_i) + \underbrace{\text{Cov}(W_i, U_i)}_0 \quad (13.4.25)$$

$$= -\beta_1 \sigma_W^2 \quad (13.4.26)$$

Then the PRF of  $V_i$  and  $\tilde{X}_i$  can be expressed as

$$\mathbb{E}[V_i | \tilde{X}_i] = \gamma_0 + \frac{\text{Cov}(\tilde{X}_i, V_i)}{\text{Var}(\tilde{X}_i)} \tilde{X}_i \quad (13.4.27)$$

where

$$\gamma_1 = \frac{-\beta_1 \sigma_W^2}{\sigma_X^2 + \sigma_W^2} \quad (13.4.28)$$

So the PRF of  $Y_i$  on  $\tilde{X}_i$  is

$$\mathbb{E}[Y_i | \tilde{X}_i] = \beta_0 + \beta_1 \mathbb{E}[\tilde{X}_i | \tilde{X}_i] + \mathbb{E}[V_i | \tilde{X}_i] \quad (13.4.29)$$

$$= \beta_0 + \beta_1 \tilde{X}_i + \gamma_0 - \beta_1 \frac{\sigma_W^2}{\sigma_X^2 + \sigma_W^2} \tilde{X}_i \quad (13.4.30)$$

$$= (\beta_0 + \gamma_0) + \beta_1 \left(1 - \frac{\sigma_W^2}{\sigma_X^2 + \sigma_W^2}\right) \tilde{X}_i \quad (13.4.31)$$

$$= (\beta_0 + \gamma_0) + \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2} \tilde{X}_i \quad (13.4.32)$$

Hence if OLS of  $Y_i$  on  $\tilde{X}_i$  is used to estimate  $\beta_1$ , we will instead have

$$\hat{\beta}_1 \xrightarrow{P} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2} \beta_1 \quad (13.4.33)$$

so the estimator will be inconsistent for  $\beta_1$  and we will fail to identify the true causal coefficient. One method to restore consistency is to correct  $\hat{\beta}_1$  by  $\frac{\sigma_X^2 + \sigma_W^2}{\sigma_X^2}$  if  $\sigma_X^2$ ,  $\sigma_W^2$  are known or can be estimated (as is done in errors-in-variables regression).

### Best Guess Measurement Error

Another model for the measurement error is that each measurement constitutes the ‘best guess’ of the explanatory variable  $X_i$  based on information  $F_i$  available to observation  $i$ . That is,

$$\tilde{X}_i = \mathbb{E}[X_i|F_i] \quad (13.4.34)$$

$F_i$  can be random and may not need to be identically distributed for each observation, be we require it to satisfy

$$\mathbb{E}[U_i|F_i] = 0 \quad (13.4.35)$$

Having the best guess gives

$$\mathbb{E}[(\tilde{X}_i - X_i)\tilde{X}_i] = 0 \quad (13.4.36)$$

using the orthogonal projection characterisation of the conditional expectation. To show this explicitly, by definition:

$$(\tilde{X}_i - X_i)\tilde{X}_i = \tilde{X}_i^2 - X_i\tilde{X}_i \quad (13.4.37)$$

$$= \mathbb{E}[X_i|F_i]^2 - X_i\mathbb{E}[X_i|F_i] \quad (13.4.38)$$

Then due to the Law of Iterated Expectations:

$$\mathbb{E}[(\tilde{X}_i - X_i)\tilde{X}_i] = \mathbb{E}\left[\mathbb{E}[(\tilde{X}_i - X_i)\tilde{X}_i|F_i]\right] \quad (13.4.39)$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}[X_i|F_i]^2 - X_i\mathbb{E}[X_i|F_i]|F_i\right]\right] \quad (13.4.40)$$

$$= \mathbb{E}\left[\mathbb{E}[X_i|F_i]^2 - \mathbb{E}[X_i|F_i]^2\right] \quad (13.4.41)$$

$$= 0 \quad (13.4.42)$$

since  $\mathbb{E}[X_i|F_i]$  is treated as non-random when conditioned on  $F_i$ , and thus can be taken out of the conditional expectation like so:

$$\mathbb{E}[X_i\mathbb{E}[X_i|F_i]|F_i] = \mathbb{E}[X_i|F_i]\mathbb{E}[X_i|F_i] \quad (13.4.43)$$

Moreover,

$$\mathbb{E}[\tilde{X}_i - X_i] = \mathbb{E}[\mathbb{E}[\mathbb{E}[X_i|F_i]|F_i] - \mathbb{E}[X_i|F_i]] \quad (13.4.44)$$

$$= \mathbb{E}[\mathbb{E}[X_i|F_i] - \mathbb{E}[X_i|F_i]] \quad (13.4.45)$$

$$= 0 \quad (13.4.46)$$

Therefore  $\text{Cov}(\tilde{X}_i - X_i, \tilde{X}_i)$ . From our assumption  $\mathbb{E}[U_i|F_i] = 0$ , it follows that

$$\text{Cov}(\tilde{X}_i, U_i) = \mathbb{E}[\tilde{X}_i U_i] - \mathbb{E}[\tilde{X}_i] \mathbb{E}[U_i] \quad (13.4.47)$$

$$= \mathbb{E}[\mathbb{E}[X_i|F_i] U_i] - \mathbb{E}[\tilde{X}_i] \mathbb{E}[\mathbb{E}[U_i|F_i]] \quad (13.4.48)$$

$$= \mathbb{E}[\mathbb{E}[\mathbb{E}[X_i|F_i] U_i|F_i]] \quad (13.4.49)$$

$$= \mathbb{E}[\mathbb{E}[X_i|F_i] \mathbb{E}[U_i|F_i]] \quad (13.4.50)$$

$$= 0 \quad (13.4.51)$$

This shows that  $\tilde{X}_i$  is uncorrelated with the overall error term  $V_i$ , meaning it is exogenous. In this case of measurement error, OLS of  $Y_i$  on  $\tilde{X}_i$  will identify the causal coefficient  $\beta_1$ . A way to think about the information  $F_i$  is that it acts like an ‘instrumental variable’ for  $X_i$ .

### 13.4.4 Simultaneous Causal Equations

Another source of endogeneity can be from simultaneity. Suppose we have a pair of simultaneous causal equations given by

$$Y_i = \delta X_i + U_i \quad (13.4.52)$$

$$X_i = \alpha Y_i + V_i \quad (13.4.53)$$

where  $\delta$  and  $\alpha$  are causal coefficients, while  $U_i$  and  $V_i$  are error terms. It is clear that  $X_i$  is endogenous, since a change in  $U_i$  causes a change in  $Y_i$  through the first causal equation, which simultaneously causes a change in  $X_i$  through the second causal equation. Thus  $X_i$  and  $U_i$  will be correlated. This poses a question, that if we estimate the coefficient in the PRF of the form

$$\mathbb{E}[Y_i|X_i] = \beta X_i \quad (13.4.54)$$

that how  $\beta$  relates to the causal coefficients  $\delta$  and  $\alpha$ . From the OLS formula (noting that there is only a single coefficient in this PRF),  $\beta$  is given by

$$\beta = \frac{\mathbb{E}[X_i Y_i]}{\mathbb{E}[X_i^2]} \quad (13.4.55)$$

If we solve the two causal equations simultaneously, we end up with

$$Y_i = \frac{1}{1 - \delta\alpha} (U_i + \delta V_i) \quad (13.4.56)$$

$$X_i = \frac{1}{1 - \delta\alpha} (\alpha U_i + V_i) \quad (13.4.57)$$

This is known as the ‘reduced form’. Let  $\mathbb{E}[U_i^2] = \sigma_U^2$ ,  $\mathbb{E}[V_i^2] = \sigma_V^2$  and for simplicity assume that  $\mathbb{E}[U_i V_i] = 0$  (i.e.  $U_i$  and  $V_i$  are uncorrelated). Then we can compute the following expectations:

$$\mathbb{E}[X_i^2] = \mathbb{E}\left[\left(\frac{1}{1 - \delta\alpha} (\alpha U_i + V_i)\right)^2\right] \quad (13.4.58)$$

$$= \frac{1}{(1 - \delta\alpha)^2} \mathbb{E}[\alpha^2 U_i^2 + 2\alpha U_i V_i + V_i^2] \quad (13.4.59)$$

$$= \frac{\alpha^2 \sigma_U^2 + \sigma_V^2}{(1 - \delta\alpha)^2} \quad (13.4.60)$$

and

$$\mathbb{E}[X_i Y_i] = \mathbb{E}\left[\left(\frac{1}{1 - \delta\alpha} (\alpha U_i + V_i)\right) \left(\frac{1}{1 - \delta\alpha} (U_i + \delta V_i)\right)\right] \quad (13.4.61)$$

$$= \frac{1}{(1 - \delta\alpha)^2} \mathbb{E}[\alpha U_i^2 + \alpha \delta U_i V_i + U_i V_i + \delta V_i^2] \quad (13.4.62)$$

$$= \frac{\alpha \sigma_U^2 + \delta \sigma_V^2}{(1 - \delta\alpha)^2} \quad (13.4.63)$$

Thus we can show

$$\beta = \frac{\alpha \sigma_U^2 + \delta \sigma_V^2}{\alpha^2 \sigma_U^2 + \sigma_V^2} \quad (13.4.64)$$

which implies that the coefficient  $\beta$  is a mixture of the causal coefficients  $\delta$  and  $\alpha$ . Therefore an unbiased estimator for  $\beta$  will be a biased estimator for either  $\delta$  or  $1/\alpha$ . This is sometimes referred to as simultaneity bias.

### 13.4.5 Systems of Simultaneous Causal Equations [72]

#### Structural Form of Simultaneous Causal Equations

Generally, we can have a system of  $M$  causal equations involving  $M$  endogenous variables ( $Y_1, \dots, Y_M$ ),  $K$  exogenous variables ( $X_1, \dots, X_K$ ) (which may include a 1 variable for the intercept coefficient), and  $M$  error terms ( $U_1, \dots, U_M$ ). A system of  $M$  simultaneous causal equations takes the form (dropping the index subscript for notational ease):

$$Y_1 = \alpha_{21}Y_2 + \dots + \alpha_{M1}Y_M + \delta_{11}X_1 + \dots + \delta_{K1}X_K + U_1 \quad (13.4.65)$$

$$Y_2 = \alpha_{12}Y_1 + \dots + \alpha_{M2}Y_M + \delta_{12}X_1 + \dots + \delta_{K2}X_K + U_2 \quad (13.4.66)$$

$$\vdots \quad (13.4.67)$$

$$Y_M = \alpha_{1M}Y_1 + \dots + \delta_{1M}X_1 + \dots + \delta_{KM}X_K + U_M \quad (13.4.68)$$

This is known as the structural form of the model.

#### Matrix Form of Simultaneous Causal Equations

We can rearrange the system of simultaneous causal equations into the matrix form:

$$\underbrace{\begin{bmatrix} Y_1 & \dots & Y_M \end{bmatrix}}_{\mathbf{Y}^\top} \underbrace{\begin{bmatrix} \gamma_{11} & \dots & \gamma_{1M} \\ \vdots & \ddots & \vdots \\ \gamma_{M1} & \dots & \gamma_{MM} \end{bmatrix}}_{\boldsymbol{\Gamma}} + \underbrace{\begin{bmatrix} X_1 & \dots & X_K \end{bmatrix}}_{\mathbf{X}^\top} \underbrace{\begin{bmatrix} \beta_{11} & \dots & \beta_{1M} \\ \vdots & \ddots & \vdots \\ \beta_{K1} & \dots & \beta_{KM} \end{bmatrix}}_{\mathbf{B}} = \underbrace{\begin{bmatrix} U_1 & \dots & U_M \end{bmatrix}}_{\mathbf{U}^\top} \quad (13.4.69)$$

or compactly:

$$\mathbf{Y}^\top \boldsymbol{\Gamma} + \mathbf{X}^\top \mathbf{B} = \mathbf{U}^\top \quad (13.4.70)$$

For normalisation, we may restrict there to be at least a 1 in each column of  $\boldsymbol{\Gamma}$  (for example, the main diagonal can be all 1s).

#### Reduced Form of Simultaneous Causal Equations

If  $\boldsymbol{\Gamma}$  is non-singular (known as the completeness condition for simultaneous causal equations), then the solution of  $\mathbf{Y}$  to the system of simultaneous causal equations is given by

$$\mathbf{Y}^\top = -\mathbf{X}^\top \mathbf{B} \boldsymbol{\Gamma}^{-1} + \mathbf{U}^\top \boldsymbol{\Gamma}^{-1} \quad (13.4.71)$$

Let  $\boldsymbol{\Pi} := -\mathbf{B} \boldsymbol{\Gamma}^{-1}$  and  $\mathbf{V}^\top := \mathbf{U}^\top \boldsymbol{\Gamma}^{-1}$ , then we have

$$\mathbf{Y}^\top = \mathbf{X}^\top \boldsymbol{\Pi} + \mathbf{V}^\top \quad (13.4.72)$$

where  $\mathbf{V}$  is known as the reduced form errors. The reduced form can be used to ‘generate’  $\mathbf{Y}$  from known  $\mathbf{X}$ ,  $\boldsymbol{\Pi}$  and  $\mathbf{V}$ .

#### Rank Condition for Simultaneous Causal Equations [1]

We consider the identifiability of the parameters in  $\mathbf{B}$  and  $\boldsymbol{\Gamma}$ , by which we mean whether there exists a consistent estimator for these parameters. From the reduced form  $\mathbf{Y}^\top = \mathbf{X}^\top \boldsymbol{\Pi} + \mathbf{V}^\top$ , we see that under standard assumptions, ordinary least squares will provide a consistent estimator for  $\boldsymbol{\Pi}$ . Thus to address identifiability we can ask whether the parameters in  $\mathbf{B}$  and  $\boldsymbol{\Gamma}$  can be uniquely determined if  $\boldsymbol{\Pi} = -\mathbf{B} \boldsymbol{\Gamma}^{-1}$  has been uniquely determined. Intuitively, we might suggest that the parameters are identifiable if there are ‘not too many’ non-zero entries in  $\mathbf{B}$  and

$\Gamma^{-1}$ . This condition can be formalised as follows. Assume that there are some zero entries in  $\mathbf{B}$  and  $\Gamma$ , and without loss of generality fix the main diagonal of  $\Gamma$  to be 1s (for normalisation):

$$\Gamma = \begin{bmatrix} 1 & \cdots & \gamma_{1M} \\ \vdots & \ddots & \vdots \\ \gamma_{M1} & \cdots & 1 \end{bmatrix} \quad (13.4.73)$$

We focus on the  $i^{\text{th}}$  simultaneous causal equation, so write out the columns of  $\mathbf{B}$  as

$$\mathbf{B} = [\mathbf{B}_1 \ \dots \ \mathbf{B}_M] \quad (13.4.74)$$

and similarly for  $\Gamma$ :

$$\Gamma = [\Gamma_1 \ \dots \ \Gamma_M] \quad (13.4.75)$$

Then the  $i^{\text{th}}$  simultaneous causal equation involves the parameters in  $\mathbf{B}_i$  and  $\Gamma_i$ . Now let  $\beta_i$  be a column vector of length  $K_i$  consisting of the subset of elements in  $\mathbf{B}_i$  which are non-zero. Then the vector

$$\mathbf{B}'_i = \begin{bmatrix} \beta_i \\ \mathbf{0}_{K-K_i} \end{bmatrix} \quad (13.4.76)$$

(where  $\mathbf{0}_{K-K_i}$  denotes a zero column vector of length  $K - K_i$ ) will be a permutation (i.e. row-swapped version) of  $\mathbf{B}_i$ . Similarly let  $\gamma_i$  be a column vector of length  $M_i$  consisting of the subset of elements in  $\Gamma_i$  which are non-zero and also not fixed to 1. Then

$$\Gamma'_i = \begin{bmatrix} 1 \\ \gamma_i \\ \mathbf{0}_{M-M_i-1} \end{bmatrix} \quad (13.4.77)$$

will be a permutation of  $\Gamma_i$ . Then because  $\Pi\Gamma = -\mathbf{B}$ , we can accordingly write using some column-swapped permutation of  $\Pi$  (the same arrangement used for  $\Gamma'_i$ ):

$$\begin{bmatrix} \pi_{i1} & \Pi_{i1} & \Pi'_{i1} \\ \pi_{i0} & \Pi_{i0} & \Pi'_{i0} \end{bmatrix} \begin{bmatrix} 1 \\ \gamma_i \\ \mathbf{0}_{M-M_i-1} \end{bmatrix} = - \begin{bmatrix} \beta_i \\ \mathbf{0}_{K-K_i} \end{bmatrix} \quad (13.4.78)$$

Here,  $\Pi'_{i1}$  and  $\Pi'_{i0}$  can be miscellaneous values which do not affect the matrix multiplication so we can compactly write

$$\begin{bmatrix} \pi_{i1} & \Pi_{i1} \\ \pi_{i0} & \Pi_{i0} \end{bmatrix} \begin{bmatrix} 1 \\ \gamma_i \end{bmatrix} = - \begin{bmatrix} \beta_i \\ \mathbf{0}_{K-K_i} \end{bmatrix} \quad (13.4.79)$$

This leads to two equations:

$$\pi_{i1} + \Pi_{i1}\gamma_i = -\beta_i \quad (13.4.80)$$

$$\pi_{i0} + \Pi_{i0}\gamma_i = \mathbf{0} \quad (13.4.81)$$

For the second equation, we have a system of linear equations in  $\gamma_i$ :

$$\Pi_{i0}\gamma_i = -\pi_{i0} \quad (13.4.82)$$

This has a unique solution for  $\gamma_i$  if the rank of  $\Pi_{i0}$  is equal to the number of variables to solve for, or in other words,

$$\text{rank}(\Pi_{i0}) = M_i \quad (13.4.83)$$

This is known as the rank condition for identifiability for the  $i^{\text{th}}$  simultaneous causal equation, because we can see that if  $\gamma_i$  is uniquely determined, it then follows that  $\beta_i$  can be uniquely determined from the first equation by

$$\beta_i = -\pi_{i1} - \Pi_{i1}\gamma_i \quad (13.4.84)$$

The rank condition is a sufficient condition for identifiability.

### Order Condition for Simultaneous Causal Equations [1]

Since the dimension of  $\Pi_{i0}$  is  $(K - K_i) \times M_i$ , then

$$\text{rank}(\Pi_{i0}) \leq \min\{K - K_i, M_i\} \quad (13.4.85)$$

Thus a necessary condition in order for the rank condition to hold is that

$$K - K_i \geq M_i \quad (13.4.86)$$

To interpret these integers,  $K - K_i$  is the number of exogenous variables which are excluded from the  $i^{\text{th}}$  simultaneous causal equation, while  $M_i$  is the number of additional endogenous variables included in the  $i^{\text{th}}$  simultaneous causal equation. For example, the order condition may be trivially satisfied if each simultaneous causal equation has its own exogenous variable which does not appear elsewhere, because then  $K - K_i \geq M - 1$  and  $M_i \leq M - 1$ . We also say that:

- If  $K - K_i < M_i$ , then the equation is under-identified.
- If  $K - K_i = M_i$ , then the equation is just-identified.
- If  $K - K_i > M_i$ , then the equation is over-identified.

#### 13.4.6 Two-Stage Least Squares

##### Instrumental Variables

Consider in the context of the simple causal equation

$$Y_i = \delta_0 + \delta_1 X_i + U_i \quad (13.4.87)$$

that we are trying to estimate the true causal coefficient  $\delta_1$  of  $X_i$ , however  $X_i$  is endogenous (for example, due to omitted variables, measurement errors or simultaneity), meaning that  $X_i$  and  $U_i$  are correlated. Thus,  $\delta_1$  cannot be identified (i.e. consistently estimated) using OLS. However, suppose we can find a variable  $Z_i$  which satisfies the *exogeneity* condition:

$$\mathbb{E}[U_i|Z_i] = \lambda_0 \quad (13.4.88)$$

for some constant  $\lambda_0$ . Recall that this condition is somewhere in between independence and uncorrelatedness. That is, independence of  $U_i$  and  $X_i$  implies exogeneity, but exogeneity does not necessarily imply uncorrelatedness. Further suppose that  $Z_i$  and  $X_i$  satisfy the *relevance* condition:

$$\mathbb{E}[X_i|Z_i] = \pi_0 + \pi_1 Z_i \quad (13.4.89)$$

with some  $\pi_1 \neq 0$ . This does express that  $Z_i$  and  $X_i$  are correlated. Then,  $Z_i$  satisfies the conditions to be an *instrumental variable* (or ‘instrument’) for  $X_i$ . To summarise these two conditions in words,  $Z_i$  can only affect the dependent variable  $Y_i$  through its effect on  $X_i$ . Conditioning the causal equation on  $Z_i$  and taking expectations yields

$$\mathbb{E}[Y_i|Z_i] = \delta_0 + \delta_1 \mathbb{E}[X_i|Z_i] + \mathbb{E}[U_i|Z_i] \quad (13.4.90)$$

$$= \delta_0 + \delta_1 \mathbb{E}[X_i|Z_i] \quad (13.4.91)$$

This shows that by specifying the population regression function as  $\widehat{Y}_i$  on  $\mathbb{E}[X_i|Z_i]$ , OLS should be able to identify  $\delta_1$ . To see why the exogeneity and relevance conditions are necessary, suppose we have irrelevance (i.e.  $\pi_1 = 0$ ). Then  $\mathbb{E}[Y_i|Z_i]$  simply becomes

$$\mathbb{E}[Y_i|Z_i] = \delta_0 + \delta_1 \pi_0 \quad (13.4.92)$$

so  $\delta_1$  can no longer be identified. Or if there is endogeneity between  $Z_i$  and  $U_i$  (i.e. we can write  $\mathbb{E}[U_i|Z_i] = \lambda_0 + \lambda_1 Z_i$  with  $\lambda_1 \neq 0$ ), then

$$\mathbb{E}[Y_i|Z_i] = \delta_0 + \delta_1 \mathbb{E}[X_i|Z_i] + \lambda_0 + \lambda_1 Z_i \quad (13.4.93)$$

Substituting  $Z_i = \frac{\mathbb{E}[X_i|Z_i] - \pi_0}{\pi_1}$  obtains

$$\mathbb{E}[Y_i|Z_i] = \delta_0 + \delta_1 \mathbb{E}[X_i|Z_i] + \lambda_0 + \lambda_1 \frac{\mathbb{E}[X_i|Z_i] - \pi_0}{\pi_1} \quad (13.4.94)$$

$$= \left( \delta_0 + \lambda_0 - \frac{\lambda_1 \pi_0}{\pi_1} \right) + \left( \delta_1 + \frac{\lambda_1}{\pi_1} \right) \mathbb{E}[X_i|Z_i] \quad (13.4.95)$$

So we also see in this case that  $\delta_1$  cannot be identified via a PRF of  $\hat{Y}_i$  on  $\mathbb{E}[X_i|Z_i]$ . In the general case with multiple regressors, we write the causal equation as

$$Y_i = X_i^\top \beta + U_i \quad (13.4.96)$$

where  $X_i$  is a vector. If at least one element of  $X_i$  is endogenous, this means that  $\mathbb{E}[U_i|X_i] \neq 0$ . We partition  $X_i$  into endogenous variables  $X_{1,i}$  and exogenous variables  $X_{2,i}$ , for instance:

$$X_i = \begin{bmatrix} 1 \\ X_{1,i} \\ X_{2,i} \end{bmatrix} \quad (13.4.97)$$

If  $Z_{1,i}$  is our vector of instruments, we can construct the instrumental variables vector  $Z_i$  by

$$Z_i = \begin{bmatrix} 1 \\ Z_{1,i} \\ X_{2,i} \end{bmatrix} \quad (13.4.98)$$

and we require that  $\dim(Z_i) = m \geq \dim(X_i) = p$  (i.e. there are at least as many instruments as endogenous variables, and each exogenous variable essentially acts as an instrument for itself). In this general case, the exogeneity condition is still

$$\mathbb{E}[U_i|Z_i] = 0 \quad (13.4.99)$$

as before, but now the relevance condition is that we can write

$$\mathbb{E}[X_i|Z_i] = \Pi^\top Z_i \quad (13.4.100)$$

where the  $m \times p$  matrix  $\Pi$  is full (column) rank, so that  $\text{rank}(\Pi) = p$ .

### Two-Stage Least Squares Estimation

The idea behind two-stage least squares estimation (also known as instrumental variables regression) is to first obtain the estimates  $\hat{X}_i = \hat{\mathbb{E}}[X_i|Z_i]$ , so that we can regress  $Y_i$  on this, since we showed that still will identify the desired causal coefficients. In the simple case for the causal equation  $Y_i = \delta_0 + \delta_1 X_i + U_i$ , the stages are summarised by:

1. Regress  $X_i$  on  $Z_i$  with the sample regression function

$$\hat{\mathbb{E}}[X_i|Z_i] = \hat{\pi}_0 + \hat{\pi}_1 Z_i \quad (13.4.101)$$

to obtain the estimates  $\hat{\pi}_0$ ,  $\hat{\pi}_1$ , and subsequently the fitted values  $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ .

2. Regress  $Y_i$  on the fitted values  $\hat{X}_i$  in the sample regression function

$$\hat{\mathbb{E}}[Y_i|Z_i] = \hat{\delta}_0 + \hat{\delta}_1 \hat{X}_i \quad (13.4.102)$$

which gives the estimate  $\hat{\delta}_1$  for  $\delta_1$ .

The stages in the general case with causal equation  $Y_i = \beta^\top X_i + U_i$  are analogous:

1. Regress  $X_i$  on  $Z_i$  with the sample regression function

$$\hat{\mathbb{E}}[X_i|Z_i] = \hat{\Pi}^\top Z_i \quad (13.4.103)$$

using the OLS estimator (multiple output)

$$\hat{\Pi} = \left( \sum_{i=1}^n Z_i Z_i^\top \right)^{-1} \sum_{i=1}^n Z_i X_i^\top \quad (13.4.104)$$

to obtain fitted values  $\hat{X}_i = \hat{\Pi}^\top Z_i$ .

2. Regress  $Y_i$  on the fitted values  $\hat{X}_i$  in the sample regression function

$$\hat{\mathbb{E}}[Y_i|Z_i] = \hat{\beta}^\top \hat{X}_i \quad (13.4.105)$$

using the two-stage least squares (2SLS) estimator for  $\beta$ :

$$\hat{\beta} = \left( \sum_{i=1}^n \hat{X}_i \hat{X}_i^\top \right)^{-1} \sum_{i=1}^n \hat{X}_i Y_i \quad (13.4.106)$$

To use 2SLS for estimating systems of simultaneous equations, we begin from the reduced form  $\mathbf{Y}_i = \boldsymbol{\Pi}^\top \mathbf{X}_i + \mathbf{V}_i$ .

1. Regress endogenous variables  $\mathbf{Y}_i$  on exogenous variables  $\mathbf{X}_i$  using multiple-output least squares to obtain coefficients  $\hat{\boldsymbol{\Pi}}$  and fitted values  $\hat{\mathbf{Y}}_i = \hat{\boldsymbol{\Pi}} \mathbf{X}_i$ .
2. In the structural form, perform OLS for each equation by regressing the endogenous variable on the exogenous variables and fitted endogenous variables to obtain the estimates  $\hat{\alpha}_{21}, \dots, \hat{\delta}_{KM}$  of the causal coefficients.

### Consistency of Two-Stage Least Squares

Consistency for the 2SLS estimator can be shown in the general case, under the assumptions of exogeneity and relevance. Since the first stage uses OLS which was shown to be consistent under the usual assumptions, then  $\hat{\boldsymbol{\Pi}} \xrightarrow{P} \boldsymbol{\Pi}$ . Write out the 2SLS estimator for  $\beta$  (by including the  $1/n$  factors which cancel each other out) as

$$\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^n \hat{X}_i \hat{X}_i^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n \hat{X}_i Y_i \quad (13.4.107)$$

$$= \left( \hat{\boldsymbol{\Pi}} \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top \hat{\boldsymbol{\Pi}} \right)^{-1} \hat{\boldsymbol{\Pi}} \frac{1}{n} \sum_{i=1}^n Z_i Y_i \quad (13.4.108)$$

Define  $\Sigma_{ZZ} := \mathbb{E}[Z_i Z_i^\top]$  and  $\Sigma_{ZY} := \mathbb{E}[Z_i Y_i]$ . Then the Law of Large Numbers combined with a Slutsky's Theorem argument allows us to determine the probability limit of  $\hat{\beta}$  as

$$\hat{\beta} \xrightarrow{P} \left( \boldsymbol{\Pi}^\top \Sigma_{ZZ} \boldsymbol{\Pi} \right)^{-1} \boldsymbol{\Pi}^\top \Sigma_{ZY} \quad (13.4.109)$$

Note that the relevance condition is required for this probability limit to exist. If  $\text{rank}(\Pi) < p$ , then

$$\text{rank}(\Pi^\top \Sigma_{ZZ} \Pi) \leq \min\{\text{rank}(\Pi), \text{rank}(\Sigma_{ZZ})\} \quad (13.4.110)$$

$$< p \quad (13.4.111)$$

as  $\text{rank}(\Sigma_{ZZ}) = m \geq p$ . Hence  $\Pi^\top \Sigma_{ZZ} \Pi$  would not be invertible. If relevance was satisfied and we ran 2SLS on random data,  $\widehat{\Pi}$  might end up being full rank, but still poorly conditioned such that  $\widehat{\Pi} \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top \widehat{\Pi}$  is poorly conditioned, leading to ‘strange’ estimates in  $\widehat{\beta}$ . Continuing on, we show that this probability limit is equivalently:

$$(\Pi^\top \Sigma_{ZZ} \Pi)^{-1} \Pi^\top \Sigma_{ZY} = (\Pi^\top \Sigma_{ZZ} \Pi)^{-1} \Pi^\top \mathbb{E}[Z_i Y_i] \quad (13.4.112)$$

$$= (\Pi^\top \Sigma_{ZZ} \Pi)^{-1} \Pi^\top \mathbb{E}[Z_i \mathbb{E}[Y_i | Z_i]] \quad (13.4.113)$$

$$= (\Pi^\top \Sigma_{ZZ} \Pi)^{-1} \Pi^\top \mathbb{E}\left[Z_i \mathbb{E}\left[X_i^\top \beta + U_i | Z_i\right]\right] \quad (13.4.114)$$

$$= (\Pi^\top \Sigma_{ZZ} \Pi)^{-1} \Pi^\top \mathbb{E}\left[Z_i \mathbb{E}[X_i | Z_i]^\top \beta\right] \quad (13.4.115)$$

$$= (\Pi^\top \Sigma_{ZZ} \Pi)^{-1} \Pi^\top \mathbb{E}\left[Z_i Z_i^\top \Pi \beta\right] \quad (13.4.116)$$

$$= (\Pi^\top \Sigma_{ZZ} \Pi)^{-1} \Pi^\top \Sigma_{ZZ} \Pi \beta \quad (13.4.117)$$

$$= \beta \quad (13.4.118)$$

where we used the Law of Iterated Expectations in the second equality, and the exogeneity assumption in the fourth equality. Therefore  $\widehat{\beta} \xrightarrow{P} \beta$ . If the exogeneity condition were not satisfied, then the probability limit would not be equal to  $\beta$ .

### Bias of Two-Stage Least Squares

In finite samples, the 2SLS estimator is generally biased. Some intuition can be provided on how 2SLS is biased. Consider a univariate regressor (plus intercept) for simplicity, and suppose we use a valid instrument  $Z_i$  that is perfectly correlated with  $X_i$ . Thus we can perfectly identify the PRF  $\mathbb{E}[X_i | Z_i] = \pi_0 + \pi_1 Z_i$  from the sample, i.e.  $\widehat{\pi}_0 = \pi_0$  and  $\widehat{\pi}_1 = \pi_1$ . Then the fitted values  $\widehat{X}_i$  in the first stage will be identical to  $Z_i$ . Moreover, the perfect correlation with a valid instrument has to imply that  $X_i$  is exogenous, hence OLS on an exogenous regressor in the second stage yields an unbiased estimate in this scenario. Now suppose the instrument is no longer perfectly correlated, and there can be multiple instruments. Then  $X_i$  is decomposed into

$$X_i = \pi_0 + \Pi^\top Z_i + V_i \quad (13.4.119)$$

where  $\pi_0 + \Pi^\top Z_i$  is the part not correlated with the error  $U_i$  (in the causal equation for  $Y_i$ ), and  $V_i$  is the part correlated with  $U_i$  (making  $X_i$  endogenous). If the instrument  $Z_i$  is ‘weak’ (i.e. weak correlation with  $X_i$ ), then most of the variation in  $X_i$  is explained by variation in  $V_i$ . Intuitively, the influence of  $V_i$  on the fitted values is stronger, which biases the second stage estimate.

Additionally, if the number of instruments is equal to the number of endogenous regressors, then the expectation of the 2SLS estimator does not exist [154]. So then it would not be sensible to discuss biasedness (although it would still be appropriate to discuss median-bias; whether the median of  $\widehat{\beta}$  is equal to  $\beta$ ). Some intuition is also available for my the expectation does not exist. For simplicity, consider a univariate regressor and instrument (with no intercept), so each

stage involves estimating just a single coefficient. Then the 2SLS estimate can be derived as

$$\hat{\beta} = \frac{\mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}}{\mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}} \quad (13.4.120)$$

which is a ratio of random variables. We know that the ratio of two zero-mean normal random variables is Cauchy distributed (which has an undefined mean), so it is reasonable to think that under particular conditions, this ratio will also have an undefined mean.

### 13.4.7 Two-Stage Least Squares Inference

#### Asymptotic Normality of Two-Stage Least Squares

In ordinary least squares, asymptotic normality of the form  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{XX}^{-1} \Omega_{XU} \Sigma_{XX}^{-1})$  was shown. Comparing the 2SLS estimator against the OLS estimator for  $\beta$ , we see that  $X_i$  has been replaced by  $\tilde{X}_i = \tilde{\Pi}^\top Z_i$ . In the large sample limit, each  $\tilde{X}_i$  converges in probability to  $\tilde{X}_i := \Pi^\top Z_i$ . So in much the same vein as for OLS, the 2SLS estimator can be shown to be asymptotically normal with the same form of asymptotic covariance, but with  $X_i$  replaced by  $\tilde{X}_i$ :

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{\tilde{X}\tilde{X}}^{-1} \Omega_{\tilde{X}U} \Sigma_{\tilde{X}\tilde{X}}^{-1}) \quad (13.4.121)$$

where  $\Sigma_{\tilde{X}\tilde{X}} := \mathbb{E}[\tilde{X}_i \tilde{X}_i^\top]$  and  $\Omega_{\tilde{X}U} := \text{Cov}(\tilde{X}_i U_i)$ . Each of these can be written out as

$$\Sigma_{\tilde{X}\tilde{X}} = \mathbb{E}[\Pi^\top Z_i Z_i^\top \Pi] \quad (13.4.122)$$

$$= \Pi^\top \Sigma_{ZZ} \Pi \quad (13.4.123)$$

and

$$\Omega_{\tilde{X}U} = \text{Cov}(\Pi^\top Z_i U_i) \quad (13.4.124)$$

$$= \Pi^\top \text{Cov}(Z_i U_i) \Pi \quad (13.4.125)$$

$$= \Pi^\top \Omega_{ZU} \Pi \quad (13.4.126)$$

Hence a ‘long-form’ version of the asymptotic distribution is

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\Pi^\top \Sigma_{ZZ} \Pi\right)^{-1} \Pi^\top \Omega_{ZU} \Pi \left(\Pi^\top \Sigma_{ZZ} \Pi\right)^{-1}\right) \quad (13.4.127)$$

#### Two-Stage Least Squares Standard Errors

The asymptotic normality of 2SLS allows us to approximate the sampling distribution of the 2SLS estimator by

$$\hat{\beta} \xrightarrow{\text{approx.}} \mathcal{N}(\beta, \mathcal{V}) \quad (13.4.128)$$

where

$$\mathcal{V} = \frac{1}{n} \left( \Pi^\top \Sigma_{ZZ} \Pi \right)^{-1} \Pi^\top \Omega_{ZU} \Pi \left( \Pi^\top \Sigma_{ZZ} \Pi \right)^{-1} \quad (13.4.129)$$

To estimate  $\mathcal{V}$  in order to compute standard errors, we take the usual plug-in approach of replacing each term in  $\mathcal{V}$  with their respective sample estimates:

$$\hat{\mathcal{V}} = \frac{1}{n} \left( \hat{\Pi}^\top \hat{\Sigma}_{ZZ} \hat{\Pi} \right)^{-1} \hat{\Pi}^\top \hat{\Omega}_{ZU} \hat{\Pi} \left( \hat{\Pi}^\top \hat{\Sigma}_{ZZ} \hat{\Pi} \right)^{-1} \quad (13.4.130)$$

where  $\widehat{\Pi}$  is of course the matrix of coefficient estimates from the first stage regression. We then take

$$\widehat{\Sigma}_{ZZ} = \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top \quad (13.4.131)$$

and

$$\widehat{\Omega}_{ZU} = \frac{1}{n} \sum_{i=1}^n \widehat{U}_i Z_i Z_i^\top \quad (13.4.132)$$

where the residuals come from the equation  $\widehat{U}_i = Y_i - \widehat{\beta}^\top X_i$ . Note that this is the proper way to compute the residuals, in contrast to using the fitted values  $\widehat{Y}_i - \widehat{\beta}^\top \widehat{X}_i$  from the first stage. To see why, we first assume  $X_i$  satisfies its own causal equation of the form

$$X_i = \Pi^\top Z_i + V_i \quad (13.4.133)$$

with error term  $V_i$ . Thus

$$U_i = Y_i - \beta^\top X_i \quad (13.4.134)$$

$$= Y_i - \beta^\top (\Pi^\top Z_i + V_i) \quad (13.4.135)$$

$$= Y_i - \beta^\top \Pi^\top Z_i - \beta^\top V_i \quad (13.4.136)$$

and then

$$Y_i - \beta^\top \Pi^\top Z_i = U_i + \beta^\top V_i \quad (13.4.137)$$

Hence computing  $Y_i - \widehat{\beta}^\top \widehat{X}_i = Y_i - \widehat{\beta}^\top \widehat{\Pi}^\top Z_i$  emulates the combined error  $U_i + \beta^\top V_i$ , so the estimated covariance will not be correct. Once the residuals are computed the proper way, the standard errors are finally given by the square roots of the diagonal elements of  $\mathcal{V}$ . The standard errors computed this way will also be heteroskedasticity-consistent, since we directly emulated the asymptotic covariance matrix.

### Testing Relevance in Instrumental Variables Regression

The relevance condition  $\mathbb{E}[X_i|Z_i] = \Pi^\top Z_i$  with  $\text{rank}(\Pi) = \dim(X_i)$  can be tested. Letting  $X_i = (1, X_{1,i}, X_{2,i})$  and  $Z_i = (1, Z_{1,i}, X_{2,i})$  with  $X_{1,i}$  being the endogenous variables,  $X_{2,i}$  being the exogenous variables and  $Z_{1,i}$  being the instruments (at least as many endogenous variables), we write out  $\mathbb{E}[X_i|Z_i] = \Pi^\top Z_i$  as

$$\mathbb{E} \left[ \begin{bmatrix} 1 \\ X_{1,i} \\ X_{2,i} \end{bmatrix} \middle| \begin{bmatrix} 1 \\ Z_{1,i} \\ X_{2,i} \end{bmatrix} \right] = \begin{bmatrix} 1 & 0 & 0 \\ \pi_0 & \Pi_1^\top & \Pi_2^\top \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix} \begin{bmatrix} 1 \\ Z_{1,i} \\ X_{2,i} \end{bmatrix} \quad (13.4.138)$$

This in order for  $\Pi$  to be full rank, we want the dimension of the space spanned by the rows of  $\Pi_1^\top$  to be equal to  $\dim(X_i)$ ; the number of endogenous variables. This condition is immediately disqualified if any row in  $\Pi_1^\top$  is made of all zeros, which is to say that no instrument is relevant for that particular endogenous variable. Then we can perform a joint test that all the coefficients in that row are equal to zero. Assume for simplicity that there is only one endogenous variable. If there is more than one endogenous variable, we can apply the test to each endogenous variable individually and use a multiple testing procedure such as Bonferroni's correction. Thus

$$\mathbb{E}[X_{1,i}|Z_{1,i}, X_{2,i}] = \pi_0 + \Pi_1^\top Z_{1,i} + \Pi_2^\top X_{2,i} \quad (13.4.139)$$

and we setup the null and alternative hypothesis as

$$H_0 : \Pi_1 = \mathbf{0} \quad (13.4.140)$$

$$H_A : \Pi_1 \neq \mathbf{0} \quad (13.4.141)$$

A heteroskedasticity-consistent test is the Wald test with test statistic

$$W = \left( R \begin{bmatrix} \widehat{\pi}_0 \\ \widehat{\Pi}_1 \\ \widehat{\Pi}_2 \end{bmatrix} - r \right)^\top \left( R \widehat{\mathcal{V}} R^\top \right)^{-1} \left( R \begin{bmatrix} \widehat{\pi}_0 \\ \widehat{\Pi}_1 \\ \widehat{\Pi}_2 \end{bmatrix} - r \right) \quad (13.4.142)$$

with  $R = [0 \ I \ \mathbf{0}]$  and  $r = \mathbf{0}$ . Here,  $\widehat{\mathcal{V}}$  is any heteroskedasticity-consistent covariance estimator for  $(\widehat{\pi}_0, \widehat{\Pi}_1, \widehat{\Pi}_2)$ . Then  $W \sim \chi_k^2$  where  $k$  is the number of restrictions being tested, which in this case is the number of instruments  $\dim(Z_{1,i})$ . Rejection of the null is evidence in support of relevance.

### Testing Exogeneity in Instrumental Variables Regression

Some evidence for the exogeneity condition  $\mathbb{E}[U_i|Z_i] = 0$  is available in the form of an overidentifying restrictions test. By posing instrumental variables regression as generalised method of moments estimation with moment condition  $\mathbb{E}[Z_i U_i] = \mathbf{0}$ , then using the Law of Iterated Expectations we can see if exogeneity is satisfied, we will have:

$$\mathbb{E}[Z_i U_i] = \mathbb{E}[Z_i \mathbb{E}[U_i|Z_i]] \quad (13.4.143)$$

$$= \mathbf{0} \quad (13.4.144)$$

Thus we can use the  $J$ -statistic to test  $\mathbb{E}[Z_i U_i] = \mathbf{0}$ , as a proxy for  $\mathbb{E}[U_i|Z_i] = 0$ . When we use the weighting matrix  $W$  that makes 2SLS identical to GMM, the  $J$ -statistic is

$$J = \left( \sqrt{n} \widehat{\mathbb{E}}[Z_i U_i] \right)^\top \widehat{\text{Cov}}(Z_i U_i)^{-1} \left( \sqrt{n} \widehat{\mathbb{E}}[Z_i U_i] \right) \quad (13.4.145)$$

$$= \left( \frac{\sqrt{n}}{n} \sum_{i=1}^n Z_i \widehat{U}_i \right)^\top \left( \frac{\widehat{\sigma}_U^2}{n} \sum_{i=1}^n Z_i Z_i^\top \right)^{-1} \left( \frac{\sqrt{n}}{n} \sum_{i=1}^n Z_i \widehat{U}_i \right) \quad (13.4.146)$$

$$= \left( \sum_{i=1}^n Z_i \widehat{U}_i \right)^\top \left( \widehat{\sigma}_U^2 \sum_{i=1}^n Z_i Z_i^\top \right)^{-1} \left( \sum_{i=1}^n Z_i \widehat{U}_i \right) \quad (13.4.147)$$

where  $\widehat{U}_i = Y_i - \widehat{\beta}^\top X_i$  and  $\widehat{\sigma}_U^2$  is the sample variance of the residuals. Note that we can show that this  $J$ -statistic is equal to  $n$  times the multiple  $R^2$  obtained by a regression of  $\widehat{U}_i$  on  $Z_i$  [205]. To show this, introduce  $\widehat{\sigma}_Z$  as the sample standard deviation of  $Z_i$ . Then  $J$  can be turned into

$$J = n \left( \frac{1}{n} \sum_{i=1}^n \frac{Z_i}{\widehat{\sigma}_Z} \frac{\widehat{U}_i}{\widehat{\sigma}_U} \right)^\top \left( \frac{1}{n} \sum_{i=1}^n \frac{Z_i}{\widehat{\sigma}_Z} \frac{Z_i^\top}{\widehat{\sigma}_Z} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \frac{Z_i}{\widehat{\sigma}_Z} \frac{\widehat{U}_i}{\widehat{\sigma}_U} \right) \quad (13.4.148)$$

$$= nR^2 \quad (13.4.149)$$

where  $R^2$  has the same form as the formula for the multiple  $R^2$  pertaining to correlations (up to a factor depending on whether Bessel's correction is used in the standard deviation estimators or not). Note that  $\widehat{U}_i$  should have sample mean of zero, but we implicitly assume  $Z_i$  is also centered to have sample mean zero in order for this to work. This relationship between  $J$  and  $R^2$  makes sense intuitively, because the less  $\widehat{U}_i$  is explained by  $Z_i$ , the smaller  $R^2$  should be, so the null is less likely to be rejected in support of exogeneity.

## 13.5 Panel Data Regression

In panel data, each observation is indexed by an individual (or *entity*) index  $i = 1, \dots, n$  and a time index  $t = 1, \dots, T$ . We may assume the causal equation

$$Y_{i,t} = \alpha_i + \lambda_t + \beta^\top X_{i,t} + V_{i,t} \quad (13.5.1)$$

where  $V_{i,t}$  is the error term, and

- $\alpha_i$  is known as an individual-fixed effect. It is an effect which can be different for each individual, but stays constant for that individual throughout time (so it is also called a time-invariant effect).
- $\lambda_t$  is known as a time-fixed effect. It is an effect which may be different at each point in time, but is the same for all individuals at that point in time. Hence, it is also called an individual-invariant effect.

### 13.5.1 Pooled Ordinary Least Squares

Pooled OLS ideally assumes that the  $\alpha_i$  and  $\lambda_t$  are each the same constant regardless of the index, i.e.  $\alpha_1 = \dots = \alpha_n$  and  $\lambda_1 = \dots = \lambda_T$  (so there is no heterogeneity across individuals or heterogeneity across time), in which case they may be absorbed into the intercept coefficient. Alternatively, we may assume they are random terms (i.e. no longer fixed) which are from some distribution not impacted by  $i, t$  or  $X_{i,t}$ , in which case they can be absorbed into the error term  $V_{i,t}$ . Then, in either case,  $\beta$  may be estimated by pooling the panel data together and running OLS on the entire sample of  $nT$  observations:

$$\hat{\beta} = \left( \sum_{i=1}^n \sum_{t=1}^T X_{i,t} X_{i,t}^\top \right)^{-1} \sum_{i=1}^n \sum_{t=1}^T X_{i,t} Y_{i,t} \quad (13.5.2)$$

### 13.5.2 Fixed Effects Models

#### Individual-Fixed Effects Regression

In individual-fixed effects, we ignore time-fixed effects and write the causal equation as

$$Y_{i,t} = \alpha_i + \beta^\top X_{i,t} + V_{i,t} \quad (13.5.3)$$

With the goal being to estimate  $\beta$ , we take the ‘within transformation’ by subtracting averages across time:

$$\tilde{Y}_{i,t} = Y_{i,t} - \frac{1}{T} \sum_{t=1}^T Y_{i,t} \quad (13.5.4)$$

$$= \beta^\top \left( X_{i,t} - \frac{1}{T} \sum_{t=1}^T X_{i,t} \right) + V_{i,t} - \frac{1}{T} \sum_{t=1}^T V_{i,t} \quad (13.5.5)$$

$$= \beta^\top \tilde{X}_{i,t} + \tilde{V}_{i,t} \quad (13.5.6)$$

where  $\tilde{X}_{i,t} = X_{i,t} - \frac{1}{T} \sum_{t=1}^T X_{i,t}$  and  $\tilde{V}_{i,t} = V_{i,t} - \frac{1}{T} \sum_{t=1}^T V_{i,t}$ . The purpose of this is to subtract out the  $\alpha_i$  term in each observation, so then  $\beta$  may be estimated using OLS on the within-transformed data:

$$\hat{\beta} = \left( \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{i,t} \tilde{X}_{i,t}^\top \right)^{-1} \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{i,t} \tilde{Y}_{i,t} \quad (13.5.7)$$

## Time-Fixed Effects Regression

In time-fixed effects, we ignore individual-fixed effects and write the causal equation as

$$Y_{i,t} = \lambda_t + \beta^\top X_{i,t} + V_{i,t} \quad (13.5.8)$$

To estimate  $\beta$ , an analogous approach to individual-fixed effects can be performed, using the within transformation but averaging across individuals.

$$\tilde{Y}_{i,t} = Y_{i,t} - \frac{1}{n} \sum_{i=1}^n Y_{i,t} \quad (13.5.9)$$

$$= \beta^\top \left( X_{i,t} - \frac{1}{n} \sum_{i=1}^n X_{i,t} \right) + V_{i,t} - \frac{1}{n} \sum_{i=1}^n V_{i,t} \quad (13.5.10)$$

$$= \beta^\top \tilde{X}_{i,t} + \tilde{V}_{i,t} \quad (13.5.11)$$

where  $\tilde{X}_{i,t} = X_{i,t} - \frac{1}{n} \sum_{i=1}^n X_{i,t}$  and  $\tilde{V}_{i,t} = V_{i,t} - \frac{1}{n} \sum_{i=1}^n V_{i,t}$ . Now  $\lambda_i$  is subtracted out, and  $\beta$  can be estimated from the within transformed data by

$$\hat{\beta} = \left( \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{i,t} \tilde{X}_{i,t}^\top \right)^{-1} \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{i,t} \tilde{Y}_{i,t} \quad (13.5.12)$$

## Two-Way Fixed Effects Regression

Consider both forms of fixed effects present in the causal equation

$$Y_{i,t} = \alpha_i + \lambda_t + \beta^\top X_{i,t} + V_{i,t} \quad (13.5.13)$$

In order to estimate  $\beta$ , we can perform a double within transformation. The order of within transformation is arbitrary, but suppose the within transformation for individual-fixed effects is taken first. Then

$$\tilde{Y}_{i,t} = \lambda_t - \bar{\lambda} + \beta^\top \tilde{X}_{i,t} + \tilde{V}_{i,t} \quad (13.5.14)$$

where  $\bar{\lambda} = \frac{1}{T} \sum_{t=1}^T \lambda_t$ ,  $\tilde{X}_{i,t} = X_{i,t} - \frac{1}{T} \sum_{t=1}^T X_{i,t}$  and  $\tilde{V}_{i,t} = V_{i,t} - \frac{1}{T} \sum_{t=1}^T V_{i,t}$ . Then subtracting averages across individuals (which cancels out the  $\lambda_t - \bar{\lambda}$  terms):

$$\tilde{\tilde{Y}}_{i,t} = \tilde{Y}_{i,t} - \frac{1}{n} \sum_{i=1}^n \tilde{Y}_{i,t} \quad (13.5.15)$$

$$= \beta^\top \left( \tilde{X}_{i,t} - \frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,t} \right) + \tilde{V}_{i,t} - \frac{1}{n} \sum_{i=1}^n \tilde{V}_{i,t} \quad (13.5.16)$$

$$= \beta^\top \tilde{\tilde{X}}_{i,t} + \tilde{\tilde{V}}_{i,t} \quad (13.5.17)$$

where  $\tilde{\tilde{X}}_{i,t} = \tilde{X}_{i,t} - \frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,t}$  and  $\tilde{\tilde{V}}_{i,t} = \tilde{V}_{i,t} - \frac{1}{n} \sum_{i=1}^n \tilde{V}_{i,t}$ . Note that the double within transformed explanatory variables can be expressed as

$$\tilde{\tilde{X}}_{i,t} = X_{i,t} - \frac{1}{T} \sum_{t=1}^T X_{i,t} - \frac{1}{n} \sum_{i=1}^n \left( X_{i,t} - \frac{1}{T} \sum_{t=1}^T X_{i,t} \right) \quad (13.5.18)$$

$$= X_{i,t} - \frac{1}{T} \sum_{t=1}^T X_{i,t} - \frac{1}{n} \sum_{i=1}^n X_{i,t} + \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{T} \sum_{t=1}^T X_{i,t} \right) \quad (13.5.19)$$

$$= X_{i,t} - \frac{1}{T} \sum_{t=1}^T X_{i,t} - \frac{1}{n} \sum_{i=1}^n X_{i,t} + \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T X_{i,t} \quad (13.5.20)$$

Similarly for  $\tilde{Y}_{i,t}$ ,

$$\tilde{\tilde{Y}}_{i,t} = X_{i,t} - \frac{1}{T} \sum_{t=1}^T Y_{i,t} - \frac{1}{n} \sum_{i=1}^n Y_{i,t} + \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T Y_{i,t} \quad (13.5.21)$$

We then apply OLS on the double within transformed data to estimate  $\beta$ :

$$\hat{\beta} = \left( \sum_{i=1}^n \sum_{t=1}^T \tilde{\tilde{X}}_{i,t} \tilde{\tilde{X}}_{i,t}^\top \right)^{-1} \sum_{i=1}^n \sum_{t=1}^T \tilde{\tilde{X}}_{i,t} \tilde{\tilde{Y}}_{i,t} \quad (13.5.22)$$

### Dummy Variable Representation of Fixed Effects

An alternative to using the double within transformation is to specify the causal equation using dummy variables:

$$Y_{i,t} = \sum_{j=1}^n \alpha_j \mathbb{I}_{\{j=i\}} + \sum_{s=1}^T \lambda_s \mathbb{I}_{\{s=t\}} + \beta^\top X_{i,t} + V_{i,t} \quad (13.5.23)$$

so that we have included  $n+T$  additional dummy regressors in the equation. Note that in order to avoid the dummy variable trap, an intercept term should not be included in  $X_{i,t}$ . Applying pooled OLS to this model, we then obtain estimates of  $\beta$  as well as of each of the individual-fixed and time-fixed effects  $\alpha_i, \lambda_t$ .

### Additively Separable Variables of Fixed Effects

If an individual-varying and time-varying variable can be written additively in terms of an individual-fixed and time-fixed effect, then it will be eliminated by the double within transformation. As a natural example, suppose there is an explanatory variable  $X_{i,t}$  that is increasing at the same rate for each individual over time. For instance, the variable may represent some quantity that naturally (or by definition) accumulates over time by everybody at the same rate. It can then be shown that the double within transformation eliminates this explanatory variable, and thus it is not possible to estimate its slope coefficient. This is because each  $X_{i,t}$  can be decomposed into

$$X_{i,t} = X_{i,1} + \gamma(t-1) \quad (13.5.24)$$

where  $X_{i,1}$  is the initial value for individual  $i$ , and  $\gamma$  represents the rate of accumulation. Here,  $X_{i,1}$  can be treated as an individual fixed effect, while the term  $\gamma(t-1)$  can be treated as a time fixed effect. Hence they will both be eliminated by the double within transformation. More explicitly, the first within transformation (subtracting time-averages) yields

$$\tilde{X}_{i,t} = X_{i,1} - \frac{1}{T} \sum_{t=1}^T X_{i,1} + \gamma(t-1) - \frac{1}{T} \sum_{t=1}^T \gamma(t-1) \quad (13.5.25)$$

$$= X_{i,1} - X_{i,1} + \gamma(t-1) - \frac{\gamma}{T} \sum_{t=1}^T (t-1) \quad (13.5.26)$$

$$= \gamma(t-1) - \frac{\gamma}{T} \cdot \frac{T(T-1)}{2} \quad (13.5.27)$$

$$= \gamma(t-1) - \frac{\gamma(T-1)}{2} \quad (13.5.28)$$

This quantity only depends on  $t$  (so it can be treated as a time-fixed effect), hence it will be removed by the second within transformation of subtracting individual-averages.

### Clustered Covariance Estimator

Introduce the notation for within-transformed data (either single or double):

$$\underbrace{\begin{bmatrix} \tilde{Y}_{i,1} \\ \vdots \\ \tilde{Y}_{i,T} \end{bmatrix}}_{\tilde{Y}_i} = \underbrace{\begin{bmatrix} \tilde{X}_{i,1}^\top \\ \vdots \\ \tilde{X}_{i,T}^\top \end{bmatrix}}_{\tilde{X}_i^\top} \beta + \underbrace{\begin{bmatrix} \tilde{V}_{i,1} \\ \vdots \\ \tilde{V}_{i,T} \end{bmatrix}}_{\tilde{V}_i} \quad (13.5.29)$$

so that we group all the time observations for a single individual together, and only need to index by individual. Assume that each observation  $(Y_i, X_i)$  is i.i.d. across all individuals. Observe that

$$\sum_{i=1}^n \tilde{X}_i \tilde{X}_i^\top = \sum_{i=1}^n \begin{bmatrix} \tilde{X}_{i,1} & \dots & \tilde{X}_{i,T} \end{bmatrix} \begin{bmatrix} \tilde{X}_{i,1}^\top \\ \vdots \\ \tilde{X}_{i,T}^\top \end{bmatrix} \quad (13.5.30)$$

$$= \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{i,t} \tilde{X}_{i,t}^\top \quad (13.5.31)$$

and

$$\sum_{i=1}^n \tilde{X}_i \tilde{Y}_i = \sum_{i=1}^n \begin{bmatrix} \tilde{X}_{i,1} & \dots & \tilde{X}_{i,T} \end{bmatrix} \begin{bmatrix} \tilde{Y}_{i,1} \\ \vdots \\ \tilde{Y}_{i,T} \end{bmatrix} \quad (13.5.32)$$

$$= \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{i,t} \tilde{Y}_{i,t} \quad (13.5.33)$$

Then pooled OLS on the within-transformed data can be written like the OLS estimator using only individual indices:

$$\hat{\beta} = \left( \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^\top \right)^{-1} \sum_{i=1}^n \tilde{X}_i \tilde{Y}_i \quad (13.5.34)$$

We can follow the same steps as OLS in deriving the asymptotic covariance matrix. We will have

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{XX}^{-1} \Omega_{XV} \Sigma_{XX}^{-1}) \quad (13.5.35)$$

where

$$\Sigma_{XX} = \mathbb{E} [\tilde{X}_i \tilde{X}_i^\top] \quad (13.5.36)$$

$$\begin{aligned} \Omega_{XV} &= \text{Cov} (\tilde{X}_i \tilde{V}_i) \\ &= \mathbb{E} [\tilde{X}_i \tilde{V}_i \tilde{V}_i^\top \tilde{X}_i] \end{aligned} \quad (13.5.37)$$

with  $\mathbb{E} [\tilde{X}_i \tilde{V}_i] = 0$  under an exogeneity and zero-mean error assumption. The usual estimate of  $\Sigma_{XX}$  is given by

$$\widehat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^\top \quad (13.5.38)$$

and after defining residuals  $\widehat{\tilde{V}}_i := \widetilde{Y}_i - \widetilde{X}_i^\top \widehat{\beta}$ , the natural estimate of  $\Omega_{XV}$  is

$$\widehat{\Omega}_{XV} = \frac{1}{n} \sum_{i=1}^n \widetilde{X}_i \widehat{\tilde{V}} \widehat{\tilde{V}}^\top \widetilde{X}_i^\top \quad (13.5.39)$$

Then the heteroskedasticity-consistent covariance estimator of  $\widehat{\beta}$  is

$$\widehat{\Sigma} = \frac{1}{n} \widehat{\Sigma}_{XX}^{-1} \widehat{\Omega}_{XV} \widehat{\Sigma}_{XX} \quad (13.5.40)$$

This is also known as a *clustered covariance estimator*, which will also be autocorrelation-consistent (hence it yields heteroskedasticity and autocorrelation-consistent standard errors). To see why, we contrast it against an *unclustered covariance estimator*, which is when  $\widehat{\Omega}_{XV}$  is computed by

$$\widehat{\Omega}_{XV} = \frac{1}{nT} \sum_{i=1}^n \widehat{\tilde{V}}_{i,t}^2 \widetilde{X}_{i,t} \widetilde{X}_{i,t}^\top \quad (13.5.41)$$

This would be valid if each  $X_{i,t} V_{i,t}$  were i.i.d., because then

$$\text{Cov}(\widetilde{X}_i \widetilde{V}_i) = \text{Cov}\left(\sum_{t=1}^T \widetilde{X}_{i,t} \widetilde{V}_{i,t}\right) \quad (13.5.42)$$

$$= T \text{Cov}(\widetilde{X}_{i,t} \widetilde{V}_{i,t}) \quad (13.5.43)$$

$$= T \mathbb{E}[\widetilde{V}_{i,t}^2 \widetilde{X}_{i,t} \widetilde{X}_{i,t}^\top] \quad (13.5.44)$$

and so the natural estimator for this is

$$\widehat{\Omega}_{XV} = T \cdot \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \widehat{\tilde{V}}_{i,t}^2 \widetilde{X}_{i,t} \widetilde{X}_{i,t}^\top \quad (13.5.45)$$

$$= \frac{1}{nT} \sum_{i=1}^n \widehat{\tilde{V}}_{i,t}^2 \widetilde{X}_{i,t} \widetilde{X}_{i,t}^\top \quad (13.5.46)$$

But if there is autocorrelation in the error terms, meaning that  $\text{Cov}(V_{i,t}, V_{i,s}) \neq 0$  for  $t \neq s$ , then each term  $\widetilde{X}_{i,t} \widehat{\tilde{V}}_{i,t}$  will not be i.i.d. and thus the Weak Law of Large Numbers will not apply to the sum  $\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \widehat{\tilde{V}}_{i,t}^2 \widetilde{X}_{i,t} \widetilde{X}_{i,t}^\top$ . Instead, we ‘cluster’ by individual so each term  $\widetilde{X}_i \widetilde{V}_i$  will be i.i.d. under our assumptions, even when there is autocorrelation in the errors between time. Note that this clustered covariance is suited to a large  $n$  and small  $T$  regime. Analogously, just by swapping the indices in the notation, we can arrive at an estimator that is clustered by time, which is consistent when  $\text{Cov}(V_{i,t}, V_{j,t}) \neq 0$  for  $i \neq j$ , and suitable in a regime of large  $T$  and small  $n$ .

### 13.5.3 Differences-in-Differences Estimation

Differences-in-differences estimation is a special case of **two-way fixed effects**. Each individual  $i$  belongs to a group, denoted  $s(i)$ , which determines the individual-fixed effect. The causal equation is given by

$$Y_{i,t} = \alpha_{s(i)} + \lambda_t + \beta X_{i,t} + V_{i,t} \quad (13.5.47)$$

Moreover, there are two groups  $s \in \{1, 2\}$ , two time periods  $t \in \{1, 2\}$ , and the variable  $X_{i,t}$  is a dummy variable:

$$X_{i,t} = \mathbb{I}_{\{s=2, t=2\}} \quad (13.5.48)$$

In this context, we can view group  $s = 1$  control group, while group  $s = 2$  is the treatment group. The coefficient  $\beta$  is the effect size of the treatment, after one time period has passed. Let the number of individuals in a group be denoted by

$$n_s = \sum_{i=1}^n \mathbb{I}_{\{s(i)=s\}} \quad (13.5.49)$$

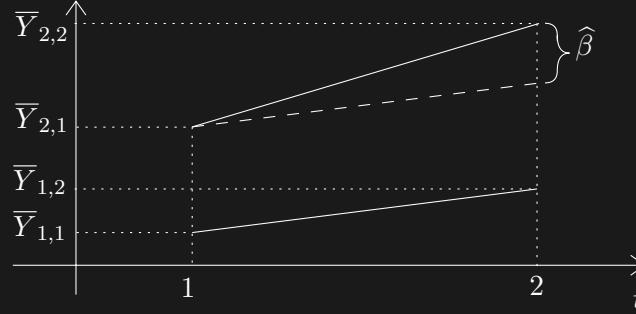
so that the average dependent variable over group  $s$  at time  $t$  is given by

$$\bar{Y}_{s,t} = \frac{1}{n_s} \sum_{i=1}^n Y_{i,t} \mathbb{I}_{\{s(i)=s\}} \quad (13.5.50)$$

We can come up with an intuitive estimator for the effect size as

$$\hat{\beta} = \bar{Y}_{2,2} - [\bar{Y}_{1,2} + (\bar{Y}_{2,1} - \bar{Y}_{1,1})] \quad (13.5.51)$$

$$= (\bar{Y}_{1,1} - \bar{Y}_{1,2}) - (\bar{Y}_{2,1} - \bar{Y}_{2,2}) \quad (13.5.52)$$



It can be demonstrated that this is indeed an appropriate estimator (in an unbiasedness sense). First, we compute the averaged fixed effects over each group:

$$\bar{\alpha}_s := \frac{1}{n_s} \sum_{i=1}^n \alpha_{s(i)} \mathbb{I}_{\{s(i)=s\}} \quad (13.5.53)$$

$$= \frac{\alpha_s}{n_s} \sum_{i:s(i)=s} 1 \quad (13.5.54)$$

$$= \alpha_s \quad (13.5.55)$$

and

$$\bar{\lambda}_{s,t} := \frac{1}{n_s} \sum_{i=1}^n \lambda_t \mathbb{I}_{\{s(i)=s\}} \quad (13.5.56)$$

$$= \frac{\lambda_t}{n_s} \sum_{i=1}^n \mathbb{I}_{\{s(i)=s\}} \quad (13.5.57)$$

$$= \lambda_t \quad (13.5.58)$$

The averaged dummy variable also reduces to itself:

$$D_{s,t} := \frac{1}{n_s} \sum_{i=1}^n X_{i,t} \mathbb{I}_{\{s(i)=s\}} \quad (13.5.59)$$

$$= \frac{1}{n_s} \sum_{i=1}^n \mathbb{I}_{\{s=2, t=2\}} \mathbb{I}_{\{s(i)=s\}} \quad (13.5.60)$$

$$= \frac{\mathbb{I}_{\{s=2,t=2\}}}{n_s} \sum_{i=1}^n \mathbb{I}_{\{s(i)=s\}} \quad (13.5.61)$$

$$= \mathbb{I}_{\{s=2,t=2\}} \quad (13.5.62)$$

Then denoted the averaged error term by

$$\bar{V}_{s,t} := \frac{1}{n_s} \sum_{i=1}^n V_{i,t} \mathbb{I}_{\{s(i)=s\}} \quad (13.5.63)$$

Thus

$$\bar{Y}_{s,t} = \alpha_s + \lambda_t + \beta D_{s,t} + \bar{V}_{s,t} \quad (13.5.64)$$

and the estimator becomes

$$\hat{\beta} = (\bar{Y}_{1,1} - \bar{Y}_{1,2}) - (\bar{Y}_{2,1} - \bar{Y}_{2,2}) \quad (13.5.65)$$

$$= [(\alpha_1 + \lambda_1 + \beta D_{1,1} + \bar{V}_{1,1}) - (\alpha_1 + \lambda_2 + \beta D_{1,2} + \bar{V}_{1,2})] \quad (13.5.66)$$

$$- [(\alpha_2 + \lambda_1 + \beta D_{2,1} + \bar{V}_{2,1}) - (\alpha_2 + \lambda_2 + \beta D_{2,2} + \bar{V}_{2,2})]$$

$$= \beta \left[ \left( \cancel{D_{1,1}}^0 - \cancel{D_{1,2}}^0 \right) - \left( \cancel{D_{2,1}}^0 - \cancel{D_{2,2}}^0 \right) \right] + \bar{V}_{1,1} - \bar{V}_{1,2} + \bar{V}_{2,2} - \bar{V}_{2,1} \quad (13.5.67)$$

$$= \beta + \bar{V}_{1,1} - \bar{V}_{1,2} + \bar{V}_{2,2} - \bar{V}_{2,1} \quad (13.5.68)$$

Under a strict exogeneity condition

$$\mathbb{E}[V_{i,t} | \mathbb{I}_{\{s(1)=s\}}, \dots, \mathbb{I}_{\{s(n)=s\}}] = 0 \quad (13.5.69)$$

for each  $i, s$  and  $t$ , then the difference-in-difference estimator is unbiased:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta} | \mathbb{I}_{\{s(1)=s\}}, \dots, \mathbb{I}_{\{s(n)=s\}}]] \quad (13.5.70)$$

$$= \mathbb{E}[\beta + \mathbb{E}[\bar{V}_{1,1} - \bar{V}_{1,2} + \bar{V}_{2,2} - \bar{V}_{2,1} | \mathbb{I}_{\{s(1)=s\}}, \dots, \mathbb{I}_{\{s(n)=s\}}]] \quad (13.5.71)$$

$$= \beta \quad (13.5.72)$$

### 13.5.4 Seemingly Unrelated Regressions

### 13.5.5 Random Effects Models [14]

In fixed effects regression, the effects  $\alpha_i$  and  $\lambda_t$  were treated as population parameters which could be estimated via the dummy variable approach. However if  $n$  or  $T$  became large, this would lead to a substantial loss in degrees of freedom trying to estimate the  $\alpha_i$  and  $\lambda_t$ . In a random effects model, it is now assumed that  $\alpha_i$  and  $\lambda_t$  are both zero-mean i.i.d. terms, with variances  $\sigma_\alpha^2$  and  $\sigma_\lambda^2$  respectively. Along with existing error term  $V_{i,t}$  with variance  $\sigma_V^2$ , they can be absorbed together to create a new error term

$$U_{i,t} = \alpha_i + \lambda_t + V_{i,t} \quad (13.5.73)$$

with

$$\text{Var}(U_{i,t}) = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_V^2 \quad (13.5.74)$$

Now the focus shifts to estimating  $\sigma_\alpha^2$  and  $\sigma_\lambda^2$ , rather than each of the  $\alpha_i$  and  $\lambda_t$ . An approach to estimate these is by maximum likelihood, for instance. This is also why the random effects model is also referred to as the *variance components model*, because we end up estimating components of the variance of the error term.

### 13.5.6 Mixed Effects Models [14]

In mixed effects models, there are a combination of fixed and random effects. For example, the  $\alpha_i$  could be fixed while the  $\lambda_t$  are random, making the causal model

$$Y_{i,t} = \alpha_i + \beta^\top X_{i,t} + W_{i,t} \quad (13.5.75)$$

where  $W_{i,t} = \lambda_t + V_{i,t}$ . Or, the  $\lambda_t$  could be fixed while the  $\alpha_i$  are random, giving the causal equation

$$Y_{i,t} = \lambda_t + \beta^\top X_{i,t} + U_{i,t} \quad (13.5.76)$$

where  $U_{i,t} = \alpha_i + V_{i,t}$ . Modelling using mixed effects may be a viable option when either of  $n$  or  $T$  (but not both) is large, meaning the smaller set of effects can be modelled with fixed effects, while the larger set of effects is modelled by random effects.

### Linear Mixed Models [64]

Mixed effects models can be rearranged into a more general form, sometimes known as linear mixed models. Recall the dummy variable representation of the causal equation as

$$Y_{i,t} = \sum_{j=1}^n \alpha_j \mathbb{I}_{\{j=i\}} + \beta^\top X_{i,t} + \sum_{s=1}^T \lambda_s \mathbb{I}_{\{s=t\}} + V_{i,t} \quad (13.5.77)$$

$$= \mathcal{I}_{\alpha,i}^\top \boldsymbol{\alpha} + X_{i,t}^\top \beta + \mathcal{I}_{\lambda,t}^\top \boldsymbol{\lambda} + V_{i,t} \quad (13.5.78)$$

where  $\boldsymbol{\alpha}$  is the vector of  $\alpha_1, \dots, \alpha_n$  and  $\boldsymbol{\lambda}$  is the vector of  $\lambda_1, \dots, \lambda_n$ . Also,  $\mathcal{I}_{\alpha,i}$  is the vector of dummies  $\mathbb{I}_{\{1=i\}}, \dots, \mathbb{I}_{\{n=i\}}$  and likewise  $\mathcal{I}_{\lambda,t}$  is the vector of dummies  $\mathbb{I}_{\{1=t\}}, \dots, \mathbb{I}_{\{T=t\}}$ . Without loss of generality, suppose the  $\alpha_i$  are fixed and  $\lambda_t$  are random. Then we can group together all the fixed parameters:

$$Y_{i,t} = [\mathcal{I}_{\alpha,i}^\top \quad X_{i,t}^\top] \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} \gamma + \mathcal{I}_{\lambda,t}^\top \boldsymbol{\lambda} + V_{i,t} \quad (13.5.79)$$

Then stacking all observations:

$$\underbrace{\begin{bmatrix} Y_{1,1} \\ \vdots \\ Y_{1,T} \\ \hline Y_{n,1} \\ \vdots \\ Y_{n,T} \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \mathcal{I}_{\alpha,1}^\top & X_{1,1}^\top \\ \vdots & \vdots \\ \mathcal{I}_{\alpha,1}^\top & X_{1,T}^\top \\ \vdots & \vdots \\ \mathcal{I}_{\alpha,n}^\top & X_{n,1}^\top \\ \vdots & \vdots \\ \mathcal{I}_{\alpha,n}^\top & X_{n,T}^\top \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}}_{\boldsymbol{\gamma}} + \underbrace{\begin{bmatrix} \mathcal{I}_{\lambda,1}^\top \\ \vdots \\ \mathcal{I}_{\lambda,T}^\top \\ \hline \mathcal{I}_{\lambda,1}^\top \\ \vdots \\ \mathcal{I}_{\lambda,T}^\top \end{bmatrix}}_{\mathbf{Z}} \boldsymbol{\lambda} + \underbrace{\begin{bmatrix} V_{1,1} \\ \vdots \\ V_{1,T} \\ \hline V_{n,1} \\ \vdots \\ V_{n,T} \end{bmatrix}}_{\mathbf{v}} \quad (13.5.80)$$

This gives the general matrix form  $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\lambda} + \mathbf{v}$  where  $\boldsymbol{\gamma}$  is treated as all the fixed effects,  $\boldsymbol{\lambda}$  is treated as the random effects, the design matrices are  $\mathbf{X}$  and  $\mathbf{Z}$ , and the error term is  $\mathbf{v}$ .

## 13.6 Time-Series Models

### 13.6.1 Autoregressive (AR) Models

An autoregressive model of order  $p$  is denoted AR( $p$ ) and is defined by

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t \quad (13.6.1)$$

where  $\varphi_1, \dots, \varphi_p$  are parameters of the model,  $c$  is a constant term and  $\varepsilon_t$  is zero-mean white noise.

## MA Representation of AR Models

Consider an AR(1) process with zero constant term:

$$X_t = \varphi X_{t-1} + \varepsilon_t \quad (13.6.2)$$

Recursion backwards in time yields the representation

$$X_t = \varphi(\varphi X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \quad (13.6.3)$$

$$= \varphi^2 X_{t-1} + \varphi \varepsilon_{t-1} + \varepsilon_t \quad (13.6.4)$$

$$= \varphi^k X_{t-k} + \sum_{j=0}^{k-1} \varphi^j \varepsilon_{t-j} \quad (13.6.5)$$

Letting  $k \rightarrow \infty$ , if  $|\varphi| < 1$ , we see that

$$X_t = \sum_{j=0}^{\infty} \varphi^j \varepsilon_{t-j} \quad (13.6.6)$$

Hence if  $|\varphi| < 1$ , an AR(1) process may be represented as an MA( $\infty$ ) process.

Now consider a general AR( $p$ ) process, which we write as

$$X_t + \sum_{i=1}^p \phi_i X_{t-i} = c + \varepsilon_t \quad (13.6.7)$$

where  $\phi_i = -\varphi_i$  in the original notation. Introduce  $L$  as the lag operator (i.e.  $L^i X_t = X_{t-i}$ ), then

$$(1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p) X_t = c + \varepsilon_t \quad (13.6.8)$$

Note that this is analogous to taking a  $z$ -transform but with  $L = z^{-1}$ . Since we can treat  $z$  as a complex number on the unit circle (hence having modulus one), it follows we can also treat  $L$  as a complex number with modulus one. Consider the inverse of the polynomial in  $L$ :

$$(1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p)^{-1} = \frac{1}{1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p} \quad (13.6.9)$$

$$= \frac{1}{(a_1 - L) \times \cdots \times (a_p - L)} \quad (13.6.10)$$

$$= \frac{1}{a_1 - L} \times \cdots \times \frac{1}{a_p - L} \quad (13.6.11)$$

where  $a_1, \dots, a_p$  are the complex roots of the characteristic polynomial. Suppose that each  $|a_j| > 1$ , then for each  $j$  we have

$$\frac{1}{a_j - L} = \frac{1}{a_j} \cdot \frac{1}{1 - L/a_j} \quad (13.6.12)$$

$$= \frac{1}{a_j} \sum_{i=0}^{\infty} \left( \frac{L}{a_j} \right)^i \quad (13.6.13)$$

by the geometric series, since  $\left| \frac{L}{a_j} \right| < 1$ . Thus the AR( $p$ ) model can be expressed as

$$X_t = (c + \varepsilon_t) \left( \frac{1}{a_1} \sum_{i=0}^{\infty} \left( \frac{L}{a_1} \right)^i \right) \times \cdots \times \left( \frac{1}{a_p} \sum_{i=0}^{\infty} \left( \frac{L}{a_p} \right)^i \right) \quad (13.6.14)$$

which is another MA( $\infty$ ) process.

## Stationarity of AR Models

Suppose that  $\text{AR}(p)$ , the roots of the characteristic polynomial, i.e. the solutions to

$$1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p = 0 \quad (13.6.15)$$

are  $L = a_j$  such that  $|a_j| > 1$  for all  $j$ . If we instead took a  $z$ -transform of the process to obtain the transfer function, we would instead be looking at solutions  $z^{-1} = a_j$ , hence the poles would be  $z = 1/a_j$  with each  $|1/a_j| < 1$ . This satisfies the stability condition, hence the impulse response (treating  $\varepsilon_t$  as an input) would be absolutely summable. It follows that if the roots of  $1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p$  are outside the unit circle and  $\varepsilon_t$  is white noise (hence weakly stationary), then the process  $X_t$  will also be weakly stationary.

## Causality of AR Models

Consider an *explosive*  $\text{AR}(p)$  process, where all the roots of the characteristic polynomial are inside the unit circle, i.e.  $|a_j| < 1$  for all  $j$  (hence all poles in the transfer function are outside the unit circle). Such a process, when simulated, would ‘blow up’ over time. However, it is possible to find a stationary solution to the process. In terms of the lag operator, an  $\text{AR}(p)$  process can be written like

$$X_t = (1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p)^{-1} (c + \varepsilon_t) \quad (13.6.16)$$

Note that the roots of  $1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p$  are also the roots of

$$\frac{1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p}{L^P} = \phi_p + \phi_{p-1} L + \cdots + \phi_1 L^{-p+1} + L^{-p} \quad (13.6.17)$$

Hence we can factorise this rational function as

$$\phi_p + \phi_{p-1} L + \cdots + \phi_1 L^{-p+1} + L^{-p} = (a_1^{-1} - L^{-1}) \times \cdots \times (a_p^{-1} - L^{-1}) \quad (13.6.18)$$

since if  $a_j - L = 0$ , then also  $\frac{1}{a_j} - \frac{1}{L} = 0$  (assuming  $a_j \neq 0$ ). Thus if each  $|a_j| < 1$ , then each  $|a_j^{-1}| > 1$ , and in the same vein as above, we have

$$(1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p)^{-1} = \frac{L^p}{\phi_p + \phi_{p-1} L + \cdots + \phi_1 L^{-p+1} + L^{-p}} \quad (13.6.19)$$

$$= \frac{L}{a_1^{-1} - L^{-1}} \times \cdots \times \frac{L}{a_p^{-1} - L^{-1}} \quad (13.6.20)$$

where each factor becomes

$$\frac{L}{a_j^{-1} - L^{-1}} = \frac{L}{a_j} \sum_{i=0}^{\infty} \left( \frac{L^{-1}}{a_j^{-1}} \right)^i \quad (13.6.21)$$

$$= \sum_{i=-1}^{\infty} \left( \frac{L^{-1}}{a_j^{-1}} \right)^i \quad (13.6.22)$$

And we can express the  $\text{AR}(p)$  process as

$$X_t = (c + \varepsilon_t) \left( \sum_{i=-1}^{\infty} \left( \frac{L^{-1}}{a_1^{-1}} \right)^i \right) \times \cdots \times \left( \sum_{i=-1}^{\infty} \left( \frac{L^{-1}}{a_p^{-1}} \right)^i \right) \quad (13.6.23)$$

This process is now weakly stationary, as the sums are convergent and  $\varepsilon_t$  is a white noise process. Intuitively, what we have done is that since an explosive process ‘blows up’ in time, we considered the process in reverse time, which ‘shrinks down’ and behaves like a weakly stationary process. However, the model is now longer *causal*, since the value of the process now depends on future values of the noise  $\varepsilon_t$ . We conclude that the roots of the characteristic polynomial being outside the unit circle is not a necessary condition for the process to be weakly stationary, but it is a necessary and sufficient condition for the process to be causal and weakly stationary (i.e. stable).

### Autocovariance Function of AR Processes [77]

Taking expectations of the canonical AR ( $p$ ) form yields

$$\mathbb{E}[X_t] = c + \sum_{i=1}^p \varphi_i \mathbb{E}[X_{t-i}] \quad (13.6.24)$$

If the process is weakly stationary, then the mean is some constant  $\mathbb{E}[X_t]$  for all  $t$ . Hence

$$c = \mu - \varphi_1\mu - \cdots - \varphi_p\mu \quad (13.6.25)$$

Substituting this into the AR ( $p$ ) form gives

$$X_t - \mu = \varphi_1(X_{t-1} - \mu) + \cdots + \varphi_p(X_{t-p} - \mu) + \varepsilon_t \quad (13.6.26)$$

Multiplying each side by  $X_{t-j} - \mu$  and taking expectations, we see

$$\begin{aligned} & \mathbb{E}[(X_t - \mu)(X_{t-j} - \mu)] \\ &= \varphi_1 \mathbb{E}[(X_{t-1} - \mu)(X_{t-j} - \mu)] + \cdots + \mathbb{E}[(X_{t-p} - \mu)(X_{t-j} - \mu)] + \mathbb{E}[\varepsilon_t(X_{t-j} - \mu)] \end{aligned} \quad (13.6.27)$$

Note that for  $j = 0$ :

$$\mathbb{E}[\varepsilon_t(X_{t-j} - \mu)] = \mathbb{E}[\varepsilon_t(\varphi_1(X_{t-1} - \mu) + \cdots + \varphi_p(X_{t-p} - \mu) + \varepsilon_t)] \quad (13.6.28)$$

$$= \mathbb{E}[\varepsilon_t^2] \quad (13.6.29)$$

$$= \sigma^2 \quad (13.6.30)$$

where  $\sigma^2$  is the variance of the zero-mean white noise  $\varepsilon_t$ . Whereas for  $j \neq 0$ :

$$\mathbb{E}[\varepsilon_t(X_{t-j} - \mu)] = \mathbb{E}[\varepsilon_t \varepsilon_{t-j}] \quad (13.6.31)$$

$$= 0 \quad (13.6.32)$$

Let  $\gamma_j := \mathbb{E}[(X_t - \mu)(X_{t-j} - \mu)]$  denote the autocovariance function (recalling that for a weakly stationary process, the autocovariance can be written just in terms of the time difference). Using the above expression for  $\mathbb{E}[(X_t - \mu)(X_{t-j} - \mu)]$ , the autocovariance function can be written recursively as

$$\gamma_j = \sum_{i=1}^p \varphi_i \gamma_{j-1} + \sigma^2 \delta_j \quad (13.6.33)$$

for  $j = 0, \dots, p$ , where  $\delta_j := \mathbb{I}_{\{j=0\}}$  is the Kronecker delta. Recall the property that  $\gamma_{-j} = \gamma_j$  for weakly stationary processes. Since  $\gamma_j$  can be expressed as a difference equation (with same coefficients as for the stable process  $X_t$ ), then  $\gamma_j$  is anticipated to decay with  $j$ .

### Autocovariance Function of AR(1) Processes

Using the recursive expression for the autocovariance function  $\gamma_j$  of an AR( $p$ ) process, we obtain a direct expression for the autocovariance function of an AR(1) process  $X_t = c + \varphi_1 X_{t-1} + \varepsilon_t$ , and assume that  $|\varphi_1| < 1$  so that the process is stationary. Taking the variance of both sides gives

$$\gamma_0 = \text{Var}(X_t) \quad (13.6.34)$$

$$= \varphi_1^2 \text{Var}(X_{t-1}) + \text{Var}(\varepsilon_t) \quad (13.6.35)$$

$$= \varphi_1^2 \gamma_0 + \sigma^2 \quad (13.6.36)$$

where  $\text{Var}(\varepsilon_t) = \sigma^2$ , and  $\text{Var}(X_t) = \text{Var}(X_{t-1})$  because of stationarity. Rearranging yields the variance function

$$\gamma_0 (1 - \varphi_1^2) = \sigma^2 \quad (13.6.37)$$

$$\gamma_0 = \frac{\sigma^2}{1 - \varphi_1^2} \quad (13.6.38)$$

Then applying the recursive form of the autocovariance function, we have  $\gamma_j = \varphi_1 \gamma_{j-1}$  thus

$$\gamma_j = \text{Cov}(X_t, X_{t-j}) \quad (13.6.39)$$

$$= \frac{\varphi_1^j \sigma^2}{1 - \varphi_1^2} \quad (13.6.40)$$

### Autocorrelation Coefficient of AR Processes

The autocorrelation coefficient  $\rho_j$  of a weakly stationary AR process is the autocovariance divided by the variance, i.e.

$$\rho_j = \frac{\gamma_j}{\gamma_0} \quad (13.6.41)$$

An alternative representation for  $\rho_j$  can be derived. Consider the conditional expectation of an AR( $p$ ) process:

$$\mathbb{E}[X_t | X_{t-j}] = \mu + \rho_j X_{t-j} \quad (13.6.42)$$

with  $j \leq p$ . Referring to the population version of the simple least squares estimator, we know that

$$\rho_j = \frac{\text{Cov}(X_t, X_{t-j})}{\text{Var}(X_{t-j})} \quad (13.6.43)$$

which gives back the definition of the autocorrelation coefficient, since  $\text{Var}(X_{t-j}) = \text{Var}(X_t)$ . Hence  $\rho_j$  can be viewed as the coefficient on  $X_{t-j}$  when  $X_t$  is regressed on  $X_{t-j}$ ; this also reveals an approach for estimating  $\rho_j$ .

### Partial Autocorrelation Function of AR Processes

A representation for the partial autocorrelation function  $\varrho_j$  of a weakly stationary AR process can be derived. For an arbitrary AR( $p$ ) process, consider the conditional expectation:

$$\mathbb{E}[X_t | X_{i-1}, \dots, X_{t-j}] = \mu + \varrho_1 X_{t-1} + \dots + \varrho_j X_{t-j} \quad (13.6.44)$$

with  $j \leq p$ . Suppose we then condition both sides on  $X_{i-1}, \dots, X_{t-j}$ , then the left-hand side looks like a conditional expectation conditioned on  $X_{t-j}$  (in addition to  $X_{i-1}, \dots, X_{t-j}$ ). In the right-hand side, each of the terms  $X_{i-1}, \dots, X_{t-j}$  can be treated as constants, while  $X_{t-j}$  is conditioned on  $X_{i-1}, \dots, X_{t-j}$ . Overall, it now appears like a simple linear PRF in  $X_{t-j}$ , all

conditioned on  $X_{i-1}, \dots, X_{t-j}$ . Then the formula for the coefficient  $\varrho_j$  in the simple linear PRF may be written in terms of conditional variance/covariances:

$$\varrho_j = \frac{\text{Cov}(X_t, X_{t-j}|X_{i-1}, \dots, X_{t-j})}{\text{Var}(X_{t-j}|X_{i-1}, \dots, X_{t-j})} \quad (13.6.45)$$

Next, we show the conditional variance of  $X_{t-j}$  is equal to that of  $X_t$  as follows. By weak stationarity, the covariance matrix of  $(Y_t, \dots, Y_{t-j})$  is equivalent to:

$$\text{Cov} \left( \begin{bmatrix} X_t \\ \vdots \\ X_{t-j} \end{bmatrix} \right) = \begin{bmatrix} \gamma_0 & \dots & \gamma_j \\ \vdots & \ddots & \vdots \\ \gamma_j & \dots & \gamma_0 \end{bmatrix} \quad (13.6.46)$$

$$= \text{Cov} \left( \begin{bmatrix} X_{t-j} \\ \vdots \\ X_t \end{bmatrix} \right) \quad (13.6.47)$$

Thus in the sense of the best linear approximation for the partial correlation, we have

$$\text{Cov} \left( \begin{bmatrix} X_t \\ X_{t-j} \end{bmatrix} \middle| X_{i-1}, \dots, X_{t-j} \right) = \text{Cov} \left( \begin{bmatrix} X_{t-j} \\ X_t \end{bmatrix} \middle| X_{i-1}, \dots, X_{t-j} \right) \quad (13.6.48)$$

which implies

$$\text{Var}(X_{t-j}|X_{i-1}, \dots, X_{t-j}) = \text{Var}(X_t|X_{i-1}, \dots, X_{t-j}) \quad (13.6.49)$$

Hence

$$\varrho_j = \frac{\text{Cov}(X_t, X_{t-j}|X_{i-1}, \dots, X_{t-j})}{\sqrt{\text{Var}(X_t|X_{i-1}, \dots, X_{t-j}) \text{Var}(X_{t-j}|X_{i-1}, \dots, X_{t-j})}} \quad (13.6.50)$$

which is the conditional correlation between  $X_t$  and  $X_{t-j}$  given  $X_{i-1}, \dots, X_{t-j}$ , thereby fulfilling the definition of the partial autocorrelation function. This means we can also estimate the partial autocorrelation function at lag  $j$  by regressing  $Y_t$  on  $Y_{t-1}, \dots, Y_{t-j}$  and then taking the estimate of the coefficient  $\varrho_j$  in front of  $Y_{t-j}$ .

### 13.6.2 Autoregressive Distributed Lag (ARDL) Models

The ARDL model can also be referred to as an autoregressive exogenous (ARX) model, and generalise AR models. An ARDL model of order  $p$  and  $b$  is denoted by ARDL( $p, b$ ), and specified by

$$X_t = \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^b \eta_i Y_{t-i} + \varepsilon_t \quad (13.6.51)$$

where  $\varepsilon_t$  is zero-mean white noise and  $Y_t$  is called an exogenous input sequence. We can also allows for lags of other exogenous series in the model (such as by introducing  $Z_t$  and its lags into the specification).

### 13.6.3 Moving Average (MA) Models

A moving average model of order  $q$  is denoted MA( $q$ ) and is defined by

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (13.6.52)$$

where  $\mu$  is the mean of the series,  $\theta_1, \dots, \theta_q$  are the parameters of the model and  $\varepsilon_t, \dots, \varepsilon_{t-q}$  are white noise error terms. With  $\mu = 0$ , an MA process is essentially a filtered (i.e. ‘smoothed’) white noise process.

## Invertibility of MA Models

If we simply swap  $X_t$  and  $\varepsilon_t$  in an MA process, notice that it then looks like an AR process (but in variable  $\varepsilon_t$  rather than  $X_t$ ). Therefore, using the same techniques for representing an AR process as an MA( $\infty$ ) process, it can be shown that any MA( $q$ ) process may be inverted into a AR( $\infty$ ) process, as long as the roots of the characteristic polynomial  $1 + \theta_1 L + \dots + \theta_q L^q$  are each outside the unit circle.

### Autocovariance Function of MA Processes [183]

The autocovariance function  $\gamma_j$  of an MA( $q$ ) process can be computed as follows.

$$\gamma_j = \text{Cov}(X_t, X_{t-j}) \quad (13.6.53)$$

$$= \text{Cov}(X_t - \mu, X_{t-j} - \mu) \quad (13.6.54)$$

$$= \text{Cov}\left(\sum_{i=0}^q \theta_i \varepsilon_{t-i}, \sum_{k=0}^q \theta_k \varepsilon_{t-j-k}\right) \quad (13.6.55)$$

$$= \sum_{i=0}^q \sum_{k=0}^q \theta_i \theta_k \text{Cov}(\varepsilon_{t-i}, \varepsilon_{t-j-k}) \quad (13.6.56)$$

where we let  $\theta_0 = 1$ . Since  $\varepsilon_t$  is white noise, then  $\text{Cov}(\varepsilon_{t-i}, \varepsilon_{t-j-k}) = \sigma^2$  when  $i = k + j$ , and zero otherwise. In the double summation, we see that the only non-zero terms produced will involve  $\varepsilon_{t-j}, \dots, \varepsilon_{t-q}$  (i.e. from  $k = 0, \dots, q - j$ ). Hence we can write the autocovariance function as

$$\gamma_j = \sigma^2 \sum_{k=0}^{q-j} \theta_k \theta_{k+j} \quad (13.6.57)$$

and so clearly,  $\gamma_j = 0$  when  $j > q$ .

### 13.6.4 Autoregressive Moving Average (ARMA) Models

An ARMA model is a generalisation of the autoregressive and moving average models, and is denoted by ARMA( $p, q$ ). The model is defined by

$$X_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (13.6.58)$$

The process can be split into

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + U_t \quad (13.6.59)$$

where  $U_t = \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$  is a zero-mean MA process. Hence an ARMA process can be interpreted as an AR process except the error terms are autocorrelated, and defined by a filtered white noise process.

### Invertibility of ARMA Models

if the roots of the characteristic polynomial of the MA part of an ARMA( $p, q$ ) model are outside the unit circle (i.e. invertible MA process), we know that the MA( $q$ ) part can be inverted into an AR( $\infty$ ) process. When combined with the existing AR( $p$ ) process, it gives an AR( $\infty$ ) process of the form

$$c' + \sum_{i=0}^{\infty} \pi_i X_{t-i} = \varepsilon_t \quad (13.6.60)$$

The relation of the coefficients  $\pi_i$  to  $\varphi_i$  and  $\theta_i$  can be determined by writing the ARMA process in lag polynomial form:

$$\varphi(L) X_t = c + \theta(L) \varepsilon_t \quad (13.6.61)$$

where

$$\varphi(L) = 1 - \varphi_1 L - \dots - \varphi_p L^p \quad (13.6.62)$$

$$\theta(L) = 1 + \theta_1 L + \dots + \theta_q L^q \quad (13.6.63)$$

Rearranging, we see

$$\frac{\varphi(L) X_t}{\theta(L)} - \frac{c}{\theta(L)} = \varepsilon_t \quad (13.6.64)$$

Equating coefficients:

$$\frac{\varphi(L)}{\theta(L)} X_t - \frac{c}{\theta(L)} \equiv c' + \pi(L) X_t \quad (13.6.65)$$

and so we require equating the coefficients of the left-hand side to the right-hand side in order to determine  $\pi_0, \pi_1, \dots$  in the polynomial  $\pi(L)$ .

### MA Representation of ARMA Models [183]

Analogous to invertibility of ARMA models, if the roots of the characteristic polynomial of the AR part of an ARMA  $(p, q)$  model are outside the unit circle (i.e. weakly stationary and causal AR process), we know that the AR  $(p)$  component can be represented as an MA  $(\infty)$  process. When added to the existing MA  $(q)$  process, it gives another MA  $(\infty)$  process. Thus an ARMA  $(p, q)$  process with stationary AR  $(p)$  part can be written as an MA  $(\infty)$  process:

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} \quad (13.6.66)$$

In a similar way to inverting an ARMA process, the coefficients  $\psi_i$  have to be determined by equating coefficients in

$$\frac{c}{\varphi(L)} + \frac{\theta(L)}{\varphi(L)} \varepsilon_t \equiv \mu + \psi(L) \varepsilon_t \quad (13.6.67)$$

Focusing on just the coefficients of  $\varepsilon_t$  and its lags (i.e. ignoring the constant term), we can equate

$$(1 + \theta_1 L + \theta_2 L^2 + \theta_3 L^3 + \dots) = (1 - \varphi_1 L - \varphi_2 L^2 - \varphi_3 L^3 - \dots) (\psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \dots) \quad (13.6.68)$$

We can see for the first few orders that

$$1 = \psi_0 \quad (13.6.69)$$

$$\theta_1 = \psi_1 - \varphi_1 \psi_0 \quad (13.6.70)$$

$$\theta_2 = \psi_2 - \varphi_1 \psi_1 - \varphi_2 \psi_0 \quad (13.6.71)$$

$$\theta_3 = \psi_3 - \varphi_1 \psi_2 - \varphi_2 \psi_1 - \varphi_3 \psi_0 \quad (13.6.72)$$

$$\vdots \quad (13.6.73)$$

Generalising this pattern, we can write

$$\psi_i - \sum_{k=1}^i \varphi_k \psi_{i-k} = \theta_i \quad (13.6.74)$$

Note that in this notation, we have gone beyond  $\varphi_p$  and  $\theta_q$ , but we can just take  $\varphi_i = 0$  for  $i > p$  and  $\theta_i = 0$  for  $i > q$ . In that case, for  $i \geq p$ , the sum can be terminated at  $p$ , and moreover if  $i > q$ , then the  $\psi_i$  coefficients can be written as the solution to a linear homogeneous difference equation satisfying

$$\psi_i - \sum_{k=1}^p \varphi_k \psi_{i-k} = 0 \quad (13.6.75)$$

for  $i > \max\{p-1, q\}$ , with the initial conditions satisfying

$$\psi_i - \sum_{k=1}^i \varphi_k \psi_{i-k} = \theta_i \quad (13.6.76)$$

for  $0 \leq i \leq \max\{p-1, q\}$ .

### Autocovariance Function of ARMA Processes

Using the MA  $(\infty)$  representation of a causal ARMA  $(p, q)$  process, we can express the autocovariance function as

$$\gamma_j = \text{Cov}(X_t, X_{t-j}) \quad (13.6.77)$$

$$= \text{Cov}(X_t - \mu, X_{t-j} - \mu) \quad (13.6.78)$$

$$= \text{Cov}\left(\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-j-k}\right) \quad (13.6.79)$$

$$= \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \psi_i \psi_k \text{Cov}(\varepsilon_{t-i}, \varepsilon_{t-j-k}) \quad (13.6.80)$$

In the same way as for the autocovariance function of MA processes, this becomes

$$\gamma_j = \sigma^2 \sum_{k=0}^{\infty} \psi_k \psi_{k+j} \quad (13.6.81)$$

which are directly in terms of the coefficients  $\psi_i$  and the error variance  $\sigma^2$ . Alternatively, using the canonical representation, we can show (with  $j \geq 0$  without loss of generality):

$$\gamma_j = \text{Cov}(X_t, X_{t+j}) \quad (13.6.82)$$

$$= \text{Cov}\left(X_t, c + \sum_{i=1}^p \varphi_i X_{t-i+j} + \sum_{i=0}^q \theta_i \varepsilon_{t-i+j}\right) \quad (13.6.83)$$

$$= \sum_{i=1}^p \varphi_i \text{Cov}(X_t, X_{t-i+j}) + \sum_{i=0}^q \theta_i \text{Cov}(X_t, \varepsilon_{t-i+j}) \quad (13.6.84)$$

$$= \sum_{i=1}^p \varphi_i \gamma_{j-i} + \sum_{i=0}^q \theta_i \text{Cov}\left(\sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}, \varepsilon_{t-i+j}\right) \quad (13.6.85)$$

$$= \sum_{i=1}^p \varphi_i \gamma_{j-i} + \sigma^2 \sum_{i=0}^q \theta_i \psi_{i-j} \quad (13.6.86)$$

$$= \sum_{i=1}^p \varphi_i \gamma_{j-i} + \sigma^2 \sum_{i=j}^q \theta_i \psi_{i-j} \quad (13.6.87)$$

since we can take  $\psi_{i-j} = 0$  when  $i < j$ . Note that if  $j > q$  then  $\sum_{i=j}^q \theta_i \psi_{i-j} = 0$ . Hence the autocovariance function can be expressed as the solution to a linear homogeneous difference equation of the form

$$\gamma_j - \sum_{i=1}^p \varphi_i \gamma_{j-i} = 0 \quad (13.6.88)$$

for  $j > q$ , with initial conditions

$$\gamma_j - \sum_{i=1}^p \varphi_i \gamma_{j-i} = \sigma^2 \sum_{i=j}^q \theta_i \psi_{i-j} \quad (13.6.89)$$

for  $0 \leq j \leq q$ .

### Long-Run Variance of ARMA Models [222]

Let  $X_t$  be a stationary ARMA model with stationary mean  $\mu$ . Consider the sample mean

$$\bar{X}_T = \frac{1}{T} \sum_{t=1}^T X_t \quad (13.6.90)$$

This is an unbiased estimator for  $\mu$ , i.e.  $\mathbb{E}[\bar{X}_T] = \mu$ . If the autocovariances  $\gamma_j$  satisfy  $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ , then by the law of large numbers for correlated sequences, we have  $\bar{X}_T \xrightarrow{P} \mu$ . Now consider the variance of the sampling distribution:

$$\text{Var}(\bar{X}_T) = \mathbb{E}[(\bar{X}_T - \mu)^2] \quad (13.6.91)$$

$$= \frac{1}{T^2} \mathbb{E} \left[ \left( \sum_{t=1}^T X_t - T\mu \right)^2 \right] \quad (13.6.92)$$

$$= \frac{1}{T^2} \mathbb{E} \left[ \left( \sum_{t=1}^T (X_t - \mu) \right)^2 \right] \quad (13.6.93)$$

$$= \frac{1}{T^2} \mathbb{E} \left[ \sum_{t=1}^T \sum_{s=1}^T (X_t - \mu)(X_s - \mu) \right] \quad (13.6.94)$$

$$= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbb{E}[(X_t - \mu)(X_s - \mu)] \quad (13.6.95)$$

Note that these summands are the autocovariances, i.e.  $\mathbb{E}[(X_t - \mu)(X_s - \mu)] = \gamma_{t-s}$ . By thinking of a  $T \times T$  array containing these summands, we can count along the diagonals (remembering  $\gamma_{-j} = \gamma_j$  because of stationarity), and find:

$$\frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbb{E}[(X_t - \mu)(X_s - \mu)] = \frac{1}{T^2} [T\gamma_0 + 2(T-1)\gamma_1 + 2(T-2)\gamma_2 + \cdots + 2\gamma_{T-1}] \quad (13.6.96)$$

$$= \frac{1}{T} \left[ \gamma_0 + \frac{T-1}{T}(2\gamma_1) + \frac{T-2}{T}(2\gamma_2) + \cdots + \frac{1}{T}(2\gamma_{T-1}) \right] \quad (13.6.97)$$

Consider what happens as  $T \rightarrow \infty$ . The earlier terms inside the square brackets with small  $j$  will tend to  $2\gamma_j$ , while the later terms with large  $j$  will tend to zero. However due to the

condition  $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ , we would have  $\gamma_j \rightarrow 0$  as  $j \rightarrow \infty$ . Intuitively, this makes it ‘safe’ to ignore the later terms in the limit, and we are able to write [77]:

$$\lim_{T \rightarrow \infty} \left[ \gamma_0 + \frac{T-1}{T} (2\gamma_1) + \frac{T-2}{T} (2\gamma_2) + \cdots + \frac{1}{T} (2\gamma_{T-1}) \right] = \gamma_0 + 2\gamma_1 + 2\gamma_2 + \dots \quad (13.6.98)$$

$$= \gamma_0 + \sum_{j=1}^{\infty} \gamma_j \quad (13.6.99)$$

Therefore the limiting scaled (by factor  $\sqrt{T}$ ) variance of  $\bar{X}_T$  is

$$\varsigma^2 := \lim_{T \rightarrow \infty} \text{Var}(\sqrt{T}\bar{X}_T) \quad (13.6.100)$$

$$= \lim_{T \rightarrow \infty} T \text{Var}(\bar{X}_T) \quad (13.6.101)$$

$$= \lim_{T \rightarrow \infty} T \mathbb{E}[(\bar{X}_T - \mu)^2] \quad (13.6.102)$$

$$= \gamma_0 + 2\gamma_1 + 2\gamma_2 + \dots \quad (13.6.103)$$

$$= \gamma_0 + \sum_{j=1}^{\infty} \gamma_j \quad (13.6.104)$$

$$= \sum_{j=-\infty}^{\infty} \gamma_j \quad (13.6.105)$$

This is called the long-run variance (which should not be confused with the stationary variance,  $\gamma_0$ ). The long-run variance for a stationary and causal ARMA( $p, q$ ) process can be derived in terms of its parameters. Recall that the autocovariance can be expressed as  $\gamma_j = \sigma^2 \sum_{k=0}^{\infty} \psi_k \psi_{k+j}$  where the  $\psi_k$  are the coefficients of the MA( $\infty$ ) representation. Substituting this yields

$$\varsigma^2 = \sum_{j=-\infty}^{\infty} \gamma_j \quad (13.6.106)$$

$$= \sum_{j=0}^{\infty} \gamma_j + \sum_{j=1}^{\infty} \gamma_j \quad (13.6.107)$$

$$= \sigma^2 \left( \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \psi_k \psi_{k+j} + \sum_{j=1}^{\infty} \sum_{k=0}^{\infty} \psi_k \psi_{k+j} \right) \quad (13.6.108)$$

$$= \sigma^2 \left( \sum_{k=0}^{\infty} \sum_{\ell=k}^{\infty} \psi_k \psi_{\ell} + \sum_{k=0}^{\infty} \sum_{\ell=k+1}^{\infty} \psi_k \psi_{\ell} \right) \quad (13.6.109)$$

$$= \sigma^2 \left( \sum_{k=0}^{\infty} \sum_{\ell=k}^{\infty} \psi_k \psi_{\ell} + \sum_{\ell=0}^{\infty} \sum_{k=\ell+1}^{\infty} \psi_k \psi_{\ell} \right) \quad (13.6.110)$$

$$= \sigma^2 \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} \psi_k \psi_{\ell} \quad (13.6.111)$$

$$= \sigma^2 \left( \sum_{k=0}^{\infty} \psi_k \right)^2 \quad (13.6.112)$$

where the sums were combined because each sum covered a triangular half of the two-dimensional array of all the  $(\psi_k, \psi_{\ell})$ . Denoting the polynomial  $\psi(L) = \psi_0 + \psi_1 L + \psi_2 L^2 + \dots$ , we then have

$$\varsigma^2 = \sigma^2 \psi(1)^2 \quad (13.6.113)$$

Recall that we can equate

$$\frac{\theta(L)}{\varphi(L)} = \psi(L) \quad (13.6.114)$$

Therefore the long-run variance in terms of the ARMA parameters is

$$\varsigma^2 = \sigma^2 \frac{\theta(1)^2}{\varphi(1)^2} \quad (13.6.115)$$

$$= \sigma^2 \left( \frac{1 + \theta_1 + \dots + \theta_q}{1 - \varphi_1 - \dots - \varphi_p} \right)^2 \quad (13.6.116)$$

### 13.6.5 Autoregressive Integrated Moving Average (ARIMA) Models

An ARIMA model is a generalisation of the ARMA model, and is denoted by ARIMA  $(p, d, q)$ . The ARIMA model can be thought of as a an ARMA model on a  $d$ -differenced series. Denoting  $L$  as the lag operator (i.e.  $L^i X_t = X_{t-i}$ ), let the first-differenced series be defined as  $X'_t = X_t - X_{t-1}$ , which can be rewritten using the lag operator:

$$X'_t = (1 - L) X_t \quad (13.6.117)$$

Let the second-differenced series be defined as  $X''_t = X'_t - X'_{t-1}$ , which can be rewritten as:

$$X''_t = (1 - L) X'_t \quad (13.6.118)$$

$$= (1 - L)^2 X_t \quad (13.6.119)$$

Hence continuing this pattern,  $(1 - L)^d X_t$  is the  $d$ -differenced series. Suppose  $d \geq 1$  (otherwise it would just be an ARMA model) so that differencing removes the constant term  $c$  in the ARMA series. Then an ARMA model on the  $d$ -differenced series can be specified as

$$(1 - L)^d X_t = \varepsilon_t + \sum_{i=1}^p \varphi_i (1 - L)^d X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (13.6.120)$$

Rearranging this and factorising the lag operators gives the formal definition for an ARIMA model:

$$\left( 1 - \sum_{i=1}^p \varphi_i L^i \right) (1 - L)^d X_t = \left( 1 + \sum_{i=1}^d \theta_i L^i \right) \varepsilon_t \quad (13.6.121)$$

The ‘integrated’ refers to it being the reversal of ‘differenced’. Additionally using the binomial expansion, we can write

$$\left( 1 - \sum_{i=1}^p \varphi_i L^i \right) (1 - L)^d = \left( 1 - \sum_{i=1}^p \varphi_i L^i \right) \sum_{j=0}^d \binom{d}{j} (-L)^{d-j} \quad (13.6.122)$$

$$= \sum_{j=0}^d \binom{d}{j} (-1)^{d-j} L^{d-j} - \sum_{i=1}^p \sum_{j=0}^d \varphi_i \binom{d}{j} (-1)^{d-j} L^{d-j+i} \quad (13.6.123)$$

Hence an alternative specification of the model is

$$\sum_{j=0}^d \binom{d}{j} (-1)^{d-j} X_{t-d+j} - \sum_{i=1}^p \sum_{j=0}^d \varphi_i \binom{d}{j} (-1)^{d-j} X_{t-d+j-i} = \varepsilon_t + \sum_{i=1}^d \theta_i \varepsilon_{t-i} \quad (13.6.124)$$

### Long-Run Component of ARIMA Models [82]

Let  $X_t$  be an ARIMA process such that the first difference  $\Delta X_t := X_t - X_{t-1}$  is a stationary ARMA process (thus, we require  $d = 1$  or  $d = 0$ ). Using the MA( $\infty$ ) representation of an ARMA process, we can write

$$\Delta X_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i} \quad (13.6.125)$$

$$= b(L) \varepsilon_t \quad (13.6.126)$$

where  $b(L)$  is the polynomial

$$b(L) = b_0 + b_1 L + b_2 L^2 + \dots \quad (13.6.127)$$

Now consider the term:

$$b(L) - b(1) = b_0 - b_0 + b_1 L - b_1 + b_2 L^2 - b_2 + \dots \quad (13.6.128)$$

$$= b_1(L-1) + b_2(L^2-1) + b_3(L^3-1) + \dots \quad (13.6.129)$$

$$= \sum_{i=0}^{\infty} b_i (L^i - 1) \quad (13.6.130)$$

Because of the roots of unity,  $L-1$  is a factor of  $L^i-1$ . And factorising leaves

$$\frac{L^i - 1}{L - 1} = L^{i-1} + L^{i-2} + \dots + 1 \quad (13.6.131)$$

which can be confirmed, for example, via polynomial long division. Thus

$$b(L) - b(1) = (L-1)[b_1 + b_2(L+1) + b_3(L^2+L+1) + \dots] \quad (13.6.132)$$

$$= (L-1)[(b_1 + b_2 + \dots) + L(b_2 + b_3 + \dots) + L^2(b_3 + b_4 + \dots) + \dots] \quad (13.6.133)$$

$$= (1-L)b^*(L) \quad (13.6.134)$$

where  $b^*(L)$  is the polynomial  $b^*(L) = b_0^* + b_1^*L + b_2^*L^2 + \dots$  with coefficients

$$b_i^* = - \sum_{j=i+1}^{\infty} b_j \quad (13.6.135)$$

Using this newly-derived expression, we have

$$\Delta X_t = b(L) \varepsilon_t \quad (13.6.136)$$

$$= [b(1) + (1-L)b^*(L)] \varepsilon_t \quad (13.6.137)$$

$$= b(1) \varepsilon_t + (1-L)b^*(L) \varepsilon_t \quad (13.6.138)$$

and putting this into the original series  $X_t$  yields

$$X_t = X_0 + \sum_{s=1}^t \Delta X_s \quad (13.6.139)$$

$$= X_0 + \sum_{s=1}^t [b(1) \varepsilon_s + (1-L)b^*(L) \varepsilon_s] \quad (13.6.140)$$

$$= X_0 + b(1) \sum_{s=1}^t \varepsilon_s + b^*(L) \sum_{s=1}^t \varepsilon_s - b^*(L) \sum_{s=1}^t L \varepsilon_s \quad (13.6.141)$$

$$= b(1) \sum_{s=1}^t \varepsilon_s + b^*(L) \left( \sum_{s=1}^t \varepsilon_s - \sum_{s=0}^{t-1} \varepsilon_s \right) + X_0 \quad (13.6.142)$$

$$= b(1) \sum_{s=1}^t \varepsilon_s + b^*(L) \varepsilon_t + X_0 - b^*(L) \varepsilon_0 \quad (13.6.143)$$

This representation decomposes the series into three components:

- The component  $b(1) \sum_{s=1}^t \varepsilon_s$  is called the long-run component, because it has ‘memory’ from previous times.
- The component  $b^*(L) \varepsilon_t$  is called the short-run component, because it is applicable only to the current time.
- The component  $X_0 - b^*(L) \varepsilon_0$  is the initial term at time zero.

### 13.6.6 Autoregressive Moving Average with Exogenous Inputs (ARMAX) Models

An ARMAX model is a generalisation of the ARMA model, and is denoted by ARMAX  $(p, q, b)$ . The model is defined by

$$X_t = \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \sum_{i=1}^b \eta_i d_{t-i} \quad (13.6.144)$$

where  $\eta_1, \dots, \eta_b$  are parameters of the exogeneous input  $d_t$ .

### 13.6.7 Vector Autoregressive (VAR) Models

A VAR model is a generalisation of the AR model, and is denoted by VAR  $(p)$ . The model is defined by

$$Y_t = c + \sum_{i=1}^p A_i Y_{t-i} + U_t \quad (13.6.145)$$

where  $Y_t$  is a stochastic vector,  $c$  is a constant vector,  $A_1, \dots, A_p$  are square matrix parameters, and  $U_t$  is a zero-mean vector error term with no correlation across time, i.e.  $\mathbb{E}[U_t U_{t-k}^\top] = 0$  for any non-zero  $k$ .

#### VMA Representation of VAR Models [136]

### 13.6.8 Vector Autoregressive Exogenous (VARX) Models

Also known as dynamic simultaneous equations models, VARX models generalise VAR models with an exogenous input. A VARX model of orders  $p$  and  $s$  (denoted VARX  $(p, s)$ ) may be written as

$$Y_t = A_1 Y_{t-1} + \dots + A_p Y_{t-p} + B_0 X_t + \dots + B_s X_{t-s} + U_t \quad (13.6.146)$$

$$= \sum_{i=1}^p A_i Y_{t-i} + B_0 X_t + \sum_{j=1}^s B_j X_{t-j} + U_t \quad (13.6.147)$$

where  $Y_t, X_t$  are vector-valued (not necessarily of the same dimension),  $A_1, \dots, A_p, B_0, \dots, B_s$  are parameter matrices of appropriate dimension and  $U_t$  is the error term.

### 13.6.9 Nonlinear Autoregressive Exogeneous (NARX) Models

A NARX model is a generalisation of the AR model, and is denoted by NARX  $(p, b)$ . The model is defined by

$$X_t = F(X_{t-1}, \dots, X_{t-p}, d_t, d_{t-1}, \dots, d_{t-b}) + \varepsilon_t \quad (13.6.148)$$

where  $F$  is a nonlinear function and  $d_t, \dots, d_{t-b}$  are exogenous variables.

### 13.6.10 Trend Models [95]

A model with a time trend can be written as

$$X_t = f(t) + \eta_t \quad (13.6.149)$$

where  $\eta_t$  is another time-series (possibly autoregressive) process, and  $f(t)$  is the trend function in time. For instance,  $f(t)$  could be a linear trend:

$$X_t = \beta_0 + \beta_1 t + \eta_t \quad (13.6.150)$$

and  $\eta_t$  could be a stationary process (in which case  $X_t$  will be trend stationary).

#### Deterministic Trends

If the process  $\eta_t$  in a trend model is a stationary ARMA process, then it is said to have a deterministic trend. An example is  $\eta_t = \varepsilon_t$  where  $\varepsilon_t$  is white noise.

#### Stochastic Trends

If the process  $\eta_t$  in a trend model is an ARIMA process with  $d = 1$  that when differenced, leaves a stationary ARMA process, then it is said to have a stochastic trend. An example is the random walk  $\eta_t = \sum_{s=1}^t \varepsilon_s$ .

### 13.6.11 Seasonal Models

#### Seasonal ARIMA (SARIMA) Models

## 13.7 Time-Series Regression

### 13.7.1 AR Estimation

#### Least Squares Estimation for AR Models

An AR  $(p)$  model for a series  $Y_t$  may be expressed as a PRF of the form

$$\mathbb{E}[Y_t | \mathcal{F}_{t-1}] = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p} \quad (13.7.1)$$

where  $\mathcal{F}_{t-1} = \{Y_{t-1}, \dots, Y_1\}$  is the information set up to and including time  $t - 1$ . Using this specification, we can estimate the coefficients  $\beta_0, \dots, \beta_p$  as well as the error variance  $\sigma^2$  using ordinary least squares (sometimes called *conditional least squares* in the context of time-series estimation). This estimator will generally be biased (because the strict exogeneity condition is no longer satisfied), however it can be shown to be consistent by similar arguments as for consistency of OLS (except that we need to use a law of large numbers for correlated sequences).

## Yule-Walker Equations

The Yule-Walker equations provides a way to estimate the parameters of an AR ( $p$ ) model based on the method of moments. Using the derived autocovariance function for an AR ( $p$ ) process:

$$\gamma_j = \sum_{i=1}^p \varphi_i \gamma_{j-1} + \sigma^2 \delta_j \quad (13.7.2)$$

we can rearrange this into a matrix form for  $(\gamma_1, \dots, \gamma_p)$ :

$$\underbrace{\begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{bmatrix}}_{\boldsymbol{\gamma}} = \begin{bmatrix} \gamma_0 & \cdots & \gamma_{1-p} \\ \vdots & \ddots & \vdots \\ \gamma_{p-1} & \cdots & \gamma_0 \end{bmatrix} \underbrace{\begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_p \end{bmatrix}}_{\boldsymbol{\varphi}} \quad (13.7.3)$$

$$= \underbrace{\begin{bmatrix} \gamma_0 & \cdots & \gamma_{p-1} \\ \vdots & \ddots & \vdots \\ \gamma_{p-1} & \cdots & \gamma_0 \end{bmatrix}}_{\boldsymbol{\Gamma}} \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_p \end{bmatrix} \quad (13.7.4)$$

Hence the estimation involves first computing the sample autocovariances  $(\hat{\gamma}_0, \dots, \hat{\gamma}_p)$ , and then taking

$$\begin{bmatrix} \hat{\varphi}_1 \\ \vdots \\ \hat{\varphi}_p \end{bmatrix} = \begin{bmatrix} \hat{\gamma}_0 & \cdots & \hat{\gamma}_{p-1} \\ \vdots & \ddots & \vdots \\ \hat{\gamma}_{p-1} & \cdots & \hat{\gamma}_0 \end{bmatrix}^{-1} \begin{bmatrix} \hat{\gamma}_1 \\ \vdots \\ \hat{\gamma}_p \end{bmatrix} \quad (13.7.5)$$

Additionally,  $\sigma^2$  may then be estimated by applying the method of moments to  $j = 0$ :

$$\hat{\sigma}^2 = \hat{\gamma}_0 - \sum_{i=1}^p \hat{\varphi}_i \hat{\gamma}_i \quad (13.7.6)$$

If a constant term  $c$  is required to be estimated, we can first take  $\hat{\mu} = \hat{\mathbb{E}}[X_t]$  from the sample and then use the relation:

$$\hat{c} = \hat{\mu} (1 - \hat{\varphi}_1 - \cdots - \hat{\varphi}_p) \quad (13.7.7)$$

### 13.7.2 ARMA Estimation

#### Two-Step ARMA Estimation

A straightforward and intuitive approach for estimating the coefficients of an ARMA  $(p, q)$  model:

$$X_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (13.7.8)$$

is via a two-step procedure. Since the errors  $\varepsilon_t$  are not known, we first regress  $X_t$  on a constant and  $X_{t-1}, \dots, X_{t-p}$  using some estimator such as OLS to obtain the initial coefficient estimates  $\hat{\varphi}'_1, \dots, \hat{\varphi}'_p$ . From the fitted values, we can then estimate the series of residuals  $\hat{\varepsilon}_t$ . The ARMA specification then suggests we should regress  $X_t$  on a constant and  $X_{t-1}, \dots, X_{t-p}, \hat{\varepsilon}_{t-1}, \dots, \hat{\varepsilon}_{t-q}$  using another estimator (such as OLS again) to obtain the final coefficient estimates  $\hat{c}, \hat{\varphi}_1, \dots, \hat{\varphi}_p, \hat{\theta}_1, \dots, \hat{\theta}_q$ .

## Yule-Walker Equations for ARMA Estimation

To apply the Yule-Walker method of moments approach for ARMA estimation, we begin with the conditions derived from the autocovariance function of an ARMA  $(p, q)$  process:

$$\gamma_j - \varphi_1 \gamma_{j-1} - \cdots - \varphi_p \gamma_{j-p} = \sigma^2 \sum_{i=j}^q \theta_i \psi_{i-j} \quad (13.7.9)$$

To perform estimation, we can first compute the first  $p+q+1$  sample autocovariances  $\hat{\gamma}_0, \dots, \hat{\gamma}_{p+q}$ , and then recognise that the right-hand side of the moment condition is zero for  $j > q$ . Hence we can solve the system of  $p$  linear equations

$$\hat{\gamma}_j - \hat{\varphi}_1 \hat{\gamma}_{j-1} - \cdots - \hat{\varphi}_p \hat{\gamma}_{j-p} = 0 \quad (13.7.10)$$

for  $q+1 \leq j \leq q+p$  to obtain  $(\hat{\varphi}_1, \dots, \hat{\varphi}_p)$ . Then for  $0 \leq j \leq q$ , we solve the system of  $q+1$  equations

$$\hat{\gamma}_j - \hat{\varphi}_1 \hat{\gamma}_{j-1} - \cdots - \hat{\varphi}_p \hat{\gamma}_{j-p} = \hat{\sigma}^2 \sum_{i=j}^q \hat{\theta}_i \hat{\psi}_{i-j} \quad (13.7.11)$$

to obtain  $\hat{\sigma}^2, \hat{\theta}_1, \dots, \hat{\theta}_q$ , where each  $\hat{\psi}_{i-j}$  will be a function of the other coefficients, governed by the solution to the difference equation governing it. Because of this dependence, the problem will be a nonlinear system of equations to solve, so this approach cannot guarantee existence nor uniqueness of solutions.

## Likelihood of ARMA Models [33]

In an ARMA  $(p, q)$  model, assume that errors are normally distributed white noise with variance  $\sigma^2$ , i.e.

$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2) \quad (13.7.12)$$

Then the process  $X_t$  will be a Gaussian process, which can be seen via the MA  $(\infty)$  representation of causal stationary ARMA models. We can then formulate a Gaussian likelihood to estimate the parameters of an ARMA model. Even if the errors are not normally distributed, the consistency of quasi-maximum likelihood estimation can still justify using a Gaussian likelihood as a measure of ‘fit’. Suppose we have collected  $n$  time-series observations

$$\mathbf{X}_n = [X_1, \dots, X_n]^\top \quad (13.7.13)$$

from a causal, stationary and invertible ARMA  $(p, q)$  process, and assume that the process is zero-mean, otherwise we can just work with the mean-subtracted data. The likelihood can be stated as

$$\mathcal{L}(\Gamma_n; \mathbf{X}_n) = \frac{1}{(2\pi)^{n/2} \det(\Gamma_n)^{1/2}} \exp\left(-\frac{1}{2} \mathbf{X}_n^\top \Gamma_n^{-1} \mathbf{X}_n\right) \quad (13.7.14)$$

where  $\text{Cov}(\mathbf{X}_n) = \Gamma_n$  is the covariance matrix, whose elements implicitly depend on the parameters  $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_p)$ ,  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)$  and  $\sigma^2$  because they can be used to compute the autocovariance function via a difference equation. A structured way to simplify the computation of the likelihood is by using the innovations algorithm. Recall that the recursive forecasts  $\hat{\mathbf{X}}_n = [\hat{X}_1 \ \dots \ \hat{X}_n]^\top$  are expressed as

$$\begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \\ \vdots \\ \hat{X}_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vartheta_{1,1} (X_1 - \hat{X}_1) \\ \vdots \\ \sum_{j=1}^{n-1} \vartheta_{n-1,j} (X_j - \hat{X}_j) \end{bmatrix} \quad (13.7.15)$$

which can be rearranged into the form

$$\begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \\ \vdots \\ \hat{X}_n \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & & & \\ \vartheta_{1,1} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \vartheta_{n-1,1} & \cdots & \vartheta_{n-1,n-1} & 0 \end{bmatrix}}_{C-I} \begin{bmatrix} X_1 - \hat{X}_1 \\ X_2 - \hat{X}_2 \\ \vdots \\ X_n - \hat{X}_n \end{bmatrix} \quad (13.7.16)$$

Hence

$$\hat{\mathbf{X}}_n = (C - I) (\mathbf{X}_n - \hat{\mathbf{X}}_n) \quad (13.7.17)$$

$$\hat{\mathbf{X}}_n = C\mathbf{X}_n - C\hat{\mathbf{X}}_n - \mathbf{X}_n + \hat{\mathbf{X}}_n \quad (13.7.18)$$

$$\mathbf{X}_n = C (\mathbf{X}_n - \hat{\mathbf{X}}_n) \quad (13.7.19)$$

Due to the orthogonality properties of the forecasts, the covariance matrix of the forecast errors  $\mathbf{X}_n - \hat{\mathbf{X}}_n$  is diagonal with elements  $\nu_0, \dots, \nu_{n-1}$  (computed via the algorithm), i.e.

$$D := \text{Cov} (\mathbf{X}_n - \hat{\mathbf{X}}_n) \quad (13.7.20)$$

$$= \begin{bmatrix} \mathbb{E} [ (X_1 - \hat{X}_1)^2 ] & & \\ & \ddots & \\ & & \mathbb{E} [ (X_n - \hat{X}_n)^2 ] \end{bmatrix} \quad (13.7.21)$$

$$= \text{diag} \{ \nu_0, \dots, \nu_{n-1} \} \quad (13.7.22)$$

Moreover,

$$\Gamma_n = C \cdot \text{Cov} (\mathbf{X}_n - \hat{\mathbf{X}}_n) \cdot C^\top \quad (13.7.23)$$

$$= CDC^\top \quad (13.7.24)$$

and thus we can compute the quadratic form

$$\mathbf{X}_n^\top \Gamma_n^{-1} \mathbf{X}_n = (\mathbf{X}_n - \hat{\mathbf{X}}_n)^\top C^\top (CDC^\top)^{-1} C (\mathbf{X}_n - \hat{\mathbf{X}}_n) \quad (13.7.25)$$

$$= (\mathbf{X}_n - \hat{\mathbf{X}}_n)^\top C^\top (C^\top)^{-1} D^{-1} C^{-1} C (\mathbf{X}_n - \hat{\mathbf{X}}_n) \quad (13.7.26)$$

$$= (\mathbf{X}_n - \hat{\mathbf{X}}_n)^\top D^{-1} (\mathbf{X}_n - \hat{\mathbf{X}}_n) \quad (13.7.27)$$

$$= \sum_{j=1}^n \frac{(X_j - \hat{X}_j)^2}{\nu_{j-1}} \quad (13.7.28)$$

We also have

$$\det(\Gamma_n) = \det(C) \cdot \det(D) \cdot \det(C^\top) \quad (13.7.29)$$

$$= \det(C)^2 \cdot \det(D) \quad (13.7.30)$$

$$= \prod_{j=1}^n \nu_{j-1} \quad (13.7.31)$$

where we used  $\det(C) = 1$ , as  $C$  is a lower triangular matrix with a leading diagonal of ones. Now introduce an ARMA  $(p, q)$  process  $W_t$ , with the same  $\boldsymbol{\varphi}$ ,  $\boldsymbol{\theta}$  coefficients as  $X_t$ , but with unit variance in the error term. Note that the standard deviation of the error  $\sigma$  just scales the process; this is evident by looking at the MA  $(\infty)$  representation. So if we divided out the data by  $\sigma$  to obtain  $\mathbf{W}_n = \sigma^{-1}\mathbf{X}_n$ , this would make the likelihood (given  $\mathbf{W}_n$ ) as

$$\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\theta}, 1; \mathbf{W}_n) = \frac{1}{(2\pi)^{n/2} \left( \prod_{j=1}^n r_{j-1} \right)^{1/2}} \exp \left( -\frac{1}{2} \sum_{j=1}^n \frac{(W_j - \widehat{W}_j)^2}{r_{j-1}} \right) \quad (13.7.32)$$

where the  $r_{t-1}$  are the mean squared prediction errors when the innovations algorithm is applied to  $W_t$ . Recognising that  $\Gamma_n = \sigma^2 \text{Cov}(\mathbf{W}_n)$  and  $\det(\Gamma_n) = \sigma^{2n} \prod_{j=1}^n r_n$ , this means we may write the likelihood (given  $\mathbf{X}_n$ ) as

$$\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2; \mathbf{X}_n) = \frac{1}{(2\pi\sigma^2)^{n/2} \left( \prod_{j=1}^n r_{j-1} \right)^{1/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{j=1}^n \frac{(X_j - \widehat{X}_j)^2}{r_{j-1}} \right) \quad (13.7.33)$$

In summary, to compute the likelihood of  $(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2)$  given  $\mathbf{X}_n$ :

1. Apply the innovations algorithm for an ARMA process with parameters  $(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2)$  and the realisations of  $\mathbf{X}_n$  to obtain the forecasts  $\widehat{X}_1, \dots, \widehat{X}_n$ .
2. Apply the innovations algorithm for an ARMA process with parameters  $(\boldsymbol{\varphi}, \boldsymbol{\theta}, 1)$  and the realisations of  $\mathbf{W}_n = \sigma^{-1}\mathbf{X}_n$  to obtain the mean squared prediction errors  $r_0, \dots, r_{n-1}$ .
3. Compute the likelihood as above, and subsequently the log-likelihood as

$$\log \mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2; \mathbf{X}_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{j=1}^n \log r_{j-1} - \frac{1}{2\sigma^2} \sum_{j=1}^n \frac{(X_j - \widehat{X}_j)^2}{r_{j-1}} \quad (13.7.34)$$

Note that the  $\vartheta_{t,j}$  coefficients in the innovations algorithm will be the same for both processes  $X_t$  and  $W_t$  (i.e. they only depend on  $\boldsymbol{\varphi}$  and  $\boldsymbol{\theta}$ ), because we can see that a scaling factor of the autocovariance will cancel out in the numerator and denominator for the computation of each  $\vartheta_{t,j}$ .

### Maximum Likelihood Estimation of ARMA Models [33]

The likelihood of ARMA models will generally need to be optimised using a numerical algorithm. However, we can derive a relationship between the parameters by differentiating the log-likelihood above with respect to  $\sigma^2$ :

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2; \mathbf{X}_n)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^n \frac{(X_j - \widehat{X}_j)^2}{r_{j-1}} \quad (13.7.35)$$

Then setting this to zero, we see that the relation

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n \frac{(X_j - \widehat{X}_j)^2}{r_{j-1}} \quad (13.7.36)$$

is satisfied at the maximum likelihood estimate  $(\hat{\varphi}, \hat{\boldsymbol{\vartheta}}, \hat{\sigma}^2)$ . Substituting this back into the log-likelihood expression, we get

$$-\log \mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\vartheta}, \hat{\sigma}^2; \mathbf{X}_n) = \frac{n}{2} \log \left[ \frac{1}{n} \sum_{j=1}^n \frac{(X_j - \hat{X}_j)^2}{r_{j-1}} \right] + \frac{n}{2} \log(2\pi) + \frac{1}{2} \sum_{j=1}^n \log r_{j-1} + \frac{n}{2} \quad (13.7.37)$$

Hence after multiplying by a factor of  $2/n$  and ignoring terms which do not depend on the parameters, a procedure for maximum likelihood estimation is to first obtain  $(\hat{\varphi}, \hat{\boldsymbol{\vartheta}})$  by solving

$$(\hat{\varphi}, \hat{\boldsymbol{\vartheta}}) = \underset{\boldsymbol{\varphi}, \boldsymbol{\vartheta}}{\operatorname{argmin}} \left\{ \log \left[ \frac{1}{n} \sum_{j=1}^n \frac{(X_j - \hat{X}_j)^2}{r_{j-1}} \right] + \frac{1}{n} \sum_{j=1}^n \log r_{j-1} \right\} \quad (13.7.38)$$

and then compute  $\hat{\sigma}^2$  using the relation above.

### 13.7.3 ARIMA Estimation

Since an ARIMA  $(p, d, q)$  model is an ARMA  $(p, q)$  model applied to a  $d$ -differenced series, then ARIMA estimation can be performed by applying ARMA estimation techniques to  $d$ -differenced time-series data.

### 13.7.4 VAR Estimation [132]

Given a multivariate time-series sample of size  $T$ , plus  $p$  pre-sample values, we can write the system as

$$\underbrace{[Y_1 \dots Y_T]}_{\mathbf{Y}} = \underbrace{[c \ A_1 \ \dots \ A_p]}_{\mathbf{B}} \underbrace{\begin{bmatrix} 1 & 1 & \dots & 1 \\ Y_0 & Y_1 & \dots & Y_{T-1} \\ \vdots & \vdots & & \vdots \\ Y_{1-p} & Y_{2-p} & \dots & Y_{T-p} \end{bmatrix}}_{\mathbf{Z}} + \underbrace{[U_1 \dots U_T]}_{\mathbf{U}} \quad (13.7.39)$$

Vectorising (i.e. stacking the vectors) gives

$$\operatorname{vec}(\mathbf{Y}) = \operatorname{vec}(\mathbf{BZ}) + \operatorname{vec}(\mathbf{U}) \quad (13.7.40)$$

Note that a general property of the Kronecker product is that  $\operatorname{vec}(\mathbf{BZ})$  can be written as  $(\mathbf{Z}^\top \otimes I_{K \times K}) \operatorname{vec}(\mathbf{B})$  where  $K$  is the number of rows of  $\mathbf{B}$  (in this case the same as the number of series). By additionally defining  $\mathbf{y} = \operatorname{vec}(\mathbf{Y})$ ,  $\boldsymbol{\beta} = \operatorname{vec}(\mathbf{B})$ ,  $\mathbf{u} = \operatorname{vec}(\mathbf{U})$ , we have

$$\mathbf{y} = (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} + \mathbf{u} \quad (13.7.41)$$

Then an ordinary least squares approach to estimating  $\boldsymbol{\beta}$  is to find the value of  $\boldsymbol{\beta}$  such that

$$V(\boldsymbol{\beta}) = \mathbf{u}^\top \mathbf{u} \quad (13.7.42)$$

$$= [\mathbf{y} - (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta}]^\top [\mathbf{y} - (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta}] \quad (13.7.43)$$

is minimised. Expanding this out:

$$V(\boldsymbol{\beta}) = [\mathbf{y} - (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta}]^\top [\mathbf{y} - (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta}] \quad (13.7.44)$$

$$= \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top (\mathbf{Z}^\top \otimes I_{K \times K})^\top \mathbf{y} + \boldsymbol{\beta}^\top (\mathbf{Z}^\top \otimes I_{K \times K})^\top (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} \quad (13.7.45)$$

$$= \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top (\mathbf{Z} \otimes I_{K \times K}) \mathbf{y} + \boldsymbol{\beta}^\top (\mathbf{Z} \otimes I_{K \times K}) (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} \quad (13.7.46)$$

where we have used the property of Kronecker products that  $(\mathbf{Z}^\top \otimes I_{K \times K})^\top = (\mathbf{Z}^\top)^\top \otimes I_{K \times K}^\top$ . Now use the property that  $(\mathbf{Z} \otimes I_{K \times K}) (\mathbf{Z}^\top \otimes I_{K \times K}) = \mathbf{Z} \mathbf{Z}^\top \otimes I_{K \times K} I_{K \times K}$  to give

$$V(\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top (\mathbf{Z} \otimes I_{K \times K}) \mathbf{y} + \boldsymbol{\beta}^\top (\mathbf{Z} \mathbf{Z}^\top \otimes I_{K \times K} I_{K \times K}) \boldsymbol{\beta} \quad (13.7.47)$$

This can be differentiated to yield

$$\nabla_{\boldsymbol{\beta}} V(\boldsymbol{\beta}) = 2 \left( \mathbf{Z} \mathbf{Z}^\top \otimes I_{K \times K} \right) \boldsymbol{\beta} - 2 (\mathbf{Z} \otimes I_{K \times K}) \mathbf{y} \quad (13.7.48)$$

and setting to zero gives the normal equations

$$\left( \mathbf{Z} \mathbf{Z}^\top \otimes I_{K \times K} \right) \hat{\boldsymbol{\beta}} = (\mathbf{Z} \otimes I_{K \times K}) \mathbf{y} \quad (13.7.49)$$

Hence the estimator is

$$\hat{\boldsymbol{\beta}} = \left( \mathbf{Z} \mathbf{Z}^\top \otimes I_{K \times K} \right)^{-1} (\mathbf{Z} \otimes I_{K \times K}) \mathbf{y} \quad (13.7.50)$$

To simplify further, use the additional property  $(\mathbf{Z} \mathbf{Z}^\top \otimes I_{K \times K})^{-1} = (\mathbf{Z} \mathbf{Z}^\top)^{-1} \otimes I_{K \times K}^{-1}$  giving

$$\hat{\boldsymbol{\beta}} = \left[ \left( \mathbf{Z} \mathbf{Z}^\top \right)^{-1} \otimes I_{K \times K}^{-1} \right] (\mathbf{Z} \otimes I_{K \times K}) \mathbf{y} \quad (13.7.51)$$

and subsequently

$$\hat{\boldsymbol{\beta}} = \left[ \left( \mathbf{Z} \mathbf{Z}^\top \right)^{-1} \mathbf{Z} \otimes I_{K \times K}^{-1} I_{K \times K} \right] \mathbf{y} \quad (13.7.52)$$

$$= \left[ \left( \mathbf{Z} \mathbf{Z}^\top \right)^{-1} \mathbf{Z} \otimes I_{K \times K} \right] \mathbf{y} \quad (13.7.53)$$

## VAR Estimation by Generalised Least Squares

In generalised least squares, the cost changes to

$$V(\boldsymbol{\beta}) = \mathbf{u}^\top \Sigma_{\mathbf{u}}^{-1} \mathbf{u} \quad (13.7.54)$$

where  $\Sigma_{\mathbf{u}}$  is the covariance of  $\text{Vec}(U_1, \dots, U_T)$ . Suppose the errors are uncorrelated with any other time and  $\text{Cov}(U_t) = \Sigma_U$ . Then

$$\text{Cov}(\text{Vec}(U_1, \dots, U_T)) = \begin{bmatrix} \Sigma_U & & \\ & \ddots & \\ & & \Sigma_U \end{bmatrix} \quad (13.7.55)$$

$$= I_{T \times T} \otimes \Sigma_U \quad (13.7.56)$$

Hence we aim to minimise

$$V(\boldsymbol{\beta}) = \mathbf{u}^\top (I_{T \times T} \otimes \Sigma_U)^{-1} \mathbf{u} \quad (13.7.57)$$

$$= \mathbf{u}^\top (I_{T \times T} \otimes \Sigma_U^{-1}) \mathbf{u} \quad (13.7.58)$$

$$= \left[ \mathbf{y} - (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} \right]^\top (I_{T \times T} \otimes \Sigma_U^{-1}) \left[ \mathbf{y} - (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} \right] \quad (13.7.59)$$

Expanding out and applying Kronecker product properties in the same way as for the ordinary least squares estimator:

$$\begin{aligned} V(\boldsymbol{\beta}) &= \mathbf{y}^\top (I_{T \times T} \otimes \Sigma_U^{-1}) \mathbf{y} - 2\boldsymbol{\beta}^\top (\mathbf{Z} \otimes I_{K \times K}) (I_{T \times T} \otimes \Sigma_U^{-1}) \mathbf{y} \\ &\quad + \boldsymbol{\beta}^\top (\mathbf{Z} \otimes I_{K \times K}) (I_{T \times T} \otimes \Sigma_U^{-1}) (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} \end{aligned} \quad (13.7.60)$$

$$= \mathbf{y}^\top (I_{T \times T} \otimes \Sigma_U^{-1}) \mathbf{y} - 2\boldsymbol{\beta}^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{y} + \boldsymbol{\beta}^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} \quad (13.7.61)$$

$$= \mathbf{y}^\top (I_{T \times T} \otimes \Sigma_U^{-1}) \mathbf{y} - 2\boldsymbol{\beta}^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{y} + \boldsymbol{\beta}^\top (\mathbf{Z} \mathbf{Z}^\top \otimes \Sigma_U^{-1}) \boldsymbol{\beta} \quad (13.7.62)$$

Differentiating yields

$$\nabla_{\boldsymbol{\beta}} V(\boldsymbol{\beta}) = 2 (\mathbf{Z} \mathbf{Z}^\top \otimes \Sigma_U^{-1}) \boldsymbol{\beta} - 2 (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{y} \quad (13.7.63)$$

and solving the resulting normal equations:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z} \mathbf{Z}^\top \otimes \Sigma_U^{-1})^{-1} (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{y} \quad (13.7.64)$$

$$= \left[ (\mathbf{Z} \mathbf{Z}^\top)^{-1} \otimes \Sigma_U \right] (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{y} \quad (13.7.65)$$

$$= \left[ (\mathbf{Z} \mathbf{Z}^\top)^{-1} \mathbf{Z} \otimes \Sigma_U \Sigma_U^{-1} \right] \mathbf{y} \quad (13.7.66)$$

$$= \left[ (\mathbf{Z} \mathbf{Z}^\top)^{-1} \mathbf{Z} \otimes I_{K \times K} \right] \mathbf{y} \quad (13.7.67)$$

We see that the cancellation of  $\Sigma_U$  with  $\Sigma_U^{-1}$  means that the generalised least squares estimator is identical to the ordinary least squares estimator for VARX estimation. This same result can also be used to show that when generalised least squares is used for multiple output least squares and the error covariance is uncorrelated between observations, then the estimator will be the same as ordinary least squares.

### VAR Estimation with Parameter Constraints

Suppose in a VAR model we know the values of some particular parameters (e.g. some elements of the  $A_1$  matrix are known to be zero, or we wish to set  $c = 0$ ). Then we can introduce linear restrictions for the parameter vector  $\boldsymbol{\beta}$  of the form

$$\boldsymbol{\beta} = R\boldsymbol{\gamma} + r \quad (13.7.68)$$

where  $\boldsymbol{\gamma}$  is a vector of remaining parameters that need to be estimated, and choice of  $R$  and  $r$  allows for the restrictions to be specified ( $R$  and  $r$  will have the same number of rows as  $\boldsymbol{\beta}$ ). To estimate  $\boldsymbol{\gamma}$ , from the compact form  $\mathbf{y} = (\mathbf{Z}^\top \otimes I_{K \times K}) \boldsymbol{\beta} + \mathbf{u}$  we can write

$$\mathbf{y} = (\mathbf{Z}^\top \otimes I_{K \times K}) (R\boldsymbol{\gamma} + r) + \mathbf{u} \quad (13.7.69)$$

$$\mathbf{y} - \underbrace{(\mathbf{Z}^\top \otimes I_{K \times K}) r}_{\mathbf{z}} = (\mathbf{Z}^\top \otimes I_{K \times K}) R\boldsymbol{\gamma} + \mathbf{u} \quad (13.7.70)$$

The generalised least squares estimator follows from minimising

$$V(\boldsymbol{\gamma}) = \mathbf{u}^\top (I_{T \times T} \otimes \Sigma_U)^{-1} \mathbf{u} \quad (13.7.71)$$

$$= \left[ \mathbf{z} - (\mathbf{Z}^\top \otimes I_{K \times K}) R\boldsymbol{\gamma} \right]^\top (I_{T \times T} \otimes \Sigma_U^{-1}) \left[ \mathbf{z} - (\mathbf{Z}^\top \otimes I_{K \times K}) R\boldsymbol{\gamma} \right] \quad (13.7.72)$$

$$= \mathbf{z}^\top (I_{T \times T} \otimes \Sigma_U^{-1}) \mathbf{z} - 2\boldsymbol{\gamma}^\top R^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{z} + \boldsymbol{\gamma}^\top R^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) (\mathbf{Z}^\top \otimes I_{K \times K}) R \boldsymbol{\gamma} \quad (13.7.73)$$

where the steps to show this are analogous to the unrestricted case above. This time, differentiating with respect to  $\boldsymbol{\gamma}$  and solving the normal equations yields the estimator

$$\hat{\boldsymbol{\gamma}} = [R^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) (\mathbf{Z}^\top \otimes I_{K \times K}) R]^{-1} R^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{z} \quad (13.7.74)$$

$$= [R^\top (\mathbf{Z} \mathbf{Z}^\top \otimes \Sigma_U^{-1}) R]^{-1} R^\top (\mathbf{Z} \otimes \Sigma_U^{-1}) \mathbf{z} \quad (13.7.75)$$

This time however, the  $\Sigma_U$  term does not cancel out so the generalised least squares estimates (of  $\boldsymbol{\gamma}$ ) will generally differ from the ordinary least squares estimates, unlike the unrestricted case. Since in practice  $\Sigma_U$  is usually not known, then it can be instead replaced by a consistent estimate  $\hat{\Sigma}_U$ , yielding an estimated generalised least squares (EGLS) estimator. One example to first compute the ordinary least squares estimates:

$$\hat{\boldsymbol{\gamma}}_{\text{LS}} = [R^\top (\mathbf{Z} \mathbf{Z}^\top \otimes I_{K \times K}) R]^{-1} R^\top (\mathbf{Z} \otimes I_{K \times K}) \mathbf{z} \quad (13.7.76)$$

to obtain  $\hat{\Sigma}_U$  via the residuals. One instance of a consistent estimator is

$$\hat{\Sigma}_U = \frac{1}{T} \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \quad (13.7.77)$$

where  $\hat{\mathbf{U}}$  is the sample version of  $\mathbf{U}$ , consisting of residuals. Then perform:

$$\hat{\boldsymbol{\gamma}}_{\text{EGLS}} = [R^\top (\mathbf{Z} \mathbf{Z}^\top \otimes \hat{\Sigma}_U^{-1}) R]^{-1} R^\top (\mathbf{Z} \otimes \hat{\Sigma}_U^{-1}) \mathbf{z} \quad (13.7.78)$$

### Multivariate Yule-Walker Equations [31]

#### 13.7.5 VARX Estimation [132]

A VARX( $p, s$ ) model can be reduced to compact notation

$$Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + B_0 X_t + \cdots + B_s X_{t-s} + U_t \quad (13.7.79)$$

$$= \sum_{i=1}^p A_i Y_{t-i} + B_0 X_t + \sum_{j=1}^s B_j X_{t-j} + U_t \quad (13.7.80)$$

$$= \mathbf{A} \mathbf{Y}_{t-1} + \mathbf{B} \mathbf{X}_{t-1} + B_0 X_t + U_t \quad (13.7.81)$$

where  $\mathbf{A} = [A_1 \ \dots \ A_p]$ ,  $\mathbf{B} = [B_1 \ \dots \ B_s]$  and  $\mathbf{Y}_{t-1}$ ,  $\mathbf{X}_{t-1}$  are the stacked vectors

$$\mathbf{Y}_{t-1} = \begin{bmatrix} Y_{t-1} \\ \vdots \\ Y_{t-p} \end{bmatrix} \quad (13.7.82)$$

$$\mathbf{X}_{t-1} = \begin{bmatrix} X_{t-1} \\ \vdots \\ X_{t-s} \end{bmatrix} \quad (13.7.83)$$

We assume that  $U_t$  is white noise (i.e. uncorrelated between different times) with covariance  $\text{Cov}(U_t) = \Sigma_u$ . Define the vector  $\boldsymbol{\beta}$  as a vector of all the parameters, i.e.  $\boldsymbol{\beta} = \text{vec}(\mathbf{A}, \mathbf{B}, B_0)$  which means to stack all the columns left to right. Moreover, suppose there may be arbitrary

linear restrictions in the model to estimate (e.g. we wish to set  $B_0 = 0$ ). Then  $\beta$  can be written as

$$\beta = R\gamma \quad (13.7.84)$$

where  $\gamma$  is the parameter vector of unknowns to be estimated and  $R$  is a matrix of appropriate dimensions (if  $Y_t \in \mathbb{R}^K$ ,  $X_t \in \mathbb{R}^M$ , then  $R$  will have  $K^2p + KMs + KM$  rows, and columns equal to the number of unknown parameters). So for example, an entire row of  $R$  can be set to zeros if the corresponding parameter is to be set to zero. For a given sample of size  $T$  (i.e. we will need data for  $T + \max\{p, s\}$  time points), we can then write out the system as

$$\begin{bmatrix} Y_1 & \dots & Y_T \end{bmatrix} = \left[ \begin{array}{c|cc} A_1 & \dots & A_p \\ \hline B_1 & \dots & B_s \\ \hline B_0 \end{array} \right] \begin{bmatrix} Y_0 & Y_1 & \dots & Y_{T-1} \\ \vdots & \vdots & & \vdots \\ \hline Y_{1-p} & Y_{2-p} & \dots & Y_{T-p} \\ \hline X_0 & X_1 & \dots & X_{T-1} \\ \vdots & \vdots & & \vdots \\ \hline X_{1-s} & X_{2-s} & \dots & X_{T-s} \\ \hline X_1 & X_2 & \dots & X_T \end{bmatrix} + [U_1 \dots U_T] \quad (13.7.85)$$

which can be denoted compactly as

$$\mathbf{Y} = [\mathbf{A} \quad \mathbf{B} \quad B_0] \mathbf{Z} + \mathbf{U} \quad (13.7.86)$$

and vectorised using Kronecker product notation by

$$\mathbf{y} = (\mathbf{Z}^\top \otimes I_{K \times K}) \beta + \mathbf{u} \quad (13.7.87)$$

$$= (\mathbf{Z}^\top \otimes I_{K \times K}) R\gamma + \mathbf{u} \quad (13.7.88)$$

where  $\mathbf{y} = \text{vec}(\mathbf{Y})$  and  $\mathbf{u} = \text{vec}(\mathbf{U})$  are stacked vectors. To see this, we write out each element of  $\mathbf{Z}^\top \in \mathbb{R}^{T \times K^p}$  explicitly using the convention  $Y_t = [y_{t,1} \dots y_{t,K}]$  as

$$\mathbf{Z}^\top = \left[ \begin{array}{ccc|ccc|ccc|ccc} y_{0,1} & \dots & y_{0,K} & \dots & y_{1-p,1} & \dots & y_{1-p,K} & x_{0,1} & \dots & x_{0,M} & \dots \\ \vdots & & & \ddots & \vdots & & \vdots & \vdots & & \vdots & \dots \\ y_{T-1,1} & \dots & y_{T-1,K} & \dots & y_{T-p,1} & \dots & y_{T-p,K} & x_{T-1,1} & \dots & x_{T-1,M} & \dots \end{array} \right] \quad (13.7.89)$$

so the Kronecker product  $\mathbf{Z}^\top \otimes I_{K \times K} \in \mathbb{R}^{TK \times K^2p}$  yields

$$\mathbf{Z}^\top \otimes I_{K \times K} = \left[ \begin{array}{ccc|ccc|ccc|ccc} y_{0,1}I_{K \times K} & \dots & y_{0,K}I_{K \times K} & \dots & y_{1-p,K}I_{K \times K} & \dots & y_{1-p,K}I_{K \times K} & \dots \\ \vdots & & \vdots & \ddots & \vdots & & \vdots & \dots \\ y_{T-1,1}I_{K \times K} & \dots & y_{T-1,K}I_{K \times K} & \dots & y_{T-p,K}I_{K \times K} & \dots & y_{T-p,K}I_{K \times K} & \dots \end{array} \right] \quad (13.7.90)$$

Denoting the columns of each parameter matrix according to  $A_1 = [a_{1,1} \dots a_{1,K}]$ , etc., we see that the Kronecker product conforms with  $\beta$ :

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_T \end{bmatrix} = \left[ \begin{array}{ccc|ccc|ccc} y_{0,1}I_{K \times K} & \dots & y_{0,K}I_{K \times K} & \dots \\ \vdots & & \vdots & \dots \\ y_{T-1,1}I_{K \times K} & \dots & y_{T-1,K}I_{K \times K} & \dots \end{array} \right] \begin{bmatrix} a_{1,1} \\ \vdots \\ a_{1,K} \end{bmatrix} + \begin{bmatrix} U_1 \\ \vdots \\ U_T \end{bmatrix} \quad (13.7.91)$$

$$= \begin{bmatrix} a_{1,1}y_{0,1} + \dots + a_{1,K}y_{0,K} \\ \vdots \\ a_{1,1}y_{T-1,1} + \dots + a_{1,K}y_{T-1,K} \end{bmatrix} + \begin{bmatrix} U_1 \\ \vdots \\ U_T \end{bmatrix} \quad (13.7.92)$$

$$= \begin{bmatrix} A_1 Y_0 + \dots \\ \vdots \\ A_1 Y_{T-1} + \dots \end{bmatrix} + \begin{bmatrix} U_1 \\ \vdots \\ U_T \end{bmatrix} \quad (13.7.93)$$

Notice from the compact form  $\mathbf{y} = (\mathbf{Z}^\top \otimes I_{K \times K}) R\boldsymbol{\gamma} + \mathbf{u}$  that it is in the same form as that considered for VAR regression with parameter constraints. Hence the same GLS estimator or EGLS estimator can be used, except we used the symbol  $\mathbf{y}$  instead of  $\mathbf{z}$ :

$$\hat{\boldsymbol{\gamma}}_{\text{EGLS}} = \left[ R^\top (\mathbf{Z}\mathbf{Z}^\top \otimes \hat{\Sigma}_U^{-1}) R \right]^{-1} R^\top (\mathbf{Z} \otimes \hat{\Sigma}_U^{-1}) \mathbf{y} \quad (13.7.94)$$

## 13.8 Time-Series Analysis

### 13.8.1 Residual Autocorrelation

Also known as *serial correlation in the residuals*, residual autocorrelation is when there is evidence of correlation between the residuals and past lags. This suggests that

$$\mathbb{E}[U_t | \mathcal{F}_{t-1}] \neq 0 \quad (13.8.1)$$

where  $U_t$  is the error term and  $\mathcal{F}_{t-1}$  is information up to and including time  $t-1$ . What this is saying that if errors are autocorrelated, then in principle it is possible to predict future errors from past information. However this goes against the definition of the error, which is

$$U_t = Y_t - \mathbb{E}[Y_t | \mathcal{F}_{t-1}] \quad (13.8.2)$$

Taking  $\mathbb{E}[\cdot | \mathcal{F}_{t-1}]$  of both sides, we get

$$\mathbb{E}[U_t | \mathcal{F}_{t-1}] = \mathbb{E}[Y_t | \mathcal{F}_{t-1}] - \mathbb{E}[Y_t | \mathcal{F}_{t-1}] \quad (13.8.3)$$

The RHS evaluates to zero, however evidence of residual autocorrelation suggests the LHS is not zero, which results in a contradiction. Hence this gives the implication that the model specification is not the correct specification for the actual conditional mean.

#### Durbin-Watson Test

#### Ljung-Box Test

#### Breusch-Godfrey Test

### 13.8.2 Structural Breaks

In time-series models, a structural break represents a change in the regime behind the data-generating process at a point in time. For simplicity, we consider a structural break in an AR( $p$ ) model, however the same concept can be applied to other models as well. We can specify a structural break at time  $s$  using the hybrid PRF:

$$\mathbb{E}[Y_t | \mathcal{F}_{t-1}] = \begin{cases} \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p}, & t < s \\ \delta_0 + \delta_1 Y_{t-1} + \dots + \delta_p Y_{t-p}, & t \geq s \end{cases} \quad (13.8.4)$$

#### Chow Test

The Chow test can be used to test for structural breaks. Consider the PRF with dummy variable interactions:

$$\mathbb{E}[Y_t | \mathcal{F}_{t-1}] = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p}$$

$$+ \gamma_0 \mathbb{I}_{\{t \geq s\}} + \gamma_1 Y_{t-1} \mathbb{I}_{\{t \geq s\}} + \cdots + \gamma_p Y_{t-p} \mathbb{I}_{\{t \geq s\}} \quad (13.8.5)$$

This can be rewritten as

$$\mathbb{E}[Y_t | \mathcal{F}_{t-1}] = (\beta_0 + \gamma_0 \mathbb{I}_{\{t \geq s\}}) + (\beta_1 + \gamma_1 \mathbb{I}_{\{t \geq s\}}) Y_{t-1} + \cdots + (\beta_p + \gamma_p \mathbb{I}_{\{t \geq s\}}) Y_{t-p} \quad (13.8.6)$$

where each  $\gamma_i$  can be thought of as  $\gamma_i = \delta_i - \beta_i$ ; the difference in coefficients before and after the structural break. If there is no structural break, then each  $\gamma_i = 0$ . Thus, a null hypothesis:

$$H_0 : \gamma_0 = \gamma_1 = \cdots = \gamma_p \quad (13.8.7)$$

can be tested against the alternative that at least one of the  $\gamma_i$  are not equal to zero. Note that the Chow test can be used to test differences in regressions outside of a time-series context (i.e. for cross-sectional data), by using an alternative dummy variables for an appropriate condition to be tested.

### 13.8.3 Unit Root

For an AR( $p$ ) process with characteristic polynomial such that there is a root  $L = 1$  with multiplicity one, then we say that the process has a unit root. The process will also be nonstationary. A simple example is an AR(1) process that is a random walk:

$$X_t = X_{t-1} + \varepsilon_t \quad (13.8.8)$$

#### Dickey-Fuller Test [136]

Sometimes we may wish to test whether a process has a unit root. For example, we may want to distinguish whether a trend model has a deterministic or stochastic trend. The Dickey-Fuller test can be applied to test whether an AR(1) process has a unit root. That is, whether the model is

$$Y_t = \varphi_1 Y_{t-1} + V_t \quad (13.8.9)$$

with  $\varphi_1 = 1$ . We assume we have data  $Y_1, \dots, Y_T$  which may have already been de-meaned and de-trended (to allow for greater flexibility of models we can test). The usual OLS estimator for  $\varphi_1$  by regressing  $Y_t$  on  $Y_{t-1}$  is

$$\hat{\varphi}_1 = \frac{\sum_{t=2}^T Y_{t-1} Y_t}{\sum_{t=2}^T Y_{t-1}^2} \quad (13.8.10)$$

$$= \frac{\sum_{t=2}^T Y_{t-1} (Y_{t-1} + V_t)}{\sum_{t=2}^T Y_{t-1}^2} \quad (13.8.11)$$

$$= 1 + \frac{\sum_{t=2}^T Y_{t-1} V_t}{\sum_{t=2}^T Y_{t-1}^2} \quad (13.8.12)$$

Consider the Dickey-Fuller test statistic

$$DF := T(\hat{\varphi}_1 - 1) \quad (13.8.13)$$

$$= \frac{\frac{1}{T} \sum_{t=2}^T Y_{t-1} V_t}{\frac{1}{T^2} \sum_{t=2}^T Y_{t-1}^2} \quad (13.8.14)$$

for the hypothesis:

$$H_0 : \varphi_1 = 1 \quad (13.8.15)$$

$$H_A : \varphi_1 < 1 \quad (13.8.16)$$

To derive the asymptotic distribution of this statistic under the assumption that the null is true, we examine the limiting forms of the numerator and denominator. If the null is true, then the process  $Y_t = \sum_{k=1}^t V_k$  is a random walk, and we can use results from the **functional central limit theorem**, i.e.

$$B_T(s) := \frac{1}{\sigma} \sum_{k=1}^{\lfloor Ts \rfloor} \frac{V_k}{\sqrt{T}} \quad (13.8.17)$$

$$\xrightarrow{d} B(s) \quad (13.8.18)$$

as  $T \rightarrow \infty$ , where  $B(s)$  is a standard Wiener process (Brownian motion), and  $\sigma^2$  is the variance of  $V_t$ . For the numerator,

$$\frac{1}{T^2} \sum_{t=2}^T Y_{t-1}^2 = \sum_{t=2}^T \left( \frac{Y_{t-1}}{\sqrt{T}} \right)^2 \cdot \frac{1}{T} \quad (13.8.19)$$

$$= \sum_{t=2}^T \left( \frac{\sum_{k=1}^{t-1} V_k}{\sqrt{T}} \right)^2 \cdot \frac{1}{T} \quad (13.8.20)$$

$$= \sigma^2 \sum_{t=2}^T \left( B_T \left( \frac{t-1}{T} \right) \right)^2 \cdot \frac{1}{T} \quad (13.8.21)$$

$$\xrightarrow{d} \sigma^2 \int_0^1 B(s)^2 ds \quad (13.8.22)$$

Note that the Riemann sum approaches a Riemann integral. In the case of the denominator, using the definition of the **Ito integral**, we have

$$\frac{1}{T} \sum_{t=2}^T Y_{t-1} V_t = \sum_{t=2}^T \frac{Y_{t-1}}{\sqrt{T}} \cdot \frac{V_t}{\sqrt{T}} \quad (13.8.23)$$

$$= \sum_{t=2}^T \frac{Y_{t-1}}{\sqrt{T}} \cdot \frac{Y_t - Y_{t-1}}{\sqrt{T}} \quad (13.8.24)$$

$$= \sum_{t=2}^T \frac{\sum_{k=1}^{t-1} V_k}{\sqrt{T}} \cdot \left( \frac{\sum_{k=1}^t V_k}{\sqrt{T}} - \frac{\sum_{k=1}^{t-1} V_k}{\sqrt{T}} \right) \quad (13.8.25)$$

$$= \sigma^2 \sum_{t=2}^T B_T \left( \frac{t-1}{T} \right) \cdot \left( B_T \left( \frac{t}{T} \right) - B_T \left( \frac{t-1}{T} \right) \right) \quad (13.8.26)$$

$$\xrightarrow{d} \sigma^2 \int_0^1 B(s) dB(s) \quad (13.8.27)$$

This integral can be evaluated, through an alternative form of  $\frac{1}{T} \sum_{t=2}^T Y_{t-1} V_t$ . Firstly squaring both sides of the random walk,

$$Y_t^2 = (Y_{t-1} + V_t)^2 \quad (13.8.28)$$

$$= Y_{t-1}^2 + 2Y_{t-1}V_t + V_t^2 \quad (13.8.29)$$

Summing from  $t = 2$  to  $t = T$ ,

$$\sum_{t=2}^T Y_t^2 = \sum_{t=2}^T Y_{t-1}^2 + 2 \sum_{t=2}^T Y_{t-1} V_t + \sum_{t=2}^T V_t^2 \quad (13.8.30)$$

which rearranges to

$$\sum_{t=2}^T Y_t^2 - \sum_{t=2}^T Y_{t-1}^2 = Y_T^2 - Y_1^2 \quad (13.8.31)$$

$$= 2 \sum_{t=2}^T Y_{t-1} V_t + \sum_{t=2}^T V_t^2 \quad (13.8.32)$$

and thus

$$\sum_{t=2}^T Y_{t-1} V_t = \frac{1}{2} \left( Y_T^2 - Y_1^2 - \sum_{t=2}^T V_t^2 \right) \quad (13.8.33)$$

Dividing out by  $T$ ,

$$\frac{1}{T} \sum_{t=2}^T Y_{t-1} V_t = \frac{1}{2} \left( \frac{Y_T^2}{T} - \frac{Y_1^2}{T} - \frac{1}{T} \sum_{t=2}^T V_t^2 \right) \quad (13.8.34)$$

$$= \frac{1}{2} \left[ \left( \frac{Y_T}{\sqrt{T}} \right)^2 - \left( \frac{Y_1}{\sqrt{T}} \right)^2 - \frac{1}{T} \sum_{t=2}^T V_t^2 \right] \quad (13.8.35)$$

We have for each of the terms

$$\left( \frac{Y_1}{\sqrt{T}} \right)^2 \xrightarrow{\text{P}} 0 \quad (13.8.36)$$

$$\left( \frac{Y_T}{\sqrt{T}} \right)^2 \xrightarrow{\text{d}} \sigma^2 B(1)^2 \quad (13.8.37)$$

$$\frac{1}{T} \sum_{t=2}^T V_t^2 \xrightarrow{\text{P}} \sigma^2 \quad (13.8.38)$$

Therefore by Slutsky's theorem,

$$\frac{1}{T} \sum_{t=2}^T Y_{t-1} V_t \xrightarrow{\text{d}} \frac{\sigma^2}{2} (B(1)^2 - 1) \quad (13.8.39)$$

and this concludes

$$\int_0^1 B(s) dB(s) = \frac{1}{2} (B(1)^2 - 1) \quad (13.8.40)$$

Therefore the Dickey-Fuller statistic under the null hypothesis converges in distribution to

$$\text{DF} = T(\hat{\varphi}_1 - 1) \quad (13.8.41)$$

$$\xrightarrow{\text{d}} \frac{\sigma^2 \int_0^1 B(s) dB(s)}{\sigma^2 \int_0^1 B(s)^2 ds} \quad (13.8.42)$$

$$= \frac{\frac{\sigma^2}{2} (B(1)^2 - 1)}{\sigma^2 \int_0^1 B(s)^2 ds} \quad (13.8.43)$$

$$= \frac{B(1)^2 - 1}{2 \int_0^1 B(s)^2 ds} \quad (13.8.44)$$

This distribution is *non-standard* because the test statistic is not asymptotically normal, however we can still use Monte-Carlo simulation to compute  $p$ -values or critical values.

An alternative Dickey-Fuller statistic that can be used is

$$\text{DF}' := \frac{\widehat{\varphi}_1 - 1}{\text{se}(\widehat{\varphi}_1)} \quad (13.8.45)$$

where

$$\text{se}(\widehat{\varphi}_1) = \sqrt{\widehat{\sigma}^2 \left( \sum_{t=2}^T Y_{t-1}^2 \right)^{-1}} \quad (13.8.46)$$

is the OLS standard error and  $\widehat{\sigma}^2$  is a consistent estimator for  $\sigma^2$ . To derive this asymptotic distribution of this statistic under the null, observe that

$$\text{DF}' = \frac{\sum_{t=2}^T Y_{t-1} V_t}{\sum_{t=2}^T Y_{t-1}^2} \cdot \frac{\sqrt{\sum_{t=2}^T Y_{t-1}^2}}{\widehat{\sigma}} \quad (13.8.47)$$

$$= \frac{\frac{1}{T} \sum_{t=2}^T Y_{t-1} V_t}{\widehat{\sigma} \sqrt{\frac{1}{T^2} \sum_{t=2}^T Y_{t-1}^2}} \quad (13.8.48)$$

$$\xrightarrow{d} \frac{\frac{\sigma^2}{2} (B(1)^2 - 1)}{\sigma \sqrt{\sigma^2 \int_0^1 B(s)^2 ds}} \quad (13.8.49)$$

$$= \frac{B(1)^2 - 1}{2 \sqrt{\int_0^1 B(s)^2 ds}} \quad (13.8.50)$$

### Fuller Reparametrisation

Consider the AR( $p$ ) model:

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_{p-1} Y_{t-p+1} + \varphi_p Y_{t-p} + \varepsilon_t \quad (13.8.51)$$

We can obtain a reparametrisation of the model in terms of the lagged differences  $\Delta Y_t := Y_t - Y_{t-1}$ . Firstly by adding and subtracting  $\varphi_p Y_{t-p+1}$  on the right-hand side,

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_{p-1} Y_{t-p+1} + \varphi_p (Y_{t-p+1} - Y_{t-p+1}) + \varphi_p Y_{t-p} + \varepsilon_t \quad (13.8.52)$$

$$= \varphi_1 Y_{t-1} + \cdots + (\varphi_{p-1} + \varphi_p) Y_{t-p+1} - \varphi_p \Delta Y_{t-p+1} + \varepsilon_t \quad (13.8.53)$$

Next, add and subtract  $(\varphi_{p-1} + \varphi_p) Y_{t-p+2}$  on the right-hand side, so that

$$Y_t = \varphi_1 Y_{t-1} + \cdots + (\varphi_{p-1} + \varphi_p) (Y_{t-p+2} - Y_{t-p+2}) + (\varphi_{p-1} + \varphi_p) Y_{t-p+1} - \varphi_p \Delta Y_{t-p+1} + \varepsilon_t \quad (13.8.54)$$

$$= \varphi_1 Y_{t-1} + \cdots + (\varphi_{p-2} + \varphi_{p-1} + \varphi_p) Y_{t-p+2} - (\varphi_{p-1} + \varphi_p) \Delta Y_{t-p+1} - \varphi_p \Delta Y_{t-p+1} + \varepsilon_t \quad (13.8.55)$$

Continuing this pattern, we will eventually get

$$Y_t = (\varphi_1 + \cdots + \varphi_p) Y_{t-1} - (\varphi_2 + \cdots + \varphi_p) \Delta Y_{t-1} - \cdots - \varphi_p \Delta Y_{t-p+1} + \varepsilon_t \quad (13.8.56)$$

Thus we get the reparametrised model

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 \Delta Y_{t-1} + \cdots + \alpha_p \Delta Y_{t-p+1} + \varepsilon_t \quad (13.8.57)$$

where

$$\alpha_1 = \varphi_1 + \cdots + \varphi_p \quad (13.8.58)$$

$$\alpha_2 = -(\varphi_2 + \cdots + \varphi_p) \quad (13.8.59)$$

$$\vdots \quad (13.8.60)$$

$$\alpha_p = -\varphi_p \quad (13.8.61)$$

This is called the Fuller reparametrisation.

### Augmented Dickey-Fuller Test [82]

The augmented Dickey-Fuller test tests for the presence of a unit root in general AR ( $p$ ) models. The characteristic equation will be given by

$$1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p = 0 \quad (13.8.62)$$

If the series has a unit root, then this literally means that  $L = 1$  is a solution to the characteristic equation. Substituting  $L = 1$ , we see that

$$1 - \varphi_1 - \varphi_2 - \cdots - \varphi_p = 1 - \alpha_1 \quad (13.8.63)$$

$$= 0 \quad (13.8.64)$$

where  $\alpha_1$  is from the Fuller reparametrisation. Thus if the series has a unit root, then  $\alpha_1 = 1$ , which we can test for using the Fuller reparametrisation. The augmented Dickey-Fuller test proceeds as follows, under a further assumption that  $Y_t$  is difference-stationary (i.e.  $\Delta Y_t$  is a stationary series). Under the null hypothesis of a unit root, express the data generating process in terms of the Fuller reparametrisation as

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 \Delta Y_{t-1} + \cdots + \alpha_p \Delta Y_{t-p+1} + V_t \quad (13.8.65)$$

with  $\alpha_1 = 1$ . Subtracting  $Y_{t-1}$  from both sides, we then have in difference form:

$$\Delta Y_t = (\alpha_1 - 1) Y_{t-1} + \alpha_2 \Delta Y_{t-1} + \cdots + \alpha_p \Delta Y_{t-p+1} + V_t \quad (13.8.66)$$

$$= \alpha_2 \Delta Y_{t-1} + \cdots + \alpha_p \Delta Y_{t-p+1} + V_t \quad (13.8.67)$$

and note that  $\Delta Y_t$  is a stationary AR ( $p - 1$ ) process. We regress  $\Delta Y_t$  on regressors  $Y_{t-1}, \Delta Y_{t-1}, \dots, \Delta Y_{t-p+1}$  using OLS, and obtain the coefficient estimates  $\hat{\alpha}_1 - 1, \hat{\alpha}_2, \dots, \hat{\alpha}_p$ . Consider the augmented Dickey-Fuller test statistic:

$$ADF = \frac{\hat{\alpha}_1 - 1}{se(\hat{\alpha}_1)} \quad (13.8.68)$$

which is similar to the Dickey-Fuller statistic. Perhaps surprisingly, this also has the same asymptotic distribution under the null:

$$ADF \xrightarrow{d} \frac{B(1)^2 - 1}{2\sqrt{\int_0^1 B(s)^2 ds}} \quad (13.8.69)$$

This can be shown as follows. Denote  $\mathbf{X}$  as the data matrix for the regressors  $Y_{t-1}, \Delta Y_{t-1}, \dots, \Delta Y_{t-p+1}$ , so the OLS estimates can be expressed as:

$$\begin{bmatrix} \hat{\alpha}_1 - 1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_p \end{bmatrix} = \begin{bmatrix} \alpha_1 - 1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{v} \quad (13.8.70)$$

$$= \begin{bmatrix} 0 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{v} \quad (13.8.71)$$

where  $\mathbf{v}$  is a vector containing the  $V_t$ . From this, we see that the statistic  $T(\hat{\alpha}_1 - 1)$  is also the first element of the vector  $\mathbf{D}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{v}$  where  $\mathbf{D}$  is the diagonal matrix

$$\mathbf{D} = \text{diag} \left\{ T, T^{1/2}, \dots, T^{1/2} \right\} \quad (13.8.72)$$

However also note that

$$\left(\mathbf{D}^{-1}\mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1}\right)^{-1} \mathbf{D}^{-1} \mathbf{X}^\top \mathbf{v} = \mathbf{D} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{D} \mathbf{D}^{-1} \mathbf{X}^\top \mathbf{v} \quad (13.8.73)$$

$$= \mathbf{D} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{v} \quad (13.8.74)$$

so we can consider the first elements of  $\mathbf{D}^{-1}\mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1}$  and  $\mathbf{D}^{-1}\mathbf{X}^\top \mathbf{v}$  separately. We claim that the (1,1) element of  $\mathbf{D}^{-1}\mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1}$  is

$$\begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \mathbf{D}^{-1} \mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \frac{1}{T^2} \sum Y_{t-1}^2 \quad (13.8.75)$$

$$\xrightarrow{\text{d}} \sigma^2 \int_0^1 B(s)^2 ds \quad (13.8.76)$$

where  $\sigma^2$  is the long-run variance of the series  $\Delta Y_t$ . To explain this, we can look again at the Fuller reparametrisation and realise that we may write

$$Y_t = Y_{t-1} + U_t \quad (13.8.77)$$

where

$$U_t = \Delta Y_t \quad (13.8.78)$$

$$= \alpha_2 \Delta Y_{t-1} + \dots + \alpha_p \Delta Y_{t-p+1} + V_t \quad (13.8.79)$$

$$= \alpha_2 U_{t-1} + \dots + \alpha_p U_{t-p+1} + V_t \quad (13.8.80)$$

So  $Y_t$  is like a random walk except  $U_t$  is a stationary process rather than an i.i.d. sequence. Then we can invoke a similar arguments as used in the Dickey-Fuller statistic, involving the functional central limit theorem, but with the following key changes:

- The stationary process central limit theorem is invoked rather than the Lindberg-Levy central limit theorem, since  $U_t$  is a martingale and we have also assumed it to be stationary.
- The term  $\sigma^2$  becomes the long-run variance of  $U_t$ , rather than the variance of  $V_t$ , because in order to standardise the sum  $\sum_{t=1}^T U_t / \sqrt{T}$ , we need to use the limiting variance:

$$\lim_{T \rightarrow \infty} \text{Var} \left( \sum_{t=1}^T U_t / \sqrt{T} \right) = \lim_{T \rightarrow \infty} \text{Var} \left( \sqrt{T} \frac{\sum_{t=1}^T U_t}{T} \right) \quad (13.8.81)$$

which is the definition of the long-run variance. Only in the random walk special case could we take this to be equal to the sum of variances, due to independence.

- Where an i.i.d. property has been used to establish equality in law for time-shifted variables, we can use the stationarity property of  $U_t$  instead.

As for the other elements of  $\mathbf{D}^{-1}\mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1}$ , any other element in the first column or row (excluding the (1,1) element) will be (e.g. for the  $p^{\text{th}}$  row, first column):

$$\begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix} \mathbf{D}^{-1} \mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \frac{1}{T^{3/2}} \sum Y_{t-1} \Delta Y_{t-p+1} \quad (13.8.82)$$

$$= \frac{1}{T^{1/2}} \left( \frac{1}{T} \sum Y_{t-1} \Delta Y_{t-p+1} \right) \quad (13.8.83)$$

The term  $\frac{1}{T} \sum Y_{t-1} \Delta Y_{t-p+1}$  converges in distribution to something like  $\sigma^2 \int_0^1 B(s) dB(s)$  because of similar reasoning to that in the Dickey-Fuller test. Thus, considering the factor  $1/T^{1/2}$  outside, the entire term converges in probability to zero. Also, the remaining  $(i, j)$  elements for  $i > 1, j > 1$  are  $\sum \Delta Y_{t-(i-1)} \Delta Y_{t-(j-1)}/T$ , which converge in probability to the autocovariances of  $U_t$ . Thus the matrix  $\mathbf{D}^{-1} \mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1}$  is asymptotically block diagonal, with the first block being  $1 \times 1$ . Therefore the inverse is also asymptotically block diagonal, with the first block also being  $1 \times 1$ , which is the reciprocal of  $\sigma^2 \int_0^1 B(s)^2 ds$ . The first element of the other vector  $\mathbf{D}^{-1} \mathbf{X}^\top \mathbf{v}$  will be

$$\begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \mathbf{D}^{-1} \mathbf{X}^\top \mathbf{v} = \frac{1}{T} \sum Y_{t-1} V_t \quad (13.8.84)$$

We will find the asymptotic distribution of this sum. Using the long-run component representation of  $Y_t$  (which is difference stationary) we have

$$Y_{t-1} = b(1) \sum_{k=1}^{t-1} V_k + b^*(L) V_{t-1} + Y_0 - b^*(L) V_0 \quad (13.8.85)$$

where

$$b(1) = 1 - \alpha_2 - \dots - \alpha_p \quad (13.8.86)$$

Thus

$$\frac{1}{T} \sum_t Y_{t-1} V_t = \frac{1}{T} \sum_t \left( b(1) \sum_{k=1}^{t-1} V_k + b^*(L) V_{t-1} + Y_0 - b^*(L) V_0 \right) V_t \quad (13.8.87)$$

$$= \frac{b(1)}{T} \sum_t \left( \sum_{k=1}^{t-1} V_k \right) V_t + \frac{1}{T} \sum_t V_t b^*(L) V_{t-1} + (Y_0 - b^*(L) V_0) \frac{1}{T} \sum_t V_t \quad (13.8.88)$$

Because  $V_t$  is zero mean, then

$$(Y_0 - b^*(L) V_0) \frac{1}{T} \sum_t V_t \xrightarrow{\text{P}} 0 \quad (13.8.89)$$

and since the polynomial  $b^*(L)$  contains only non-negative powers of  $L$ , then

$$\frac{1}{T} \sum_t V_t b^*(L) V_{t-1} \xrightarrow{\text{P}} 0 \quad (13.8.90)$$

due to  $V_t$  being white. Then note that  $W_t := \sum_{k=1}^t V_k$  is a random walk, so we can apply the previous result in the Dickey-Fuller test to  $\frac{1}{T} \sum_t W_{t-1} V_t$ :

$$\frac{b(1)}{T} \sum_t \left( \sum_{k=1}^{t-1} V_k \right) V_t = b(1) \cdot \frac{1}{T} \sum_t W_{t-1} V_t \quad (13.8.91)$$

$$\xrightarrow{\text{d}} b(1) \sigma_V^2 \int_0^1 B(s) dB(s) \quad (13.8.92)$$

where  $\sigma_V^2$  is the variance of  $V_t$ . Putting everything together,

$$\frac{1}{T} \sum_t Y_{t-1} V_t \xrightarrow{\text{d}} b(1) \sigma_V^2 \int_0^1 B(s) dB(s) \quad (13.8.93)$$

$$= \frac{b(1)\sigma_V^2}{2} \left( B(1)^2 - 1 \right) \quad (13.8.94)$$

and therefore the asymptotic distribution of the statistic  $T(\hat{\alpha}_1 - 1)$  is

$$T(\hat{\alpha}_1 - 1) = \left( [1 \ 0 \ \dots \ 0] \mathbf{D}^{-1} \mathbf{X}^\top \mathbf{X} \mathbf{D}^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right)^{-1} \left( [1 \ 0 \ \dots \ 0] \mathbf{D}^{-1} \mathbf{X}^\top \mathbf{v} \right) \quad (13.8.95)$$

$$= \frac{\frac{1}{T} \sum Y_{t-1} V_t}{\frac{1}{T^2} \sum Y_{t-1}^2} \quad (13.8.96)$$

$$\xrightarrow{d} \frac{b(1)\sigma_V^2 \left( B(1)^2 - 1 \right)}{2\sigma^2 \int_0^1 B(s)^2 ds} \quad (13.8.97)$$

$$= \frac{b(1)\sigma_V^2 \left( B(1)^2 - 1 \right)}{2b(1)^2 \sigma_V^2 \int_0^1 B(s)^2 ds} \quad (13.8.98)$$

$$= \frac{B(1)^2 - 1}{2b(1) \int_0^1 B(s)^2 ds} \quad (13.8.99)$$

Relating this to the augmented Dickey-Fuller statistic:

$$\text{ADF} = \frac{\hat{\alpha}_1 - 1}{\text{se}(\hat{\alpha}_1)} \quad (13.8.100)$$

$$= \frac{T(\hat{\alpha}_1 - 1)}{\sqrt{T^2 \hat{\sigma}_V^2 (\sum Y_{t-1}^2)^{-1}}} \quad (13.8.101)$$

$$= \frac{1}{\hat{\sigma}_V} \cdot T(\hat{\alpha}_1 - 1) \cdot \sqrt{\frac{1}{T} \sum Y_{t-1}^2} \quad (13.8.102)$$

$$\xrightarrow{d} \frac{1}{\sigma_V} \cdot \frac{B(1)^2 - 1}{2b(1) \int_0^1 B(s)^2 ds} \cdot \sqrt{\sigma^2 \int_0^1 B(s)^2 ds} \quad (13.8.103)$$

$$= \frac{B(1)^2 - 1}{2\sqrt{\int_0^1 B(s)^2 ds}} \quad (13.8.104)$$

where we have used the fact that

$$\sigma = b(1)\sigma_V \quad (13.8.105)$$

$$= (1 - \alpha_2 - \dots - \alpha_p)\sigma_V \quad (13.8.106)$$

for the long-run variance.

### 13.8.4 Cointegration

#### Granger Representation Theorem

### 13.8.5 Spurious Regression

In time-series, a spurious regression is a phenomenon that occurs when we regress one time series  $Y_t$  on another independent series  $X_t$  using a simple linear regression of the form:

$$\mathbb{E}[Y_t | X_t] = \beta_0 + \beta_1 X_t \quad (13.8.107)$$

and then we get statistically significant estimates for  $\beta_1$  at a rate well above the level of significance, even though one should think that  $\beta_1 = 0$  if the series are independent. For this phenomenon to occur, the series  $Y_t$  and  $X_t$  need to be non-stationary (for example, random walks). To explain why non-stationarity might lead to over-rejection of the null, first suppose  $Y_t$  and  $X_t$  were both independent stationary series. Then the quantity

$$\beta_1 = \frac{\text{Cov}(X_t, Y_t)}{\text{Var}(X_t)} \quad (13.8.108)$$

is well-defined, and in fact  $\beta_1 = 0$  because the two series are independent. However, in the case that  $Y_t$  and  $X_t$  are non-stationary, then although  $\text{Cov}(X_t, Y_t) = 0$  for any time  $t$ , it no longer makes sense to estimate  $\beta_1$  from the sample. This is because the sample contains non-stationary data, so we cannot rely on sample averages to converge to their intended population values like we can for stationary ergodic data.

Instead, supposing that  $Y_t$  and  $X_t$  are independent random walks, then  $(X_t, Y_t)$  is a random walk on a two-dimensional plane, so  $\hat{\beta}_1$  captures the ‘eccentricity’ of the sample path. Intuitively, since random walks typically exhibit some eccentricity, we would believe that  $\hat{\beta}_1$  is rarely close to zero, even if we were to take a very large sample. Yet the computed OLS standard errors would be typically small in such a large sample, so we would end up concluding statistical significance, thus producing a spurious regression. The message is that just because we find a statistically significant linear relationship between time-series, this is not as strong evidence that the series are correlated.

A crude characterisation for  $\hat{\beta}_1$  is also available. Suppose we collect  $T$  observations from  $Y_t$  and  $X_t$ , which are independent and identical zero-mean random walks. Then  $Y_T$  and  $X_T$  are both approximately normal via the central limit theorem. Using the fact that a ratio of identical normal random variables is a standard Cauchy distribution, then

$$\frac{Y_T}{X_T} \xrightarrow{d} \text{Cauchy}(0, 1) \quad (13.8.109)$$

We can use the quantity  $\frac{Y_T}{X_T}$  as a loose approximation of  $\hat{\beta}_1$ . This is because a random walk that starts at zero and ends up at  $Y_T$  appears to have a trend, so taking the ratio of trends  $\frac{Y_T}{X_T}$  gives a rough representation of the slope  $\hat{\beta}_1$ .

### 13.8.6 Seasonality [33, 42, 200, 222]

### 13.8.7 Box-Jenkins Method

### 13.8.8 Cholesky Impulse

### 13.8.9 Slutsky-Yule Effect

A rolling average of a series  $X_t$  is taken as the arithmetic mean of the past  $m$  observations:

$$Z_t = \frac{1}{m} \sum_{i=0}^{m-1} X_{t-i} \quad (13.8.110)$$

which can be applied as a method to smooth the series when plotting. However, in the case that  $X_t$  is white noise, then the rolling average  $Z_t$  will be a (scaled) moving average process. Thus taking the rolling average of a white process will induce autocorrelation in the rolling average. This is known as the Slutsky-Yule effect. Importantly, if a rolling average is seen to exhibit autocorrelation through patterns or trends, this does not necessarily imply that the underlying series being averaged is also autocorrelated.

## 13.9 Time-Series Forecasting

### 13.9.1 Granger Causality

Let  $Y_t$  and  $X_t$  be stochastic processes. Let  $\mathcal{F}_t = \{Y_t, Y_{t-1}, \dots\}$  and  $\mathcal{I}_t = \{X_t, X_{t-1}, \dots\}$  denote information sets for  $Y_t$  and  $X_t$  respectively. Then  $X_t$  is said to Granger-cause  $Y_t$  if

$$\Pr(Y_{t+1} \in A | \mathcal{F}_t \cup \mathcal{I}_t) \neq \Pr(Y_{t+1} \in A | \mathcal{F}_t) \quad (13.9.1)$$

for some non-empty set  $A$ . Intuitively, this says that having information about  $X_t$  is more useful for forecasting  $Y_{t+1}$ , compared to not having any information at all. Note that Granger causality does not necessarily imply true causality, because the definition still allows for correlation. A more suitable name would perhaps be Granger predictability. There are also stronger definitions of Granger causality [218], which are restricted to conditional expectation:

$$\mathbb{E}[Y_{t+1} | \mathcal{F}_t \cup \mathcal{I}_t] \neq \mathbb{E}[Y_{t+1} | \mathcal{F}_t] \quad (13.9.2)$$

It is stronger in the sense that if the conditional expectations are not equal, this implies the conditional distributions are not equal, but not necessarily in reverse.

### Testing for Granger Causality

A test is possible for the stronger definition of Granger causality. We consider two linear population regression functions specified by

$$\mathbb{E}[Y_t | \mathcal{F}_{t-1} \cup \mathcal{I}_{t-1}] = \delta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \alpha_1 X_{t-1} + \dots + \alpha_2 X_{t-2} + \dots \quad (13.9.3)$$

and

$$\mathbb{E}[Y_t | \mathcal{F}_{t-1}] = \delta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \quad (13.9.4)$$

for a select number of lags. Under the null hypothesis that there is no Granger causality, we should test the hypothesis that  $\alpha_1 = \alpha_2 = \dots = 0$  against the alternative that at least one of  $\alpha_1, \alpha_2, \dots$  is not equal to zero.

### 13.9.2 Innovations Algorithm [33]

The innovations algorithm is a recursive method for forecasting  $X_{n+1}$  from observations  $(X_1, \dots, X_n)$ . Denote the forecast of  $X_{n+1}$ , using all the information  $(X_1, \dots, X_n)$ , as  $\hat{X}_{n+1}$ . We call the forecast error using information up to time  $n$ , denoted  $U_n = X_{n+1} - \hat{X}_{n+1}$ , the *innovation*. Assume that  $X_t$  is a second-order process, and without loss of generality, zero-mean (otherwise we can consider the mean-subtracted process). Also, the innovations algorithm traditionally requires that all the parameters of the process are known, so that the covariances:

$$\kappa_{t,s} = \mathbb{E}[X_t X_s] \quad (13.9.5)$$

can be computed. The forecast involves recursively computing  $\hat{X}_1, \dots, \hat{X}_{n+1}$  using a linear combination of previous innovations of the form:

$$\hat{X}_{t+1} = \sum_{j=1}^t \vartheta_{t,j} U_{t-j} \quad (13.9.6)$$

$$= \sum_{j=1}^t \vartheta_{t,j} (X_{t+1-j} - \hat{X}_{t+1-j}) \quad (13.9.7)$$

where we take  $\hat{X}_1 = 0$ . The coefficients  $\vartheta_{t,j}$  can be found from the parameters of the process, in a way that will be shown. Intuitively, the forecast is recursively ‘correcting’ itself based on

previous forecast errors. Also, by treating  $\hat{X}_{n+1}$  as the ‘best’ (minimum mean squared prediction error) forecast using the information available, we can apply the projection characterisation of the best linear predictor, and write

$$X_{n+1} = \hat{X}_{n+1} + U_n \quad (13.9.8)$$

where  $\hat{X}_{n+1}$  is orthogonal to  $U_n$ , i.e.  $\mathbb{E}[\hat{X}_{n+1}U_n] = 0$ . We now derive formulas that can be used to find the  $\vartheta_{t,j}$  coefficients. First, we claim that  $\{U_0, \dots, U_n\}$  is an orthogonal set. The argument is that for any  $j \geq 1$ ,  $\hat{X}_j$  is a linear combination of  $(X_1 - \hat{X}_1, \dots, X_{j-1} - \hat{X}_{j-1})$ . Since  $X_j - \hat{X}_j$  is orthogonal to  $\hat{X}_j$  as established, then  $X_j - \hat{X}_j$  must be orthogonal to each of  $\{X_1 - \hat{X}_1, \dots, X_{j-1} - \hat{X}_{j-1}\}$ . Due to this, if we multiply both sides of the forecasting equation for  $\hat{X}_{t+1}$  by  $X_{k+1} - \hat{X}_{k+1}$  and take expectations, we get

$$\mathbb{E}[\hat{X}_{t+1}(X_{k+1} - \hat{X}_{k+1})] = \mathbb{E}\left[\sum_{j=1}^n \vartheta_{t,j}(X_{t+1-j} - \hat{X}_{t+1-j})(X_{k+1} - \hat{X}_{k+1})\right] \quad (13.9.9)$$

$$= \vartheta_{t,t-k} \mathbb{E}\left[(X_{k+1} - \hat{X}_{k+1})^2\right] \quad (13.9.10)$$

$$= \vartheta_{t,t-k} \nu_k \quad (13.9.11)$$

where we use  $\nu_k := \mathbb{E}\left[(X_{k+1} - \hat{X}_{k+1})^2\right]$  to denote the mean squared prediction error using information up to time  $k$ . We also have

$$\mathbb{E}[\hat{X}_{t+1}(X_{k+1} - \hat{X}_{k+1})] = \mathbb{E}\left[(X_{t+1} - (X_{t+1} - \hat{X}_{t+1}))(X_{k+1} - \hat{X}_{k+1})\right] \quad (13.9.12)$$

$$= \mathbb{E}[X_{t+1}(X_{k+1} - \hat{X}_{k+1})] \quad (13.9.13)$$

by orthogonality between  $X_{t+1} - \hat{X}_{t+1}$  and  $X_{k+1} - \hat{X}_{k+1}$ . Combining and rearranging for  $\vartheta_{t,t-k}$ :

$$\vartheta_{t,t-k} = \frac{1}{\nu_k} \mathbb{E}[X_{t+1}(X_{k+1} - \hat{X}_{k+1})] \quad (13.9.14)$$

$$= \frac{1}{\nu_k} (\mathbb{E}[X_{t+1}X_{k+1}] - \mathbb{E}[X_{t+1}\hat{X}_{k+1}]) \quad (13.9.15)$$

$$= \frac{1}{\nu_k} \left( \kappa_{t+1,k+1} - \sum_{j=0}^{k-1} \vartheta_{k,k-j} \mathbb{E}[X_{t+1}(X_{j+1} - \hat{X}_{j+1})] \right) \quad (13.9.16)$$

$$= \frac{1}{\nu_k} \left( \kappa_{t+1,k+1} - \sum_{j=0}^{k-1} \vartheta_{k,k-j} \vartheta_{t,t-j} \nu_j \right) \quad (13.9.17)$$

because  $\mathbb{E}[X_{t+1}(X_{j+1} - \hat{X}_{j+1})] = \vartheta_{t,t-j} \nu_j$ . This formula is recursive for  $k = 0, \dots, t-1$ . A recursive equation for  $\nu_t$  can be found by

$$\nu_t = \mathbb{E}\left[(X_{t+1} - \hat{X}_{t+1})^2\right] \quad (13.9.18)$$

$$= \mathbb{E}[X_{t+1}^2 - 2X_{t+1}\hat{X}_{t+1} + \hat{X}_{t+1}^2] \quad (13.9.19)$$

$$= \mathbb{E}[X_{t+1}^2] - \mathbb{E}[\hat{X}_{t+1}^2] \quad (13.9.20)$$

since

$$\mathbb{E} [2X_{t+1}\hat{X}_{t+1}] = 2\mathbb{E} [\left(\hat{X}_{t+1} + U_t\right)\hat{X}_{t+1}] \quad (13.9.21)$$

$$= 2\mathbb{E} [\hat{X}_{t+1}^2] \quad (13.9.22)$$

by orthogonality between  $\hat{X}_{t+1}$  and  $U_t$ . Continuing,

$$\nu_t = \kappa_{t+1,t+1} - \mathbb{E} \left[ \left( \sum_{j=1}^t \vartheta_{t,j} (X_{t+1-j} - \hat{X}_{t+1-j}) \right)^2 \right] \quad (13.9.23)$$

$$= \mathbb{E} \left[ \left( \sum_{k=0}^{t-1} \vartheta_{t,t-k} (X_{k+1} - \hat{X}_{k+1}) \right)^2 \right] \quad (13.9.24)$$

$$= \sum_{k=0}^{t-1} \vartheta_{t,t-k}^2 \nu_k \quad (13.9.25)$$

by orthogonality between terms in the sum, noting that we have also used the change of variables  $j = t - k$ . As  $\hat{X}_1 = 0$ , the initial mean squared prediction error  $\nu_0 = \mathbb{E} [\left(X_1 - \hat{X}_1\right)^2]$  is given by

$$\nu_0 = \mathbb{E} [X_1^2] \quad (13.9.26)$$

$$= \kappa_{1,1} \quad (13.9.27)$$

To illustrate the recursive procedure using these formulas, the first few terms can be found by

$$\vartheta_{1,1} = \frac{\kappa_{2,1}}{\nu_0} \quad (13.9.28)$$

$$\nu_1 = \kappa_{2,2} - \vartheta_{1,1}^2 \nu_0 \quad (13.9.29)$$

$$\vartheta_{2,2} = \frac{\kappa_{3,1}}{\nu_0} \quad (13.9.30)$$

$$\vartheta_{2,1} = \frac{1}{\nu_1} (\kappa_{3,2} - \vartheta_{1,1} \vartheta_{2,1} \nu_0) \quad (13.9.31)$$

$$\nu_2 = \kappa_{3,3} - \vartheta_{2,2}^2 \nu_0 - \vartheta_{2,1}^2 \nu_1 \quad (13.9.32)$$

$$\vdots \quad (13.9.33)$$

## 13.10 Generalised Method of Moments

Let  $X_1, \dots, X_n$  be a sample of observations, where each  $X_i$  may be a multivariate random variable. Let  $\theta$  be a vector for parameters of interest to be identified. Suppose we can form the following ‘moment conditions’ for the observation  $X_i$ :

$$\mathbb{E} [m (X_i, \theta)] = 0 \quad (13.10.1)$$

where  $m (X_i, \theta)$  is a function of the observation and parameters, and is allowed to be vector-valued. In the method of moments,  $\mathbb{E} [m (X_i, \theta)]$  was explicitly a function of the moments  $\mathbb{E} [X]$ ,  $\mathbb{E} [X^2]$ , etc. In the generalised method of moments (GMM),  $m (X_i, \theta)$  could be a much more general function. The idea behind GMM is to work with the empirical analog of expectation - the sample mean of  $m (X, \theta)$ :

$$\bar{m} (\theta) = \frac{1}{n} \sum_{i=1}^n m (X, \theta) \quad (13.10.2)$$

We know that the true value of  $\theta$  causes  $\mathbb{E}[m(X_i, \theta)] = 0$  to hold. So values of  $\bar{m}(\theta)$  which are closer to zero are deemed to be ‘more probable’. Thus, the estimate of  $\theta$  is obtained by solving an optimisation problem which minimises the norm (or equivalently, squared norm) of  $\bar{m}(\theta)$ , as in:

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \left\{ \left( \frac{1}{n} \sum_{i=1}^n m(X_i, \theta) \right)^\top \left( \frac{1}{n} \sum_{i=1}^n m(X_i, \theta) \right) \right\} \quad (13.10.3)$$

where  $\Theta$  is the parameter space. A more general estimator can be considered, by minimising the weighted norm of positive weighting matrix  $W$ :

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \left\{ \left( \frac{1}{n} \sum_{i=1}^n m(X_i, \theta) \right)^\top W \left( \frac{1}{n} \sum_{i=1}^n m(X_i, \theta) \right) \right\} \quad (13.10.4)$$

### 13.10.1 Ordinary Least Squares as Generalised Method of Moments

Ordinary least squares can be cast as a generalised method of moments problem, with the moment condition

$$\mathbb{E}[X_i(Y_i - X_i^\top \beta)] = 0 \quad (13.10.5)$$

Note that this moment condition is then responsible for the property of orthogonality between the regressors  $X_i$  and residuals  $Y_i - X_i^\top \beta$ . To solve for the estimator, we write the sample version of the moment condition as

$$\bar{m}(\beta) = \frac{1}{n} \sum_{i=1}^n X_i(Y_i - X_i^\top \beta) \quad (13.10.6)$$

which notice has a solution for  $\bar{m}(\beta) = 0$ . So rather than minimise  $\bar{m}(\beta)^\top \bar{m}(\beta)$ , we can instead find the roots of  $\bar{m}(\beta)$ . This gives

$$\frac{1}{n} \sum_{i=1}^n X_i(Y_i - X_i^\top \hat{\beta}) \quad (13.10.7)$$

Upon rearranging, this yields

$$\frac{1}{n} \sum_{i=1}^n X_i Y_i - \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \hat{\beta} = 0 \quad (13.10.8)$$

$$\hat{\beta} = \left( \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left( \sum_{i=1}^n X_i Y_i \right) \quad (13.10.9)$$

which is the OLS estimator.

### 13.10.2 Maximum Likelihood as Generalised Method of Moments

The maximum likelihood estimator can also be cast as a special case of generalised method of moments estimation. The moment condition here is

$$\mathbb{E}[\nabla_\theta \log f(X_i; \theta)] = 0 \quad (13.10.10)$$

where  $f(X_i; \theta) = \mathcal{L}(\theta; X_i)$  is the density probability of  $X_i$  given  $\theta$ , or equivalently the likelihood of  $\theta$  given  $X_i$ . Recall that the score function was shown to be equal to zero; the moment condition represents this. Hence sample version of this moment condition is

$$\bar{m}(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla_\theta \log \mathcal{L}(\theta; X_i) \quad (13.10.11)$$

$$= \nabla_{\theta} \left( \frac{1}{n} \sum_{i=1}^n \log \mathcal{L}(\theta; X_i) \right) \quad (13.10.12)$$

$$= 0 \quad (13.10.13)$$

Solving this amounts to finding where the gradient of the overall log-likelihood  $\sum_{i=1}^n \log \mathcal{L}(\theta; X_i)$  is equal to zero, which is the first order condition for finding the minimum of the overall log-likelihood.

### 13.10.3 Instrumental Variables Regression as Generalised Method of Moments

Instrumental variables regression can be cast as GMM estimation. Consider the causal equation  $Y_i = X_i^\top \beta + U_i$  where  $X_i$  is exogenous, and  $Z_i$  is an instrumental variable. The moment condition posed is

$$\mathbb{E}[Z_i U_i] = \mathbb{E} \left[ Z_i (Y_i - X_i^\top \beta) \right] \quad (13.10.14)$$

$$= 0 \quad (13.10.15)$$

which is that the instruments  $Z_i$  are orthogonal with the errors  $U_i$  (conveying almost the same meaning as the exogeneity condition). The sample moment condition is

$$\bar{m}(\beta) = \frac{1}{n} \sum_{i=1}^n Z_i (Y_i - X_i^\top \beta) \quad (13.10.16)$$

$$= 0 \quad (13.10.17)$$

Rearranging and rewriting in matrix form, this gives

$$\frac{1}{n} \sum_{i=1}^n Z_i Y_i = \frac{1}{n} \sum_{i=1}^n Z_i X_i^\top \beta \quad (13.10.18)$$

$$\mathbf{Z}^\top \mathbf{y} = \mathbf{Z}^\top \mathbf{X} \beta \quad (13.10.19)$$

where the observations in  $\mathbf{X}$ ,  $\mathbf{y}$  and  $\mathbf{Z}$  are row-wise (i.e. they will be tall matrices with  $n$  rows). Since the dimension of  $Z_i$  can be greater than that of  $X_i$ , then generally we will not be able to solve the sample moment equations. One option is to take the Moore-Penrose pseudoinverse  $(\mathbf{Z}^\top \mathbf{X})^\dagger$  and make the GMM estimator

$$\hat{\beta} = (\mathbf{Z}^\top \mathbf{X})^\dagger \mathbf{Z}^\top \mathbf{y} \quad (13.10.20)$$

Since  $\mathbf{Z}^\top \mathbf{X}$  will have at least as many rows as columns, then assuming  $\mathbf{Z}^\top \mathbf{X}$  is full column rank, its pseudoinverse is explicitly computed by

$$(\mathbf{Z}^\top \mathbf{X})^\dagger = (\mathbf{X}^\top \mathbf{Z} \mathbf{Z}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z} \quad (13.10.21)$$

so that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{Z} \mathbf{Z}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z} \mathbf{Z}^\top \mathbf{y} \quad (13.10.22)$$

This actually corresponds to minimising the weighted quadratic criterion for the GMM estimator with weighting matrix  $W = I$ . More generally, we define the GMM estimator as

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left\{ (\mathbf{Z}^\top \mathbf{y} - \mathbf{Z}^\top \mathbf{X} \beta)^\top W (\mathbf{Z}^\top \mathbf{y} - \mathbf{Z}^\top \mathbf{X} \beta) \right\} \quad (13.10.23)$$

Differentiating yields the first order condition

$$\mathbf{X}^\top \mathbf{Z} \cdot 2W (\mathbf{Z}^\top \mathbf{y} - \mathbf{Z}^\top \mathbf{X} \hat{\beta}) = 0 \quad (13.10.24)$$

and solving reveals

$$\mathbf{X}^\top \mathbf{Z} W \mathbf{Z}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{Z} W \mathbf{Z}^\top \mathbf{X} \hat{\beta} \quad (13.10.25)$$

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{Z} W \mathbf{Z}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z} W \mathbf{Z}^\top \mathbf{y} \quad (13.10.26)$$

The GMM estimator is asymptotically efficient with choice of  $W$  being the inverse of the asymptotic covariance of  $\sqrt{n}\bar{m}(\beta)$ , which in this case is equivalently  $n$  times the asymptotic covariance of  $\frac{1}{n} \sum_{i=1}^n Z_i U_i$ . Thus applying the multivariate central limit theorem, the optimal weighting matrix satisfies

$$W^{-1} = n \times \frac{1}{n} \text{Cov}(Z_i U_i) \quad (13.10.27)$$

$$= \text{Cov}(Z_i U_i) \quad (13.10.28)$$

$$= \mathbb{E}[Z_i U_i^2 Z_i^\top] \quad (13.10.29)$$

since  $\mathbb{E}[Z_i U_i] = 0$  under the moment condition. So using the Law of Iterated Expectations,

$$W^{-1} = \mathbb{E}[\mathbb{E}[U_i^2 | Z_i] Z_i Z_i^\top] \quad (13.10.30)$$

If we assume that  $\mathbb{E}[U_i^2 | Z_i] = \sigma^2$ , i.e. homoskedasticity of the errors with respect to  $Z_i$ , then the optimal weighting is

$$W = \frac{1}{\sigma^2} \mathbb{E}[Z_i Z_i^\top]^{-1} \quad (13.10.31)$$

As this quantity is not known in practice, then a consistent estimate (which still preserves the asymptotic efficiency property) is

$$\widehat{W} = \frac{1}{\sigma^2} (\mathbf{Z}^\top \mathbf{Z})^{-1} \quad (13.10.32)$$

Hence the GMM estimator becomes (after cancelling out the  $1/\sigma^2$  factors):

$$\hat{\beta} = \left[ \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X} \right]^{-1} \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \quad (13.10.33)$$

We can show that this version of the GMM estimator is then identical to the two-stage least squares estimator. In 2SLS, the first stage estimate is

$$\hat{\Pi} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X} \quad (13.10.34)$$

hence the fitted values  $\hat{\mathbf{X}}^\top = \hat{\Pi}^\top \mathbf{Z}^\top$  in the first stage is given by

$$\hat{\mathbf{X}}^\top = \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \quad (13.10.35)$$

In the second stage, the 2SLS estimate is

$$\hat{\beta} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top \mathbf{y} \quad (13.10.36)$$

$$= \left[ \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \left[ \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \right]^\top \right]^{-1} \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \quad (13.10.37)$$

$$= \left[ \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X} \right]^{-1} \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \quad (13.10.38)$$

$$= \left[ \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X} \right]^{-1} \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \quad (13.10.39)$$

which is the same as the GMM estimator derived above.

### 13.10.4 Consistency of Generalised Method of Moments [72]

The GMM estimator can be shown to be consistent. We consider the quadratic criterion indexed in sample size  $n$ :

$$q_n(\theta) = \bar{m}_n(\theta)^\top W_n \bar{m}_n(\theta) \quad (13.10.40)$$

where  $W_n$  is a positive definite matrix also indexed in  $n$  (it could depend on the data, but not on  $\theta$ ). Suppose that  $\theta^*$  is the true parameter value. By the Law of Large Numbers,  $\bar{m}_n(\theta^*) \xrightarrow{P} 0$  which also implies  $q_n(\theta^*) \xrightarrow{P} 0$ . The GMM estimator  $\hat{\theta}$  satisfies

$$q_n(\hat{\theta}) = \min_{\theta} q_n(\theta) \quad (13.10.41)$$

$$\leq q_n(\theta^*) \quad (13.10.42)$$

for any  $\theta^*$ . But since  $q_n(\theta^*) \xrightarrow{P} 0$  and moreover  $q_n(\theta) \geq 0$  due to positive definiteness of  $W_n$ , we deduce that  $q_n(\hat{\theta}) \xrightarrow{P} 0$ . Additionally due to positive definiteness of  $W_n$ , if  $q_n(\hat{\theta}) = 0$  then it must be that  $\bar{m}_n(\hat{\theta}) = 0$ . Therefore  $\bar{m}_n(\hat{\theta}) \xrightarrow{P} 0$  as well. Under an ‘identifiability’ regularity condition (which requires that  $\bar{m}_n(\hat{\theta}) = 0$  implies  $\bar{m}_n(\theta^*) = 0$ ), we have consistency:

$$\hat{\theta} \xrightarrow{P} \theta^* \quad (13.10.43)$$

Strong consistency ( $\hat{\theta} \xrightarrow{\text{a.s.}} \theta^*$ ) can also be analogously shown.

### 13.10.5 Asymptotic Normality of Generalised Method of Moments [72]

In addition to consistency, the GMM estimator can be shown to be asymptotically normal. The first order condition can be written out as

$$\nabla_{\theta} q_n(\hat{\theta}) = 2\bar{G}_n(\hat{\theta})^\top W_n \bar{m}_n(\hat{\theta}) \quad (13.10.44)$$

$$= 0 \quad (13.10.45)$$

where  $\bar{G}_n(\hat{\theta})$  denotes the Jacobian of  $\bar{m}_n(\theta)$  with respect to  $\theta$ , evaluated at  $\hat{\theta}$ . So for  $L$  moment conditions and  $K$  parameters, it will have dimension  $L \times K$ . The mean value theorem then says we can Taylor expand  $\bar{m}_n(\hat{\theta})$  as

$$\bar{m}_n(\hat{\theta}) = \bar{m}_n(\theta^*) + \bar{G}_n(\bar{\theta})(\hat{\theta} - \theta^*) \quad (13.10.46)$$

where  $\bar{\theta}$  is some value between  $\hat{\theta}$  and  $\theta^*$ . Substituting this expansion into the first order condition, we get:

$$\bar{G}_n(\hat{\theta})^\top W_n (\bar{m}_n(\theta^*) + \bar{G}_n(\bar{\theta})(\hat{\theta} - \theta^*)) = 0 \quad (13.10.47)$$

$$\bar{G}_n(\hat{\theta})^\top W_n \bar{m}_n(\theta^*) + \bar{G}_n(\hat{\theta})^\top W_n \bar{G}_n(\bar{\theta})(\hat{\theta} - \theta^*) = 0 \quad (13.10.48)$$

$$\hat{\theta} - \theta^* = - \left( \bar{G}_n(\hat{\theta})^\top W_n \bar{G}_n(\bar{\theta}) \right)^{-1} \bar{G}_n(\hat{\theta})^\top W_n \bar{m}_n(\theta^*) \quad (13.10.49)$$

Let

$$G = \mathbb{E} \left[ \frac{\partial}{\partial \theta} m(X_i, \theta^*) \right] \quad (13.10.50)$$

By consistency of  $\hat{\theta}$ , we have  $\bar{\theta} \xrightarrow{P} \theta^*$  as well. Under some regularity relating to continuity and differentiability, it follows that  $\bar{G}_n(\hat{\theta}) \xrightarrow{P} G$  and  $\bar{G}_n(\bar{\theta}) \xrightarrow{P} G$ . Assume also that  $W_n \xrightarrow{P} W$  for some positive definite  $W$ . Then in the probability limit,

$$\hat{\theta} - \theta^* \xrightarrow{d} -\left(G^\top WG\right)^{-1} G^\top W \bar{m}_n(\theta^*) \quad (13.10.51)$$

Using the central limit theorem (under the appropriate qualifying assumptions),  $\sqrt{n}\bar{m}_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, \Omega)$  for some asymptotic covariance matrix  $\Omega$ . Then from

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \left[-\left(G^\top WG\right)^{-1} G^\top W\right] \sqrt{n}\bar{m}_n(\theta^*) \quad (13.10.52)$$

we see that

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \left(G^\top WG\right)^{-1} G^\top W \Omega \left[\left(G^\top WG\right)^{-1} G^\top W\right]^\top\right) \quad (13.10.53)$$

After carrying the  $\sqrt{n}$  factor to the right and simplifying the asymptotic covariance, we get

$$\hat{\theta} - \theta^* \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n} \left(G^\top WG\right)^{-1} G^\top W \Omega W G \left(G^\top WG\right)^{-1}\right) \quad (13.10.54)$$

### 13.10.6 Asymptotic Efficiency of Generalised Method of Moments

If the weighting matrix  $W$  is chosen as  $\Omega^{-1}$  (or even if  $W_n$  is chosen such that it converges in probability to  $\Omega^{-1}$ ), then the GMM can be shown to be asymptotically efficient. That is, we let  $V(W)$  be the asymptotic covariance of the GMM estimator with weighting matrix  $W$ , and show that  $V(\Omega^{-1}) \preceq V(W)$  for any choice of  $W$ . First observe that choosing  $W = \Omega^{-1}$  causes the asymptotic covariance to collapse to

$$V(\Omega^{-1}) = \left(G^\top \Omega^{-1} G\right)^{-1} G^\top \Omega^{-1} \Omega \Omega^{-1} G \left(G^\top \Omega^{-1} G\right)^{-1} \quad (13.10.55)$$

$$= \left(G^\top \Omega^{-1} G\right)^{-1} G^\top \Omega^{-1} G \left(G^\top \Omega^{-1} G\right)^{-1} \quad (13.10.56)$$

$$= \left(G^\top \Omega^{-1} G\right)^{-1} \quad (13.10.57)$$

Thus the difference  $V(W) - V(\Omega^{-1})$  may be factorised as

$$V(W) - V(\Omega^{-1}) = \left(G^\top WG\right)^{-1} G^\top W \Omega W G \left(G^\top WG\right)^{-1} - \left(G^\top \Omega^{-1} G\right)^{-1} \quad (13.10.58)$$

$$= \left(G^\top WG\right)^{-1} \left[ G^\top W \Omega W G - \left(G^\top WG\right) \left(G^\top \Omega^{-1} G\right)^{-1} \left(G^\top WG\right) \right] \left(G^\top WG\right)^{-1} \quad (13.10.59)$$

Factoring out an additional  $G^\top W \Omega^{1/2}$ :

$$V(W) - V(\Omega^{-1}) = \underbrace{\left(G^\top WG\right)^{-1} G^\top W \Omega^{1/2}}_{\Psi} \left[ I - \underbrace{\Omega^{-1/2} G \left(G^\top \Omega^{-1} G\right)^{-1} G^\top \Omega^{-1/2}}_{\Xi} \right] \underbrace{\Omega^{1/2} W G \left(G^\top WG\right)^{-1}}_{\Phi^\top} \quad (13.10.60)$$

$$= \Psi (1 - \Xi) \Psi^\top \quad (13.10.61)$$

We may quickly verify  $\Xi$  is idempotent because

$$\Xi^2 = \Omega^{-1/2} G \left( G^\top \Omega^{-1} G \right)^{-1} G^\top \Omega^{-1/2} \Omega^{-1/2} G \left( G^\top \Omega^{-1} G \right)^{-1} G^\top \Omega^{-1/2} \quad (13.10.62)$$

$$= \Omega^{-1/2} G \left( G^\top \Omega^{-1} G \right)^{-1} G^\top \Omega^{-1} G \left( G^\top \Omega^{-1} G \right)^{-1} G^\top \Omega^{-1/2} \quad (13.10.63)$$

$$= \Omega^{-1/2} G \left( G^\top \Omega^{-1} G \right)^{-1} G^\top \Omega^{-1/2} \quad (13.10.64)$$

$$= \Xi \quad (13.10.65)$$

Hence  $I - \Xi$  is also idempotent since

$$(I - \Xi)^2 = I - 2\Xi + \Xi^2 \quad (13.10.66)$$

$$= I - \Xi \quad (13.10.67)$$

Therefore using the fact that  $I - \Xi$  is also symmetric

$$V(W) - V(\Omega^{-1}) = \Psi(I - \Xi)(I - \Xi)\Psi^\top \quad (13.10.68)$$

$$= [\Psi(I - \Xi)][\Psi(I - \Xi)]^\top \quad (13.10.69)$$

$$\succeq 0 \quad (13.10.70)$$

for any  $\Psi, \Xi$ . It follows that  $V(\Omega^{-1}) \preceq V(W)$ . In practice, a consistent estimate of  $\Omega^{-1}$  may be obtained via a sum of outer products estimator:

$$\widehat{W}_n = \left( \frac{1}{n} \sum_{i=1}^n \nabla_\theta m(X_i, \widehat{\theta}) [\nabla_\theta m(X_i, \widehat{\theta})]^\top \right)^{-1} \quad (13.10.71)$$

with similar justification to that as used in the estimator for the Hessian in maximum likelihood.

### 13.10.7 Sargan-Hansen Overidentifying Restrictions $J$ -Test

Let  $L$  be the number of moment equations in GMM estimation, and let  $K$  be the number of parameters to estimate. If  $K = L$ , then the sample moment equations  $\bar{m}(\widehat{\theta}) = \mathbf{0}$  can be solved exactly, thus the quadratic criterion  $q(\widehat{\theta}) = \bar{m}(\widehat{\theta})^\top W \bar{m}(\widehat{\theta}) = \mathbf{0}$ . If  $L > K$ , this is called the overidentified case. Then we may not be able to fixed when exactly  $\bar{m}(\widehat{\theta}) = \mathbf{0}$ . But when  $L > K$ , we can perform an *overidentifying restrictions* test (also called the Sargan-Hansen  $J$ -test) to test whether the specification of the moment conditions are correct. The null and alternative hypotheses can be posed as:

- $H_0$ : There are  $L$  moment equations satisfying  $\mathbb{E}[m(\theta)] = \mathbf{0}$ .
- $H_A$ : The  $L$  moment equations are not valid; i.e.  $\mathbb{E}[m(\theta)] \neq \mathbf{0}$ .

We consider the following ‘ $J$ -statistic’ [192] which is computed as  $J = nq(\widehat{\theta})$ , with weighting matrix  $W$  chosen as the ‘optimal’ inverse of the estimate of the asymptotic covariance of  $\sqrt{n}\bar{m}(\widehat{\theta})$ , which we denote  $\widehat{\Omega}_n^{-1}$ . This  $J$ -statistic can be written out as [72]:

$$J = \sqrt{n}\bar{m}(\widehat{\theta})^\top \cdot \widehat{\Omega}_n^{-1} \cdot \sqrt{n}\bar{m}(\widehat{\theta}) \quad (13.10.72)$$

Intuitively, if the null is true, then this statistic should be ‘small’, especially in large samples. Since  $\sqrt{n}\bar{m}(\widehat{\theta}) \xrightarrow{d} \mathcal{N}(\sqrt{n}\theta^*, \Omega)$  (by the Central Limit Theorem), we can see that the  $J$ -statistic

is trying to whiten  $\sqrt{n}\bar{m}(\hat{\theta})$  so that  $J \approx \mathbf{z}^\top \mathbf{z}$  where  $\mathbf{z}$  is a standard normal random vector. Hence the  $J$ -statistic will also be a Wald statistic, and  $J$  will be asymptotically chi-squared distributed. To determine the degrees of freedom of  $J$ , we need to look at the first-order condition in the GMM estimation:

$$\nabla_{\theta} q(\hat{\theta}) = \underbrace{\nabla_{\theta} \bar{m}(\hat{\theta})}_{K \times L} \cdot \underbrace{W}_{L \times L} \underbrace{\bar{m}(\hat{\theta})}_{L \times 1} \quad (13.10.73)$$

$$= \underbrace{\mathbf{0}}_{K \times 1} \quad (13.10.74)$$

The matrix  $\nabla_{\theta} \bar{m}(\hat{\theta}) W$  can be viewed as a linear transformation from  $\mathbb{R}^L$  to  $\mathbb{R}^K$ . By definition, the null space of  $\nabla_{\theta} \bar{m}(\hat{\theta}) W$  is the set of solutions that  $\bar{m}(\hat{\theta})$  can take satisfying the first-order condition. That is to say,  $\bar{m}(\hat{\theta})$  will lie in the null space of  $\nabla_{\theta} \bar{m}(\hat{\theta}) W$ . Recall from the Rank-Nullity Theorem that

$$\text{rank}(\nabla_{\theta} \bar{m}(\hat{\theta}) W) + \text{null}(\nabla_{\theta} \bar{m}(\hat{\theta}) W) = \dim(\mathbb{R}^L) \quad (13.10.75)$$

where the nullity  $\text{null}(\nabla_{\theta} \bar{m}(\hat{\theta}) W)$  is the dimension of the null space of  $\nabla_{\theta} \bar{m}(\hat{\theta}) W$ . Implicitly assuming  $\nabla_{\theta} \bar{m}(\hat{\theta}) W$  is full rank, we have

$$\text{null}(\nabla_{\theta} \bar{m}(\hat{\theta}) W) = L - K \quad (13.10.76)$$

Hence  $\bar{m}(\hat{\theta})$  actually lies in an  $(L - K)$ -dimensional subspace. This is why the degrees of freedom of  $J$  is actually  $L - K$ , not  $L$ . We can imagine  $K$  of the  $L$  moment equations being ‘used up’ to fit the parameters (hence they will equal zero), while  $L - K$  of them will be ‘free’ to vary [205]. Thus we test using the  $J \stackrel{\text{asympt.}}{\sim} \chi^2_{L-K}$  distribution, and rejection of the null is evidence in the direction of the moment condition specifications being incorrect.

## Chapter 14

# Machine Learning

### 14.1 Concepts in Machine Learning

#### 14.1.1 Machine Learning Datasets

##### Training Dataset

The training dataset is the dataset of examples used directly in learning, for instance used to fit the parameters of a model by optimising a cost function.

##### Validation Dataset

The validation dataset is the dataset of examples that are used while training (and should be from the same probability distribution as the training dataset), but for the purpose of tuning the hyperparameters of the learning algorithm. The validation dataset is sometimes also known as the development dataset.

##### Test Dataset

The test dataset is a dataset that should be independent from the training and validation datasets, but also drawn from the same probability distribution. The purpose of the test dataset is to provide an unbiased evaluation of the model performance after all training and hyperparameter tuning has taken place.

##### Imbalanced Datasets

##### Data Preprocessing

##### Bias-Variance Decomposition

##### Bias-Variance Decomposition of Mean Square Estimation Error

##### Bias-Variance Decomposition of Mean Square Prediction Error

In a regression problem, suppose we have a training set consisting of the inputs  $(x_1, \dots, x_n)$  and associated labels  $(y_1, \dots, y_n)$ . Denote this combined training data by  $\mathcal{D}$ . Assume that the labels are generated by

$$y_i = f(x_i) + \varepsilon_i \quad (14.1.1)$$

where  $\varepsilon_i$  is noise which satisfies for any given test point  $X_*$ :

$$\mathbb{E}[\varepsilon|X_*] = 0 \quad (14.1.2)$$

i.e. exogeneity, and

$$\text{Var}(\varepsilon|X_*) = \sigma^2 \quad (14.1.3)$$

i.e. homoskedasticity. Note that the exogeneity condition implies

$$\mathbb{E}[\varepsilon] = \mathbb{E}[\mathbb{E}[\varepsilon|X_*]] \quad (14.1.4)$$

$$= 0 \quad (14.1.5)$$

and moreover

$$\mathbb{E}[Y_*|X_*] = \mathbb{E}[f(X_*) + \varepsilon|X_*] \quad (14.1.6)$$

$$= f(X_*) \quad (14.1.7)$$

while the homoskedasticity condition implies

$$\text{Var}(\varepsilon) = \sigma^2 \quad (14.1.8)$$

which together with  $\mathbb{E}[\varepsilon] = 0$ , means that

$$\mathbb{E}[\varepsilon^2] = \sigma^2 \quad (14.1.9)$$

We may choose to apply any learning algorithm on the data  $\mathcal{D}$  to obtain an approximation of the model, denoted  $\hat{f}_{\mathcal{D}}(x)$ . If we use a squared error loss function:

$$L(y, \hat{f}_{\mathcal{D}}(x)) = (y - \hat{f}_{\mathcal{D}}(x))^2 \quad (14.1.10)$$

then our goal is to decompose the expected loss (which is the mean square prediction error) into a bias term and a variance term. To do so, first consider the expected loss on an independent test point  $(X_*, Y_*)$  conditional on the training data:

$$\mathbb{E}_{X_*, Y_*} [L(Y_*, \hat{f}_{\mathcal{D}}(X_*)) | \mathcal{D}] = \mathbb{E}_{X_*, Y_*} [(Y_* - \hat{f}_{\mathcal{D}}(X_*))^2 | \mathcal{D}] \quad (14.1.11)$$

$$= \mathbb{E}_{X_*, Y_*} [(Y_* - f(X_*) + f(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 | \mathcal{D}] \quad (14.1.12)$$

$$= \mathbb{E}_{X_*, Y_*} [(Y_* - f(X_*))^2 | \mathcal{D}] \quad (14.1.13)$$

$$\begin{aligned} &+ 2\mathbb{E}_{X_*, Y_*} [(Y_* - f(X_*)) (f(X_*) - \hat{f}_{\mathcal{D}}(X_*)) | \mathcal{D}] \\ &\quad + \mathbb{E}_{X_*, Y_*} [(f(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 | \mathcal{D}] \\ &= \sigma^2 + \mathbb{E}_{X_*, Y_*} [(f(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 | \mathcal{D}] \end{aligned} \quad (14.1.14)$$

where the cross-term vanishes because

$$\mathbb{E}_{X_*, Y_*} [(Y_* - f(X_*)) (f(X_*) - \hat{f}_{\mathcal{D}}(X_*)) | \mathcal{D}] = \mathbb{E}_{X_*, Y_*} [\mathbb{E}[(Y_* - f(X_*)) (f(X_*) - \hat{f}_{\mathcal{D}}(X_*)) | X_*, \mathcal{D}] | \mathcal{D}] \quad (14.1.15)$$

$$= \mathbb{E}_{X_*, Y_*} [(\mathbb{E}[Y_*|X_*] - f(X_*)) (f(X_*) - \hat{f}_{\mathcal{D}}(X_*)) | \mathcal{D}] \quad (14.1.16)$$

$$= 0 \quad (14.1.17)$$

Also, the  $\sigma^2$  term appears as  $\mathcal{D}$  and  $(X_*, Y_*)$  are assumed independent:

$$\mathbb{E}_{X_*, Y_*} [(Y_* - f(X_*))^2 | \mathcal{D}] = \mathbb{E}_{X_*, Y_*} [(Y_* - f(X_*))^2] \quad (14.1.18)$$

$$= \mathbb{E} [\varepsilon^2] \quad (14.1.19)$$

$$= \sigma^2 \quad (14.1.20)$$

Define  $\bar{f}(X_*)$  as the prediction averaged over the distribution which the training data was drawn from:

$$\bar{f}(X_*) := \mathbb{E}_{\mathcal{D}} [\hat{f}_{\mathcal{D}}(X_*) | X_*] \quad (14.1.21)$$

Now from the decomposition

$$(f(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 = (f(X_*) - \bar{f}(X_*) + \bar{f}(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 \quad (14.1.22)$$

$$= (f(X_*) - \bar{f}(X_*))^2 + 2(f(X_*) - \bar{f}(X_*))(\bar{f}(X_*) - \hat{f}_{\mathcal{D}}(X_*)) + (\bar{f}(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 \quad (14.1.23)$$

and taking the expectation over the cross-term, we see that this also vanishes:

$$\mathbb{E}_{\mathcal{D}, X_*} [(f(X_*) - \bar{f}(X_*))(\bar{f}(X_*) - \hat{f}_{\mathcal{D}}(X_*))] = \mathbb{E}_{X_*} [\mathbb{E}_{\mathcal{D}} [(f(X_*) - \bar{f}(X_*))(\bar{f}(X_*) - \hat{f}_{\mathcal{D}}(X_*)) | X_*]] \quad (14.1.24)$$

$$= \mathbb{E}_{X_*} \left[ (f(X_*) - \bar{f}(X_*)) \left( \cancel{\bar{f}(X_*)} - \mathbb{E}_{\mathcal{D}} [\hat{f}_{\mathcal{D}}(X_*) | X_*] \right) \right] \quad (14.1.25)$$

$$= 0 \quad (14.1.26)$$

Therefore the expected loss over both the training data and the test point is given by

$$\mathbb{E}_{\mathcal{D}, X_*, Y_*} [(Y_* - \hat{f}_{\mathcal{D}}(X_*))^2] = \mathbb{E}_{\mathcal{D}} \left[ \mathbb{E}_{X_*, Y_*} \left[ (Y_* - \hat{f}_{\mathcal{D}}(X_*))^2 | \mathcal{D} \right] \right] \quad (14.1.27)$$

$$= \sigma^2 + \mathbb{E}_{\mathcal{D}} \left[ \mathbb{E}_{X_*} \left[ (f(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 | \mathcal{D} \right] \right] \quad (14.1.28)$$

$$= \sigma^2 + \mathbb{E}_{\mathcal{D}, X_*} \left[ (f(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 \right] \quad (14.1.29)$$

$$= \sigma^2 + \mathbb{E}_{\mathcal{D}, X_*} [(f(X_*) - \bar{f}(X_*))^2] + \mathbb{E}_{\mathcal{D}, X_*} [(\bar{f}(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2] \quad (14.1.30)$$

If we condition both sides on a test input  $X_*$ , then

$$\mathbb{E}_{\mathcal{D}, Y_*} \left[ (Y_* - \hat{f}_{\mathcal{D}}(X_*))^2 | X_* \right] = \sigma^2 + \mathbb{E}_{\mathcal{D}} \left[ (f(X_*) - \bar{f}(X_*))^2 | X_* \right] + \mathbb{E}_{\mathcal{D}} \left[ (\bar{f}(X_*) - \hat{f}_{\mathcal{D}}(X_*))^2 | X_* \right] \quad (14.1.31)$$

$$= \sigma^2 + (f(X_*) - \bar{f}(X_*))^2 + \mathbb{E}_{\mathcal{D}} \left[ (\hat{f}_{\mathcal{D}}(X_*) - \mathbb{E}_{\mathcal{D}} [\hat{f}_{\mathcal{D}}(X_*) | X_*])^2 | X_* \right] \quad (14.1.32)$$

$$= \sigma^2 + \underbrace{(f(X_*) - \mathbb{E}_{\mathcal{D}} [\hat{f}_{\mathcal{D}}(X_*) | X_*])^2}_{b^2} + \underbrace{\text{Var}_{\mathcal{D}} (\hat{f}_{\mathcal{D}}(X_*) | X_*)}_{\nu} \quad (14.1.33)$$

This shows that the expected loss of a prediction conditioned at  $X_*$  can be decomposed into the bias  $b$  squared, the variance of the prediction  $\nu$ , and an irreducible error  $\sigma^2$ .

## Bias-Variance Tradeoff

### 14.1.2 Cross-Validation

#### Leave-One-Out Cross-Validation

#### *K*-fold Cross-Validation

In *K*-fold cross-validation, we partition the sample into *K* equally-sized (or nearly equally-sized) subsamples. For each subsample  $k = 1, \dots, K$ , we do the following. We train the model on all except the  $k^{\text{th}}$  subsample, and then we evaluate model fit on the  $k^{\text{th}}$  subsample to obtain an estimate  $E_k$  of model performance. Averaging over all *K* folds, the overall estimate of model performance is given by

$$E = \sum_{k=1}^K E_k \quad (14.1.34)$$

### 14.1.3 Machine Learning Models [101]

#### Discriminative Models

A discriminative probability model for data drawn from the distribution of  $(\mathbf{X}, \mathbf{y})$  can be thought of as a model which only models the conditional distribution  $p(\mathbf{y}|\mathbf{X})$ . Thus, a discriminative model is only interested in the mapping from  $\mathbf{X}$  to  $\mathbf{y}$ . Given  $\mathbf{X}$ , we can use the discriminative model to predict or generate  $\mathbf{y}$ .

#### Generative Models

In contrast to discriminative probability models, a generative probability model for data drawn from the distribution of  $(\mathbf{X}, \mathbf{y})$  can be thought of as a model for the entire joint distribution  $p(\mathbf{X}, \mathbf{y})$ . In a way, this supersedes a discriminative model because the conditional distribution  $p(\mathbf{y}|\mathbf{X})$  can be obtained from the joint distribution. Thus we can use a generative model to not only predict/generate  $\mathbf{y}$  given  $\mathbf{X}$ , but we can also use it to generate whole samples of  $(\mathbf{X}, \mathbf{y})$ .

## 14.2 Statistical Classification

### 14.2.1 Performance Metrics in Classification

When we use a classifier to classify a data set of size  $n$  with binary ('positive' and 'negative') labels, there is a distinction between the classified positive/negative values and the 'actual' positive/negative values. In terms of actual values, we can decompose

$$n = N + P \quad (14.2.1)$$

where

- $P$  is the number of actual positive cases in the dataset.
- $N$  is the number of actual negative cases in the dataset.

Similarly, we can alternatively decompose in terms of classified values:

$$n = N' + P' \quad (14.2.2)$$

where

- $P'$  is the number of classified positive cases in the dataset.

- $N'$  is the number of classified negative cases in the dataset.

For the positive and negative classified cases, we can further decompose:

$$P' = TP + FP \quad (14.2.3)$$

$$N' = TN + FN \quad (14.2.4)$$

where

- TP is the number of *true positives* - where the case was classified positive and it was actually positive (the correct classification decision).
- FP is the number of *false positives* - where the case was classified positive when it was actually negative (analogous to making a Type I error).
- TN is the number of *true negatives* - where the case was classified negative and it was actually negative (the correct classification decision).
- FN is the number of *false negatives* - where the case was classified negative when it was actually positive (analogous to making a Type II error).

In terms of actual values, the appropriate decomposition is given by

$$P = TP + FN \quad (14.2.5)$$

$$N = TN + FP \quad (14.2.6)$$

We can define several performance metrics for classification using these elements.

## Recall

The recall is also known as the *sensitivity*, or true positive rate (TPR). This is defined as

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (14.2.7)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the recall is analogous to

$$\frac{\Pr(D \cap E)}{\Pr(E)} = \Pr(D|E) \quad (14.2.8)$$

which is also analogous to the power of a statistical test. A higher recall for a classifier is regarded as better (all else equal).

## Specificity

Specificity is the true negative rate (TNR), defined by

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (14.2.9)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the specificity is analogous to

$$\frac{\Pr(\overline{D} \cap \overline{E})}{\Pr(\overline{E})} = \Pr(\overline{D}|\overline{E}) \quad (14.2.10)$$

A higher specificity for a classifier is regarded as better (all else equal).

### Positive Predictive Value

The positive predictive value (PPV), also known as the *precision* of a classifier (not to be confused with precision as the inverse of variance), is defined as

$$\text{PPV} = \frac{\text{TP}}{\text{P}'} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (14.2.11)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the precision is analogous to

$$\frac{\Pr(D \cap E)}{\Pr(D)} = \Pr(E|D) \quad (14.2.12)$$

A higher precision for a classifier is regarded as better (all else equal).

### Negative Predictive Value

The negative predictive value (NPV), is defined as

$$\text{NPV} = \frac{\text{TN}}{\text{N}'} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (14.2.13)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the negative predictive value is analogous to

$$\frac{\Pr(\overline{D} \cap \overline{E})}{\Pr(\overline{D})} = \Pr(\overline{E}|\overline{D}) \quad (14.2.14)$$

A higher negative predictive value for a classifier is regarded as better (all else equal).

### False Negative Rate

The false negative rate FNR is also known as the miss rate, and is defined by

$$\text{FNR} = \frac{\text{FN}}{\text{P}} = \frac{\text{FN}}{\text{TP} + \text{FN}} \quad (14.2.15)$$

Also note that FNR is the complement of TPR:

$$\text{FNR} = 1 - \text{TPR} \quad (14.2.16)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the false negative rate is analogous to

$$\frac{\Pr(\overline{D} \cap E)}{\Pr(E)} = \Pr(\overline{D}|E) \quad (14.2.17)$$

which is also analogous to the probability of Type II error of a statistical test. A smaller false negative rate for a classifier is regarded as better (all else equal).

### False Positive Rate

The false positive rate FPR is also known as the *fall-out*, and is defined by

$$\text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (14.2.18)$$

Also note that FPR is the complement of TNR:

$$\text{FPR} = 1 - \text{TNR} \quad (14.2.19)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the false positive rate is analogous to

$$\frac{\Pr(D \cap \bar{E})}{\Pr(\bar{E})} = \Pr(D|\bar{E}) \quad (14.2.20)$$

which is also analogous to the probability of Type I error of a statistical test. A smaller false positive rate for a classifier is regarded as better (all else equal).

### False Discovery Rate

The false discovery rate FDR is defined as

$$\text{FDR} = \frac{\text{FP}}{\text{P}'} = \frac{\text{FP}}{\text{TP} + \text{FP}} \quad (14.2.21)$$

Also note that FDR is the complement of PPV:

$$\text{FDR} = 1 - \text{PPV} \quad (14.2.22)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the false positive rate is analogous to

$$\frac{\Pr(D \cap \bar{E})}{\Pr(D)} = \Pr(\bar{E}|D) \quad (14.2.23)$$

A smaller false omission rate for a classifier is regarded as better (all else equal).

### False Omission Rate

The false discovery rate FOR is defined as

$$\text{FOR} = \frac{\text{FN}}{\text{N}'} = \frac{\text{FN}}{\text{TN} + \text{FN}} \quad (14.2.24)$$

Also note that FOR is the complement of NPV:

$$\text{FOR} = 1 - \text{NPV} \quad (14.2.25)$$

If we let  $E$  represent the event to be detected and  $D$  the event of a detection, then the false positive rate is analogous to

$$\frac{\Pr(\bar{D} \cap E)}{\Pr(\bar{D})} = \Pr(E|\bar{D}) \quad (14.2.26)$$

A smaller false discovery rate for a classifier is regarded as better (all else equal).

### Accuracy

The accuracy ACC of a classifier is the proportion of examples that the classifier gets correct:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{n} \quad (14.2.27)$$

which is analogous to the probability  $\Pr((D \cap E) \cup (\bar{D} \cap \bar{E}))$ . A higher accuracy is regarded as better, all else equal.

### Error Rate

The error rate of a classifier is the proportion of examples that the classifier gets incorrect, which is the complement of accuracy, i.e.  $1 - \text{ACC}$ .

## Prevalence

The *prevalence* of a classifier refers to the proportion of actual positive cases in the data, i.e.  $\frac{P}{n}$ . It is analogous to the probability  $\Pr(E)$ .

## Positive Likelihood Ratio

The positive likelihood ratio  $\text{LR}_+$  is defined as

$$\text{LR}_+ = \frac{\text{TPR}}{\text{FPR}} \quad (14.2.28)$$

This is analogous to the odds ratio (or likelihood ratio):

$$\frac{\Pr(D|E)}{\Pr(D|\bar{E})} = \frac{\mathcal{L}(E; D)}{\mathcal{L}(\bar{E}; D)} \quad (14.2.29)$$

This is a way to combine the TPR and FPR metrics such that a higher positive likelihood ratio for a classifier is better, all else equal.

## Negative Likelihood Ratio

The negative likelihood ratio  $\text{LR}_-$  is defined as

$$\text{LR}_- = \frac{\text{FNR}}{\text{TNR}} \quad (14.2.30)$$

This is analogous to the odds ratio (or likelihood ratio):

$$\frac{\Pr(\bar{D}|E)}{\Pr(\bar{D}|\bar{E})} = \frac{\mathcal{L}(E; \bar{D})}{\mathcal{L}(\bar{E}; \bar{D})} \quad (14.2.31)$$

This is a way to combine the FNR and TNR metrics such that a smaller negative likelihood ratio for a classifier is better, all else equal.

## Diagnostic Odds Ratio

The diagnostic odds ratio DOR combines the  $\text{LR}_+$  and  $\text{LR}_-$  into

$$\text{DOR} = \frac{\text{LR}_+}{\text{LR}_-} \quad (14.2.32)$$

such that the higher diagnostic odds ratio for a classifier is better, all else equal.

## $F_1$ -score

Given precision PPV and recall TPR, an ' $F_1$ -score' can be calculated by the harmonic mean of PPV and TPR (since the harmonic mean is suitable for averaging over ratios), that is:

$$\frac{1}{F_1} = \frac{1}{2} \left( \frac{1}{\text{PPV}} + \frac{1}{\text{TPR}} \right) \quad (14.2.33)$$

$$F_1 = \frac{2\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}} \quad (14.2.34)$$

This is a way to quantitatively trade between precision and recall, where a model with higher  $F_1$ -score is preferred.

**$F_\beta$ -score [179]**

The  $F_\beta$ -score generalises the  $F_1$ -score, where  $\beta^2$  times more weighting is attached to recall than to precision. It is calculated by

$$\frac{1}{F_\beta} = \frac{1}{1 + \beta^2} \left( \frac{1}{\text{PPV}} + \frac{\beta^2}{\text{TPR}} \right) \quad (14.2.35)$$

$$F_\beta = \frac{1 + \beta^2}{1/\text{PPV} + \beta^2/\text{TPR}} \quad (14.2.36)$$

**Duality with Performance Metrics in Binary Classification**

If the positive and negative labels are swapped around, then for the new dataset (denoted by overlines) in terms of the old dataset:

$$\bar{P} = N \quad (14.2.37)$$

$$\bar{N} = P \quad (14.2.38)$$

$$\bar{P}' = N' \quad (14.2.39)$$

$$\bar{N}' = P' \quad (14.2.40)$$

and

$$\bar{\text{TP}} = \text{TN} \quad (14.2.41)$$

$$\bar{\text{TN}} = \text{TP} \quad (14.2.42)$$

$$\bar{\text{FP}} = \text{FN} \quad (14.2.43)$$

$$\bar{\text{FN}} = \text{FP} \quad (14.2.44)$$

In this way, we can show that the new  $F_1$  score after swapping around positive and negative labels is the same as taking the harmonic mean of the original TNR and NPV, since:

$$\bar{\text{PPV}} = \frac{\bar{\text{TP}}}{\bar{P}'} \quad (14.2.45)$$

$$= \frac{\text{TN}}{\text{N}'} \quad (14.2.46)$$

$$= \text{NPV} \quad (14.2.47)$$

and similarly

$$\bar{\text{TPR}} = \frac{\bar{\text{TP}}}{\bar{P}} \quad (14.2.48)$$

$$= \frac{\text{TN}}{\text{N}} \quad (14.2.49)$$

$$= \text{TNR} \quad (14.2.50)$$

**Coverage [69]**

A classifier may also be coded so that it can refuse to make a decision (e.g. if it is not certain enough about a particular example). The coverage is defined as the fraction of examples for which the classifier is able to make a decision.

### 14.2.2 Confusion Matrices

A confusion matrix is a special type of  $2 \times 2$  contingency table which tabulates the true positives, false positives, true negatives and false negatives.

|                              | Actual condition positive | Actual condition negative |    |
|------------------------------|---------------------------|---------------------------|----|
| Predicted condition positive | TP                        | FP                        | P' |
| Predicted condition negative | FN                        | TN                        | N' |
|                              | P                         | N                         | n  |

### 14.2.3 Receiver Operating Characteristic

The receiver operating characteristic (ROC), sometimes also referred to as a receiver/relative operator/operating curve, is a graph of TPR against FPR for a classifier. Suppose a classifier uses a classification rule that classifies an example as positive if some statistic  $X > t$  where  $t$  is a threshold (for example,  $X$  could be a likelihood ratio as in **maximum likelihood binary hypothesis testing**). Assume that the conditional distribution given actual positive label (the conditional distribution of  $[X|Y = 1]$ ) has density  $f_1(x)$ . Let the conditional distribution given actual negative label (the conditional distribution of  $[X|Y = -1]$ ) have density  $f_0(x)$ . Then the recall (i.e. true positive rate) written as a probability parametrised in terms of  $t$  is

$$\text{TPR}(t) = \Pr(X > t|Y = 1) \quad (14.2.51)$$

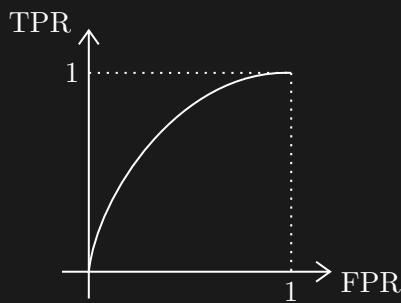
$$= \int_t^\infty f_1(x) dx \quad (14.2.52)$$

Likewise, the false positive rate (i.e. analogous to probability of Type I error) parametrised in terms of  $t$  is

$$\text{FPR}(t) = \Pr(X > t|Y = -1) \quad (14.2.53)$$

$$= \int_t^\infty f_0(x) dx \quad (14.2.54)$$

Observe that as  $t \rightarrow \infty$ , we see  $\text{TPR}(t) \rightarrow 0$  and  $\text{FPR}(t) \rightarrow 0$ . Conversely as  $t \rightarrow -\infty$ , then  $\text{TPR}(t) \rightarrow 1$  and  $\text{FPR}(t) \rightarrow 1$ . Sketching out this parametric curve gives the ROC.



### Area Under Receiver Operating Characteristic

The area under the receiver operator curve (AUC) can be specified as

$$\text{AUC} = \int_{p=0}^{p=1} \text{TPR}(\text{FPR}^{-1}(p)) dp \quad (14.2.55)$$

After a change of variables  $t = \text{FPR}^{-1}(p)$  with  $\text{FPR}(t) = p$  and  $\frac{dp}{dt} = \text{FPR}'(t)$ , we get

$$\text{AUC} = \int_{\infty}^{-\infty} \text{TPR}(t) \text{FPR}'(t) dt \quad (14.2.56)$$

$$= \int_{-\infty}^{-\infty} \text{TPR}(t) \frac{d}{dt} \left( \int_t^{\infty} f_0(s) ds \right) dt \quad (14.2.57)$$

$$= \int_{-\infty}^{-\infty} \text{TPR}(t) \frac{d}{dt} \left( 1 - \int_{-\infty}^t f_0(s) ds \right) dt \quad (14.2.58)$$

$$= - \int_{-\infty}^{-\infty} \text{TPR}(t) f_0(t) dt \quad (14.2.59)$$

after applying the definition of  $\text{FPR}(t)$  above and the Fundamental Theorem of Calculus. Then swapping the terminals applying the definition of  $\text{TPR}(t)$ :

$$\text{AUC} = \int_{-\infty}^{\infty} \left( \int_t^{\infty} f_1(r) dr \right) f_0(t) dt \quad (14.2.60)$$

$$= \int_{-\infty}^{\infty} \int_t^{\infty} f_1(r) f_0(t) dr dt \quad (14.2.61)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}_{\{r>t\}} f_1(r) f_0(t) dr dt \quad (14.2.62)$$

$$= \mathbb{E} [\mathbb{I}_{\{X_1 > X_0\}}] \quad (14.2.63)$$

where we define  $X_1$  and  $X_0$  as independent random variables such that  $X_1$  has density  $f_1(x)$  and  $X_0$  has density  $f_0(x)$ . Thus the AUC can be characterised by

$$A = \Pr(X_1 > X_0) \quad (14.2.64)$$

That is, the AUC represents the probability that the statistic for a positive example is greater than a negative example. Hence a classifier with a larger AUC should be regarded as better, all else equal.

### Receiver Operating Characteristic in Statistical Hypothesis Testing [220]

Receiver operator curves can also be used in the context of statistical hypothesis testing, using the analogy between binary hypothesis testing and binary classification. We consider null and alternative hypotheses  $H_0$  and  $H_1$  respectively, and for decision rule that rejects the null when the test statistic  $X > t$  for some threshold  $t$ , recall that we can write the ‘miss probability’  $P_{\text{MISS}}$  (probability of Type II error) and ‘false alarm’ probability  $P_{\text{FA}}$  (probability of Type I error) as

$$P_{\text{MISS}} = \Pr(X \leq t | H_1) \quad (14.2.65)$$

$$P_{\text{FA}} = \Pr(X > t | H_0) \quad (14.2.66)$$

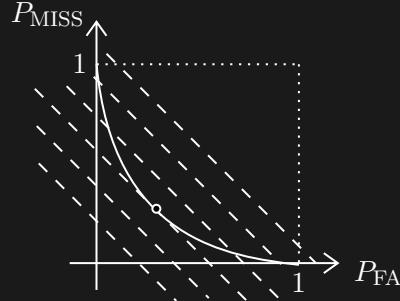
Recognise that  $P_{\text{FA}}$  is analogous to  $\text{FPR}$ , and  $P_{\text{MISS}}$  is analogous to  $\text{FNR}$ , which is the complement of  $\text{TNR}$ . Hence a receiver operating characteristic can be constructed in a similar way to above, except the vertical-axis values will be reflected about 0.5. Recall that each decision rule corresponds to a point on the ROC. Moreover, using the minimum-cost hypothesis testing framework, a decision rule can be derived by minimising a linear combination of  $P_{\text{MISS}}$  and  $P_{\text{FA}}$ . Consider linear curves of the form

$$c = P_{\text{MISS}} + m \cdot P_{\text{FA}} \quad (14.2.67)$$

where the intercept  $c$  represents the cost (up to a positive scaling). By deriving the minimum-cost threshold, this can be graphically interpreted as finding the curve from the family of isolines (constant slope)

$$P_{\text{MISS}} = c - m \cdot P_{\text{FA}} \quad (14.2.68)$$

with the lowest intercept. This curve must still intersect the ROC, and it will do so at a tangent (typically where the ROC has a slope of  $-m$ ).



#### 14.2.4 $k$ -Nearest Neighbours [80]

The  $k$ -nearest neighbours algorithm is a nonparametric approach for classification. The estimate of target  $\hat{Y}$  for feature vector  $\mathbf{x}$  is defined by

$$\hat{Y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i \quad (14.2.69)$$

where  $N_k(\mathbf{x})$  denotes a neighbourhood of  $\mathbf{x}$  consisting of the  $k$  closest points in the training set to  $\mathbf{x}$  with respect to some metric for distance (such as Euclidean distance), and  $y_i$  are the corresponding labels. Essentially, this takes the average (i.e. proportion) of positive labels after finding the  $k$  closest points to  $\mathbf{x}$ . A hard classification can be made using a threshold rule, such as  $\hat{Y} > 0.5$ . The value of  $k$  is a hyperparameter that may be set by the practitioner.

#### 14.2.5 Linear Discriminant Analysis

Suppose we have random variable  $Y$  belonging to two classes, and observe a random vector  $\mathbf{X}$  which is conditionally Gaussian distributed as:

$$[\mathbf{X}|Y=0] \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) \quad (14.2.70)$$

$$[\mathbf{X}|Y=1] \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1) \quad (14.2.71)$$

A classification rule can be formed by thresholding the log-likelihood ratio given  $\mathbf{X} = \mathbf{x}$  as:

$$(\mathbf{x} - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + \log \det(\Sigma_0) - (\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - \log \det(\Sigma_1) > t \quad (14.2.72)$$

for some threshold  $t$ , where we classify  $Y = 1$  if the log-likelihood ratio is greater than  $t$ . In linear discriminant analysis, we assume homoskedasticity (i.e.  $\text{Cov}(\mathbf{X}|Y=0) = \text{Cov}(\mathbf{X}|Y=1)$ ), and we denote

$$\Sigma_0 = \Sigma_1 = \Sigma \quad (14.2.73)$$

After some cancellation in the log-likelihood ratio above, we see that

$$\begin{aligned} & (\mathbf{x} - \boldsymbol{\mu}_0)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + \log \det(\Sigma) - (\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - \log \det(\Sigma) \\ &= -2\boldsymbol{\mu}_0^\top \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}_0^\top \Sigma^{-1} \boldsymbol{\mu}_0 + 2\boldsymbol{\mu}_1^\top \Sigma^{-1} \mathbf{x} - \boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1 \end{aligned} \quad (14.2.74)$$

and the classification rule can be rearranged to

$$2(\boldsymbol{\mu}_0^\top - \boldsymbol{\mu}_1^\top) \Sigma^{-1} \mathbf{x} > t - \boldsymbol{\mu}_0^\top \Sigma^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1 \quad (14.2.75)$$

and simplified as

$$\mathbf{w}^\top \mathbf{x} > c \quad (14.2.76)$$

where

$$\mathbf{w} = \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \quad (14.2.77)$$

$$c = \frac{1}{2} \left( t - \boldsymbol{\mu}_0^\top \Sigma^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1 \right) \quad (14.2.78)$$

To make this classifier operational in practice, we would require estimates of  $\boldsymbol{\mu}_0$ ,  $\boldsymbol{\mu}_1$  and  $\Sigma$  from a training dataset.

### Fisher's Linear Discriminant

The linear discriminant analysis classifier can be derived in another way using Fisher's linear discriminant. The idea is to project the observed  $\mathbf{x}$  onto a line, giving  $u = \mathbf{w}^\top \mathbf{x}$ , and then perform the classification on the scalar  $u$  using a threshold rule. We would like to choose  $\mathbf{w}$  so that the two groups of projected data have the greatest separation in their means relative to their spread. Denote

$$\boldsymbol{\mu}_0 = \mathbb{E}[\mathbf{X}|Y=0] \quad (14.2.79)$$

$$\boldsymbol{\mu}_1 = \mathbb{E}[\mathbf{X}|Y=1] \quad (14.2.80)$$

and

$$\Sigma = \text{Cov}(\mathbf{X}) \quad (14.2.81)$$

Note that we have dropped the assumption of conditional normality of  $\mathbf{X}$ , and we no longer require homoskedasticity;  $\Sigma$  is simply the unconditional covariance of  $\mathbf{X}$ . Letting  $U = \mathbf{w}^\top \mathbf{X}$ , we define the amount of separation by

$$R(\mathbf{w}) = \frac{(\mathbb{E}[U|Y=1] - \mathbb{E}[U|Y=0])^2}{\text{Var}(U)} \quad (14.2.82)$$

$$= \frac{(\mathbf{w}^\top \boldsymbol{\mu}_1 - \mathbf{w}^\top \boldsymbol{\mu}_0)^2}{\mathbf{w}^\top \Sigma \mathbf{w}} \quad (14.2.83)$$

$$= \frac{[\mathbf{w}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)]^2}{\mathbf{w}^\top \Sigma \mathbf{w}} \quad (14.2.84)$$

$$= \frac{\mathbf{w}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \mathbf{w}}{\mathbf{w}^\top \Sigma \mathbf{w}} \quad (14.2.85)$$

Defining  $\Sigma_\boldsymbol{\mu} := (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top$ , we have

$$R(\mathbf{w}) = \frac{\mathbf{w}^\top \Sigma_\boldsymbol{\mu} \mathbf{w}}{\mathbf{w}^\top \Sigma \mathbf{w}} \quad (14.2.86)$$

which is a ratio of quadratic forms. This is sometimes called the Rayleigh quotient. We seek to find the optimal projection  $\mathbf{w}^*$  which is the maximiser of  $R(\mathbf{w})$ :

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} R(\mathbf{w}) \quad (14.2.87)$$

We can take the gradient using the quotient rule:

$$\nabla_{\mathbf{w}} R(\mathbf{w}) = \frac{2\Sigma_\boldsymbol{\mu} \mathbf{w} (\mathbf{w}^\top \Sigma \mathbf{w}) - 2\Sigma \mathbf{w} (\mathbf{w}^\top \Sigma_\boldsymbol{\mu} \mathbf{w})}{(\mathbf{w}^\top \Sigma \mathbf{w})^2} \quad (14.2.88)$$

Setting the gradient to zero, the optimal  $\mathbf{w}^*$  satisfies

$$\Sigma_\boldsymbol{\mu} \mathbf{w}^* (\mathbf{w}^{*\top} \Sigma \mathbf{w}^*) = \Sigma \mathbf{w}^* (\mathbf{w}^{*\top} \Sigma_\boldsymbol{\mu} \mathbf{w}^*) \quad (14.2.89)$$

$$\Sigma_\boldsymbol{\mu} \mathbf{w}^* = R(\mathbf{w}^*) \Sigma \mathbf{w}^* \quad (14.2.90)$$

This takes the form of the generalised eigenvalue problem. It can be solved by rearranging

$$\Sigma^{-1} \Sigma_\boldsymbol{\mu} \mathbf{w}^* = R(\mathbf{w}^*) \mathbf{w}^* \quad (14.2.91)$$

where we can now see that the maximum separation occurs when  $\mathbf{w}^*$  is the eigenvector corresponding to the largest eigenvalue of  $\Sigma^{-1}\Sigma_{\boldsymbol{\mu}}$ , and this largest eigenvalue is equal to  $R(\mathbf{w}^*)$ . Then note that by definition

$$\Sigma_{\boldsymbol{\mu}}\mathbf{w}^* = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\top} \mathbf{w}^* \quad (14.2.92)$$

which is in the same direction as  $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ , since  $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\top} \mathbf{w}^*$  evaluates to a scalar. Since as an eigenvector the choice of scale for  $\mathbf{w}^*$  is arbitrary, then using  $\Sigma^{-1}\Sigma_{\boldsymbol{\mu}}\mathbf{w}^* = R(\mathbf{w}^*)\mathbf{w}^*$  we can simply take

$$\mathbf{w}^* = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \quad (14.2.93)$$

This yields the Fisher linear discriminant function

$$f(\mathbf{x}) = u \quad (14.2.94)$$

$$= \mathbf{w}^{*\top} \mathbf{x} \quad (14.2.95)$$

$$= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\top} \Sigma^{-1} \quad (14.2.96)$$

and the classification rule to classify  $Y = 1$  if  $f(\mathbf{x})$  is greater than some threshold  $c$  (since the way  $R(\mathbf{w})$  is setup means that a higher  $f(\mathbf{x})$  means that the projected feature vector is closer to the projected mean  $\boldsymbol{\mu}_1$  than it is to the projected mean  $\boldsymbol{\mu}_0$ ). Reversing the signs in the term  $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$  would also reverse the classification rule. Finally recognise that if we set the threshold  $c$  to

$$c = \frac{1}{2} \left( t - \boldsymbol{\mu}_0^{\top} \Sigma^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1^{\top} \Sigma^{-1} \boldsymbol{\mu}_1 \right) \quad (14.2.97)$$

then we recover the linear discriminant analysis classification rule. As usual, we can replace  $\boldsymbol{\mu}_0$ ,  $\boldsymbol{\mu}_1$  and  $\Sigma$  by estimates from a training dataset to obtain an operational classifier.

### 14.2.6 Quadratic Discriminant Analysis

Quadratic discriminant analysis proceeds from the setup of linear discriminant analysis, except we do not assume homoskedasticity. In that case, the classification rule for  $Y = 1$  is left as

$$(\mathbf{x} - \boldsymbol{\mu}_0)^{\top} \Sigma_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + \log \det(\Sigma_0) - (\mathbf{x} - \boldsymbol{\mu}_1)^{\top} \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - \log \det(\Sigma_1) > t \quad (14.2.98)$$

### 14.2.7 Support Vector Machines

#### Support Vector Machines for Linearly Separable Data [25]

Suppose our training data consists of  $n$  pairs of  $(\mathbf{x}, y)$  with  $\mathbf{x} \in \mathbb{R}^d$  (which can be in some appropriate basis) and labels  $y \in \{-1, 1\}$ . Learning a support vector machine (SVM) involves learning a ‘linear discriminant function’

$$f(\mathbf{x}) = \mathbf{w}^{\top} \mathbf{x} + b \quad (14.2.99)$$

so that a classification is made by taking the sign:

$$\hat{y}(\mathbf{x}) = \text{sgn}(f(\mathbf{x})) \quad (14.2.100)$$

For now, we assume that the data is linearly separable, which means that there does exist vectors  $\mathbf{w}$ ,  $\mathbf{b}$  such that  $y_i \cdot \hat{y}(\mathbf{x}_i) > 0$  or equivalently  $y_i \cdot f(\mathbf{x}_i) > 0$  for all  $i = 1, \dots, n$  (i.e. the signs of  $y_i$  and  $\hat{y}(\mathbf{x}_i)$  agree for the entire dataset). We claim that the perpendicular distance between any point  $\mathbf{x}$  and the  $(d-1)$ -dimensional hyperplane  $\mathbf{w}^{\top} \mathbf{x} + b = 0$  is given by  $\frac{|f(\mathbf{x})|}{\|\mathbf{x}\|}$ . This can

be explained as follows. Consider distinct points  $\mathbf{x}_1$  and  $\mathbf{x}_2$  which lie on the hyperplane. Then the vector  $\mathbf{x}_2 - \mathbf{x}_1$  is lying on the hyperplane, and by  $f(\mathbf{x}_1) = f(\mathbf{x}_2) = 0$  we have

$$f(\mathbf{x}_2) - f(\mathbf{x}_1) = \mathbf{w}^\top \mathbf{x}_2 + b - \mathbf{w}^\top \mathbf{x}_1 - b \quad (14.2.101)$$

$$= \mathbf{w}^\top (\mathbf{x}_2 - \mathbf{x}_1) \quad (14.2.102)$$

$$= 0 \quad (14.2.103)$$

Hence this means the vector  $\mathbf{w}$  is normal to the hyperplane  $f(\mathbf{x}) = 0$ . Then let  $\mathbf{x}$  be an arbitrary point and let  $\mathbf{x}_\perp$  be its orthogonal projection on the hyperplane. That is to say,

$$\mathbf{x} = \mathbf{x}_\perp + r \frac{\mathbf{w}}{\|\mathbf{w}\|} \quad (14.2.104)$$

where  $\mathbf{x}_\perp$  lies on the hyperplane and  $r \frac{\mathbf{w}}{\|\mathbf{w}\|}$  is the perpendicular vector from the hyperplane to  $\mathbf{x}$ . Thus our aim is to find the absolute perpendicular distance  $|r|$ . Left-multiplying both sides by  $\mathbf{w}^\top$  and adding  $b$ , we see that

$$\mathbf{w}^\top \mathbf{x} + b = \mathbf{w}^\top \mathbf{x}_\perp + b + r \frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|} \quad (14.2.105)$$

$$f(\mathbf{x}) = 0 + r \frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \quad (14.2.106)$$

$$r = \frac{f(\mathbf{x})}{\|\mathbf{w}\|} \quad (14.2.107)$$

$$|r| = \frac{|f(\mathbf{x})|}{\|\mathbf{w}\|} \quad (14.2.108)$$

For a given separating hyperplane, we define the ‘margin’ as the minimum perpendicular distance from the hyperplane to a point, given by  $\min_i \frac{|f(\mathbf{x}_i)|}{\|\mathbf{w}\|}$ . This we have assumed for the separating hyperplane that  $y_i f(\mathbf{x}_i) > 0$  for all  $i$ , we can modify the margin by rewriting it as  $\min_i \frac{y_i (\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|}$ . The objective in learning an SVM classifier then becomes to find the  $\mathbf{w}, b$  which maximises the margin:

$$(\mathbf{w}^*, \mathbf{b}^*) = \underset{\mathbf{w}, b}{\operatorname{argmax}} \left\{ \frac{1}{\|\mathbf{w}\|} \min_i \left\{ y_i (\mathbf{w}^\top \mathbf{x}_i + b) \right\} \right\} \quad (14.2.109)$$

This optimisation problem can be difficult to solve, however it can be reformulated as follows. Note that scaling  $\mathbf{w}$  and  $b$  by some constant  $\kappa > 0$  does not change the hyperplane  $f(\mathbf{x}) = 0$  nor the classification decision. Hence choose  $\kappa$  to be the value such that the numerator of the margin is normalised to one, i.e.

$$\min_i \left\{ y_i (\kappa \mathbf{w}^\top \mathbf{x}_i + \kappa b) \right\} = 1 \quad (14.2.110)$$

Denote  $\tilde{\mathbf{w}} = \kappa \mathbf{w}$  and  $\tilde{b} = \kappa b$ , then we have that

$$y_i (\tilde{\mathbf{w}}^\top \mathbf{x}_i + \tilde{b}) \geq 1 \quad (14.2.111)$$

for all  $i = 1, \dots, n$ , where the points for which this holds with equality are called the ‘support vectors’. The optimisation problem now becomes

$$\begin{aligned} \max_{\tilde{\mathbf{w}}, \tilde{b}} \quad & \frac{1}{\|\tilde{\mathbf{w}}\|} \\ \text{s.t.} \quad & y_i (\tilde{\mathbf{w}}^\top \mathbf{x}_i + \tilde{b}) \geq 1, \quad i = 1, \dots, n \end{aligned} \quad (14.2.112)$$

which is equivalent to solving the quadratic program

$$\begin{aligned} \min_{\tilde{\mathbf{w}}, \tilde{b}} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 \\ \text{s.t.} \quad & y_i (\tilde{\mathbf{w}}^\top \mathbf{x}_i + \tilde{b}) \geq 1, \quad i = 1, \dots, n \end{aligned} \quad (14.2.113)$$

An alternative specification of this problem is

$$\min_{\tilde{\mathbf{w}}, \tilde{b}} \left\{ \sum_{i=1}^n \zeta \left( y_i (\tilde{\mathbf{w}}^\top \mathbf{x}_i + \tilde{b}) - 1 \right) + \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 \right\} \quad (14.2.114)$$

where  $\zeta(\cdot)$  plays the role of a barrier function such that  $\zeta(z) = 0$  if  $z \geq 0$  and  $\zeta(z) = \infty$  otherwise. In this way, SVM methods can be compared to other learning methods in terms of an error function and a regularisation term.

### Support Vector Machines for Non-Linearly Separable Data [25, 143]

In the case where the data is non-linearly separable, we can take the approach of softening the barrier function in the formulation above. Introduce a vector of slack variables  $\xi = (\xi_1, \dots, \xi_n)$  where each  $\xi_i \geq 0$ , such that  $\xi_i = 0$  if example  $i$  is correctly classified and  $y_i f(\mathbf{x}_i) \geq 1$ , and  $\xi_i = |y_i - f(\mathbf{x}_i)|$  otherwise, where  $f(\mathbf{x}) = \tilde{\mathbf{w}}^\top \mathbf{x} + \tilde{b}$ . We see that:

- $\xi_i = 0$  if  $\mathbf{x}_i$  is on or beyond the margin and correctly classified.
- $0 < \xi_i < 1$  if  $\mathbf{x}_i$  is inside the margin and correctly classified (when for example  $y_i = 1$  and  $0 < f(\mathbf{x}_i) < 1$ ).
- $\xi_i = 1$  if  $\mathbf{x}_i$  is on the decision boundary (since  $f(\mathbf{x}_i) = 0$ ).
- $\xi_i > 0$  if  $\mathbf{x}_i$  is misclassified (when for example  $y_i = 1$  and  $f(\mathbf{x}_i) < 0$ ).

The above requirements can be fully specified by writing the inequalities

$$y_i \tilde{f}(\mathbf{x}_i) \geq 1 - \xi_i \quad (14.2.115)$$

$$\xi_i \geq 0 \quad (14.2.116)$$

for each  $i = 1, \dots, n$ . This leads to the quadratic optimisation problem

$$\begin{aligned} \min_{\tilde{\mathbf{w}}, \tilde{b}, \xi} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0, \quad i = 1, \dots, n \\ & y_i \tilde{f}(\mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1, \dots, n \end{aligned} \quad (14.2.117)$$

where the term  $C$  is a parameter which can be used as the (inverse of) a regularisation coefficient. Increasing  $C$  will result in fewer training errors, but will result in a larger  $\|\tilde{\mathbf{w}}\|$ .

### Kernel Support Vector Machines [25]

Support vector machines can be extended to use a basis transformation on the features  $\phi(\mathbf{x})$ . The discriminant function is now

$$f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b \quad (14.2.118)$$

and the quadratic program to solve in the overlapping (i.e. non-separable) case is

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0, \quad i = 1, \dots, n \\ & y_i (\mathbf{w}^\top \phi(\mathbf{x}_i) + b) - 1 + \xi_i \geq 0, \quad i = 1, \dots, n \end{aligned} \quad (14.2.119)$$

We derive the dual of this problem. The Lagrangian is

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \mu_i \xi_i - \sum_{i=1}^n \lambda_i [y_i (\mathbf{w}^\top \phi(\mathbf{x}_i) + b) - 1 + \xi_i] \quad (14.2.120)$$

with Lagrange multiplier vectors  $\boldsymbol{\lambda} \geq 0$  and  $\boldsymbol{\mu} \geq 0$ . Differentiating the Lagrangian yields

$$\nabla_{\mathbf{w}} \mathcal{L} = \mathbf{w} - \sum_{i=1}^n \lambda_i y_i \phi(\mathbf{x}_i) \quad (14.2.121)$$

$$\frac{\partial \mathcal{L}}{\partial b} = - \sum_{i=1}^n \lambda_i y_i \quad (14.2.122)$$

$$\nabla_{\boldsymbol{\xi}} \mathcal{L} = C \mathbf{1} - \boldsymbol{\lambda} - \boldsymbol{\mu} \quad (14.2.123)$$

Setting these to zero gives the conditions

$$\mathbf{w} = \sum_{i=1}^n \lambda_i y_i \phi(\mathbf{x}_i) \quad (14.2.124)$$

$$\sum_{i=1}^n \lambda_i y_i = 0 \quad (14.2.125)$$

and

$$\lambda_i = C - \mu_i \quad (14.2.126)$$

for each  $i = 1, \dots, n$ . Applying the definition of the Lagrangian dual function produces

$$\mathcal{L}^*(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\mathbf{w}, b, \boldsymbol{\xi}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad (14.2.127)$$

$$= \inf_{\mathbf{w}, b, \boldsymbol{\xi}} \left\{ \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \mathbf{1}^\top \boldsymbol{\xi} - \sum_{i=1}^n \lambda_i y_i \mathbf{w}^\top \phi(\mathbf{x}_i) - b \sum_{i=1}^n \lambda_i y_i + \mathbf{1}^\top \boldsymbol{\lambda} + \boldsymbol{\lambda}^\top \boldsymbol{\xi} - \boldsymbol{\mu}^\top \boldsymbol{\xi} \right\} \quad (14.2.128)$$

Substituting the condition for  $\nabla_{\mathbf{w}} \mathcal{L} = 0$ , we get

$$\begin{aligned} \frac{1}{2} \mathbf{w}^\top \mathbf{w} - \sum_{i=1}^n \lambda_i y_i \mathbf{w}^\top \phi(\mathbf{x}_i) &= \frac{1}{2} \left( \sum_{i=1}^n \lambda_i y_i \phi(\mathbf{x}_i) \right) \left( \sum_{i=1}^n \lambda_i y_i \phi(\mathbf{x}_i)^\top \right) - \sum_{i=1}^n \lambda_i y_i \left( \sum_{j=1}^n \lambda_j y_j \phi(\mathbf{x}_j)^\top \right) \phi(\mathbf{x}_i) \\ &\quad (14.2.129) \end{aligned}$$

$$= -\frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \lambda_j \lambda_i y_j y_i \phi(\mathbf{x}_j)^\top \phi(\mathbf{x}_i) \quad (14.2.130)$$

Substituting the condition for  $\frac{\partial \mathcal{L}}{\partial b} = 0$ , we get

$$b \sum_{i=1}^n \lambda_i y_i = 0 \quad (14.2.131)$$

And substituting the condition for  $\nabla_{\xi} \mathcal{L}$ , we have

$$C\mathbf{1}^\top - \boldsymbol{\lambda}^\top \boldsymbol{\xi} - \boldsymbol{\mu}^\top \boldsymbol{\xi} = (C\mathbf{1} - \boldsymbol{\lambda} - \boldsymbol{\mu})^\top \boldsymbol{\xi} \quad (14.2.132)$$

$$= 0 \quad (14.2.133)$$

Hence the Lagrangian dual function simplifies so

$$\mathcal{L}^*(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{1}^\top \boldsymbol{\lambda} - \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \lambda_j \lambda_i y_j y_i \phi(\mathbf{x}_j)^\top \phi(\mathbf{x}_i) \quad (14.2.134)$$

$$= \mathbf{1}^\top \boldsymbol{\lambda} - \frac{1}{2} \boldsymbol{\lambda}^\top \begin{bmatrix} y_1^2 \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_1) & \dots & y_1 y_n \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ y_n y_1 \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_1) & \dots & y_n^2 \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_n) \end{bmatrix} \boldsymbol{\lambda} \quad (14.2.135)$$

Using the kernel trick, we can introduce the kernel  $k(\mathbf{x}_j, \mathbf{x}_i)$  to replace  $\phi(\mathbf{x}_j)^\top \phi(\mathbf{x}_i)$ , giving

$$\mathcal{L}^*(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{1}^\top \boldsymbol{\lambda} - \frac{1}{2} \underbrace{\boldsymbol{\lambda}^\top \begin{bmatrix} y_1^2 k(\mathbf{x}_1, \mathbf{x}_1) & \dots & y_1 y_n k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ y_n y_1 k(\mathbf{x}_n, \mathbf{x}_1) & \dots & y_n^2 k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}}_{\mathbf{K}} \boldsymbol{\lambda} \quad (14.2.136)$$

where now the basis transform need not be finite-dimensional. The constraints for the dual problem are  $\lambda_i \geq 0$  for the Lagrange multipliers, and  $\sum_{i=1}^n \lambda_i y_i = 0$  from the condition on  $\frac{\partial \mathcal{L}}{\partial b} = 0$ . Also note that since  $\mu_i \geq 0$ , then from the condition  $\lambda_i = C - \mu_i$  this implies that  $\lambda_i \leq C$ . We can write the dual problem as

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & -\frac{1}{2} \boldsymbol{\lambda}^\top \mathbf{K} \boldsymbol{\lambda} + \mathbf{1}^\top \boldsymbol{\lambda} \\ \text{s.t.} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \dots, n \\ & \sum_{i=1}^n \lambda_i y_i = 0, \end{aligned} \quad (14.2.137)$$

As the original problem is convex, we can solve the original problem by solving the dual problem, which is another quadratic program. After solving and obtaining the maximiser  $\boldsymbol{\lambda}^*$ , we can then find  $\mathbf{w}^*$  by the condition on  $\nabla_{\mathbf{w}} \mathcal{L} = 0$ :

$$\mathbf{w}^* = \sum_{i=1}^n \lambda_i^* y_i \phi(\mathbf{x}_i) \quad (14.2.138)$$

To find  $b^*$ , observe the following. If  $\lambda_i^* > 0$ , this implies that  $y_i (\phi(\mathbf{x}_i)^\top \mathbf{w}^* + b^*) - 1 + \xi_i^* = 0$ , because  $\lambda_i$  is the Lagrange multiplier for the constraint  $y_i (\mathbf{w}^\top \phi(\mathbf{x}) + b) - 1 + \xi_i \geq 0$  (i.e. either the constraint is active or the Lagrange multiplier is zero). Also if  $\lambda_i^* < C$ , then  $\mu_i^* > 0$  from the condition  $\lambda_i^* = C - \mu_i^*$ , which implies  $\xi_i^*$  (since  $\mu_i$  is the Lagrange multiplier for the constraint  $\xi_i \geq 0$ ). Thus,  $0 < \lambda_i^* < C$  together implies

$$y_i (\phi(\mathbf{x}_i)^\top \mathbf{w}^* + b^*) - 1 = 0 \quad (14.2.139)$$

$$y_i \left( \sum_{j=1}^n \lambda_j^* y_j \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) + b^* \right) = 1 \quad (14.2.140)$$

$$y_i \left( \sum_{j=1}^n \lambda_j^* y_j k(\mathbf{x}_i, \mathbf{x}_j) + b^* \right) = 1 \quad (14.2.141)$$

We can solve this with any  $i$  for which  $0 < \lambda_i^* < C$  to obtain  $b^*$ . Alternatively, we could average the solutions over all  $i = 1, \dots, n$  where  $0 < \lambda_i^* < C$  in order for numerical stability. Once we have found  $\lambda^*$  and  $b^*$ , we can use the SVM to make predictions at a given test point  $\mathbf{x}_*$  by

$$\hat{y} = \text{sgn}(f^*(\mathbf{x}_*)) \quad (14.2.142)$$

$$= \text{sgn}(\phi(\mathbf{x}_*)^\top \mathbf{w}^* + b^*) \quad (14.2.143)$$

$$= \text{sgn} \left( \sum_{i=1}^n \lambda_i^* y_i \phi(\mathbf{x}_*)^\top \phi(\mathbf{x}_i) + b^* \right) \quad (14.2.144)$$

$$= \text{sgn} \left( \sum_{i=1}^n \lambda_i^* y_i k(\mathbf{x}_*, \mathbf{x}_i) + b^* \right) \quad (14.2.145)$$

where  $k(\cdot, \cdot)$  is any positive definite kernel.

### Probability Class Estimation with Support Vector Machines [25, 143]

SVM predictions produce a hard labelling  $\hat{y} \in \{-1, 1\}$ , but can be augmented to produce class probabilities  $\Pr(y = 1|\mathbf{x})$ . This is done by assuming that the linear discriminant function  $\tilde{f}(\mathbf{x})$  can be interpreted as a log-odds ratio:

$$\tilde{f}(\mathbf{x}) = \log \left( \frac{\Pr(y = 1|\mathbf{x})}{\Pr(y = -1|\mathbf{x})} \right) \quad (14.2.146)$$

$$= \log \left( \frac{\Pr(y = 1|\mathbf{x})}{1 - \Pr(y = 1|\mathbf{x})} \right) \quad (14.2.147)$$

Recalling that the logistic function  $\sigma(\cdot)$  is the inverse of the log-odds ratio (logit):

$$\Pr(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\tilde{f}(\mathbf{x}))} \quad (14.2.148)$$

$$= \sigma(\tilde{f}(\mathbf{x})) \quad (14.2.149)$$

Or more generally,

$$\theta_1 \tilde{f}(\mathbf{x}) + \theta_2 = \log \left( \frac{\Pr(y = 1|\mathbf{x}, \theta_1, \theta_2)}{\Pr(y = -1|\mathbf{x}, \theta_1, \theta_2)} \right) \quad (14.2.150)$$

which leads to

$$\Pr(y = 1|\mathbf{x}, \theta_1, \theta_2) = \sigma(\theta_1 \tilde{f}(\mathbf{x}) + \theta_2) \quad (14.2.151)$$

where hyperparameters  $\theta_1$  and  $\theta_2$  can be optimised on some validation dataset.

## 14.2.8 Multiclass Classification

### One-vs-Rest Classification

The one-vs-rest technique allows a binary classifier to be transformed into a multiclass classifier for output classes  $\{1, \dots, K\}$ . For this to work, the binary classifier needs to be able to compute some score  $f(x)$  for a given feature  $x$ , where a higher score indicates  $x$  is more likely to be in a positive class. For example,  $f(x)$  could be a thresholding function (like in support vector machines) or a probability a positive class. Then from a training dataset, we construct  $K$

binary classifiers  $f_1(x), \dots, f_K(x)$  where for the  $k^{\text{th}}$  classifier, all class  $k$  samples have been relabelled as the positive class, and all other samples have been relabelled as the negative class. Multiclass classification can then be performed by

$$\hat{y}(x) = \operatorname{argmax}_{k \in \{1, \dots, K\}} f_k(x) \quad (14.2.152)$$

i.e. the class which gives the strongest score when compared against the rest in a binary classifier. There may be an arbitrary method used to account for ties.

### One-vs-One Classification

The one-vs-one technique is another way (like one-vs-rest) to transform a binary classifier into a multiclass classifier for output classes  $\{1, \dots, K\}$ . Unlike one-vs-rest however, the binary classifier does not necessarily need to provide a score; it just needs to output the predicted binary class. As there are  $\binom{K}{2} = K(K - 1)/2$  pairs of classes, from the training dataset we train  $K(K - 1)$  binary classifiers denoted  $\hat{y}_{1,2}(x), \dots, \hat{y}_{K-1,K}(x)$  where the classifier  $\hat{y}_{k,\kappa}(x) \in \{k, \kappa\}$  with is trained using all samples in either class  $k$  or  $\kappa$ . For multiclass prediction for a feature  $x$ , we predict with all  $K(K - 1)$  binary classifiers and apply the following reasoning:

- If we are currently using  $\hat{y}_{k,\kappa}(x)$  and  $x$  is actually in class  $k^*$  which is neither class  $k$  nor  $\kappa$ , then there should be no discernible pattern in the classification. That is, for all  $k, \kappa \in \{1, \dots, K\} \setminus \{k^*\}$  with  $k \neq \kappa$ , there should be no pattern in the predictions of  $\hat{y}_{k,\kappa}(x)$ .
- If we are currently using  $\hat{y}_{k,\kappa}(x)$  and  $x$  is actually in class  $k$  or  $\kappa$  (say  $k$ ), then there should be a pattern where  $x$  is usually classified as  $k$ . That is, for all  $\kappa \in \{1, \dots, K\} \setminus \{k\}$ , there should be a pattern of  $\hat{y}_{k,\kappa}(x) = k$ .

From this, we can apply the classification rule

$$\hat{y}(x) = \operatorname{argmax}_{k \in \{1, \dots, K\}} \sum_{\kappa \in \{1, \dots, K\} \setminus \{k\}} \mathbb{I}_{\{\hat{y}_{k,\kappa}(x)=k\}} \quad (14.2.153)$$

That is, we put  $x$  in the class that gets predicted most often (with some arbitrary way to account for ties).

## 14.3 Unsupervised Learning

### 14.3.1 $k$ -means Clustering

The  $k$ -means clustering problem aims to partition the multivariate dataset  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  into  $k$  clusters (where  $k$  is a hyperparameter that can be chosen by the practitioner). These  $k$  clusters are represented by the disjoint sets  $\mathbf{C} = \{C_1, \dots, C_k\}$ . The objective is to partition the data so that the within-cluster sum of squared deviation (in Euclidean distance) from the cluster center is minimised:

$$\mathbf{C}^* = \operatorname{argmin}_{C_1, \dots, C_k} \left\{ \sum_{i=1}^k \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{m}_i\|^2 \right\} \quad (14.3.1)$$

where each cluster centre  $\mathbf{m}_i$  is the mean of points in the cluster:

$$\mathbf{m}_i = \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j \quad (14.3.2)$$

Because the squared Euclidean distance (inner-product) is equal to the trace of the outer-product), we can interpret this objective as minimising the trace of the sample covariance matrix within each cluster, averaged over all the cluster. So a low objective should represent in a sense the points within each cluster being ‘clustered together’. This objective also admits an equivalent alternative representation in terms of squared pairwise distances within each cluster:

$$\mathbf{C}^* = \operatorname{argmin}_{C_1, \dots, C_k} \left\{ \sum_{i=1}^k \frac{1}{|C_i|} \sum_{\mathbf{x}_j, \mathbf{x}'_j \in C_i} \|\mathbf{x}_j - \mathbf{x}'_j\|^2 \right\} \quad (14.3.3)$$

Another equivalent characterisation of  $k$ -means clustering involves maximising the variance between cluster centres. This is more immediate after writing down a sample-version of the Law of Total Variance:

$$\widehat{\operatorname{Var}}(\mathbf{X}) = \widehat{\mathbb{E}} \left[ \widehat{\operatorname{Var}}(\mathbf{X}|C) \right] + \widehat{\operatorname{Var}} \left( \widehat{\mathbb{E}}[\mathbf{X}|C] \right) \quad (14.3.4)$$

The first term in the right-hand side is analogous to the objective we are trying to minimise. Since the total variance in the dataset stays constant regardless of partitioning, minimising the first term is equivalent to maximising the second term in the right-hand side. The quantity  $\widehat{\mathbb{E}}[\mathbf{X}|C]$  is analogous to the cluster centre, so this second term is the variance of the cluster centres.

### Lloyd’s Algorithm

Lloyd’s algorithm is a method for minimising the  $k$ -means clustering objective. In each iteration of the algorithm, we begin with cluster centres  $\mathbf{m}_1, \dots, \mathbf{m}_k$ . Then we perform the following two steps:

1. We assign each point  $\mathbf{x}_j$  to its closest cluster centre and cluster them that way. That is, for each  $i = 1, \dots, k$ ,

$$C_i \leftarrow \left\{ \mathbf{x} : \|\mathbf{x} - \mathbf{m}_i\|^2 \leq \|\mathbf{x} - \mathbf{m}_j\|^2, j = 1, \dots, k \right\} \quad (14.3.5)$$

In other words, we perform a Voronoi partitioning of the data using the cluster centres. Since the partitions must be disjoint, we can ensure this by splitting any ties randomly.

2. Next, update the cluster centres using the new assignments. So for each  $i = 1, \dots, k$ ,

$$\mathbf{m}_i \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j \quad (14.3.6)$$

This procedure is then iterated until the assignments stop changing (i.e. the objective converges). As for how to choose the initial cluster centres, some options are:

- Set the cluster centres as  $k$  randomly chosen points.
- Randomly assign a cluster to each point, and compute the cluster centres as the means.

Lloyd’s algorithm guarantees convergence to a local optima, as the objective can never increase. To show this, we can show that each step in every iteration does not increase the objective.

1. In the first step, by virtue of the Voronoi partitioning, each point is assigned to its cluster which minimises its distance from the cluster centre. So for fixed  $\mathbf{m}_1, \dots, \mathbf{m}_k$ , the objective is guaranteed to not increase after this step.

2. In the second step, we can use the fact that the mean is the minimiser of squared deviations from the data. So for fixed  $C_1, \dots, C_k$ , then for each cluster  $C_i$  we have

$$\sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{m}_i\|^2 \leq \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{m}'_i\|^2 \quad (14.3.7)$$

where  $\mathbf{m}'_i$  is anything other than the cluster mean.

### *k*-medians Clustering

The *k*-medians clustering approach is similar in idea to *k*-means clustering, except we use the  $\ell_1$  distance in the objective:

$$\mathbf{C}^* = \operatorname{argmin}_{C_1, \dots, C_k} \left\{ \sum_{i=1}^k \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \check{\mathbf{m}}_i\|_1^2 \right\} \quad (14.3.8)$$

where the cluster centre is now a multivariate generalisation of the median, being the spacial median with 1-norm:

$$\check{\mathbf{m}}_i = \operatorname{argmin}_{\mathbf{m}} \left\{ \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{m}\|_1 \right\} \quad (14.3.9)$$

### *k*-medoids Clustering

The *k*-medoids clustering approach is similar in idea to *k*-means and *k*-medians clustering, except we now allow for an arbitrary distance function  $d(\cdot, \cdot)$  in the objective:

$$\mathbf{C}^* = \operatorname{argmin}_{C_1, \dots, C_k} \left\{ \sum_{i=1}^k \sum_{\mathbf{x}_j \in C_i} d(\mathbf{x}_j, \check{\mathbf{m}}_i) \right\} \quad (14.3.10)$$

and use the mediod with distance function  $d(\cdot, \cdot)$  as the cluster centre:

$$\check{\mathbf{m}}_i = \operatorname{argmin}_{\mathbf{m} \in \{\mathbf{x}_1, \dots, \mathbf{x}_n\}} \left\{ \sum_{\mathbf{x}_j \in C_i} d(\mathbf{x}_j, \mathbf{m}) \right\} \quad (14.3.11)$$

This additionally means that each cluster centre will be an actual point in the dataset.

## 14.3.2 Mode-Seeking Algorithms

Mode-seeking algorithms attempt to find the modes (i.e. local maxima) of a density function  $f(\mathbf{x})$ . The density function  $f(\mathbf{x})$  itself can be obtained as an estimate from some data points (e.g. via kernel density estimation). The modes found can then be deemed as the clusters of the distribution.

### Mean-Shift Algorithm

#### Simulated Annealing [115]

Simulated annealing algorithms can be used to locate the maximum (i.e. find the mode) of a target density  $p(x)$ . Somewhat more generally, consider the maximisation problem  $\max_{x \in \mathcal{X}} f(x)$ . Define the Boltzmann PDF by

$$p(x) \propto e^{f(x)} \quad (14.3.12)$$

where we have implicitly assumed that  $e^{f(x)}$  is normalisable. Also assume that  $f(x)$  has a unique maximum, then so does  $p(x)$ , which is thus unimodal. Then the maximiser of  $f(x)$  will be the maximiser of  $p(x)$ , i.e.

$$\operatorname{argmax}_{x \in \mathcal{X}} f(x) = \operatorname{argmax}_{x \in \mathcal{X}} p(x) \quad (14.3.13)$$

Introduce a ‘temperature’ variable  $T$ , and consider the unnormalised PDF

$$p(x)^{1/T} \propto e^{f(x)/T} \quad (14.3.14)$$

For small  $T > 0$ , the effect of the exponent  $1/T$  is to make the shape of  $p(x)^{1/T}$  much more leptokurtic compared to that of  $p(x)$ . Thus if we were to draw a sample from the distribution proportional to  $p(x)^{1/T}$ , then it is ‘very likely’ that the sample will be ‘close’ to the mode, and thus the maximiser of  $f(x)$ . A simulated annealing algorithm then applies a variant of Markov chain Monte-Carlo sampling to approximately draw a sample from target  $p(x)^{1/T}$ , with  $T$  small.

As an example, let the temperature *cooling schedule* be given by the sequence  $T_t > 0$  such that  $T_{t+1} < T_t$ . Suppose we use the Metropolis algorithm, which uses a symmetric proposal distribution so that the acceptance probability only involves the ratio of target densities. Then beginning from some initial temperature  $T_0$  and initial point  $x_0$ , each iteration of the simulated annealing algorithm is given by:

- Sample  $y$  from the symmetric proposal density  $q(y|x_t)$ , where we can think of  $y$  as being some ‘local’ candidate nearby  $x_t$ .
- Let  $x_{t+1} = y$  with acceptance probability

$$\alpha(x_t, y) = \min \left\{ \frac{p(y)^{1/T_t}}{p(x_t)^{1/T_t}}, 1 \right\} \quad (14.3.15)$$

$$= \min \left\{ \frac{\exp(f(y)/T_t)}{\exp(f(x_t)/T_t)}, 1 \right\} \quad (14.3.16)$$

$$= \min \left\{ \frac{\exp(f(y)/T_t)}{\exp(f(x_t)/T_t)}, 1 \right\} \quad (14.3.17)$$

$$= \min \left\{ \exp \left( \frac{f(y) - f(x_t)}{T_t} \right), 1 \right\} \quad (14.3.18)$$

Thus, the way this algorithm works is that if a proposal is found such that  $f(y) \geq f(x_t)$ , it will always accept the proposal. However if  $f(y) < f(x_t)$ , then in order to allow the algorithm to escape local optima, there is a chance that the proposal will be accepted, but with probability becoming increasingly small as  $T_t$  decreases.

### 14.3.3 Gaussian Mixture Models [25]

In a Gaussian mixture model, the data is assumed to be generated from a mixture of multivariate Gaussians, with density

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \quad (14.3.19)$$

with  $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$ ,  $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_K\}$  and  $\boldsymbol{\Sigma} = \{\Sigma_1, \dots, \Sigma_K\}$ . Each  $\mu_k$  and  $\Sigma_k$  denotes the mean vector and covariance matrix respectively of the  $k^{\text{th}}$  component, where there are  $K$  components in total and each component constitutes a multivariate Gaussian, with density

denoted  $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ . The  $\{\pi_1, \dots, \pi_K\}$  are known as the mixing weights/coefficients, such that each  $\pi_k \geq 0$  and

$$\sum_{k=1}^K \pi_k = 1 \quad (14.3.20)$$

The centre of each component is meant to represent a cluster of data. To learn the model parameters from  $n$  i.i.d. observations of data denoted  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ , we obtain the log-likelihood as

$$\log \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}) = \log \left( \prod_{i=1}^n p(\mathbf{x}_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right) \quad (14.3.21)$$

$$= \sum_{i=1}^n \log p(\mathbf{x}_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (14.3.22)$$

$$= \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k) \right) \quad (14.3.23)$$

So a maximum likelihood solution is given by

$$\begin{aligned} (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) &= \underset{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}}{\operatorname{argmin}} \{-\log \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X})\} \\ \text{s.t. } &\sum_{k=1}^K \pi_k = 1 \\ &\pi_k \geq 0, \quad k = 1, \dots, K \\ &\Sigma_k \succeq \mathbf{0}, \quad k = 1, \dots, K \end{aligned} \quad (14.3.24)$$

This problem is much harder to solve than for single-component Gaussian maximum likelihood because there is no analytical solution. However, we can use iterative procedures for maximum likelihood estimation such as the Expectation Maximisation algorithm to obtain an estimate. Furthermore, we can couple this with a higher-level method to choose the number of components  $K$ , such as by information criteria.

#### 14.3.4 Dirichlet Process Mixtures

### 14.4 Artificial Neural Networks

#### 14.4.1 Multi-Layer Perceptrons

For an input vector  $x \in \mathbb{R}^{N_x}$  with  $x = [x_1 \ \dots \ x_{N_x}]^\top$ , we have an associated (target/observed) output vector  $y \in \mathbb{R}^{N_y}$  with  $y = [y_1 \ \dots \ y_{N_y}]^\top$ . Let there be  $L + 1$  layers in the network, with structure given by

$$\{N_x, N_1, N_2, \dots, N_\ell, \dots, N_L\} \in \mathbb{N}^{L+1} \quad (14.4.1)$$

which denotes the number of nodes (size) of each layer. Layers  $1, \dots, L - 1$  are known as the hidden layer. A multi-layer perception network is often called a *deep neural network*. Note that  $N_L \equiv N_y$  in the final layer. For each layer excluding the input layer, there is an ‘intermediate output’  ${}^\ell z \in \mathbb{R}^{N_\ell}$ , an activation  ${}^\ell a \in \mathbb{R}^{N_\ell}$ , a bias  ${}^\ell b \in \mathbb{R}^{N_\ell}$ , some weights  ${}^\ell w \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$  and an activation function  ${}^\ell \sigma(\cdot)$ . Their relationship in each layer from the one before it is given by the affine transformation

$${}^\ell z = {}^{\ell-1} w {}^{\ell-1} a + {}^\ell b \quad (14.4.2)$$

and  ${}^0a$  can be taken to equal  $x$ . Denote

$${}^\ell\boldsymbol{\sigma}\left({}^\ell z\right) = \begin{bmatrix} {}^\ell\sigma({}^\ell z_1) \\ \vdots \\ {}^\ell\sigma({}^\ell z_N) \end{bmatrix} \quad (14.4.3)$$

so that

$${}^\ell a = {}^\ell\boldsymbol{\sigma}\left({}^\ell z\right) \quad (14.4.4)$$

The feedforward function is

$$f(x) = {}^L a = {}^L\boldsymbol{\sigma}\left({}^L w^{L-1} a + {}^L b\right) \quad (14.4.5)$$

Suppose we have a training data set indexed by

$$X = \{x[1], \dots, x[n]\} \quad (14.4.6)$$

$$Y = \{y[1], \dots, y[n]\} \quad (14.4.7)$$

Use the cost function

$$C = \frac{1}{2n} \sum_{i=1}^n \|f(x[i]) - y[i]\|^2 \quad (14.4.8)$$

This can be broken up into a cost for each training point:

$$C_i = \frac{1}{2} \|f(x[i]) - y[i]\|^2 \quad (14.4.9)$$

$$= \frac{1}{2} (f_1(x[i]) - y_1[i])^2 + \dots + \frac{1}{2} (f_{N_y}(x[i]) - y_{N_y}[i])^2 \quad (14.4.10)$$

so that

$$C = \frac{1}{n} \sum_{i=1}^n C_i \quad (14.4.11)$$

Using this cost function, the gradient descent updates for the weights and biases in each layer are

$${}^{\ell-1}w \leftarrow {}^{\ell-1}w - \eta \frac{\partial C}{\partial {}^{\ell-1}w}^\top \quad (14.4.12)$$

$${}^\ell b \leftarrow {}^\ell b - \eta \frac{\partial C}{\partial {}^\ell b}^\top \quad (14.4.13)$$

where  $\eta$  is the learning rate. For ease of notation we will use for a given  $i$ :  $c := C_i$ ,  $f := f(x)[i]$  and  $y := y[i]$ . For example, instead of using the cumbersome notation for the gradient vector

$$\nabla_f C_i = \frac{\partial C_i}{\partial f}^\top = \begin{bmatrix} \frac{\partial C_i}{\partial f_1} & \dots & \frac{\partial C_i}{\partial f_{N_y}} \end{bmatrix}^\top = \begin{bmatrix} f_1(x[i]) - y_1[i] \\ \vdots \\ f_{N_y}(x[i]) - y_{N_y}[i] \end{bmatrix} \quad (14.4.14)$$

we can instead write

$$\nabla_f c = \begin{bmatrix} f_1 - y_1 \\ \vdots \\ f_{N_y} - y_{N_y} \end{bmatrix} \quad (14.4.15)$$

## Backpropagation

To develop a method for computing  $\frac{\partial C}{\partial_{\ell-1} w}$  and  $\frac{\partial C}{\partial_{\ell} b}$ , first consider  $\nabla_{Lz} c$ . We have from the chain rule

$$\nabla_{Lz} c = \begin{bmatrix} \frac{\partial c}{\partial^L z_1} \\ \vdots \\ \frac{\partial c}{\partial^L z_{N_L}} \end{bmatrix} = \begin{bmatrix} \frac{\partial c}{\partial f_1} \times \frac{\partial f_1}{\partial^L z_1} \\ \vdots \\ \frac{\partial c}{\partial f_{N_L}} \times \frac{\partial f_{N_L}}{\partial^L z_{N_L}} \end{bmatrix} = \begin{bmatrix} \frac{\partial c}{\partial f_1} \cdot {}^L \sigma'({}^L z) \\ \vdots \\ \frac{\partial c}{\partial f_{N_L}} \cdot {}^L \sigma'({}^L z) \end{bmatrix} = \nabla_f c \odot {}^L \sigma'({}^L z) \quad (14.4.16)$$

where  $\odot$  is the Hadamard (element-wise) product. Denote the ‘error’ in each layer as  ${}^\ell \delta := \nabla_{\ell z} c$ . For the final layer, we have

$${}^L z = {}^{L-1} w^{L-1} a + {}^L b \quad (14.4.17)$$

It can be easily seen that  $\frac{\partial^L z}{\partial^L b} = I$ , so this gives

$$\frac{\partial C^\top}{\partial^L b} = \nabla_{Lz} c = {}^L \delta \quad (14.4.18)$$

and in general for any layer

$$\frac{\partial C^\top}{\partial^{\ell} b} = \nabla_{\ell z} c = {}^{\ell} \delta \quad (14.4.19)$$

For the following steps use for ease of notation:  $z := {}^\ell z$ ,  $b := {}^\ell b$ ,  $a := {}^{\ell-1} a$  and  $w := {}^{\ell-1} w$ . Then write out the following sums:

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} w_{11}a_1 + w_{12}a_2 + \cdots + w_{1N}a_N + b_1 \\ w_{21}a_1 + w_{22}a_2 + \cdots + w_{2N}a_N + b_2 \\ \vdots \\ w_{N1}a_1 + w_{N2}a_2 + \cdots + w_{NN}a_N + b_N \end{bmatrix} \quad (14.4.20)$$

This is helpful for seeing that if we compute the derivative of scalar by matrix  $\frac{\partial z_1}{\partial w}$ , this gives

$$\frac{\partial z_1}{\partial w} = \begin{bmatrix} \frac{\partial z_1}{\partial w_{11}} & \frac{\partial z_1}{\partial w_{21}} & \cdots \\ \frac{\partial z_1}{\partial w_{12}} & \ddots & \\ \vdots & & \ddots \end{bmatrix} = \begin{bmatrix} a_1 & 0 & \cdots \\ a_2 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad (14.4.21)$$

The derivative of the cost with respect to the weights are

$$\frac{\partial c}{\partial w} = \begin{bmatrix} \frac{\partial c}{\partial w_{11}} & \frac{\partial c}{\partial w_{21}} & \cdots \\ \frac{\partial c}{\partial w_{12}} & \ddots & \\ \vdots & & \ddots \end{bmatrix} \quad (14.4.22)$$

Consider the first element, this will consist of all the contributions from elements of  $z$  as so

$$\frac{\partial c}{\partial w_{11}} = \frac{\partial c}{\partial z_1} \frac{\partial z_1}{\partial w_{11}} + \frac{\partial c}{\partial z_2} \frac{\partial z_2}{\partial w_{11}} + \underbrace{[\dots]}_0 = \frac{\partial c}{\partial z_1} a_1 \quad (14.4.23)$$

The element  $\frac{\partial c}{\partial w_{12}}$  is similarly given by

$$\frac{\partial c}{\partial w_{12}} = \frac{\partial c}{\partial z_1} \frac{\partial z_1}{\partial w_{12}} + \frac{\partial c}{\partial z_2} \frac{\partial z_2}{\partial w_{12}} + \dots \stackrel{0}{=} \frac{\partial c}{\partial z_1} a_2 \quad (14.4.24)$$

Thus

$$\frac{\partial c}{\partial w}^\top = \begin{bmatrix} \frac{\partial c}{\partial w_{11}} & \frac{\partial c}{\partial w_{12}} & \dots \\ \frac{\partial c}{\partial w_{21}} & \ddots & \\ \vdots & & \end{bmatrix} = \begin{bmatrix} \frac{\partial c}{\partial z_1} a_1 & \frac{\partial c}{\partial z_1} a_2 & \dots \\ \frac{\partial c}{\partial z_2} a_1 & \ddots & \\ \vdots & & \end{bmatrix} = \begin{bmatrix} \frac{\partial c}{\partial z_1} \\ \frac{\partial c}{\partial z_2} \\ \vdots \end{bmatrix} [a_1 \ a_2 \ \dots] \quad (14.4.25)$$

Reverting back to notation which specifies the layers, we have

$$\frac{\partial c}{\partial^{\ell-1} w}^\top = {}^\ell \delta \cdot {}^{\ell-1} a^\top \quad (14.4.26)$$

Next, consider the relationship between  ${}^{\ell-1} \delta$  and  ${}^\ell \delta$ . Computing  ${}^{\ell-1} \delta$  from  ${}^\ell \delta$  is known as backpropagation. First write

$${}^\ell z = {}^{\ell-1} w^{\ell-1} \sigma({}^{\ell-1} z) + {}^\ell b \quad (14.4.27)$$

For ease of notation, use  $\sigma(\cdot) := {}^{\ell-1} \sigma(\cdot)$ . Then writing out the sums arising from the matrix multiplication gives

$$\begin{bmatrix} {}^\ell z_1 \\ {}^\ell z_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} {}^{\ell-1} w_{11} \sigma({}^{\ell-1} z_1) + {}^{\ell-1} w_{12} \sigma({}^{\ell-1} z_2) + \dots + {}^\ell b_1 \\ {}^{\ell-1} w_{21} \sigma({}^{\ell-1} z_1) + {}^{\ell-1} w_{22} \sigma({}^{\ell-1} z_2) + \dots + {}^\ell b_2 \\ \vdots \end{bmatrix} \quad (14.4.28)$$

We want to find

$${}^{\ell-1} \delta = \begin{bmatrix} \frac{dc}{d^{\ell-1} z_1} \\ \frac{dc}{d^{\ell-1} z_2} \\ \vdots \end{bmatrix} \quad (14.4.29)$$

Consider the first element  $\frac{dc}{d^{\ell-1} z_1}$  as the total derivative with contribution from all  ${}^\ell z_1$ ,  ${}^\ell z_2$ , etc.

$$\frac{dc}{d^{\ell-1} z_1} = \frac{\partial c}{\partial^{\ell-1} z_1} \cdot \frac{\partial^{\ell} z_1}{\partial^{\ell-1} z_1} + \frac{\partial c}{\partial^{\ell-1} z_2} \cdot \frac{\partial^{\ell} z_2}{\partial^{\ell-1} z_1} + \dots \quad (14.4.30)$$

Calculating some of these terms using the chain rule gives

$$\frac{\partial^{\ell} z_1}{\partial^{\ell-1} z_1} = {}^{\ell-1} w_{11} \sigma'({}^{\ell-1} z_1) \quad (14.4.31)$$

$$\frac{\partial^{\ell} z_2}{\partial^{\ell-1} z_1} = {}^{\ell-1} w_{21} \sigma'({}^{\ell-1} z_1) \quad (14.4.32)$$

$$\vdots \quad (14.4.33)$$

So then

$$\frac{dc}{d^{\ell-1} z_1} = \left( \frac{\partial c}{\partial^{\ell-1} z_1} \cdot {}^{\ell-1} w_{11} + \frac{\partial c}{\partial^{\ell-1} z_2} \cdot {}^{\ell-1} w_{21} + \dots \right) \sigma'({}^{\ell-1} z_1) \quad (14.4.34)$$

and more generally,

$$\begin{bmatrix} \frac{dc}{d^{\ell-1}z_1} \\ \frac{dc}{d^{\ell-1}z_2} \\ \vdots \end{bmatrix} = \begin{bmatrix} \left( \frac{\partial c}{\partial^{\ell}z_1} \cdot {}^{\ell-1}w_{11} + \frac{\partial c}{\partial^{\ell}z_2} \cdot {}^{\ell-1}w_{21} + \dots \right) \sigma'({}^{\ell-1}z_1) \\ \left( \frac{\partial c}{\partial^{\ell}z_1} \cdot {}^{\ell-1}w_{12} + \frac{\partial c}{\partial^{\ell}z_2} \cdot {}^{\ell-1}w_{22} + \dots \right) \sigma'({}^{\ell-1}z_2) \\ \vdots \end{bmatrix} \quad (14.4.35)$$

Notice that we may then write this as

$${}^{\ell-1}\delta = \left( {}^{\ell-1}w^\top \cdot {}^{\ell}\delta \right) \odot {}^{\ell-1}\sigma'({}^{\ell-1}z) \quad (14.4.36)$$

Now reverting back to notation that is indexed by each training point  $i$ :  ${}^{\ell}\delta[i] = \frac{\partial C_i}{\partial^{\ell}z}$ . Since we have found the derivative of each individual cost with respect to the weights, the derivative of the total cost with respect to the weights is

$$\frac{\partial C}{\partial^{\ell-1}w}^\top = \frac{1}{n} \sum_{i=1}^n \frac{\partial C_i}{\partial^{\ell-1}w}^\top = \frac{1}{n} \sum_{i=1}^n {}^{\ell}\delta[i] \cdot {}^{\ell-1}a[i]^\top \quad (14.4.37)$$

$$\frac{\partial C}{\partial^{\ell}b}^\top = \frac{1}{n} \sum_{i=1}^n \frac{\partial C_i}{\partial^{\ell}b}^\top = \frac{1}{n} \sum_{i=1}^n {}^{\ell}\delta[i] \quad (14.4.38)$$

So the gradient descent equations using the entire data set are

$${}^{\ell-1}w \leftarrow {}^{\ell-1}w - \eta \cdot \frac{1}{n} \sum_{i=1}^n {}^{\ell}\delta[i] \cdot {}^{\ell-1}a[i]^\top \quad (14.4.39)$$

$${}^{\ell}b \leftarrow {}^{\ell}b - \eta \cdot \frac{1}{n} \sum_{i=1}^n {}^{\ell}\delta[i] \quad (14.4.40)$$

### Mini-Batch Gradient Descent

If the number of training points is very large, then gradient descent can take a long time, so learning occurs slowly. An idea is to use *stochastic gradient descent*, where a random mini-batch of  $m$  inputs is chosen to approximate the gradient of  $C$  and update the weights according to

$${}^{\ell-1}w \leftarrow {}^{\ell-1}w - \eta \cdot \frac{1}{m} \sum_{i=1}^n {}^{\ell}\delta[i] \cdot {}^{\ell-1}a[i]^\top \quad (14.4.41)$$

$${}^{\ell}b \leftarrow {}^{\ell}b - \eta \cdot \frac{1}{m} \sum_{i=1}^n {}^{\ell}\delta[i] \quad (14.4.42)$$

Then another mini-batch is chosen to update the weights. This occurs until the entirety of the training set is exhausted, in which this is called one training epoch. We may then start over again with a new epoch to continue training.

### Cross Entropy Loss Function

In a classification model, we have data with training examples from  $1, \dots, n$ , and the outputs are the class probabilities for each of  $1, \dots, M$  different classes. Denote by  $y_{i,j}$  the  $i^{\text{th}}$  class probability of the  $j^{\text{th}}$  training example. Each training example target vector  $\mathbf{y}_j$  should really be (but in a more general setting is not restricted to) a ‘one-hot’ vector, e.g.  $(0, 1, 0)$ . Similarly

denote the estimates of class probabilities by the model as  $\hat{y}_{i,j}$ . A choice of loss function  $L$  is the cross entropy loss function, given by

$$L(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{i=1}^M y_i \log \hat{y}_i \quad (14.4.43)$$

which is just the cross entropy between the ‘true’ distribution  $\mathbf{y}$  and the approximating distribution  $\hat{\mathbf{y}}$ . Then because the difference between the cross entropy and the KL divergence is only an additive constant (which is the entropy of  $\mathbf{y}$ , which we treat as fixed in the data), then minimising the cross entropy is the same as minimising the KL divergence between the estimated distribution and the actual distribution. Over the entire dataset, by assuming independence we can just sum over the cross entropy for each training example, leading to the cost function

$$J = - \sum_{j=1}^n \sum_{i=1}^M y_{i,j} \log \hat{y}_{i,j} \quad (14.4.44)$$

Now suppose the estimated vector  $\hat{\mathbf{y}}$  is applied via a softmax of some values  $z_1, \dots, z_M$ , i.e.

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{l=1}^M \exp(z_l)} \quad (14.4.45)$$

Then to derive the derivative of the loss with respect with the values  $z_1, \dots, z_M$  (for use in backpropagation as an example), we get (for a particular  $z_k$ ):

$$\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial z_k} = \frac{\partial}{\partial z_k} \left( - \sum_{i=1}^M y_i \log \hat{y}_i \right) \quad (14.4.46)$$

$$= \frac{\partial}{\partial z_k} \left( - \sum_{i=1}^M y_i \log \left( \frac{\exp(z_i)}{\sum_{l=1}^M \exp(z_l)} \right) \right) \quad (14.4.47)$$

$$= \frac{\partial}{\partial z_k} \left( - \sum_{i=1}^M y_i z_i + \sum_{i=1}^M y_i \log \left( \sum_{l=1}^M \exp(z_l) \right) \right) \quad (14.4.48)$$

$$= -y_k + \sum_{i=1}^M y_i \frac{\exp(z_k)}{\sum_{l=1}^M \exp(z_l)} \quad (14.4.49)$$

$$= -y_k + \hat{y}_k \sum_{i=1}^M y_i \quad (14.4.50)$$

$$= -y_k + \hat{y}_k \quad (14.4.51)$$

Hence the gradient vector is

$$\nabla_{\mathbf{z}} L(\mathbf{y}, \hat{\mathbf{y}}) = -\mathbf{y} + \hat{\mathbf{y}} \quad (14.4.52)$$

and the gradient vector of the cost function is the sum over the examples

$$\nabla_{\mathbf{z}} J = \sum_{j=1}^n (\hat{\mathbf{y}}_j - \mathbf{y}_j) \quad (14.4.53)$$

We can arrive at the cross entropy cost function by considering a maximum likelihood approach. Suppose  $\bar{\mathbf{y}}_j$  gives the actual probability distribution for example  $j$ . The likelihood of the data  $\mathbf{y}_1, \dots, \mathbf{y}_n$  given  $\bar{\mathbf{y}}_j$  is denoted  $\mathcal{L}(\bar{\mathbf{y}}_1, \dots, \bar{\mathbf{y}}_n | \mathbf{y}_1, \dots, \mathbf{y}_n)$  and represented by

$$\mathcal{L}(\bar{\mathbf{y}}_1, \dots, \bar{\mathbf{y}}_n | \mathbf{y}_1, \dots, \mathbf{y}_n) = \Pr(\mathbf{y}_1, \dots, \mathbf{y}_n | \bar{\mathbf{y}}_1, \dots, \bar{\mathbf{y}}_n) \quad (14.4.54)$$

In the case  $\mathbf{y}_j$  are one-hot vectors and all independent, this becomes

$$\mathcal{L}(\bar{\mathbf{y}}_1, \dots, \bar{\mathbf{y}}_n | \mathbf{y}_1, \dots, \mathbf{y}_n) = \prod_{j=1}^n \bar{y}_{\{i:y_{i,j}=1\},j} \quad (14.4.55)$$

where  $\{i : y_{i,j} = 1\}$  express the index of the class for which was actually observed in the  $j^{\text{th}}$  training example. Hence  $\bar{y}_{\{i:y_{i,j}=1\},j}$  is just the probability of the class that was actually observed. Note that if  $\mathbf{y}_j$  were not one-hot vectors, this still generalises well to

$$\mathcal{L}(\bar{\mathbf{y}}_1, \dots, \bar{\mathbf{y}}_n | \mathbf{y}_1, \dots, \mathbf{y}_n) = \prod_{j=1}^n \prod_{i=1}^M \bar{y}_{i,j}^{y_{i,j}} \quad (14.4.56)$$

To illustrate, consider a  $n = 100$  trials in an i.i.d. multinomial experiment with  $M = 3$  for example. If we obtained 40 in class 1, 25 in class 2 and 35 in class 3, then the likelihood would be written as

$$\mathcal{L}(\bar{y}_1, \bar{y}_2, \bar{y}_3 | \mathbf{y}_1, \dots, \mathbf{y}_{100}) = \bar{y}_1^{40} \bar{y}_2^{25} \bar{y}_3^{35} \quad (14.4.57)$$

$$= (\bar{y}_1^{0.4} \bar{y}_2^{0.25} \bar{y}_3^{0.35})^{100} \quad (14.4.58)$$

Thus we can more generally view the likelihood as the product over all the training examples of the weighted geometric mean of the true class probabilities  $\bar{y}_j$  (weighted by the class probabilities for that training example). Taking the negative log likelihood gives

$$-\log \mathcal{L}(\bar{\mathbf{y}}_1, \dots, \bar{\mathbf{y}}_n | \mathbf{y}_1, \dots, \mathbf{y}_n) = -\sum_{j=1}^n \sum_{i=1}^M \bar{y}_{i,j} \log y_{i,j} \quad (14.4.59)$$

If we find the estimates  $\hat{y}_j$  (in terms of the parameters of the model) of  $\bar{y}_j$  which minimise this negative log likelihood, then it becomes the same problem as minimising the cross entropy.

#### 14.4.2 Convolutional Neural Networks

#### 14.4.3 Recurrent Neural Networks [70]

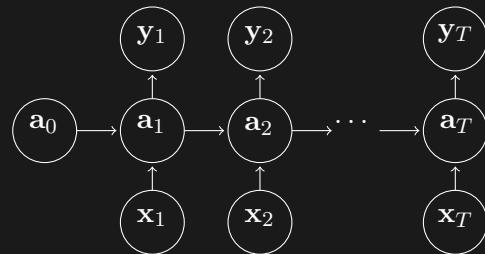
A recurrent neural network (RNN) can be thought of as a generalisation of a multi-layer perception network with cyclical connections, and now takes as input a sequence of  $T$  vectors

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T) \quad (14.4.60)$$

and outputs a sequence of  $T$  vectors

$$\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_T) \quad (14.4.61)$$

An RNN also keeps ‘in memory’ a sequence of hidden states  $(\mathbf{a}_0, \dots, \mathbf{a}_T)$  which are the activations in the hidden layers. The value of  $\mathbf{a}_0$  acts like an initial condition to the RNN, and may sometimes be set to the zero vector for simplicity.



For the structure of the RNN, we use the convention  $\{N_x = N_0, N_1, \dots, N_L = N_y\}$ , where  $N_0$  and  $N_L$  are the number of nodes in the input and output layer respectively, and  $N_\ell$  is the number of nodes in the  $\ell^{\text{th}}$  hidden layer. We consider ‘fully self-connected’ hidden layers, where each node in a hidden layer has a connection to every other node in that layer. A forward pass of an RNN is described as follows. For each  $t = 1, \dots, T$ , we compute the affine transformations for each layer  $\ell = 1, \dots, L - 1$ :

$${}^\ell z_t = {}_{\ell-1}^L W \times {}^{\ell-1} a_t + {}_{\ell-1}^r W \times {}^\ell a_{t-1} + {}^\ell b \quad (14.4.62)$$

where we can take  ${}^0 a_t = \mathbf{x}_t$ . The difference here from a standard neural network is that we have the matrix of ‘recurrent’ weights  ${}_{\ell-1}^r W$  on the previous activations  ${}^\ell a_{t-1}$  at the same layer, which arises from the self-connections. The activations for each layer are then computed by

$${}^\ell a_t = {}^\ell \sigma({}^\ell z_t) \quad (14.4.63)$$

with any general (differentiable) activation function of choice  $\sigma(\cdot)$ . The outputs in the final output layer (which doesn’t have self-connections) are then computed by

$$\mathbf{y}_t = {}^L \sigma({}^L z_t) \quad (14.4.64)$$

$$= {}^L \sigma({}^L_{L-1} W \times {}^L a_t + {}^L b) \quad (14.4.65)$$

From a state-space dynamical systems perspective with stacked state vector  $\mathbf{a}_t = ({}^1 a_t, \dots, {}^{L-1} a_t)$ , input  $\mathbf{x}_t$  and output  $\mathbf{y}_t$ , we can write the forward pass as

$$\mathbf{a}_t = f(\mathbf{a}_{t-1}, \mathbf{x}_t) \quad (14.4.66)$$

$$\mathbf{y}_t = h(\mathbf{a}_t) \quad (14.4.67)$$

where the state update equation is

$$f(\mathbf{a}_{t-1}, \mathbf{x}_t) = \begin{bmatrix} {}^1 \sigma({}^1_0 W \mathbf{x}_t + {}^1_1 W {}^1 a_{t-1} + {}^1 b) \\ {}^2 \sigma({}^2_1 W {}^1 \sigma(\dots) + {}^2_2 W {}^2 a_{t-1} + {}^2 b) \\ \vdots \\ {}^{L-1} \sigma({}^{L-1}_{L-2} W {}^{L-2} \sigma(\dots) + {}^{L-1}_{L-1} W {}^{L-1} a_{t-1} + {}^{L-1} b) \end{bmatrix} \quad (14.4.68)$$

and the output equation is

$$h(\mathbf{a}_t) = {}^L \sigma({}^L_{L-1} W \times {}^L a_t + {}^L b) \quad (14.4.69)$$

## Backpropagation Through Time

The same principles for backpropagation in multi-layer perceptrons can be used to derive the backpropagation rules in RNNs. However there are a few key differences, the first of which is defining an appropriate cost function for sequence data. Suppose the training data consists of  $n$  training examples, each with an input and output sequence indexed by  $i$ :

$$\mathbf{X}[i] = (\mathbf{x}_1[i], \dots, \mathbf{x}_{T_i}[i]) \quad (14.4.70)$$

$$\mathbf{Y}[i] = (\mathbf{y}_1[i], \dots, \mathbf{y}_{T_i}[i]) \quad (14.4.71)$$

Then for each training example, an appropriate loss (for a regression problem) is:

$$\mathcal{L}[i] = \sum_{t=1}^{T_i} \|\hat{\mathbf{y}}_t[i] - \mathbf{y}_t[i]\|^2 \quad (14.4.72)$$

and we take the total cost as

$$C = \sum_{i=1}^n \mathcal{L}[i] \quad (14.4.73)$$

Define  ${}^\ell\delta_t[i] := \nabla_{{}^\ell z_t} \mathcal{L}[i]$  as the gradient of the loss of the  $i^{\text{th}}$  training example for time  $t$  with respect to the affine transformation of the  $\ell^{\text{th}}$  layer. For an arbitrary layer we can write

$${}^\ell z_t = {}^{\ell-1}W^{\ell-1}\boldsymbol{\sigma}\left({}^{\ell-1}z_t\right) + {}^{\ell}W^\ell\boldsymbol{\sigma}\left({}^\ell z_{t-1}\right) + {}^\ell b \quad (14.4.74)$$

We see that  ${}^\ell z_t$  is used in the calculation of  ${}^{\ell+1}z_t$  and  ${}^\ell z_{t+1}$ . Hence it follows (in similar way as derived for multi-layer perceptrons) that the backpropagation calculation of  ${}^\ell\delta_t[i]$  should take the form

$${}^\ell\delta_t[i] = \left( \left( {}^{\ell+1}W \right)^\top {}^{\ell+1}\delta_t[i] \right) \odot {}^\ell\boldsymbol{\sigma}'\left({}^\ell z_t\right) + \left( ({}^\ell W)^\top {}^\ell\delta_{t+1}[i] \right) \odot {}^\ell\boldsymbol{\sigma}'\left({}^\ell z_t\right) \quad (14.4.75)$$

$$= \left( \left( {}^{\ell+1}W \right)^\top {}^{\ell+1}\delta_t[i] + ({}^\ell W)^\top {}^\ell\delta_{t+1}[i] \right) \odot {}^\ell\boldsymbol{\sigma}'\left({}^\ell z_t\right) \quad (14.4.76)$$

for  $\ell = 1, \dots, L-1$ , and

$${}^L\delta_t[i] = \nabla_{{}^L z_t} \mathcal{L}[i] \quad (14.4.77)$$

$$= \nabla_{{}^L y_t} \mathcal{L}[i] \odot {}^L\boldsymbol{\sigma}'\left({}^L z_t\right) \quad (14.4.78)$$

by the chain rule. Thus, this means that in order to compute  ${}^\ell\delta_t[i]$  for all  $\ell = 1, \dots, L$  and  $t = 1, \dots, T_i$ , we need to start at  $t = T_i$  and recursively compute backwards through the layers (taking  ${}^\ell\delta_{T_i+1}[i] = 0$ ). Then we work backwards through time for  $t = T_i - 1$ , etc. Once this is done, the gradient of  $\mathcal{L}[i]$  with respect to the weights and biases takes the same form as in multi-layer perceptrons, except we need to sum across all time (since the weights are used in every calculation in the sequence during the forward pass):

$$\frac{\partial \mathcal{L}[i]}{\partial {}^{\ell-1}W}^\top = \sum_{t=1}^{T_i} {}^\ell\delta_t[i] \cdot {}^{\ell-1}a_t[i]^\top \quad (14.4.79)$$

$$\frac{\partial \mathcal{L}[i]}{\partial {}^\ell b}^\top = \sum_{t=1}^{T_i} {}^\ell\delta_t[i] \quad (14.4.80)$$

This gradient can then be used in an appropriate algorithm to optimise the parameters of the network.

### Long Short-Term Memory Unit

A particular architecture of recurrent neural network can be constructed using long short-term memory (LSTM) units. A single LSTM unit consists of  $h$  LSTM ‘cells’. We abstract away from the higher level architecture, and consider the input-output mapping of a single LSTM unit. Let  $\mathbf{x}_t \in \mathbb{R}^d$  be the input and let  $\mathbf{a}_t \in \mathbb{R}^h$  be the output. Note that  $\mathbf{x}_t$  need not necessarily be the network’s input; it could be the output activations from a previous hidden layer, and  $\mathbf{a}_t$  are the activations into the next hidden layer. A forward pass inside an LSTM unit consists of interactions between four different intermediate activation vectors:

- $\mathbf{i}_t \in \mathbb{R}^h$  is called the *input gate* activation vector.
- $\bar{\mathbf{c}}_t \in \mathbb{R}^h$  is called the *cell-input* activation vector.
- $\mathbf{o}_t \in \mathbb{R}^h$  is called the *output gate* activation vector.

- $\mathbf{f}_t \in \mathbb{R}^h$  is called the *forget gate* activation vector.

Each LSTM unit also possess a cell-state vector  $\mathbf{c}_t$ . To conduct a forward pass, we need network weights  $W_i, W_{\bar{c}}, W_o, W_f \in \mathbb{R}^{h \times d}$  and  $U_i, U_{\bar{c}}, U_o, U_f \in \mathbb{R}^{h \times h}$  as well as biases  $b_i, b_{\bar{c}}, b_o, b_f \in \mathbb{R}^{h \times 1}$ . To keep notation simple, we suppress the layer indices. A forward pass consists of the following four intermediate activations:

$$\mathbf{i}_t = \sigma_i (W_i \mathbf{x}_t + U_i \mathbf{a}_{t-1} + b_i) \quad (14.4.81)$$

$$\bar{\mathbf{c}}_t = \sigma_{\bar{c}} (W_{\bar{c}} \mathbf{x}_t + U_{\bar{c}} \mathbf{a}_{t-1} + b_{\bar{c}}) \quad (14.4.82)$$

$$\mathbf{o}_t = \sigma_o (W_o \mathbf{x}_t + U_o \mathbf{a}_{t-1} + b_o) \quad (14.4.83)$$

$$\mathbf{f}_t = \sigma_f (W_f \mathbf{x}_t + U_f \mathbf{a}_{t-1} + b_f) \quad (14.4.84)$$

where  $\sigma_i, \sigma_{\bar{c}}, \sigma_o, \sigma_f : \mathbb{R}^h \rightarrow \mathbb{R}^h$  can be chosen as arbitrary element-wise activation functions (such as the sigmoid or tanh). Then the cell-state update is governed by

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \bar{\mathbf{c}}_t \quad (14.4.85)$$

where  $\circ$  is the Hadamard product. The LSTM unit's output activation  $\mathbf{a}_t$  is then determined by

$$\mathbf{a}_t = \mathbf{o}_t \circ \sigma_h (\mathbf{c}_t) \quad (14.4.86)$$

with another activation function  $\sigma_h : \mathbb{R}^h \rightarrow \mathbb{R}^h$ . From a systems theory perspective, we can view each LSTM unit as a dynamical system with states  $(\mathbf{a}_t, \mathbf{c}_t)$  and the update equation

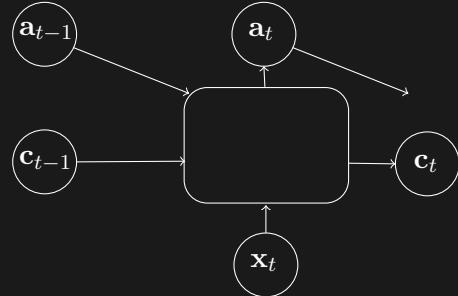
$$\begin{bmatrix} \mathbf{a}_t \\ \mathbf{c}_t \end{bmatrix} = g(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{x}_t) \quad (14.4.87)$$

$$= \begin{bmatrix} g_a(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{x}_t) \\ g_c(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{x}_t) \end{bmatrix} \quad (14.4.88)$$

where

$$g_a(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{x}_t) = \sigma_o (W_o \mathbf{x}_t + U_o \mathbf{a}_{t-1} + b_o) \circ \sigma_h (\sigma_f (W_f \mathbf{x}_t + U_f \mathbf{a}_{t-1} + b_f) \circ \mathbf{c}_{t-1}) \quad (14.4.89)$$

$$g_c(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{x}_t) = \sigma_f (W_f \mathbf{x}_t + U_f \mathbf{a}_{t-1} + b_f) \circ \mathbf{c}_{t-1} + \sigma_i (W_i \mathbf{x}_t + U_i \mathbf{a}_{t-1} + b_i) \circ \sigma_{\bar{c}} (W_{\bar{c}} \mathbf{x}_t + U_{\bar{c}} \mathbf{a}_{t-1} + b_{\bar{c}}) \quad (14.4.90)$$



#### 14.4.4 Mixture Density Networks

#### 14.4.5 Generative Adversarial Networks [39]

### 14.5 Gaussian Process Regression

#### 14.5.1 Gaussian Process Classification

#### Relevance Vector Machines [25]

### 14.6 Ensemble Methods

#### 14.6.1 Bagging

A portmanteau of ‘bootstrap aggregation’, the bagging technique trains several different models separately, and then ‘votes’ for the output on test examples. Consider  $k$  different models. In bagging, each of the models has been trained on a dataset which is resampled from the original training data with replacement. In a regression example, suppose the  $i^{\text{th}}$  model has prediction error  $\epsilon_i$  on the test data, where  $\epsilon_i$  has zero-mean. Let the expected squared error be  $\mathbb{E}[\epsilon_i^2] = \text{Var}(\epsilon) = v$ , and let the covariances between errors on different models be  $\text{Cov}(\epsilon_i, \epsilon_j) = \mathbb{E}[\epsilon_i \epsilon_j] = c$ . By averaging predictions across the ensemble, the ensemble prediction errors is given by  $\frac{1}{k} \sum_{i=1}^k \epsilon_i$ . Then the expected squared error of the ensemble is

$$\mathbb{E} \left[ \left( \frac{1}{k} \sum_{i=1}^k \epsilon_i \right)^2 \right] = \frac{1}{k^2} \mathbb{E} \left[ \sum_{i=1}^k \epsilon_i \sum_{j=1}^k \epsilon_j \right] \quad (14.6.1)$$

$$= \frac{1}{k^2} \mathbb{E} \left[ \sum_{i=1}^k \left( \epsilon_i^2 + \sum_{j:j \neq i}^k \epsilon_i \epsilon_j \right) \right] \quad (14.6.2)$$

$$= \frac{1}{k} \mathbb{E} \left[ \epsilon_i^2 + \sum_{j:j \neq i}^k \epsilon_i \epsilon_j \right] \quad (14.6.3)$$

$$= \frac{1}{k} \text{Var}(\epsilon) + \frac{k-1}{k} \text{Cov}(\epsilon_i, \epsilon_j) \quad (14.6.4)$$

$$= \frac{1}{k} v + \frac{k-1}{k} c \quad (14.6.5)$$

Suppose errors are perfectly correlated and  $c = v$  (e.g. the exact same training data is used for each model). Then the mean squared error of the ensemble is  $v$ . On the other hand if errors are perfectly uncorrelated such that  $c = 0$ , then the mean squared error by the ensemble is only  $\frac{1}{k}v$ . It shows that in this case, the ensemble can never perform worse than any of its members [69].

#### 14.6.2 Boosting

A generic boosting algorithm takes a ‘weak learner’ (e.g. only slightly better than a random 50-50 guess on a classification problem), and produces a sequence of weak learners  $G_m(x)$  for  $m = 1, \dots, M$  to form a ‘strong’ committee of learners that is a weighting of each weak learner. An example is that for a classification problem into classes  $\{-1, 1\}$ , this weighting is given by

$$G(x) = \text{sign} \left( \sum_{m=1}^M \alpha_m G_m(x) \right) \quad (14.6.6)$$

At each step  $m$ , the training data is reweighted to put emphasis on the examples that the previous learner got wrong.

## AdaBoost [80, 221]

In AdaBoost (short for Adaptive Boosting), the weightings for each observation  $(x_i, y_i)$  out of  $N$  observations total start off at  $w_i = \frac{1}{N}$ . At each step  $m$ , the classifier  $G_m(x)$  is fitted from the weighted data and the weighted classification error rate  $e_m$  is computed. The learner weighted is calculated by

$$\alpha_m = \log \left( \frac{1 - e_m}{e_m} \right) \quad (14.6.7)$$

and then each weight is updated according to

$$w_i \leftarrow w_i \exp [\alpha_m \mathbb{I}(y_i \neq G_m(x_i))] \quad (14.6.8)$$

where  $\mathbb{I}(y_i \neq G_m(x_i))$  is an indicator function. The weightings can be re-normalised if needed. Thus we can see that the weight for an observation which the learner got wrong is increased, while the weight for an observation which the learner got correct stays the same. Also note that the learner weighting  $\alpha_m$  is decreasing in the error rate  $e_m$ .

## Gradient Boosting

### 14.6.3 Stacking [80, 221]

In stacking, there are  $M$  ‘first-level learners’, and one ‘second-level learner’. These learners may be not necessarily be from the same class of learning algorithm. Let the estimate from the  $m^{\text{th}}$  learner be denoted  $\hat{f}_m(x)$ . Then out of  $N$  training observations, let  $\hat{f}_m^{(-i)}(x)$  denote the estimate with the  $i^{\text{th}}$  training observation removed. Thus we can view  $\hat{f}_m^{(-i)}(x_i)$  as a ‘prediction’ on  $x_i$  (since  $x_i$  was not used in training the learner). The second learner (denoted by  $\hat{f}(z)$ ) is then trained over a dataset created by the first learners. Each observation of the second learner consists of the ‘leave-one-out’ outputs of the first learners as its inputs, and the original label as its output. Concretely, denote the  $i^{\text{th}}$  training observation of the second learner by  $(z_i, y_i)$ , where

$$z_i = \left( \hat{f}_1^{(-i)}(x_i), \dots, \hat{f}_M^{(-i)}(x_i) \right) \quad (14.6.9)$$

The leave-one-out approach is to prevent overfitting. Thus, a stacked prediction  $\hat{h}(x)$  is given by  $\hat{h}(x) = \hat{f}(z)$  where

$$z = \left( \hat{f}_1(x), \dots, \hat{f}_M(x) \right) \quad (14.6.10)$$

### 14.6.4 Condorcet’s Jury Theorem

Condorcet’s jury theorem is a result which can partly be used to explain the strength of ensembling. Suppose there are  $n$  voters who vote on a decision. Each voter votes independently and votes for the correct decision with probability  $p$ . If  $p > \frac{1}{2}$ , then the probability that the majority decision is the correct decision approaches 1 as  $n \rightarrow \infty$ .

*Proof.* Treating each vote as a Bernoulli random variable, then by the strong law of large numbers, the sample mean of votes (or alternatively, the sample proportion of correct votes) converges to  $p > \frac{1}{2}$  almost surely as  $n \rightarrow \infty$ . Hence the probability that the majority vote is correct as  $n \rightarrow \infty$  is equal to one.  $\square$

This suggests that ensembling more models should lead to greater predictive accuracy, provided that each model is better than a coin flip. The assumption that each model is independent will usually not be satisfied since each model is trained on the same data. However, in an approach such as bagging with independent resampling, each model trained will be conditionally independent given the data.

## 14.7 Decision Trees [99]

### 14.7.1 Regression Trees

A regression tree consists of a partition of the predictor space  $\mathcal{X} \subseteq \mathbb{R}^d$  into  $J$  distinct and non-overlapping rectangular regions  $R_1 \subset \mathcal{X}, \dots, R_J \subset \mathcal{X}$ . For each  $j = 1, \dots, J$ , we assign the prediction for the corresponding region  $R_j$  to be the mean of observations in that region. That is, with sample  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ , we set

$$c_j = \frac{1}{|\{i : \mathbf{x}_i \in R_j\}|} \sum_{i:\mathbf{x}_i \in R_j} y_i \quad (14.7.1)$$

Then a prediction on a test example  $\mathbf{x}_*$  is made by

$$f(\mathbf{x}_*) = \sum_{j=1}^J c_j \mathbb{I}_{\mathbf{x}_* \in R_j} \quad (14.7.2)$$

### Regression Tree Learning

One method to obtain the regions  $R_1, \dots, R_J$  is known as binary recursive splitting. To begin, we split  $\mathcal{X}$  into two half-planes along a single predictor at  $s$ . That is, we designate

$$R_1(p, s) = \{\mathbf{x} = (x_1, \dots, x_p, \dots, x_d) \in \mathcal{X} : x_p < s\} \quad (14.7.3)$$

$$R_2(p, s) = \{\mathbf{x} = (x_1, \dots, x_p, \dots, x_d) \in \mathcal{X} : x_p \geq s\} \quad (14.7.4)$$

In order to choose  $p$  and  $s$ , we minimise the residual sum of squares (RSS) that would result in making that split:

$$\min_{p,s} \left\{ \sum_{i:\mathbf{x}_i \in R_1(p,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\mathbf{x}_i \in R_2(p,s)} (y_i - \hat{y}_{R_2})^2 \right\} \quad (14.7.5)$$

where  $\hat{y}_{R_1}$  is the mean response of training observations contained in  $R_1$ , and  $\hat{y}_{R_2}$  is the mean response of training observations contained in  $R_2$ . To make the third split, we further split one of the existing half-planes and decide on a predictor  $p$  and split  $s$  to minimise the RSS criterion. Explicitly, suppose we are currently at  $J$  regions. To ‘grow’ the tree and produce the  $(J+1)^{\text{th}}$  region, we solve

$$(j^*, p^*, s^*) = \operatorname{argmin}_{j,p,s} \left\{ \sum_{i:\mathbf{x}_i \in R_j^-(p,s)} (y_i - \hat{y}_{R_j^-})^2 + \sum_{i:\mathbf{x}_i \in R_j^+(p,s)} (y_i - \hat{y}_{R_j^+})^2 + \sum_{j' \neq j} \sum_{i:\mathbf{x}_i \in R_{j'}} (y_i - \hat{y}_{R_{j'}})^2 \right\} \quad (14.7.6)$$

where for shorthand:

$$R_j^-(p,s) := R_j \cap \{\mathbf{x} \in \mathcal{X} : x_p < s\} \quad (14.7.7)$$

$$R_j^+(p,s) := R_j \cap \{\mathbf{x} \in \mathcal{X} : x_p \geq s\} \quad (14.7.8)$$

Then perform the split by modifying current region:

$$R_{j^*} \leftarrow R_{j^*} \cap \{\mathbf{x} \in \mathcal{X} : x_{p^*} < s^*\} \quad (14.7.9)$$

and adding new region

$$R_{J+1} = R_{j^*} \cap \{\mathbf{x} \in \mathcal{X} : x_{p^*} \geq s^*\} \quad (14.7.10)$$

This procedure is repeated and the tree is grown until some specified termination condition, such as when each region has fewer than some number of observations. The end result is a binary tree where each terminal node corresponds to a partitioned region.

## Cost Complexity Pruning

To prevent overfitting, nodes can be ‘pruned’ from the tree. To achieve this, we introduce some hyperparameter  $\alpha \geq 0$ . Denote the fully-fitted tree by  $T_0 = \{R_1, R_2, \dots, R_J\}$ . Then a subtree  $T$  is such that  $T \subset T_0$ . For each  $\alpha \geq 0$ , we can find the subtree

$$T_\alpha = \operatorname{argmin}_{T \subset T_0} \left\{ \sum_{j=1}^{|T|} \sum_{i: \mathbf{x}_i \in R_j} (y_i - \hat{y}_{R_j})^2 + \alpha |T| \right\} \quad (14.7.11)$$

In this sense,  $\alpha$  acts like a regularisation parameter on having too many nodes. The value of  $\alpha$  may be chosen manually, or it may be chosen via a cross-validation method such as  $K$ -fold cross-validation.

### 14.7.2 Classification Trees

Decision trees can be used for multi-class classification with  $K$  classes. The main difference to regression trees is that for each region, we classify using the most commonly occurring class in the training sample for that region. The learning process also occurs in a similar manner, using binary recursive splitting. Rather than RSS however, the misclassification rate can be used as a performance criterion. Denoting  $\hat{p}_{R_j,k}$  as the proportion of observations from the  $k^{\text{th}}$  class in region  $R_j$ , the missclassification rate for terminal node  $j$  is given by  $1 - \max_k \hat{p}_{R_j,k}$ , and the general binary recursive splitting rule is defined by

$$(j^*, p^*, s^*) = \operatorname{argmin}_{j,p,s} \left\{ \left( 1 - \max_k \hat{p}_{R_j^-(p,s),k} \right) + \left( 1 - \max_k \hat{p}_{R_j^+(p,s),k} \right) + \sum_{j' \neq j} \left( 1 - \max_k \hat{p}_{R_{j'},k} \right) \right\} \quad (14.7.12)$$

#### Gini Index

An alternative criterion is to use the Gini index  $\sum_{k=1}^K \hat{p}_{R_j,k} (1 - \hat{p}_{R_j,k})$ , which is a measure of the total variance across the  $K$  classes (note that the variance of a single class is  $\hat{p}_{R_j,k} (1 - \hat{p}_{R_j,k})$  by the binomial distribution). A sketch of the graph of  $\hat{p}_{R_j,k} (1 - \hat{p}_{R_j,k})$  will reveal that it is smaller when  $\hat{p}_{R_j,k}$  is closer to zero or to one. Hence the Gini index takes on smaller values if all the proportions in a region are close to zero or one, as would be the case if the majority of examples in a region belonged to a single class. In this way, the Gini index is sometimes referred to as a measure of node ‘purity’. The binary recursive splitting rule using the Gini index is

$$(j^*, p^*, s^*) = \operatorname{argmin}_{j,p,s} \left\{ \sum_{k=1}^K \hat{p}_{R_j^-(p,s),k} \left( 1 - \hat{p}_{R_j^-(p,s),k} \right) + \sum_{k=1}^K \hat{p}_{R_j^+(p,s),k} \left( 1 - \hat{p}_{R_j^+(p,s),k} \right) + \sum_{j' \neq j} \sum_{k=1}^K \hat{p}_{R_{j'},k} \left( 1 - \hat{p}_{R_{j'},k} \right) \right\} \quad (14.7.13)$$

#### Information Gain in Decision Trees

Another criterion is the entropy of the distribution among the classes for a region, given by  $-\sum_{k=1}^K \hat{p}_{R_j,k} \log \hat{p}_{R_j,k}$ . From the characterisation of entropy it is clear that having the majority of examples in a single class for a region will result in smaller entropy. A sketch of the graph of  $-\hat{p}_{R_j,k} \log \hat{p}_{R_j,k}$  will also reveal that this performance criterion behaves similarly to the Gini index. This performance criterion is sometimes referred to as the information gain. The binary recursive splitting rule with information gain is

$$(j^*, p^*, s^*) = \underset{j,p,s}{\operatorname{argmin}} \left\{ - \sum_{k=1}^K \widehat{p}_{R_j^-(p,s),k} \log \widehat{p}_{R_j^-(p,s),k} - \sum_{k=1}^K \widehat{p}_{R_j^+(p,s),k} \log \widehat{p}_{R_j^+(p,s),k} - \sum_{j' \neq j} \sum_{k=1}^K \widehat{p}_{R_{j'},k} \log \widehat{p}_{R_{j'},k} \right\} \quad (14.7.14)$$

### 14.7.3 Random Forests

Random forests are an ensemble learning method using decision trees, built upon bagging. As in bagging, a ‘forest’ of trees are grown from bootstrapped training samples. However, each time a split is considered, a random sample of the  $d$  predictors are used as the split candidates. The rationale for this is that under conventional bagging, a strong predictor may cause the prediction from each tree to be highly correlated. Using a subset of predictors at each split has the effect of decorrelating the trees, making the ensemble average (or majority vote, in the case of classification) more reliable.

## 14.8 Dimensionality Reduction

### 14.8.1 Principal Component Analysis

Consider a random feature vector  $\mathbf{X} \in \mathbb{R}^d$ , with covariance matrix  $\operatorname{Cov}(\mathbf{X}) = \Sigma$ . By eigendecomposition, we can write the covariance matrix as

$$\Sigma = W\Lambda W^\top \quad (14.8.1)$$

where  $W$  is an orthonormal basis and  $\Lambda$  is a diagonal matrix of eigenvalues. Assume that the diagonal elements of  $\Lambda$  are ordered in decreasing magnitude:

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \quad (14.8.2)$$

with  $\lambda_1 \geq \dots \geq \lambda_d$ . Now suppose we want to find some way to reduce the dimension of  $\mathbf{X}$  while maintaining as much of the original variation in  $\mathbf{X}$  as possible. Recall that we can show

$$\operatorname{trace}(\Sigma) = \operatorname{trace}(W\Lambda W^\top) \quad (14.8.3)$$

$$= \operatorname{trace}(\Lambda W^\top W) \quad (14.8.4)$$

$$= \operatorname{trace}(\Lambda) \quad (14.8.5)$$

$$= \sum_{j=1}^d \lambda_j \quad (14.8.6)$$

If we wanted to reduce the dimension of the feature vector to  $L$ , then to maximise the trace of the covariance of the reduced-dimension feature vector, we should take the components containing the  $L$  highest eigenvalues as the ‘principal components’, of the ‘principal axis’ given by the basis  $W$ . Thus this amounts to taking the first  $L$  columns (i.e. eigenvectors) of  $W$ , denoted  $W_L \in \mathbb{R}^{d \times L}$ , and defining the reduced-dimension feature vector  $\mathbf{X}' \in \mathbb{R}^L$  as a result of the linear transformation

$$\mathbf{X}' = W_L^\top \mathbf{X} \quad (14.8.7)$$

with covariance  $\Sigma' = W_L^\top \Sigma W_L$ , which has trace

$$\operatorname{trace}(\Sigma') = \operatorname{trace}(W_L^\top \Sigma W_L) \quad (14.8.8)$$

Decomposing  $W = \begin{bmatrix} W_L & \widetilde{W} \end{bmatrix}$  and  $\Lambda = \text{diag}\{\Lambda_L, \widetilde{\Lambda}\}$ , we can write out  $\Sigma$  as

$$\Sigma = \begin{bmatrix} W_L & \widetilde{W} \end{bmatrix} \begin{bmatrix} \Lambda_L & \\ & \widetilde{\Lambda} \end{bmatrix} \begin{bmatrix} W_L^\top \\ \widetilde{W}^\top \end{bmatrix} \quad (14.8.9)$$

$$= W_L \Lambda_L W_L^\top + \widetilde{W} \widetilde{\Lambda} \widetilde{W}^\top \quad (14.8.10)$$

Hence

$$\text{trace}(\Sigma') = \text{trace}\left(W_L^\top (W_L \Lambda_L W_L^\top + \widetilde{W} \widetilde{\Lambda} \widetilde{W}^\top) W_L\right) \quad (14.8.11)$$

$$= \text{trace}\left(W_L^\top W_L \Lambda_L W_L^\top W_L + W_L^\top \widetilde{W} \widetilde{\Lambda} \widetilde{W}^\top W_L\right) \quad (14.8.12)$$

$$= \text{trace}(\Lambda_L) \quad (14.8.13)$$

$$= \sum_{j=1}^L \lambda_j \quad (14.8.14)$$

where we note that  $W_L^\top W_L = I_{L \times L}$  since each column in  $W_L$  is orthogonal to each other, and  $W_L^\top \widetilde{W} = \mathbf{0}_{L \times (d-L)}$  since each column in  $W_L$  is orthogonal to every other column in  $\widetilde{W}$ . Therefore this shows that  $\mathbf{X}'$  retains as much of the variance as it can after being reduced to  $L$  dimensions. This property motivates the following algorithm for dimension reduction on sample data. Suppose we have feature vectors  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  from a dataset of size  $n$ . Then use the procedure:

1. Obtain an estimate of the covariance matrix, denoted  $C$ . This estimate could be the unbiased estimate (with Bessel's correction), or the maximum likelihood estimator (with factor  $1/n$ ) can also be valid.
2. Perform eigendecomposition of  $C = V D V^\top$  where  $D$  is in order of decreasing magnitude by eigenvalues.
3. Take  $V_L$  as the first  $L$  columns of  $V$ , where  $L$  is the desired number of dimensions after reduction.
4. Project the features using the linear transformation  $V_L^\top$  so that each reduced-dimension feature satisfies  $\mathbf{x}'_i = V_L^\top \mathbf{x}_i$  for  $i = 1, \dots, n$ .

### Principal Component Regression

In principal component regression, the idea is to use principal component analysis on the features before running a regression. For a linear regression model

$$Y_i = \boldsymbol{\beta}^\top X_i + \varepsilon_i \quad (14.8.15)$$

where feature vector  $X_i \in \mathbb{R}^d$ , our goal is to obtain an estimator for  $\boldsymbol{\beta}$ . Using principal component analysis, we find the transformation  $V_L^\top$  so that the reduced-dimension feature vector in  $L$  dimensions is given by  $Z_i = V_L^\top X_i$ . Then  $Y_i$  can be regressed on  $Z_i$  (using an estimator such as ordinary least squares) to obtain the estimate  $\hat{\gamma} \in \mathbb{R}^L$ . Then to 'invert' this to obtain an estimator for  $\boldsymbol{\beta}$ , we can use

$$\hat{\boldsymbol{\beta}} = V_L \hat{\gamma} \quad (14.8.16)$$

so that

$$\hat{Y}_i = \hat{\boldsymbol{\beta}}^\top X_i \quad (14.8.17)$$

$$= \hat{\gamma}^\top V_L^\top X_i \quad (14.8.18)$$

$$= \hat{\gamma}^\top Z_i \quad (14.8.19)$$

Performing principal component regression can have the dual purpose of regularisation as well as dimensionality reduction.

## Maximum Likelihood Principal Component Analysis [25]

### 14.8.2 Factor Analysis [103]

In factor analysis, we begin with  $p$  observed random variables  $X_1, \dots, X_p$ , and assume that these can be expressed as linear functions of  $m$  (with  $m < p$ ) unobserved common factors, denoted  $Z_1, \dots, Z_m$ , written as

$$X_1 = \lambda_{1,1}Z_1 + \dots + \lambda_{1,m}Z_m + \varepsilon_1 \quad (14.8.20)$$

$$\vdots \quad (14.8.21)$$

$$X_p = \lambda_{p,1}Z_1 + \dots + \lambda_{p,m}Z_m + \varepsilon_p \quad (14.8.22)$$

where  $\varepsilon_1, \dots, \varepsilon_p$  are zero-mean stochastic error terms and the coefficients  $\lambda_{1,1}, \dots, \lambda_{p,m}$  are called the *factor loadings*. The system of equations can be compactly written in matrix form by

$$\mathbf{X} = \Lambda\mathbf{Z} + \boldsymbol{\varepsilon} \quad (14.8.23)$$

with  $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$  and some covariance  $\text{Cov}(\boldsymbol{\varepsilon}) = \Psi$ . We assume without loss of generality that  $\mathbb{E}[\mathbf{X}] = \mathbf{0}$ , because otherwise we can use  $\mathbf{X}' = \mathbf{X} - \mathbb{E}[\mathbf{X}]$ . We also assume that the factors are zero-mean and have covariance equal to the identity matrix:

$$\mathbb{E}[\mathbf{Z}] = \mathbf{0} \quad (14.8.24)$$

$$\text{Cov}(\mathbf{Z}) = I_{m \times m} \quad (14.8.25)$$

Note that the latter assumption implies that different factors are uncorrelated, and variances being equal to one is not restrictive since scalings can be absorbed into the factor loadings. Lastly, we assume that the errors are uncorrelated with the factors, i.e.  $\mathbb{E}[\mathbf{Z}\boldsymbol{\varepsilon}^\top] = \mathbf{0}$ . The problem is to find the random vector  $\mathbf{Z}$  and matrices  $\Lambda$  and  $\Psi$  such that all assumptions above are satisfied. Note that this solution will not be unique. To see why, first suppose  $\mathbf{X} = \Sigma$ , then we have

$$\Sigma = \text{Cov}(\mathbf{X}) \quad (14.8.26)$$

$$= \text{Cov}(\Lambda\mathbf{Z} + \boldsymbol{\varepsilon}) \quad (14.8.27)$$

$$= \Lambda \text{Cov}(\mathbf{Z}) \Lambda^\top + \text{Cov}(\boldsymbol{\varepsilon}) \quad (14.8.28)$$

$$= \Lambda\Lambda^\top + \Psi \quad (14.8.29)$$

For any solution  $\Lambda^*$ , then  $\Lambda^*W$  (where  $W$  is any orthogonal matrix) will also satisfy the assumptions, since

$$\Lambda^*W(\Lambda^*W)^\top = \Lambda^*WW^\top(\Lambda^*)^\top \quad (14.8.30)$$

$$= \Lambda^*(\Lambda^*)^\top \quad (14.8.31)$$

Although factor analysis seems similar to principal components analysis, the notable difference is that factor analysis prescribes a particular causal model to the observations  $\mathbf{X}$ , whereas PCA does not. There are many different methods to obtain the factors [141]. One approach is to use PCA with the first  $m$  principal components to obtain an initial estimate for  $\Lambda$ , which can then be iterated upon.

### 14.8.3 Canonical Correlation Analysis

Given two random variables  $X \in \mathbb{R}^q$  and  $Y \in \mathbb{R}^p$ , canonical correlation analysis aims to find the ‘most interesting’ linear combination [79] with some vectors  $a \in \mathbb{R}^q$  and  $b \in \mathbb{R}^p$ , such that

the correlation  $\text{Corr}(a^\top X, b^\top Y)$  is maximised. Firstly to simplify notation, let  $U = a^\top X$  and  $V = b^\top Y$ . Then denote

$$\text{Cov}(X) = \Sigma_{XX} \quad (14.8.32)$$

$$\text{Cov}(Y) = \Sigma_{YY} \quad (14.8.33)$$

$$\text{Cov}(X, Y) = \Sigma_{XY} \quad (14.8.34)$$

where we remind ourselves that  $\Sigma_{XY} \in \mathbb{R}^{q \times p}$ . Then we have

$$\text{Cov}(U) = a^\top \Sigma_{XX} a \quad (14.8.35)$$

$$\text{Cov}(V) = b^\top \Sigma_{YY} b \quad (14.8.36)$$

$$\text{Cov}(U, V) = a^\top \Sigma_{XY} b \quad (14.8.37)$$

Hence the correlation  $\rho = \text{Corr}(U, V)$  is given by

$$\rho = \frac{\text{Cov}(U, V)}{\sqrt{\text{Cov}(U)} \cdot \sqrt{\text{Cov}(V)}} \quad (14.8.38)$$

$$= \frac{a^\top \Sigma_{XY} b}{\sqrt{a^\top \Sigma_{XX} a} \cdot \sqrt{b^\top \Sigma_{YY} b}} \quad (14.8.39)$$

Now define  $c = \Sigma_{XX}^{1/2} a$  and  $d = \Sigma_{YY}^{1/2} b$  so that

$$\rho = \frac{c^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} d}{\sqrt{c^\top c} \cdot \sqrt{d^\top d}} \quad (14.8.40)$$

where for clarity we have obtained the denominator by

$$a^\top \Sigma_{XX} a = c^\top \Sigma_{XX}^{-1/2} \left( \Sigma_{XX}^{1/2} \Sigma_{XX}^{1/2} \right) \Sigma_{XX}^{-1/2} c \quad (14.8.41)$$

$$= c^\top c \quad (14.8.42)$$

and analogously for  $d$ . Then using the Cauchy-Schwarz inequality in the numerator, this gives

$$\left( c^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \right) \cdot d \leq \left\| \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} c \right\| \cdot \|d\| \quad (14.8.43)$$

$$\leq \left( c^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} c \right)^{1/2} \left( d^\top d \right)^{1/2} \quad (14.8.44)$$

Hence

$$\rho \leq \frac{\left( c^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2} c \right)^{1/2} (d^\top d)^{1/2}}{\sqrt{c^\top c} \cdot \sqrt{d^\top d}} \quad (14.8.45)$$

$$\leq \frac{\left( c^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2} c \right)^{1/2}}{\sqrt{c^\top c}} \quad (14.8.46)$$

Note that the Cauchy-Schwarz inequality holds with equality if and only if the vectors  $\Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} c$  and  $d$  are collinear, i.e. the angle between the two angles is zero, meaning the cosine is one and the dot product is maximised. Hence the correlation is maximised by first finding

$$c^* = \underset{c}{\operatorname{argmax}} \left\{ \frac{\left( c^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2} c \right)^{1/2}}{\sqrt{c^\top c}} \right\} \quad (14.8.47)$$

and then setting  $d^*$  collinear to  $\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}c^*$ . Squaring the numerator and denominator, we can instead solve

$$c^* = \operatorname{argmax}_c \left\{ \frac{c^\top \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}c}{c^\top c} \right\} \quad (14.8.48)$$

$$= \operatorname{argmax}_c \left\{ \frac{c^\top \Sigma c}{c^\top I c} \right\} \quad (14.8.49)$$

where  $\Sigma = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ . Recognising that this problem takes the same form as that encountered in the derivation of Fisher's linear discriminant (involving the Rayleigh quotient), we thus find that  $c^*$  is given by the eigenvector corresponding to the largest eigenvalue for  $\Sigma$  (i.e.  $\Sigma c^* = \lambda_{\max} c^*$ ). The following steps summarise the overall procedure.

1. Let  $c^*$  be the eigenvector corresponding to the largest eigenvalue for  $\Sigma = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ .
2. Then  $d^*$  can be set as any vector proportional to  $\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}c^*$ , which does not affect the resulting correlation.
3. Obtain  $a^* = \Sigma_{XX}^{-1/2}c^*$  and  $b^* = \Sigma_{YY}^{-1/2}d^*$ .

This ensures that the correlation between the random variables  $U$  and  $V$  is maximised. Note that if we simply set  $d^* = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}c^*$ , then  $b^*$  will be related to  $a^*$  by

$$b^* = \Sigma_{YY}^{-1/2}\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}c^* \quad (14.8.50)$$

$$= \Sigma_{YY}^{-1}\Sigma_{YX}a^* \quad (14.8.51)$$

#### 14.8.4 Random Projections

Consider a high dimensional feature space  $\mathbb{R}^D$ . The goal is to reduce the dimensionality of data in this feature space onto  $\mathbb{R}^d$  with  $d < D$ , and ideally  $d \ll D$ . We can do so by finding an appropriate mapping  $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$  such that the structure of the data is preserved in some sense. In the technique of random projections, this mapping is generated randomly, with some notion of guarantee that the structure of the data will be preserved.

##### Johnson-Lindenstrauss Lemma [30]

Let  $A$  be a finite set of points in  $\mathbb{R}^D$ :

$$A = \{a_1, \dots, a_n\} \quad (14.8.52)$$

such that  $a_1, \dots, a_n \in \mathbb{R}^D$ . For  $\varepsilon \in (0, 1)$ , a mapping  $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$  is called an  $\varepsilon$ -isometry on  $A$  if for every pair  $a, a' \in A$ , we have:

$$(1 - \varepsilon) \|a - a'\|^2 \leq \|f(a) - f(a')\|^2 \leq (1 + \varepsilon) \|a - a'\|^2 \quad (14.8.53)$$

That is, an  $\varepsilon$ -isometry mapping with small  $\varepsilon$  approximately preserves distances. Of course, this should not seem too hard if  $d$  is large, but the goal is to find an  $\varepsilon$ -isometry where  $d$  is relatively small. The Johnson-Lindenstrauss lemma shows that it is possible to find a linear  $\varepsilon$ -isometry whenever

$$d \geq \kappa \varepsilon^{-2} \log(n) \quad (14.8.54)$$

where  $\kappa$  is an absolute constant. Critically, this lower bound does not depend on  $D$  (in fact we could be projecting from an infinite dimensional space, such as a Hilbert space). In particular,

consider the linear isometry  $W : \mathbb{R}^D \rightarrow \mathbb{R}^d$  represented by the  $d \times D$  matrix

$$\mathbf{W} = \frac{1}{\sqrt{d}} \begin{bmatrix} X_{11} & \dots & X_{1D} \\ \vdots & \ddots & \vdots \\ X_{d1} & \dots & X_{dD} \end{bmatrix} \quad (14.8.55)$$

where each  $X_{ij}$  for  $i = 1, \dots, d$  and  $j = 1, \dots, D$  are sub-Gaussian random variables with variance factor  $v \geq 1$  and

$$\mathbb{E}[X_{ij}] = 0 \quad (14.8.56)$$

$$\text{Var}(X_{ij}) = 1 \quad (14.8.57)$$

If  $d$  is chosen such that  $d \geq 100v^2\varepsilon^{-2} \log(n/\sqrt{\delta})$ , then with probability at least  $1 - \delta$ , the random projection  $W$  will be an  $\varepsilon$ -isometry on  $A$ .

*Proof.* Treat  $\alpha = [a_1 \ \dots \ a_D]^\top \in \mathbb{R}^D$  as an arbitrary vector between two points  $a, a' \in A$ . Let

$$W(\alpha) = \mathbf{W}\alpha \quad (14.8.58)$$

$$= \frac{1}{\sqrt{d}} \begin{bmatrix} X_{11} & \dots & X_{1D} \\ \vdots & \ddots & \vdots \\ X_{d1} & \dots & X_{dD} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_D \end{bmatrix} \quad (14.8.59)$$

Denote the  $i^{\text{th}}$  element of  $W(\alpha)$  without the normalisation of factor  $d^{-1/2}$  by

$$W_i(\alpha) = \sum_{j=1}^D \alpha_j X_{ij} \quad (14.8.60)$$

Using the facts that the  $X_{ij}$  are independent (implying zero covariance), and that  $\mathbb{E}[X_{ij}^2] = \text{Var}(X_{ij}) - \mathbb{E}[X_{ij}]^2 = 1$ , we can show that  $W$  is an exact isometry in expectation. Firstly,

$$\mathbb{E}[W_i(\alpha)^2] = \mathbb{E}\left[\left(\sum_{j=1}^D \alpha_j X_{ij}\right)^2\right] \quad (14.8.61)$$

$$= \mathbb{E}\left[\left(\sum_{j=1}^D \alpha_j X_{ij}\right)\left(\sum_{j=1}^D \alpha_j X_{ij}\right)\right] \quad (14.8.62)$$

$$= \mathbb{E}\left[\left(\sum_{j=1}^D \alpha_j^2 X_{ij}^2\right)\right] \quad (14.8.63)$$

$$= \sum_{j=1}^D \alpha_j^2 \mathbb{E}[X_{ij}^2] \quad (14.8.64)$$

$$= \sum_{j=1}^D \alpha_j^2 \quad (14.8.65)$$

$$= \|\alpha\|^2 \quad (14.8.66)$$

so that by returning the normalisation factor, we have

$$\mathbb{E}[\|W(\alpha)\|^2] = \mathbb{E}\left[\sum_{i=1}^d \left(\frac{1}{\sqrt{d}} W_i(\alpha)\right)^2\right] \quad (14.8.67)$$

$$= \frac{1}{d} \sum_{i=1}^d \mathbb{E} [W_i(\alpha)^2] \quad (14.8.68)$$

$$= \|\alpha\|^2 \quad (14.8.69)$$

Now to show that  $W$  is an  $\varepsilon$ -isometry with high probability, consider the set

$$T = \left\{ \frac{a - a'}{\|a - a'\|} : a, a' \in A, a \neq a' \right\} \quad (14.8.70)$$

which is the set of all vectors between pairs of points in  $A$ , projected onto the unit hypersphere in  $\mathbb{R}^D$ . Note that  $T$  has cardinality

$$N := |T| \quad (14.8.71)$$

$$= \frac{n(n-1)}{2} \quad (14.8.72)$$

$$\leq \frac{n^2}{2} \quad (14.8.73)$$

Then it suffices to show that  $\sup_{\alpha \in T} \left\{ \left| \|W(a)\|^2 - 1 \right| \right\} \leq \varepsilon$  with high probability, because this condition implies for all pairs  $a, a' \in A$

$$\varepsilon \geq \left| \left\| W \left( \frac{a - a'}{\|a - a'\|} \right) \right\|^2 - 1 \right| \quad (14.8.74)$$

$$= \left| \frac{1}{\|a - a'\|^2} \|W(a - a')\|^2 - 1 \right| \quad (14.8.75)$$

as  $W$  is a linear mapping. Then multiplying out by  $\|a - a'\|^2$ ,

$$\left| \|W(a - a')\|^2 - \|a - a'\|^2 \right| \leq \varepsilon \|a - a'\|^2 \quad (14.8.76)$$

Then rewriting as a two-sided inequality:

$$(1 - \varepsilon) \|a - a'\|^2 \leq \|W(a - a')\|^2 \leq (1 + \varepsilon) \|a - a'\|^2 \quad (14.8.77)$$

which recovers the definition of an  $\varepsilon$ -isometry. We proceed by observing that as a linear combination of centered sub-Gaussian random variables,  $W_i(\alpha)$  for all  $\alpha \in T$  will also be sub-Gaussian, with variance factor computed as follows

$$\mathbb{E} [e^{\lambda W_i(\alpha)}] = \mathbb{E} \left[ \exp \left( \lambda \sum_{j=1}^D \alpha_j X_{ij} \right) \right] \quad (14.8.78)$$

$$= \prod_{j=1}^D \mathbb{E} [\exp (\lambda \alpha_j X_{ij})] \leq \prod_{j=1}^D \exp \left( \lambda^2 \alpha_j^2 \frac{v}{2} \right) \quad (14.8.79)$$

$$= \exp \left( \frac{\lambda^2 v}{2} \sum_{j=1}^D \alpha_j^2 \right) \quad (14.8.80)$$

$$= \exp \left( \frac{\lambda^2 v}{2} \|\alpha\|^2 \right) \quad (14.8.81)$$

$$= \exp \left( \frac{\lambda^2 v}{2} \right) \quad (14.8.82)$$

as  $\|\alpha\|^2 = 1$  for all  $\alpha \in T$  by definition, and we have used the bound on the moment generating function for each of the sub-Gaussian  $X_{ij}$ . Thus,  $W_i(\alpha)$  has variance factor  $v$ . Using the moment bound characterisation of sub-Gaussians, we get for every integer  $q \geq 1$

$$\mathbb{E} [W_i(\alpha)^{2q}] \leq \frac{q!}{2} (4v)^q \quad (14.8.83)$$

Introduce the centered sum

$$S = \sum_{i=1}^d \left( W_i(\alpha)^2 - \mathbb{E} [W_i(\alpha)^2] \right) \quad (14.8.84)$$

with  $\mathbb{E} [W_i(\alpha)^2] = \|\alpha\|^2 = 1$  for all  $\alpha \in T$ , so

$$S = \sum_{i=1}^d \left( W_i(\alpha)^2 - 1 \right) \quad (14.8.85)$$

We establish the following moment bound for  $W_i(\alpha)^2$ :

$$\mathbb{E} \left[ \sum_{i=1}^d W_i(\alpha)^{2q} \right] = \sum_{i=1}^d \mathbb{E} [W_i(\alpha)^{2q}] \quad (14.8.86)$$

$$\leq d \frac{q!}{2} (4v)^q \quad (14.8.87)$$

$$= d \frac{q!}{2} (4v)^2 (4v)^{q-2} \quad (14.8.88)$$

and the following bound on the sum of second moments:

$$\sum_{i=1}^d \mathbb{E} \left[ (W_i(\alpha)^2)^2 \right] = \sum_{i=1}^d \mathbb{E} [W_i(\alpha)^4] \quad (14.8.89)$$

$$\leq \sum_{i=1}^d \frac{2!}{2} (4v)^2 \quad (14.8.90)$$

$$= d (4v)^2 \quad (14.8.91)$$

Also note  $W_1(\alpha)^2, \dots, W_d(\alpha)^2$  are independent (given  $\alpha$ ), and also non-negative. Therefore, Bernstein's inequality can be used with constant  $c = 4v$  and  $d(4v)^2$  as the constant bound on the sum of second moments. This yields

$$\Pr \left( S \geq \sqrt{2d(4v)^2 t} + 4vt \right) \leq e^{-t} \quad (14.8.92)$$

for all  $t > 0$ . Moreover, we can apply Bernstein's inequality to the non-positive random variables  $-W_1(\alpha)^2, \dots, -W_d(\alpha)^2$  (where the left-truncated moment bound is satisfied trivially since  $\max \{-W_i(\alpha)^2, 0\} = 0$ ) to obtain

$$\Pr \left( -S \geq \sqrt{2d(4v)^2 t} + 4vt \right) \leq e^{-t} \quad (14.8.93)$$

This gives the two-sided inequality

$$\Pr \left( \left| \sum_{i=1}^d \left( W_i(\alpha)^2 - 1 \right) \right| \geq \sqrt{2d(4v)^2 t} + 4vt \right) = \Pr \left( |S| \geq \sqrt{2d(4v)^2 t} + 4vt \right) \quad (14.8.94)$$

$$\leq 2e^{-t} \quad (14.8.95)$$

Applying the union bound over  $\alpha \in T$ ,

$$\Pr \left( \sup_{\alpha \in T} \left\{ \left| \sum_{i=1}^d (W_i(\alpha)^2 - 1) \right| \right\} \geq \sqrt{2d(4v)^2 t} + 4vt \right) = \Pr \left( \bigcup_{\alpha \in T} \left\{ |S| \geq \sqrt{2d(4v)^2 t} + 4vt \right\} \right) \quad (14.8.96)$$

$$\leq \sum_{\alpha \in T} \Pr \left( |S| \geq \sqrt{2d(4v)^2 t} + 4vt \right) \quad (14.8.97)$$

$$\leq 2 \sum_{\alpha \in T} e^{-t} \quad (14.8.98)$$

$$= 2N e^{-t} \quad (14.8.99)$$

$$\leq n^2 e^{-t} \quad (14.8.100)$$

where the last inequality is since  $N \leq n^2/2$  as established above. If we set  $t = \log(n^2/\delta)$  with  $\delta \in (0, 1)$ , we have

$$\delta \geq \Pr \left( \sup_{\alpha \in T} \left\{ \left| \sum_{i=1}^d (W_i(\alpha)^2 - 1) \right| \right\} \geq \sqrt{2d(4v)^2 \log(n^2/\delta)} + 4v \log(n^2/\delta) \right) \quad (14.8.101)$$

$$= \Pr \left( \sup_{\alpha \in T} \left\{ \left| \sum_{i=1}^d (W_i(\alpha)^2 - 1) \right| \right\} \geq 4v \sqrt{2d \cdot 2 \log(n/\sqrt{\delta})} + 4v \cdot 2 \log(n/\sqrt{\delta}) \right) \quad (14.8.102)$$

$$= \Pr \left( d \sup_{\alpha \in T} \left\{ |\|W(a)\|^2 - 1| \right\} \geq 8v \sqrt{d \log(n/\sqrt{\delta})} + 8v \log(n/\sqrt{\delta}) \right) \quad (14.8.103)$$

$$= \Pr \left( \sup_{\alpha \in T} \left\{ |\|W(a)\|^2 - 1| \right\} \geq \frac{8v \sqrt{d \log(n/\sqrt{\delta})}}{d} + \frac{8v \log(n/\sqrt{\delta})}{d} \right) \quad (14.8.104)$$

where we recalled the fact  $\|W(a)\|^2 = \sum_{i=1}^d W_i(\alpha)^2 / d$ . Consider a choice of  $d \geq 100v^2 \varepsilon^{-2} \log(n/\sqrt{\delta})$ . Then

$$\frac{8v \sqrt{d \log(n/\sqrt{\delta})}}{d} + \frac{8v \log(n/\sqrt{\delta})}{d} = 8v \sqrt{\frac{\log(n/\sqrt{\delta})}{d}} + \frac{8v \log(n/\sqrt{\delta})}{d} \quad (14.8.105)$$

$$\leq 8v \sqrt{\frac{\varepsilon^2}{100v^2}} + \frac{8\varepsilon^2}{100} \quad (14.8.106)$$

$$= \frac{16\varepsilon}{25} + \frac{2\varepsilon^2}{25} \quad (14.8.107)$$

$$= \varepsilon \left( \frac{16+2\varepsilon}{25} \right) \quad (14.8.108)$$

$$< \varepsilon \quad (14.8.109)$$

for  $\varepsilon \in (0, 1)$ . Hence

$$\Pr \left( \sup_{\alpha \in T} \left\{ |\|W(a)\|^2 - 1| \right\} > \varepsilon \right) \leq \delta \quad (14.8.110)$$

or

$$\Pr \left( \sup_{\alpha \in T} \left\{ |\|W(a)\|^2 - 1| \right\} \leq \varepsilon \right) \geq 1 - \delta \quad (14.8.111)$$

as required.  $\square$

### 14.8.5 Multidimensional Scaling [48]

#### Metric Multidimensional Scaling

Given a set of  $n$  objects (which could be very high-dimensional objects), suppose we some some notion of ‘dissimilarity’ between objects. That is, we can write out symmetric dissimilarity matrix

$$\Delta = \begin{bmatrix} \delta_{1,1} & \dots & \delta_{1,n} \\ \vdots & \ddots & \vdots \\ \delta_{n,1} & \dots & \delta_{n,n} \end{bmatrix} \quad (14.8.112)$$

where  $\delta_{i,j}$  represents the dissimilarity between the  $i^{\text{th}}$  and  $j^{\text{th}}$  objects. The goal of metric multidimensional scaling is to embed the objects into a  $p$ -dimensional space (typically Euclidean space) such that each object is represented by a  $p$ -dimensional vector and the dissimilarities between vectors (in terms of a distance metric) is preserved as well as possible. Explicitly (in the case of embedding in a Euclidean space), we wish to find  $n$  vectors  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  where each  $\mathbf{x}_i \in \mathbb{R}^p$  where the distance between vectors is the Euclidean distance:

$$d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\| \quad (14.8.113)$$

and the distances approximate the dissimilarities:

$$d_{i,j} \approx \delta_{i,j} \quad (14.8.114)$$

for all  $i, j$ . One approach for finding the set of vectors is to solve the least squares optimisation problem over the  $\frac{n(n-1)}{2}$  pairs of objects:

$$(\mathbf{x}_1^*, \dots, \mathbf{x}_n^*) = \underset{(\mathbf{x}_1, \dots, \mathbf{x}_n)}{\operatorname{argmin}} \sum_{i < j} (d_{i,j} - \delta_{i,j})^2 \quad (14.8.115)$$

Note that the optimal set of vectors  $(\mathbf{x}_1^*, \dots, \mathbf{x}_n^*)$  for this problem will be invariant to rotations and translations, so additional constraints should be added to ensure a unique solution. If  $p$  is chosen as 1, 2 or 3, then the resulting vectors can be plotted and easily visualised.

### 14.8.6 $t$ -Distributed Stochastic Neighbour Embedding

#### 14.8.7 Autoencoders

An autoencoder is a particular type of deep neural network which aims to learn the identity function (that is, it should ideally output  $f^*(x) = x$ ). This is achieved by training the network on identical input-output pairs  $(x_1, x_1), \dots, (x_n, x_n)$ . To use an autoencoder for dimensionality reduction, the structure of the network should begin with an ‘encoder’ and end with a ‘decoder’. The output of the encoder  $z = f_E(x)$  should be a smaller dimension than  $x$ , and the decoder  $\hat{x} = f_D(z)$  attempts to reconstruct  $x$  from  $z$ . The full network output is given by the composition

$$\hat{x} = f_D(z) \quad (14.8.116)$$

$$= f_D \circ f_E(x) \quad (14.8.117)$$

In principle, once trained, the autoencoder should have learned to ‘compress’ the vector  $x$  into the smaller dimensional  $z$ .

## Variational Autoencoders

### 14.9 Statistical Learning Theory

#### 14.9.1 Agnostic Probably Approximately Correct Learning [179]

Consider a sample of data pairs  $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$  with sample size  $m$ . Assume that each data pair is i.i.d.  $(x_i, y_i) \sim \mathcal{D}$ , where  $\mathcal{D}$  denotes an unknown distribution on support  $\mathcal{X} \times \mathcal{Y}$ . We refer to  $\mathcal{X}$  as the *instance space* and  $\mathcal{Y}$  as the *label space*. For simplicity, suppose that  $\mathcal{Y} = \{0, 1\}$  (i.e. we are considering binary classification). In a learning problem, we are given a hypothesis class  $\mathcal{H}$  which is a set such that for each  $h \in \mathcal{H}$ , the mapping  $h : \mathcal{X} \rightarrow \mathcal{Y}$  is a classifier. We define the risk (or generalisation error) of a classifier with respect to  $\mathcal{D}$  to be:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbb{I}_{\{h(x) \neq y\}}] \quad (14.9.1)$$

$$= \Pr(h(x) \neq y) \quad (14.9.2)$$

That is,  $L_{\mathcal{D}}(h)$  is the expected loss of  $h$  over the distribution  $\mathcal{D}$  for a 0-1 loss function. Hence the optimal classifier  $h^*$  from the hypothesis class would be

$$h^* = \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \quad (14.9.3)$$

Resigning to the fate that  $h^*$  is not attainable, as a ‘next-best’ substitute we may consider the empirical risk defined by

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}_{\{h(x_i) \neq y_i\}} \quad (14.9.4)$$

Note that this is effectively the same as taking the risk with respect to a distribution which is the empirical distribution. With this, we can perform empirical risk minimisation

$$\hat{h}_S = \operatorname{argmin}_{h \in \mathcal{H}} L_S(h) \quad (14.9.5)$$

In the probability approximately correct (PAC) framework, we consider learners (or learning algorithms) which are functions  $\mathcal{L} : \bigcup_{m=1}^{\infty} \{\mathcal{X} \times \mathcal{Y}\}^m \rightarrow \mathcal{H}$  capable of taking a sample of any size, and outputting a classifier from the hypothesis class  $\mathcal{H}$ . A hypothesis class is said to be *agnostic PAC learnable* if for all  $\varepsilon \in (0, 1)$ ,  $\delta \in (0, 1)$  and any distribution  $\mathcal{D}$ , there exists learner  $\mathcal{L}$  and an integer (i.e. dependent on  $\delta$  and  $\varepsilon$ ) such that for all  $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ , the learner returns a classifier  $\hat{h}$  that with probability at least  $1 - \delta$ ,

$$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \varepsilon \quad (14.9.6)$$

That is, the learner is ‘probably’ (with probability no less than  $1 - \delta$ ) ‘approximately correct’ (with allowable error up to  $\varepsilon$  from the optimal risk). The “agnostic” qualifier refers to the fact that nothing is assumed about the distribution  $\mathcal{D}$ . We usually treat  $\mathcal{H}$  to be reasonably restricted, i.e. a learner should not be allowed to simply memorise the sample because although this causes  $L_S(\hat{h}_S) = 0$ , this will lead to overfitting when tested with examples from  $\mathcal{D}$ .

#### Sample Complexity

The function  $m_{\mathcal{H}}(\varepsilon, \delta)$  is known as the sample complexity of learning the hypothesis class  $\mathcal{H}$ . That is, it is the minimum number of examples guaranteed to guarantee a probably approximately correct solution.

### Agnostic PAC Learnability of Finite Hypothesis Classes

**Theorem 14.1.** Suppose the hypothesis class  $\mathcal{H}$  is finite, i.e.  $|\mathcal{H}| < \infty$ . Then  $\mathcal{H}$  is agnostically PAC learnable.

*Proof.* We show that the choice of empirical risk minimisation learning algorithm is enough to satisfy the condition for agnostic PAC learnability. First, we define  $\varepsilon$ -representativeness as a property of a sample  $S$  which satisfies

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon \quad (14.9.7)$$

for all  $h \in \mathcal{H}$ . This formalises the notion that the empirical risk is close to the true risk. We can then show that if  $S$  is  $\frac{\varepsilon}{2}$ -representative, then the empirical risk minimiser  $\hat{h}_S$  satisfies for any  $h \in \mathcal{H}$ :

$$L_{\mathcal{D}}(\hat{h}_S) \leq L_S(\hat{h}_S) + \frac{\varepsilon}{2} \quad (14.9.8)$$

$$\leq L_S(h) + \frac{\varepsilon}{2} \quad (14.9.9)$$

$$\leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \quad (14.9.10)$$

$$= L_{\mathcal{D}}(h) + \varepsilon \quad (14.9.11)$$

where the first and third inequalities follow from the definition of  $\varepsilon$ -representativeness, and the second inequality is due to the characterisation/definition of the empirical risk minimiser. Hence, it holds for the optimal  $h^* = \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$  that:

$$L_{\mathcal{D}}(\hat{h}_S) \leq L_{\mathcal{D}}(h^*) + \varepsilon \quad (14.9.12)$$

Thus to show that  $\mathcal{H}$  is agnostic PAC learnable, it suffices to show that there exists a minimum number of examples  $m_{\mathcal{H}}(\varepsilon, \delta)$  such that  $S$  will be  $\frac{\varepsilon}{2}$ -representative with probability of at least  $1 - \delta$ . We first show a similar property of  $\mathcal{H}$  (called the uniform convergence property) that there exists a minimum number of examples such that  $S$  will be  $\varepsilon$ -representative with probability of at least  $1 - \delta$ . That is, we find the condition for

$$\Pr \left( \bigcap_{h \in \mathcal{H}} \{|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\} \right) \geq 1 - \delta \quad (14.9.13)$$

By DeMorgan's laws, an equivalent condition is

$$\Pr \left( \bigcup_{h \in \mathcal{H}} \{|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \right) \leq \delta \quad (14.9.14)$$

From the union bound (Boole's inequality):

$$\Pr \left( \bigcup_{h \in \mathcal{H}} \{|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \right) \leq \sum_{h \in \mathcal{H}} \Pr(|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon) \quad (14.9.15)$$

Recognise that  $L_S(h)$  is a sample mean of indicator random variables  $0 \leq \mathbb{I}_{\{h(x) \neq y\}} \leq 1$  with expectation  $\mathbb{E}[L_S(h)] = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbb{I}_{\{h(x) \neq y\}}] = L_{\mathcal{D}}(h)$ , so it satisfies the criteria to apply Hoeffding's inequality. Applying said inequality to each summand gives

$$\sum_{h \in \mathcal{H}} \Pr(|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon) \leq \sum_{h \in \mathcal{H}} 2 \exp(-2m\varepsilon^2) \quad (14.9.16)$$

$$= 2 |\mathcal{H}| \exp(-2m\epsilon^2) \quad (14.9.17)$$

We can then see that for  $m \geq \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\epsilon^2}$ , we will have

$$\Pr\left(\bigcup_{h \in \mathcal{H}} \{|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}\right) \leq \delta \quad (14.9.18)$$

Therefore by choosing  $m \geq \left\lceil \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\epsilon^2} \right\rceil$ , we have shown that  $\mathcal{H}$  satisfies the uniform convergence property. To find the condition for  $\frac{\epsilon}{2}$ -representativeness, simply repeat the above steps with  $\frac{\epsilon}{2}$ , which yields a sample complexity of

$$m_{\mathcal{H}}(\epsilon, \delta) = \left\lceil \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2(\epsilon/2)^2} \right\rceil \quad (14.9.19)$$

$$= \left\lceil \frac{2 \log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{\epsilon^2} \right\rceil \quad (14.9.20)$$

□

### Bias-Complexity Tradeoff

Different from the **bias-variance tradeoff** is the bias-complexity tradeoff, which describes the tradeoff of increasing the complexity of a hypothesis class. For an empirical risk minimiser  $\hat{h}_S = \operatorname{argmin}_{h \in \mathcal{H}} L_s(h)$ , the expected loss  $L_{\mathcal{D}}(\hat{h}_S) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbb{I}_{\{\hat{h}_S(x) \neq y\}}]$  can be decomposed into:

$$L_{\mathcal{D}}(\hat{h}_S) = \epsilon^* + \underbrace{\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - \epsilon^*}_{\epsilon} + \underbrace{\left(L_{\mathcal{D}}(\hat{h}_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)\right)}_{\epsilon} \quad (14.9.21)$$

where

- $\epsilon^*$  is the **Bayes error**, i.e. the best possible with any classifier (essentially only obtainable if  $\mathcal{D}$  were known).
- $\epsilon > 0$  is the **approximation error** for restricting ourselves to the hypothesis class  $\mathcal{H}$ , i.e. it is the best we can do in excess of the Bayes error.
- $\epsilon > 0$  is the **estimation error** that arises from having to learn from an empirical sample  $S$ , rather than the distribution  $\mathcal{D}$ .

Consider the effect of increasing the size of the hypothesis class  $|\mathcal{H}|$ , which is a way of increasing the complexity. By increasing the complexity, this may very well reduce the **inductive bias**  $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$  from restricting the hypothesis class (note that this meaning of bias is different from the **usual meaning** in a statistical sense). However, increasing the complexity can come at a tradeoff of increasing the estimation error  $\epsilon$ , given a fixed sample  $S$ . Recall the sample complexity for finite hypothesis classes:

$$m \geq \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\epsilon^2} \quad (14.9.22)$$

Rearranging gives the inequality required for  $\varepsilon$  as

$$\varepsilon \geq \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2m}} \quad (14.9.23)$$

What this shows is that for a fixed  $m$  and fixed  $\delta$ , increasing  $|\mathcal{H}|$  means that we also need to increase  $\varepsilon$  to say that the learner is  $\varepsilon$ -approximately  $\delta$ -probably correct. Hence the presumed tradeoff is that we can reduce  $\epsilon$  at the cost of increasing  $\varepsilon$ .

### No-Free-Lunch Theorem in Learning [179]

A No-Free-Lunch theorem in learning can be stated, which roughly says that for any binary classification algorithm  $\mathcal{A}$  (which for example is observed to perform well on a particular problem), there is going to be another distribution  $\mathcal{D}_*$  (i.e. another class of problem) for which a different learner will perform better than it.

**Theorem 14.2.** *For every  $\{0, 1\}$  binary classification learning algorithm  $\mathcal{A}$  applied to a training sample of size  $m < |\mathcal{X}|/2$  from the domain  $\mathcal{X} \times \{0, 1\}$ , there exists a distribution  $\mathcal{D}_*$  on support  $\mathcal{X} \times \{0, 1\}$  such that:*

1. *There is a classification function  $f^* : \mathcal{X} \rightarrow \{0, 1\}$  that achieves perfect loss  $L_{\mathcal{D}_*}(f^*) = 0$  with respect to the 0-1 loss function.*
2. *For i.i.d. samples of size  $m$  from  $\mathcal{D}_*$ , i.e.  $S \sim \mathcal{D}_*^m$ , we have that  $\Pr_{\mathcal{D}_*^m}(L_{\mathcal{D}_*}(\mathcal{A}(S)) \geq 1/8) \geq 1/7$ .*

*Proof.* Consider a subset  $\mathcal{C} \subset \mathcal{X}$  with  $|\mathcal{C}| = 2m < |\mathcal{X}|$ . There are  $T = 2^{2m}$  different functions from  $\mathcal{C}$  to  $\{0, 1\}$ , because each element of  $\mathcal{C}$  can either map to 0 or 1. Denote each of these functions by  $f_1, \dots, f_T$ . For each possible function, we can trivially construct a distribution  $\mathcal{D}_i$  so that  $f_i$  achieves perfect loss, by having the distribution be supported only over the examples that  $f_i$  would get correct. One such way is to make  $\mathcal{D}_i$  uniform over  $\mathcal{C}$  so that:

$$\Pr_{\mathcal{D}_i}(X = x, Y = y) = \begin{cases} 1/|\mathcal{C}|, & f_i(x) = y \\ 0, & \text{otherwise} \end{cases} \quad (14.9.24)$$

Then this gives  $L_{\mathcal{D}_i}(f_i) = 0$ . Now consider i.i.d. samples of size  $m$  from  $\mathcal{D}_i$ , which will be on the reduced domain  $\mathcal{C}$ . There are  $K = (2m)^m$  possible samples, which we denote by  $S_{i,1}, \dots, S_{i,K}$ . As per our construction of  $\mathcal{D}_i$ , the corresponding labels for each of the instances in these samples will be defined by  $f_i$ . Because samples are i.i.d. and  $\mathcal{D}_i$  is uniform, each of  $S_{i,1}, \dots, S_{i,K}$  has equal probability of being received, so the expected loss for a learning algorithm  $\mathcal{A}$  over the distribution of i.i.d. samples of size  $m$  from  $\mathcal{D}_i$  is calculated by:

$$\mathbb{E}_{S \sim \mathcal{D}_i^m}[L_{\mathcal{D}_i}(\mathcal{A}(S))] = \frac{1}{K} \sum_{j=1}^K L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j})) \quad (14.9.25)$$

There is going to be a function what yields the worst loss for  $\mathcal{A}$ , which is going to be worse than the average loss over all possible functions:

$$\max_{i \in \{1, \dots, T\}} \left\{ \frac{1}{K} \sum_{j=1}^K L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j})) \right\} \geq \frac{1}{T} \sum_{i=1}^T \left[ \frac{1}{K} \sum_{j=1}^K L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j})) \right] \quad (14.9.26)$$

$$= \frac{1}{K} \sum_{j=1}^K \left[ \frac{1}{T} \sum_{i=1}^T L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j})) \right] \quad (14.9.27)$$

where the second line is due to switching around the order of sums. Similarly, the average loss is worse than than the best loss (over all possible samples, this time):

$$\frac{1}{K} \sum_{j=1}^K \left[ \frac{1}{T} \sum_{i=1}^T L_{D_i}(\mathcal{A}(S_{i,j})) \right] \geq \min_{j \in \{1, \dots, K\}} \left\{ \frac{1}{T} \sum_{i=1}^T L_{D_i}(\mathcal{A}(S_{i,j})) \right\} \quad (14.9.28)$$

For a particular sample  $S_{i,j}$ , denote  $v_i, \dots, v_p$  to be the leftover instances in  $\mathcal{C}$  which do not appear in  $S_{i,j}$ . Because there may be some repeated instances in  $S_{i,j}$ , it will always be the case that  $p \geq m$ . Hence for any arbitrary function  $f : \mathcal{C} \rightarrow \{0, 1\}$  we can write using the definition for the 0-1 loss function for each  $i$ :

$$L_{D_i}(f) = \mathbb{E}_{D_i} [\mathbb{I}_{\{f(X) \neq f_i(X)\}}] \quad (14.9.29)$$

$$= \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \mathbb{I}_{\{f(x) \neq f_i(x)\}} \quad (14.9.30)$$

$$\geq \frac{1}{2p} \sum_{x \in \mathcal{C}} \mathbb{I}_{\{f(x) \neq f_i(x)\}} \quad (14.9.31)$$

$$\geq \frac{1}{2p} \sum_{r=1}^p \mathbb{I}_{\{f(v_r) \neq f_i(v_r)\}} \quad (14.9.32)$$

So substitute  $f$  with  $\mathcal{A}(S_{i,j})$  and put this into the average loss over all the distributions to give

$$\frac{1}{T} \sum_{i=1}^T L_{D_i}(\mathcal{A}(S_{i,j})) \geq \frac{1}{T} \sum_{i=1}^T \left[ \frac{1}{2p} \sum_{r=1}^p \mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} \right] \quad (14.9.33)$$

$$= \frac{1}{2p} \sum_{r=1}^p \left[ \frac{1}{T} \sum_{i=1}^T \mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} \right] \quad (14.9.34)$$

Again lower bounding the average by the minimum yields

$$\frac{1}{T} \sum_{i=1}^T L_{D_i}(\mathcal{A}(S_{i,j})) \geq \frac{1}{2} \min_{r \in \{1, \dots, p\}} \left\{ \frac{1}{T} \sum_{i=1}^T \mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} \right\} \quad (14.9.35)$$

Now for any particular  $r \in \{1, \dots, p\}$ , we are able to split all the functions  $f_1, \dots, f_T$  into two halves (or  $T/2$  pairs), where all the functions in one half map  $v_r$  to 0, while all the functions in the other half map  $v_r$  to 1. Since  $v_r$  is not in the sample, then for each pair  $i, i'$  the correct labels of  $S_{i,j}$  will be the same as the labels of  $S_{i',j}$  (so the learned function will be identical). However if the learner correctly classifies instance  $v_r$  over distribution  $D_i$ , it must make a mistake over distribution  $D_{i'}$ , and vice-versa. This means that the loss for instance  $v_r$  satisfies

$$\mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} + \mathbb{I}_{\{\mathcal{A}(S_{i',j})(v_r) \neq f_{i'}(v_r)\}} = 1 \quad (14.9.36)$$

Taking the sum over all pairs, we see that

$$\sum_{i=1}^T \mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} = \frac{T}{2} \quad (14.9.37)$$

$$\frac{1}{T} \sum_{i=1}^T \mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} = \frac{1}{2} \quad (14.9.38)$$

Putting all the arguments above together, we obtain:

$$\max_{i \in \{1, \dots, T\}} \left\{ \mathbb{E}_{S \sim \mathcal{D}_i^m} [L_{D_i}(\mathcal{A}(S))] \right\} = \max_{i \in \{1, \dots, T\}} \left\{ \frac{1}{K} \sum_{j=1}^K L_{D_i}(\mathcal{A}(S_{i,j})) \right\} \quad (14.9.39)$$

$$\geq \min_{j \in \{1, \dots, K\}} \left\{ \frac{1}{T} \sum_{i=1}^T L_{D_i}(\mathcal{A}(S_{i,j})) \right\} \quad (14.9.40)$$

$$\geq \min_{j \in \{1, \dots, K\}} \left\{ \frac{1}{2} \min_{r \in \{1, \dots, p\}} \left\{ \frac{1}{T} \sum_{i=1}^T \mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} \right\} \right\} \quad (14.9.41)$$

$$\geq \min_{j \in \{1, \dots, K\}} \left\{ \frac{1}{2} \min_{r \in \{1, \dots, p\}} \left\{ \frac{1}{2} \right\} \right\} \quad (14.9.42)$$

since we showed  $\frac{1}{T} \sum_{i=1}^T \mathbb{I}_{\{\mathcal{A}(S_{i,j})(v_r) \neq f_i(v_r)\}} = \frac{1}{2}$  for arbitrary  $r$ . Hence we arrive at

$$\max_{i \in \{1, \dots, T\}} \{ \mathbb{E}_{S \sim \mathcal{D}_i^m} [L_{D_i}(\mathcal{A}(S))] \} \geq \frac{1}{4} \quad (14.9.43)$$

which implies there exists a function  $f^*$  and distribution  $\mathcal{D}_*$  for every learner  $\mathcal{A}$  such that  $L_{D_*}(f^*) = 0$  and

$$\mathbb{E}_{S \sim \mathcal{D}_*^m} [L_{D_*}(\mathcal{A}(S))] \geq \frac{1}{4} \quad (14.9.44)$$

The last steps are to apply the reverse Markov inequality  $\Pr(Z \geq c) \geq \frac{\mathbb{E}[X] - c}{b - c}$  applicable for any random variable upper bounded by  $b$ . Since  $L_{D_*}(\mathcal{A}(S))$  takes on values in  $[0, 1]$ , we can set  $c = 1/8$  and  $b = 1$  to get

$$\Pr_{\mathcal{D}_*^m} (L_{D_*}(\mathcal{A}(S)) \geq 1/8) \geq \frac{\mathbb{E}[X] - 1/8}{1 - 1/8} \quad (14.9.45)$$

$$\geq \frac{1/4 - 1/8}{1 - 1/8} \quad (14.9.46)$$

$$= \frac{1/8}{7/8} \quad (14.9.47)$$

$$= \frac{1}{7} \quad (14.9.48)$$

□

Intuitively, what this theorem shows is that if a learning algorithm has seen fewer than half of all the possible training examples from a domain, it is possible to construct an ‘adversarial’ distribution such that the learning algorithm will perform well on the training set, but poorly when generalising.

### 14.9.2 Vapnik-Chervonenkis Dimension

Let  $\mathcal{H}$  be a hypothesis class of functions from the feature space  $\mathcal{X}$  to  $\{0, 1\}$ . Let

$$\mathcal{C} = \{c_1, \dots, c_m\} \quad (14.9.49)$$

be a finite subset  $\mathcal{C} \subset \mathcal{X}$ , which we can think of as our training features. We define  $\mathcal{H}_C$  to be the *restriction* of  $\mathcal{H}$  to  $\mathcal{C}$ . That is,  $\mathcal{H}_C$  is the set of all the possible functions from  $\mathcal{H}$  but restricted on the domain  $\mathcal{C}$ . Then each function from  $\mathcal{H}_C$  is a member of  $\{0, 1\}^m$ . All the elements of  $\{0, 1\}^m$  (each of which is a function from  $\mathcal{C}$  to  $\{0, 1\}$ ) may or may not be represented in  $\mathcal{H}_C$ . If  $\mathcal{H}_C$  does contain all functions from  $\mathcal{C}$  to  $\{0, 1\}$ , then we say that  $\mathcal{H}$  *shatters*  $\mathcal{C}$ , i.e. if

$$|\mathcal{H}_C| = 2^m \quad (14.9.50)$$

The Vapnik-Chervonenkis (VC) dimension of a hypothesis class  $\mathcal{H}$  is the largest size of a set  $\mathcal{C} \subset \mathcal{X}$  such that  $\mathcal{H}$  shatters  $\mathcal{C}$ . This VC dimension is denoted

$$\text{VCdim}(\mathcal{H}) = d \quad (14.9.51)$$

If  $\text{VCdim}(\mathcal{H}) = d$ , then this implies that for each  $m \leq d$  there exists a subset  $\mathcal{C} \subset \mathcal{X}$  with  $|\mathcal{C}| = m$  such that  $\mathcal{H}$  shatters  $\mathcal{C}$ , i.e.  $|\mathcal{H}_C| = 2^m$ . This can be trivially shown by construction since any subset of size  $m$  out of the  $\mathcal{C}$  with  $|\mathcal{C}| = d$  which  $\mathcal{H}$  shatters  $\mathcal{C}$  will also be shattered by  $\mathcal{H}$ . However note that this does not necessarily mean that all subsets  $\mathcal{C} \subset \mathcal{X}$  of size  $d$  or less must be shattered by  $\mathcal{H}$ .

**Theorem 14.3.** *If  $\text{VCdim}(\mathcal{H}) = \infty$ , then  $\mathcal{H}$  is not agnostic PAC learnable.*

*Proof.* Recall that  $\mathcal{H}$  shatters  $\mathcal{C}$  if  $|\mathcal{H}_C| = 2^m$ . Combining this with the No Free Lunch theorem in learning, we have the following. For a hypothesis class  $\mathcal{H}$ , suppose we have  $m$  training examples and there exists a subset  $\mathcal{C} \subset \mathcal{X}$  with  $|\mathcal{C}| = 2m$  that is shattered by  $\mathcal{H}$ . This means it is possible to ‘build’ any function from  $\mathcal{C}$  to  $\{0, 1\}$  from the hypothesis class  $\mathcal{H}$ . For any learning algorithm  $\mathcal{A}$  applied to these  $m$  training examples, then from the No Free Lunch theorem there exists a distribution  $\mathcal{D}_*$  on support  $\mathcal{X} \times \{0, 1\}$  and a function  $h^* \in \mathcal{H}$  (constructed in the same way as in the No Free Lunch theorem proof, which is allowable since  $\mathcal{H}_C$  contains any function from  $\mathcal{C}$  to  $\{0, 1\}$ ) such that  $\mathcal{L}_{\mathcal{D}_*}(h^*) = 0$  but for samples  $S \sim \mathcal{D}_*^m$ , we have that

$$\Pr_{\mathcal{D}_*^m}(L_{\mathcal{D}_*}(\mathcal{A}(S)) \geq 1/8) \geq 1/7 \quad (14.9.52)$$

Now since  $\mathcal{H}$  has infinite VC dimension, then for any finite training set size  $m$  there exists a subset  $\mathcal{C} \subset \mathcal{X}$  of size  $2m$  that is shattered by  $\mathcal{H}$ . This provides a counterexample, since for the definition of agnostic PAC learnability we require for any distribution  $\mathcal{D}$  that that exists a large enough  $m$  such that

$$\Pr_{\mathcal{D}^m}(L_{\mathcal{D}}(\mathcal{A}(S)) \leq \mathcal{L}_{\mathcal{D}}(h^*) + \varepsilon) \geq 1 - \delta \quad (14.9.53)$$

for all  $\varepsilon \in (0, 1]$ ,  $\delta \in (0, 1)$  and for at least one learner  $\mathcal{A}$ . But from above we know that there is a distribution and a function such that  $\mathcal{L}_{\mathcal{D}_*}(h^*) = 0$  and

$$\Pr_{\mathcal{D}_*^m}(L_{\mathcal{D}_*}(\mathcal{A}(S)) < 1/8) \leq 1 - 1/7 \quad (14.9.54)$$

for all learners  $\mathcal{A}$  and for any sample size  $m$ . Putting any  $\varepsilon < 1/8$  and any  $\delta < 1/7$  gives

$$\Pr_{\mathcal{D}_*^m}(L_{\mathcal{D}_*}(\mathcal{A}(S)) \leq \mathcal{L}_{\mathcal{D}_*}(h^*) + \varepsilon) \leq \Pr_{\mathcal{D}_*^m}\left(L_{\mathcal{D}_*}(\mathcal{A}(S)) < \mathcal{L}_{\mathcal{D}_*}(h^*) + \frac{1}{8}\right) \quad (14.9.55)$$

$$\leq 1 - \frac{1}{7} \quad (14.9.56)$$

$$< 1 - \delta \quad (14.9.57)$$

which completes the counterexample.  $\square$

Hence, the VC dimension is an appropriate characterisation of agnostic PAC learnability for a hypothesis class.

### 14.9.3 Rademacher Complexity [140, 179]

Let  $\mathcal{H}$  be a hypothesis class,  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  be a domain and  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  be a class of loss function. For each  $h \in \mathcal{H}$ , we have a mapping from  $z = (x, y) \in \mathcal{Z}$  to  $\ell(h(x), y)$ , which is the loss of example  $z$  on function  $h$ . Define  $\mathcal{G}$  to be the collection of these mappings  $g : \mathcal{Z} \rightarrow \mathbb{R}$  over all  $h \in \mathcal{H}$ , i.e.

$$\mathcal{G} = \{g : h \in \mathcal{H}\} \quad (14.9.58)$$

We can interpret  $\mathcal{G}$  as the family of all loss functions from the class  $\ell$  that are possible under hypothesis class  $\mathcal{H}$ . For some  $g \in \mathcal{G}$ , the expected loss over distribution  $\mathcal{D}$  on support  $\mathcal{Z}$  is given by

$$L_{\mathcal{D}}(g) = \mathbb{E}_{z \sim \mathcal{D}}[g(z)] \quad (14.9.59)$$

while the empirical loss over sample  $S = (z_1, \dots, z_m) \sim \mathcal{D}^m$  is given by

$$L_S(g) = \frac{1}{m} \sum_{i=1}^m g(z_i) \quad (14.9.60)$$

One way to define how ‘representative’  $S$  is of  $\mathcal{D}$  with respect to the choice of hypothesis class  $\mathcal{H}$  and class of loss function  $\ell$  is by taking the supremum of the error in expected loss over  $\mathcal{G}$ :

$$\text{Rep}_{S,\mathcal{D}}(\mathcal{G}) = \sup_{g \in \mathcal{G}} \{L_{\mathcal{D}}(g) - L_S(g)\} \quad (14.9.61)$$

where a smaller  $\text{Rep}_{S,\mathcal{D}}(\mathcal{G})$  is ‘more representative’. Suppose  $\mathcal{D}$  is not available and we wish to estimate  $\text{Rep}_{S,\mathcal{D}}(\mathcal{G})$  using only sample  $S$ . Then one approach is to split  $S$  into disjoint validation set  $S_1$  and training set  $S_2$ . Suppose  $S$  is of even size, so that it can be split such that  $S_1$  and  $S_2$  are of the same size. Then this ‘empirical representativeness’ is given by replacing  $\mathcal{D}$  with the validation set  $S_1$ :

$$\widehat{\text{Rep}}_S(\mathcal{G}) = \sup_{g \in \mathcal{G}} \{L_{S_1}(g) - L_{S_2}(g)\} \quad (14.9.62)$$

If we define the vector  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_m)$  such that  $\sigma_i = 1$  if  $z_i$  is in  $S_1$  and  $\sigma_i = -1$  otherwise, then the empirical representativeness can alternatively be written as

$$\widehat{\text{Rep}}_S(\mathcal{G}) = \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m/2} \sum_{i:z_i \in S_1} g(z_i) - \frac{1}{m/2} \sum_{i:z_i \in S_2} g(z_i) \right\} \quad (14.9.63)$$

$$= \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m/2} \sum_{i:z_i \in S_1} \sigma_i g(z_i) + \frac{1}{m/2} \sum_{i:z_i \in S_2} \sigma_i g(z_i) \right\} \quad (14.9.64)$$

$$= \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i) \right\} \quad (14.9.65)$$

This leads us to the *empirical Rademacher complexity* of  $\mathcal{G}$  with respect to  $S$ , where we now treat  $\boldsymbol{\sigma}$  as a random vector with independent Rademacher-distributed components (i.e. uniformly distributed on support  $\{-1, 1\}$ ), and then take the expectation of the empirical representativeness:

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[ \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i) \right\} \right] \quad (14.9.66)$$

Let  $\mathbf{g}_S$  denote the vector  $(g(z_1), \dots, g(z_m))$ , then the empirical Rademacher complexity becomes

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{m} \sup_{g \in \mathcal{G}} \mathbb{E}_{\boldsymbol{\sigma}} [\boldsymbol{\sigma} \cdot \mathbf{g}_S] \quad (14.9.67)$$

This way, the empirical Rademacher complexity can be interpreted intuitively as follows. Regard  $\boldsymbol{\sigma}$  as (zero-mean) random noise. Then  $\mathbb{E}_{\boldsymbol{\sigma}} [\boldsymbol{\sigma} \cdot \mathbf{g}_S]$  corresponds to how well the ‘worst-case’  $g$  from the class  $\mathcal{G}$  correlates with random noise. Thus a more complex/richer class  $\mathcal{G}$  (induced by the choice of classes  $\mathcal{H}$  and  $\ell$ ) will lead to better correlation with random noise. Another interpretation is that richer classes of  $\mathcal{G}$  will be estimated to be less representative, on average. The *Rademacher complexity* of class  $\mathcal{G}$  for sample size  $m$  from distribution  $\mathcal{D}$  then characterises the overall richness of  $\mathcal{G}$  by taking the expectation of the empirical Rademacher complexity over all samples of size  $m$  drawn from  $\mathcal{D}$ :

$$\mathfrak{R}_{\mathcal{D},m}(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^m} [\widehat{\mathfrak{R}}_S(\mathcal{G})] \quad (14.9.68)$$

## Generalisation Bounds with Rademacher Complexity

Let  $\mathcal{G}$  be a class of functions from  $\mathcal{Z}$  to  $[0, 1]$  (induced by a choice of hypothesis class  $\mathcal{H}$  and appropriate loss function  $\ell$ , for example a 0-1 loss for binary classification), which Rademacher complexity  $\mathfrak{R}_{\mathcal{D},m}(\mathcal{G})$  with respect to the distribution  $\mathcal{D}$  over  $\mathcal{Z}$ . Let  $S$  be an i.i.d. sample of size  $m$  from  $\mathcal{D}$ .

**Theorem 14.4** ([140]). *For all  $g \in \mathcal{G}$  (i.e. for all  $h \in \mathcal{H}$  with  $g(z) = \ell(y, h(x))$ ), with probability at least  $1 - \delta$ ,*

$$L_{\mathcal{D}}(g) - L_S(g) \leq 2\mathfrak{R}_{\mathcal{D},m}(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2m}} \quad (14.9.69)$$

Recall that  $L_{\mathcal{D}}(g) = \mathbb{E}_{z \sim \mathcal{D}}[g(z)]$  and  $L_S(g) = \frac{1}{m} \sum_{i=1}^m g(z_i)$ , so these can be alternatively written as

$$\mathbb{E}_{z \sim \mathcal{D}}[g(z)] \leq \frac{1}{m} \sum_{i=1}^m g(z_i) + 2\mathfrak{R}_{\mathcal{D},m}(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2m}} \quad (14.9.70)$$

Note that  $L_{\mathcal{D}}(g) - L_S(g)$  can be interpreted in terms of generalisation error as how much the expected loss on unseen examples exceeds the training loss, and this can be upper bounded using the Rademacher complexity regardless of the learned function  $h$ .

*Proof.* We introduce a function of the sample  $S$  as

$$f(S) = f(z_1, \dots, z_m) \quad (14.9.71)$$

$$= \sup_{g \in \mathcal{G}} \{L_{\mathcal{D}}(g) - L_S(g)\} \quad (14.9.72)$$

Denote  $\tilde{S}$  as a perturbed version of sample  $S$ , where only a single example is changed. We seek to bound the distance

$$|f(S) - f(\tilde{S})| = \left| \sup_{g \in \mathcal{G}} \{L_{\mathcal{D}}(g) - L_S(g)\} - \sup_{g \in \mathcal{G}} \{L_{\mathcal{D}}(g) - L_{\tilde{S}}(g)\} \right| \quad (14.9.73)$$

By regarding the supremum operator as the infinity-norm over a function space, we have

$$|f(S) - f(\tilde{S})| = \|L_{\mathcal{D}}(g) - L_S(g)\|_{\infty} - \|L_{\mathcal{D}}(g) - L_{\tilde{S}}(g)\|_{\infty} \quad (14.9.74)$$

$$\leq \|(L_{\mathcal{D}}(g) - L_S(g)) - (L_{\mathcal{D}}(g) - L_{\tilde{S}}(g))\|_{\infty} \quad (14.9.75)$$

$$= \sup_{g \in \mathcal{G}} \{L_{\tilde{S}}(g) - L_S(g)\} \quad (14.9.76)$$

where we have used the reverse triangle inequality to obtain the upper bound. Now note that since  $g$  maps to  $[0, 1]$ , and  $L_S(g) = \frac{1}{m} \sum_{i=1}^m g(z_i)$  is the empirical loss (i.e. with  $m$  in the denominator), then the deviation in  $f(S)$  is upper bounded by

$$|f(S) - f(\tilde{S})| \leq \frac{1}{m} \quad (14.9.77)$$

Thus, this satisfies the conditions of McDiarmid's inequality, which we can apply to the function  $f(S)$  of the independent sample  $S$ . Using the upper tail bound at  $t = \sqrt{\log(1/\delta)/(2m)}$ , this gives

$$\Pr \left( f(S) - \mathbb{E}_S[f(S)] > \sqrt{\frac{\log(1/\delta)}{2m}} \right) \leq \exp \left[ - \frac{2 \left( \sqrt{\log(1/\delta)/(2m)} \right)^2}{\sum_{i=1}^m (1/m)^2} \right] \quad (14.9.78)$$

$$= \exp\left(-\frac{\log(1/\delta)/m}{1/m}\right) \quad (14.9.79)$$

$$= \exp(\log \delta) \quad (14.9.80)$$

$$= \delta \quad (14.9.81)$$

Hence

$$\Pr\left(f(S) - \mathbb{E}_S[f(S)] \leq \sqrt{\frac{\log(1/\delta)}{2m}}\right) \geq 1 - \delta \quad (14.9.82)$$

Let  $S' = (z'_1, \dots, z'_m)$  be another sample of size  $m$  that is i.i.d. with  $S$ . Then we can write  $L_{\mathcal{D}}(g) = \mathbb{E}_{S'}[L_{S'}(g)]$  because

$$\mathbb{E}_{S'}[L_{S'}(g)] = \mathbb{E}_{S'}\left[\frac{1}{m} \sum_{i=1}^m g(z'_i)\right] \quad (14.9.83)$$

$$= \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{S'}[g(z'_i)] \quad (14.9.84)$$

$$= \mathbb{E}_{z \sim \mathcal{D}}[g(z)] \quad (14.9.85)$$

$$= L_{\mathcal{D}}(g) \quad (14.9.86)$$

So using the definition of  $f(S)$ :

$$\mathbb{E}_S[f(S)] = \mathbb{E}_S\left[\sup_{g \in \mathcal{G}}\{L_{\mathcal{D}}(g) - L_S(g)\}\right] \quad (14.9.87)$$

$$= \mathbb{E}_S\left[\sup_{g \in \mathcal{G}}\{\mathbb{E}_{S'}[L_{S'}(g)] - L_S(g)\}\right] \quad (14.9.88)$$

$$= \mathbb{E}_S\left[\sup_{g \in \mathcal{G}}\{\mathbb{E}_{S'}[L_{S'}(g) - L_S(g)]\}\right] \quad (14.9.89)$$

$$\leq \mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{g \in \mathcal{G}}\{L_{S'}(g) - L_S(g)\}\right]\right] \quad (14.9.90)$$

$$= \mathbb{E}_{S,S'}\left[\sup_{g \in \mathcal{G}}\{L_{S'}(g) - L_S(g)\}\right] \quad (14.9.91)$$

$$= \mathbb{E}_{S,S'}\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{m} \sum_{i=1}^m (g(z'_i) - g(z_i))\right\}\right] \quad (14.9.92)$$

where the upper bound arises because the supremum is subadditive (i.e.  $\sup\{a + b\} \leq \sup a + \sup b$ ), and we can treat the expectation operator like a sum. Due to identicality and independence of  $S$  and  $S'$ , we see that swapping the signs in the sum does not matter:

$$\mathbb{E}_{S,S'}\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{m} \sum_{i=1}^m (g(z'_i) - g(z_i))\right\}\right] = \mathbb{E}_{S,S'}\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{m} \sum_{i=1}^m (g(z_i) - g(z'_i))\right\}\right] \quad (14.9.93)$$

In fact, this will hold even if signs are arbitrarily swapped (equivalent to swapping arbitrary corresponding elements between  $S$  and  $S'$ , which just introduces another pair of samples with the same distribution as  $S$  and  $S'$ ). Because of this property, we can show the following. If we introduce  $m$  i.i.d. Rademacher random variables  $\sigma = (\sigma_1, \dots, \sigma_m)$ , this yields

$$\mathbb{E}_{\sigma, S, S'}\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{m} \sum_{i=1}^m \sigma_i (g(z'_i) - g(z_i))\right\}\right] = \mathbb{E}_{S, S'}\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{m} \sum_{i=1}^m (g(z'_i) - g(z_i))\right\}\right] \quad (14.9.94)$$

Again using the subadditivity of the supremum, our upper bound becomes

$$\mathbb{E}_S [f(S)] \leq \mathbb{E}_{\sigma, S, S'} \left[ \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m \sigma_i g(z'_i) \right\} + \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m (-\sigma_i) g(z_i) \right\} \right] \quad (14.9.95)$$

$$\leq \mathbb{E}_{\sigma, S'} \left[ \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m \sigma_i g(z'_i) \right\} \right] + \mathbb{E}_{\sigma, S} \left[ \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m (-\sigma_i) g(z_i) \right\} \right] \quad (14.9.96)$$

Since each  $\sigma_i$  and  $-\sigma_i$  are identically distributed due to the symmetry of the Rademacher distribution, then

$$\mathbb{E}_S [f(S)] \leq 2\mathbb{E}_{\sigma, S} \left[ \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i) \right\} \right] \quad (14.9.97)$$

$$= 2\mathbb{E}_S \left[ \mathbb{E}_{\sigma} \left[ \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i) \right\} \right] \right] \quad (14.9.98)$$

$$= 2\mathbb{E}_S [\widehat{\mathfrak{R}}_S(\mathcal{G})] \quad (14.9.99)$$

$$= 2\mathfrak{R}_{D,m}(\mathcal{G}) \quad (14.9.100)$$

where we have recognised and utilised the definitions of Rademacher complexity. Lastly it follows that

$$\Pr \left( \sup_{g \in \mathcal{G}} \{L_D(g) - L_S(g)\} \leq 2\mathfrak{R}_{D,m}(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2m}} \right) \geq 1 - \delta \quad (14.9.101)$$

$$\Pr \left( L_D(g) - L_S(g) \leq 2\mathfrak{R}_{D,m}(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2m}} \right) \geq 1 - \delta \quad (14.9.102)$$

for all  $g \in \mathcal{G}$ . □

**Corollary 14.1.** *Moreover for all  $g \in \mathcal{G}$ :*

$$L_D(g) - L_S(g) \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \quad (14.9.103)$$

or equivalently

$$\mathbb{E}_{z \sim D}[g(z)] \leq \frac{1}{m} \sum_{i=1}^m g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \quad (14.9.104)$$

*Proof.* Re-introduce the perturbed sample  $\tilde{S}$  same as above. The empirical Rademacher complexity (which is a function of  $S$ ) has bounded deviations given by

$$|\widehat{\mathfrak{R}}_S(\mathcal{G}) - \widehat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G})| \leq \frac{1}{m} \quad (14.9.105)$$

using similar reasoning as above, since the image of  $g$  is  $[0, 1]$  and  $\sigma_i$  can take on either 1 or  $-1$ . As above, we can derive the bound (using  $\delta/2$  instead of  $\delta$ ):

$$\Pr \left( f(S) - \mathbb{E}_S[f(S)] > \sqrt{\frac{\log(2/\delta)}{2m}} \right) \leq \frac{\delta}{2} \quad (14.9.106)$$

Then applying McDiarmid's inequality to the function  $\widehat{\mathfrak{R}}_S(\mathcal{G})$  with  $\mathbb{E}_S[\widehat{\mathfrak{R}}_S(\mathcal{G})] = \mathfrak{R}_{D,m}(\mathcal{G})$  (this time using the lower tail bound at the tail value  $t = \sqrt{\log(2/\delta)/(2m)}$ ), we obtain

$$\Pr \left( \widehat{\mathfrak{R}}_S(\mathcal{G}) - \mathfrak{R}_{D,m}(\mathcal{G}) < -\sqrt{\frac{\log(2/\delta)}{2m}} \right) \leq \frac{\delta}{2} \quad (14.9.107)$$

or equivalently

$$\Pr \left( 2\widehat{\mathfrak{R}}_S(\mathcal{G}) - 2\mathfrak{R}_{\mathcal{D},m}(\mathcal{G}) < -2\sqrt{\frac{\log(2/\delta)}{2m}} \right) \leq \frac{\delta}{2} \quad (14.9.108)$$

Combining these bounds using Boole's inequality, this gives

$$\Pr \left( \left\{ f(S) - \mathbb{E}_S[f(S)] > \sqrt{\frac{\log(2/\delta)}{2m}} \right\} \cup \left\{ 2\widehat{\mathfrak{R}}_S(\mathcal{G}) - 2\mathfrak{R}_{\mathcal{D},m}(\mathcal{G}) < -2\sqrt{\frac{\log(2/\delta)}{2m}} \right\} \right) \leq \delta \quad (14.9.109)$$

Taking complements and using DeMorgan's laws:

$$\Pr \left( \left\{ f(S) - \mathbb{E}_S[f(S)] \leq \sqrt{\frac{\log(2/\delta)}{2m}} \right\} \cap \left\{ 2\widehat{\mathfrak{R}}_S(\mathcal{G}) - 2\mathfrak{R}_{\mathcal{D},m}(\mathcal{G}) \geq -2\sqrt{\frac{\log(2/\delta)}{2m}} \right\} \right) \geq 1 - \delta \quad (14.9.110)$$

which can be rearranged as

$$\Pr \left( \left\{ f(S) \leq \mathbb{E}_S[f(S)] + \sqrt{\frac{\log(2/\delta)}{2m}} \right\} \cap \left\{ 2\mathfrak{R}_{\mathcal{D},m}(\mathcal{G}) \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 2\sqrt{\frac{\log(2/\delta)}{2m}} \right\} \right) \geq 1 - \delta \quad (14.9.111)$$

Using the previously derived bound  $\mathbb{E}_S[f(S)] \leq 2\mathfrak{R}_{\mathcal{D},m}(\mathcal{G})$  and the definition of  $f(S)$ , this implies

$$\Pr \left( \sup_{g \in \mathcal{G}} \{L_{\mathcal{D}}(g) - L_S(g)\} \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \right) \geq 1 - \delta \quad (14.9.112)$$

and in the same way as before

$$\Pr \left( L_{\mathcal{D}}(g) - L_S(g) \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \right) \geq 1 - \delta \quad (14.9.113)$$

for all  $g \in \mathcal{G}$ . □

#### 14.9.4 Growth Function [140, 179]

The *growth function* of a hypothesis class  $\mathcal{H}$  of functions from  $\mathcal{X}$  to  $\{0, 1\}$  is a function in argument  $m$  which gives the maximum number of functions in the restriction of  $\mathcal{H}$  on  $\mathcal{C}$  (which recall is denoted  $\mathcal{H}_{\mathcal{C}}$  from the definition of the VC dimension), with  $|\mathcal{C}| = m$ . The growth function  $\mathcal{G}_{\mathcal{H}} : \mathbb{N} \rightarrow \mathbb{N}$  can be defined by:

$$\mathcal{G}_{\mathcal{H}}(m) = \max_{\{\mathcal{C} \subset \mathcal{X} : |\mathcal{C}|=m\}} |\mathcal{H}_{\mathcal{C}}| \quad (14.9.114)$$

The growth function provides a combinatorial measure of the richness of class  $\mathcal{H}$  in terms of a training set of size  $m$ . We can also uncover the following relations between the growth function and the VC dimension.

- The definition of the VC dimension can be written in terms of the growth function by

$$\text{VCdim}(\mathcal{H}) = \max_{\{m : \mathcal{G}_{\mathcal{H}}(m)=2^m\}} m \quad (14.9.115)$$

- If  $\mathcal{G}_{\mathcal{H}}(m) = 2^m$ , this implies there exists a  $\mathcal{C} \subset \mathcal{X}$  with  $|\mathcal{C}| = m$  such that  $|\mathcal{H}_{\mathcal{C}}| = 2^m$ . Then from the definition of the VC dimension, it follows that the VC dimension is at least  $m$ , i.e.  $\text{VCdim}(\mathcal{H}) \geq m$ .
- Recall that if  $\text{VCdim}(\mathcal{H}) = d$ , then for any  $m \leq d$  there exists a subset  $\mathcal{C} \subset \mathcal{X}$  of size  $|\mathcal{C}| = m$  such that  $|\mathcal{H}_{\mathcal{C}}| = 2^m$ . Hence this also means that  $\mathcal{G}_{\mathcal{H}}(m) = 2^m$  for all  $m \leq d$ .

**Sauer's Lemma**

Sauer's lemma bounds the growth function for hypotheses classes with finite VC dimension.

**Lemma 14.1.** *Let  $\mathcal{H}$  be a hypothesis class with  $\text{VCdim}(\mathcal{H}) = d < \infty$ . Then for all  $m \in \mathbb{N}$ :*

$$\mathcal{G}_{\mathcal{H}}(m) \leq \sum_{i=1}^d \binom{m}{i} \quad (14.9.116)$$

*Proof.* □

**Generalisation Bounds with Growth Function****Generalisation Bounds with VC Dimension****14.9.5 Fundamental Theorem of Statistical Learning [179]**

It was established that an infinite VC dimension hypothesis class is not agnostic PAC learnable, but the converse is also true. Any hypothesis class  $\mathcal{H}$  with finite VC dimension  $\text{VCdim}(\mathcal{H}) = d < \infty$  is agnostic PAC learnable. In particular, the sample complexity  $m_{\mathcal{H}}(\varepsilon, \delta)$  is bounded by

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d \log(d/\varepsilon) + \log(1/\delta)}{\varepsilon^2} \quad (14.9.117)$$

where  $C_1, C_2$  are absolute constants. This is known as the Fundamental Theorem of Statistical Learning.

*Proof.* □

**14.9.6 Regression Generalisation Bounds [81]**

## Chapter 15

# Statistical Signal Processing

### 15.1 Random Signals and Systems

#### 15.1.1 Random Linear Time Invariant Systems

Consider a linear time invariant (LTI) system in discrete time. For a sequence of inputs  $u(k)$ , there is a corresponding sequence of outputs  $y(k)$ . The linearity property means that for a superposition of inputs  $u_1(k) + u_2(k)$ , the output is also a superposition  $y_1(k) + y_2(k)$ . We can construct any input sequence with a superposition of Kronecker deltas

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad (15.1.1)$$

like so for  $\{\dots, u(0), u(1), u(2), \dots\} = \{\dots, a_0, a_1, a_2, \dots\}$ :

$$u(k) = \sum_{j=-\infty}^{\infty} a_k \delta_{jk} = a_k \quad (15.1.2)$$

An impulse response  $h(k)$  is defined as the output  $y(k)$  for a input of a unit pulse at  $k = 0$ , i.e.  $u(k) = \delta_{k,0}$ . Then for a ‘lagged’ unit pulse input  $u(k) = \delta_{k,1}$ , the output will be  $h(k - 1)$  because it is just the impulse response lagged by 1. By superposition, if we apply the input  $u(k) = \delta_{k,0} + \delta_{k,1}$ , the output should be  $y(k) = h(k) + h(k - 1)$ . Generalising, if the input is  $u(k) = \sum_{j=-\infty}^{\infty} a_k \delta_{jk}$ , then the output should be

$$y(k) = \dots + a_0 h(k) + a_1 h(k - 1) + \dots \quad (15.1.3)$$

$$= \dots + u(0) h(k) + u(1) h(k - 1) + \dots \quad (15.1.4)$$

$$= \sum_{j=-\infty}^{\infty} h(k - j) u(j) \quad (15.1.5)$$

which is the convolution sum. Note that we can rewrite this as

$$y(k) = \sum_{j=-\infty}^{\infty} h(j) u(k - j) \quad (15.1.6)$$

because this would only change the sequence of summation. It follows that if  $U(k)$  is a random sequence input into an LTI system, the random sequence output  $Y(k)$  would be given by

$$Y(k) = \sum_{j=-\infty}^{\infty} h(j) U(k - j) \quad (15.1.7)$$

Now if we considered a continuous time LTI system with impulse response  $h(t)$ , then by similar arguments (except considering Dirac deltas instead of Kronecker deltas and integrals instead of sums), for a stochastic process input  $U(t)$  the stochastic process output  $Y(t)$  is given by the convolution integral

$$Y(t) = \int_{-\infty}^{\infty} h(\tau) U(t - \tau) d\tau \quad (15.1.8)$$

### Sampled Systems with Random Signals [130]

Suppose we have a sampled system, where we observe output  $Y(t)$  at time instants separated by interval  $T$ , and thus define the random sequence  $Y_k = Y(kT)$  for  $k = \dots, 0, 1, \dots$ . Also suppose the random sequence of inputs  $U_k$  is held constant over the interval  $T$  such that

$$U(t) = U_k \quad (15.1.9)$$

for  $kT \leq t < (k+1)T$ . Then for an LTI system defined by  $g(t)$ , the output at instant  $t = kT$  is

$$Y(kT) = \int_{-\infty}^{\infty} h(\tau) U(kT - \tau) d\tau \quad (15.1.10)$$

$$= \sum_{j=-\infty}^{\infty} \int_{(j-1)T}^{jT} h(\tau) U(kT - \tau) d\tau \quad (15.1.11)$$

Note that by definition  $U(kT - \tau) = U_{k-j}$  for  $(j-1)T < \tau \leq jT$  so

$$Y(kT) = \sum_{j=-\infty}^{\infty} \int_{(j-1)T}^{jT} h(\tau) d\tau U_{k-j} \quad (15.1.12)$$

Define  $h_j = \int_{(j-1)T}^{jT} h(\tau) d\tau$ , then

$$Y_k = \sum_{j=-\infty}^{\infty} h_j U_{k-j} \quad (15.1.13)$$

which is a convolution sum in terms of the discrete LTI system defined by  $h_k$ .

### Transfer Functions with Random Signals

For a causal system, we should have  $h_k = 0$  for  $k < 0$  (and analogously  $h(t) = 0$  for  $t < 0$ ) otherwise this would imply a response has been ‘caused’ by an input in the future. So a convolution sum can be written as

$$Y_k = \sum_{j=0}^{\infty} h_j U_{k-j} \quad (15.1.14)$$

Define the shift operator  $z$  so that  $z^j U_k := U_{k+j}$  and  $z^{-j} U_k := U_{k-j}$ . Thus

$$Y_k = \sum_{j=0}^{\infty} h_j z^{-j} U_k \quad (15.1.15)$$

$$\frac{Y_k}{U_k} = \sum_{j=0}^{\infty} h_j z^{-j} \quad (15.1.16)$$

We define  $H(z) := \frac{Y_k}{U_k} = \sum_{j=0}^{\infty} h_j z^{-j}$  and this is the  $z$ -transform of  $h_k$ . Analogously, a continuous system with impulse response  $h(t)$  has a transfer function which is the Laplace

transform  $H(s) = \mathcal{L}[h(t)]$ . Note that we can also define ‘strictly causal’ systems where  $h_k = 0$  for  $k \leq 0$  and  $h(t) = 0$  for  $t \leq 0$ , then the transfer function (in the discrete case) becomes

$$H(z) = \sum_{j=1}^{\infty} h_j z^{-j} \quad (15.1.17)$$

### Wide Sense Stationarity in LTI Systems

Suppose the input  $X(t)$  to an LTI system with impulse response  $h(t)$  is a wide sense stationary process. That is,  $X(t)$  has mean function  $\mu_X$  and autocorrelation function  $R_X(\tau)$ . Then the output  $Y(t)$  will also be a wide sense stationary process with mean function given by

$$\mu_Y = \mathbb{E}[Y(t)] \quad (15.1.18)$$

$$= \mathbb{E}\left[\int_{-\infty}^{\infty} h(u) X(t-u) du\right] \quad (15.1.19)$$

$$= \int_{-\infty}^{\infty} h(u) \mathbb{E}[X(t-u)] du \quad (15.1.20)$$

$$= \mu_X \int_{-\infty}^{\infty} h(u) du \quad (15.1.21)$$

and autocorrelation function given by

$$R_Y(\tau) = \mathbb{E}[Y(t) Y(t-\tau)] \quad (15.1.22)$$

$$= \mathbb{E}\left[\int_{-\infty}^{\infty} h(u) X(t-u) du \int_{-\infty}^{\infty} h(v) X(t+\tau-v) dv\right] \quad (15.1.23)$$

$$= \mathbb{E}\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) X(t-u) X(t+\tau-v) dudv\right] \quad (15.1.24)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) \mathbb{E}[X(t-u) X(t+\tau-v)] dudv \quad (15.1.25)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) R_X(\tau+u-v) dudv \quad (15.1.26)$$

The cross-correlation function between  $X(t)$  and  $Y(t)$  is given by

$$R_{XY}(\tau) = \mathbb{E}[X(t) Y(t-\tau)] \quad (15.1.27)$$

$$= \mathbb{E}\left[X(t) \int_{-\infty}^{\infty} h(u) X(t+\tau-u) du\right] \quad (15.1.28)$$

$$= \int_{-\infty}^{\infty} h(u) \mathbb{E}[X(t) X(t+\tau-u)] du \quad (15.1.29)$$

$$= \int_{-\infty}^{\infty} h(u) R_X(\tau-u) du \quad (15.1.30)$$

and the autocorrelation function of  $Y(t)$  can be related to the cross-correlation function by the substitution  $w = -u$ :

$$R_Y(\tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) R_X(\tau+u-v) dudv \quad (15.1.31)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(-w) h(v) R_X(\tau-w-v) dw dv \quad (15.1.32)$$

$$= \int_{-\infty}^{\infty} h(-w) \underbrace{\int_{-\infty}^{\infty} h(v) R_X(\tau-w-v) dv}_{R_{XY}(\tau-w)} dw \quad (15.1.33)$$

$$= \int_{-\infty}^{\infty} h(-w) R_{XY}(\tau - w) dw \quad (15.1.34)$$

By analogous arguments, if  $X_n$  is a wide sense stationary random sequence and  $h_n$  is the impulse response of a discrete LTI system, then the output sequence  $Y_n$  is also a wide sense stationary random sequence which satisfies

$$\mu_Y = \mu_X \sum_{i=-\infty}^{\infty} h_i \quad (15.1.35)$$

$$R_Y(k) = \sum_{i=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} h_i h_{\ell} R_X(k+i-\ell) \quad (15.1.36)$$

$$R_{XY}(k) = \sum_{i=-\infty}^{\infty} h_i R_X(k-i) \quad (15.1.37)$$

$$R_Y(k) = \sum_{i=-\infty}^{\infty} h_{-i} R_{XY}(k-i) \quad (15.1.38)$$

### Random Discrete-Time Finite Impulse Response Systems

Consider a causal discrete-time system which has a  $N^{\text{th}}$ -order finite impulse response (FIR), i.e.  $h_k = 0$  for  $k < 0$  and  $k \geq N$ . Via convolution, the output can be written as

$$Y_k = \sum_{j=0}^{N-1} h_j X_{k-j} \quad (15.1.39)$$

and we can also represent the impulse response using the vector

$$\mathbf{h} = [h_0 \ \dots \ h_{N-1}]^\top \quad (15.1.40)$$

Let the first  $L$  inputs to the system be given by  $\mathbf{X} = (X_0, \dots, X_L)$ , and assume  $X_k = 0$  when  $k < 0$ . The sequence of outputs can be expressed in matrix form by

$$\underbrace{\begin{bmatrix} Y_0 \\ \vdots \\ Y_{N-1} \\ \vdots \\ Y_{L-N} \\ \vdots \\ Y_{L-1} \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} h_0 & & & & & & \\ \vdots & \ddots & & & & & \\ h_{N-1} & \dots & h_0 & & & & \\ & \ddots & & \ddots & & & \\ & & \ddots & & \ddots & & \\ & & & \ddots & & h_0 & \\ & & & & \ddots & & \ddots \\ & & & & & h_{N-1} & \dots & h_0 \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} X_0 \\ \vdots \\ X_{N-1} \\ \vdots \\ X_{L-N} \\ \vdots \\ X_{L-1} \end{bmatrix}}_{\mathbf{X}} \quad (15.1.41)$$

Observe that  $\mathbf{H}$  has a *band-Toeplitz* matrix structure. Suppose  $X_k$  is wide sense stationary with mean  $\mu_X$  and autocorrelation function  $R_X(m)$ , then we can form the matrix  $\mathbf{R}_X$  as

$$\mathbf{R}_X = \mathbf{E} [\mathbf{XX}^\top] \quad (15.1.42)$$

$$= \begin{bmatrix} R_X(0) & \dots & R_X(L-1) \\ \vdots & \ddots & \vdots \\ R_X(L-1) & \dots & R_X(0) \end{bmatrix} \quad (15.1.43)$$

which also has a Toeplitz structure. By the definition of the covariance  $\mathbf{C}_X := \text{Cov}(\mathbf{X})$ , we have

$$\mathbf{C}_X = \mathbf{R}_X - \mu_X \mu_X^\top \quad (15.1.44)$$

where  $\mu_X = \mu_X \mathbf{1}$ . To find the covariance of  $\mathbf{Y}$  and the cross-covariance between  $\mathbf{X}$  and  $\mathbf{Y}$ , we can use the joint representation

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} I \\ \mathbf{H} \end{bmatrix} \mathbf{X} \quad (15.1.45)$$

Hence

$$\begin{bmatrix} \mathbf{C}_X & \mathbf{C}_{XY} \\ \mathbf{C}_{YX} & \mathbf{C}_Y \end{bmatrix} := \begin{bmatrix} \text{Cov}(\mathbf{X}) & \text{Cov}(\mathbf{X}, \mathbf{Y}) \\ \text{Cov}(\mathbf{Y}, \mathbf{X}) & \text{Cov}(\mathbf{Y}) \end{bmatrix} \quad (15.1.46)$$

$$= \text{Cov}\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}\right) \quad (15.1.47)$$

$$= \begin{bmatrix} I \\ \mathbf{H} \end{bmatrix} \mathbf{C}_X [I \ \mathbf{H}^\top] \quad (15.1.48)$$

$$= \begin{bmatrix} \mathbf{C}_X & \mathbf{C}_X \mathbf{H}^\top \\ \mathbf{H} \mathbf{C}_X & \mathbf{H} \mathbf{C}_X \mathbf{H}^\top \end{bmatrix} \quad (15.1.49)$$

$$= \begin{bmatrix} \mathbf{R}_X - \mu_X \mu_X^\top & (\mathbf{R}_X - \mu_X \mu_X^\top) \mathbf{H}^\top \\ \mathbf{H}(\mathbf{R}_X - \mu_X \mu_X^\top) & \mathbf{H}(\mathbf{R}_X - \mu_X \mu_X^\top) \mathbf{H}^\top \end{bmatrix} \quad (15.1.50)$$

and by noting  $\mathbb{E}[\mathbf{Y}] = \mathbf{H}\mu_X$ , we also find that the analogous matrices  $\mathbf{R}_Y$  and  $\mathbf{R}_{YX}$  for the autocorrelations and cross-correlations of  $Y_k$ , and between  $Y_k$  and  $X_k$  respectively are given by

$$\mathbf{R}_Y = \mathbf{H} \mathbf{R}_X \mathbf{H}^\top \quad (15.1.51)$$

$$\mathbf{R}_{YX} = \mathbf{H} \mathbf{R}_X \quad (15.1.52)$$

### 15.1.2 Discrete-Time Stochastic State-Space Models [114]

#### Nonlinear Discrete-Time Stochastic State-Space Models

A general nonlinear time-invariant specification of a discrete-time stochastic state-space model is the stochastic difference equation

$$x_{k+1} = f(x_k, w_k) \quad (15.1.53)$$

for the *state equation*, where  $f : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a nonlinear function of the state  $x_k$  and process noise  $w_k$ . Usual assumptions for  $w_k$  are either white noise or i.i.d. noise. This is followed by the *observation equation*

$$y_k = h(x_k, v_k) \quad (15.1.54)$$

where  $h : \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}^p$  with measurement noise  $v_k$  which is usually assumed to be i.i.d. or white. We can allow for a time-varying specification by indexing  $f$  and  $h$  by  $k$ , i.e. with  $f_k(x_k, w_k)$  and  $h_k(x_k, v_k)$ . We could also allow for an exogenous input  $u_k$  so that the state update equation is  $f(x_k, u_k, w_k)$ . If the input sequence is known, then this just induces a time-varying model.

### Transition Density Representation of State-Space Models

We can show that if the noise is i.i.d. and the functions  $f, h$  are sufficiently well-behaved, then a nonlinear discrete-time stochastic state-space model can be equivalently represented by a hidden Markov model. Assume that  $f$  and  $h$  are invertible with respect to  $w_k$  and  $v_k$  respectively, so that we have

$$w_k = f_w^{-1}(x_{k+1}, x_k) \quad (15.1.55)$$

$$v_k = h_v^{-1}(x_{k+1}, y_k) \quad (15.1.56)$$

Let  $p_w(\cdot)$  and  $p_v(\cdot)$  denote the densities of  $w_k$  and  $v_k$  respectively. Then using the relation between densities by invertible transformations, the state transition density can be expressed as

$$p(x_{k+1}|x_k) = p_w(f_w^{-1}(x_{k+1}, x_k)|x_k) \left| \det \left( \frac{\partial f_w^{-1}(x_{k+1}, x_k)}{\partial x_{k+1}} \right) \right| \quad (15.1.57)$$

assuming that  $f_w^{-1}(x_{k+1}, x_k)$  is continuously differentiable. By independence of  $w_k$ , this becomes

$$p(x_{k+1}|x_k) = p_w(f_w^{-1}(x_{k+1}, x_k)) \left| \det \left( \frac{\partial f_w^{-1}(x_{k+1}, x_k)}{\partial x_{k+1}} \right) \right| \quad (15.1.58)$$

Similarly, by assuming that  $h_v^{-1}(x_{k+1}, x_k)$  is continuously differentiable:

$$p(y_k|x_k) = p_v(h_v^{-1}(x_{k+1}, y_k)) \left| \det \left( \frac{\partial h_v^{-1}(x_{k+1}, y_k)}{\partial x_{k+1}} \right) \right| \quad (15.1.59)$$

gives the observation likelihood.

### Additive Noise State-Space Models

An additive noise state-space model is a special case of the nonlinear state-space model, and is given in the time-invariant case by

$$x_{k+1} = f(x_k) + \Gamma(x_k) w_k \quad (15.1.60)$$

$$y_k = h(x_k) + \Psi(x_k) v_k \quad (15.1.61)$$

where  $\Gamma : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$  and  $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^{p \times p}$ . In this case, we have the inverses

$$w_k = \Gamma(x_k)^{-1}(x_{k+1} - f(x_k)) \quad (15.1.62)$$

$$v_k = \Psi(x_k)^{-1}(y_k - h(x_k)) \quad (15.1.63)$$

So the state transition density is

$$p(x_{k+1}|x_k) = p_w\left(\Gamma(x_k)^{-1}(x_{k+1} - f(x_k))\right) \left| \det \left( \frac{\partial \Gamma(x_k)^{-1}(x_{k+1} - f(x_k))}{\partial x_{k+1}} \right) \right| \quad (15.1.64)$$

$$= p_w\left(\Gamma(x_k)^{-1}(x_{k+1} - f(x_k))\right) \left| \det \left( \Gamma(x_k)^{-1} \right) \right| \quad (15.1.65)$$

and observation likelihood is similarly

$$p(y_k|x_k) = p_v\left(\Psi(x_k)^{-1}(y_k - h(x_k))\right) \left| \det \left( \Psi(x_k)^{-1} \right) \right| \quad (15.1.66)$$

### Linear Discrete-Time Stochastic State-Space Models

A special case of the additive noise state-space model is with  $f(x_k) = Ax_k$ ,  $\Gamma(x_k) = I$ ,  $h(x_k) = Cx_k$  and  $\Psi(x_k) = I$  for matrices  $A$  and  $C$  of appropriate dimension. Note that we can assume  $\Gamma(x_k)$  and  $\Psi(x_k)$  as identity matrices without loss of generality (as opposed to some non-identity constant matrix), otherwise it would just scale the noise. This yields the linear stochastic state-space model

$$x_{k+1} = Ax_k + w_k \quad (15.1.67)$$

$$y_k = Cx_k + v_k \quad (15.1.68)$$

with the transition density and observation likelihood

$$p(x_{k+1}|x_k) = p_w(x_{k+1} - Ax_k) \quad (15.1.69)$$

$$p(y_k|x_k) = p_v(y_k - Cx_k) \quad (15.1.70)$$

because the determinant of the identity matrix is 1.

### Linear Gaussian State-Space Models

A special case of the linear stochastic state-space model is when  $w_k \sim \mathcal{N}(0, Q)$  and  $v_k \sim \mathcal{N}(0, R)$ , giving the linear Gaussian state-space model. Thus

$$p(x_{k+1}|x_k) = \mathcal{N}(Ax_k, Q) \quad (15.1.71)$$

$$p(y_k|x_k) = \mathcal{N}(Cx_k, R) \quad (15.1.72)$$

The process  $x_k$  will also be a Gaussian Markov process.

### Jump Markov Linear Systems

A jump Markov linear system has dynamics like a linear system, except these dynamics can ‘jump’ around in a fashion which follows a Markov chain. The system can be written as

$$z_{k+1} = A(s_k) z_k + \Gamma(s_k) w_k \quad (15.1.73)$$

$$y_k = C(s_k) z_k + \Psi(s_k) v_k \quad (15.1.74)$$

where  $w_k$ ,  $v_k$  are noise processes, and  $s_k$  is an (unobserved) finite-state Markov chain on  $\{1, \dots, N\}$  with some transition matrix  $P$ . The actual state here is the combined  $x_k = (z_k, s_k)$ , thus there are some states which can be continuous, and other state which are discrete.

Jump Markov linear systems generalise both linear systems and several variants of hidden Markov models. If the Markov chain  $s_k$  is known/observed, then the dynamics reduce to a linear time-varying system

$$z_{k+1} = A_k z_k + \Gamma_k w_k \quad (15.1.75)$$

$$y_k = C_k z_k + \Psi_k v_k \quad (15.1.76)$$

or further into a linear time-variant system (which is a continuous-state hidden Markov model) if  $N = 1$ . If  $A(s_k) = I$ ,  $\Gamma(s_k) = 0$ , then  $z_k = \bar{z}$  is a constant for all time. With  $\Psi(s_k) = I$ , then the output

$$y_k = C(s_k) \bar{z} + v_k \quad (15.1.77)$$

becomes that of a continuous-observation HMM with additive noise. Alternatively, jump Markov linear systems can also express finite-state HMMs with finite observation symbols. To do this, keep  $A(s_k) = I$ ,  $\Gamma(s_k) = 0$  and now have  $\Psi(s_k) = 0$ , then write the output as

$$y_k = \underbrace{\begin{bmatrix} \mathbb{I}_{\{s_k=1\}} & \cdots & \mathbb{I}_{\{s_k=N\}} \end{bmatrix}}_{\Psi(s_k)} \underbrace{\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}}_{v_k} \quad (15.1.78)$$

where  $v_k$  is an i.i.d. vector process such that its components mimics the observation likelihoods:

$$\Pr(Y_i = j) = \Pr(y_k = j | s_k = i) \quad (15.1.79)$$

for all  $j \in \{1, \dots, M\}$  and  $i \in \{1, \dots, N\}$ .

To obtain the transition densities for a jump Markov linear system, we first factorise using the Markov property

$$p(x_{k+1}|x_k) = p(z_{k+1}, s_{k+1}|z_k, s_k) \quad (15.1.80)$$

$$= p(z_{k+1}|s_{k+1}, z_k, s_k) p(s_{k+1}|z_k, s_k) \quad (15.1.81)$$

$$= p(z_{k+1}|s_{k+1}, z_k) p(s_{k+1}|s_k) \quad (15.1.82)$$

For  $p(z_{k+1}|s_{k+1}, z_k)$  we can refer to the transition density for linear systems and additive noise systems. For  $p(s_{k+1}|s_k)$ , this comes from the transition matrix  $P$ . If a right-stochastic convention is used, then

$$p(x_{k+1}|x_k) = \left| \det \left( \Gamma(s_k)^{-1} \right) \right| p_w \left( \Gamma(s_k)^{-1} (z_{k+1} - A(s_k) z_k) \right) P_{s_k, s_{k+1}} \quad (15.1.83)$$

In obtaining the observation likelihood, it becomes a similar form to the additive noise case:

$$p(y_k|x_k) = p(y_k|z_k, s_k) \quad (15.1.84)$$

$$= \left| \det \left( \Psi(s_k)^{-1} \right) \right| p_v \left( \Psi(s_k)^{-1} (y_k - C(s_k) z_k) \right) \quad (15.1.85)$$

### 15.1.3 Continuous-Time Stochastic State-Space Models

## 15.2 Power Spectral Density

Consider sample functions  $x(t)$  of a stationary process  $X(t)$ . The Fourier transform of  $x(t)$  usually does not exist because it is not integrable on an infinite domain. However, if we define the truncated sample function

$$x_T(t) = \begin{cases} x(t), & -T \leq t \leq T \\ 0, & \text{elsewhere} \end{cases} \quad (15.2.1)$$

then this truncated sample function has a Fourier transform. Denote this Fourier transform by

$$\tilde{x}_T(f) = \int_{-\infty}^{\infty} x(t) e^{j2\pi ft} dt \quad (15.2.2)$$

$$= \int_{-T}^{T} x(t) e^{j2\pi ft} dt \quad (15.2.3)$$

where  $j = \sqrt{-1}$ . The power spectral density of  $X(t)$  is a function of frequency  $f$  which may be roughly interpreted as the average density of power of frequency content  $f$  for sample functions of  $X(t)$ . It is defined by

$$S_X(f) = \lim_{T \rightarrow \infty} \frac{1}{2T} \mathbb{E} \left[ |\tilde{x}_T(f)|^2 \right] \quad (15.2.4)$$

$$= \lim_{T \rightarrow \infty} \frac{1}{2T} \mathbb{E} \left[ \left| \int_{-T}^T x(t) e^{-j2\pi f t} dt \right|^2 \right] \quad (15.2.5)$$

The power spectral density  $S_X(\phi)$  of a wide-sense stationary random sequence  $X_n$  is also defined similarly as

$$S_X(\phi) = \lim_{L \rightarrow \infty} \frac{1}{2L+1} \mathbb{E} \left[ \left| \sum_{n=-L}^L X_n e^{-j2\pi \phi n} \right|^2 \right] \quad (15.2.6)$$

### 15.2.1 Wiener-Khintchine Theorem [219]

The Wiener-Khintchine theorem states that the autocorrelation function and power spectral density of a wide sense stationary stochastic process  $X(t)$  are Fourier transform pairs. That is,

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f \tau} d\tau \quad (15.2.7)$$

$$R_X(\tau) = \int_{-\infty}^{\infty} S_X(f) e^{-j2\pi f \tau} df \quad (15.2.8)$$

*Proof.* Begin by writing the expression for  $\mathbb{E} \left[ |\tilde{x}_T(f)|^2 \right]$ .

$$\mathbb{E} \left[ |\tilde{x}_T(f)|^2 \right] = \mathbb{E} \left[ \left| \int_{-T}^T x(t) e^{-j2\pi f t} dt \right|^2 \right] \quad (15.2.9)$$

The Fourier transform  $\tilde{x}_T(f)$  is a complex valued function of frequency, but since  $x(t)$  is real valued, then the complex conjugate only affects the complex exponential, i.e.  $\tilde{x}_T^*(f) = \int_{-T}^T x(t) e^{j2\pi f t} dt$ . Then use the fact that for complex numbers,  $|\tilde{x}_T(f)|^2 = \tilde{x}_T(f) \tilde{x}_T^*(f)$ . Hence

$$\mathbb{E} \left[ |\tilde{x}_T(f)|^2 \right] = \mathbb{E} \left[ \left( \int_{-T}^T x(t) e^{-j2\pi f t} dt \right) \left( \int_{-T}^T x(t') e^{j2\pi f t'} dt' \right) \right] \quad (15.2.10)$$

$$= \mathbb{E} \left[ \int_{-T}^T \int_{-T}^T x(t) x(t') e^{-j2\pi f(t-t')} dt dt' \right] \quad (15.2.11)$$

$$= \int_{-T}^T \int_{-T}^T \mathbb{E} [x(t) x(t')] e^{-j2\pi f(t-t')} dt dt' \quad (15.2.12)$$

$$= \int_{-T}^T \int_{-T}^T R_X(t-t') e^{-j2\pi f(t-t')} dt dt' \quad (15.2.13)$$

Introduce the change of variables  $\tau = t - t'$ , noting that the integrand then only depends on  $\tau$ . To determine the limits of integration, we reason that if  $-T \leq t \leq T$  and  $-T \leq t' \leq T$ , then  $-2T \leq \tau \leq 2T$ . Then a graphical sketch on the  $t$ - $t'$  plane of the region  $\Omega$  for  $t'$  given by the intersection

$$\Omega = \{t' : t - t' = \tau \cap -T \leq t \leq T \cap -T \leq t' \leq T \cap -2T \leq \tau \leq 2T\} \quad (15.2.14)$$

will reveal

$$\Omega = \left\{ t' : \begin{cases} -T \leq t' \leq T - \tau, & \tau \geq 0 \\ -T - \tau \leq t' \leq T, & \tau < 0 \end{cases} \right\} \quad (15.2.15)$$

Combining these two cases gives  $\Omega = \{t' : -T \leq t' \leq T - |\tau|\}$ . Hence the integral can be written as

$$\mathbb{E} [|\tilde{x}_T(f)|^2] = \int_{-2T}^{2T} \int_{-T}^{T-|\tau|} R_X(\tau) e^{-j2\pi f\tau} dt' d\tau \quad (15.2.16)$$

$$= \int_{-2T}^{2T} (2T - |\tau|) R_X(\tau) e^{-j2\pi f\tau} d\tau \quad (15.2.17)$$

This can be visually imagined as integrating along the positive (upward sloping) diagonals for all intercepts  $\tau$  in the box bounded by  $-T \leq t \leq T$  and  $-T \leq t' \leq T$ . Then

$$\frac{1}{2T} \mathbb{E} [|\tilde{x}_T(f)|^2] = \frac{1}{2T} \int_{-2T}^{2T} (2T - |\tau|) R_X(\tau) e^{-j2\pi f\tau} d\tau \quad (15.2.18)$$

$$= \int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T}\right) R_X(\tau) e^{-j2\pi f\tau} d\tau \quad (15.2.19)$$

Define the function

$$h_T(\tau) = \begin{cases} 1 - \frac{|\tau|}{2T}, & |\tau| \leq 2T \\ 0, & \text{elsewhere} \end{cases} \quad (15.2.20)$$

This gives  $\lim_{T \rightarrow \infty} h_T(\tau) = 1$ . The power spectral density is therefore

$$S_X(f) = \lim_{T \rightarrow \infty} \frac{1}{2T} \mathbb{E} [|\tilde{x}_T(f)|^2] \quad (15.2.21)$$

$$= \lim_{T \rightarrow \infty} \int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T}\right) R_X(\tau) e^{-j2\pi f\tau} d\tau \quad (15.2.22)$$

$$= \int_{-\infty}^{\infty} \lim_{T \rightarrow \infty} (h_T(\tau)) R_X(\tau) e^{-j2\pi f\tau} d\tau \quad (15.2.23)$$

$$= \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau} d\tau \quad (15.2.24)$$

which is the Fourier transform of  $R_X(\tau)$ .  $\square$

Note that one may define average power of a stochastic process as the second moment  $\mathbb{E} [X(t)^2] = R_X(0)$ , in which case

$$R_X(0) = \int_{-\infty}^{\infty} S_X(f) df \quad (15.2.25)$$

In this sense, the power spectral density can be viewed as taking the role of a density function. So for a given frequency  $f$ , the power spectral density gives an indication of the contribution of that frequency to average power (or second moment).

### 15.2.2 Discrete-Time Wiener-Khintchine Theorem

The discrete-time Wiener-Khintchine theorem (sometimes called Wold's theorem) is the discrete-time analog to the Wiener-Khintchine theorem, involving the connection between the autocorrelation and power spectral density of random sequences by the discrete-time Fourier transform. For a wide-sense stationary random sequence  $X_n$ , the autocorrelation function  $R_X(k)$  and power spectral density  $S_X(\phi)$  form a discrete-time Fourier transform pair:

$$S_X(\phi) = \sum_{k=-\infty}^{\infty} R_X(k) e^{-j2\pi\phi k} \quad (15.2.26)$$

$$R_X(k) = \int_{-1/2}^{1/2} S_X(\phi) e^{j2\pi\phi k} d\phi \quad (15.2.27)$$

The proof is analogous to continuous time.

*Proof.* First write

$$\mathbb{E} \left[ \left| \sum_{n=-L}^L X_n e^{-j2\pi\phi n} \right|^2 \right] = \mathbb{E} \left[ \left( \sum_{n=-L}^L X_n e^{-j2\pi\phi n} \right) \left( \sum_{n'=-L}^L X_{n'} e^{j2\pi\phi n'} \right) \right] \quad (15.2.28)$$

$$= \mathbb{E} \left[ \sum_{n=-L}^L \sum_{n'=-L}^L X_n X_{n'} e^{-j2\pi\phi(n-n')} \right] \quad (15.2.29)$$

$$= \sum_{n=-L}^L \sum_{n'=-L}^L \mathbb{E}[X_n X_{n'}] e^{-j2\pi\phi(n-n')} \quad (15.2.30)$$

$$= \sum_{n=-L}^L \sum_{n'=-L}^L R_X(n - n') e^{-j2\pi\phi(n-n')} \quad (15.2.31)$$

Analogous to continuous time, introduce a change of variables  $\eta = n - n'$  so that

$$\mathbb{E} \left[ \left| \sum_{n=-L}^L X_n e^{-j2\pi\phi n} \right|^2 \right] = \sum_{\eta=-2L}^{2L} \sum_{n'=-L}^{L-|\eta|} R_X(\eta) e^{-j2\pi\phi\eta} \quad (15.2.32)$$

$$= \sum_{\eta=-2L}^{2L} (2L + 1 - |\eta|) R_X(\eta) e^{-j2\pi\phi\eta} \quad (15.2.33)$$

noting that  $\sum_{n'=-L}^{L-|\eta|} = 2L + 1 - |\eta|$ . So now

$$\frac{1}{2L+1} \mathbb{E} \left[ \left| \sum_{n=-L}^L X_n e^{-j2\pi\phi n} \right|^2 \right] = \sum_{\eta=-2L}^{2L} \left( 1 - \frac{|\eta|}{2L+1} \right) R_X(\eta) e^{-j2\pi\phi\eta} \quad (15.2.34)$$

Introduce the function

$$h_L(\eta) = \begin{cases} 1 - \frac{|\eta|}{2L+1}, & |\eta| \leq 2L+1 \\ 0, & \text{elsewhere} \end{cases} \quad (15.2.35)$$

where we have that  $\lim_{L \rightarrow \infty} h_L(\eta) = 1$ . Then

$$S_X(\phi) = \lim_{L \rightarrow \infty} \frac{1}{2L+1} \mathbb{E} \left[ \left| \sum_{n=-L}^L X_n e^{-j2\pi\phi n} \right|^2 \right] \quad (15.2.36)$$

$$= \lim_{L \rightarrow \infty} \sum_{\eta=-2L}^{2L} h_L(\eta) R_X(\eta) e^{-j2\pi\phi\eta} \quad (15.2.37)$$

$$= \sum_{\eta=-\infty}^{\infty} R_X(\eta) e^{-j2\pi\phi\eta} \quad (15.2.38)$$

which is the discrete-time Fourier transform of  $R_X(\eta)$ , the autocorrelation function of  $X_n$ .  $\square$

### 15.2.3 Cross Spectral Density

Appealing to the Wiener-Kintchine theorem gives us a way to appropriately define a notion of cross spectral density. The cross spectral density  $S_{XY}(f)$  of two jointly wide-sense stationary processes  $X(t)$  and  $Y(t)$  is the Fourier transform of the cross-correlation  $R_{XY}(\tau)$ :

$$S_{XY}(f) = \int_{-\infty}^{\infty} R_{XY}(\tau) e^{-j2\pi f\tau} d\tau \quad (15.2.39)$$

Analogously in discrete-time, the cross spectral density  $S_{XY}(\phi)$  of two jointly wide-sense random sequences  $X_n$  and  $Y_n$  is the discrete-time Fourier transform of the cross-correlation  $R_{XY}(k)$ :

$$S_{XY}(\phi) = \sum_{k=-\infty}^{\infty} R_{XY}(k) e^{-j2\pi\phi k} \quad (15.2.40)$$

### 15.2.4 Spectral Density Characterisation of Filters

#### Spectral Density Characterisation of Filters in Continuous-Time

The frequency response of linear filters have a spectral density interpretation. The following result establishes the connection between the spectral density of the input and the spectral density of the output for a LTI system (or filter). Let  $h(t)$  be the impulse response of a stable LTI system. Then it has a Fourier transform  $H(f)$ , which is called the frequency response. If a wide-sense stationary stochastic process  $X(t)$  with spectral density  $S_X(f)$  is the input to a LTI system with transfer function  $H(f)$ , then the spectral density of the output  $Y(t)$  is

$$S_Y(f) = |H(f)|^2 S_X(f) \quad (15.2.41)$$

*Proof.* As  $X(t)$  is wide-sense stationary, then  $Y(t)$  is also wide-sense stationary with autocorrelation function

$$R_Y(\tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) R_X(\tau + u - v) du dv \quad (15.2.42)$$

Then using the property of the spectral density as the Fourier transform of the autocorrelation function:

$$S_Y(f) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) R_X(\tau + u - v) du dv \right) e^{-j2\pi f\tau} d\tau \quad (15.2.43)$$

Perform the change of variables  $\tau' = \tau + v - u$ . Then

$$S_Y(f) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) R_X(\tau') du dv \right) e^{-j2\pi f(\tau+u-v)} d\tau' \quad (15.2.44)$$

$$= \int_{-\infty}^{\infty} h(u) e^{-j2\pi fu} du \int_{-\infty}^{\infty} h(v) e^{j2\pi fv} dv \int_{-\infty}^{\infty} R_X(\tau') d\tau' \quad (15.2.45)$$

$$= H(f) H^*(f) S_X(f) \quad (15.2.46)$$

$$= |H(f)|^2 S_X(f) \quad (15.2.47)$$

where  $H^*(f)$  denotes the complex conjugate of  $H(f)$ .  $\square$

Rearranging:

$$|H(f)|^2 = \frac{S_Y(f)}{S_X(f)} \quad (15.2.48)$$

Thus, the squared frequency response is characterised by the ratio of output to input spectral densities. Recall that the continuous-time Fourier transform is equivalent to the two-sided

Laplace transform with the substitution  $s = e^{j2\pi f}$ , so by abusing notation, we can alternatively write

$$|H(s)|^2 = \frac{S_Y(s)}{S_X(s)} \quad (15.2.49)$$

This characterises the squared transfer function of the filter as the ratio of the Laplace transforms of the autocorrelation functions. It can be similarly shown that the cross spectral density is given by

$$S_{XY}(f) = H(f) S_X(f) \quad (15.2.50)$$

*Proof.* From the cross-correlation:

$$R_{XY}(\tau) = \int_{-\infty}^{\infty} h(u) R_X(\tau - u) du \quad (15.2.51)$$

By the property of the cross spectral density:

$$S_{XY}(f) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} h(u) R_X(\tau - u) du \right) e^{-j2\pi f \tau} d\tau \quad (15.2.52)$$

Make the substitution  $\tau' = \tau - u$ :

$$S_{XY}(f) = \int_{-\infty}^{\infty} h(u) e^{-j2\pi f u} du \int_{-\infty}^{\infty} R_X(\tau') e^{-j2\pi f \tau'} d\tau' \quad (15.2.53)$$

$$= H(f) S_X(f) \quad (15.2.54)$$

□

Hence

$$H(f) = \frac{S_{XY}(f)}{S_X(f)} \quad (15.2.55)$$

An alternative derivation is to use the fact that convolution in time domain becomes multiplication in frequency domain. Therefore the frequency response of a filter can be characterised as the ratio of the input-output cross spectral density to the input spectral density (with a wide sense stationary input). Alternatively in terms of the Laplace transform, we can state

$$H(s) = \frac{S_{XY}(s)}{S_X(s)} \quad (15.2.56)$$

### Spectral Density Characterisation of Filters in Discrete-Time

Analogous results can be derived in the discrete-time case. If a wide-sense stationary random sequence  $X_n$  with spectral density  $S_X(\phi)$  is the input to a stable LTI system with pulse response  $h(\phi)$  and frequency response  $H(\phi)$  as the discrete-time Fourier transform of  $h(\phi)$ , then the output  $Y_n$  has spectral density

$$S_Y(\phi) = |H(\phi)|^2 S_X(\phi) \quad (15.2.57)$$

*Proof.* The autocorrelation of  $Y_n$  in terms of the autocorrelation of  $X_n$  is

$$R_Y(k) = \sum_{i=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} h(i) h(\ell) R_X(k + i - \ell) \quad (15.2.58)$$

Using the property of the spectral density:

$$S_Y(\phi) = \sum_{k=-\infty}^{\infty} \left( \sum_{i=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} h(i) h(\ell) R_X(k + i - \ell) \right) e^{-j2\pi \phi k} \quad (15.2.59)$$

Make the change of variables  $k' = k + i - \ell$ :

$$S_Y(\phi) = \sum_{k'=-\infty}^{\infty} \left( \sum_{i=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} h(i) h(\ell) R_X(k') \right) e^{-j2\pi\phi(k'-i+\ell)} \quad (15.2.60)$$

$$= \sum_{i=-\infty}^{\infty} h_i e^{-j2\pi\phi i} \sum_{\ell=-\infty}^{\infty} h(\ell) e^{-j2\pi\phi\ell} \sum_{k'=-\infty}^{\infty} e^{-j2\pi\phi k'} \quad (15.2.61)$$

$$= H(\phi) H^*(\phi) S_X(\phi) \quad (15.2.62)$$

$$= |H(\phi)|^2 S_X(\phi) \quad (15.2.63)$$

□

Then

$$|H(\phi)|^2 = \frac{S_Y(\phi)}{S_X(\phi)} \quad (15.2.64)$$

with analogous characterisation to the continuous-time case. Since the  $z$ -transform is related to the discrete-time Fourier transform by the substitution  $z = e^{j2\pi\phi}$ , we have, with abuse of notation:

$$|H(z)|^2 = \frac{S_Y(z)}{S_X(z)} \quad (15.2.65)$$

where  $H(z)$  is the  $z$ -transform of the filter, while  $S_Y(z)$ ,  $S_X(z)$  are  $z$ -transforms of the output and input autocorrelation functions, respectively. In a similar fashion to above and the continuous-time case, the cross-spectral density is

$$S_{XY}(\phi) = H(\phi) S_X(\phi) \quad (15.2.66)$$

Or in terms of  $z$ -transforms:

$$H(z) = \frac{S_{XY}(z)}{S_X(z)} \quad (15.2.67)$$

### 15.2.5 Coloured Noise

We can view noise with different spectral densities as the outputs of a LTI system (or filter) with white noise as an input. Thus, the filter ‘colours’ the noise.

#### Spectral Density of White Noise

White noise may be defined as the stochastic process which has a constant spectral density, in both the discrete-time and continuous-time case. In discrete-time, let the spectral density  $S_X(\phi) = c$  be a constant. Then the autocorrelation function is

$$R_X(k) = \int_{-1/2}^{1/2} c \cdot e^{j2\pi\phi k} d\phi \quad (15.2.68)$$

$$= \int_{-1/2}^{1/2} c \cdot \cos(2\pi\phi k) d\phi + \int_{-1/2}^{1/2} c \cdot j \sin(2\pi\phi k) d\phi \quad (15.2.69)$$

Since  $\sin(2\pi\phi k)$  is an odd function, then

$$R_X(k) = \int_{-1/2}^{1/2} c \cdot \cos(2\pi\phi k) d\phi \quad (15.2.70)$$

$$= \frac{c}{2\pi k} [\sin(\pi k) + \sin(-\pi k)] \quad (15.2.71)$$

$$= \frac{c \sin(\pi k)}{\pi k} \quad (15.2.72)$$

where  $\frac{c \sin(\pi k)}{\pi k}$  is known as the (normalised) sinc( $k$ ) function. Since  $k$  is integer-valued, we can work with the limit  $\lim_{k \rightarrow 0} \frac{\sin k}{k} = 1$  (which can be shown through L'Hôpital's rule) and state that the autocorrelation function is

$$R_X(k) = \begin{cases} c, & k = 0 \\ 0, & k = \dots, -2, -1, 1, 2, \dots \end{cases} \quad (15.2.73)$$

Hence discrete-time white noise can be referred to as an uncorrelated process, because it is a sequence of uncorrelated random variables (by considering zero-mean stochastic processes). In continuous-time, we use the fact that the continuous-time Fourier transform of 1 is the Dirac-delta function, so the autocorrelation function of a continuous-time stochastic process with constant spectral density  $c$  is

$$R_X(\tau) = c\delta(\tau) \quad (15.2.74)$$

In this sense, continuous-time white noise can also be referred to as an uncorrelated process when  $\tau \neq 0$ , however note that it also has infinite variance.

### Band-Limited White Noise

A signal with infinite variance has ‘infinite power’ so this is not physically realisable. It is convenient to introduce band-limited white noise, defined by having the spectral density:

$$S_X(f) = \begin{cases} c, & |f| \leq b/2 \\ 0, & |f| > b/2 \end{cases} \quad (15.2.75)$$

Then in following the discrete-time case its autocorrelation is calculated by

$$R_X(\tau) = \int_{-b/2}^{b/2} c \cdot e^{j2\pi f \tau} df \quad (15.2.76)$$

$$= \frac{c \sin(b\pi\tau)}{\pi\tau} \quad (15.2.77)$$

We can analyse what happens as  $b \rightarrow \infty$  by considering the integral of  $R_X(\tau)$ :

$$\int_{-\infty}^{\tau} R_X(s) ds = c \int_{-\infty}^{\tau} \frac{\sin(b\pi s)}{\pi s} ds \quad (15.2.78)$$

$$= c \int_{-\infty}^{b\pi\tau} \frac{\sin u}{u/b} \frac{du}{b\pi} \quad (15.2.79)$$

$$= \frac{c}{\pi} \int_{-\infty}^{b\pi\tau} \frac{\sin u}{u} du \quad (15.2.80)$$

where we have made the substitution  $u = b\pi s$ . Then using the fact that  $\int_{-\infty}^0 \frac{\sin u}{u} du = \int_0^\infty \frac{\sin u}{u} du = \pi$ , we have the limit

$$\lim_{b \rightarrow \infty} \int_{-\infty}^{\tau} R_X(s) ds = \begin{cases} 0, & \tau < 0 \\ c, & \tau = 0 \\ 2c, & \tau > 0 \end{cases} \quad (15.2.81)$$

which is a step function, therefore the limiting case of  $R_X(\tau)$  as  $b \rightarrow \infty$  is the Dirac-delta function, the same as white noise. So formally, whenever we need to get around the fact that the spectral density of white noise is not integrable on  $(-\infty, \infty)$ , we may treat white noise as band-limited white noise (which is integrable on  $(-\infty, \infty)$ ), and then take the limit as the bandwidth  $b \rightarrow \infty$ .

### Non-Stationary White Noise

A white noise stochastic process  $X(t)$  with spectral density  $S_X(f) = c$  has autocorrelation

$$\mathbb{E}[X(t)X(t+\tau)] = \begin{cases} c\delta(t), & \tau = 0 \\ 0, & \tau \neq 0 \end{cases} \quad (15.2.82)$$

We can make a generalisation of this to stochastic processes with the autocorrelation

$$\mathbb{E}[X(t)X(t+\tau)] = \begin{cases} c(t)\delta(t), & \tau = 0 \\ 0, & \tau \neq 0 \end{cases} \quad (15.2.83)$$

where the process will be uncorrelated, however it will no longer be stationary and a spectral density is no longer defined for it. We can refer to this process as non-stationary white noise. Analogously in discrete-time, a non-stationary white noise process  $X_n$  can be defined as a process with autocorrelation

$$\mathbb{E}[X_nX_{n+k}] = \begin{cases} c_n, & k = 0 \\ 0, & k \neq 0 \end{cases} \quad (15.2.84)$$

### Brown Noise

Consider a continuous-time LTI system which is an integrator, i.e. it has a transfer function of  $G(s) = \frac{1}{s}$ . Alternatively, since the Fourier transform is equal to the two-sided Laplace transform with the substitution  $s = j2\pi f$ , and for causal systems (i.e. impulse response  $h(t) = 0$  for  $t < 0$ ) this is the same as the one-sided Laplace transform, then this LTI system has frequency response

$$H(f) = \frac{1}{j2\pi f} \quad (15.2.85)$$

Therefore for a white noise input  $X(t)$  (i.e. with ‘flat’ spectral density  $S_X(f) = c$ ), the spectral density of the output  $Y(t)$  will be

$$S_Y(f) = \left| \frac{1}{j2\pi f} \right|^2 c \quad (15.2.86)$$

$$\propto \frac{1}{f^2} \quad (15.2.87)$$

Hence brown noise (or Brownian noise) may be defined as noise with power spectral density proportional to  $\frac{1}{f^2}$ . This is also consistent with the characterisation of Brownian motion being the integral of white noise.

### Violet Noise

Consider a continuous-time LTI system which is a differentiator, i.e. it has a transfer function of  $G(s) = s$ . Its frequency response will be

$$H(f) = j2\pi f \quad (15.2.88)$$

Hence for a white noise input  $X(t)$  with spectral density  $S_X(f) = c$ , the spectral density of the output  $Y(t)$  will be

$$S_Y(f) = |j2\pi f|^2 c \quad (15.2.89)$$

$$\propto f^2 \quad (15.2.90)$$

Hence violet noise may be defined as noise with power spectral density proportional to  $f^2$ , and characterised as the derivative of white noise.

### 15.2.6 Spectral Factorisation Theorem [9]

#### Spectral Factorisation Theorem in Continuous-Time

The spectral factorisation theorem (or more generally known as the polynomial matrix spectral factorisation theorem or the Fejer-Riesz lemma in the  $1 \times 1$  case) can be applied to the spectral density. For convenience, we work with the spectral density  $S_X(\omega)$  where  $\omega$  is in units rad/s, and  $\omega = 2\pi f$ . Using the spectral factorisation theorem, we can show that a rational spectral density (that is, a spectral density that is a rational function) can be factorised into:

$$S_X(\omega) = G(s)G(-s) \quad (15.2.91)$$

where  $s = j\omega$  and  $G(s)$  is the transfer function of a non-minimum phase system (i.e. roughly speaking it will have poles and zeros in the left-half plane). It follows that  $G(-s)$  is a non-minimum phase system, as it will have poles and zeros in the right-half plane.

*Proof.* To show this theorem, we will begin with establishing some properties about  $S_X(\omega)$ . Firstly,  $S_X(\omega)$  is an even function because  $S_X(\omega) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau = S_X(-\omega)$ . Since  $S_X(\omega)$  is also a rational function (i.e. its numerator and denominator are polynomials), it can be written as

$$S_X(\omega) = c \frac{\prod_{k=1}^m (\omega^2 - z_k^2)}{\prod_{\ell=1}^n (\omega^2 - p_\ell^2)} \quad (15.2.92)$$

$$= c \frac{\prod_{k=1}^m (\omega + z_k)(\omega - z_k)}{\prod_{\ell=1}^n (\omega + p_\ell)(\omega - p_\ell)} \quad (15.2.93)$$

where  $z_k$  are complex-valued zeros and  $p_\ell$  are complex-valued poles. From this form we can check that indeed  $S_X(\omega) = S_X(-\omega)$ . A further property of  $S_X(\omega)$  is that it is non-negative (this follows from the definition  $S_X(f) = \lim_{T \rightarrow \infty} \frac{1}{2T} \mathbb{E} [\tilde{x}_T(f)^2]$ ) and it is also integrable on  $(-\infty, \infty)$  (we know this because  $\int_{-\infty}^{\infty} S_X(f) df = R_X(0)$ ). This latter fact reveals that  $m < n$  (the function does not ‘blow up’) and that there are no real poles (otherwise  $S_X(\omega) \rightarrow \infty$  as  $\omega \rightarrow p_\ell$ ). The former fact additionally means that all the factors corresponding to real zeroes  $z_k$  must have even multiplicity (i.e. the zeros appear in pairs) to prevent a sign change. For these factors, we can always factorise them as follows:

$$(\omega^2 - z_k^2)^2 = \left[ -(j\omega)^2 - z_k^2 \right]^2 \quad (15.2.94)$$

$$= \left[ (j\omega)^2 + z_k^2 \right]^2 \quad (15.2.95)$$

$$= (s^2 + z_k^2)^2 \quad (15.2.96)$$

$$= (s + jz_k)(s - jz_k) \quad (15.2.97)$$

Let  $\overline{S_X(\omega)}$  denote the complex conjugate of  $S_X(\omega)$ . As  $S_X(\omega)$  is real, then

$$S_X(\omega) = \overline{S_X(\omega)} \quad (15.2.98)$$

$$= \bar{c} \frac{\prod_{k=1}^m (\omega^2 - \bar{z}_k^2)}{\prod_{\ell=1}^n (\omega^2 - \bar{p}_\ell^2)} \quad (15.2.99)$$

$$= \bar{c} \frac{\prod_{k=1}^m (\omega + \bar{z}_k)(\omega - \bar{z}_k)}{\prod_{\ell=1}^n (\omega + \bar{p}_\ell)(\omega - \bar{p}_\ell)} \quad (15.2.100)$$

This shows that if  $z_k$  is a zero of  $S_X(\omega)$ , then  $-z_k$ ,  $\bar{z}_k$ ,  $-\bar{z}_k$  are also zeroes. For factors corresponding to purely imaginary  $z_k$ , these can be factored as

$$(\omega^2 - z_k^2) = -[(j\omega)^2 - (jz_k^2)^2] \quad (15.2.101)$$

$$= -(j\omega + jz_k)(j\omega - jz_k) \quad (15.2.102)$$

$$= -(s + jz_k)(s - jz_k) \quad (15.2.103)$$

The factors corresponding to generally complex  $z_k$  can be factored as

$$(\omega^2 - z_k^2)(\omega^2 - \bar{z}_k^2) = (\omega + z_k)(\omega - z_k)(\omega + \bar{z}_k)(\omega - \bar{z}_k) \quad (15.2.104)$$

$$= j^4 (\omega + z_k)(\omega - z_k)(\omega + \bar{z}_k)(\omega - \bar{z}_k) \quad (15.2.105)$$

$$= (j\omega + jz_k)(j\omega - jz_k)(j\omega + j\bar{z}_k)(j\omega - j\bar{z}_k) \quad (15.2.106)$$

$$= (s + jz_k)(s - jz_k)(s + j\bar{z}_k)(s - j\bar{z}_k) \quad (15.2.107)$$

$$= (s + jz_k)(s - jz_k)(s + \bar{jz}_k)(s - \bar{jz}_k) \quad (15.2.108)$$

where to see the last step, note that for  $z_k = a + bj$ , we have  $j\bar{z}_k = j(a - bj) = aj + b = j\overline{(a + bj)} = \overline{jz_k}$ . Hence overall we have shown that for any factor  $(s + jz_k)$  with root in the left-half plane, there exists a factor  $(s - jz_k)$  with root in the right-half plane. The same arguments apply to the poles of  $S_X(\omega)$ . Therefore we can write

$$S_X(\omega) = \frac{A(s)A(-s)}{B(s)B(-s)} \quad (15.2.109)$$

where we choose the polynomial  $A(s)$  to have all roots in the left-half plane, polynomial  $A(-s)$  to have all roots in the right-half plane, and similarly for polynomials  $B(s)$  and  $B(-s)$ . Then  $G(s) = \frac{A(s)}{B(s)}$  is a minimum phase system and  $G(-s) = \frac{A(-s)}{B(-s)}$  is a non-minimum phase system.  $\square$

The implication of this theorem is that if we have a process with rational spectral density  $S_X(\omega)$ , we can find a minimum phase dynamical system  $G(s)$  (which would be physically realisable) and interpret this process as having been generated by sending white noise through  $G(s)$ . This in many cases justifies modelling noise as white noise.

### Spectral Factorisation Theorem in Discrete-Time

An analogous spectral factorisation theorem exists for discrete-time processes. We work with the spectral density  $S_X(\omega)$ , where  $\omega$  is normalised frequency, which can typically be taken to be in  $-\pi \leq \omega < \pi$  or  $0 \leq \omega < 2\pi$ . Note that to convert between the  $z$ -transform and the discrete-time Fourier transform of a causal system, one may use the substitution  $z = e^{j\omega}$ , which is a point on the unit circle on the complex plane. The spectral factorisation theorem states that if  $S_X(\omega)$  is rational in  $e^{j\omega}$ , we can factorise

$$S_X(\omega) = H(e^{j\omega})H(e^{-j\omega}) \quad (15.2.110)$$

$$= H(e^{j\omega})\overline{H(e^{j\omega})} \quad (15.2.111)$$

$$= |H(e^{j\omega})|^2 \quad (15.2.112)$$

where  $H(z)$  is the transfer function of a minimum phase discrete-time system (i.e. all its poles and zeros lie on or inside the unit circle).

*Proof.* Similar ideas are used as in the continuous-time case. Since  $S_X(\omega)$  is rational in  $e^{j\omega}$ , we can write

$$S_X(\omega) = ce^{j\omega\lambda} \frac{\prod_{k=1}^m (e^{j\omega} - \alpha_k)}{\prod_{\ell=1}^n (e^{j\omega} - \beta_\ell)} \quad (15.2.113)$$

for some  $c$ ,  $\lambda$ , and  $\alpha_k$  are the zeros while  $\beta_\ell$  are poles. Note that in order for  $S_X(\omega)$  to be integrable, there can be no pole with modulus such that  $|\beta_\ell| = 1$  (for similar reasons as in the continuous-time case). For each factor  $e^{j\omega} - \alpha_k$  in the numerator we may write

$$\overline{e^{j\omega} - \alpha_k} = e^{-j\omega} - \overline{\alpha_k} \quad (15.2.114)$$

$$= e^{-j\omega} \overline{\alpha_k} (1/\overline{\alpha_k} - e^{j\omega}) \quad (15.2.115)$$

and similarly for the denominator. As  $S_X(\omega)$  is real:

$$S_X(\omega) = \overline{S_X(\omega)} \quad (15.2.116)$$

$$= \overline{ce^{-j\omega\lambda} \frac{\prod_{k=1}^m e^{-j\omega} \overline{\alpha_k} (1/\overline{\alpha_k} - e^{j\omega})}{\prod_{\ell=1}^n e^{-j\omega} \overline{\beta_\ell} (1/\overline{\beta_\ell} - e^{j\omega})}} \quad (15.2.117)$$

$$= \overline{ce^{-j\omega(\lambda+m-n)}} \frac{\prod_{k=1}^m \overline{\alpha_k} (1/\overline{\alpha_k} - e^{j\omega})}{\prod_{\ell=1}^n \overline{\beta_\ell} (1/\overline{\beta_\ell} - e^{j\omega})} \quad (15.2.118)$$

Hence for each zero  $\alpha_k$ , there exists another zero  $1/\overline{\alpha_k}$ . If  $\alpha_k$  has modulus less than 1, then  $1/\overline{\alpha_k}$  has modulus greater than 1 and vice-versa. The same properties exists for the poles. Therefore  $m$  and  $n$  must necessarily be even and we can write

$$S_X(\omega) = \overline{ce^{-j\omega(\lambda+m-n)}} \frac{\prod_{k=1}^{m/2} \overline{\alpha_k} (1/\overline{\alpha_k} - e^{j\omega}) \prod_{k=1}^{m/2} |e^{j\omega} - \alpha_k|}{\prod_{\ell=1}^{n/2} \overline{\beta_\ell} (1/\overline{\beta_\ell} - e^{j\omega}) \prod_{\ell=1}^{n/2} |e^{j\omega} - \beta_\ell|} \quad (15.2.119)$$

Taking the modulus:

$$S_X(\omega) = |S_X(\omega)| \quad (15.2.120)$$

$$= \left| \overline{ce^{-j\omega(\lambda+m-n)}} \right| \frac{\prod_{k=1}^{m/2} |\overline{\alpha_k}| \cdot |1/\overline{\alpha_k} - e^{j\omega}| \prod_{k=1}^{m/2} |e^{j\omega} - \alpha_k|}{\prod_{\ell=1}^{n/2} |\overline{\beta_\ell}| \cdot |1/\overline{\beta_\ell} - e^{j\omega}| \prod_{\ell=1}^{n/2} |e^{j\omega} - \beta_\ell|} \quad (15.2.121)$$

$$= C \frac{\prod_{k=1}^{m/2} |\overline{\alpha_k}| \cdot |1/\overline{\alpha_k} - e^{j\omega}| \prod_{k=1}^{m/2} |e^{j\omega} - \alpha_k|}{\prod_{\ell=1}^{n/2} |\overline{\beta_\ell}| \cdot |1/\overline{\beta_\ell} - e^{j\omega}| \prod_{\ell=1}^{n/2} |e^{j\omega} - \beta_\ell|} \quad (15.2.122)$$

for some  $C > 0$ . Furthermore,

$$|e^{j\omega} - \alpha_k| = \left| \overline{e^{j\omega} - \alpha_k} \right| \quad (15.2.123)$$

$$= \left| e^{-j\omega} \overline{\alpha_k} (1/\overline{\alpha_k} - e^{j\omega}) \right| \quad (15.2.124)$$

$$= \left| e^{-j\omega} \right| \cdot |\overline{\alpha_k}| \cdot \left| 1/\overline{\alpha_k} - e^{j\omega} \right| \quad (15.2.125)$$

$$= |\overline{\alpha_k}| \cdot \left| 1/\overline{\alpha_k} - e^{j\omega} \right| \quad (15.2.126)$$

So

$$S_X(\omega) = C \frac{\left( \prod_{k=1}^{m/2} |e^{j\omega} - \alpha_k| \right)^2}{\left( \prod_{\ell=1}^{n/2} |e^{j\omega} - \beta_\ell| \right)^2} \quad (15.2.127)$$

$$= C \left| \frac{\prod_{k=1}^{m/2} (e^{j\omega} - \alpha_k)}{\prod_{\ell=1}^{n/2} (e^{j\omega} - \beta_\ell)} \right|^2 \quad (15.2.128)$$

where each  $a_k$  can be chosen as either of  $\alpha_k$  or  $1/\overline{\alpha_k}$  which has modulus less than or equal to 1, and likewise for  $b_\ell$ . We have thus found a rational function

$$H(z) = \frac{A(z)}{B(z)} \quad (15.2.129)$$

$$= \sqrt{C} \frac{\prod_{k=1}^{m/2} (z - a_k)}{\prod_{\ell=1}^{n/2} (z - b_\ell)} \quad (15.2.130)$$

with poles and zeros within the unit circle. From this we can also see that if  $a_k$  is a zero, then  $\overline{a_k}$  is also a zero, and likewise for the poles. Then from the Fundamental Theorem of Algebra, the polynomials  $A(z)$  and  $B(z)$  will have real coefficients. Hence  $H(z)$  is minimum phase and physically realisable.  $\square$

The implication of the spectral factorisation theorem in discrete-time is analogous to continuous-time, except we do not need to deal with the complication of continuous-time white noise as the variance of discrete-time white noise is finite. The theorem ensures that if we have a rational spectral density  $S_X(\omega)$  of a discrete-time random sequence, then there exists a physically realisable dynamical system  $H(z)$  such that sending discrete-time white noise through that system will generate the random sequence. This justifies modelling discrete-time noise as white noise.

### 15.2.7 Wold Decomposition Theorem [197]

### 15.2.8 Discretisation of Continuous-Time Stochastic Systems

## 15.3 Spectral Density Estimation [194]

### 15.3.1 Periodograms

### 15.3.2 Correlograms

### 15.3.3 Whittle Likelihood

### 15.3.4 Maximum Entropy Spectral Density Estimation [47]

## 15.4 Linear Filtering

### 15.4.1 Linear Prediction Filter [219]

Let  $\mathbf{X}_n = (X_{n-N+1}, \dots, X_n)$  denote the vector of the past  $N$  observations of a random sequence at time  $n$ . The objective is to design an  $N^{\text{th}}$  order linear finite impulse response filter that predicts the value  $m^{\text{th}}$ -step ahead in the sequence,  $X_{n+m}$ . Recall that because we can represent the output of an FIR filter via matrix multiplication, the output of the prediction filter can be written as

$$\hat{X}_{n+m} = \mathbf{a}^\top \mathbf{X}_n \quad (15.4.1)$$

where  $\mathbf{a}$  is a weighting vector which consists of the filter finite impulse response, but in reverse time (due to the nature of convolution). Suppose the sequence  $X_k$  is wide sense stationary with autocorrelation function  $R_X(m)$ . Denote the prediction error by  $\varepsilon = X_{n+m} - \hat{X}_{n+m}$ . We know that the minimum mean square error prediction will satisfy the orthogonality condition  $\mathbb{E}[\mathbf{X}_n \varepsilon] = 0$ . This can be used to derive the  $\mathbf{a}$  which yields the MMSE:

$$0 = \mathbb{E} \left[ \mathbf{X}_n \left( X_{n+m} - \hat{X}_{n+m} \right) \right] \quad (15.4.2)$$

$$= \mathbb{E} \left[ \mathbf{X}_n X_{n+m} - \mathbf{X}_n \mathbf{X}_n^\top \mathbf{a} \right] \quad (15.4.3)$$

$$= \mathbb{E}[\mathbf{X}_n X_{n+m}] - \mathbb{E}[\mathbf{X}_n \mathbf{X}_n^\top] \mathbf{a} \quad (15.4.4)$$

Hence

$$\mathbf{a} = \mathbf{R}_{\mathbf{X}_n}^{-1} \mathbf{R}_{\mathbf{X}_n X_{n+m}} \quad (15.4.5)$$

where

$$\mathbf{R}_{\mathbf{X}_n} = \begin{bmatrix} \mathbb{E}[X_{n-N+1}^2] & \dots & \mathbb{E}[X_{n-N+1} X_n] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_n X_{n-N+1}] & \dots & \mathbb{E}[X_n^2] \end{bmatrix} \quad (15.4.6)$$

$$= \begin{bmatrix} R_X(0) & \dots & R_X(N) \\ \vdots & \ddots & \vdots \\ R_X(N) & \dots & R_X(0) \end{bmatrix} \quad (15.4.7)$$

and

$$\mathbf{R}_{\mathbf{X}_n X_{n+m}} = \mathbb{E} \left[ \begin{bmatrix} X_{n-N+1} \\ \vdots \\ X_n \end{bmatrix} X_{n+m} \right] \quad (15.4.8)$$

$$= \begin{bmatrix} R_X(m+N-1) \\ \vdots \\ R_X(m) \end{bmatrix} \quad (15.4.9)$$

### 15.4.2 Matched Filtering [57, 197]

Suppose there is a ‘template’ of a signal that we wish to detect the presence of. This template is deterministic, which we represent using  $N$  values in discrete time:  $\mathbf{s} = (s_1, \dots, s_N) \neq \mathbf{0}$ . We observe the discrete-time process:

$$X_k = Y_k + W_k \quad (15.4.10)$$

where  $W_k$  is zero-mean wide sense stationary noise with autocorrelation function  $R_W(m)$ , and  $Y_k$  is the signal process, whereby if the template were to appear, then it would appear in  $Y_k$  (and for all other times, we can take  $Y_k = 0$ ). The goal of a matched filter is to design an  $N^{\text{th}}$  order linear FIR filter on  $X_k$  that ‘helps’ us detect  $\mathbf{s}$ . Let  $\mathbf{a}$  be a vector of the filter coefficients. At time  $n$ , the filter output is given by

$$\mathbf{a}^\top \mathbf{X}_n = \mathbf{a}^\top \mathbf{Y}_n + \mathbf{a}^\top \mathbf{W}_n \quad (15.4.11)$$

where

$$\mathbf{X}_n = [X_{n-N+1} \ \dots \ X_n]^\top \quad (15.4.12)$$

$$\mathbf{Y}_n = [Y_{n-N+1} \ \dots \ Y_n]^\top \quad (15.4.13)$$

$$\mathbf{W}_n = [W_{n-N+1} \ \dots \ W_n]^\top \quad (15.4.14)$$

denote vectors of the past  $N$  observations of the respective signals. The idea is that since  $W_k$  is zero-mean and  $Y_k$  is otherwise zero, then the filtered signal  $\mathbf{a}^\top \mathbf{X}_n$  should be ‘close’ to zero, apart from the time when the template  $\mathbf{s}$  is occurring or has occurred. However, the problem is that random noise can cause us to believe the signal has occurred, when it is actually a false alarm. Thus, we want to design the filter  $\mathbf{a}$  in such a way that the filtered noise  $\mathbf{a}^\top \mathbf{W}_n$  is ‘small’

while the filtered signal  $\mathbf{a}^\top \mathbf{Y}_n$  is ‘large’ directly after the presence of the signal. To make this more precise, introduce the signal-to-noise ratio

$$\text{SNR} = \frac{(\mathbf{a}^\top \mathbf{s})^2}{\mathbb{E}[(\mathbf{a}^\top \mathbf{W}_n)^2]} \quad (15.4.15)$$

which is the ratio of the signal power  $(\mathbf{a}^\top \mathbf{s})^2$  to the noise power  $\mathbb{E}[(\mathbf{a}^\top \mathbf{W}_n)^2]$  in the filtered signal. The intuition is that the larger  $(\mathbf{a}^\top \mathbf{s})^2$  is, the easier it is to detect the signal right after it occurs. On the other hand, the smaller  $\mathbb{E}[(\mathbf{a}^\top \mathbf{W}_n)^2]$  is (corresponding to the variance of the filtered noise), the harder it is for the noise to cause false alarms. Therefore we should choose the filter coefficients to be

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmax}} \text{SNR} \quad (15.4.16)$$

$$= \underset{\mathbf{a}}{\operatorname{argmax}} \frac{\mathbf{a}^\top \mathbf{s} \mathbf{s}^\top \mathbf{a}}{\mathbf{a}^\top \mathbb{E}[\mathbf{W}_n \mathbf{W}_n^\top] \mathbf{a}} \quad (15.4.17)$$

$$= \underset{\mathbf{a}}{\operatorname{argmax}} \frac{\mathbf{a}^\top \mathbf{s} \mathbf{s}^\top \mathbf{a}}{\mathbf{a}^\top \mathbf{R}_W \mathbf{a}} \quad (15.4.18)$$

where the matrix

$$\mathbf{R}_W = \begin{bmatrix} R_W(0) & \dots & R_W(N-1) \\ \vdots & \ddots & \vdots \\ R_W(N-1) & \dots & R_W(0) \end{bmatrix} \quad (15.4.19)$$

is constant since  $W_k$  is wide sense stationary. Recognise that this optimisation problem is the same as the generalised eigenvalue problem as encountered in Fisher’s linear discriminant and canonical correlation analysis. Hence the solution takes the property

$$\mathbf{R}_W^{-1} \mathbf{s} \mathbf{s}^\top \mathbf{a}^* = \lambda \mathbf{a}^* \quad (15.4.20)$$

where  $\lambda$  is the largest eigenvalue of  $\mathbf{R}_W^{-1} \mathbf{s} \mathbf{s}^\top$ , and  $\mathbf{a}^*$  is the corresponding eigenvector. Note that in this case, there will be exactly  $N - 1$  eigenvalues that are zero, and one remaining eigenvalue that is non-zero, since  $\mathbf{s} \mathbf{s}^\top$  is of rank one. Using analogous arguments to Fisher’s linear discriminant, we can then take  $\mathbf{a}^*$  to be a vector that is proportional to  $\mathbf{R}_W^{-1} \mathbf{s}$ , i.e.

$$\mathbf{a}^* \propto \mathbf{R}_W^{-1} \mathbf{s} \quad (15.4.21)$$

To obtain a unique solution, we can impose a constraint on  $\mathbf{a}$ . To implement a detection rule for  $\mathbf{s}$ , we would compare the filtered signal  $\mathbf{X}_n^\top \mathbf{a}^*$  against a threshold value (where the tradeoff is that a smaller threshold value increases the probability of false alarm, but reduces the probability of miss).

### Matched Filtering for Random Signals

A matched filter can also be applied to detect the presence of a signal when the template is no longer deterministic. Suppose the template is described by a wide sense stationary signal  $Y_k$  with autocorrelation function  $R_Y(m)$ . The signal we observe is defined by

$$X_k = Y_k \cdot \mathbb{I}_k + W_k \quad (15.4.22)$$

where  $\mathbb{I}_k$  is an indicator for whether the signal is present. The signal-to-noise ratio now becomes

$$\text{SNR} = \frac{\mathbf{a}^\top \mathbb{E}[\mathbf{Y}_n \mathbf{Y}_n^\top] \mathbf{a}}{\mathbf{a}^\top \mathbb{E}[\mathbf{W}_n \mathbf{W}_n^\top] \mathbf{a}} \quad (15.4.23)$$

$$= \frac{\mathbf{a}^\top \mathbf{R}_Y \mathbf{a}}{\mathbf{a}^\top \mathbf{R}_W \mathbf{a}} \quad (15.4.24)$$

where

$$\mathbf{R}_Y = \begin{bmatrix} R_Y(0) & \dots & R_Y(N-1) \\ \vdots & \ddots & \vdots \\ R_Y(N-1) & \dots & R_Y(0) \end{bmatrix} \quad (15.4.25)$$

The generalised eigenvalue problem now resolves to  $\mathbf{a}^*$  becoming the eigenvector corresponding to the largest eigenvalue of  $\mathbf{R}_W^{-1} \mathbf{R}_Y$ .

### 15.4.3 Wiener-Kolmogorov Filtering

#### Finite Impulse Response Wiener Filter

Let  $Y_k$  be a discrete-time signal that we wish to estimate, but suppose we have noisy observations  $X_k$  of the form

$$X_k = Y_k + W_k \quad (15.4.26)$$

where the noise process  $W_k$  is orthogonal to  $Y_k$  (e.g.  $W_k$  could be zero-mean and independent of  $Y_k$ ). Using the past  $N$  observations at time  $n$ , denoted  $\mathbf{X}_n = (X_{n-N+1}, \dots, X_n)$ , we wish design an optimal linear estimator for  $Y_n$ , denoted  $\hat{Y}_n$ . Analogous to the linear prediction filter, this amounts to designing an  $N^{\text{th}}$  order finite impulse response filter on  $X_k$  so that its output  $\hat{Y}_k$  estimates  $Y_k$  with minimum mean square error. As the estimator is linear, we express it as

$$\hat{Y}_n = \mathbf{a}^\top \mathbf{X}_n \quad (15.4.27)$$

Due to the projection theorem, and analogous to the linear prediction filter, the MMSE weighting vector for  $\mathbf{a}$  can be found to satisfy

$$\mathbb{E}[\mathbf{X}_n Y_n] - \mathbb{E}[\mathbf{X}_n \mathbf{X}_n^\top] \mathbf{a} = \mathbf{0} \quad (15.4.28)$$

These are known as the *Wiener-Hopf* equations. Denote  $\mathbf{Y}_n$  and  $\mathbf{W}_n$  as

$$\mathbf{Y}_n = [Y_{n-N+1} \ \dots \ Y_n] \quad (15.4.29)$$

$$\mathbf{W}_n = [W_{n-N+1} \ \dots \ W_n] \quad (15.4.30)$$

respectively. Then we can write the Wiener-Hopf equations as

$$\mathbb{E}[\mathbf{Y}_n Y_n + \mathbf{W}_n Y_n] - \mathbb{E}[(\mathbf{Y}_n + \mathbf{W}_n)(\mathbf{Y}_n + \mathbf{W}_n)^\top] \mathbf{a} = \mathbf{0} \quad (15.4.31)$$

Since  $W_k$  is orthogonal to  $Y_k$ , this simplifies to

$$\mathbb{E}[\mathbf{Y}_n Y_n] - (\mathbb{E}[\mathbf{Y}_n \mathbf{Y}_n^\top] + \mathbb{E}[\mathbf{W}_n \mathbf{W}_n^\top]) \mathbf{a} = \mathbf{0} \quad (15.4.32)$$

Assuming  $Y_k$  and  $W_k$  are both wide sense stationary with autocorrelation functions  $R_Y(m)$  and  $R_W(m)$  respectively, this gives the solution

$$\mathbf{a} = (\mathbf{R}_{\mathbf{Y}_n} + \mathbf{R}_{\mathbf{W}_n})^{-1} \mathbf{R}_{\mathbf{Y}_n Y_n} \quad (15.4.33)$$

where

$$\mathbf{R}_{\mathbf{Y}_n} = \begin{bmatrix} R_Y(0) & \dots & R_Y(N-1) \\ \vdots & \ddots & \vdots \\ R_Y(N-1) & \dots & R_Y(0) \end{bmatrix} \quad (15.4.34)$$

$$\mathbf{R}_{\mathbf{W}_n} = \begin{bmatrix} R_W(0) & \dots & R_W(N-1) \\ \vdots & \ddots & \vdots \\ R_W(N-1) & \dots & R_W(0) \end{bmatrix} \quad (15.4.35)$$

and

$$\mathbf{R}_{\mathbf{Y}_n Y_n} = \begin{bmatrix} R_Y(N-1) \\ \vdots \\ R_Y(0) \end{bmatrix} \quad (15.4.36)$$

Furthermore, note that if  $W_k$  is white noise, then  $\mathbf{R}_{\mathbf{W}_n}$  becomes a diagonal matrix.

### Infinite Impulse Response Discrete-Time Non-Causal Wiener Filter [85]

With the same problem setup as the FIR Wiener filter (estimating  $Y_n$  from measurements  $X_n$ ), we derive a MMSE filter which has an infinite impulse response. Denote the estimation error at time  $n$  as

$$\varepsilon_n = Y_n - \hat{Y}_n \quad (15.4.37)$$

$$= Y_n - \sum_{j=-\infty}^{\infty} h_j X_{n-j} \quad (15.4.38)$$

where  $h_j$  is the impulse response of the filter, and the estimate  $\hat{Y}_n$  is the convolution of the signal  $X_n$  with the filter. To minimise  $\mathbb{E}[\varepsilon_n^2]$ , we differentiate it with respect to each filter coefficient. For the  $k^{\text{th}}$  coefficient,

$$\frac{\partial \mathbb{E}[\varepsilon_n^2]}{\partial h_k} = \mathbb{E}\left[2\varepsilon_n \cdot \frac{\partial \varepsilon_n}{\partial h_k}\right] \quad (15.4.39)$$

$$= \mathbb{E}[-2\varepsilon_n X_{n-k}] \quad (15.4.40)$$

Setting the derivative to zero, the solution satisfies  $\mathbb{E}[\varepsilon_n X_{n-k}] = 0$  for all  $-\infty < k < \infty$ . This is essentially the [projection theorem](#), except we are now performing infinite-dimensional optimisation because the filter impulse response is allowed to be infinite-duration. Putting the definition of the estimation error:

$$\mathbb{E}\left[\left(Y_n - \hat{Y}_n\right) X_{n-k}\right] = 0 \quad (15.4.41)$$

$$\mathbb{E}[Y_n X_{n-k}] = \mathbb{E}\left[\hat{Y}_n X_{n-k}\right] \quad (15.4.42)$$

In the right-hand side, in terms of the optimal filter coefficients  $h_j^*$ , we have

$$\mathbb{E}\left[\hat{Y}_n X_{n-k}\right] = \mathbb{E}\left[\sum_{j=-\infty}^{\infty} h_j^* X_{n-j} X_{n-k}\right] \quad (15.4.43)$$

$$= \sum_{j=-\infty}^{\infty} h_j^* \mathbb{E}[X_{n-j} X_{n-k}] \quad (15.4.44)$$

Thus if  $X_n$  and  $Y_n$  are jointly wide sense stationary, the condition becomes

$$R_{XY}(k) = \sum_{j=-\infty}^{\infty} h_j^* R_X(k-j) \quad (15.4.45)$$

since  $\mathbb{E}[Y_n X_{n-k}] = R_{YX}(-k) = R_{XY}(k)$ . From this, we see that  $R_{XY}(k)$  is the convolution of the filter with  $R_X(k)$ . Taking the discrete-time transform of both sides, this can be expressed in frequency domain as

$$H(\phi) = \frac{S_{XY}(\phi)}{S_X(\phi)} \quad (15.4.46)$$

where  $H(\phi)$  is the frequency response of the optimal filter, and  $S_{XY}(\phi)$ ,  $S_X(\phi)$  are the respective spectral densities. Thus, solution to the Wiener filter can be expressed in frequency domain, in terms of the properties of  $X_n$  and  $Y_n$ . Furthermore, if we assume the additive noise process  $X_n = Y_n + W_n$  where  $Y_n$  and  $W_n$  are orthogonal, then we have

$$R_X(k) = \mathbb{E}[X_n X_{n+k}] \quad (15.4.47)$$

$$= \mathbb{E}[(Y_n + W_n)(Y_{n+k} + W_{n+k})] \quad (15.4.48)$$

$$= \mathbb{E}[Y_n Y_{n+k}] + \mathbb{E}[W_n W_{n+k}] \quad (15.4.49)$$

$$= R_Y(k) + R_W(k) \quad (15.4.50)$$

and

$$R_{XY}(k) = \mathbb{E}[X_n Y_{n+k}] \quad (15.4.51)$$

$$= \mathbb{E}[Y_n Y_{n+k} + W_n Y_{n+k}] \quad (15.4.52)$$

$$= R_Y(k) \quad (15.4.53)$$

The filter is generally a non-causal filter because the estimate  $\hat{Y}_n$  at time  $n$  requires information from the future. Therefore it cannot be implemented online, however it can be implemented on a batch of data to estimate  $Y_n$  in retrospect.

### Infinite Impulse Discrete-Time Causal Wiener Filter [85]

A causal solution for the Wiener filter (thus implementable online) can be derived. Initially following the same steps as for the non-causal case, we end up with the Wiener-Hopf equations

$$R_{XY}(k) = \sum_{j=0}^{\infty} h_j^* R_X(k-j) \quad (15.4.54)$$

for all  $0 \leq k < \infty$ , since a causal solution requires  $h_k = 0$  for all  $k \leq 0$ . Assume that  $X_n = \epsilon_n$  is white noise with unit variance. This is allowed by the spectral factorisation theorem, since white noise can be characterised by passing non-white noise through a *whitening filter*. Since  $X_n$  is white noise, then  $R_X(k) = \delta(k)$  is the Kronecker delta function and therefore

$$h_k^* = R_{\epsilon Y}(k) u(k) \quad (15.4.55)$$

where  $u(k)$  is the unit step function. In frequency domain, the filter looks like

$$H(z) = S_{\epsilon Y}^+(z) \quad (15.4.56)$$

where  $S_{\epsilon Y}^+(z)$  denotes the  $z$ -transform of  $R_{\epsilon Y}(k) u(k)$ , or explicitly:

$$S_{\epsilon Y}^+(z) = \mathcal{Z}[\mathcal{Z}^{-1}[S_{\epsilon Y}(z)] u(k)] \quad (15.4.57)$$

where we can refer to  $S_{\epsilon Y}^+(z)$  as the ‘causal component’ of  $S_{\epsilon Y}(z)$ . Now assume  $X_n$  is non-white, and more precisely has a rational spectral density (i.e. the spectral density can be written as a ratio of polynomials in  $z$ -domain). By the spectral factorisation theorem, we can factorise  $S_X(z)$  as

$$S_X(z) = Q(z) Q(1/z) \quad (15.4.58)$$

where  $Q(z)$  is minimum phase. Define the pre-whitening filter  $F(z) = \frac{1}{Q(z)}$ . We can show that the output of  $X_n$  through this filter is white noise because, its spectral density is given by

$$S_\epsilon(z) = S_X(z)|F(z)|^2 \quad (15.4.59)$$

$$= S_X(z)F(z)\overline{F(z)} \quad (15.4.60)$$

$$= S_X(z)F(z)F(1/z) \quad (15.4.61)$$

$$= Q(z)Q(1/z) \cdot \frac{1}{Q(z)} \cdot \frac{1}{Q(1/z)} \quad (15.4.62)$$

$$= 1 \quad (15.4.63)$$

which is flat (i.e. a constant). Note that we have used  $\overline{F(z)} = \overline{F(e^{j\omega})} = F(e^{-j\omega}) = F(1/z)$ . Therefore the Wiener filter can be formed by cascading a whitening filter with the optimal filter for white noise, i.e.

$$H(z) = F(z)S_{\epsilon Y}^+(z) \quad (15.4.64)$$

To determine  $S_{\epsilon Y}(z)$  in term of  $S_{XY}(z)$ , begin with

$$R_{\epsilon Y}(k) = \mathbb{E}[\epsilon_n Y_{n+k}] \quad (15.4.65)$$

$$= \mathbb{E}\left[\sum_{j=-\infty}^{\infty} f_j X_{n-j} Y_{n+k}\right] \quad (15.4.66)$$

since by convolution with the whitening filter,  $\epsilon_n = \sum_{j=-\infty}^{\infty} f_j X_{n-j}$  where  $f_j$  is the impulse response of  $F(z)$ . Continuing,

$$R_{\epsilon Y}(k) = \sum_{j=-\infty}^{\infty} f_j \mathbb{E}[X_{n-j} Y_{n+k}] \quad (15.4.67)$$

$$= \sum_{j=-\infty}^{\infty} f_j R_{XY}(k+j) \quad (15.4.68)$$

$$= \sum_{\ell=-\infty}^{\infty} f_{-\ell} R_{XY}(k-\ell) \quad (15.4.69)$$

by a change of variables  $\ell = -j$ . Taking the  $z$  transform of both sides, we have

$$S_{\epsilon Y}(z) = F(1/z)S_{XY}(z) \quad (15.4.70)$$

$$= \frac{S_{XY}(z)}{Q(1/z)} \quad (15.4.71)$$

where the  $\mathcal{Z}[f_{-\ell}] = F(1/z)$  due to the time reversal property of the  $z$ -transform. The explicit solution for the Wiener filter in frequency domain is given by

$$H(z) = \frac{1}{Q(z)}\mathcal{Z}\left[\mathcal{Z}^{-1}\left[\frac{S_{XY}(z)}{Q(1/z)}\right]u(k)\right] \quad (15.4.72)$$

### Continuous-Time Non-Causal Wiener Filter

Consider the continuous-time process

$$X(t) = Y(t) + W(t) \quad (15.4.73)$$

where  $X(t)$  is observed,  $W(t)$  is noise and  $Y(t)$  is the signal we wish to estimate. A minimum mean square error filter on  $X(t)$  can be designed to estimate  $Y(t)$ . Analogous to the discrete-time case, a non-causal solution is expressed in frequency domain (via Laplace transforms) by

$$H(s) = \frac{S_{XY}(s)}{S_X(s)} \quad (15.4.74)$$

where  $S_X(s)$  and  $S_{XY}(s)$  are the respective spectral and cross-spectral densities.

### Continuous-Time Causal Wiener Filter

Analogous to the discrete-time case, a solution exists for a causal filter. Under the appropriate conditions for the [spectral factorisation theorem](#), we can decompose the spectral density of  $X(t)$  as

$$S_X(s) = Q(s)Q(-s) \quad (15.4.75)$$

where  $G(s)$  is minimum phase. Then the explicit solution for the filter in frequency domain is given by

$$H(s) = \frac{1}{Q(s)} \mathcal{L} \left[ \mathcal{L}^{-1} \left[ \frac{S_{XY}(s)}{Q(-s)} \right] u(t) \right] \quad (15.4.76)$$

where  $u(t)$  denotes the unit step function.

#### 15.4.4 Recursive Least Squares Filter [85, 189]

Let  $x_k$  and  $y_k$  be discrete-time signals, which could be deterministic. Suppose the signal  $x_k$  is filtered through a causal linear filter and noise  $v_k$  is added, for which we observe the output  $y_k$ . Explicitly,

$$y_k = \sum_{j=0}^{\infty} b_j x_{k-j} + v_k \quad (15.4.77)$$

where the  $b_j$  are the values of the filter impulse response (which does not necessarily need to be infinite-duration). We wish to design an  $N^{\text{th}}$  order finite impulse response filter on the measurement sequence  $x_k$  that emulates  $y_k$ , without knowing  $b_k$  or any specific details about  $v_k$ . Like with the [linear prediction filter](#), we can express the estimate at time  $n$  as

$$\hat{y}_n = \hat{\mathbf{b}}^\top \mathbf{x}_n \quad (15.4.78)$$

where  $\mathbf{x}_n = [x_n \dots x_{n-N+1}]^\top$  is a vector of the last  $N$  observations. Note that  $\hat{\mathbf{b}}$  now contains the values of the designed filter's impulse response (in chronological order, since the ordering in  $\mathbf{x}_n$  is reversed). For the recursive least squares filter, the approach is to combine elements of both the [weighted least squares](#) and [recursive least squares](#) estimators. More precisely, suppose the weighting matrix  $\mathbf{W}$  on the first  $n+1$  observations (beginning from  $x_0$ ) is given by

$$\mathbf{W} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & \lambda^n \end{bmatrix} \quad (15.4.79)$$

for some  $0 < \lambda \leq 1$ . That is, we weight the  $j^{\text{th}}$  most recent observation by  $\lambda^{j-1}$ , where  $\lambda$  is known as a ‘forgetting factor’, and can also be thought of as a ‘discounting factor’. Then using the weighted least squares estimator in the derivation of the recursive least squares estimator, we see that (with analogous notation):

$$\mathbf{R}_n = \mathbf{X}_n^\top \mathbf{W} \mathbf{X}_n \quad (15.4.80)$$

$$= \sum_{k=0}^n \lambda^{n-k} \mathbf{x}_n \mathbf{x}_n^\top \quad (15.4.81)$$

Hence  $\mathbf{R}_{n+1}$  is related to  $\mathbf{R}_n$  by:

$$\mathbf{R}_{n+1} = \sum_{k=0}^{n+1} \lambda^{n+1-k} \mathbf{x}_n \mathbf{x}_n^\top \quad (15.4.82)$$

$$= \lambda \sum_{k=0}^n \lambda^{n-k} \mathbf{x}_n \mathbf{x}_n^\top + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \quad (15.4.83)$$

$$= \lambda \mathbf{R}_n + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \quad (15.4.84)$$

Applying the matrix inversion lemma in a similar way (with  $B = (\lambda \mathbf{R}_n)^{-1}$ ), we get

$$\mathbf{R}_{n+1}^{-1} = \frac{\mathbf{R}_n^{-1}}{\lambda} - \frac{\mathbf{R}_n^{-1} \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1}}{\lambda^2} \left( 1 + \frac{\mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1} \mathbf{x}_{n+1}}{\lambda} \right)^{-1} \quad (15.4.85)$$

$$= \frac{\mathbf{R}_n^{-1}}{\lambda} - \frac{\mathbf{R}_n^{-1} \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1}}{\lambda^2 (1 + \mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1} \mathbf{x}_{n+1} / \lambda)} \quad (15.4.86)$$

$$= \lambda^{-1} \left( \mathbf{R}_n^{-1} - \frac{\mathbf{R}_n^{-1} \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1}}{\lambda + \mathbf{x}_{n+1}^\top \mathbf{R}_n^{-1} \mathbf{x}_{n+1}} \right) \quad (15.4.87)$$

Recall we can put  $\mathbf{P}_n = \mathbf{R}_n^{-1}$  and get

$$\mathbf{P}_{n+1} = \lambda^{-1} \left( \mathbf{P}_n - \frac{\mathbf{P}_n \mathbf{x}_{n+1} \mathbf{x}_{n+1}^\top \mathbf{P}_n}{\lambda + \mathbf{x}_{n+1}^\top \mathbf{P}_n \mathbf{x}_{n+1}} \right) \quad (15.4.88)$$

Hence the gain  $\mathbf{K}_{n+1}$  is given by

$$\mathbf{K}_{n+1} = \mathbf{P}_{n+1} \mathbf{x}_{n+1} \quad (15.4.89)$$

And the recursive update for  $\hat{\mathbf{b}}$  at time  $n+1$  is

$$\hat{\mathbf{b}}_{n+1} = \hat{\mathbf{b}}_n + \mathbf{K}_{n+1} (y_{n+1} - \mathbf{x}_{n+1}^\top \hat{\mathbf{b}}_n) \quad (15.4.90)$$

$$= \hat{\mathbf{b}}_n + \mathbf{P}_{n+1} \mathbf{x}_{n+1} (y_{n+1} - \mathbf{x}_{n+1}^\top \hat{\mathbf{b}}_n) \quad (15.4.91)$$

Note that there does not need to be any element of  $\mathbf{W}$  between  $\mathbf{x}_{n+1}$  and  $y_{n+1}$ , because the most recent observation is not discounted. Also, we do not require the true underlying filter/system to necessarily be time-invariant, as the recursive least squares filter can ‘adapt’ to changes of the system over time.

### 15.4.5 Least Mean Squares Filter [85]

Consider the same setup as the recursive least squares filter, except the signals  $x_n$  and  $y_n$  could also be non-stationary. We wish to design an FIR filter with coefficients  $\hat{\mathbf{b}}_n$ , that approximates the true filter on  $x_n$ , so that we can estimate  $y_n$  by  $\hat{y}_n = \hat{\mathbf{b}}_n^\top \mathbf{x}_n$ . Denoting the estimation error  $\varepsilon_n = y_n - \hat{y}_n$  and using the mean squared error criterion  $\mathbb{E} \left[ \frac{1}{2} \varepsilon_n^2 \right]$ , the gradient with respect to  $\hat{\mathbf{b}}_n$  is

$$\nabla_{\hat{\mathbf{b}}_n} \mathbb{E} \left[ \frac{1}{2} \varepsilon_n^2 \right] = \mathbb{E} \left[ \varepsilon_n \nabla_{\hat{\mathbf{b}}_n} \varepsilon_n \right] \quad (15.4.92)$$

$$= -\mathbb{E} [\varepsilon_n \mathbf{x}_n] \quad (15.4.93)$$

As the signals are potentially non-stationary, an adaptive approach for computing  $\hat{\mathbf{b}}_n$  is via gradient descent updates:

$$\hat{\mathbf{b}}_{n+1} = \hat{\mathbf{b}}_n - \mu \nabla_{\hat{\mathbf{b}}_n} \mathbb{E} \left[ \frac{1}{2} \varepsilon_n^2 \right] \quad (15.4.94)$$

$$= \hat{\mathbf{b}}_n + \mu \mathbb{E} [\varepsilon_n \mathbf{x}_n] \quad (15.4.95)$$

where  $\mu > 0$  is the step size. As the quantity  $\mathbb{E} [\varepsilon_n \mathbf{x}_n]$  may not be known, we can instead take a stochastic approximation approach (or more precisely, stochastic gradient descent) and replace  $\mathbb{E} [\varepsilon_n \mathbf{x}_n]$  by the instantaneous quantity  $\varepsilon_n \mathbf{x}_n$ . This yields the least mean squares filter update rule:

$$\hat{\mathbf{b}}_{n+1} = \hat{\mathbf{b}}_n + \mu \varepsilon_n \mathbf{x}_n \quad (15.4.96)$$

### 15.4.6 Deconvolution

#### Wiener Deconvolution

## 15.5 Kalman Filtering

### 15.5.1 Linear Kalman Filter

The Kalman filter can be considered a variant of the Bayes filter when the dynamic model is linear and the noise is white, and the filter returns only the posterior mean and covariance. Additionally, if the noise is Gaussian, then the exact posterior distribution is specified by this posterior mean and covariance. We consider the discrete-time stochastic dynamical system

$$x_{k+1} = Ax_k + Bu_k + Ew_k \quad (15.5.1)$$

where the state to be estimated is  $x \in \mathbb{R}^n$  and the known input is  $u \in \mathbb{R}^m$ . This means that the matrices have dimensions  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times m}$ . Assume that  $w_k \in \mathbb{R}^d$  is a vector of independent zero-mean white noise with unit variance, i.e.

$$\text{Cov}(w_k) = \mathbb{E} [w_k w_k^\top] - \mathbb{E}[w_k] \mathbb{E}[w_k^\top] \quad (15.5.2)$$

$$= \mathbb{E} [w_k w_k^\top] \quad (15.5.3)$$

and known autocorrelation

$$\mathbb{E} [w_k w_i^\top] = \begin{cases} I, & k = i \\ 0, & k \neq i \end{cases} \quad (15.5.4)$$

so that  $\text{Cov}(Ew_k) = EE^\top$  and  $\mathbb{E} [Ew_k (Ew_i)^\top] = 0$  when  $k \neq i$ . Alternatively, we can write the system as

$$x_{k+1} = Ax_k + Bu_k + w'_k \quad (15.5.5)$$

where  $w'_k$  is zero-mean and has autocorrelation

$$\mathbb{E} [w'_k w_i'^\top] = \begin{cases} Q, & k = i \\ 0, & k \neq i \end{cases} \quad (15.5.6)$$

which gives equivalence to the specification above if  $Q = EE^\top$ . The measurement process can be written as

$$z_k = Cx_k + D + Fv_k \quad (15.5.7)$$

where the measured output  $z_k \in \mathbb{R}^p$  is known and the noise is of dimension  $v_k \in \mathbb{R}^{d'}$ , matrices  $C, F$  are of appropriate dimension, and  $D$  is a vector. Note that we could instead make  $D$  a matrix have have a feedforward term  $Du_k$ , however this would then disallow the ability for  $u_k$  to depend on the state estimate (i.e. feedback). We again assume that  $v_k$  is independent zero-mean unit variance white noise with  $\text{Cov}(Fv_k) = FF^\top$ , or alternatively we can let  $R = FF^\top$  and write the process as

$$z_k = Cx_k + D + v'_k \quad (15.5.8)$$

where  $v'_k$  is zero-mean and has known autocorrelation

$$\mathbb{E}[v'_k v_i'^\top] = \begin{cases} R, & k = i \\ 0, & k \neq i \end{cases} \quad (15.5.9)$$

Additionally assume that noise processes  $w'_k$  and  $v'_k$  are uncorrelated such that  $\mathbb{E}[w'_k v_i'^\top] = 0$  for all  $k, i$ . The system begins at initial condition  $x_0$ , and assume it is known that  $\mathbb{E}[x_0] = 0$  with initial estimate  $\hat{x}_0 = 0$ , so that the initial estimation error covariance is  $P_0$ , i.e.

$$\mathbb{E}[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^\top] = \mathbb{E}[x_0 x_0^\top] \quad (15.5.10)$$

$$= P_0 \quad (15.5.11)$$

Use  $\hat{x}_{k+1}^-$  to denote the prior estimate at step  $k+1$ , which is the mean of the prior distribution. Use  $P_{k+1}^-$  to similarly denote the prior covariance. A single iteration of the Kalman filter algorithm at step  $k+1$  is given by:

1. Perform the ‘predict’ step by propagating the previous estimate according to the dynamic model by

$$\hat{x}_{k+1}^- = A\hat{x}_k + Bu_k \quad (15.5.12)$$

$$P_{k+1}^- = AP_k A^\top + Q \quad (15.5.13)$$

2. Compute the Kalman gain, which is a  $n \times p$  matrix, by

$$K_{k+1} = P_{k+1}^- C^\top \left( CP_{k+1}^- C^\top + R \right)^{-1} \quad (15.5.14)$$

3. Perform the update step to obtain the new posterior estimates:

$$\hat{x}_{k+1} = \hat{x}_{k+1}^- + K_{k+1} (z_{k+1} - C\hat{x}_{k+1}^- - D) \quad (15.5.15)$$

$$P_{k+1} = (I - K_{k+1} C) P_{k+1}^- \quad (15.5.16)$$

If noises  $w'_k$  and  $v'_k$  are Gaussian, then the posterior estimates specify the full (Gaussian) posterior distribution (since the linear operations preserve the Gaussian distributions). Note that the algorithm is also applicable to a time-varying system with known time-varying matrices  $A_k, B_k, C_k, D_k$  and known non-stationary white noise covariances  $Q_k, R_k$ .

### Unbiasedness of Kalman Filter

We prove that the Kalman filter estimate is unbiased, i.e.  $\mathbb{E}[\hat{x}_{k+1}] = \mathbb{E}[x_{k+1}]$  through induction. Beginning with the induction step, we assume that  $\mathbb{E}[\hat{x}_k] = \mathbb{E}[x_k]$ . Hence the expected prior estimate is

$$\mathbb{E}[\hat{x}_{k+1}^-] = \mathbb{E}[A\hat{x}_k + Bu_k] \quad (15.5.17)$$

$$= A\mathbb{E}[\hat{x}_k] + Bu_k \quad (15.5.18)$$

The expectation of the true state is

$$\mathbb{E}[x_{k+1}] = \mathbb{E}[Ax_k + Bu_k + w'_k] \quad (15.5.19)$$

$$= A\mathbb{E}[x_k] + Bu_k + \mathbb{E}[w'_k] \quad (15.5.20)$$

$$= A\mathbb{E}[x_k] + Bu_k \quad (15.5.21)$$

as we assumed  $\mathbb{E}[\hat{x}_k] = \mathbb{E}[x_k]$ , it then follows that  $\mathbb{E}[\hat{x}_{k+1}^-] = \mathbb{E}[x_{k+1}]$ . Then the expected posterior estimate is

$$\mathbb{E}[\hat{x}_{k+1}] = \mathbb{E}[\hat{x}_{k+1}^- + K_{k+1}(z_{k+1} - C\hat{x}_{k+1}^- - D)] \quad (15.5.22)$$

$$= \mathbb{E}[\hat{x}_{k+1}^- + K_{k+1}(Cx_{k+1} + D + v'_{k+1} - C\hat{x}_{k+1}^- - D)] \quad (15.5.23)$$

$$= \mathbb{E}[\hat{x}_{k+1}^- + K_{k+1}Cx_{k+1} + K_{k+1}v'_{k+1} - K_{k+1}C\hat{x}_{k+1}^-] \quad (15.5.24)$$

$$= \mathbb{E}[\hat{x}_{k+1}^-] + K_{k+1}C\mathbb{E}[x_{k+1}] - K_{k+1}\mathbb{E}[v'_{k+1}] - K_{k+1}C\mathbb{E}[\hat{x}_{k+1}^-] \quad (15.5.25)$$

$$= \mathbb{E}[\hat{x}_{k+1}^-] + K_{k+1}C\mathbb{E}[x_{k+1}] - K_{k+1}C\mathbb{E}[\hat{x}_{k+1}^-] \quad (15.5.26)$$

by the fact  $\mathbb{E}[v'_{k+1}] = 0$ . Then using  $\mathbb{E}[\hat{x}_{k+1}^-] = \mathbb{E}[x_{k+1}]$  we have

$$\mathbb{E}[\hat{x}_{k+1}] = \mathbb{E}[x_{k+1}] \quad (15.5.27)$$

Hence it only requires that  $\mathbb{E}[\hat{x}_0] = \mathbb{E}[x_0] = 0$  (which can be assumed true in a Bayesian setting) in order for the Kalman filter to be unbiased by induction.

### Minimum Mean Square Error of Kalman Filter

We show that the Kalman filter is the minimum mean square error estimator. Denote the prior estimation error  $e_k^- = x_k - \hat{x}_k^-$ . Since the unbiasedness means  $\mathbb{E}[e_k^-] = 0$ , the prior error covariance matrix is

$$P_k^- = \text{Cov}(e_k^-) \quad (15.5.28)$$

$$= \mathbb{E}[e_k^-(e_k^-)^\top] - \mathbb{E}[e_k^-]\mathbb{E}[e_k^-]^\top \quad (15.5.29)$$

$$= \mathbb{E}[e_k^-(e_k^-)^\top] \quad (15.5.30)$$

$$= \mathbb{E}[(x_k - \hat{x}_k^-)(x_k - \hat{x}_k^-)^\top] \quad (15.5.31)$$

Denote the posterior estimation error  $e_k = x_k - \hat{x}_k$ . Given the posterior error covariance matrix at step  $k-1$ , we can show that the formula in the propagation step is indeed correct by

$$P_k^- = \text{Cov}(e_k^-) \quad (15.5.32)$$

$$= \text{Cov}(x_k - \hat{x}_k^-) \quad (15.5.33)$$

$$= \text{Cov}(Ax_{k-1} + Bu_{k-1} + w'_k - A\hat{x}_{k-1} - Bu_{k-1}) \quad (15.5.34)$$

$$= \text{Cov}(A(x_{k-1} - \hat{x}_{k-1}) + w'_k) \quad (15.5.35)$$

$$= A \text{Cov}(e_{k-1}) A^\top + \text{Cov}(w'_k) \quad (15.5.36)$$

$$= AP_{k-1}A^\top + Q \quad (15.5.37)$$

For the posterior error covariance matrix at step  $k$ , we can rewrite this using  $x_k - \hat{x}_k = (x_k - \hat{x}_k^-) - (\hat{x}_k - \hat{x}_k^-)$  to give

$$P_k = \mathbb{E}[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^\top] \quad (15.5.38)$$

$$= \mathbb{E}\left[((x_k - \hat{x}_k^-) - (\hat{x}_k - \hat{x}_k^-))((x_k - \hat{x}_k^-) - (\hat{x}_k - \hat{x}_k^-))^\top\right] \quad (15.5.39)$$

From the update equation:

$$\hat{x}_k = \hat{x}_k^- + K_k (z_k - C\hat{x}_k^- - D) \quad (15.5.40)$$

which can be rearranged so that

$$\hat{x}_k - \hat{x}_k^- = K_k (z_k - C\hat{x}_k^- - D) \quad (15.5.41)$$

$$= K_k (Cx_k + D + v'_k - C\hat{x}_k^- - D) \quad (15.5.42)$$

$$= K_k (Cx_k + v'_k - C\hat{x}_k^-) \quad (15.5.43)$$

Hence

$$P_k = \mathbb{E} \left[ ((x_k - \hat{x}_k^-) - K_k (Cx_k + v'_k - C\hat{x}_k^-)) ((x_k - \hat{x}_k^-) - K_k (Cx_k + v'_k - C\hat{x}_k^-))^{\top} \right] \quad (15.5.44)$$

Factoring out  $(x_k - \hat{x}_k^-)$ , we see that

$$P_k = \mathbb{E} \left[ ((I - K_k C) (x_k - \hat{x}_k^-) - K_k v'_k) ((I - K_k C) (x_k - \hat{x}_k^-) - K_k v'_k)^{\top} \right] \quad (15.5.45)$$

The quadratic can be expanded into

$$\begin{aligned} P_k = \mathbb{E} & \left[ (I - K_k C) (x_k - \hat{x}_k^-) (x_k - \hat{x}_k^-)^{\top} (I - K_k C)^{\top} - (I - K_k C) (x_k - \hat{x}_k^-) v'_k K_k^{\top} \right. \\ & \left. - K_k v'_k (x_k - \hat{x}_k^-)^{\top} (I - K_k C)^{\top} + K_k v'_k v'^{\top} K_k^{\top} \right] \end{aligned} \quad (15.5.46)$$

The expectations of the cross terms evaluate to zero, since the noise and state are uncorrelated, i.e.  $\mathbb{E} [v'_k x_k^{\top}] = 0$ , as is the prior estimate and noise, i.e.  $\mathbb{E} [v'_k (\hat{x}_k^-)^{\top}] = 0$  since  $\hat{x}_k^-$  is determined using information up to and including step  $k-1$ . Therefore

$$P_k = (I - K_k C) \mathbb{E} \left[ (x_k - \hat{x}_k^-) (x_k - \hat{x}_k^-)^{\top} \right] (I - K_k C)^{\top} + K_k \mathbb{E} \left[ v'_k v'^{\top} \right] K_k^{\top} \quad (15.5.47)$$

$$= (I - K_k C) P_k^- (I - K_k C)^{\top} + K_k R K_k^{\top} \quad (15.5.48)$$

The diagonals of  $P_k$  are the expected squared estimation errors for each component of  $x_k$ . Hence to minimise the mean squared error of the estimate, we can minimise the trace of  $P_k$ . We drop the subscript in  $P_k$  for ease of notation, and write  $P$  as

$$P = (I - KC) P^- (I - KC)^{\top} + KRK^{\top} \quad (15.5.49)$$

$$= P^- - KCP^- - P^- C^{\top} K^{\top} + KCP^- C^{\top} K^{\top} + KRK^{\top} \quad (15.5.50)$$

$$= P^- - KCP^- - P^- C^{\top} K^{\top} + K(CP^- C^{\top} + R) K^{\top} \quad (15.5.51)$$

This is convex in  $K$  since  $CP^- C^{\top} + R$  is positive semi-definite. Using the properties of the trace that for arbitrary (commuting) matrices:  $\text{trace}(G + H) = \text{trace}(G) + \text{trace}(H)$ ,  $\frac{\partial \text{trace}(XH)}{\partial X} = H^{\top}$ ,  $\frac{\partial \text{trace}(HX^{\top})}{\partial X} = H$  and  $\frac{\partial \text{trace}(XHX^{\top})}{\partial X} = XH^{\top} + XH$ , we differentiate the trace of  $P$  with respect to  $K$  to obtain:

$$\frac{\partial \text{trace}(P)}{\partial K} = -P^- C^{\top} - P^- C^{\top} + K(CP^- C^{\top} + R)^{\top} + K(CP^- C^{\top} + R) \quad (15.5.52)$$

Note that  $CP^- C^{\top}$  and  $R$  are symmetric matrices so  $(CP^- C^{\top} + R)^{\top} = (CP^- C^{\top} + R)$ . Then

$$\frac{\partial \text{trace}(P)}{\partial K} = -2P^- C^{\top} + 2K(CP^- C^{\top} + R) \quad (15.5.53)$$

Setting the derivative to zero, we arrive at the Kalman gain  $K^*$ :

$$K^* (CP^- C^{\top} + R) = P^- C^{\top} \quad (15.5.54)$$

$$K^* = P^- C^{\top} (CP^- C^{\top} + R)^{-1} \quad (15.5.55)$$

## Kalman Prediction Filter [206]

Another variant of the Kalman filter predicts the state at time  $k + 1$  using observations up to time  $k$  (instead of up to time  $k + 1$ , as in the conventional Kalman filter). It is analogous to the linear prediction filter, and can be referred to as the one-step-ahead Kalman prediction filter. Consider the linear state-space system

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (15.5.56)$$

$$z_k = Cx_k + v_k \quad (15.5.57)$$

with zero-mean white noise, i.e.

$$\mathbb{E}[w_k] = 0 \quad (15.5.58)$$

$$\mathbb{E}[v_k] = 0 \quad (15.5.59)$$

and

$$\text{Cov}\left(\begin{bmatrix} w_k \\ v_k \end{bmatrix}\right) = \begin{bmatrix} \text{Cov}(w_k) & \text{Cov}(w_k, v_k) \\ \text{Cov}(v_k, w_k) & \text{Cov}(v_k) \end{bmatrix} \quad (15.5.60)$$

$$= \begin{bmatrix} \mathbb{E}[w_k w_k^\top] & \mathbb{E}[w_k v_k^\top] \\ \mathbb{E}[v_k w_k^\top] & \mathbb{E}[v_k v_k^\top] \end{bmatrix} \quad (15.5.61)$$

$$= \begin{bmatrix} Q & S \\ S^\top & R \end{bmatrix} \quad (15.5.62)$$

while

$$\text{Cov}\left(\begin{bmatrix} w_k \\ v_k \end{bmatrix}, \begin{bmatrix} w_\kappa \\ v_\kappa \end{bmatrix}\right) = 0 \quad (15.5.63)$$

for all  $k \neq \kappa$ . Note that in comparison to the conventional Kalman filter, we have allowed for some cross-covariance between  $w_k$  and  $v_k$ . We could also allow for time-varying matrices  $A_k$ ,  $B_k$ ,  $C_k$ ,  $Q_k$ ,  $R_k$ ,  $S_k$  and also a feedforward component  $D_k u_k$  in the output  $z_k$ , however we ignore these for the sake of simplicity in the notation. At time  $k$ , the prediction for the state at time  $k + 1$  is computed as follows:

1. Update the error covariance matrix by

$$P_k = AP_{k-1}A^\top + Q - (S + AP_{k-1}C^\top)(R + CP_{k-1}C^\top)^{-1}(S + AP_{k-1}C^\top) \quad (15.5.64)$$

2. Compute the Kalman gain as

$$K_k = (S + AP_kC^\top)(R + CP_kC^\top)^{-1} \quad (15.5.65)$$

3. Update the state prediction by

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k-1} + Bu_k + K_k(z_k - C\hat{x}_{k|k-1}) \quad (15.5.66)$$

We can show that this estimate is unbiased, analogously to the conventional Kalman filter. Using induction, we assume  $\mathbb{E}[\hat{x}_{k|k-1}] = \mathbb{E}[x_k]$  and take expectations of the estimate for  $k + 1$ :

$$\mathbb{E}[\hat{x}_{k+1|k}] = A\mathbb{E}[\hat{x}_{k|k-1}] + Bu_k + K_k\mathbb{E}[z_k - C\hat{x}_{k|k-1}] \quad (15.5.67)$$

$$= A\mathbb{E}[\hat{x}_{k|k-1}] + Bu_k + K_k\mathbb{E}[Cx_k + v_k - C\hat{x}_{k|k-1}] \quad (15.5.68)$$

$$= A\mathbb{E}[\hat{x}_{k|k-1}] + Bu_k + \cancel{K_k C \mathbb{E}[x_k]}^0 + \cancel{K_k \mathbb{E}[v_k]}^0 - \cancel{K_k C \mathbb{E}[\hat{x}_{k|k-1}]}^0 \quad (15.5.69)$$

$$= A\mathbb{E}[\hat{x}_{k|k-1}] + Bu_k = A\mathbb{E}[x_k] + Bu_k \quad (15.5.70)$$

On the other hand,

$$\mathbb{E}[x_{k+1}] = \mathbb{E}[Ax_k + Bu_k + w_k] \quad (15.5.71)$$

$$= A\mathbb{E}[x_k] + Bu_k + \mathbb{E}[w_k]^0 \quad (15.5.72)$$

which shows unbiasedness by  $\mathbb{E}[\hat{x}_{k+1|k}] = \mathbb{E}[x_{k+1}]$ , under assumption of the base case  $\mathbb{E}[\hat{x}_0] = \mathbb{E}[x_0]$  for the initial condition. Note that this property does not depend on the Kalman gain  $K_k$ ; for example we could set  $K_k = 0$  and obtain an unbiased estimate (essentially the same as the prior estimate in the conventional Kalman filter). However, this would not be the minimum mean square error estimate. To show that the stated choice of the Kalman gain yields the minimum mean square error estimate, we first define the prediction error

$$\epsilon_k := x_k - \hat{x}_{k|k-1} \quad (15.5.73)$$

$$= Ax_{k-1} + Bu_{k-1} + w_{k-1} - A\hat{x}_{k-1|k-2} - Bu_{k-1} - K_{k-1}(z_{k-1} - C\hat{x}_{k-1|k-2}) \quad (15.5.74)$$

$$= A(x_{k-1} - \hat{x}_{k-1|k-2}) + w_{k-1} - K_{k-1}(Cx_{k-1} + v_{k-1} - C\hat{x}_{k-1|k-2}) \quad (15.5.75)$$

$$= A\epsilon_{k-1} + w_{k-1} - K_{k-1}(C\epsilon_{k-1} + v_{k-1}) \quad (15.5.76)$$

Then the error covariance matrix for the prediction at  $k+1$  is

$$P_{k+1} = \text{Cov}(\epsilon_{k+1}) \quad (15.5.77)$$

$$= \mathbb{E}[\epsilon_{k+1}\epsilon_{k+1}^\top] \quad (15.5.78)$$

$$= \mathbb{E}[(A\epsilon_k + w_k - K_k(C\epsilon_k + v_k))(A\epsilon_k + w_k - K_k(C\epsilon_k + v_k))^\top] \quad (15.5.79)$$

$$\begin{aligned} &= A\mathbb{E}[\epsilon_k\epsilon_k^\top]A^\top - A\mathbb{E}[\epsilon_k\epsilon_k^\top]C^\top K_k^\top + \mathbb{E}[w_kw_k^\top] - \mathbb{E}[w_kv_k^\top]K_k^\top \\ &\quad - K_kC\mathbb{E}[\epsilon_k\epsilon_k^\top]A^\top - K_k\mathbb{E}[v_kw_k^\top] + K_kC\mathbb{E}[\epsilon_k\epsilon_k^\top]C^\top K_k^\top + K_k\mathbb{E}[v_kv_k^\top]K_k^\top \end{aligned} \quad (15.5.80)$$

$$\begin{aligned} &= AP_kA^\top - AP_kC^\top K_k^\top + Q - SK_k^\top \\ &\quad - K_kCP_kA^\top - K_kS^\top + K_kCP_kC^\top K_k^\top + K_kRK_k^\top \end{aligned} \quad (15.5.81)$$

Treating this as a function of an arbitrary Kalman gain  $K$ , we can follow the same steps as in the conventional Kalman filter, and differentiate the trace of  $P_{k+1}$  to obtain

$$\frac{\partial \text{trace}(P_{k+1})}{\partial K} = -2AP_kC^\top - 2S + 2K(R + CP_kC^\top) \quad (15.5.82)$$

Setting this to zero yields the optimal Kalman gain:

$$K_k(R + CP_kC^\top) = S + AP_kC^\top \quad (15.5.83)$$

$$K_k = (S + AP_kC^\top)(R + CP_kC^\top)^{-1} \quad (15.5.84)$$

Back-substituting this into the error covariance matrix, we can also confirm the update equation for  $P_{k+1}$ :

$$P_{k+1} = AP_kA^\top - (AP_kC^\top + S)K_k^\top + Q - K_k(CP_kA^\top + S^\top) + K_k(R + CP_kC^\top)K_k^\top$$

$$(15.5.85)$$

$$\begin{aligned}
&= AP_k A^\top - \left( AP_k C^\top + S \right) \left( R + CP_k C^\top \right)^{-1} \left( S + AP_k C^\top \right)^\top + Q \\
&\quad - \left( S + AP_k C^\top \right) \left( R + CP_k C^\top \right)^{-1} \left( CP_k A^\top + S^\top \right) \\
&\quad + \left( S + AP_k C^\top \right) \left( R + CP_k C^\top \right)^{-1} \left( R + CP_k C^\top \right) \left( R + CP_k C^\top \right)^{-1} \left( S + AP_k C^\top \right)^\top
\end{aligned} \tag{15.5.86}$$

$$= AP_k A^\top + Q - \left( AP_k C^\top + S \right) \left( R + CP_k C^\top \right)^{-1} \left( S + AP_k C^\top \right)^\top \tag{15.5.87}$$

### Innovation Form of State-Space Models [130]

Suppose in the Kalman prediction filter the error covariance matrix is initialised as  $P_0 = \bar{P}$ , where  $\bar{P}$  is the steady-state error covariance, which is a solution to the Riccati equation

$$\bar{P} = A\bar{P}A^\top + Q - \left( A\bar{P}C^\top + S \right) \left( R + C\bar{P}C^\top \right)^{-1} \left( S + A\bar{P}C^\top \right)^\top \tag{15.5.88}$$

i.e. essentially when the error covariance update stops changing. Then the Kalman gain will remain a constant, given by

$$K = \left( S + A\bar{P}C^\top \right) \left( R + C\bar{P}C^\top \right)^{-1} \tag{15.5.89}$$

Define the innovation as

$$e_k := z_k - C_k \hat{x}_{k|k-1} \tag{15.5.90}$$

which is so-called because it is the part of  $z_k$  which cannot be predicted from past data. Recognising that this term appears in the prediction update for  $\hat{x}_{k+1|k}$ , we have

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k-1} + Bu_k + Ke_k \tag{15.5.91}$$

and rearranging the definition of the innovation gives

$$z_k = C\hat{x}_{k|k-1} + e_k \tag{15.5.92}$$

This is known as the innovations representation of a state-space model, in state variable  $\hat{x}_{k|k-1}$  and noise  $e_k$ . Furthermore, the innovations sequence  $e_k$  is white noise. This follows from analogous arguments to the [innovations algorithm](#), since in the same way, each prediction  $\hat{x}_{k|k-1}$  can be written as a linear form in terms of the previous innovations.

### 15.5.2 Linearised Kalman Filter [105]

Approximate variants of the Kalman filter can be applied to nonlinear time-varying stochastic dynamical systems. Suppose that the state-space dynamics are given by the stochastic difference equation

$$x_{k+1} = f_k(x_k, u_k, w_k) \tag{15.5.93}$$

where  $f(\cdot, \cdot, \cdot)$  can be a general vector-valued differentiable nonlinear function in the time  $k$ , state  $x_k$ , input  $u_k$  and noise/disturbance  $w_k$ . The measurement process can also be treated as nonlinear time-varying differentiable mapping:

$$z_k = h_k(x_k, v_k) \tag{15.5.94}$$

where  $v_k$  is measurement noise. As with the Kalman filter we assume that the noises  $w_k$  and  $v_k$  are zero-mean and white (this can be relaxed to non-stationary white):

$$\mathbb{E} \left[ w_k w_i^\top \right] = \begin{cases} Q, & k = i \\ 0, & k \neq i \end{cases} \tag{15.5.95}$$

$$\mathbb{E} [v_k v_i^\top] = \begin{cases} R, & k = i \\ 0, & k \neq i \end{cases} \quad (15.5.96)$$

and uncorrelated with each other:

$$\mathbb{E} [w_k v_i^\top] = 0 \quad (15.5.97)$$

for all  $k, i$ . We also require an initial estimate  $\bar{x}_0$  and initial estimation error covariance matrix  $P_0$ . The premise of the linearised Kalman filter is to find a deterministic reference trajectory (given  $u_k$ ) with which we can linearise the dynamics about. A suitable choice are trajectories generated by the dynamics and measurement process in the absence of noise. Then the reference trajectories  $\bar{x}_k$  and  $\bar{z}_k$  evolve by the following deterministic difference equations:

$$\bar{x}_{k+1} = f_k(\bar{x}_k, u_k, 0) \quad (15.5.98)$$

$$\bar{z}_k = h_k(\bar{x}_k, 0) \quad (15.5.99)$$

This allows us to introduce a change of coordinate for the state  $\Delta x_k = x_k - \bar{x}_k$ , which has dynamics

$$\Delta x_{k+1} = x_{k+1} - \bar{x}_{k+1} \quad (15.5.100)$$

$$= f_k(x_k, u_k, w_k) - f_k(\bar{x}_k, u_k, 0) \quad (15.5.101)$$

$$= f_k(\bar{x}_k + \Delta x_k, u_k, w_k) - f_k(\bar{x}_k, u_k, 0) \quad (15.5.102)$$

These dynamics are linearised by performing a first-order Taylor expansion approximation about  $(\bar{x}_k, u_k, 0)$ :

$$f_k(\bar{x}_k + \Delta x_k, u_k, w_k) \approx f_k(\bar{x}_k, u_k, 0) + \frac{\partial f_k}{\partial x} \Big|_{(\bar{x}_k, u_k, 0)} \Delta x_k + \frac{\partial f_k}{\partial w} \Big|_{(\bar{x}_k, u_k, 0)} w_k \quad (15.5.103)$$

and this gives the approximate dynamics

$$\Delta x_{k+1} \approx \frac{\partial f_k}{\partial x} \Big|_{(\bar{x}_k, u_k, 0)} \Delta x_k + \frac{\partial f_k}{\partial w} \Big|_{(\bar{x}_k, u_k, 0)} w_k \quad (15.5.104)$$

$$= A_k \Delta x_k + E_k w_k \quad (15.5.105)$$

where we denote  $A_k = \frac{\partial f_k}{\partial x} \Big|_{(\bar{x}_k, u_k, 0)}$  and  $E_k = \frac{\partial f_k}{\partial w} \Big|_{(\bar{x}_k, u_k, 0)}$  (note that these are Jacobian matrices). Similarly, a change of coordinate for the output  $\Delta z_k = z_k - \bar{z}_k$  gives the augmented measurement process

$$\Delta z_k = z_k - \bar{z}_k \quad (15.5.106)$$

$$= h_k(x_k, v_k) - h_k(\bar{x}_k, 0) \quad (15.5.107)$$

$$= h_k(\bar{x}_k + \Delta x_k, v_k) - h_k(\bar{x}_k, 0) \quad (15.5.108)$$

Again, a first order Taylor approximation about  $(\bar{x}_k, 0)$  yields:

$$h_k(\bar{x}_k + \Delta x_k, v_k) \approx h_k(\bar{x}_k, 0) + \frac{\partial h_k}{\partial x} \Big|_{(\bar{x}_k, 0)} \Delta x_k + \frac{\partial h_k}{\partial v} \Big|_{(\bar{x}_k, 0)} v_k \quad (15.5.109)$$

giving the approximation

$$\Delta z_k \approx \frac{\partial h_k}{\partial x} \Big|_{(\bar{x}_k, 0)} \Delta x_k + \frac{\partial h_k}{\partial v} \Big|_{(\bar{x}_k, 0)} v_k \quad (15.5.110)$$

$$= C_k \Delta x_k + F_k v_k \quad (15.5.111)$$

with Jacobian matrices  $C_k = \left. \frac{\partial h_k}{\partial x} \right|_{(\bar{x}_k, 0)}$  and  $F_k = \left. \frac{\partial h_k}{\partial v} \right|_{(\bar{x}_k, 0)}$ . We can then work with and apply standard Kalman filter techniques to the linear time-varying stochastic dynamical system:

$$\Delta x_{k+1} = A_k \Delta x_k + w'_k \quad (15.5.112)$$

$$\Delta z_k = C_k \Delta x_k + v'_k \quad (15.5.113)$$

where  $w'_k = E_k w_k$  is still zero-mean and non-stationary white with covariance  $Q'_k$  given by

$$Q'_k = \text{Cov}(w'_k) \quad (15.5.114)$$

$$= E_k \text{Cov}(w_k) E_k^\top \quad (15.5.115)$$

$$= E_k Q E_k^\top \quad (15.5.116)$$

and  $v'_k = F_k v_k$  is also still zero-mean and non-stationary white with covariance  $R'_k$  given by

$$R'_k = \text{Cov}(v'_k) \quad (15.5.117)$$

$$= F_k \text{Cov}(v_k) F_k^\top \quad (15.5.118)$$

$$= F_k R F_k^\top \quad (15.5.119)$$

and  $w'_k$  will still be uncorrelated with  $v'_k$ . From an initial estimate  $\Delta \hat{x}_0 = \hat{x}_0 - \bar{x}_0$  and covariance  $P_0$ , the linearised Kalman filter algorithm at time  $k + 1$  is thus completed in a similar manner to the linear Kalman filter, summarised by:

1. Propagate the previous estimate to obtain the prior estimate and covariance:

$$\Delta \hat{x}_{k+1}^- = A_k \Delta \hat{x}_k \quad (15.5.120)$$

$$P_{k+1}^- = A_k P_k A_k^\top + Q'_k \quad (15.5.121)$$

2. Compute the Kalman gain by

$$K_{k+1} = P_{k+1}^- C_{k+1}^\top \left( C_{k+1} P_{k+1}^- C_{k+1}^\top + R'_k \right)^{-1} \quad (15.5.122)$$

3. Update the estimates using the Kalman gain to obtain the posterior estimates

$$\Delta \hat{x}_{k+1} = \Delta \hat{x}_{k+1}^- + K_{k+1} (\Delta z_{k+1} - C_{k+1} \Delta \hat{x}_{k+1}^-) \quad (15.5.123)$$

$$P_{k+1} = (I - K_{k+1} C_{k+1}) P_{k+1}^- \quad (15.5.124)$$

4. The state estimate in original coordinates can be obtained by

$$\hat{x}_{k+1} = \Delta \hat{x}_{k+1} + \bar{x}_{k+1} \quad (15.5.125)$$

### 15.5.3 Extended Kalman Filter

The extended Kalman filter (EKF) is also another variant of the Kalman filter applicable to nonlinear systems. However unlike the linearised Kalman filter, rather than linearising about a reference trajectory, a linearisation is taken about the most recent estimates. Again we can introduce the general nonlinear time-varying differentiable stochastic dynamical system in discrete-time:

$$x_{k+1} = f_k(x_k, u_k, w_k) \quad (15.5.126)$$

$$z_k = h_k(x_k, v_k) \quad (15.5.127)$$

with the same assumptions placed on noises  $w_k$ ,  $v_k$  as for the linearised Kalman filter, and the initial estimate  $\hat{x}_0$  and error covariance  $P_0$  provided. The following Jacobian matrices are defined:

$$A_k = \left. \frac{\partial f_k}{\partial x} \right|_{(\hat{x}_{k-1}, u_k, 0)} \quad (15.5.128)$$

$$E_k = \left. \frac{\partial f_k}{\partial w} \right|_{(\hat{x}_{k-1}, u_k, 0)} \quad (15.5.129)$$

$$C_k = \left. \frac{\partial h_k}{\partial x} \right|_{(\hat{x}_k^-, 0)} \quad (15.5.130)$$

$$F_k = \left. \frac{\partial h_k}{\partial v} \right|_{(\hat{x}_k^-, 0)} \quad (15.5.131)$$

Note that Jacobian matrices  $A_k$ ,  $E_k$  are linearised about the posterior estimate at the previous step, and matrices  $C_k$ ,  $F_k$  are linearised about the prior estimate at the current step. Also define the covariance matrices

$$Q'_k = E_k Q E_k^\top \quad (15.5.132)$$

$$R'_k = F_k R F_k^\top \quad (15.5.133)$$

as with the linearised Kalman filter. With these, the EKF algorithm at step  $k + 1$  is:

1. Propagate the previous estimate to obtain the prior estimate and covariance using:

$$\hat{x}_{k+1}^- = f_k(\hat{x}_k, u_k, 0) \quad (15.5.134)$$

$$P_{k+1}^- = A_k P_k A_k^\top + Q'_k \quad (15.5.135)$$

2. Compute the Kalman gain by

$$K_{k+1} = P_{k+1}^- C_{k+1}^\top \left( C_{k+1} P_{k+1}^- C_{k+1}^\top + R'_k \right)^{-1} \quad (15.5.136)$$

3. Update the prior estimates using the Kalman gain to obtain the posterior estimates

$$\hat{x}_{k+1} = \hat{x}_{k+1}^- + K_{k+1} (z_{k+1} - h_{k+1}(\hat{x}_{k+1}^-, 0)) \quad (15.5.137)$$

$$P_{k+1} = (I - K_{k+1} C_{k+1}) P_{k+1}^- \quad (15.5.138)$$

These estimates are generally only approximate, however they can be justified to be reasonable approximations. For the propagation of the state, by Gauss' approximation theorem we will get

$$\mathbb{E}[x_{k+1}] = \mathbb{E}[f_k(x_k, u_k, w_k)] \quad (15.5.139)$$

$$\approx f_k(\mathbb{E}[x_k], \mathbb{E}[u_k], \mathbb{E}[w_k]) \quad (15.5.140)$$

$$= f_k(\hat{x}_k, u_k, 0) \quad (15.5.141)$$

For the prior estimate error covariance  $P_{k+1}^-$  we define the prior error  $e_k^- = x_k - \hat{x}_k^-$  and posterior error  $e_k = x_k - \hat{x}_k$ , then

$$P_{k+1}^- = \text{Cov}(e_{k+1}^-) \quad (15.5.142)$$

$$= \text{Cov}(x_{k+1} - \hat{x}_{k+1}^-) \quad (15.5.143)$$

$$= \text{Cov}(f_k(x_k, u_k, w_k) - f_k(\hat{x}_k, u_k, 0)) \quad (15.5.144)$$

$$= \text{Cov}(f_k(\hat{x}_k + e_k, u_k, w_k) - f_k(\hat{x}_k, u_k, 0)) \quad (15.5.145)$$

Performing a Taylor approximation about  $(\hat{x}_k, u_k, 0)$ :

$$f_k(\hat{x}_k + e_k, u_k, w_k) \approx f_k(\hat{x}_k, u_k, 0) + \frac{\partial f_k}{\partial x} \Big|_{(\hat{x}_{k-1}, u_k, 0)} e_k + \frac{\partial f_k}{\partial w} \Big|_{(\hat{x}_{k-1}, u_k, 0)} w_k \quad (15.5.146)$$

$$= f_k(\hat{x}_k, u_k, 0) + A_k e_k + E_k w_k \quad (15.5.147)$$

Hence

$$P_{k+1}^- \approx \text{Cov}(A_k e_k + E_k w_k) \quad (15.5.148)$$

$$= A_k \text{Cov}(e_k) A_k^\top + E_k \text{Cov}(w_k) E_k^\top \quad (15.5.149)$$

$$= A_k P_k A_k^\top + E_k Q E_k^\top \quad (15.5.150)$$

$$= A_k P_k A_k^\top + Q'_k \quad (15.5.151)$$

As for the posterior state update, this can be rewritten as

$$\hat{x}_{k+1} - \hat{x}_{k+1}^- = K_{k+1}(z_{k+1} - h_{k+1}(\hat{x}_{k+1}^-, 0)) \quad (15.5.152)$$

$$= K_{k+1}(h_{k+1}(x_{k+1}, v_{k+1}) - h_{k+1}(\hat{x}_{k+1}^-, 0)) \quad (15.5.153)$$

$$= K_{k+1}(h_{k+1}(\hat{x}_{k+1}^- + e_{k+1}^-, v_{k+1}) - h_{k+1}(\hat{x}_{k+1}^-, 0)) \quad (15.5.154)$$

Taking a Taylor approximation about  $(\hat{x}_{k+1}^-, 0)$ :

$$h_{k+1}(\hat{x}_{k+1}^- + e_{k+1}^-, v_{k+1}) \approx h_{k+1}(\hat{x}_{k+1}^-, 0) + \frac{\partial h_{k+1}}{\partial x} \Big|_{(\hat{x}_{k+1}^-, 0)} e_{k+1}^- + \frac{\partial h_{k+1}}{\partial v} \Big|_{(\hat{x}_{k+1}^-, 0)} v_{k+1} \quad (15.5.155)$$

$$= h_{k+1}(\hat{x}_{k+1}^-, 0) + C_{k+1} e_{k+1}^- + F_{k+1} v_{k+1} \quad (15.5.156)$$

Hence

$$\hat{x}_{k+1} - \hat{x}_{k+1}^- \approx K_{k+1}(C_{k+1} e_{k+1}^- + F_{k+1} v_{k+1}) \quad (15.5.157)$$

This approximation is in the same form as the linear case, from which we can derive the ‘near-optimal’ Kalman gain  $K_{k+1} = P_{k+1}^- C_{k+1}^\top (C_{k+1} P_{k+1}^- C_{k+1}^\top + R'_k)^{-1}$  with the same steps.

#### 15.5.4 Unscented Kalman Filter

The extended Kalman filter estimates may be poor particularly if the linearisation is poor. Also, even if the nonlinear mappings are differentiable, the Jacobian matrices may not always be easy to obtain or implement. An alternative approach to dealing with nonlinearity is with the unscented Kalman filter (UKF). The UKF uses the unscented transform to estimate the mean and covariance in the predict step.

##### Unscented Transform [104, 184]

The unscented transform can generally be used to estimate the parameters of a distribution after having undergone a nonlinear transformation (we are typically interested in the mean and covariance). Consider the nonlinear function  $f(\cdot)$  of a random vector  $\mathbf{X}$  to another random variable  $\mathbf{Y} = f(\mathbf{X})$ . From the mean  $\mu_X$  and covariance  $\mathbf{P}_X$  of  $\mathbf{X}$  (or estimates thereof), we seek to estimate the  $\mu_Y$  and covariance  $\mathbf{P}_Y$  of  $\mathbf{Y}$ . The unscented transform does this using a deterministic sampling approach. If  $\mathbf{X}$  is  $n$ -dimensional, then  $2n + 1$  sampling points (one for the mean and two in each dimension) are taken, referred to as ‘sigma points’. The  $n$  sigmas are determined as

$$[\sigma_1 \ \dots \ \sigma_n] = \mathbf{L} \sqrt{n + \kappa} \quad (15.5.158)$$

ie. the sigmas are the columns of  $\mathbf{L}$ , where  $\mathbf{L}$  can be any square root of  $\mathbf{P}_X$  such that  $\mathbf{LL}^\top = \mathbf{P}_X$ . A common choice in implementation is to use  $\mathbf{L}$  as the Cholesky decomposition. The role of  $\kappa$  is a parameter which scales the sigma point distribution. The sigma points are then computed as

$$x_i = \boldsymbol{\mu}_X + \boldsymbol{\sigma}_i \quad (15.5.159)$$

for  $i = 1, \dots, n$ ,

$$x_{i+n} = \boldsymbol{\mu}_X - \boldsymbol{\sigma}_i \quad (15.5.160)$$

for  $i = 1, \dots, n$ , and

$$x_0 = \boldsymbol{\mu}_X \quad (15.5.161)$$

The transformed sigma points are given by

$$y_i = f(x_i) \quad (15.5.162)$$

for  $i = 0, \dots, 2n$ . We choose weightings  $w_0, \dots, w_{2n}$  by

$$w_i = \begin{cases} \frac{\kappa}{n + \kappa}, & i = 0 \\ \frac{1}{2(n + \kappa)}, & i = 1, \dots, 2n \end{cases} \quad (15.5.163)$$

so that the estimate of  $\boldsymbol{\mu}_Y$  is given by the weighted sum

$$\hat{\boldsymbol{\mu}}_Y = \sum_{i=0}^{2n} w_i x_i \quad (15.5.164)$$

$$= \frac{1}{n + \kappa} \left( \kappa y_0 + \frac{1}{2} \sum_{i=1}^{2n} x_i \right) \quad (15.5.165)$$

and the estimated of the covariance  $\mathbf{P}_Y$  is

$$\hat{\mathbf{P}}_Y = \sum_{i=0}^{2n} w_i (y_i - \hat{\boldsymbol{\mu}}_Y)(y_i - \hat{\boldsymbol{\mu}}_Y)^\top \quad (15.5.166)$$

$$= \frac{1}{n + \kappa} \left( \kappa (y_0 - \hat{\boldsymbol{\mu}}_Y)(y_0 - \hat{\boldsymbol{\mu}}_Y)^\top + \frac{1}{2} \sum_{i=1}^{2n} (y_i - \hat{\boldsymbol{\mu}}_Y)(y_i - \hat{\boldsymbol{\mu}}_Y)^\top \right) \quad (15.5.167)$$

### Unscented Kalman Filter Algorithm

In the setup for the UKF algorithm, we require slightly stricter assumptions. The stochastic dynamic model is

$$x_{k+1} = f_k(x_k, u_k) + w_k \quad (15.5.168)$$

where  $f(x_k, u_k)$  can be generally time-varying and nonlinear and not necessarily differentiable, however the noise  $w_k$  is additive. The measurement model is

$$z_{k+1} = h_k(x_k) + v_k \quad (15.5.169)$$

where  $h(x_k)$  can again be time-varying, nonlinear and not necessarily differentiable, but the noise  $v_k$  is also additive. The noises are non-stationary white, but assumed to be Gaussian with covariances:

$$\mathbb{E} [w_k w_i^\top] = \begin{cases} Q_k, & k = i \\ 0, & k \neq i \end{cases} \quad (15.5.170)$$

$$\mathbb{E} [v_k v_i^\top] = \begin{cases} R_k, & k = i \\ 0, & k \neq i \end{cases} \quad (15.5.171)$$

and as usual uncorrelated with each other (this can be relaxed, although it would then need to be accounted for in the calculation of the cross-covariance):

$$\mathbb{E} [w_k v_i^\top] = 0 \quad (15.5.172)$$

Using the unscented transform, the UKF algorithm at step  $k + 1$  is summarised by:

- From the previous estimates  $\hat{x}_k$  and  $P_{xx,k}$ , form the  $2n + 1$  sigma points  $\chi_0, \dots, \chi_{2n}$  and using the weights  $w_0, \dots, w_{2n}$ , propagate the estimates to form the prior state estimates:

$$\hat{x}_{k+1}^- = \sum_{i=0}^{2n} w_i f_k(\chi_i, u_k) \quad (15.5.173)$$

$$P_{xx,k+1}^- = \sum_{i=0}^{2n} w_i (f_k(\chi_i, u_k) - \hat{x}_{k+1}^-) (f_k(\chi_i, u_k) - \hat{x}_{k+1}^-)^\top + Q_{k+1} \quad (15.5.174)$$

- From the prior state estimates  $\hat{x}_{k+1}^-$  and  $P_{xx,k+1}^-$ , resample the sigma points  $\chi_0^-, \dots, \chi_{2n}^-$  and compute the prior output estimates:

$$\hat{z}_{k+1}^- = \sum_{i=0}^{2n} w_i h_k(\chi_i^-) \quad (15.5.175)$$

$$P_{zz,k+1}^- = \sum_{i=0}^{2n} w_i (h_k(\chi_i^-) - \hat{z}_{k+1}^-) (h_k(\chi_i^-) - \hat{z}_{k+1}^-)^\top + R_{k+1} \quad (15.5.176)$$

and compute the prior cross-covariances by:

$$P_{xz,k+1}^- = \sum_{i=0}^{2n} w_i (f_k(\chi_i, u_k) - \hat{x}_{k+1}^-) (h_k(\chi_i^-) - \hat{z}_{k+1}^-)^\top \quad (15.5.177)$$

- Calculate the Kalman gain by

$$K_{k+1} = P_{xz,k+1}^- (P_{zz,k+1}^-)^{-1} \quad (15.5.178)$$

- Use the update equations from the Gaussian filter to obtain the posterior estimates:

$$\hat{x}_{k+1} = \hat{x}_{k+1}^- + K_{k+1} (z_{k+1} - \hat{z}_{k+1}^-) \quad (15.5.179)$$

$$P_{xx,k+1} = P_{xx,k+1}^- - K_{k+1} P_{zz,k+1}^- K_{k+1}^\top \quad (15.5.180)$$

Note that we are necessarily assuming that the prior distribution is Gaussian (which is one reason for assuming Gaussian noise) in order to use the update equations from the Gaussian filter. This implicitly requires us to approximate the distribution returned by the unscented transform as a Gaussian. As a result, the posterior may also be interpreted as a Gaussian.

---

### 15.5.5 Information Filter

### 15.5.6 Kalman-Bucy Filter

### 15.5.7 Kalman Smoother [114, 184]

## 15.6 Particle Filtering

Also known as *sequential Monte-Carlo* filters, the particle filters use Monte-Carlo sampling to approximate the posterior in Bayes filtering, where the setup is that of a hidden Markov Model. It is suitable as an alternative to the extended Kalman Filter and unscented Kalman filter, for systems with nonlinear dynamics and even non-Gaussian noise. Following properties of hidden Markov models, the dynamics are described by an initial distribution of the state  $p(x_0)$ , and are required to satisfy the Markov property so that the transition distribution may be specified as  $p(x_{k+1}|x_k)$ . The observation  $z_k$  is conditionally independent with  $\mathbf{x}_{0:(k-1)}$  given  $x_k$ , so that the distribution of  $z_k$  is modelled with  $p(z_k|x_k)$ .

### 15.6.1 Bootstrap Filter [53]

A particular variant of the particle filter called the Bootstrap filter is described below. Initially, we first sample  $N$  particles from the initial distribution  $p(x_0)$ , and denote the  $i^{\text{th}}$  particle by  $\chi_0[i]$ . Then at the  $(k+1)^{\text{th}}$  iteration, we perform the following steps:

1. Sample  $N$  particles from the transition density using the previous posterior particles, so that each

$$\chi_{k+1}^-[i] \sim p(x_{k+1}|\chi_k[i]) \quad (15.6.1)$$

2. For each particle evaluate the *importance weights* according to the observation likelihood by

$$w_{k+1,i} = p(z_{k+1}|\chi_{k+1}^-[i]) \quad (15.6.2)$$

3. Normalise the weights by

$$\tilde{w}_{k+1,i} = \frac{w_{k+1,i}}{\sum_{j=1}^N w_{k+1,j}} \quad (15.6.3)$$

The approximate posterior may be represented using the weighted point masses:

$$\hat{p}(x_{k+1}|\mathbf{z}_{1:(k+1)}) = \sum_{i=1}^N \tilde{w}_{k+1,i} \delta(\|x_{k+1} - \chi_{k+1}^-[i]\|) \quad (15.6.4)$$

where  $\delta(\cdot)$  is the Dirac delta function.

4. Bootstrap this posterior distribution by resampling with replacement  $N$  particles (i.e. each particle  $\chi_{k+1}^-[i]$  will be resampled with probability  $\tilde{w}_{k+1,i}$ ). Denote the resampled particles by  $\chi_{k+1}[i]$  and treat the updated posterior approximation as

$$\hat{p}'(x_{k+1}|\mathbf{z}_{1:(k+1)}) = \frac{1}{N} \sum_{i=1}^N \delta(\|x_{k+1} - \chi_{k+1}[i]\|) \quad (15.6.5)$$

From the particles at time  $k$ , we can estimate statistics of interest such as the posterior mean or posterior mode. The posterior mean is the mean over the resampled particles:

$$\hat{\mathbb{E}}[x_k] = \frac{1}{N} \sum_{i=1}^N \chi_k[i] \quad (15.6.6)$$

Since each particle effectively has uniform weight in the bootstrapped posterior, the posterior mode corresponds to the particle with the largest importance weight from the un-bootstrapped posterior:

$$\check{x}_k = \chi_k^- \left[ \operatorname{argmax}_i \{ \tilde{w}_{k,i} \} \right] \quad (15.6.7)$$

The posterior covariance may also be estimated as (noting that this is the population covariance of the posterior approximation):

$$\widehat{\operatorname{Cov}}(x_k) = \frac{1}{N} \sum_{i=1}^N \left( \chi_k[i] - \widehat{\mathbb{E}}[x_k] \right) \left( \chi_k[i] - \widehat{\mathbb{E}}[x_k] \right)^\top \quad (15.6.8)$$

### 15.6.2 Sequential Importance Sampling [53, 196]

Consider the density  $p_n(\mathbf{x}_{0:n})$ , which can be approximated using importance sampling. We seek an importance sampling technique that allows us to form a particle approximation of  $p_{n+1}(\mathbf{x}_{0:(n+1)})$  using the unnormalised target density  $\phi_{n+1}(\mathbf{x}_{0:(n+1)}) \propto p_{n+1}(\mathbf{x}_{0:(n+1)})$ , and given that we already have a particle approximation at time  $n$ :

$$\widehat{p}_n(\mathbf{x}_{0:n}) = \sum_{i=1}^N \tilde{w}_{n,i} \delta(\|\mathbf{x}_{0:n} - \chi_{0:n}[i]\|) \quad (15.6.9)$$

where  $\chi_{0:n}[i]$  denotes the  $i^{\text{th}}$  particle trajectory. This is done by choosing an importance density  $q_n(\mathbf{x}_{0:n})$  that factorises as

$$q_n(\mathbf{x}_{0:n}) = q_n(x_n | \mathbf{x}_{0:(n-1)}) q_{n-1}(\mathbf{x}_{0:(n-1)}) \quad (15.6.10)$$

so that by iterating,

$$q_n(\mathbf{x}_{0:n}) = q_0(x_0) \prod_{k=1}^n q_k(x_k | \mathbf{x}_{0:(k-1)}) \quad (15.6.11)$$

That is, we can generate particles at time  $n+1$  using the particles from time  $n$ . This allows us to recursively compute importance weights, as

$$w_{n+1}(\mathbf{x}_{0:(n+1)}) = \frac{\phi_{n+1}(\mathbf{x}_{0:(n+1)})}{q_{n+1}(\mathbf{x}_{0:(n+1)})} \quad (15.6.12)$$

$$= \frac{\phi_n(\mathbf{x}_{0:n})}{q_{n+1}(\mathbf{x}_{0:(n+1)})} \cdot \frac{\phi_{n+1}(\mathbf{x}_{0:(n+1)})}{\phi_n(\mathbf{x}_{0:n})} \quad (15.6.13)$$

$$= \frac{\phi_n(\mathbf{x}_{0:n})}{q_{n+1}(x_{n+1} | \mathbf{x}_{0:n}) q_n(\mathbf{x}_{0:n})} \cdot \frac{\phi_{n+1}(\mathbf{x}_{0:(n+1)})}{\phi_n(\mathbf{x}_{0:n})} \quad (15.6.14)$$

$$= \underbrace{w_n(\mathbf{x}_{0:n})}_{w_n(\mathbf{x}_{0:n})} \cdot \frac{\phi_{n+1}(\mathbf{x}_{0:(n+1)})}{\phi_n(\mathbf{x}_{0:n}) q_{n+1}(x_{n+1} | \mathbf{x}_{0:n})} \quad (15.6.15)$$

Now consider the choice of importance density  $q_{n+1}(x_{n+1} | \mathbf{x}_{0:n})$  that minimises the variance of the approximation in some sense. Naturally, as the weights appear in the approximation, we should try to minimise the variance of the weights, conditional on the previous weights. Since the previous weights  $w_n(\mathbf{x}_{0:n})$  are given, we should aim to make the ratio

$$\frac{\phi_{n+1}(\mathbf{x}_{0:(n+1)})}{\phi_n(\mathbf{x}_{0:n}) q_{n+1}(x_{n+1} | \mathbf{x}_{0:n})} = \frac{\phi_{n+1}(x_{n+1} | \mathbf{x}_{0:n})}{q_{n+1}(x_{n+1} | \mathbf{x}_{0:n})} \quad (15.6.16)$$

a constant, which is possible by choosing

$$q_{n+1}(x_{n+1}|\mathbf{x}_{0:n}) = p_{n+1}(x_{n+1}|\mathbf{x}_{0:n}) \quad (15.6.17)$$

$$\propto \phi_{n+1}(x_{n+1}|\mathbf{x}_{0:n}) \quad (15.6.18)$$

Of course, we may not be able to obtain  $p_{n+1}(x_{n+1}|\mathbf{x}_{0:n})$  exactly in the first place, however in practice we could attempt to use some approximation of  $p_{n+1}(x_{n+1}|\mathbf{x}_{0:n})$ .

### Importance Sampling in Particle Filtering

The principles behind the bootstrap filter can be derived from sequential importance sampling. Suppose at the start of time  $k + 1$  we have particle trajectories  $\chi_k[i]$  and weightings  $\tilde{w}_{k,i}$  from the previous iteration, and have approximated the posterior trajectory at time  $k$  using the weighted point masses

$$\hat{p}(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) = \sum_{i=1}^N \tilde{w}_{k,i} \delta(\|\mathbf{x}_{0:k} - \chi_{0:k}[i]\|) \quad (15.6.19)$$

After observing measurement  $z_{k+1}$ , our goal is to draw samples from the posterior  $p(x_{k+1}|\mathbf{z}_{1:(k+1)})$  so that we may approximate that distribution using weighted point masses at the samples. From sequential importance sampling, this gives the idea that if we can sample a particle  $\chi_{k+1}^-[i]$  from a proposal/importance distribution  $q(x_{k+1}|\mathbf{z}_{1:(k+1)})$ , then we can weight that sample by  $\tilde{w}_{k+1,i}$  when representing the posterior  $p(x_{k+1}|\mathbf{z}_{1:(k+1)})$ . Let the target density be

$$\phi(\mathbf{x}_{0:k}) = p(\mathbf{x}_{0:k}, \mathbf{z}_{1:k}) \quad (15.6.20)$$

which is proportional to our desired density, since

$$\phi(\mathbf{x}_{0:k}) = p(\mathbf{x}_{0:k}, \mathbf{z}_{1:k}) \quad (15.6.21)$$

$$\propto p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \quad (15.6.22)$$

$$= \frac{p(\mathbf{x}_{0:k}, \mathbf{z}_{1:k})}{p(\mathbf{z}_{1:k})} \quad (15.6.23)$$

The formula for updating the weights for the  $i^{\text{th}}$  particle from sequential importance sampling is

$$w_{k+1}(\mathbf{x}_{0:(k+1)}) = \tilde{w}_k(\mathbf{x}_{0:k}) \frac{\phi(\mathbf{x}_{0:(k+1)})}{\phi(\mathbf{x}_{0:n}) q(x_{k+1}|\mathbf{x}_{0:k})} \quad (15.6.24)$$

$$= \tilde{w}_{k,i} \frac{p(\mathbf{x}_{0:(k+1)}, \mathbf{z}_{1:(k+1)})}{p(\mathbf{x}_{0:k}, \mathbf{z}_{1:k}) q(x_{k+1}|\mathbf{x}_{0:k})} \quad (15.6.25)$$

$$= \tilde{w}_{k,i} \frac{p(x_{k+1}, z_{k+1}|\mathbf{x}_{0:k}, \mathbf{z}_{1:k})}{q(x_{k+1}|\mathbf{x}_{0:k})} \quad (15.6.26)$$

$$= \tilde{w}_{k,i} \frac{p(z_{k+1}|x_{k+1}) p(x_{k+1}|x_k)}{q(x_{k+1}|\mathbf{x}_{0:k})} \quad (15.6.27)$$

using conditional independence properties of hidden Markov models. For the importance density  $q(x_{n+1}|\mathbf{x}_{0:n})$ , it is desired to make it ‘close’ to the optimal importance density which is proportional to

$$\phi(x_{k+1}|\mathbf{x}_{0:k}) \propto p(x_{k+1}|\mathbf{x}_{0:k}, \mathbf{z}_{1:(k+1)}) \quad (15.6.28)$$

$$= p(x_{k+1}|x_k, z_{k+1}) \quad (15.6.29)$$

One choice is to set the importance density as the transition density

$$q(x_{k+1}|\mathbf{x}_{0:k}) = p(x_{k+1}|x_k) \quad (15.6.30)$$

so that the weight update simplifies to

$$w_{k+1}(\mathbf{x}_{0:(k+1)}) = \tilde{w}_k(\mathbf{x}_{0:k}) p(z_{k+1}|x_{k+1}) \quad (15.6.31)$$

This justifies the observation likelihood used in computing the importance weights for the bootstrap filter.

### Bootstrap Filter for Posterior Trajectory

Using sequential importance sampling, an algorithm can that approximates the posterior trajectory  $p(\mathbf{x}_{0:(k+1)}|\mathbf{z}_{1:(k+1)})$  using particle trajectories denoted  $\chi_{0:(k+1)}[i]$  can be developed. To summarise this algorithm:

1. Sample  $N$  particles from the importance density  $q(x_{k+1}|\mathbf{x}_{0:k}) = p(x_{k+1}|x_k)$  using the most recent particles in the particle trajectories, so that each

$$\chi_{k+1}^-[i] \sim p(x_{k+1}|\chi_k[i]) \quad (15.6.32)$$

2. For each particle, evaluate the importance weights according to the observation likelihood by

$$w_{k+1,i} = p(z_{k+1}|\chi_{k+1}^-[i]) \quad (15.6.33)$$

These importance weights are effectively treating the previous importance weights to be uniform.

3. Normalise the weights with

$$\tilde{w}_{k+1,i} = \frac{w_{k+1,i}}{\sum_{j=1}^N w_{k+1,j}} \quad (15.6.34)$$

The approximate posterior trajectory may be represented using the weighted point masses:

$$\hat{p}(\mathbf{x}_{0:(k+1)}|\mathbf{z}_{1:(k+1)}) = \sum_{i=1}^N \tilde{w}_{k+1,i} \delta\left(\|\mathbf{x}_{0:(k+1)} - \chi_{0:(k+1)}^-[i]\|\right) \quad (15.6.35)$$

where  $\delta(\cdot)$  is the Dirac delta function.

4. Bootstrap this posterior trajectory by resampling with replacement  $N$  particle trajectories (i.e. each particle trajectory  $\chi_{0:(k+1),i}$  will be resampled with probability  $\tilde{w}_{k+1}[i]$ ). Denote the resampled particle trajectories by  $\chi_{0:(k+1),i}$  and treat the updated posterior trajectory approximation as

$$\tilde{p}'(\mathbf{x}_{0:(k+1)}|\mathbf{z}_{1:(k+1)}) = \frac{1}{N} \sum_{i=1}^N \delta\left(\|\mathbf{x}_{0:(k+1)} - \chi_{0:(k+1)}[i]\|\right) \quad (15.6.36)$$

This resampling step gives each particle trajectory a uniform weighting of  $1/N$ , which validates the method of setting the importance weights equal to the observation likelihood.

Although bootstrapping adds another layer of approximation, by resetting the weights uniformly, this is to avoid the ‘degeneracy problem’ [163] where the variance of the weights become too large (meaning some weights may become too large and other weights become too small over time), causing the posterior to get closer to a degenerate distribution.

Note that if we discard the history of particle trajectories  $\chi_{0:k}^-[i]$ , then this reduces to the bootstrap filter for the marginal posterior  $p(x_{k+1}|\mathbf{z}_{1:(k+1)})$ . Having the posterior trajectory over  $\mathbf{x}_{0:(k+1)}$  can be useful in some situations, for instance when we are interested in the posterior mode (interpreted as the maximum a posteriori estimate) of the trajectory. The posterior mode over the trajectory (denoted  $\check{\mathbf{x}}_{0:k}$ ) will generally be different to the sequence of marginal posterior modes over the state, denoted  $(\check{x}_0, \check{x}_1, \dots, \check{x}_k)$ . The former will be in better agreement with the transition dynamics, as the estimate consists the evolution of a single particle over time, whereas the latter estimate may be comprised of several particles.

### 15.6.3 Rao-Blackwellised Particle Filter [53, 114]

Suppose we can partition the state of the system into  $(x_k, s_k)$ . The filtering density of the trajectory  $(\mathbf{x}_{0:k}, \mathbf{s}_{0:k})$  given observations  $\mathbf{z}_{1:k}$  factorises into

$$p(\mathbf{x}_{0:k}, \mathbf{s}_{0:k} | \mathbf{z}_{1:k}) = p(\mathbf{x}_{0:k} | \mathbf{z}_{1:k}, \mathbf{s}_{0:k}) p(\mathbf{s}_{0:k} | \mathbf{z}_{1:k}) \quad (15.6.37)$$

Suppose that when conditioned on the states  $\mathbf{s}_{0:k}$ , the filtering distribution  $\mathbf{x}_{0:k}$  becomes analytically tractable. For instance, the transition dynamics  $p(x_k | x_{k-1})$  and the observation likelihood  $p(z_k | x_k)$  could be conditionally Gaussian given  $s_k$ . In that case, a Gaussian filter could be used to obtain  $p(\mathbf{x}_{0:k} | \mathbf{z}_{1:k}, \mathbf{s}_{0:k})$ . As an example, the system could take the form

$$x_{k+1} = f(x_k, s_k) + w_k \quad (15.6.38)$$

$$z_k = h(x_k, s_k) + v_k \quad (15.6.39)$$

with additive i.i.d. Gaussian noise  $w_k, v_k$ . A further example is that the system becomes conditionally linear Gaussian given  $s_k$ , in which case a Kalman filter can be used to obtain  $p(x_k | \mathbf{z}_{1:k}, \mathbf{s}_{0:k})$ . Here, the role of  $s_k$  may be thought of as a parametrisation of the linear dynamics, so a jump Markov linear system satisfies this characterisation. As we have an analytical filtering distribution for  $x_k$ , a particle filter can be employed only for  $s_k$ , which gives us an approximate filtering distribution of the form

$$\hat{p}(\mathbf{s}_{0:k} | \mathbf{z}_{1:k}) = \sum_{i=1}^N \tilde{w}_{k,i} \delta(\|\mathbf{s}_{0:k} - \varsigma_{0:k}[i]\|) \quad (15.6.40)$$

with particle trajectories  $\varsigma_{0:k}[i]$  and corresponding normalised weightings  $\tilde{w}_{k,i}$ . Since particles are being used only to track  $s_k$  which is a smaller dimension than the state  $(x_k, s_k)$ , this improves the efficiency of the filter. The approximate filtering distribution for  $x_k$  can then be taken to be

$$\hat{p}(x_k | \mathbf{z}_{1:k}) = \int p(x_k | \mathbf{z}_{1:k}, \mathbf{s}_{0:k}) \hat{p}(\mathbf{s}_{0:k} | \mathbf{z}_{1:k}) d\mathbf{s}_{0:k} \quad (15.6.41)$$

$$= \sum_{i=1}^N \tilde{w}_{k,i} p(x_k | \mathbf{z}_{1:k}, \varsigma_{0:k}[i]) \quad (15.6.42)$$

That is,  $N$  analytical filters are used, one for each particle, where the trajectory  $\varsigma_{0:k}[i]$  is treated as the ground truth. If the analytical filter is a Gaussian filter, then  $\hat{p}(x_k | \mathbf{z}_{1:k})$  is a Gaussian mixture. A point estimate can also be obtained with the posterior mean:

$$\hat{\mathbb{E}}[x_k | \mathbf{z}_{1:k}] = \hat{\mathbb{E}}_{\mathbf{s}_{0:k}} [\mathbb{E}[x_k | \mathbf{z}_{1:k}, \mathbf{s}_{0:k}] | \mathbf{z}_{1:k}] \quad (15.6.43)$$

$$= \sum_{i=1}^N \tilde{w}_{k,i} \mathbb{E}[x_k | \mathbf{z}_{1:k}, \varsigma_{0:k}[i]] \quad (15.6.44)$$

It is here where it becomes clearer why the filter is called a Rao-Blackwellised particle filter. Pretend a particle filter was being used to estimate  $x_k$  as well, so that we have the joint particle trajectory  $(\chi_{0:k}[i], \varsigma_{0:k}[i])$ , and  $x_k$  is being estimated (using observations  $\mathbf{z}_{1:k}$ ) as

$$\hat{x}_{k|k} = \sum_{i=1}^N \tilde{w}_{k,i} \chi_{k|k}[i] \quad (15.6.45)$$

Rao Blackwellisation of an estimator means to take a conditional expectation, and this reduces the variance. Here, we take the conditional expectation given the  $\mathbf{s}_{0:k}$  particles, so  $\chi_{k|k}[i]$  is instead replaced with  $\mathbb{E}[x_k | \mathbf{z}_{1:k}, \varsigma_{0:k}[i]]$ , with the aim of reducing the variance.

15.6.4 Particle Smoothing [54]

15.7 Independent Component Analysis [96, 196]

15.8 Wavelets [134, 210, 211]

15.8.1 Wavelet Denoising

15.9 Compressed Sensing [62, 81]

# Chapter 16

## Stochastic Control

### 16.1 System Identification

#### 16.1.1 Quasistationary Signals [130]

A discrete-time stochastic process  $X_k$  is said to be quasistationary if it is a second-order process, i.e.

$$\mathbb{E}[X_k] = \mu_X(k) \quad (16.1.1)$$

$$\mathbb{E}[X_k X_\ell] = R_X(k, \ell) \quad (16.1.2)$$

with bounded  $|\mu_X(k)| \leq C$ ,  $|R_X(k, \ell)| \leq C$  for all  $k, \ell$  and some  $C$ , and additionally  $X_k$  satisfies the following condition:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N R_X(k, k - \tau) = R_X(\tau) \quad (16.1.3)$$

That is, the time-average of the autocorrelation function  $R_X(k, k - \tau)$  tends to a function  $R_X(\tau)$  which only depends on the time difference. Note that all ergodic wide-sense stationary processes automatically satisfy this definition and are hence also quasistationary. A quasistationary signal can also be constructed by adding a bounded deterministic sequence to a wide-sense stationary process (to create the mean function  $\mu_X(k)$ ).

The purpose of defining quasistationary signals is that in system identification, the input signals may be deterministic signals, which are generally not wide-sense stationary. However, deterministic signals can still be considered quasistationary, which is how we can handle the analysis of outputs to systems with stochastic noise and deterministic inputs.

#### Quasistationary Deterministic Signals

If a discrete-time signal  $x_k$  is a bounded deterministic sequence, then the first two conditions for quasistationarity are trivially satisfied, and the only condition required is that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N x_k x_{k-\tau} = R_x(\tau) \quad (16.1.4)$$

If  $x_k$  is the realisation of an ergodic wide-sense stationary process  $X_k$ , then by the ergodicity property,  $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N x_k x_{k-\tau} = \mathbb{E}[X_k X_{k-\tau}] = R_x(\tau)$  and we can see that realisations of wide-sense stationary processes satisfy the definition of quasistationary deterministic signals.

## Joint Quasistationarity

Two signals  $X_k$  and  $Y_k$  are said to be jointly quasistationary if  $X_k$  and  $Y_K$  are both quasistationary and

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N R_{YX}(k, k - \tau) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \mathbb{E}[X_k Y_{k-\tau}] \quad (16.1.5)$$

$$= R_{XY}(\tau) \quad (16.1.6)$$

This is analogous to joint wide-sense stationarity.

## Power Spectral Density of Quasistationary Signals

The power spectral density of quasistationary signals can be analogously defined. Written in units of angular frequency  $\omega$ , the power spectrum of  $X_k$  is

$$\Phi_X(\omega) = \sum_{\tau=-\infty}^{\infty} R_X(\tau) e^{-j\tau\omega} \quad (16.1.7)$$

while the cross-spectrum between  $X_k$  and  $Y_k$  is

$$\Phi_{XY}(\omega) = \sum_{\tau=-\infty}^{\infty} R_{XY}(\tau) e^{-j\tau\omega} \quad (16.1.8)$$

## Quasistationary Spectra under Linear Filters

Quasistationary signals passed through linear filters are also quasistationary, analogously to wide-sense stationary signals, and their power spectra are transformed accordingly. Suppose we have a quasistationary signal  $X_k$  with power spectral density  $\Phi_X(\omega)$ . Let this signal be an input into a stable transfer function  $G(q)$  (where  $q$  is a shift operator), such that the output is  $Y_k = G(q) X_k$ . Then  $Y_k$  is also quasistationary with

$$\Phi_Y(\omega) = |G(e^{j\omega})|^2 \Phi_X(\omega) \quad (16.1.9)$$

$$\Phi_{XY}(\omega) = G(e^{j\omega}) \Phi_X(\omega) \quad (16.1.10)$$

This can be generalised to multivariate systems. Suppose now that  $\Phi_X(\omega)$  is a jointly quasistationary multivariate signal with power spectral density matrix  $\Phi_X(\omega)$  (the off-diagonals give the cross-spectra), and  $G(q)$  is a  $p \times m$  transfer function matrix (hence there are  $p$  outputs and  $m$  inputs). Then  $Y_k = G(q) X_k$  is also jointly quasistationary with  $p \times p$  power spectral density matrix

$$\Phi_Y(\omega) = G(e^{j\omega}) \Phi_X(\omega) G(e^{-j\omega})^\top \quad (16.1.11)$$

and  $m \times p$  cross-spectral density matrix

$$\Phi_{XY}(\omega) = \Phi_X(\omega) G(e^{-j\omega})^\top \quad (16.1.12)$$

*Proof.* First, we assume that the limit

$$R_Y(\tau) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \mathbb{E}[Y_k Y_{k-\tau}^\top] \quad (16.1.13)$$

exists and that subsequently  $Y_k$  is jointly quasistationary. We represent the impulse response of  $G(q)$  by the sequence  $\{g(0), g(1), \dots\}$  where each  $g(k)$  is a  $p \times n$  matrix. Then by discrete-time convolution:

$$Y_k = \sum_{i=0}^{\infty} g(i) X_{k-i} \quad (16.1.14)$$

$$Y_{k-\tau} = \sum_{\ell=0}^{\infty} g(\ell) X_{k-\tau-\ell} \quad (16.1.15)$$

Hence

$$R_Y(\tau) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \mathbb{E} \left[ \left( \sum_{i=0}^{\infty} g(i) X_{k-i} \right) \left( \sum_{\ell=0}^{\infty} g(\ell) X_{k-\tau-\ell} \right)^{\top} \right] \quad (16.1.16)$$

$$= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \sum_{i=0}^{\infty} \sum_{\ell=0}^{\infty} g(i) \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^{\top}] g(\ell)^{\top} \quad (16.1.17)$$

$$= \sum_{i=0}^{\infty} \sum_{\ell=0}^{\infty} g(i) \left( \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^{\top}] \right) g(\ell)^{\top} \quad (16.1.18)$$

Note that the index  $k - i$  leads  $k - \tau - \ell$  by  $\tau + \ell - i$  timesteps, hence by a change of variables  $k - i = s$ ,

$$\sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^{\top}] = \sum_{s=1-i}^N \mathbb{E} [X_s X_{s-\tau-\ell+i}^{\top}] \quad (16.1.19)$$

So if we take  $X_k = 0$  for  $k < 0$ , the starting index of the sum is inconsequential and then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^{\top}] = R_X(\tau + \ell - i) \quad (16.1.20)$$

so

$$R_Y(\tau) = \sum_{i=0}^{\infty} \sum_{\ell=0}^{\infty} g(i) R_X(\tau + \ell - i) g(\ell)^{\top} \quad (16.1.21)$$

This mirrors an analogous result for wide-sense stationarity. Now using the definition of the power spectral density:

$$\Phi_Y(\omega) = \sum_{\tau=-\infty}^{\infty} R_Y(\tau) e^{-j\tau\omega} \quad (16.1.22)$$

$$= \sum_{\tau=-\infty}^{\infty} \left( \sum_{i=0}^{\infty} \sum_{\ell=0}^{\infty} g(i) R_X(\tau + \ell - i) g(\ell)^{\top} \right) e^{-j\tau\omega} \quad (16.1.23)$$

Splitting up the exponent into  $\tau = i - \ell + \tau + \ell - i$ :

$$\Phi_Y(\omega) = \sum_{\tau=-\infty}^{\infty} \left( \sum_{i=0}^{\infty} \sum_{\ell=0}^{\infty} g(i) R_X(\tau + \ell - i) g(\ell)^{\top} \right) e^{-j(i-\ell+\tau+\ell-i)\omega} \quad (16.1.24)$$

$$= \left( \sum_{i=0}^{\infty} g(i) e^{-ji\omega} \right) \left( \sum_{\tau=-\infty}^{\infty} R_X(\tau + \ell - i) e^{-j(\tau+\ell-i)\omega} \right) \left( \sum_{\ell=0}^{\infty} g(\ell)^{\top} e^{j\ell\omega} \right) \quad (16.1.25)$$

Using a change of variables  $t = \tau + \ell - i$ :

$$\Phi_Y(\omega) = \left( \sum_{i=0}^{\infty} g(i) e^{-ji\omega} \right) \left( \sum_{t=-\infty}^{\infty} R_X(t) e^{-jt\omega} \right) \left( \sum_{\ell=0}^{\infty} g(\ell)^{\top} e^{j\ell\omega} \right) \quad (16.1.26)$$

$$= G(e^{j\omega}) \Phi_X(\omega) G(e^{-j\omega})^{\top} \quad (16.1.27)$$

where we recognise that the  $z$ -transform of the impulse response with  $z = e^{j\omega}$  gives the definition of the transfer function. Now all that is left to establish is that  $R_Y(\tau)$  exists, meaning that  $Y_k$  is jointly quasistationary. Denote the finite average:

$$R_{Y,N}(\tau) = \frac{1}{N} \sum_{k=1}^N \mathbb{E} [Y_k Y_{k-\tau}^\top] \quad (16.1.28)$$

From above, and also using the fact that  $X_k = 0$  for  $k < 0$ :

$$R_{Y,N}(\tau) = \frac{1}{N} \sum_{k=1}^N \sum_{i=0}^k \sum_{\ell=0}^{k-\tau} g(i) \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] g(\ell)^\top \quad (16.1.29)$$

Because of this fact, also note that changing the upper terminal of the second sum from  $k$  to  $N$  will be inconsequential since  $X_{k-i} = 0$  for  $i > k$ :

$$R_{Y,N}(\tau) = \frac{1}{N} \sum_{k=1}^N \sum_{i=0}^N \sum_{\ell=0}^{k-\tau} g(i) \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] g(\ell)^\top \quad (16.1.30)$$

Additionally, we have the ability to change the upper terminal on the third sum to  $N$ :

$$R_{Y,N}(\tau) = \frac{1}{N} \sum_{k=1}^N \sum_{i=0}^N \sum_{\ell=0}^N g(i) \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] g(\ell)^\top \quad (16.1.31)$$

This is because without loss of generality we can assume  $\tau \geq 0$ , as

$$R_Y(-\tau) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \mathbb{E} [Y_k Y_{k+\tau}^\top] \quad (16.1.32)$$

$$= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k'=1+\tau}^N \mathbb{E} [Y_{k'-\tau} Y_k^\top] \quad (16.1.33)$$

$$= \left( \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k'=1+\tau}^N \mathbb{E} [Y_{k'} Y_{k-\tau}^\top] \right)^\top \quad (16.1.34)$$

$$= R_Y(\tau)^\top \quad (16.1.35)$$

So if  $R_Y(\tau)$  exists for  $\tau \geq 0$ , then  $R_Y(\tau)$  will also exist for  $\tau < 0$ . From the argument  $\tau \geq 0$  we know  $k - \tau \leq N$ , so then by the same reasoning as above,  $X_{k-\tau-\ell} = 0$  for  $\ell > k - \tau$  and we can extend the sum to  $N$ . Doing this ‘trick’ allows us to swap the order of sums so that

$$R_{Y,N}(\tau) = \sum_{i=0}^N \sum_{\ell=0}^N g(i) \left( \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] \right) g(\ell)^\top \quad (16.1.36)$$

In the same way as for  $Y$ , denote for  $X$  the finite average:

$$R_{X,N}(\tau) = \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_k X_{k-\tau}^\top] \quad (16.1.37)$$

By a change of variables  $\kappa = k - i$ , we can write for the inner sum from above:

$$\frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] = \frac{1}{N} \sum_{\kappa=i}^{N+i} \mathbb{E} [X_\kappa X_{\kappa-\tau-\ell+i}^\top] \quad (16.1.38)$$

$$= \frac{1}{N} \sum_{\kappa=0}^N \mathbb{E} [X_\kappa X_{\kappa-\tau-\ell+i}^\top] + \frac{1}{N} \sum_{\kappa=N+1}^{N+i} \mathbb{E} [X_\kappa X_{\kappa-\tau-\ell+i}^\top] - \frac{1}{N} \sum_{\kappa=1}^{i-1} \mathbb{E} [X_\kappa X_{\kappa-\tau-\ell+i}^\top] \quad (16.1.39)$$

Hence the difference between this term and  $R_{X,N}(\tau - \ell + i) = \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_k X_{k-\tau-\ell+i}^\top]$  is given by

$$R_{X,N}(\tau - \ell + i) - \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] = - \frac{1}{N} \sum_{\kappa=N+1}^{N+i} \mathbb{E} [X_\kappa X_{\kappa-\tau-\ell+i}^\top] + \frac{1}{N} \sum_{\kappa=1}^{i-1} \mathbb{E} [X_\kappa X_{\kappa-\tau-\ell+i}^\top] \quad (16.1.40)$$

In the multivariate generalisation of quasistationarity, we will assume the norm of the autocorrelation  $\|E[X_k X_{k-\tau}^\top]\|$  is bounded above by some constant  $C$  for all  $k, \tau$ . Seeing that there are  $2i$  terms in the right-hand side directly above, we thus arrive at the bound:

$$\left\| R_{X,N}(\tau - \ell + i) - \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] \right\| \leq \frac{2i}{N} C \quad (16.1.41)$$

Now consider the difference between  $R_Y(\tau)$  and  $R_{Y,N}(\tau)$ :

$$R_Y(\tau) - R_{Y,N}(\tau) = \sum_{i=0}^{\infty} \sum_{\ell=0}^{\infty} g(i) R_X(\tau + \ell - i) g(\ell)^\top - \sum_{i=0}^N \sum_{\ell=0}^N g(i) \left( \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] \right) g(\ell)^\top \quad (16.1.42)$$

We can split this difference into four terms:

$$\begin{aligned} R_Y(\tau) - R_{Y,N}(\tau) &= \sum_{i>N}^{\infty} \sum_{\ell>N}^{\infty} g(i) R_X(\tau + \ell - i) g(\ell)^\top \\ &\quad + \sum_{i>N}^{\infty} \sum_{\ell=0}^{\infty} g(i) R_X(\tau + \ell - i) g(\ell)^\top + \sum_{i=0}^{\infty} \sum_{\ell>N}^{\infty} g(i) R_X(\tau + \ell - i) g(\ell)^\top \\ &\quad + \sum_{i=0}^N \sum_{\ell=0}^N g(i) \left( R_X(\tau + \ell - i) - \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] \right) g(\ell)^\top \end{aligned} \quad (16.1.43)$$

Since  $G(q)$  is stable,  $\|g(k)\| \rightarrow 0$  as  $k \rightarrow \infty$  and so the norm of the first three terms will tend to zero as  $N \rightarrow \infty$ . Hence

$$\lim_{N \rightarrow \infty} \|R_Y(\tau) - R_{Y,N}(\tau)\| = \lim_{N \rightarrow \infty} \left\| \sum_{i=0}^N \sum_{\ell=0}^N g(i) \left( R_X(\tau + \ell - i) - \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] \right) g(\ell)^\top \right\| \quad (16.1.44)$$

$$\leq \lim_{N \rightarrow \infty} \sum_{i=0}^N \sum_{\ell=0}^N \|g(i)\| \|g(\ell)\| \left\| R_{X,N}(\tau - \ell + i) - \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_{k-i} X_{k-\tau-\ell}^\top] \right\| \quad (16.1.45)$$

Using the bound we derived from above,

$$\lim_{N \rightarrow \infty} \|R_Y(\tau) - R_{Y,N}(\tau)\| \leq \lim_{N \rightarrow \infty} \sum_{i=0}^N \sum_{\ell=0}^N \|g(i)\| \|g(\ell)\| \frac{2i}{N} C \quad (16.1.46)$$

$$\leq \lim_{N \rightarrow \infty} \frac{2C}{N} \left( \sum_{i=0}^N i \|g(i)\| \right) \left( \sum_{\ell=0}^N \|g(\ell)\| \right) \quad (16.1.47)$$

By stability of  $G(q)$ , the sums  $\sum_{i=0}^N i \|g(i)\|$  and  $\sum_{\ell=0}^N \|g(\ell)\|$  are finite as  $N \rightarrow \infty$  (note that from the definition of a linear system, the impulse responses will exhibit exponential convergence). Therefore

$$\lim_{N \rightarrow \infty} \frac{2C}{N} \left( \sum_{i=0}^N i \|g(i)\| \right) \left( \sum_{\ell=0}^N \|g(\ell)\| \right) = 0 \quad (16.1.48)$$

which proves that

$$\lim_{N \rightarrow \infty} \|R_Y(\tau) - R_{Y,N}(\tau)\| = 0 \quad (16.1.49)$$

meaning that  $R_Y(\tau)$  exists and is equal to  $\lim_{N \rightarrow \infty} R_{Y,N}(\tau)$ . The steps to show that  $\Phi_{XY}(\omega) = \Phi_X(\omega) G(e^{-j\omega})^\top$  are analogous to this.  $\square$

We can use this result to analyse outputs of linear systems with quasistationary inputs and disturbances. Denote the output sequence  $y_k$ , input sequence  $u_k$  and disturbance sequence  $w_k$ . Suppose  $u_k$  and  $w_k$  are uncorrelated so that their cross-spectral density is zero. The system is described by:

$$y_k = G(q) u_k + H(q) w_k \quad (16.1.50)$$

We can write this in terms of the augmented input  $(u_k, w_k)$ :

$$y_k = [G(q) \quad H(q)] \begin{bmatrix} u_k \\ w_k \end{bmatrix} \quad (16.1.51)$$

Then the spectral density of the output is

$$\Phi_y(\omega) = [G(e^{j\omega}) \quad H(e^{j\omega})] \begin{bmatrix} \Phi_u(\omega) & 0 \\ 0 & \Phi_w(\omega) \end{bmatrix} \begin{bmatrix} G(e^{-j\omega})^\top \\ H(e^{-j\omega})^\top \end{bmatrix} \quad (16.1.52)$$

$$= G(e^{j\omega}) \Phi_u(\omega) G(e^{-j\omega})^\top + H(e^{j\omega}) \Phi_w(\omega) H(e^{-j\omega})^\top \quad (16.1.53)$$

### 16.1.2 Persistency of Excitation [206]

Persistency of excitation characterises whether an excitation signal excites enough of the frequency spectrum to be able to identify a model. To illustrate with an extreme example, a constant excitation signal will not excite the system dynamics so a model cannot be identified. Suppose we have an input signal  $u_k$  for  $k = 0, 1, 2, \dots$ . Then formally the sequence is said to be persistently exciting of order  $n$  if and only if there exists some integer  $N$  such that the  $n \times N$  matrix

$$U_{n,N} = \begin{bmatrix} u_0 & u_1 & \dots & u_{N-1} \\ u_1 & u_2 & \dots & u_N \\ \vdots & \vdots & \ddots & \vdots \\ u_{n-1} & u_n & \dots & u_{N+n-2} \end{bmatrix} \quad (16.1.54)$$

has full row rank  $n$ . This means that  $U_{n,N}$  has no linearly dependent rows. Periodic signals will not have persistency of excitation for all orders. For example, suppose we have a periodic input signal with period  $T$ , i.e.  $u_k = u_{k+T}$ . Then the signal will not be persistently exciting of order  $T + 1$  because the matrix

$$U_{T+1,N} = \begin{bmatrix} u_0 & u_1 & \dots & u_{N-1} \\ u_1 & u_2 & \dots & u_N \\ \vdots & \vdots & \ddots & \vdots \\ u_T & u_{T+1} & \dots & u_{N+T-1} \end{bmatrix} \quad (16.1.55)$$

will have first row equal to last row for all  $N$ , so it is automatically not full row rank.

For ergodic input signals, there is an equivalent condition for persistency of excitation. Note that

$$U_{n,N} U_{n,N}^\top = \begin{bmatrix} u_0 & u_1 & \dots & u_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n-1} & u_n & \dots & u_{N+n-2} \end{bmatrix} \begin{bmatrix} u_0 & \dots & u_{n-1} \\ u_1 & \dots & u_n \\ \vdots & \ddots & \vdots \\ u_{N-1} & \dots & u_{N+n-2} \end{bmatrix} \quad (16.1.56)$$

$$= \begin{bmatrix} \sum_{k=0}^{N-1} u_k^2 & \dots & \sum_{k=0}^{N-1} u_k u_{k+n-1} \\ \vdots & \ddots & \vdots \\ \sum_{k=0}^{N-1} u_k u_{k+n-1} & \dots & \sum_{k=n-1}^{n+N-2} u_k^2 \end{bmatrix} \quad (16.1.57)$$

Thus by the properties of ergodicity,

$$\lim_{N \rightarrow \infty} U_{n,N} U_{n,N}^\top = \begin{bmatrix} R_u(0) & R_u(1) & \dots & R_u(n-1) \\ R_u(1) & R_u(0) & \dots & R_u(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_u(n-1) & R_u(n-2) & \dots & R_u(0) \end{bmatrix} \quad (16.1.58)$$

where  $R_u(\kappa)$  is the autocorrelation function of  $u_k$ , if this limit exists (i.e.  $u_k$  is a second-order process). If  $U_{n,N}$  is full row rank with some  $N$  large enough, then this matrix will be nonsingular. Thus  $u_k$  is persistently exciting of order  $n$  if the  $n \times n$  matrix of autocorrelations is nonsingular. From this condition, we can also deduce that if  $u_k$  is white noise (i.e. its autocorrelation function is  $R_u(\kappa) = c\delta_\kappa$ ), then the autocorrelation matrix will be  $cI_{n \times n}$  which is nonsingular for all  $n$ . Thus white noise is persistently exciting for all orders  $n$ .

### 16.1.3 Frequency Domain Identification [98]

#### 16.1.4 Identifiability of Dynamic Systems

Loosely speaking, identifiability refers to whether unknown model parameters can be uniquely estimated from data [107]. There are various different notions of identifiability that can be introduced [147], however whether or not a system and its parameters are identifiable will generally depend on:

- The class of model structure/parametrisation.
- The actual values of the parameters themselves.
- The estimation method used to identify the parameters.
- Properties of the input signal used to excite the system (known as the experimental condition), and the resulting data.

#### Identifiability in the Parameter-Sense [130]

Consider a parametrised model structure  $\mathcal{M}(\theta)$ , and suppose  $\theta^*$  is the ‘true value’ of the parameter. Then the model structure  $\mathcal{M}$  is said to be *globally identifiable at  $\theta^*$  in the parameter-sense* if for some  $\theta_1$ , then  $\mathcal{M}(\theta_1) = \mathcal{M}(\theta^*)$  implies that  $\theta_1 = \theta_*$  (where model ‘equality’ roughly means that they would make identical predictions). Intuitively, this concept of identifiability relates to the model invertibility - whether or not  $\mathcal{M}$  is injective (i.e. one-to-one) with the parameters.

We can extend this definition over the entire parameter set and say that  $\mathcal{M}$  is *strictly globally identifiable in the parameter-sense* if it is globally identifiable at  $\theta^*$  for all  $\theta^*$  in the parameter set. A relaxation of this definition is to say that  $\mathcal{M}$  is *globally identifiable in the parameter-sense* if it is globally identifiable at  $\theta^*$  for almost all  $\theta^*$  in the parameter set (i.e. the set of parameters at which  $\theta^*$  is not globally identifiable has Lebesgue measure of zero).

A local version of this can also be defined, where we would say that model structure  $\mathcal{M}$  is *locally identifiable at  $\theta^*$  in the parameter-sense* if for any  $\theta_1$  in a small neighbourhood about  $\theta^*$ , then  $\mathcal{M}(\theta_1) = \mathcal{M}(\theta^*)$  implies that  $\theta_1 = \theta^*$ . Local identifiability and strict local identifiability of model structures can then be analogously defined in the same way as global identifiability. Another extension can be to additionally condition the model structure on the data  $\mathcal{D}$  used [155].

### Identifiability in the Criterion-Sense [107]

For estimation techniques that involve optimising a criterion (e.g. least squares or maximum likelihood), identifiability in the criterion-sense can be defined. Suppose  $J(\theta)$  is the criterion to be minimised to estimate the parameters  $\theta$  of a model structure, with true parameters  $\theta^*$ . Then the model structure is said to be *locally identifiable at  $\theta^*$  in the criterion-sense* if  $J(\theta)$  has a local minimum at  $\theta = \theta^*$ . This definition can also be refined to be conditioned on the data used, i.e. using  $J(\theta; \mathcal{D})$ .

### Identifiability in the Consistency-Sense [187]

Some definitions of identifiability of a model structure are determined by the existence of a consistent estimator, where model structure  $\mathcal{M}(\theta)$  with true parameter  $\theta^*$  is said to be *identifiable in the consistency-sense* with respect to a data generating process if there exists an estimator  $\hat{\theta}$  that is consistent for  $\theta^*$ .

Another broad notion of identifiability does not necessarily require the true system  $\mathcal{S}$  to be contained in the class of parametrised model structures  $\mathcal{M}(\theta)$ . This results in two separate concerns of identifiability - firstly whether or not the true system can be identified using the model structure, and secondly whether or not the parameters can then be uniquely determined. We would say that the system  $\mathcal{S}$  is *system-identifiable in the consistency-sense* under  $\mathcal{M}$ , the experimental condition  $\mathcal{E}$  and the estimator  $\hat{\theta}$  if  $\mathcal{M}(\hat{\theta}) \rightarrow \mathcal{S}$  as the sample size  $N \rightarrow \infty$  under  $\mathcal{E}$ . Here, the  $\rightarrow$  symbol can be taken to mean that  $\mathcal{M}(\hat{\theta})$  converges to a set of models that ‘perfectly describes’ true system  $\mathcal{S}$ , under some appropriate sense of probabilistic convergence, i.e. convergence in probability (for standard asymptotic consistency) or almost sure convergence (for strong convergence).

We can then further say that  $\mathcal{S}$  is *parameter-identifiable in the consistency-sense* under  $\mathcal{M}$ ,  $\mathcal{E}$  and  $\hat{\theta}$  if it is system-identifiable under  $\mathcal{M}$ ,  $\mathcal{E}$  and  $\hat{\theta}$ , and the model set that  $\mathcal{M}(\hat{\theta})$  converges to consists of only a single model.

### Identifiability of Linear Systems [155]

#### 16.1.5 Closed Loop Identification [116]

#### 16.1.6 Subspace Identification [130]

Consider the state-space system

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (16.1.59)$$

$$y_k = Cx_k + Du_k + v_k \quad (16.1.60)$$

with dimensions  $x_k \in \mathbb{R}^n$ ,  $u_k \in \mathbb{R}^m$  and  $y_k \in \mathbb{R}^p$ . The noise terms  $w_k$  and  $v_k$  are considered to be zero-mean and white. Also assume the system is observable and controllable, so that this means it is a minimal realisation of the system (i.e. there are no ‘redundant’ states), however the order of the system  $n$  may be not known. The subspace identification approach finds estimates for the system matrices  $A$ ,  $B$ ,  $C$  and  $D$  from only input and output observations, and is formulated as follows. Define the stacked output and input vectors

$$Y_{r,k} = \begin{bmatrix} y_k \\ y_{k+1} \\ \vdots \\ y_{k+r-1} \end{bmatrix} \quad (16.1.61)$$

and

$$U_{r,k} = \begin{bmatrix} u_k \\ u_{k+1} \\ \vdots \\ u_{k+r-1} \end{bmatrix} \quad (16.1.62)$$

where  $r$  is known as the maximal prediction horizon. It can be checked that  $Y_{r,k}$  can be written as

$$Y_{r,k} = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{r-1} \end{bmatrix}}_{\mathcal{O}_r} x_k + \underbrace{\begin{bmatrix} D & & & \\ CB & D & & \\ \vdots & \vdots & \ddots & \\ \vdots & \vdots & & \ddots \\ CA^{r-2}B & CA^{r-3}B & \dots & CB & D \end{bmatrix}}_{\mathcal{S}_r} U_{r,k} + V_k \quad (16.1.63)$$

where  $V_k$  is another stacked vector of integrated noise terms given by

$$V_k = \begin{bmatrix} v_k \\ Cw_k + v_{k+1} \\ CAw_k + Cw_{k+1} + v_{k+2} \\ \vdots \\ CA^{r-2}w_k + \dots + Cw_{k+r-2} + v_{k+r-1} \end{bmatrix} \quad (16.1.64)$$

The matrix  $\mathcal{O}_r$  is called the extended observability matrix, and is the matrix of interest in the first step of estimating the system matrices, because it contains information about  $A$  and  $C$ . Suppose we have  $N+r-1$  input and output observations, indexed by  $(u_1, y_1), \dots, (u_{N+r-1}, y_{N+r-1})$ . Then concatenating horizontally, we can write

$$\underbrace{\begin{bmatrix} Y_{r,1} & \dots & Y_{r,N} \end{bmatrix}}_{\mathbf{Y}} = \mathcal{O}_r \underbrace{\begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}}_{\mathbf{X}} + \mathcal{S}_r \underbrace{\begin{bmatrix} U_{r,1} & \dots & U_{r,N} \end{bmatrix}}_{\mathbf{U}} + \underbrace{\begin{bmatrix} V_1 & \dots & V_N \end{bmatrix}}_{\mathbf{V}} \quad (16.1.65)$$

where  $\mathbf{Y}$ ,  $\mathbf{X}$ ,  $\mathbf{U}$ ,  $\mathbf{V}$  are known as block Hankel matrices because their blockwise anti-diagonals contain all the same elements. Define  $\mathbf{M} = I - \mathbf{U}^\top (\mathbf{U}\mathbf{U}^\top)^{-1} \mathbf{U}$  which is a matrix that performs an orthogonal projection to the  $\mathbf{U}$  matrix, since we see that  $\mathbf{U}\mathbf{M} = \mathbf{0}$ . Then multiplying the formula for  $\mathbf{Y}$  above from the right by  $\mathbf{M}$ , the term containing  $\mathbf{U}$  will be eliminated, which leaves

$$\mathbf{Y}\mathbf{M} = \mathcal{O}_r \mathbf{X}\mathbf{M} + \mathbf{V}\mathbf{M} \quad (16.1.66)$$

where  $\mathbf{V}\mathbf{M}$  contains noise terms. If the noise is small, then  $\mathbf{Y}\mathbf{M} \approx \mathcal{O}_r \mathbf{X}\mathbf{M}$  and so information about  $\mathcal{O}_r$  can be found within  $\mathbf{Y}\mathbf{M}$  (which consists of only input/output measurements).

### Instrumental Variable Methods in Subspace Identification [206]

Note that  $V_k$  contains coloured noise terms, and moreover  $\mathbf{X}$  and  $\mathbf{V}$  both contain the influence from process noise  $w_k$ . Thus, estimation of  $\mathcal{O}_r$  suffers from the same problem as that encountered by endogeneity in instrumental variables regression, and we will not be able to consistently estimate  $\mathcal{O}_r$  from  $\mathbf{Y}\mathbf{M}$  alone. We seek a choice of instrumental variables that can de-correlate the terms containing  $\mathbf{X}$  and  $\mathbf{V}$ . Define a ‘suitable’ (to be elaborated later on) instrumental variables matrix  $\mathbf{Z}$  of dimension  $s \times N$ , where  $s \geq n$ . It consists of vectors

$$\mathbf{Z} = [z_{s,1} \ \dots \ z_{s,N}] \quad (16.1.67)$$

Then let

$$G = \frac{1}{N} \mathbf{Y}\mathbf{M}\mathbf{Z}^\top \quad (16.1.68)$$

$$= \frac{1}{N} \mathcal{O}_r \mathbf{X}\mathbf{M}\mathbf{Z}^\top + \frac{1}{N} \mathbf{V}\mathbf{M}\mathbf{Z}^\top \quad (16.1.69)$$

$$= \mathcal{O}_r \underbrace{\frac{1}{N} \mathbf{X}\mathbf{M}\mathbf{Z}^\top}_{\widetilde{T}_N} + \underbrace{\frac{1}{N} \mathbf{V}\mathbf{M}\mathbf{Z}^\top}_{E_N} \quad (16.1.70)$$

where  $\widetilde{T}_N \in \mathbb{R}^{n \times s}$  and  $E_N \in \mathbb{R}^{pr \times s}$ . Also note that the dimension of  $G$  is not growing with  $N$ , which was not the case before with just  $\mathbf{Y}\mathbf{M}$ . Thus another advantage of the instrumental variables method is that we relax the memory requirements by collecting more data. Now suppose that  $\mathbf{Z}$  is suitable in the sense that as  $N \rightarrow \infty$ , we have

$$E_N \xrightarrow{\text{P}} \mathbf{0} \quad (16.1.71)$$

$$\widetilde{T}_N \xrightarrow{\text{P}} \widetilde{T} \quad (16.1.72)$$

where  $\widetilde{T}$  is a full (row) rank matrix of rank  $n$ . To examine what makes a suitable choice of  $\mathbf{Z}$ , we first write out  $E_N$  in terms of sums.

$$E_N = \frac{1}{N} [V_1 \ \dots \ V_N] \left[ I - \mathbf{U}^\top (\mathbf{U}\mathbf{U}^\top)^{-1} \mathbf{U} \right] \begin{bmatrix} z_{s,1}^\top \\ \vdots \\ z_{s,N}^\top \end{bmatrix} \quad (16.1.73)$$

$$= \frac{1}{N} \sum_{k=1}^N V_k z_{s,k}^\top - \frac{1}{N} \sum_{k=1}^N V_k U_{r,k}^\top \left( \frac{1}{N} \sum_{k=1}^N U_{r,k} U_{r,k}^\top \right)^{-1} \frac{1}{N} \sum_{k=1}^N U_{r,k} z_{s,k}^\top \quad (16.1.74)$$

Under the conditions of wide sense ergodicity, the probability limit (assuming an asymptotically stable system under zero-noise) is

$$E_N \xrightarrow{\text{P}} \lim_{k \rightarrow \infty} \left( \mathbb{E} [V_k z_{s,k}^\top] - \mathbb{E} [V_k U_{r,k}^\top] \left( \mathbb{E} [U_{r,k} U_{r,k}^\top] \right)^{-1} \mathbb{E} [U_{r,k} z_{s,k}^\top] \right) \quad (16.1.75)$$

As the noise is zero-mean, then if the noise and input are independent (as would be the case if the input is open-loop), then

$$\mathbb{E} [V_k U_{r,k}^\top] = \mathbf{0} \quad (16.1.76)$$

Also assume that  $\mathbb{E} [U_{r,k} U_{r,k}^\top]$  is invertible (which means the input should be persistently exciting). Then in order for the probability limit to be zero, we require  $V_k$  and  $z_{s,k}$  to be uncorrelated. Since  $V_k$  contains only terms for times on or after time  $k$ , then an example choice of  $z_{s,k}$  which

satisfies uncorrelatedness are the inputs and outputs which happen before time  $k$ :

$$z_{s,k} = \begin{bmatrix} y_{k-1} \\ \vdots \\ y_{k-s_1} \\ u_{k-1} \\ \vdots \\ u_{k-s_2} \end{bmatrix} \quad (16.1.77)$$

with  $s_1, s_2$  such that  $ps_1 + ms_2 = s$ . Note that this means we now need a total of  $N + r + s - 1$  observations. A simple choice in implementation is to let  $s_1 = s_2$ , and then also fix  $r = s$ . Diverting our attention to the probability limit of  $\tilde{T}_N$ , we similarly find that

$$\tilde{T}_N = \frac{1}{N} \sum_{k=1}^N x_k z_{s,k}^\top - \frac{1}{N} \sum_{k=1}^N x_k U_{r,k}^\top \left( \frac{1}{N} \sum_{k=1}^N U_{r,k} U_{r,k}^\top \right)^{-1} \frac{1}{N} \sum_{k=1}^N U_{r,k} z_{s,k}^\top \quad (16.1.78)$$

and under the conditions of the law of large numbers for correlated sequences,

$$\tilde{T}_N \xrightarrow{P} \lim_{k \rightarrow \infty} \left( \mathbb{E}[x_k z_{s,k}^\top] - \mathbb{E}[x_k U_{r,k}^\top] (\mathbb{E}[U_{r,k} U_{r,k}^\top])^{-1} \mathbb{E}[U_{r,k} z_{s,k}^\top] \right) \quad (16.1.79)$$

It can be established that  $\tilde{T}$  is full-rank with the following sufficient conditions:

- $\mathbb{E}[z_{s,k} z_{s,k}^\top]$  is positive definite (hence full-rank).
- The horizons  $s_1$  and  $s_2$  are sufficiently large such that the states can be well-estimated from the past input and output as  $\mathbb{E}[x_k | z_{s,k}] = L z_{s,k}$  for some full rank matrix  $L$ . The matrix  $L$  can be thought of as the coefficients in a linear predictor.
- The input sequence is zero-mean white noise.

To show this, use the Law of Iterated Expectations:

$$\mathbb{E}[x_k z_{s,k}^\top] = \mathbb{E}\left[\mathbb{E}[x_k z_{s,k}^\top | z_{s,k}]\right] \quad (16.1.80)$$

$$= \mathbb{E}\left[\mathbb{E}[x_k | z_{s,k}] z_{s,k}^\top\right] \quad (16.1.81)$$

$$= L \mathbb{E}[z_{s,k} z_{s,k}^\top] \quad (16.1.82)$$

Furthermore by the input sequence assumption,  $\mathbb{E}[U_{r,k} z_{s,k}^\top] = 0$  so

$$\mathbb{E}[x_k z_{s,k}^\top] - \mathbb{E}[x_k U_{r,k}^\top] (\mathbb{E}[U_{r,k} U_{r,k}^\top])^{-1} \mathbb{E}[U_{r,k} z_{s,k}^\top] = L \mathbb{E}[z_{s,k} z_{s,k}^\top] \quad (16.1.83)$$

As  $L$  and  $\mathbb{E}[z_{s,k} z_{s,k}^\top]$  are full rank,  $\tilde{T}_N$  converges in probability to a full rank matrix.

### Estimation with Subspace Identification

Using the instrumental variables method, we have established that  $G = \frac{1}{N} \mathbf{Y} \mathbf{M} \mathbf{Z}^\top$  is a noisy estimate of the matrix  $\mathcal{O}_r \tilde{T}_N$ , that is related to the extended observability matrix  $\mathcal{O}_r$ , with the noise converging in probability to zero and  $\tilde{T}_N$  converging in probability to a full rank matrix as  $N \rightarrow \infty$ . To recover an estimate of  $\mathcal{O}_r$ , we can perform a singular value decomposition on  $G$ , which gives

$$G = \mathcal{U} \Sigma \mathcal{V}^\top \quad (16.1.84)$$

which takes the form (with  $r = s$  and  $p \geq 1$ ):

$$G = \mathcal{U} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ 0 & \dots & \sigma_s & \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix} \mathcal{V}^\top \quad (16.1.85)$$

where  $\mathcal{U} \in \mathbb{R}^{pr \times pr}$  and  $\mathcal{V} \in \mathbb{R}^{s \times s}$  are orthogonal. If the order of the system is known (or predetermined), then we can further write this as

$$G = [\mathcal{U}_1 \quad \mathcal{U}_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} \mathcal{V}_1^\top \\ \mathcal{V}_2^\top \end{bmatrix} \quad (16.1.86)$$

with dimensions  $\mathcal{U}_1 \in \mathbb{R}^{pr \times n}$ ,  $\mathcal{V}_1 \in \mathbb{R}^{n \times s}$  and  $\Sigma_1 \in \mathbb{R}^{n \times n}$ . By convention,  $\Sigma_1$  contains the  $n$  largest singular values of  $G$ . If the true order of the system is  $n$  and the noise is small, it is reasonable to think that the singular values in  $\Sigma_2$  will also be small, leading to

$$G \approx \widehat{G} = \mathcal{U}_1 \Sigma_1 \mathcal{V}_1^\top \quad (16.1.87)$$

From another characterisation of SVD, this can be viewed as taking a low-rank approximation of  $G$ , with the objective to minimise the Frobenius norm between  $G$  and  $\widehat{G}$ , with the constraint that  $\text{rank}(\widehat{G}) = n$ . If the order of the system is not known or predetermined, then it may be estimated by examining the first  $n$  singular values which are ‘significantly’ non-zero. Thus for large  $N$  we have approximately

$$G \approx \mathcal{O}_r \widetilde{T} \quad (16.1.88)$$

$$\mathcal{U}_1 \Sigma_1 \mathcal{V}_1^\top \approx \mathcal{O}_r \widetilde{T} \quad (16.1.89)$$

where  $\widetilde{T}$  is full rank. Multiplying from the right by  $\mathcal{V}_1$ , we use that  $\mathcal{V}_1^\top \mathcal{V}_1 = I$  by orthogonality of  $\mathcal{V}$  to give

$$\mathcal{U}_1 \Sigma_1 \approx \mathcal{O}_r \widetilde{T} \mathcal{V}_1 \quad (16.1.90)$$

where  $T = \widetilde{T} \mathcal{V}_1$  is a square matrix that is also invertible, since it will have rank  $n$  (preserved by multiplication with  $\mathcal{V}_1$  which has linearly independent columns). Thus  $T$  is a similarity transform which only affects the choice of basis of the state-space representation. We can estimate the extended observability matrix up to a similarity transform, and can simply take

$$\widehat{\mathcal{O}}_r = \mathcal{U}_1^\top \Sigma_1 \quad (16.1.91)$$

For further generality, we could postmultiply  $G$  by some matrix  $\mathcal{W}$  so that  $G\mathcal{W} \approx \mathcal{O}_r \widetilde{T}\mathcal{W}$ . Then by SVD to take the reduced-rank approximation:  $G\mathcal{W} \approx \mathcal{U}_1 \Sigma_1 \mathcal{V}_1^\top$  will still recover an estimate of the extended observability matrix up to a similarity transform (under the appropriate rank conditions on  $\mathcal{W}$ ), because  $\mathcal{O}_r \widetilde{T}\mathcal{W}\mathcal{V}_1 \approx \mathcal{U}_1 \Sigma_1$ . For even further generality, we can use  $\widehat{\mathcal{O}}_r = \mathcal{U}_1 \mathcal{R}$  where  $\mathcal{R}$  is any invertible matrix.

From the structure of the extended observability matrix, the estimate  $\widehat{C}$  can be obtained from inspection by taking the first  $p$  rows of  $\widehat{\mathcal{O}}_r$ . To estimate  $A$ , notice that

$$\underbrace{\begin{bmatrix} CA \\ \vdots \\ CA^{r-1} \end{bmatrix}}_{\mathcal{O}_{2:r}} = \underbrace{\begin{bmatrix} C \\ \vdots \\ CA^{r-2} \end{bmatrix}}_{\mathcal{O}_{1:(r-1)}} A \quad (16.1.92)$$

where  $\mathcal{O}_{2:r}$  and  $\mathcal{O}_{1:(r-1)}$  may be estimated respectively as  $\widehat{\mathcal{O}}_{2:r}$  and  $\widehat{\mathcal{O}}_{1:(r-1)}$  by inspection of  $\widehat{\mathcal{O}}_r$ . So an estimate of  $A$  in the least-squares sense may be given by

$$\widehat{A} = \widehat{\mathcal{O}}_{1:(r-1)}^\dagger \widehat{\mathcal{O}}_{2:r} \quad (16.1.93)$$

where  $\widehat{\mathcal{O}}_{1:(r-1)}^\dagger$  is the Moore-Penrose pseudoinverse of  $\widehat{\mathcal{O}}_{1:(r-1)}$ . Subsequently,  $B$  and  $D$  may be estimated by forming the equation

$$\widehat{y}_k = \widehat{C} \left( qI - \widehat{A} \right)^{-1} Bu_k + Du_k \quad (16.1.94)$$

by recalling that  $C(qI - A)^{-1}B + D$  is the formula for the discrete-time transfer function, where  $q$  is the forward shift in time operator. Optionally, the initial condition  $x_0$  (where it is originally presumed to be zero) can also be estimated by constructing the form

$$\widehat{y}_k = \widehat{C} \left( qI - \widehat{A} \right)^{-1} x_0 \delta_k + \widehat{C} \left( qI - \widehat{A} \right)^{-1} Bu_k + Du_k \quad (16.1.95)$$

where  $\delta_k$  is the Kronecker delta (unit pulse at time  $k = 0$ ). To reformulate this into a form suitable for linear regression, we put the data into ‘predictor form’ in terms of our already obtained estimates  $\widehat{C}$ ,  $\widehat{A}$ , and our to-be estimated  $x_0$ ,  $B$  and  $D$ :

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \widehat{C}\widehat{A} \\ \widehat{C}\widehat{A}^2 \\ \vdots \\ \widehat{C}\widehat{A}^N \end{bmatrix}}_{\Xi} x_0 + \underbrace{\begin{bmatrix} \widehat{C} & & & \\ \widehat{C}\widehat{A} & \widehat{C} & & \\ \vdots & \vdots & \ddots & \\ \widehat{C}\widehat{A}^{N-1} & \widehat{C}\widehat{A}^{N-2} & \dots & \widehat{C} \end{bmatrix}}_{\Omega} \begin{bmatrix} Bu_0 \\ Bu_1 \\ \vdots \\ Bu_{N-1} \end{bmatrix} + \begin{bmatrix} Du_1 \\ Du_2 \\ \vdots \\ Du_N \end{bmatrix} + \varepsilon \quad (16.1.96)$$

where  $\varepsilon$  contains arbitrary noise terms. Note that

$$\begin{bmatrix} Bu_0 \\ \vdots \\ Bu_{N-1} \end{bmatrix} = \text{vec} (I_{n \times n} B [u_0 \ \dots \ u_{N-1}]) \quad (16.1.97)$$

and so using the ‘vec trick’ which says  $(\mathbf{B}^\top \otimes \mathbf{A}) \text{vec}(\mathbf{C}) = \text{vec}(\mathbf{ACB})$ , we can write

$$\Omega \begin{bmatrix} Bu_0 \\ \vdots \\ Bu_{N-1} \end{bmatrix} = \Omega \left( [u_0 \ \dots \ u_{N-1}]^\top \otimes I_{n \times n} \right) \text{vec}(B) \quad (16.1.98)$$

A similar thing can be done for  $\text{vec}(D)$ , which gives us the system of equations

$$\mathbf{y} = \Xi x_0 + \underbrace{\Omega \left( [u_0 \ \dots \ u_{N-1}]^\top \otimes I_{n \times n} \right)}_{\Gamma} \text{vec}(B) + \underbrace{\left( [u_1 \ \dots \ u_N]^\top \otimes I_{p \times p} \right)}_{\Delta} \text{vec}(D) + \varepsilon \quad (16.1.99)$$

Collecting all the unknown parameters to be estimated, this becomes

$$\mathbf{y} = \underbrace{[\Xi \ \Gamma \ \Delta]}_{\Psi} \underbrace{\begin{bmatrix} x_0 \\ \text{vec}(B) \\ \text{vec}(D) \end{bmatrix}}_{\Theta} + \varepsilon \quad (16.1.100)$$

with ordinary least squares solution

$$\widehat{\Theta} = (\Psi^\top \Psi)^{-1} \Psi^\top \mathbf{y} \quad (16.1.101)$$

from which we can extract the estimates  $\widehat{B}$ ,  $\widehat{D}$ ,  $\widehat{x}_0$ . To summarise the steps for subspace identification:

1. From input and output data, form  $G = \frac{1}{N} \mathbf{Y} \mathbf{M} \mathbf{Z}^\top$ .
2. Take the SVD of  $G$  and approximate it by  $\widehat{G} = \mathcal{U}_1 \Sigma_1 \mathcal{V}_1^\top$  using some choice of the model order  $n$ .
3. Estimate the extended observability matrix by  $\widehat{\mathcal{O}}_r = \mathcal{U}_1 \Sigma_1$ .
4. From  $\widehat{\mathcal{O}}_r$ , estimate  $C$  by inspection followed by estimation of  $A$  using least squares.
5. Using  $\widehat{A}$  and  $\widehat{C}$  together with the input and output data, estimate  $B$ ,  $D$  and optionally  $x_0$  by least squares.

### Consistency of Subspace Identification

Under the stated conditions, the extended observability matrix  $\mathcal{O}_r$  can be consistently estimated up to a similarity transform. This implies that  $A$  and  $C$  can also be consistently estimated (the former relies on the consistency properties of least squares). Then if consistent estimates for  $\widehat{A}$  and  $\widehat{C}$  are used, then consistency for  $\widehat{B}$  and  $\widehat{D}$  are guaranteed by assuming that the input is run in open-loop (which causes  $u_k$  to be independent with the noise terms). However note that consistency for  $x_0$  is harder to achieve because intuitively, collecting more data in the future will not give much information about the initial condition.

### Noise Statistics with Subspace Identification

The covariances of the noise terms  $w_k$  and  $v_k$  can be obtained from the subspace identification approach. Firstly, construct a predictor of the form

$$Y_{r,k} = \Pi z_{s,k} + \Phi U_{r,k} + E_k \quad (16.1.102)$$

where  $Y_{r,k}$ ,  $U_{r,k}$ ,  $z_{s,k}$  are as above,  $E_k$  is a stacked vector of error terms and  $\Pi$ ,  $\Phi$  contain parameters. Collecting these vectors horizontally for  $k = 1, \dots, N$  gives the model

$$\mathbf{Y} = \Pi \mathbf{Z} + \Phi \mathbf{U} + \mathbf{E} \quad (16.1.103)$$

$$= [\Pi \quad \Phi] \begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \end{bmatrix} + \mathbf{E} \quad (16.1.104)$$

The least squares estimates for  $\Pi$  and  $\Phi$  are obtained from the Moore-Penrose pseudoinverse of the partitioned matrix:

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \end{bmatrix}^\dagger = [\mathbf{Z}^\top \quad \mathbf{U}^\top] \left( \begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \end{bmatrix} [\mathbf{Z}^\top \quad \mathbf{U}^\top] \right)^{-1} \quad (16.1.105)$$

$$= [\mathbf{Z}^\top \quad \mathbf{U}^\top] \begin{bmatrix} \mathbf{Z} \mathbf{Z}^\top & \mathbf{Z} \mathbf{U}^\top \\ \mathbf{U} \mathbf{Z}^\top & \mathbf{U} \mathbf{U}^\top \end{bmatrix}^{-1} \quad (16.1.106)$$

giving

$$\begin{bmatrix} \widehat{\Pi} & \widehat{\Phi} \end{bmatrix} = \mathbf{Y} [\mathbf{Z}^\top \quad \mathbf{U}^\top] \begin{bmatrix} \mathbf{Z} \mathbf{Z}^\top & \mathbf{Z} \mathbf{U}^\top \\ \mathbf{U} \mathbf{Z}^\top & \mathbf{U} \mathbf{U}^\top \end{bmatrix}^{-1} \quad (16.1.107)$$

$$= [\mathbf{Y} \mathbf{Z}^\top \quad \mathbf{Y} \mathbf{U}^\top] \begin{bmatrix} \mathbf{Z} \mathbf{Z}^\top & \mathbf{Z} \mathbf{U}^\top \\ \mathbf{U} \mathbf{Z}^\top & \mathbf{U} \mathbf{U}^\top \end{bmatrix}^{-1} \quad (16.1.108)$$

Using block matrix inversion formulae and isolating only the blocks which are used in the calculation of  $\widehat{\Pi}$ :

$$\begin{bmatrix} \mathbf{Z} \mathbf{Z}^\top & \mathbf{Z} \mathbf{U}^\top \\ \mathbf{U} \mathbf{Z}^\top & \mathbf{U} \mathbf{U}^\top \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{Z} \mathbf{Z}^\top - \mathbf{Z} \mathbf{U}^\top (\mathbf{U} \mathbf{U}^\top)^{-1} \mathbf{U} \mathbf{Z}^\top)^{-1} & * \\ -(\mathbf{U} \mathbf{U}^\top)^{-1} \mathbf{U} \mathbf{Z}^\top (\mathbf{Z} \mathbf{Z}^\top - \mathbf{Z} \mathbf{U}^\top (\mathbf{U} \mathbf{U}^\top)^{-1} \mathbf{U} \mathbf{Z}^\top)^{-1} & * \end{bmatrix} \quad (16.1.109)$$

$$= \begin{bmatrix} (\mathbf{ZM}\mathbf{Z}^\top)^{-1} & * \\ -(\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{Z}^\top(\mathbf{ZM}\mathbf{Z}^\top)^{-1} & * \end{bmatrix} \quad (16.1.110)$$

where  $*$  represents arbitrary elements we are not concerned about, and in the second step we have used the definition of the orthogonal projection matrix  $\mathbf{M} = I - \mathbf{U}^\top(\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}$  defined above. Thus the computation of  $\widehat{\Pi}$  is given by

$$\widehat{\Pi} = [\mathbf{Y}\mathbf{Z}^\top \quad \mathbf{Y}\mathbf{U}^\top] \begin{bmatrix} (\mathbf{ZM}\mathbf{Z}^\top)^{-1} \\ -(\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{Z}^\top(\mathbf{ZM}\mathbf{Z}^\top)^{-1} \end{bmatrix} \quad (16.1.111)$$

$$= [\mathbf{Y}\mathbf{Z}^\top(\mathbf{ZM}\mathbf{Z}^\top)^{-1} - \mathbf{Y}\mathbf{U}^\top(\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{Z}^\top(\mathbf{ZM}\mathbf{Z}^\top)^{-1}] \quad (16.1.112)$$

$$= \mathbf{Y}\mathbf{M}\mathbf{Z}^\top(\mathbf{ZM}\mathbf{Z}^\top)^{-1} \quad (16.1.113)$$

where we have again used the definition of  $\mathbf{M}$ . Note that the form of this estimate bears similarity to the form of the estimates for the partitioned parameters in the Frisch-Waugh-Lovell theorem from econometrics (which also involves an orthogonal projection matrix). Then  $\mathbf{Y}$  can be predicted in the following way:

$$\widehat{\mathbf{Y}} = \widehat{\Pi}\mathbf{Z} \quad (16.1.114)$$

$$= \mathbf{Y}\mathbf{M}\mathbf{Z}^\top(\mathbf{ZM}\mathbf{Z}^\top)^{-1}\mathbf{Z} \quad (16.1.115)$$

We explain that  $\widehat{\Phi}$  drops out of the prediction because it is not sensible to use future values of the input to predict past values of the output (e.g. using  $u_{k+r-1}$  to predict  $y_k$ ). Moreover, the ‘best prediction’ in some sense can be viewed as when the expected future inputs are taken to be zero (for instance if the input sequence is zero-mean white noise). We then see that  $\widehat{\mathbf{Y}}$  is a case of the post-multiplication  $G\mathcal{W}$  where  $\mathcal{W} = N(\mathbf{ZM}\mathbf{Z}^\top)^{-1}\mathbf{Z}$  which is a  $s \times N$  matrix. Thus  $\widehat{\mathbf{Y}}$  can be decomposed by SVD to recover  $\widehat{A}$ ,  $\widehat{B}$ ,  $\widehat{C}$ ,  $\widehat{D}$  and optionally  $\widehat{x}_0$ . Then through  $\mathbf{Y} = \mathcal{O}_r\mathbf{X} + \mathcal{S}_r\mathbf{U} + \mathbf{V}$ , we obtain

$$\widehat{\mathbf{Y}} \approx \mathcal{O}_r\widehat{\mathbf{X}} \quad (16.1.116)$$

where  $\mathbf{U}$  drops out in the prediction due to the reasoning explained above. As  $\widehat{\mathcal{O}}_r$  can be written as  $\mathcal{U}_1\mathcal{R}$  for some invertible  $\mathcal{R}$ , then exploiting the orthogonality  $\mathcal{U}_1^\top\mathcal{U}_1 = I$  gives

$$\widehat{\mathbf{X}} = \mathcal{R}^{-1}\mathcal{U}_1^\top\widehat{\mathbf{Y}} \quad (16.1.117)$$

which gives the state estimates in the same state-space basis as that used to estimate the system matrices. This also shows the state estimates can be written in the form  $\widehat{\mathbf{X}} = L\mathbf{Z}$  as

$$\widehat{\mathbf{X}} = \mathcal{R}^{-1}\mathcal{U}_1^\top\mathbf{Y}\mathbf{M}\mathbf{Z}^\top(\mathbf{ZM}\mathbf{Z}^\top)^{-1}\mathbf{Z} \quad (16.1.118)$$

which is relatively to one of the requirements for  $\widetilde{T}$  to be full rank in the instrumental variables method. Lastly from the state estimates, the noises can be estimated by

$$\widehat{w}_k = \widehat{x}_{k+1} - \widehat{A}\widehat{x}_k - \widehat{B}u_k \quad (16.1.119)$$

$$\widehat{v}_k = y_k - \widehat{C}\widehat{x}_k - \widehat{D}u_k \quad (16.1.120)$$

from which their respective covariance estimates  $\widehat{Q}$  and  $\widehat{R}$  can be computed (for example, by taking the sample covariance).

### Subspace Identification for Specified Realisations

Suppose we want to estimate a model of the form

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (16.1.121)$$

where the state-space realisation  $x_k$  is pre-specified. Then we require data for  $x_k$  and  $u_k$  (essentially,  $x_k$  is treated as the output). Subspace identification algorithms can be adapted in the following way to estimate  $A$  and  $B$ . Firstly  $n$  is fixed as the dimension of  $x_k$ . We then obtain the usual estimates for  $(A', B', C', D')$  via subspace identification for some arbitrary state-space realisation

$$\tilde{x}_{k+1} = A'\tilde{x}_k + B'u_k + \tilde{w}_k \quad (16.1.122)$$

$$y_k = C'\tilde{x}_k + D'u_k + v_k \quad (16.1.123)$$

Recognise that treating  $x_k$  as the output means that we want to find a change of coordinates such that  $C = I$  in the new coordinate system. It turns out the appropriate thing to do is to use the coordinate transform  $x_k = C'\tilde{x}_k$  or  $\tilde{x}_k = (C')^{-1}x_k$ . Then this yields

$$(C')^{-1}x_k = A'(C')^{-1}\tilde{x}_k + B'u_k + \tilde{w}_k \quad (16.1.124)$$

$$x_k = C'A'(C')^{-1}\tilde{x}_k + C'B'u_k + C'\tilde{w}_k \quad (16.1.125)$$

This shows that we should take our estimates of  $A$  and  $B$  as

$$\hat{A} = \hat{C}'\hat{A}'(\hat{C}')^{-1} \quad (16.1.126)$$

$$\hat{B} = \hat{C}'\hat{\tilde{B}} \quad (16.1.127)$$

## 16.2 Queueing Theory [127]

### 16.3 Stochastic Stability

#### 16.3.1 Moment Stability [5]

##### Stability in Mean

Let  $X(t)$  be the stochastic processes which is the solution to a system of stochastic difference equations (so  $X(t)$  can generally be vector valued). We assume that  $X(t) = 0$  is the unique solution to the undisturbed differential equation with zero initial condition, so that the origin is an equilibrium point. Then the system is said to be *stable in mean* (or in expectation) if for each  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that

$$\sup_{t \geq 0} \mathbb{E} [\|X(t)\|] \leq \varepsilon \quad (16.3.1)$$

for initial conditions  $\|x_0\| \leq \delta$ . This definition stays true to the ‘start close, remain close’ characterisation of stability for ordinary differential equations (albeit ‘on average’ for the stochastic analogue). Moreover, we say the system is *asymptotically stable in mean* if

$$\lim_{t \rightarrow \infty} \mathbb{E} [\|X(t)\|] \quad (16.3.2)$$

for all initial conditions  $x_0$  in the neighbourhood of the origin. Additionally, when the above condition holds for all  $x_0 \in \mathbb{R}^n$ , then we can say that the system is *globally asymptotically stable in mean*, or also *asymptotically stable in the large*.

### Stability in Mean-Square

A system is *stable in mean-square* if for each  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that

$$\sup_{t \geq 0} \mathbb{E} [\|X(t)\|^2] \leq \varepsilon \quad (16.3.3)$$

for all initial conditions  $\|x_0\| \leq \delta$ . Moreover, it is *asymptotically stable in mean-square* if

$$\lim_{t \rightarrow \infty} \mathbb{E} [\|X(t)\|^2] \quad (16.3.4)$$

for all initial conditions  $x_0$  in the neighbourhood of the origin. In the same way that convergence in mean-square implies convergence in mean, we have that stability in mean-square implies stability in mean.

### Stability in $p$ -Mean

Stability in mean ( $p = 1$ ) and mean-square ( $p = 2$ ) can be generalised to stability in  $p$ -mean, for  $p > 0$ , with analogous stability condition

$$\sup_{t \geq 0} \mathbb{E} [\|X(t)\|^p] \leq \varepsilon \quad (16.3.5)$$

and asymptotic stability condition

$$\lim_{t \rightarrow \infty} \mathbb{E} [\|X(t)\|^p] \quad (16.3.6)$$

Likewise with convergence in  $p$ -mean, it can be shown that for  $1 \leq q \leq p$ , stability in  $p$ -mean implies stability in  $q$ -mean.

### Stability in Expectation Value

A related but not identical concept to stability in mean is stability in expectation value. Define the mean process  $\mu(t) = \mathbb{E}[X(t)]$ . Then the system is *stable in expectation value* if for each  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that for all  $\|x_0\| \leq \delta$ , this implies

$$\sup_{t \geq 0} \|\mu(t)\| \leq \varepsilon \quad (16.3.7)$$

Since  $\|\mathbb{E}[X(t)]\| \leq \mathbb{E}[\|X(t)\|]$  by Jensen's inequality, stability in mean implies stability in expectation value, however the converse is not necessarily true.

### Stability in Second Moment

Define the second moment process  $M(t) = \mathbb{E}[X(t) X(t)^\top]$ . Then the system is *stable in second moment* if for each  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that for all  $\|x_0\| \leq \delta$ , this implies

$$\sup_{t \geq 0} \|M(t)\| \leq \varepsilon \quad (16.3.8)$$

Stability in mean-square can be shown to be equivalent to stability in second moment. On one hand, we have

$$\mathbb{E} [\|X(t)\|^2] = \mathbb{E} [X(t)^\top X(t)] \quad (16.3.9)$$

$$= \mathbb{E} [\text{trace}(X(t) X(t)^\top)] \quad (16.3.10)$$

$$= \text{trace}(M(t)) \quad (16.3.11)$$

We know  $\|M(t)\| \leq \text{trace}(M(t))$  because  $\|M(t)\|$  is the largest eigenvalue of  $M(t)$  while  $\text{trace}(M(t))$  is the sum of eigenvalues. Hence this says that stability in mean-square implies stability in second moment. On the other hand, using the same characterisation via the sum of eigenvalues, we can also upper bound

$$\text{trace}(M(t)) \leq n \|M(t)\| \quad (16.3.12)$$

Hence stability in second moment also implies stability in mean-square. This completes showing their equivalence (so that the definition of stability in second moment can be used in place of the definition of stability in mean-square [188]).

### 16.3.2 Stability in Probability

#### Strong Stability in Probability

Let  $X(t)$  be the vector-valued stochastic processes which is the solution to a system of stochastic difference equations. Assume that  $X(t) = 0$  is the unique solution to the undisturbed differential equation with zero initial condition, so that the origin is an equilibrium point. Then the system is said to be (strongly) *stable in probability* [117] if for any  $\varepsilon > 0$  and  $\rho > 0$ , there exists some  $\delta(\rho, \varepsilon) > 0$  such that

$$\Pr\left(\sup_{t \geq 0} \|X(t)\| \geq \varepsilon\right) \leq \rho \quad (16.3.13)$$

for all initial conditions  $\|x_0\| \leq \delta(\rho, \varepsilon)$ . Because  $\rho$  can be made arbitrarily small, this leads to an alternative definition [5] that for all  $\varepsilon > 0$ :

$$\lim_{x_0 \rightarrow 0} \Pr\left(\sup_{t \geq 0} \|X(t)\| \geq \varepsilon\right) = 0 \quad (16.3.14)$$

A worded characterisation [110] of this is that all sample paths of  $X(t)$  issuing from point  $x_0$  at time zero will remain in a neighbourhood of the origin with probability converging to one as  $x_0 \rightarrow 0$ . The system is then said to be *unstable in probability* if it is not stable in probability. In the same was as for convergence in probability, stability in  $p$ -mean for  $p \geq 1$  implies stability in probability.

#### Asymptotic Stability in Probability

A system stable in probability is said to be (strongly) *asymptotically stable in probability* [5] if

$$\lim_{x_0 \rightarrow 0} \Pr\left(\lim_{t \rightarrow \infty} \|X(t)\| = 0\right) = 1 \quad (16.3.15)$$

Moreover, it is *globally asymptotically stable in probability* or *asymptotically stable in probability in the large* if

$$\Pr\left(\lim_{t \rightarrow \infty} \|X(t)\| = 0\right) = 1 \quad (16.3.16)$$

for all  $x_0 \in \mathbb{R}^n$ .

#### Weak Stability in Probability

A system is said to be *weakly stable in probability* [110] if for any  $\varepsilon > 0$  and  $\rho > 0$ , there exists some  $\delta(\rho, \varepsilon) > 0$  such that

$$\sup_{t \geq 0} \Pr(\|X(t)\| \geq \varepsilon) \leq \rho \quad (16.3.17)$$

for all initial conditions  $\|x_0\| \leq \delta(\rho, \varepsilon)$ . Note that the supremum is now outside the probability rather than inside. Since  $\rho$  can be made arbitrarily small, we get the alternative definition [5] that

$$\lim_{x_0 \rightarrow 0} \left( \sup_{t \geq 0} \Pr (\|X(t)\| \geq \varepsilon) \right) = 0 \quad (16.3.18)$$

To see why this notion is weaker than strong stability in probability, note that

$$\Pr (\|X(t)\| \geq \varepsilon) \leq \Pr \left( \sup_{t \geq 0} \|X(t)\| \geq \varepsilon \right) \quad (16.3.19)$$

for any  $t \geq 0$ . Hence

$$\sup_{t \geq 0} \Pr (\|X(t)\| \geq \varepsilon) \leq \Pr \left( \sup_{t \geq 0} \|X(t)\| \geq \varepsilon \right) \quad (16.3.20)$$

### 16.3.3 Stochastic Input-Output Stability

It is possible to construct stochastic definitions of stability for input-output systems. Let  $y(t)$  be the output of a stochastic system to the input  $u(t)$ . Assume that it is possible to take the  $L_p$  norm of sample paths of these, so that

$$\|y(t)\|_p = \left( \int_0^t \|y(\tau)\|^p d\tau \right)^{1/p} \quad (16.3.21)$$

$$\|u(t)\|_p = \left( \int_0^t \|u(\tau)\|^p d\tau \right)^{1/p} \quad (16.3.22)$$

for  $p \geq 1$  (so we may need to impose the requirement that  $\|y(t)\|^p$  and  $\|u(t)\|^p$  are mean-square integrable). Then the system is said to be  $L_p$  input-output stable in mean with gain  $\gamma$  if there exists a  $K \geq 0$  such that

$$\mathbb{E} [\|y(t)\|_p] \leq K \|x_0\| + \gamma \mathbb{E} [\|u(t)\|_p] \quad (16.3.23)$$

where  $x_0$  is the initial state of the system.

### 16.3.4 Almost Sure Stability [110]

Any deterministic notion of stability can be generalised into a probabilistic notion of almost sure stability. A solution  $X(t)$  to a system is said to be almost surely stable in an appropriate sense if almost all sample paths of  $X(t)$  are stable in the same deterministic sense.

### 16.3.5 Markov Chain Stability [138]

## 16.4 Stochastic Games

### 16.4.1 Non-Sequential Decision under Uncertainty [20]

Consider a decision-making framework in which we have the following sets:

- A set  $\mathcal{D}$  of decisions.
- A set  $\mathcal{N}$  representing all the possible actions of ‘nature’.
- A set  $\Omega$  of outcomes.

These sets are related by a function  $f : \mathcal{D} \times \mathcal{N} \rightarrow \Omega$  which maps each possible decision and action of nature to an outcome. We use  $\preceq$  to denote a binary relation on  $\Omega$ , which is totally ordered (i.e. any two outcomes can be compared and any three outcomes satisfy transitivity). This relation  $\preceq$  induces a *payoff function*  $G : \Omega \rightarrow \mathbb{R}$  which assigns a numerical value (payoff) to each outcome. This payoff function must satisfy the property that  $G(\omega_1) \leq G(\omega_2)$  implies  $\omega_1 \preceq \omega_2$  and vice-versa for any two  $\omega_1, \omega_2 \in \Omega$ . Denote the compound function  $J = G \circ f$ :

$$J(d, n) := G(f(d, n)) \quad (16.4.1)$$

which maps decisions and actions of nature to payoffs. We see that if  $\mathcal{N}$  has a single element, then  $\mathcal{D}$  is totally ordered (it is always possible to compare any two decisions). However if  $\mathcal{N}$  has multiple elements, then  $\mathcal{D}$  will in general be only partially ordered (it is not always possible to compare any two decisions because it might depend on the value of  $d$ ).

### Min-Max Approach of Ranking Decisions

We can obtain a notion of the ‘best’ decision with a min-max formulation. We say that for two decisions  $d_1, d_2 \in \mathcal{D}$ , we have  $d_1 \preceq d_2$  if and only if

$$\inf_{n \in \mathcal{N}} J(d_1, n) \leq \inf_{n \in \mathcal{N}} J(d_2, n) \quad (16.4.2)$$

That is, we evaluate each decision by the payoff the worst possible action of nature that could occur given that decision.

### Expected Utility Approach of Ranking Decisions

Suppose there is some probability distribution for the actions of nature  $\Pr(N = n)$  on support  $\mathcal{N}$  (assume that  $\mathcal{N}$  is finite or countable). Then this induces a probability distribution over outcomes for a given decisions. Denote this by

$$P_d(\omega) = \Pr(N \in \{n : f(d, n) = \omega\} | D = d) \quad (16.4.3)$$

where  $\{n : f(d, n) = \omega\}$  is the subset of all actions of nature which lead to outcome  $\omega$  given decision  $d$ . Introduce the utility function  $U : \Omega \rightarrow \mathbb{R}$  which takes the same role as the payoff function as above. Then a ranking of all decisions via an expected utility approach is to say that  $d_1 \preceq d_2$  if and only if

$$\mathbb{E}[U(f(d_1, N)) | D = d_1] \leq \mathbb{E}[U(f(d_2, N)) | D = d_2] \quad (16.4.4)$$

where expected utility is computed by

$$\mathbb{E}[U(f(d, N)) | D = d] = \sum_{\omega \in \Omega} U(\omega) P_d(\omega) \quad (16.4.5)$$

The expected utility maximisation problem is to find the decision which yields the highest expected utility:

$$d^* = \max_{d \in \mathcal{D}} \mathbb{E}[U(f(d, N)) | D = d] \quad (16.4.6)$$

### 16.4.2 Multi-Player Stochastic Games [146]

Minimax Theorem [17]

Minimax Algorithm [173]

### 16.4.3 Stochastic Dynamic Games

## 16.5 Stochastic Dynamic Programming

### 16.5.1 Stochastic Programming [180]

Consider a function  $f(u, w)$  where  $u$  is a decision vector we can control, and  $w$  is a random vector that models noise or uncertainty, and whose distribution may or may not be known. Thus, evaluating  $f(u, w)$  requires a realisation of  $w$ . The goal is to find  $u$  that makes  $f(u, w)$  ‘small’ in some sense. To make this more precise, we can aim to optimise the function

$$J(u) = \mathbb{E}[f(u, w)] \quad (16.5.1)$$

where we have averaged over the uncertainty. We can also impose an *admissible set* for  $u$ . This set can be explicitly defined as

$$\mathbb{U} = \{u : c(u) \leq 0\} \quad (16.5.2)$$

Hence this leads to the following stochastic programming problem:

$$u^* = \underset{u \in \mathbb{U}}{\operatorname{argmin}} J(u) \quad (16.5.3)$$

### Chance Constrained Optimisation

Suppose the admissible set also depends on the realisation of  $w$ , so that

$$\mathbb{U} = \{u : c(u, w) \leq 0\} \quad (16.5.4)$$

To address this in stochastic programming, we can either satisfy the constraint in expectation:

$$\begin{aligned} & \underset{u}{\min} && J(u) \\ & \text{s.t.} && \mathbb{E}_w[c(u, w)] = 0 \end{aligned} \quad (16.5.5)$$

or we could instead satisfy the constraint with high probability:

$$\begin{aligned} & \underset{u}{\min} && J(u) \\ & \text{s.t.} && \Pr(c(u, w) \leq 0) \geq 1 - \alpha \end{aligned} \quad (16.5.6)$$

where  $\alpha$  is the risk of constraint violation.

### Newsvendor Model

The newsvendor model is a model for an optimal inventory problem. Suppose on a given day, an item is to be sold at price  $p$ , at a production cost of  $c$ . A total of  $q$  items are to be sold. However, the actual number of items sold will also depend on the demand  $d$  for the item on that given day. Thus, the actual number of items sold will be whichever is the limiting factor of  $q$  or  $d$ . The profit function can be written as

$$f(q, d) = p \min\{q, d\} - cq \quad (16.5.7)$$

To formulate this as a stochastic programming problem, we let  $q \geq 0$  be our decision variable and allow it to take continuous values. We also treat the demand as being a non-negative

continuous random variable  $D$ , with CDF  $H(x)$  and PDF  $h(x)$ . Thus we aim to maximise the expected profit:

$$q^* = \underset{q \geq 0}{\operatorname{argmin}} \mathbb{E}_D [f(q, D)] \quad (16.5.8)$$

which is equivalent to minimising the negative expected profit (i.e. loss). The expected profit may be written as

$$\mathbb{E}_D [f(q, D)] = p \mathbb{E}_D [\min \{q, D\}] - cq \quad (16.5.9)$$

Using the Law of Iterated Expectations,

$$\begin{aligned} \mathbb{E}_D [\min \{q, D\}] &= \mathbb{E}_D [\min \{q, D\} | D \leq q] \Pr(D \leq q) + \mathbb{E}_D [\min \{q, D\} | D > q] \Pr(D > q) \\ &\quad (16.5.10) \end{aligned}$$

$$= \mathbb{E}_D [D | D \leq q] H(q) + q(1 - H(q)) \quad (16.5.11)$$

since  $D \leq q$  implies  $\min \{q, D\} = D$ , while  $D > q$  implies  $\min \{q, D\} = q$ . The conditional expectation  $\mathbb{E}_D [D | D \leq q]$  can be computed as

$$\mathbb{E}_D [D | D \leq q] = \frac{\int_0^q x h(x) dx}{\int_0^q h(x) dx} \quad (16.5.12)$$

$$= \int_0^q x \frac{h(x)}{\Pr(D \leq q)} dx \quad (16.5.13)$$

$$= \frac{1}{H(q)} \int_0^q x h(x) dx \quad (16.5.14)$$

Hence

$$\mathbb{E}_D [\min \{q, D\}] = \frac{H(q)}{H(q)} \int_0^q x h(x) dx + q(1 - H(q)) \quad (16.5.15)$$

$$= \int_0^q x h(x) dx + q(1 - H(q)) \quad (16.5.16)$$

So substituting this back into the expected profit,

$$\mathbb{E}_D [f(q, D)] = p \int_0^q x h(x) dx + pq(1 - H(q)) - cq \quad (16.5.17)$$

Now differentiating with respect to  $q$ :

$$\frac{\partial}{\partial q} \mathbb{E}_D [f(q, D)] = p \frac{\partial}{\partial q} \int_0^q x h(x) dx + p(1 - H(q)) + pq(-h(q)) - c \quad (16.5.18)$$

$$= pqh(q) + p(1 - H(q)) - pqh(q) - c \quad (16.5.19)$$

$$= p(1 - H(q)) - c \quad (16.5.20)$$

We set the derivative equal to zero and solve for the optimal inventory:

$$p(1 - H(q^*)) = c \quad (16.5.21)$$

$$1 - H(q^*) = \frac{c}{p} \quad (16.5.22)$$

$$\frac{p - c}{p} = H(q^*) \quad (16.5.23)$$

$$q^* = H^{-1} \left( \frac{p - c}{p} \right) \quad (16.5.24)$$

This solution is also known as the critical fractile, since it involves the quantile function of  $D$ . We can also show that this solution is a maximum (rather than a minimum) by taking the second derivative, from which we have:

$$\frac{\partial^2}{\partial q^2} \mathbb{E}_D [f(q, D)] = -ph(q) \quad (16.5.25)$$

$$\leq 0 \quad (16.5.26)$$

for all  $q \geq 0$ , since the price  $p$  is implicitly positive. Therefore the objective function is concave. The newsvendor model is a particular example of a stochastic program which admits an analytical solution, however most stochastic programs need to be (approximately) solved using iterative algorithms.

### 16.5.2 Stochastic Dynamic Programming over Finite Horizons [20]

Consider the discrete-time dynamical system which terminates in finite time:

$$x_{k+1} = f_k(x_k, u_k, w_k) \quad (16.5.27)$$

for  $k = 0, 1, \dots, N-1$  with states  $x_k \in \mathbb{X}_k$ , inputs  $u_k \in \mathbb{U}_k(x_k)$  (where the input constraint set  $\mathbb{U}_k(x_k)$  is possibly dependent on  $x_k$ ), and random noise/disturbance  $w_k \in \mathbb{W}_k$ . Assume that the noise  $w_k$  may depend on  $x_k, u_k$ , but is independent of  $w_0, w_1, \dots, w_{k-1}$ . A policy  $\pi$  is given by

$$\pi = \{\mu_0(x_0), \mu_1(x_1), \dots, \mu_{N-1}(x_{N-1})\} \quad (16.5.28)$$

which defines for each  $k = 0, 1, \dots, N-1$  the input  $\mu_k(x_k)$  that should be applied given the current state  $x_k$ . A policy is admissible if it satisfies the input constraints  $u_k \in \mathbb{U}_k(x_k)$  for all  $x_k \in \mathbb{X}_k$ . The problem is to find an admissible policy that minimises the following cost functional:

$$J_\pi(x_0) = \mathbb{E} \left[ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \right] \quad (16.5.29)$$

subject to the dynamics  $x_{k+1} = f_k(x_k, u_k, w_k)$  for each  $k = 0, \dots, N-1$ . The functions  $g_N(\cdot)$  and  $g_j(\cdot, \cdot, \cdot)$  are real-valued and treated as given. The *optimal value function* is defined as

$$J^*(x_0) = \min_{\pi \in \Pi} J_\pi(x_0) \quad (16.5.30)$$

which assigns an optimal value to the optimal policy  $\pi^*$  starting from each initial condition  $x_0$ . Here,  $\Pi$  is the set of all admissible policies. More generally, if the optimal policy does not exist, the optimal value function can be defined as

$$J^*(x_0) = \inf_{\pi \in \Pi} J_\pi(x_0) \quad (16.5.31)$$

We assume that the random variable  $g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k)$  is well-defined with a well-defined expectation. In order for this to hold, additional measurability assumptions may be required for some of the functions. One way to satisfy this condition is to assume that the support of the disturbances  $\mathbb{W}_k$  is a finite set and that the expected values of all terms in the cost functional are finite for all admissible policies. If we proceed with this assumption, then it becomes easier to write expectations in terms of summations. For a given policy  $\pi$ , we write the expected cost incurred at the beginning of time  $N-1$  as

$$J_\pi^{N-1}(x_{N-1}) = \sum_{w_{N-1,i} \in \mathbb{W}_{N-1}} p(w_{N-1,i} | x_{N-1}, \mu_{N-1}(x_{N-1}))$$

$$\times [g_N(f_{N-1}(x_{N-1}, \mu_{N-1}(x_{N-1}), w_{N-1,i})) + g_{N-1}(x_{N-1}, \mu_{N-1}(x_{N-1}), w_{N-1,i})] \quad (16.5.32)$$

where  $p(w_{k,i}|x_k, \mu_k(x_k))$  denotes the probability of disturbance  $w_{k,i}$  given  $x_k, \mu_k(x_k)$ . We see that the expected future cost at time  $k$  is recursively related to the expected future cost at time  $k+1$  by

$$\begin{aligned} J_\pi^k(x_k) &= \sum_{w_{k,i} \in \mathbb{W}_k} p(w_{k,i}|x_k, \mu_k(x_k)) \\ &\quad \times \left[ J_\pi^{k+1}(f_k(x_k, \mu_k(x_k), w_{k,i})) + g_k(x_k, \mu_k(x_k), w_{k,i}) \right] \end{aligned} \quad (16.5.33)$$

for which we can recursively compute for  $k = N-2, \dots, 1, 0$  to obtain the value function

$$J_\pi(x_0) = J_\pi^0(x_0) \quad (16.5.34)$$

An alternative way to compute the value function is to first notice that this problem induces a transition distribution on the states  $p(x_{k+1}|x_k)$ , driven by the randomness of  $w_k$  and because the policy is state-dependent. So the value function can be written using the Law of Iterated Expectations as

$$J_\pi(x_0) = \mathbb{E} \left[ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \right] \quad (16.5.35)$$

$$= \mathbb{E}_{x_1, \dots, x_N} \left[ g_N(x_N) + \sum_{k=0}^{N-1} \mathbb{E}_{w_k} [g_k(x_k, \mu_k(x_k), w_k)|x_k, \mu(x_k)] \middle| x_0 \right] \quad (16.5.36)$$

where the expectation has always been implicitly conditioned on  $x_0$ , but we only explicitly denote it now. For brevity, define

$$\bar{g}_k(x_k, \mu_k(x_k)) = \mathbb{E}_{w_k} [g_k(x_k, \mu_k(x_k), w_k)|x_k, \mu(x_k)] \quad (16.5.37)$$

so that

$$J_\pi(x_0) = \mathbb{E}_{x_1, \dots, x_N} \left[ g_N(x_N) + \sum_{k=0}^{N-1} \bar{g}_k(x_k, \mu_k(x_k)) \middle| x_0 \right] \quad (16.5.38)$$

As  $\bar{g}_0(x_0, \mu_0(x_0))$  is a constant with respect to the expectation (taken over the distribution  $p(x_1, \dots, x_N|x_0)$ ), we take it out:

$$J_\pi(x_0) = \bar{g}_0(x_0, \mu_0(x_0)) + \mathbb{E}_{x_1, \dots, x_N} \left[ g_N(x_N) + \sum_{k=1}^{N-1} \bar{g}_k(x_k, \mu_k(x_k)) \middle| x_0 \right] \quad (16.5.39)$$

Again using the Law of Iterated Expectations and taking out what is known with respect to the expectation:

$$J_\pi(x_0) = \bar{g}_0(x_0, \mu_0(x_0)) + \mathbb{E}_{x_1} \left[ \mathbb{E}_{x_2, \dots, x_N} \left[ g_N(x_N) + \sum_{k=1}^{N-1} \bar{g}_k(x_k, \mu_k(x_k)) \middle| x_1 \right] \middle| x_0 \right] \quad (16.5.40)$$

$$= \bar{g}_0(x_0, \mu_0(x_0)) + \mathbb{E}_{x_1} \left[ \bar{g}_1(x_1, \mu_1(x_1)) + \mathbb{E}_{x_2, \dots, x_N} \left[ g_N(x_N) + \sum_{k=2}^{N-1} \bar{g}_k(x_k, \mu_k(x_k)) \middle| x_1 \right] \middle| x_0 \right] \quad (16.5.41)$$

Repeating this procedure yields

$$\begin{aligned} J_\pi(x_0) &= \bar{g}_0(x_0, \mu_0(x_0)) + \mathbb{E}_{x_1} [\bar{g}_1(x_1, \mu_1(x_1))] + \mathbb{E}_{x_2} [\dots \\ &\quad + \mathbb{E}_{x_{N-1}} [\bar{g}_{N-1}(x_{N-1}, \mu_{N-1}(x_{N-1})) + \mathbb{E}_{x_N} [g_N(x_N)|x_{N-1}]|x_{N-2}] \dots |x_1]|x_0] \end{aligned} \quad (16.5.42)$$

### Bellman's Principle of Optimality

The principle of optimality states that if  $\pi^* = \{\mu_0^*, \mu_1^*, \dots, \mu_{N-1}^*\}$  is an optimal control law, then the truncated policy  $\{\mu_i^*, \mu_{i+1}^*, \dots, \mu_{N-1}^*\}$  is also an optimal policy for the subproblem beginning at time  $i$  with ‘cost-to-go’

$$J_\pi^i(x_i) = \mathbb{E} \left[ g_N(x_N) + \sum_{k=i}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \right] \quad (16.5.43)$$

### Stochastic Dynamic Programming Backwards Induction Algorithm

Consider the following algorithm in which we proceed backward in time, starting from

$$J_N(x_N) = g_N(x_N) \quad (16.5.44)$$

followed by finding

$$J_k(x_k) = \inf_{u_k \in \mathbb{U}_k(x_k)} \mathbb{E}_{w_k} [g_k(x_k, u_k, w_k) + J_{k+1}(f_k(x_k, u_k, w_k))] \quad (16.5.45)$$

for each  $k = N-1, \dots, 1, 0$ . Then the value obtained at the end:

$$J^*(x_0) = J_0(x_0) \quad (16.5.46)$$

is the optimal value function, and moreover the optimal policy consists of the sequence of control laws  $\{\mu_0^*(x_0), \dots, \mu_{N-1}^*(x_{N-1})\}$  which are defined as the minimisers of the above minimisation problem for each  $k = 0, \dots, N-1$ , given the respective  $x_0, \dots, x_{N-1}$ .

*Proof.* From the definition of the optimal policy:

$$J^*(x_0) = \inf_{\mu_0, \dots, \mu_{N-1}} \mathbb{E}_{w_0, \dots, w_{N-1}} \left[ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \right] \quad (16.5.47)$$

Because  $w_k$  is independent with  $w_{k-1}, \dots, w_0$ , we can write  $\mathbb{E}_{w_0, \dots, w_{N-1}} [\cdot] = \mathbb{E}_{w_0} [\mathbb{E}_{w_1, \dots, w_{N-1}} [\cdot | w_0]] = \mathbb{E}_{w_0} [\mathbb{E}_{w_1, \dots, w_{N-1}} [\cdot]]$  and thus

$$\begin{aligned} J^*(x_0) &= \inf_{\mu_0, \dots, \mu_{N-1}} \mathbb{E}_{w_0} [g_0(x_0, \mu_0(x_0), w_0) + \mathbb{E}_{w_1} [g_1(x_1, \mu_1(x_1), w_1) + \dots \\ &\quad + \mathbb{E}_{w_{N-1}} [g_{N-1}(x_{N-1}, \mu_{N-1}(x_{N-1}), w_{N-1}) + g_N(x_N)] \dots]] \end{aligned} \quad (16.5.48)$$

subject to dynamics  $x_{k+1} = f_k(x_k, u_k, w_k)$ . From Bellman's optimality principle, we can split up the infimum so that we find the optimal control law for each sub-problem:

$$\begin{aligned} J^*(x_0) &= \inf_{\mu_0} \left\{ \mathbb{E}_{w_0} \left[ g_0(x_0, \mu_0(x_0), w_0) + \inf_{\mu_1} \{ \mathbb{E}_{w_1} [g_1(x_1, \mu_1(x_1), w_1) + \dots \right. \right. \\ &\quad \left. \left. + \inf_{\mu_{N-1}} \{ \mathbb{E}_{w_{N-1}} [g_{N-1}(x_{N-1}, \mu_{N-1}(x_{N-1}), w_{N-1}) + g_N(x_N)] \} \dots \} \right] \right\} \end{aligned} \quad (16.5.49)$$

Then observe that  $\inf_{\mu_k} \{\cdot\} = \inf_{u_k \in \mathbb{U}_k} \{\cdot\}$ , i.e. the optimal control law on the left-hand side by definition encapsulates all the optimal inputs for each given state. The form of the optimal value function then becomes identical to that obtained by the algorithm above.  $\square$

### 16.5.3 Stochastic Dynamic Programming over Infinite Horizons

In the problem where the dynamic programming does not terminate in finite time, but instead continues on indefinitely, we assume stationarity of the system. This means that the transition function  $f(x_k, u_k, w_k)$  and input constraint  $\mathbb{U}(x_k)$  do not change over time, and the conditional distribution of disturbances  $p(w_k|x_k, u_k)$  also does not change over time. We also assume that the stage cost  $g(x_k, u_k, w_k)$  is bounded and also does not change over time. The cost functional is now a discounted infinite-horizon cost:

$$J_\pi(x_0) = \mathbb{E} \left[ \sum_{k=0}^{\infty} \alpha^k g(x_k, \mu_k(x_k), w_k) \middle| x_0 \right] \quad (16.5.50)$$

where  $0 < \alpha < 1$  is the discount factor. Note that the conditions of stationarity are satisfied if we assume the system to be a finite or countable state Markov decision process. Due to stationarity, the optimal value function is no longer time-dependent, only state-dependent, and can be written as

$$J^*(x) = \min_{\pi \in \Pi} J_\pi(x) \quad (16.5.51)$$

where we drop the time index.

#### Bellman Equation

The optimal value function satisfies what is known as the Bellman equation:

$$J^*(x) = \min_{u \in \mathbb{U}(x)} \{ \mathbb{E}_w [g(x, u, w)|x, u] + \alpha \mathbb{E}_w [J^*(f(x, u, w))|x, u] \} \quad (16.5.52)$$

This form is similar to that found in the backwards induction algorithm above for finite horizons. Here however, when we find ourselves in the next state  $f(x, u^*, w)$  where  $u^* = \mu^*(x)$  from the optimal policy, we are in another situation with an infinite horizon cost, except all costs are discounted by an additional factor of  $\alpha$  from the perspective of the current time. To solve the Bellman equation means to find the function  $J^*(x)$ .

#### Bellman Operator

The Bellman equation is written in terms of what is known as the Bellman operator. For an arbitrary policy  $\pi = \mu(x)$ , the Bellman operator can be denoted as  $\mathcal{T}_\pi$  and is an operator on the value function  $J_\pi(x)$ . It is defined as

$$\mathcal{T}\mathcal{J}_\pi(x) = \mathbb{E}_w [g(x, \mu(x), w)|x, \mu(x)] + \alpha \mathbb{E}_w [J_\pi(f(x, \mu(x), w))|x, \mu(x)] \quad (16.5.53)$$

Thus it can be seen that the Bellman equation is the Bellman operator applied to the optimal value function  $J^*(x)$ .

#### Value Iteration Algorithm [156, 159, 195]

The value iteration algorithm is a method for solving discounted stochastic dynamic programming problems by turning the Bellman equation into an update rule. We begin with an initial value function  $\widehat{J}_0^*(x)$ . Then at each iteration we perform a pass through all the states and update the value function by

$$\widehat{J}_{n+1}^*(x) = \min_{u \in \mathbb{U}(x)} \left\{ \mathbb{E}_w [g(x, u, w)|x, u] + \alpha \mathbb{E}_w [\widehat{J}_n^*(f(x, u, w))|x, u] \right\} \quad (16.5.54)$$

for each  $x \in \mathbb{X}$  (implicitly,  $\mathbb{X}$  should be finite). This can be iterated until some convergence criterion is satisfied, such as the change between  $\widehat{J}_{n+1}^*(\cdot)$  and  $\widehat{J}_n^*(\cdot)$  is sufficiently ‘small’. At

termination, we end up with  $\widehat{J}^*(\cdot) \approx J^*(\cdot)$  which can be made arbitrarily close with the number of iterations. The solved policy is to then apply the control law  $\widehat{\mu}^*(x)$  for each state, where

$$\widehat{\mu}^*(x) = \operatorname{argmin}_{u \in \mathbb{U}(x)} \left\{ \mathbb{E}_w [g(x, u, w)|x, u] + \alpha \mathbb{E}_w [\widehat{J}^*(f(x, u, w))|x, u] \right\} \quad (16.5.55)$$

### Policy Evaluation for Markov Decision Processes [156, 159]

Suppose the system dynamics are governed by a finite-state (with  $M$  states enumerated  $\{1, \dots, M\}$ ) and finite-action Markov decision process. For a stationary policy  $\pi$  which specifies

$$\pi(x|u) = \Pr(u_k = u|x_k = x) \quad (16.5.56)$$

for all time  $k$ , this induces a finite-state Markov chain with row-stochastic transition matrix

$$P_\pi = \begin{bmatrix} \Pr_\pi(x_{k+1} = 1|x_k = 1) & \dots & \Pr_\pi(x_{k+1} = M|x_k = 1) \\ \vdots & \ddots & \vdots \\ \Pr_\pi(x_{k+1} = 1|x_k = M) & \dots & \Pr_\pi(x_{k+1} = M|x_k = M) \end{bmatrix} \quad (16.5.57)$$

$$= \begin{bmatrix} \sum_{u \in \mathbb{U}} p(1|1, u) \pi(u|1) & \dots & \sum_{u \in \mathbb{U}} p(M|1, u) \pi(u|1) \\ \vdots & \ddots & \vdots \\ \sum_{u \in \mathbb{U}} p(M|M, u) \pi(u|M) & \dots & \sum_{u \in \mathbb{U}} p(M|M, u) \pi(u|M) \end{bmatrix} \quad (16.5.58)$$

Denote the expected stage cost

$$\bar{g}(x, u) = \mathbb{E}_w [g(x, u, w)|x, u] \quad (16.5.59)$$

Then given policy  $\pi$ , we can average over all inputs to obtain

$$\bar{g}_\pi(x) = \mathbb{E}_{\pi, w} [g(x, u, w)|x] \quad (16.5.60)$$

$$= \sum_{u \in \mathbb{U}} \bar{g}(x, u) \pi(u|x) \quad (16.5.61)$$

Applying the Bellman operator to each state then gives the system of equations

$$\underbrace{\begin{bmatrix} J_\pi(1) \\ \vdots \\ J_\pi(M) \end{bmatrix}}_{\mathbf{J}_\pi} = \underbrace{\begin{bmatrix} \bar{g}_\pi(1) \\ \vdots \\ \bar{g}_\pi(M) \end{bmatrix}}_{\mathbf{g}_\pi} + \alpha \underbrace{\begin{bmatrix} \sum_{i=1}^M [P_\pi]_{1i} J_\pi(i) \\ \vdots \\ \sum_{i=1}^M [P_\pi]_{Mi} J_\pi(i) \end{bmatrix}}_{P_\pi \mathbf{J}_\pi} \quad (16.5.62)$$

Rearranging, we have

$$(I - \alpha P_\pi) \mathbf{J}_\pi = \mathbf{g}_\pi \quad (16.5.63)$$

$$\mathbf{J}_\pi = (I - \alpha P_\pi)^{-1} \mathbf{g}_\pi \quad (16.5.64)$$

Thus for a given policy  $\pi$ , we can use this formula to evaluate the value function at each state. Finding an optimal policy for an initial starting state then amounts to finding the  $\pi$  which minimises the value function of the corresponding starting state, e.g. suppose without loss of generality that state 1 is the starting state, then

$$J^*(1) = \min_{\pi} \left\{ \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \mathbf{J}_\pi \right\} \quad (16.5.65)$$

### Iterative Policy Evaluation [195]

Another method of evaluating the value function  $J_\pi(x)$  for a provided stationary policy  $\pi = \mu(x)$  is by iterative evaluation. Beginning with some initial estimate  $\hat{J}_{\pi,0}(\cdot)$ , at each iteration we perform a pass through the states and update the value function by

$$\hat{J}_{\pi,n+1}(x) = \mathbb{E}_{u \sim \pi} [\mathbb{E}_w [g(x, u, w)|x, u] + \alpha \mathbb{E}_w [J_{\pi,n}(f(x, u, w))|x, u]|x] \quad (16.5.66)$$

for each  $x \in \mathbb{X}$ . This can be iterated until some desired tolerance is reached.

### Policy Improvement Step [195]

Suppose we already have a method of evaluating the value function  $J_\pi(x)$  for a given stationary policy  $\pi = \mu(x)$  (such as via [iterative policy evaluation](#) or by [solving a system of equations](#)). Now consider two policies  $\pi$  and  $\pi'$  that are to be compared. Let  $\tilde{\pi}$  be a ‘perturbed’ non-stationary policy that applies  $\pi'$  at the first time-instant, and  $\pi$  thereafter. If it holds that  $J_{\tilde{\pi}}(x) \leq J_\pi(x)$  for all  $x \in \mathbb{X}$ , then it naturally follows that  $J_{\pi'}(x) \leq J_\pi(x)$  for all  $x \in \mathbb{X}$ . The reasoning for this is that we can form a sequence of perturbed policies  $\{\tilde{\pi}_n\}$  which applies  $\pi'$  for the first two, three, etc. time instants, each successive policy improving on the last. Then as  $n \rightarrow \infty$ ,  $\tilde{\pi}_n$  is identical to  $\pi'$ .

Based on this principle, then for an existing stationary policy  $\pi = \mu(x)$  we can define an improved stationary policy  $\pi' = \mu'(x)$  that is found based on a greedy approach (i.e. only one-step lookahead) for each state. This greedy policy is defined by:

$$\mu'(x) = \operatorname{argmin}_{u \in \mathbb{U}(x)} \{\mathbb{E}_w [g(x, u, w)|x, u] + \alpha \mathbb{E}_w [J_\pi(f(x, u, w))|x, u]\} \quad (16.5.67)$$

for each  $x \in \mathbb{X}$ . This greedy policy will satisfy  $J_{\pi'}(x) \leq J_\pi(x)$  for all  $x \in \mathbb{X}$ .

### Policy Iteration Algorithm [195]

The policy iteration algorithm is another method for solving stochastic dynamic programming problems by involving the policy improvement step. Beginning with an initial policy estimate  $\pi_0$ , we perform

$$\pi_{n+1} = \pi'_n \quad (16.5.68)$$

where  $\pi'_n$  denotes the greedy policy improvement step performed on policy  $\pi_n$ . This is then iterated until the policy stops changing (i.e. it stops improving), which indicates we have found the optimal policy. This optimality applies in the case that the value function for a given policy can be evaluated perfectly, otherwise the found policy will be approximately optimal.

### Linear Programming for Stochastic Dynamic Programming [168]

Using the [Markov decision process representation](#) of the dynamics with transition probability  $p(x_{k+1}|x, u)$ , suppose there is a function  $\underline{J}(x)$  such that

$$\underline{J}(x) \leq \min_{u \in \mathbb{U}} \left\{ \bar{g}(x, u) + \alpha \sum_{i=1}^M p(i|x, u) \underline{J}(i) \right\} \quad (16.5.69)$$

for all states  $x \in \mathbb{X}$ . Recursively applying this property gives

$$\underline{J}(x) \leq \min_{u \in \mathbb{U}} \left\{ \bar{g}(x, u) + \alpha \sum_{i=1}^M p(i|x, u) \min_{u' \in \mathbb{U}} \left\{ \bar{g}(i, u) + \alpha \sum_{\ell=1}^M p(\ell|i, u') \underline{J}(\ell) \right\} \right\} \quad (16.5.70)$$

$$\leq \min_{u \in \mathbb{U}} \left\{ \bar{g}(x, u) + \alpha \sum_{i=1}^M p(i|x, u) \min_{u' \in \mathbb{U}} \left\{ \bar{g}(i, u') + \alpha \sum_{\ell=1}^M p(\ell|i, u') \min_{u'' \in \mathbb{U}} \{\dots\} \right\} \right\} \quad (16.5.71)$$

$$= J^*(x) \quad (16.5.72)$$

which is the optimal value function. In fact,  $J^*(x)$  satisfies the inequality required by  $\underline{J}(x)$  except with equality. Thus,  $J^*(x)$  can be characterised as the largest function  $\underline{J}(x)$  such that

$$\underline{J}(x) \leq \bar{g}(x, u) + \alpha \sum_{i=1}^M p(i|x, u) \underline{J}(i) \quad (16.5.73)$$

holds for all pairs  $(x, u)$ . For finite-state and finite-input Markov decision processes, this gives rise to the linear program

$$\begin{aligned} & \max_{\underline{J}(1), \dots, \underline{J}(M)} \sum_{i=1}^M \underline{J}(i) \\ \text{s.t. } & \underline{J}(x) \leq \bar{g}(x, u) + \alpha \sum_{i=1}^M p(i|x, u) \underline{J}(i), \quad x \in \mathbb{X}, u \in \mathbb{U} \end{aligned} \quad (16.5.74)$$

Then the maximiser to this problem is the optimal value function.

## 16.6 Stochastic Optimal Control

### 16.6.1 Hamilton-Jacobi-Bellman Equation [191]

### 16.6.2 Linear Quadratic Gaussian Control

#### Separation Principle

### 16.6.3 Stochastic Model Predictive Control

## 16.7 Stochastic Approximation [7, 189]

In stochastic approximation, we are interested in finding the roots of a generally vector-valued nonlinear function  $\bar{g}(\theta)$  based on noisy observations of  $\bar{g}(\theta)$ . More concretely, suppose

$$\bar{g}(\theta) = \mathbb{E}_W[g(\theta, W)] \quad (16.7.1)$$

for some function  $g(\theta, w)$ , where  $w$  takes the role of noise. Then we wish to find one or more  $\theta^*$  such that

$$\mathbb{E}_W[g(\theta^*, W)] = \mathbf{0} \quad (16.7.2)$$

based on noisy observations  $g(\theta, W)$ . We may not know the distribution of  $W$  or the structure of  $g$ , however we usually assume that we can at least (implicitly) sample  $W$  in order to evaluate  $g$ . Mean estimation of a function  $f(\cdot)$  of a random variable  $X$  can be formulated as a stochastic approximation problem, by letting  $g(\theta, x) = f(x) - \theta$ . Then finding  $\mathbb{E}_X[g(\theta, X)] = \mathbf{0}$  amounts to finding the mean vector  $\theta = \mathbb{E}[f(X)]$ .

### 16.7.1 Robbins-Monro Algorithm

If  $\theta$  and  $g(\theta)$  are of the same dimension, then an iterative algorithm for approximately solving the stochastic approximation problem is the update of the form

$$\theta_{t+1} = \theta_t + \alpha_t g(\theta_t, W_{t+1}) \quad (16.7.3)$$

where  $\alpha_t$  is a sequence of positive step sizes determined in advance, and  $W_t$  are sequential samples of noise  $W$ . To give intuition why this algorithm might work, consider the ordinary differential equation

$$\dot{\theta} = \bar{g}(\theta) \quad (16.7.4)$$

Then  $\theta^*$  is an equilibrium point, since  $\bar{g}(\theta^*) = \mathbf{0}$  by definition. In order for this to work, assume that this equilibrium point is stable. Discretising this differential equation with time step  $\alpha$ , we have

$$\frac{\theta_{t+1} - \theta_t}{\alpha} = \bar{g}(\theta_t) \quad (16.7.5)$$

which rearranges to

$$\theta_{t+1} = \theta_t + \alpha \bar{g}(\theta_t) \quad (16.7.6)$$

Thus if this sequence converges, it will converge to an equilibrium point  $\theta^*$ . We then introduce the step size sequence  $\alpha_t$  and replace  $\bar{g}(\theta_t)$  with a noisy but unbiased estimate  $g(\theta_t, W_{t+1})$ , and reason that we will be able to approximately find  $\theta^*$ .

### Robbins-Monro Conditions

The step size  $\alpha_t$  should be appropriately chosen to ensure convergence. The Robbins-Monro conditions are that  $\alpha_t$  is a sequence decreasing to zero such that

$$\sum_{t=1}^{\infty} \alpha_t = \infty \quad (16.7.7)$$

and

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty \quad (16.7.8)$$

To see why we would need the first condition, we write out the value of the iterate at time  $n$  by

$$\theta_n = \theta_0 + \sum_{t=1}^n \alpha_t g(\theta_{t-1}, W_t) \quad (16.7.9)$$

Suppose that  $g(\theta_{t-1}, W_t)$  is a uniformly bounded sequence, i.e.  $\|g(\theta_{t-1}, W_t)\| \leq c$  for some constant  $c$ . Then

$$\left\| \sum_{t=1}^n \alpha_t g(\theta_{t-1}, W_t) \right\| \leq \sum_{t=1}^n \alpha_t \|g(\theta_{t-1}, W_t)\| \quad (16.7.10)$$

$$\leq c \sum_{t=1}^n \alpha_t \quad (16.7.11)$$

So unless  $\sum_{t=1}^{\infty} \alpha_t = \infty$ , the sequence  $\|\theta_n - \theta_0\|$  would also be bounded, and we would not be able to reach the desired solution  $\theta_*$  for  $\theta_0$  sufficiently far away from  $\theta_*$ . On the other hand, for the sake of simplicity assume  $\theta$  and  $\bar{g}(\theta)$  are one-dimensional and consider the variance of  $\theta_n$  from fixed initial point  $\theta_0$ :

$$\text{Var}(\theta_n) = \text{Var} \left( \sum_{t=1}^n \alpha_t g(\theta_{t-1}, W_t) \right) \quad (16.7.12)$$

$$= \sum_{t=1}^n \alpha_t^2 \text{Var}(g(\theta_{t-1}, W_t)) + \sum_{s \neq t} \alpha_t \alpha_s \text{Cov}(g(\theta_{t-1}, W_t), g(\theta_{s-1}, W_s)) \quad (16.7.13)$$

$$\geq v \sum_{t=1}^n \alpha_t^2 + \sum_{s \neq t} \alpha_t \alpha_s \text{Cov}(g(\theta_{t-1}, W_t), g(\theta_{s-1}, W_s)) \quad (16.7.14)$$

for some constant  $v$  such that  $v \leq \text{Var}(g(\theta_{t-1}, W_t))$  for all  $t$ . Hence we see that if  $\sum_{t=1}^{\infty} \alpha_t^2 = \infty$ , then  $\theta_n$  will have infinite variance as  $n \rightarrow \infty$ , which will spoil convergence. An example of a step size sequence that satisfies the Robbins-Monroe conditions is  $\alpha_t = \frac{1}{t}$ .

### Monte-Carlo Estimation as Robbins-Monro Algorithm

Recursive Monte-Carlo estimation of an expectation can be viewed as a Robbins-Monro algorithm. Recall that estimation  $\mathbb{E}[f(X)]$  can be considered a stochastic approximation problem with  $g(\theta, x) = f(x) - \theta$ . The Monte-Carlo estimate from  $n$  samples can be written as

$$\theta_n = \frac{1}{n} \sum_{t=1}^n f(X_i) \quad (16.7.15)$$

We can show that

$$n\theta_n = \sum_{t=1}^n f(X_i) \quad (16.7.16)$$

$$= \sum_{t=1}^{n-1} f(X_i) + f(X_n) \quad (16.7.17)$$

$$= (n-1)\theta_{n-1} + f(X_n) \quad (16.7.18)$$

$$= n\theta_{n-1} + f(X_n) - \theta_{n-1} \quad (16.7.19)$$

$$= n\theta_{n-1} + g(\theta_{n-1}, X_n) \quad (16.7.20)$$

Hence we have the recursive update

$$\theta_{n+1} = \theta_n + \frac{1}{n+1} g(\theta_n, X_{n+1}) \quad (16.7.21)$$

which is a Robbins-Monro update with step size  $\alpha_n = \frac{1}{n+1}$  satisfying the Robbins-Monro conditions.

### 16.7.2 Stochastic Gradient Descent

Recognising the connection between the stochastic approximation and stochastic programming problem formulations, we can formulate a class of stochastic programming problems as stochastic approximation problems. Assuming differentiability, the minimiser

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathbb{E}[f(\theta, W)] \quad (16.7.22)$$

also satisfies

$$-\nabla_{\theta} \mathbb{E}[f(\theta, W)] = \mathbb{E}[-\nabla_{\theta} f(\theta, W)] \quad (16.7.23)$$

$$= 0 \quad (16.7.24)$$

where we also assume we can pass the differentiation through the expectation. Thus, this class of stochastic programming problem becomes a stochastic approximation problem with  $g(\theta, w) = \nabla_{\theta} f(\theta, w)$ . The reason we take the negative sign is to associate it with stable dynamics (in a maximisation problem, we would take a positive sign). Ideally, a gradient descent update for  $\theta$  would be

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_{\theta} \mathbb{E}[f(\theta, W)] \quad (16.7.25)$$

Since we treat this expectation as being unobtainable, we can replace it with an unbiased estimate, namely a single realisation of  $f(\theta, W)$ . This turns out to be a Robbins-Monro algorithm. Applying this update for  $\theta$  gives

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta f(\theta_t, W_{t+1}) \quad (16.7.26)$$

from independent samples  $W_1, W_2, \dots$  drawn online. This is known as stochastic gradient descent, and note that  $f(\theta_t, W_t)$  is indeed an unbiased estimator of  $\mathbb{E}[f(\theta_t, W_t)]$ . If we wanted to find a local maximum instead, we would then use a positive sign (this would be known as *stochastic gradient ascent*).

### Kiefer-Wolfowitz Algorithm [22]

In a Kiefer-Wolfowitz algorithm, the estimate of  $\nabla_\theta \mathbb{E}[f(\theta, W)]$  is instead replaced by a noisy finite difference estimate, which requires two function evaluations per dimension. Assuming  $\theta$  is one-dimensional, this estimate is given by

$$\widehat{\nabla}_\theta \mathbb{E}[f(\theta, W)] = \frac{f(\theta - \delta, W) - f(\theta + \delta, W')}{2\delta} \quad (16.7.27)$$

for some small  $\delta > 0$ , and where  $W'$  is an independent copy of  $W$ . Thus, a Kiefer-Wolfowitz update can be written as

$$\theta_{t+1} = \theta_t - \alpha_t \frac{f(\theta_t - \delta_t, W_{t+1}) - f(\theta_t + \delta_t, W'_{t+1})}{2\delta_t} \quad (16.7.28)$$

where we allow for  $\delta_t$  to also be a sequence.

### 16.7.3 Stochastic Average Approximation

An alternative approach for approximately solving stochastic programming problems is to first generate a sample  $W_1, \dots, W_n$ , and then take the sample average approximation:

$$\widehat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n f(\theta, W_i) \quad (16.7.29)$$

Then we minimise this approximation to obtain our solution:

$$\widehat{\theta} = \operatorname{argmin}_\theta \widehat{f}(\theta) \quad (16.7.30)$$

This is known as stochastic average approximation (or alternatively the *stochastic counterpart method*). The paradigm of empirical risk minimisation can be considered a particular case of performing stochastic average approximation. Let  $\theta$  be the parameter of a hypothesis class. For a labelled observation  $(X, Y)$ , let  $L(X, Y; \theta)$  be the loss function parametrised in  $\theta$ . Then we wish to find the  $\theta$  which minimises the expected loss (i.e. risk  $R(\theta)$ ):

$$\theta^* = \operatorname{argmin}_\theta R(\theta) \quad (16.7.31)$$

$$= \operatorname{argmin}_\theta \mathbb{E}_{X,Y} [L(X, Y; \theta)] \quad (16.7.32)$$

By taking the approach of stochastic average approximation, we draw a sample of i.i.d. observations  $(X_1, Y_1), \dots, (X_n, Y_n)$  and get the empirical loss

$$\widehat{L}(\theta) = \frac{1}{n} \sum_{i=1}^n L(X_i, Y_i; \theta) \quad (16.7.33)$$

from which we compute the empirical risk minimiser

$$\widehat{\theta} = \operatorname{argmin}_\theta \left\{ \frac{1}{n} \sum_{i=1}^n L(X_i, Y_i; \theta) \right\} \quad (16.7.34)$$

### 16.7.4 Asymptotic Normality of Stochastic Approximation

## 16.8 Multi-Armed Bandits

### 16.8.1 Stochastic Bandits

A stochastic bandit is a collection of distributions

$$\mathcal{P} = \{\mathbb{P}_a : a \in \mathcal{A}\} \quad (16.8.1)$$

where  $\mathcal{A}$  is a set of actions (or ‘arms’). For simplicity,  $\mathcal{A}$  is usually treated as being a finite set. A stochastic bandit is played over  $n$  rounds, with time index  $t \in \{1, \dots, n\}$ . In round  $t$ , let  $A_t$  be the action chosen and let  $X_t$  be the observed reward, drawn from distribution  $\mathbb{P}_{A_t}$ . The goal of a learner is to maximise the total cumulative reward, denoted by

$$S_n = \sum_{i=1}^n X_t \quad (16.8.2)$$

The learner may have some knowledge about  $\mathcal{P}$ , and uses a policy denoted by

$$\Pr(A_t = a | A_{t-1}, X_{t-1}, \dots, A_1, X_1) = \pi_t(a | A_{t-1}, X_{t-1}, \dots, A_1, X_1) \quad (16.8.3)$$

which is a distribution over actions, given all observed knowledge up to and including round  $t$ . A policy will typically tradeoff between ‘exploration’ (finding the arms which give better rewards) against ‘exploitation’ (choosing arms which are already known to give good rewards).

### 16.8.2 Regret

Denote the average reward of arm  $a$  by

$$\mu_a = \int_{-\infty}^{\infty} x d\mathbb{P}_a(x) \quad (16.8.4)$$

The best arm  $a^*$  is then defined as the arm with the highest average reward:

$$\mu^* = \max_{a \in \mathcal{A}} \mu_a \quad (16.8.5)$$

so that

$$a^* = \operatorname{argmax}_{a \in \mathcal{A}} \mu_a \quad (16.8.6)$$

The (pseudo) regret at round  $n$  of policy  $\pi$  (with respect to some underlying bandit  $\mathcal{P}$ ) is

$$R_n(\pi) = n\mu^* - \mathbb{E}[S_n] \quad (16.8.7)$$

$$= n\mu^* - \mathbb{E}\left[\sum_{i=1}^n X_t\right] \quad (16.8.8)$$

Thus we can see that minimising the regret leads to maximising the expected total cumulative reward. Also if the average rewards of the arms were known in advance, then the optimal policy is such that  $\Pr(A_t = a^*) = 1$  for all  $t$ . Since the best arm is not known in advance however, it must be learned (in a sense, like an online version of multiple comparisons).

### Regret Decomposition Lemma

Let the *immediate regret* of picking arm  $a$  be

$$\Delta_a = \mu^* - \mu_a \quad (16.8.9)$$

which gives a notion of an optimality gap for arm  $a$ . Define

$$T_a(t) = \sum_{s=1}^t \mathbb{I}_{\{A_s=a\}} \quad (16.8.10)$$

which simply counts the number of times that arm  $a$  was chosen up to and including round  $t$ . Then for any policy  $\pi$ , we have

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \quad (16.8.11)$$

*Proof.* First write  $R_n$  in terms of the immediate regret as

$$R_n = n\mu^* - \mathbb{E} \left[ \sum_{i=1}^n X_t \right] \quad (16.8.12)$$

$$= \sum_{i=1}^n \mu^* - \sum_{i=1}^n \mathbb{E}[\mathbb{E}[X_t | A_t]] \quad (16.8.13)$$

$$= \sum_{i=1}^n \mu^* - \sum_{i=1}^n \mathbb{E}[\mu_{A_t}] \quad (16.8.14)$$

$$= \sum_{i=1}^n \mathbb{E}[\Delta_{A_t}] \quad (16.8.15)$$

Now note that  $\sum_{a \in \mathcal{A}} \mathbb{I}_{\{A_t=a\}} = 1$ , so

$$R_n = \sum_{i=1}^n \mathbb{E} \left[ \Delta_{A_t} \sum_{a \in \mathcal{A}} \mathbb{I}_{\{A_t=a\}} \right] \quad (16.8.16)$$

$$= \mathbb{E} \left[ \sum_{i=1}^n \sum_{a \in \mathcal{A}} \Delta_{A_t} \mathbb{I}_{\{A_t=a\}} \right] \quad (16.8.17)$$

$$= \mathbb{E} \left[ \sum_{i=1}^n \sum_{a \in \mathcal{A}} \Delta_a \mathbb{I}_{\{A_t=a\}} \right] \quad (16.8.18)$$

where we could let  $\Delta_{A_t} = \Delta_a$  since  $\mathbb{I}_{\{A_t=a\}} = 1$  only when  $A_t = a$ . Finally, we have

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E} \left[ \sum_{i=1}^n \mathbb{I}_{\{A_t=a\}} \right] \quad (16.8.19)$$

$$= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \quad (16.8.20)$$

□

This lemma intuitively says that in order to keep regret small, the more suboptimal an arm is, the fewer times it should be pulled on average.

### Stochastic Regret [37]

Let  $X_{t|a}$  be the reward at round  $t$ , had arm  $a$  been pulled. We can then define a stochastic version of regret as

$$\tilde{R}_n = \max_{a \in \mathcal{A}} \left\{ \sum_{t=1}^n X_{t|a} \right\} - \sum_{t=1}^n X_{t|A_t} \quad (16.8.21)$$

This gives the difference in total cumulative reward, as compared to knowing the best single arm for each round in advance, and pulling those. Note that the expected stochastic regret  $\mathbb{E}[\tilde{R}_n]$  will generally not be equal to the pseudo-regret  $R_n$ .

### Simple Regret

The simple regret can be defined as

$$\check{R}_n = \mathbb{E}[\Delta_{A_{n+1}}] \quad (16.8.22)$$

which can be interpreted as the expected suboptimality of the arm that would have been picked after  $n$  rounds. This version of regret is useful for analysing algorithms where the goal is only to identify the best arm.

### 16.8.3 Bandit Algorithms [121]

#### Explore-First Algorithm

Consider a bandit with  $k$  arms, given by  $\mathcal{A} = \{1, \dots, k\}$ . In the explore-first algorithm (also known as the ‘explore-then-commit’ algorithm), we first pull each of the arms  $m$  times. Then after  $mk$  rounds, the mean reward of each arm is estimated via the sample mean, and denoted  $\hat{\mu}_i$ . For each round onwards, we then commit to pulling the arm with the highest estimated mean (ties can be broken arbitrarily). Succinctly, the policy can be characterised by

$$A_t = \begin{cases} (t \bmod k) + 1, & t \leq mk \\ \operatorname{argmax}_i \hat{\mu}_i, & t > mk \end{cases} \quad (16.8.23)$$

Assume that rewards are sub-Gaussian distributed with factor  $\sigma = 1$  (note that this factor is more for simplicity, because it just scales the rewards). If  $n \geq mk$ , then the regret of the explore-first algorithm is upper bounded by

$$R_n \leq m \sum_{i=1}^k \Delta_i + (n - mk) \sum_{i=1}^k \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right) \quad (16.8.24)$$

*Proof.* Since each arm is pulled a guaranteed  $m$  times, then the expected number of times pulled is

$$\mathbb{E}[T_i(n)] = m + \sum_{s=mk+1}^n \mathbb{E}[\mathbb{I}_{\{\operatorname{argmax}_j \hat{\mu}_j = i\}}] \quad (16.8.25)$$

$$= m + (n - mk) \Pr\left(\operatorname{argmax}_j \hat{\mu}_j = i\right) \quad (16.8.26)$$

$$\leq m + (n - mk) \Pr\left(\hat{\mu}_i \geq \max_{j \neq i} \hat{\mu}_j\right) \quad (16.8.27)$$

where the inequality is from allowing for ties. Denote the best arm by  $a^*$ . The probability above can be upper bounded by

$$\Pr \left( \hat{\mu}_i \geq \max_{j \neq i} \hat{\mu}_j \right) \leq \Pr \left( \hat{\mu}_i \geq \hat{\mu}_{a^*} \right) \quad (16.8.28)$$

$$= \Pr \left( \hat{\mu}_i - \hat{\mu}_{a^*} + \mu_{a^*} - \mu_i \geq \Delta_i \right) \quad (16.8.29)$$

$$= \Pr \left( \hat{\mu}_i - \mu_i - (\hat{\mu}_{a^*} - \mu_{a^*}) \geq \Delta_i \right) \quad (16.8.30)$$

Recognise that  $\hat{\mu}_i - \mu_i$  and  $\hat{\mu}_{a^*} - \mu_{a^*}$  are each both zero-mean and  $\sqrt{1/m}$ -sub-Gaussian because they are sample averages of rewards. Hence  $\hat{\mu}_i - \mu_i - (\hat{\mu}_{a^*} - \mu_{a^*})$  will be  $\sqrt{2/m}$ -sub-Gaussian, and by the single-tailed characterisation of sub-Gaussian variables, we can write

$$\Pr \left( \hat{\mu}_i - \mu_i - (\hat{\mu}_{a^*} - \mu_{a^*}) \geq \Delta_i \right) \leq \exp \left( -\frac{\Delta_i^2}{2(\sqrt{2/m})^2} \right) \quad (16.8.31)$$

$$= \exp \left( -\frac{m\Delta_i^2}{4} \right) \quad (16.8.32)$$

Therefore we have

$$\mathbb{E}[T_i(n)] \leq m + (n - mk) \exp \left( -\frac{m\Delta_i^2}{4} \right) \quad (16.8.33)$$

and putting this bound together with the regret decomposition lemma, we get

$$R_n \leq m \sum_{i=1}^k \Delta_i + (n - mk) \sum_{i=1}^k \Delta_i \exp \left( -\frac{m\Delta_i^2}{4} \right) \quad (16.8.34)$$

□

### $\varepsilon$ -Greedy Algorithm

The  $\varepsilon$ -greedy bandit algorithm maintains an estimate of the mean reward  $\hat{\mu}_i$  for each arm. For exploitation, the arm with the highest  $\hat{\mu}_i$  (with ties settled arbitrarily) is pulled with probability  $1 - \varepsilon$ . To encourage exploration, a random arm is pulled with probability  $\varepsilon$ , where  $\varepsilon$  is an algorithm parameter. For this algorithm, it can be shown that

$$\lim_{n \rightarrow \infty} \frac{R_n}{n} = \frac{\varepsilon}{k} \sum_{i=1}^k \Delta_i \quad (16.8.35)$$

This result is intuitive, because in the limit, we will have correctly estimated the best arm and will be pulling that with probability  $1 - \varepsilon$ . In the  $\varepsilon$  probability that we pick a random (potentially sub-optimal) arm, each arm contributes  $\Delta_i$  to regret that round, weighted by the  $\varepsilon/k$  probability of being picked. Therefore, the algorithm does not achieve sublinear regret, which would otherwise be the case if  $\lim_{n \rightarrow \infty} \frac{R_n}{n} \rightarrow 0$ .

### Upper Confidence Bound Algorithm

The upper confidence bound (UCB) algorithm attempts to store a estimate of a mean reward as well as an upper confidence of the estimate for each arm as they are pulled. Using the Hoeffding inequality for independent 1-sub-Gaussian rewards, an upper deviation inequality for the mean estimate  $\hat{\mu}_{a,n}$  of arm  $a$  after  $n$  pulls is

$$\Pr \left( \hat{\mu}_{a,n} - \mu_a \geq \varepsilon \right) \leq \exp \left( -\frac{n\varepsilon^2}{2} \right) \quad (16.8.36)$$

for any  $\varepsilon \geq 0$ . Choosing  $\varepsilon = \sqrt{-2 \log(\delta)/n}$  for some  $\delta \in (0, 1)$ , we can write

$$\Pr \left( \mu_a \geq \hat{\mu}_{a,n} + \sqrt{\frac{2 \log(1/\delta)}{n}} \right) \leq \delta \quad (16.8.37)$$

So an upper confidence bound for  $\mu_a$  with confidence of at least  $1 - \delta$  is

$$\Pr \left( \mu_a < \hat{\mu}_{a,n} + \sqrt{\frac{2 \log(1/\delta)}{n}} \right) \geq 1 - \delta \quad (16.8.38)$$

In the UCB algorithm, we denote the upper confidence bound (at level  $1 - \delta$ ) of the  $i^{\text{th}}$  arm after round  $t$  (after which arm  $i$  has been pulled  $T_i(t)$  times) by:

$$\text{UCB}_i(t) = \begin{cases} \infty, & T_i(t) = 0 \\ \hat{\mu}_i(t) + \sqrt{\frac{2 \log(1/\delta)}{T_i(t)}}, & T_i(t) > 0 \end{cases} \quad (16.8.39)$$

where  $\hat{\mu}_i(t)$  is the estimate of the mean reward for arm  $i$  at round  $t$ . Hence we start off completely agnostic about rewards, and then refine our estimates as we explore more. The UCB algorithm follows the arm selection rule for round  $t$  (using information up to round  $t - 1$ ):

$$A_t = \underset{i}{\operatorname{argmax}} \text{UCB}_i(t - 1) \quad (16.8.40)$$

where ties can be settled arbitrarily. After observing reward  $X_t$ , the mean estimates and upper confidence bounds are updated. This algorithm can be thought of as satisfying the exploitation-exploration tradeoff in the following way. If an arm's upper confidence bound is high relative to others, it might mean either:

1. It has a high mean reward, so we should exploit it.
2. We are uncertain about its reward (having not pulled it enough times), so we should explore it.

Under a suitable choice of  $\delta$ , we can show the following regret bound.

**Theorem 16.1.** *Consider the UCB algorithm on a stochastic  $k$ -armed bandit with 1-sub-Gaussian rewards. For any  $n$ , if  $\delta = 1/n^2$ , then the regret is bounded by:*

$$R_n \leq 3 \sum_{i=1}^k \Delta_i + \sum_{i: \Delta_i > 0} \frac{16 \log n}{\Delta_i} \quad (16.8.41)$$

where the  $\Delta_i$  are the suboptimality gaps (immediate regrets) for each arm.

*Proof.* Without loss of generality (because the numbering of arms is arbitrary), suppose the first arm is optimal, so  $\mu^* = \mu_1$ . To simplify notation, let  $\hat{\mu}_{i,s}$  denote the mean estimate of arm  $i$  after it has been pulled  $s$  times (for explicit dependence on  $t$ , we could write  $\hat{\mu}_i(t) = \hat{\mu}_{i,T_i(t)}$ ). For each arm, and given some constant  $\eta_i$  to be determined later, define the event

$$G_i = \left\{ \hat{\mu}_{i,\eta_i} + \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} < \mu_1 \right\} \cap \left\{ \mu_1 < \min_{t \leq n} \text{UCB}_1(t) \right\} \quad (16.8.42)$$

The first event expresses that after  $\eta_i$  observations for arm  $i$ , our upper confidence bound for that arm is below the optimal reward  $\mu_1$ . The second event expresses that the upper confidence bound never underestimates the best reward throughout the run of the algorithm. Hence,  $G_i$

is intuitively a ‘good’ event that we want to happen to for a low regret. Thus, we focus on upper bounding the probability of its complement,  $\Pr(\overline{G}_i)$ . First, consider the complement of the first event, which is a subset of:

$$\left\{ \min_{t \leq n} \text{UCB}_1(t) \leq \mu_1 \right\} \subset \left\{ \min_{s \leq n} \left\{ \widehat{\mu}_{1,s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \right\} \leq \mu_1 \right\} \quad (16.8.43)$$

The left is the event that the upper confidence bound is underestimated at least once during the run, and the right is like the event where the upper confidence bound is underestimated at least once, in the case that the algorithm picked the best arm each time (so it has more ‘chances’ to occur than the condition on the left). Equivalently,

$$\bigcup_{t=1}^{T_i(n)} \left\{ \widehat{\mu}_{1,t} + \sqrt{\frac{2 \log(1/\delta)}{t}} \leq \mu_1 \right\} \subset \bigcup_{s=1}^n \left\{ \widehat{\mu}_{1,s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \leq \mu_1 \right\} \quad (16.8.44)$$

So taking probabilities and using the union bound,

$$\Pr \left( \left\{ \min_{t \leq n} \text{UCB}_1(t) \leq \mu_1 \right\} \right) \leq \Pr \left( \bigcup_{s=1}^n \left\{ \widehat{\mu}_{1,s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \leq \mu_1 \right\} \right) \quad (16.8.45)$$

$$\leq \sum_{s=1}^n \Pr \left( \widehat{\mu}_{1,s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \leq \mu_1 \right) \quad (16.8.46)$$

$$\leq n\delta \quad (16.8.47)$$

where the last inequality is from applying the sub-Gaussian upper deviation property. Now for the complement of the other event, since  $\Delta_i \geq 0$ , then for a large enough  $\eta_i$ , there exists a constant  $c \in (0, 1)$ , to be determined later, such that

$$\Delta_i - \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} \geq c\Delta_i \quad (16.8.48)$$

Combine this with the definition  $\Delta_i = \mu_1 - \mu_i$  and apply it to the probability of the complement of the event  $\left\{ \widehat{\mu}_{i,\eta_i} + \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} < \mu_1 \right\}$  so that

$$\Pr \left( \mu_1 \leq \widehat{\mu}_{i,\eta_i} + \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} \right) = \Pr \left( \widehat{\mu}_{i,\eta_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} \right) \quad (16.8.49)$$

$$\leq \Pr(\widehat{\mu}_{i,\eta_i} - \mu_i \geq c\Delta_i) \quad (16.8.50)$$

$$\leq \exp \left( \frac{-\eta_i c^2 \Delta_i^2}{2} \right) \quad (16.8.51)$$

using the upper deviation property again. Thus, for the event  $\overline{G}_i$  given by

$$\overline{G}_i = \left\{ \min_{t \leq n} \text{UCB}_1(t) \leq \mu_1 \right\} \cup \left\{ \mu_1 \leq \widehat{\mu}_{i,\eta_i} + \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} \right\} \quad (16.8.52)$$

we have its probability upper bounded by

$$\Pr(\overline{G}_i) \leq \Pr \left( \min_{t \leq n} \text{UCB}_1(t) \leq \mu_1 \right) + \Pr \left( \mu_1 \leq \widehat{\mu}_{i,\eta_i} + \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} \right) \quad (16.8.53)$$

$$\leq n\delta + \exp\left(-\frac{\eta_i c^2 \Delta_i^2}{2}\right) \quad (16.8.54)$$

Next, we claim that  $G_i$  implies  $T_i(n) \leq \eta_i$ . Suppose  $G_i$  holds and  $T_i(n) > \eta_i$ . Then there must exist some  $t \leq n$  during the run of the algorithm such that the number of pulls was  $\eta_i$  after round  $t - 1$  (meaning  $T_i(t - 1) = \eta_i$ ), and we pulled  $i$  at round  $t$  (meaning  $A_t = i$ ). By the upper confidence bound calculation after round  $t - 1$ :

$$\text{UCB}_i(t - 1) = \hat{\mu}_i(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_i(t)}} \quad (16.8.55)$$

$$= \hat{\mu}_{i,\eta_i} + \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} \quad (16.8.56)$$

$$< \mu_1 \quad (16.8.57)$$

$$< \text{UCB}_1(t - 1) \quad (16.8.58)$$

where the latter inequalities follow from event  $G_i$ . This leads to a contradiction in the behaviour of the algorithm because  $A_t \neq i$  if  $\text{UCB}_i(t - 1) < \text{UCB}_1(t - 1)$ , so our claim is substantiated. By an indicator variable expansion:

$$\mathbb{E}[T_i(n)] = \mathbb{E}\left[\mathbb{I}_{G_i} T_i(n) + \mathbb{I}_{\bar{G}_i} T_i(n)\right] \quad (16.8.59)$$

$$\leq \eta_i + n\mathbb{E}\left[\mathbb{I}_{\bar{G}_i}\right] \quad (16.8.60)$$

$$= \eta_i + n\Pr(\bar{G}_i) \quad (16.8.61)$$

$$\leq \eta_i + n^2\delta + n\exp\left(-\frac{\eta_i c^2 \Delta_i^2}{2}\right) \quad (16.8.62)$$

since  $G_i$  implies  $T_i(n) \leq \eta_i$  while  $T_i(n) \leq n$  always holds. As for  $\eta_i$ , choose it to be the smallest integer which  $\Delta_i - \sqrt{\frac{2 \log(1/\delta)}{\eta_i}} \geq c\Delta_i$  holds, i.e.

$$[\Delta_i(1 - c)]^2 \geq \frac{2 \log(1/\delta)}{\eta_i} \quad (16.8.63)$$

$$\eta_i \geq \frac{2 \log(1/\delta)}{[\Delta_i(1 - c)]^2} \quad (16.8.64)$$

Hence pick

$$\eta_i = \left\lceil \frac{2 \log(1/\delta)}{[\Delta_i(1 - c)]^2} \right\rceil \quad (16.8.65)$$

Note that this does not ensure that  $\eta_i \leq n$ , but if  $\eta_i > n$  then the bound  $\mathbb{E}[T_i(n)] \leq n < \eta_i$  trivially holds anyway. Substituting this choice of  $\eta_i$  as well as  $\delta = 1/n^2$ :

$$\mathbb{E}[T_i(n)] \leq \left\lceil \frac{2 \log(n^2)}{[\Delta_i(1 - c)]^2} \right\rceil + 1 + n \exp\left(-\left\lceil \frac{2 \log(n^2)}{[\Delta_i(1 - c)]^2} \right\rceil \frac{c^2 \Delta_i^2}{2}\right) \quad (16.8.66)$$

$$\leq \frac{2 \log(n^2)}{[\Delta_i(1 - c)]^2} + 2 + n \exp\left(-\frac{2 \log(n^2)}{[\Delta_i(1 - c)]^2} \cdot \frac{c^2 \Delta_i^2}{2}\right) \quad (16.8.67)$$

$$= \frac{2 \log(n^2)}{[\Delta_i(1 - c)]^2} + 2 + n \exp\left(-\frac{c^2}{(1 - c)^2} \log(n^2)\right) \quad (16.8.68)$$

$$= \frac{2 \log(n^2)}{[\Delta_i(1 - c)]^2} + 2 + n \cdot n^{-2c^2/(1-c)^2} \quad (16.8.69)$$

$$= \frac{2 \log(n^2)}{[\Delta_i(1-c)]^2} + 2 + n^{1-2c^2/(1-c)^2} \quad (16.8.70)$$

If we choose  $c = 1/2$ , this further simplifies to

$$\mathbb{E}[T_i(n)] \leq \frac{8 \log(n^2)}{\Delta_i^2} + 2 + n^{-1} \quad (16.8.71)$$

$$\leq \frac{16 \log(n)}{\Delta_i^2} + 3 \quad (16.8.72)$$

Finally, using the regret decomposition lemma  $R_n = \sum_{i=1}^k \Delta_i \mathbb{E}[T_i(n)]$ , we can show

$$R_n \leq 3 \sum_{i=1}^k \Delta_i + \sum_{i:\Delta_i>0} \frac{16 \log(n)}{\Delta_i^2} \quad (16.8.73)$$

where we need to restrict the summation in the second term to avoid dividing by zero.  $\square$

From this, another bound can be obtained which does not contain the reciprocal of  $\Delta_i$  (as a small  $\Delta_i$  can lead to a very loose bound).

**Corollary 16.1.** *With 1-sub-Gaussian rewards and  $\delta = 1/n^2$ , the regret of the UCB algorithm is bounded by*

$$R_n \leq 8\sqrt{nk \log n} + 3 \sum_{i=1}^k \Delta_i \quad (16.8.74)$$

*Proof.* Split the sum in the regret decomposition lemma into

$$R_n = \sum_{i:\Delta_i < d} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq d} \Delta_i \mathbb{E}[T_i(n)] \quad (16.8.75)$$

where  $d$  is a constant we can choose later. The first term can be upper bounded by  $nd$  since  $T_i(n) \leq n$  and  $\Delta_i < d$ . We then use the bound  $\mathbb{E}[T_i(n)] \leq \frac{16 \log(n)}{\Delta_i^2} + 3$  from above for the second term to give

$$R_n \leq nd + \sum_{i:\Delta_i \geq d} \Delta_i \left( \frac{16 \log(n)}{\Delta_i^2} + 3 \right) \quad (16.8.76)$$

$$\leq nd + \frac{16k \log n}{d} + 3 \sum_{i=1}^k \Delta_i \quad (16.8.77)$$

If we choose  $d = \sqrt{16k \log(n)/n}$ , this yields

$$R_n \leq \sqrt{16nk \log n} + \sqrt{16nk \log n} + 3 \sum_{i=1}^k \Delta_i \quad (16.8.78)$$

$$= 8\sqrt{nk \log n} + 3 \sum_{i=1}^k \Delta_i \quad (16.8.79)$$

$\square$

This is what is known as a *sublinear regret* bound because  $R_n = o(n)$ , i.e.  $\sqrt{n \log n}$  grows much slower than  $n$ . Note that the  $\sum_{i=1}^k \Delta_i$  term is present because the algorithm cannot avoid picking each arm at least once (since the upper confidences begin at infinity).

### Gradient Bandit Algorithm [195]

In the gradient bandit algorithm we maintain ‘preferences’ for each arm  $a$ , denoted by  $H_t(a)$  at round  $t$ . Choosing an arm consists of picking one from the soft-max distribution

$$\pi_t(a) := \Pr(A_t = a | H_t(1), \dots, H_t(k)) \quad (16.8.80)$$

$$= \frac{e^{H_t(a)}}{\sum_{i=1}^k e^{H_t(i)}} \quad (16.8.81)$$

where we suppress the dependence on preferences in  $\pi_t(a)$  for ease of notation. Hence every arm has a chance to be chosen, but highly preferred arms will be chosen more often, satisfying the exploitation-exploration tradeoff. After receiving reward  $X_t$ , we then update the preferences by

$$H_{t+1}(A_t) = H_t(A_t) + \alpha(X_t - \bar{X}_t)(1 - \pi_t(A_t)) \quad (16.8.82)$$

for the arm that was picked, and for every other arm,

$$H_{t+1}(a) = H_t(a) + \alpha(X_t - \bar{X}_t)(\mathbb{I}_{\{a=A_t\}} - \pi_t(a)) \quad (16.8.83)$$

where  $\bar{X}_t$  is the average of all rewards received so far:

$$\bar{X}_t = \frac{1}{t} \sum_{s=1}^t X_s \quad (16.8.84)$$

and  $\alpha > 0$  is a step-size/learning rate parameter. Compactly, we can write the update as

$$H_{t+1}(a) = H_t(a) + \alpha(X_t - \bar{X}_t)(\mathbb{I}_{\{a=A_t\}} - \pi_t(a)) \quad (16.8.85)$$

Intuitively, the algorithm works as follows. If  $X_t$  is above average, then the preference for  $A_t$  will increase, whereas all others will decrease. If  $X_t$  is below average, then the preferences of all other arms will increase, while the preference of  $A_t$  will decrease. The initial preferences are set to be all equal (e.g. zero), however it does not matter what number in particular because for any constant  $c$ ,

$$\frac{e^{H_t(a)+c}}{\sum_{i=1}^k e^{H_t(i)+c}} = \frac{e^c e^{H_t(a)+c}}{\sum_{i=1}^k e^c e^{H_t(i)}} \quad (16.8.86)$$

$$= \frac{e^{H_t(a)}}{\sum_{i=1}^k e^{H_t(i)}} \quad (16.8.87)$$

We can show that the gradient bandit algorithm update is actually performing stochastic gradient ascent (analogous to stochastic gradient descent). Let  $\mathbf{H}_t$  denote the vector of  $k$  preferences at round  $t$ . We wish to find the vector that maximises the expected reward given the preferences, i.e.  $\mathbb{E}[X_t | \mathbf{H} = \mathbf{H}_t]$ . Ideally, a gradient descent update would be given by

$$\mathbf{H}_{t+1} = \mathbf{H}_t + \alpha \nabla_{\mathbf{H}} \mathbb{E}[X_t | \mathbf{H} = \mathbf{H}_t] \quad (16.8.88)$$

or component-wise,

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial \mathbb{E}[X_t | \mathbf{H} = \mathbf{H}_t]}{\partial H_t(a)} \quad (16.8.89)$$

The expected reward can be written as

$$\mathbb{E}[X_t | \mathbf{H}_t] = \sum_{i=1}^k \Pr((A_t = i | \mathbf{H}_t) \mu_i) \quad (16.8.90)$$

$$= \sum_{i=1}^k \pi_t(i) \mu_i \quad (16.8.91)$$

Thus the derivative becomes

$$\frac{\partial \mathbb{E}[R_t | \mathbf{H}_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left( \sum_{i=1}^k \pi_t(i) \mu_i \right) \quad (16.8.92)$$

$$= \sum_{i=1}^k \mu_i \frac{\partial \pi_t(i)}{\partial H_t(a)} \quad (16.8.93)$$

where the  $\mu_i$  are the mean rewards. We argue that

$$\sum_{i=1}^k \frac{\partial \pi_t(i)}{\partial H_t(a)} = 0 \quad (16.8.94)$$

since the soft-max is always a valid probability distribution, so by changing one preference, the sum of changes in probabilities is always zero. Therefore we introduce an arbitrary baseline  $B_t$  (which will play a role later) into the sum:

$$\frac{\partial \mathbb{E}[X_t | \mathbf{H}_t]}{\partial H_t(a)} = \sum_{i=1}^k (\mu_i - B_t) \frac{\partial \pi_t(i)}{\partial H_t(a)} \quad (16.8.95)$$

$$= \sum_{i=1}^k \pi_t(i) (\mu_i - B_t) \frac{\partial \pi_t(i)}{\partial H_t(a)} / \pi_t(i) \quad (16.8.96)$$

$$= \mathbb{E} \left[ (\mu_{A_t} - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \middle| \mathbf{H}_t \right] \quad (16.8.97)$$

Since  $\mathbb{E}[X_t | A_t, \mathbf{H}_t] = \mu_{A_t}$ , then this becomes

$$\frac{\partial \mathbb{E}[R_t | \mathbf{H}_t]}{\partial H_t(a)} = \mathbb{E} \left[ (\mathbb{E}[R_t | A_t, \mathbf{H}_t] - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \middle| \mathbf{H}_t \right] \quad (16.8.98)$$

$$= \mathbb{E} \left[ \mathbb{E} \left[ (R_t - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \middle| A_t, \mathbf{H}_t \right] \middle| \mathbf{H}_t \right] \quad (16.8.99)$$

$$= \mathbb{E} \left[ (R_t - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \middle| \mathbf{H}_t \right] \quad (16.8.100)$$

via the law of iterated expectations. Now we focus on computing the partial derivative, which can be done using the quotient rule:

$$\frac{\partial \pi_t(A_t)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left( \frac{e^{H_t(A_t)}}{\sum_{i=1}^k e^{H_t(i)}} \right) \quad (16.8.101)$$

$$= \frac{\sum_{i=1}^k e^{H_t(i)} \frac{\partial e^{H_t(A_t)}}{\partial H_t(a)} - e^{H_t(A_t)} \frac{\partial}{\partial H_t(a)} \left( \sum_{i=1}^k e^{H_t(i)} \right)}{\left( \sum_{i=1}^k e^{H_t(i)} \right)^2} \quad (16.8.102)$$

$$= \frac{\mathbb{I}_{\{a=A_t\}} e^{H_t(A_t)} \sum_{i=1}^k e^{H_t(i)} - e^{H_t(A_t)} e^{H_t(a)}}{\left( \sum_{i=1}^k e^{H_t(i)} \right)^2} \quad (16.8.103)$$

where we have used

$$\frac{\partial e^{H_t(A_t)}}{\partial H_t(a)} = \mathbb{I}_{\{a=A_t\}} e^{H_t(A_t)} \quad (16.8.104)$$

Continuing,

$$\frac{\partial \pi_t(A_t)}{\partial H_t(a)} = \frac{\mathbb{I}_{\{a=A_t\}} e^{H_t(A_t)} \sum_{i=1}^k e^{H_t(i)}}{\left(\sum_{i=1}^k e^{H_t(i)}\right)^2} - \frac{e^{H_t(A_t)} e^{H_t(a)}}{\left(\sum_{i=1}^k e^{H_t(i)}\right)^2} \quad (16.8.105)$$

$$= \frac{\mathbb{I}_{\{a=A_t\}} e^{H_t(A_t)}}{\sum_{i=1}^k e^{H_t(i)}} - \frac{e^{H_t(A_t)} e^{H_t(a)}}{\left(\sum_{i=1}^k e^{H_t(i)}\right)^2} \quad (16.8.106)$$

$$= \mathbb{I}_{\{a=A_t\}} \pi_t(A_t) - \pi_t(A_t) \pi_t(a) \quad (16.8.107)$$

$$= \pi_t(A_t) (\mathbb{I}_{\{a=A_t\}} - \pi_t(a)) \quad (16.8.108)$$

Putting this back into above, we have

$$\frac{\partial \mathbb{E}[X_t | \mathbf{H}_t]}{\partial H_t(a)} = \mathbb{E}[(X_t - B_t) \pi_t(A_t) (\mathbb{I}_{\{a=A_t\}} - \pi_t(a)) / \pi_t(A_t) | \mathbf{H}_t] \quad (16.8.109)$$

$$= \mathbb{E}[(X_t - B_t) (\mathbb{I}_{\{a=A_t\}} - \pi_t(a)) | \mathbf{H}_t] \quad (16.8.110)$$

Let the baseline  $B_t = \bar{X}_t$  be the average reward. It can be chosen more generally, however this choice lends itself to a nice interpretation of the update. We then find

$$\frac{\partial \mathbb{E}[X_t | \mathbf{H}_t]}{\partial H_t(a)} = \mathbb{E}[(X_t - \bar{X}_t) (\mathbb{I}_{\{a=A_t\}} - \pi_t(a)) | \mathbf{H}_t] \quad (16.8.111)$$

so that the update

$$H_{t+1}(a) = H_t(a) + \alpha (X_t - \bar{X}_t) (\mathbb{I}_{\{a=A_t\}} - \pi_t(a)) \quad (16.8.112)$$

is a case of stochastic gradient ascent.

#### 16.8.4 Contextual Bandits

In a contextual bandit environment, the notion of a context  $C_t$  (which could be random) is made available to the algorithm before an arm is chosen, and the reward now depends on the context as well as the arm. To describe a single round:

- At the start of round  $t$ , the algorithm observes  $C_t \in \mathcal{C}$ .
- The algorithm chooses an arm  $A_t$  based on the available information up to round  $t$ , and the context  $C_t$ .
- The algorithm receives reward  $X_t$  drawn from the distribution  $\mathbb{P}_{A_t|C_t}$ .

We may write the reward as

$$X_t = r(C_t, A_t) + \eta_t \quad (16.8.113)$$

where  $r : \mathcal{C} \times \mathcal{A} \rightarrow \mathbb{R}$  is the reward function and  $\eta_t$  is noise, which we may assume to be zero-mean without loss of generality. Now, the best arm is defined as

$$A_t^* = \operatorname{argmax}_{a \in \mathcal{A}} r(C_t, A_t) \quad (16.8.114)$$

which will be random, since it depends on  $C_t$ . Because of this, the regret needs to be re-defined as

$$R_n = \sum_{t=1}^n \mathbb{E}[r(C_t, A_t^*) - X_t] \quad (16.8.115)$$

Intuitively, an algorithm should learn the reward mapping  $r(c, a)$  if it hopes to achieve low regret.

### Linear Contextual Bandits

In a linear contextual bandit setting, it is known that the reward function takes the form

$$r(c, a) = \theta^\top \psi(c, a) \quad (16.8.116)$$

where  $\theta \in \mathbb{R}^d$  is a parameter vector and  $\psi(c, a)$  can be thought of as a feature map or basis expansion. Hence the conditional mean rewards  $r(c, a)$  are linear in the basis expansion. A upper confidence bound strategy for this problem would involve keeping a point estimate and confidence set for  $\theta$ , instead of the mean rewards.

#### 16.8.5 Adversarial Bandits

In an adversarial bandit environment, we now consider there to be ‘adversary’ who can choose the rewards ahead of time. We can assume very little about how the adversary chooses rewards, and can even generally assume that the adversary has access to the bandit algorithm. In this sense, the adversary acts like a second-mover. After seeing the bandit algorithm, the adversary selects rewards

$$x = (x_1, \dots, x_n) \quad (16.8.117)$$

where each  $x_i \in \mathbb{X} \subseteq \mathbb{R}^k$  in a  $k$ -armed bandit setting. For simplicity, we can sometimes take  $\mathbb{X} = [0, 1]^k$  to prevent the adversary from choosing arbitrarily low rewards (as this would complicate regret analysis). Since the adversary could be actively trying to maximise the regret, a natural way to limit the second-mover advantage of the adversary is to include randomness in the bandit algorithm. Thus in round  $t$ :

- The algorithm samples action  $A_t$  from the randomised policy  $\pi_t(a|A_{t-1}, X_{t-1}, \dots, A_1, X_1)$ .
- The reward  $X_t = x_{t,A_t}$  is observed from the adversary’s choice.

The appropriate notion of regret here is to use the expected stochastic regret:

$$\mathbb{E} [\tilde{R}_n] = \max_{a \in \mathcal{A}} \left\{ \sum_{t=1}^n x_{t,a} \right\} - \mathbb{E} \left[ \sum_{t=1}^n X_t \right] \quad (16.8.118)$$

which may be interpreted as the expected regret in hindsight, from knowing the single best arm. Note that the expectation is over the randomness induced by the algorithm; we treat the adversary’s rewards as non-random.

#### Adversarial Linear Bandits

In an adversarial linear bandit environment, the adversary chooses rewards as before, and we assume that the action set  $\mathcal{A}$  spans  $\mathbb{R}^k$ , and the reward is given by

$$X_t = x_t^\top A_t \quad (16.8.119)$$

Note that this generalises the  $k$ -armed bandit setting, because we can take  $\mathcal{A} = \{\mathbf{e}_1, \dots, \mathbf{e}_k\}$  as the unit basis vectors of  $\mathbb{R}^k$ , but it also opens up the possibility where there are infinitely or even uncountably many arms.

#### 16.8.6 Non-Stationary Bandits

In a non-stationary bandit environment, the environment is changing over time, which generally means that the mean rewards are also changing over time. A simple model for the rewards of a non-stationary bandit is given by

$$X_t = \mu_{A_t}(t) + \eta_t \quad (16.8.120)$$

where  $\eta_t$  is an i.i.d. sequence of zero-mean random variables. To handle this, we can introduce the *non-stationary regret* with respect to horizon  $n$ :

$$R_n = \sum_{t=1}^n \mu^*(t) - \mathbb{E} \left[ \sum_{t=1}^n \mu_{A_t}(t) \right] \quad (16.8.121)$$

where the optimal arm is changing with each round:  $\mu^*(t) = \max_a \mu_a(t)$ . For a bandit algorithm to achieve low regret, it should try to ‘track’ the best arm.

### 16.8.7 Markovian Bandits

In a Markovian bandit environment with  $k$  arms  $\mathcal{A} = \{1, \dots, k\}$ , each arm  $a$  has its own state  $x_{a,t} \in \mathcal{X}_a$  at round  $t$ . The following describes the interaction with a Markovian bandit at round  $t$ :

- The algorithm selects arm  $a = A_t \in \mathcal{A}$  at round  $t$ .
- A reward  $r_a(x_{a,t})$  is received, which is just a function of the current state. There is a discounting factor of  $\gamma < 1$ , so the discounted present value of the reward at round  $t = 0$  is  $\gamma^t r_a(x_{a,t})$ .
- The state for that arm then transitions according to a Markov process defined by  $p_a(x_{a,t+1}|x_{a,t})$ .
- All the other arms are ‘frozen’ for round  $t$  in the sense that their states do not change, and no rewards are received from them. That is,  $x_{a',t+1} = x_{a',t}$  for all  $a' \neq A_t$ .

Note that this bandit environment induces a Markov decision process over the state-space  $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_k$  and state

$$x_t = (x_{1,t}, \dots, x_{k,t}) \quad (16.8.122)$$

Hence stochastic dynamic programming could be applied to find an optimal policy  $\pi^*$  for an infinite-horizon problem (finite-horizon problems are also possible), with the value function for the former being:

$$V(x) = \sup_{\pi} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_{A_t}(x_{A_t,t}) \middle| x_0 = x \right] \quad (16.8.123)$$

The Bellman equation can also be written as

$$V(x_t) = \max_{a \in \mathcal{A}} \left\{ r_a(x_{a,t}) + \gamma \sum_{x' \in \mathcal{X}_a} p_a(x'|x_{a,t}) V(x_{1,t}, \dots, x', \dots, x_{k,t}) \right\} \quad (16.8.124)$$

### Gittins Index [157]

For simplicity, consider a single-armed Markovian bandit. Introduce  $V_{\bar{r}}(x_t)$ , which is defined as

$$V_{\bar{r}}(x_t) = \max \{ \bar{r} + \gamma V_{\bar{r}}(x_t), \mathbb{E}[r(x_t) + \gamma V_{\bar{r}}(x_{t+1})|x_t] \} \quad (16.8.125)$$

This quantity represents the value at state  $x_t$  where we are also given the option to receive the reward  $\bar{r}$  with certainty, without transitioning the state. For infinite horizons, observe that if  $\bar{r}$  were chosen at round  $t$ , then we would also want to choose  $\bar{r}$  at round  $t+1$  because the state has remained unchanged. Hence the discounted value  $\bar{r} + \gamma V_{\bar{r}}(x_t)$  can be computed using the geometric series as

$$\bar{r} + \gamma V_{\bar{r}}(x_t) = \frac{\bar{r}}{1-\gamma} \quad (16.8.126)$$

Hence we can rewrite  $V_{\bar{r}}(x_t)$  as

$$V_{\bar{r}}(x_t) = \max \left\{ \frac{\bar{r}}{1-\gamma}, \mathbb{E}[r(x_t) + \gamma V_{\bar{r}}(x_{t+1})|x_t] \right\} \quad (16.8.127)$$

Let  $r^*(x_t)$  be the reward received with certainty that would make us indifferent between  $r^*(x_t)$  and pulling the arm, i.e. the value  $r^*(x_t)$  solves

$$\frac{r^*(x_t)}{1-\gamma} = \mathbb{E}[r(x_t) + \gamma V_{\bar{r}}(x_{t+1})|x_t] \quad (16.8.128)$$

for a given value of the current state  $x_t$ . Then  $r^*(x_t)$  is known as the Gittins index for the arm.

### Gittins Index Theorem [67]

Consider a multi-armed Markovian bandit, and let  $r_a^*(x_{a,t})$  denote the Gittins index computed for each arm  $a$ . The Gittins index theorem says that, rather than solving a full stochastic dynamic programming problem, the optimal policy is obtained by pulling the arm with the highest Gittins index, i.e.

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}} r_a^*(x_{a,t}) \quad (16.8.129)$$

This result feels intuitive, because the Gittins index captures the discounted certainty-equivalent value of an arm. Thus it makes sense for us to pick the arm with the highest certainty-equivalent value.

### Restless Bandits [37]

A restless bandit environment takes the form of a Markovian bandit environment, however the states of the arms are no longer frozen. Pulling any arm also generates an unobserved transition for every unchosen arm (but their rewards are not received).

#### 16.8.8 Bayesian Bandits [185]

In Bayesian bandits, we begin with a prior distribution  $\mathcal{D}$  over the reward distributions  $\mathcal{P}$ , and write  $\mathcal{P} \sim \mathcal{D}$ , where we recall  $\mathcal{P} = \{\mathbb{P}_a : a \in \mathcal{A}\}$ . For simplicity, we usually assume that  $\mathcal{P}$  consists of a known parametric class of distributions (with fixed and known number of arms), so that we are effectively only drawing the parameters from the prior  $\mathcal{D}$ . Examples of a parametric class are:

- Rewards are Bernoulli distributed, where mean rewards are themselves random and drawn from the prior.
- Rewards are Gaussian distributed, with known fixed variance (usually taken to be unit variance), and mean rewards drawn from the prior.

### Bayesian Regret

In the analysis of algorithms for Bayesian bandits, it is customary to assume that the reward distributions are actually drawn from the prior, i.e. we assert  $\mathcal{P} \sim \mathcal{D}$ . Introduce Bayesian regret as the expected conventional regret  $R_n$ , averaged over all possible reward distributions:

$$\mathfrak{B}_n = \mathbb{E}_{\mathcal{P} \sim \mathcal{D}} [R_n] \quad (16.8.130)$$

$$= \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - \mu_{A_t}) \right] \quad (16.8.131)$$

where the best arm  $A^*$  is now a random variable.

## Thompson Sampling [174]

Thompson sampling is a bandit algorithm for Bayesian bandit environments, based on Bayesian updating principles. Let

$$\mathcal{F}_t = \{A_t, X_t, \dots, A_1, X_1\} \quad (16.8.132)$$

denote the history of actions and rewards up to round  $t$ . Then the policy under Thompson sampling is

$$\pi_t(a|\mathcal{F}_t) = \Pr(A^* = a|\mathcal{F}_t) \quad (16.8.133)$$

$$= \Pr_{\mathcal{D}|\mathcal{F}_t}(A^* = a) \quad (16.8.134)$$

What this means is that we maintain the posterior distribution for  $\mathcal{D}$  given the history  $\mathcal{F}_t$ , denoted  $\mathcal{D}|\mathcal{F}_t$ , and sample the action  $A_t$  according to the probability that is the best arm under the posterior. This satisfies the exploration-exploitation tradeoff by:

- Exploration by allowing for other arms to be pulled by sampling randomly (rather than always choosing the arm with the greatest posterior probability to be the best arm).
- Exploitation by weighting the arms so that the best posterior arm is the most likely to be pulled.

An equivalent characterisation of Thompson sampling (which is perhaps simpler to implement) is to sample a vector of the mean rewards  $\hat{\mu}$  from the posterior  $\mathcal{D}|\mathcal{F}_t$ , and then choose the arm by

$$A_t = \operatorname{argmax}_a \hat{\mu}_a \quad (16.8.135)$$

To show why this is equivalent, we can write for each  $a$ :

$$\Pr(A_t = a|\mathcal{F}_t) = \Pr\left(\operatorname{argmax}_i \hat{\mu}_i = a \middle| \mathcal{F}_t\right) \quad (16.8.136)$$

$$= \Pr_{\mathcal{D}|\mathcal{F}_t}(A^* = a) \quad (16.8.137)$$

$$= \Pr(A^* = a|\mathcal{F}_t) \quad (16.8.138)$$

$$= \pi_t(a|\mathcal{F}_t) \quad (16.8.139)$$

where here,  $A^*$  specifically denotes a random variable for the best arm if mean rewards were drawn from  $\mathcal{D}|\mathcal{F}_t$ . Intuitively, since we sample  $\hat{\mu}$  from  $\mathcal{D}|\mathcal{F}_t$ , then by construction each arm has the desired probability of being the greatest.

The posterior from Thompson sampling is generally intractable analytically, thus the posterior typically needs to be approximated using a variety of approaches, or otherwise a conjugate pair needs to be imposed on the distributions.

## 16.9 Reinforcement Learning

### 16.9.1 Markov Decision Problems

To formulate a reinforcement learning problem, we augment a Markov decision process with rewards. Consider a Markov decision process with state-space  $\mathcal{S}$ , action space  $\mathcal{A}(s) \subseteq \mathcal{A}$  for  $s \in \mathcal{S}$  and transition probability:

$$p(s'|s, a) = \Pr(S_{t+1} = s' | S_t = s, A_t = a) \quad (16.9.1)$$

We use the term *agent* to describe the party that takes actions based on the current state (i.e. another name for controller). For a given pair  $(S_t, A_t)$  which is the action  $A_t$  taken at state  $S_t$ ,

the agent receives a (possibly random) reward  $R_{t+1} \in \mathcal{R}$ . Note that whether to assign a time index of  $t + 1$  or  $t$  to this reward is arbitrary, however we use  $t + 1$  for convention. The agent's decisions are governed by the policy:

$$\pi(a|s) = \Pr(A_t = a \in \mathcal{A}(s)|S_t = s) \quad (16.9.2)$$

The agent together with the Markov decision process gives rise to a *trajectory* of random variables:  $S_0, A_0, R_1, S_1, A_1, R_2, \dots$  etc. For simplicity, we will assume that  $\mathcal{S}, \mathcal{A}, \mathcal{R}$  are finite sets, which means the random variables  $S_t, A_t, R_t$  are discrete. Additionally, we introduce an augmented transition probability which also encompasses rewards:

$$p(s', r|s, a) = \Pr(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a) \quad (16.9.3)$$

This defines the dynamics of the *environment* (i.e. system). An alternative specification of the rewards is to write

$$R_{t+1} = r(S_t, A_t) + \eta_t \quad (16.9.4)$$

for some reward function  $r(s, a)$  and noise  $\eta$ . This representation allows us to relax the assumption that  $\mathcal{R}$  is a finite set, and instead  $R_t$  could be a continuous random variable with a conditional density  $p(r|s, a)$  which is conditionally independent with  $s'$ , given  $(s, a)$ .

Denote  $G_t$  as the discounted sum of rewards from time  $t$ :

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (16.9.5)$$

$$= \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} \quad (16.9.6)$$

with discount rate  $0 \leq \gamma \leq 1$ . Note that for *episodic tasks* corresponding to an episodic Markov decision process (where there is some absorbing state which signifies the termination of the process), this can be handled by setting the rewards from the terminal state to zero. The goal of the Markov decision problem is to find the policy  $\pi^*$  which minimises the expected discounted sum of future rewards from each state:

$$V_{\pi}(s) = \mathbb{E}_{\pi}[G_t|S_t = s] \quad (16.9.7)$$

The function  $V_{\pi}(s)$  is called the *state-value function* (or just the value function). If the transition probabilities are known, then the Markov decision problem can in principle be solved using stochastic dynamic programming, and the optimal value function  $V^*(s) = V_{\pi^*}(s)$  follows the Bellman equation:

$$V^*(s) = \max_{a \in \mathcal{A}(s)} \mathbb{E}[R_{t+1} + \gamma V^*(S_{t+1})|S_t = s, A_t = a] \quad (16.9.8)$$

Reinforcement learning (also known as approximate dynamic programming) is a collection of strategies and techniques to address situations where standard stochastic dynamic programming cannot be applied straightforwardly. This is typically when the transition dynamics are not given, and/or the state and action spaces are large.

## Action-Value Functions

The action-value function  $Q_{\pi}(s, a)$  is defined as the expected discounted sum of rewards for choosing action  $a$  at state  $s$ , the following policy  $\pi$  thereafter. It can be written as

$$Q_{\pi}(s, a) = \mathbb{E}[R_{t+1} + \gamma V_{\pi}(S_{t+1})|S_t = s, A_t = a] \quad (16.9.9)$$

The optimal action-value function  $Q^*(s, a)$  is the action-value function for the optimal policy:

$$Q^*(s, a) = \mathbb{E}[R_{t+1} + \gamma V^*(S_{t+1})|S_t = s, A_t = a] \quad (16.9.10)$$

From this we can see that the optimal value function can be also written in terms of the optimal action-value function by

$$V^*(s) = \max_{a \in \mathcal{A}(s)} Q^*(s, a) \quad (16.9.11)$$

### $\varepsilon$ -Greedy Policies

The action-value function is useful for deriving policies ‘on-the-spot’. An  $\varepsilon$ -greedy policy for an action-value function  $Q_\pi(s, a)$  is to choose action  $a^* = \max_{a \in \mathcal{A}(s)} Q_\pi(s, a)$  with probability  $1 - \varepsilon$ , and then with probability  $\varepsilon$  choose a random action from  $\mathcal{A}(s)$  uniformly. The selection of  $\varepsilon$  determines the exploration-exploitation tradeoff.

From the relation  $V^*(s) = \max_{a \in \mathcal{A}(s)} Q^*(s, a)$ , this shows that the  $\varepsilon$ -greedy policy derived from the optimal action-value function  $Q^*(s, a)$  approaches the optimal policy  $\pi^*$  as  $\varepsilon \rightarrow 0$ . This then motivates strategies for learning the optimal action-value function  $Q^*(s, a)$ .

## 16.9.2 Monte-Carlo Approximate Dynamic Programming

In Monte-Carlo approaches for approximate dynamic programming, we assume that the model is unknown (i.e. we do not know the probability transitions of all possible transitions), however we do assume that sample transitions from the model can be generated.

### Monte-Carlo Policy Evaluation [195]

Suppose we have a policy  $\pi$  that we wish to evaluate the value function of, for an episodic Markov decision process. Then a ‘first-visit’ Monte-Carlo approach is to generate an episode following  $\pi$  for  $T$  time-steps which results in the sequence  $S_0, A_0, R_1, S_1, A_2, R_2, \dots, S_{T-1}, A_{T-1}, R_T$ . From this sampled episode, we can find the first-occurring instance of each state  $s \in \mathcal{S}$  that was visited, and compute the discounted sum of total rewards from that first occurrence. The estimated value function  $\widehat{V}_\pi(s)$  for each  $s \in \mathcal{S}$  is taken as the averaged discounted sum of total rewards over all the simulation replications for that state.

### Monte-Carlo Action-Value Function Evaluation

Monte-Carlo approaches can be used in the same manner as above to evaluate the action-value function  $Q_\pi(s, a)$  for every state-action pair. However, in order for the method to be effective, the policy  $\pi$  should allow for every state-action pair to be encountered throughout the simulation (i.e. be sufficiently ‘explorative’).

### Monte-Carlo Tree Search [142]

## 16.9.3 Temporal Differences

### One-Step Temporal Differences [195]

From the Bellman operator, we can write

$$V_\pi(s) = \mathbb{E}_\pi [R + \gamma V_\pi(S')] \quad (16.9.12)$$

where  $R$  is the reward from taking an action following  $\pi$  at  $s$ , and  $S'$  is the respective transition. This motivates the following weighted average update rule to estimate the value function for a given policy:

$$\widehat{V}_\pi(s) \leftarrow (1 - \alpha) \widehat{V}_\pi(s) + \alpha (R + \gamma \widehat{V}_\pi(S')) \quad (16.9.13)$$

where  $\alpha \in (0, 1]$  is the learning rate. This can be repeated for every state encountered over as many episodes as desired.

### SARSA Algorithm

The SARSA algorithm applies one-step temporal differences to find the optimal action-value function, rather than evaluating the state-value function. Each update operates on the observed quintuple  $(S, A, R, S', A')$ , hence its namesake. This update is given by (assuming some initial  $\widehat{Q}^*(\cdot, \cdot)$ ):

$$\widehat{Q}^*(S, A) \leftarrow (1 - \alpha) \widehat{Q}^*(S, A) + \alpha \left( R + \gamma \widehat{Q}^*(S', A') \right) \quad (16.9.14)$$

where  $A$  is the action taken at  $S$  by following a policy derived from the current  $\widehat{Q}^*(\cdot, \cdot)$ , while  $R$  is the respective reward for the transition to  $S'$ . Then  $A'$  is the action taken at  $S'$  by following a policy also derived from the current  $\widehat{Q}^*(\cdot, \cdot)$ . SARSA is known as an ‘on-policy’ approach because updates are made using the same action as the policy derived from the current action-value function.

### $Q$ -Learning

$Q$ -learning can be thought of as a ‘greedier’ version of SARSA, with the updates

$$\widehat{Q}^*(S, A) \leftarrow (1 - \alpha) \widehat{Q}^*(S, A) + \alpha \left( R + \gamma \widehat{Q}^*(S', a^*) \right) \quad (16.9.15)$$

where  $S, A, R$  and  $S'$  as before with SARSA, however now  $a^*$  is chosen as  $a^* = \operatorname{argmax}_a \widehat{Q}^*(S', a)$  using the current estimate of the action-value function. In this sense,  $Q$ -learning is regarded as an ‘off-policy’ approach because the update is made using a different policy than the one used to generate  $A$ .

$Q$ -learning can be seen as stochastic approximation. Using the Bellman equation for the action-value function, we have

$$Q^*(s, a) = \mathbb{E} [\bar{r}(s, a) + \gamma V^*(s^+) | s, a] \quad (16.9.16)$$

$$= Q^*(s, a) = \mathbb{E} \left[ \bar{r}(s, a) + \gamma \max_{a' \in \mathcal{A}(s^+)} Q^*(s^+, a') \middle| s, a \right] \quad (16.9.17)$$

where to simplify notation, we use  $s^+$  to denote the state transitioned to after taking action  $a$  from state  $s$ . Also,  $\bar{r}(s, a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$  is the average reward from  $(s, a)$ . Rearranging the Bellman equation,

$$\mathbb{E} \left[ \bar{r}(s, a) + \gamma \max_{a' \in \mathcal{A}(s^+)} Q^*(s^+, a') - Q^*(s, a) \middle| s, a \right] = 0 \quad (16.9.18)$$

Thus the problem of stochastic dynamic programming amounts to finding a function  $Q^*(s, a)$  satisfying this equality pointwise for all pairs  $(s, a)$ . Although this is a conditional expectation, we can still derive a stochastic approximation algorithm if we condition on  $(s, a)$ . Thus given  $(s, a)$ , we observe a random  $s^+$  and perform a Robbins-Monro update of the form

$$\widehat{Q}^*(s, a) \leftarrow \widehat{Q}^*(s, a) + \eta \left( \bar{r}(s, a) + \gamma \max_{a' \in \mathcal{A}(s^+)} \widehat{Q}^*(s^+, a') - \widehat{Q}^*(s, a) \right) \quad (16.9.19)$$

If we do not have access to  $\bar{r}(s, a)$ , it is valid to replace it by the observed reward  $R$ , since we still have unbiased estimate:

$$\mathbb{E} \left[ R + \gamma \max_{a' \in \mathcal{A}(s^+)} Q^*(s^+, a') - Q^*(s, a) \middle| s, a \right] = \mathbb{E} \left[ \mathbb{E}[R | s, a] + \gamma \max_{a' \in \mathcal{A}(s^+)} Q^*(s^+, a') - Q^*(s, a) \middle| s, a \right] \quad (16.9.20)$$

$$= \mathbb{E} \left[ \bar{r}(s, a) + \gamma \max_{a' \in \mathcal{A}(s^+)} Q^*(s^+, a') - Q^*(s, a) \middle| s, a \right] \quad (16.9.21)$$

The policy used to generate the actions can be somewhat arbitrary, however it should ideally satisfy a reasonable exploration-exploitation tradeoff (e.g. use  $\varepsilon$ -greedy with the current estimate  $\hat{Q}^*(s, a)$ ). This is so that high rewards will still be received in the process of learning, but the algorithm will still have the ability to sample all state-action pairs.

#### 16.9.4 Value Function Approximation

When it is infeasible to store the value function for every single state (such as when the state-space is very large), a parametric approach can be used to approximate the value function for policy  $\pi$ , parametrised on some weights  $\mathbf{w}$ . The approximation is denoted

$$\hat{V}_\pi(s; \mathbf{w}) \approx V_\pi(s) \quad (16.9.22)$$

A reasonable metric to quantify the closeness of approximation is the weighted sum of squared errors:

$$J(\mathbf{w}) = \frac{1}{2} \sum_{s \in \mathcal{S}} \mu(s) (V_\pi(s) - \hat{V}_\pi(s; \mathbf{w}))^2 \quad (16.9.23)$$

where the weighting distribution  $\mu(s)$  can for example be chosen as the typical fraction of time spent at  $s$  under policy  $\pi$ .

#### Stochastic Gradient Value Function Approximation

Gradient descent can be used to derive a stochastic gradient descent update rule to learn  $\hat{V}_\pi(s; \mathbf{w})$ . The usual gradient descent step for minimising  $J(\mathbf{w})$  is to perform

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} J(\mathbf{w}_t) \quad (16.9.24)$$

with positive step size  $\alpha$ . As  $V_\pi(s)$  is treated as unattainable, performing an update of this form is not possible. Rather, we can replace  $\nabla_{\mathbf{w}} J(\mathbf{w}_t)$  with stochastic approximations. Suppose state  $S_t$  is visited at time  $t$ . Now consider the gradient approximation

$$\hat{\nabla}_{\mathbf{w}} J(\mathbf{w}_t) = \nabla_{\mathbf{w}} \left( \frac{1}{2} (U_\pi(S_t) - \hat{V}_\pi(S_t; \mathbf{w}))^2 \right) \quad (16.9.25)$$

where  $U_\pi(s)$  is an unbiased estimate for  $V_\pi(s)$  for all  $s \in \mathcal{S}$ , and this implicitly assumes that  $\hat{V}_\pi(S_t; \mathbf{w})$  is differentiable with respect to  $\mathbf{w}$ . Taking expectations of this approximation, we can show it is unbiased, by assuming that the distribution of time spent at state  $S_t$  is indeed  $\mu(S_t)$ :

$$\mathbb{E} [\hat{\nabla}_{\mathbf{w}} J(\mathbf{w}_t)] = \mathbb{E} \left[ \nabla_{\mathbf{w}} \left( \frac{1}{2} (U_\pi(S_t) - \hat{V}_\pi(S_t; \mathbf{w}_t))^2 \right) \right] \quad (16.9.26)$$

$$= \nabla_{\mathbf{w}} \mathbb{E} \left[ \frac{1}{2} (U_\pi(S_t) - \hat{V}_\pi(S_t; \mathbf{w}_t))^2 \right] \quad (16.9.27)$$

$$= \nabla_{\mathbf{w}} \left( \frac{1}{2} \sum_{s \in \mathcal{S}} \Pr(S_t = s) \mathbb{E} \left[ (U_\pi(S_t) - \hat{V}_\pi(S_t; \mathbf{w}_t))^2 \middle| S_t = s \right] \right) \quad (16.9.28)$$

$$= \nabla_{\mathbf{w}} \left( \frac{1}{2} \sum_{s \in \mathcal{S}} \Pr(S_t = s) (V_\pi(s) - \hat{V}_\pi(s; \mathbf{w}_t))^2 \right) \quad (16.9.29)$$

$$= \nabla_{\mathbf{w}} \left( \frac{1}{2} \sum_{s \in \mathcal{S}} \mu(s) \left( V_{\pi}(s) - \widehat{V}_{\pi}(s; \mathbf{w}_t) \right)^2 \right) \quad (16.9.30)$$

$$= \nabla_{\mathbf{w}} J(\mathbf{w}_t) \quad (16.9.31)$$

Hence we have the stochastic gradient descent update rule:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{1}{2} \alpha \nabla_{\mathbf{w}} \left( \frac{1}{2} \left( U_{\pi}(S_t) - \widehat{V}_{\pi}(S_t; \mathbf{w}_t) \right)^2 \right) \quad (16.9.32)$$

$$= \mathbf{w}_t + \alpha \left( U_{\pi}(S_t) - \widehat{V}_{\pi}(S_t; \mathbf{w}_t) \right) \nabla_{\mathbf{w}} \widehat{V}_{\pi}(S_t; \mathbf{w}_t) \quad (16.9.33)$$

by the chain rule. An example of a choice for  $U_{\pi}(S_t)$  as an unbiased estimate of  $V_{\pi}(S_t)$  (conditional on  $S_t$ ) is a Monte-Carlo policy evaluation on simulated episodes. These stochastic gradient descent updates can be iterated over episodes until convergence of  $\mathbf{w}$ .

### Stochastic Semi-Gradient Value Function Approximation

Instead of requiring an unbiased estimate  $U_{\pi}(S_t)$ , one option is to estimate  $V_{\pi}(S_t)$  using the current value function approximation as  $V_{\pi}(S_t) \approx R_{t+1} + \gamma \widehat{V}_{\pi}(S_{t+1}; \mathbf{w}_t)$ , where  $R_t$  is the reward observed for following policy  $\pi$ , and  $S_{t+1}$  is the observed next state transitioned to. This yields the update rule

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left( R_{t+1} + \gamma \widehat{V}_{\pi}(S_{t+1}; \mathbf{w}_t) - \widehat{V}_{\pi}(S_t; \mathbf{w}_t) \right) \nabla_{\mathbf{w}} \widehat{V}_{\pi}(S_t; \mathbf{w}_t) \quad (16.9.34)$$

Note that in general this estimate is biased since it depends on the current weights  $\mathbf{w}_t$ , hence the method is often referred to as a stochastic semi-gradient method. The advantage of using this method is that it provides an algorithm for online learning.

### Semi-Gradient SARSA

Value function approximation learning methods can be extended to approximation of the optimal action-value function. This parametrised approximation is denoted

$$\widehat{Q}^*(s, a; \mathbf{w}) \approx Q^*(s, a) \quad (16.9.35)$$

Analogous to stochastic semi-gradient value function approximation, a semi-gradient SARSA method using the update rule

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left( R_{t+1} + \gamma \widehat{Q}^*(S_{t+1}, A_{t+1}; \mathbf{w}_t) - \widehat{Q}^*(S_t, A_t; \mathbf{w}_t) \right) \nabla_{\mathbf{w}} \widehat{Q}^*(S_t, A_t; \mathbf{w}_t) \quad (16.9.36)$$

### 16.9.5 Policy Gradients

Policy gradient methods attempt to learn the optimal policy via a parametric approximation

$$\widehat{\pi}^*(a|s; \boldsymbol{\theta}) \approx \pi^*(a|s) \quad (16.9.37)$$

To ensure the policy is ‘explorative’ enough (i.e. the policy is not deterministic per se, but may be allowed to approach a deterministic policy), one option when the action space is finite is to parametrise a function  $h(s, a; \boldsymbol{\theta})$ , and then take a softmax to determine action selection probabilities:

$$\widehat{\pi}^*(a|s; \boldsymbol{\theta}) = \frac{e^{h(s, a; \boldsymbol{\theta})}}{\sum_{a' \in \mathcal{A}(s)} e^{h(s, a'; \boldsymbol{\theta})}} \quad (16.9.38)$$

An appropriate performance metric to evaluate the optimal policy approximation is the value function under the policy:

$$J(\boldsymbol{\theta}) = V_{\pi_{\boldsymbol{\theta}}}(s_0) \quad (16.9.39)$$

where  $s_0$  is the starting state (assumed non-random), and  $\pi_{\boldsymbol{\theta}} = \pi(a|s; \boldsymbol{\theta})$  is the policy under  $\boldsymbol{\theta}$ . Hence we aim to find

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} V_{\pi_{\boldsymbol{\theta}}}(s_0) \quad (16.9.40)$$

### Policy Gradient Theorem

#### REINFORCE Algorithm

Applying policy  $\hat{\pi}^*(a|s; \boldsymbol{\theta})$  from starting state  $s_0$  induces a distribution over the trajectories  $\boldsymbol{\tau} = (s_0, A_0, R_1, S_1, A_1, \dots)$ . This probability distribution can be written as

$$p(\boldsymbol{\tau}; \boldsymbol{\theta}) = \prod_{t=0}^{L(\boldsymbol{\tau})} p(S_{t+1}, R_{t+1}|S_t, A_t) \hat{\pi}^*(A_t|S_t; \boldsymbol{\theta}) \quad (16.9.41)$$

where  $L(\boldsymbol{\tau})$  is the length of trajectory  $\boldsymbol{\tau}$  (until a terminal state is reached). Denote the support of this distribution by the set  $\mathcal{T}$ . Denote  $g(\boldsymbol{\tau})$  as the sum of discounted rewards from trajectory  $\boldsymbol{\tau}$ :

$$g(\boldsymbol{\tau}) = \sum_{t=1}^{L(\boldsymbol{\tau})} \gamma^{t-1} R_t \quad (16.9.42)$$

Then the performance criterion can be expressed as

$$J(\boldsymbol{\theta}) = V_{\pi_{\boldsymbol{\theta}}}(s_0) \quad (16.9.43)$$

$$= \mathbb{E}_{\pi_{\boldsymbol{\theta}}} [G_0 | S_0 = s_0] \quad (16.9.44)$$

$$= \mathbb{E}_{\pi_{\boldsymbol{\theta}}} [g(\boldsymbol{\tau})] \quad (16.9.45)$$

$$= \sum_{\boldsymbol{\tau} \in \mathcal{T}} g(\boldsymbol{\tau}) p(\boldsymbol{\tau}; \boldsymbol{\theta}) \quad (16.9.46)$$

Hence taking the gradient yields

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \sum_{\boldsymbol{\tau} \in \mathcal{T}} g(\boldsymbol{\tau}) p(\boldsymbol{\tau}; \boldsymbol{\theta}) \quad (16.9.47)$$

$$= \sum_{\boldsymbol{\tau} \in \mathcal{T}} g(\boldsymbol{\tau}) \nabla_{\boldsymbol{\theta}} p(\boldsymbol{\tau}; \boldsymbol{\theta}) \quad (16.9.48)$$

$$= \sum_{\boldsymbol{\tau} \in \mathcal{T}} g(\boldsymbol{\tau}) \nabla_{\boldsymbol{\theta}} p(\boldsymbol{\tau}; \boldsymbol{\theta}) \times \frac{p(\boldsymbol{\tau}; \boldsymbol{\theta})}{p(\boldsymbol{\tau}; \boldsymbol{\theta})} \quad (16.9.49)$$

$$= \sum_{\boldsymbol{\tau} \in \mathcal{T}} g(\boldsymbol{\tau}) \left( \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\tau}; \boldsymbol{\theta})}{p(\boldsymbol{\tau}; \boldsymbol{\theta})} \right) p(\boldsymbol{\tau}; \boldsymbol{\theta}) \quad (16.9.50)$$

Note by the chain rule that  $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\tau}; \boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\tau}; \boldsymbol{\theta})}{p(\boldsymbol{\tau}; \boldsymbol{\theta})}$ , hence

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{\boldsymbol{\tau} \in \mathcal{T}} (g(\boldsymbol{\tau}) \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\tau}; \boldsymbol{\theta})) p(\boldsymbol{\tau}; \boldsymbol{\theta}) \quad (16.9.51)$$

$$= \mathbb{E}_{\pi_{\boldsymbol{\theta}}} [g(\boldsymbol{\tau}) \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\tau}; \boldsymbol{\theta})] \quad (16.9.52)$$

Moreover  $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\tau}; \boldsymbol{\theta})$  can be evaluated without knowing the transition probabilities because:

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\tau}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \left( \prod_{t=0}^{L(\boldsymbol{\tau})} p(S_{t+1}, R_{t+1}|S_t, A_t) \hat{\pi}^*(A_t|S_t; \boldsymbol{\theta}) \right) \quad (16.9.53)$$

$$= \nabla_{\boldsymbol{\theta}} \sum_{t=0}^{L(\boldsymbol{\tau})} (\log p(S_{t+1}, R_{t+1} | S_t, A_t) + \log \hat{\pi}^*(A_t | S_t; \boldsymbol{\theta})) \quad (16.9.54)$$

$$= \sum_{t=0}^{L(\boldsymbol{\tau})} \nabla_{\boldsymbol{\theta}} \log \hat{\pi}^*(A_t | S_t; \boldsymbol{\theta}) \quad (16.9.55)$$

Therefore

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[ g(\boldsymbol{\tau}) \sum_{t=0}^{L(\boldsymbol{\tau})} \nabla_{\boldsymbol{\theta}} \log \hat{\pi}^*(A_t | S_t; \boldsymbol{\theta}) \right] \quad (16.9.56)$$

From this we derive the REINFORCE algorithm, which is to generate an episode following  $\pi_{\boldsymbol{\theta}}$ , observe the trajectory  $\boldsymbol{\tau}$  and the return  $g(\boldsymbol{\tau})$ , then update the parameters  $\boldsymbol{\theta}$  by stochastic gradient ascent:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k g(\boldsymbol{\tau}_k) \sum_{t=0}^{L(\boldsymbol{\tau}_k)} \nabla_{\boldsymbol{\theta}} \log \hat{\pi}^*(A_{t,k} | S_{t,k}; \boldsymbol{\theta}_k) \quad (16.9.57)$$

where  $\{\alpha_k\}$  is the stepsize sequence.

### 16.9.6 Actor-Critic Methods

Actor-critic methods learn both a policy (actor) and a state-value function (critic), so that incremental online learning is possible. An approximation of the optimal policy  $\hat{\pi}^*(a|s; \boldsymbol{\theta}) \approx \pi^*(a|s)$  and of the value function  $\hat{V}_{\pi_{\boldsymbol{\theta}}}(s; \mathbf{w}) \approx V_{\pi_{\boldsymbol{\theta}}}(s)$  are simultaneously updated. In the update for the policy parameters  $\boldsymbol{\theta}$ , the full return of a trajectory  $g(\boldsymbol{\tau})$  is replaced by a discounted one-step return plus a value using the learned value function. That is,

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_{\boldsymbol{\theta}} \gamma^{t-1} \left( R_{t+1} + \gamma \hat{V}_{\pi_{\boldsymbol{\theta}}}(S_{t+1}; \mathbf{w}_t) \right) \nabla_{\boldsymbol{\theta}} \log \hat{\pi}^*(A_t | S_t; \boldsymbol{\theta}_t) \quad (16.9.58)$$

using stepsize  $\alpha_{\boldsymbol{\theta}}$ . The factor of  $\gamma^{t-1}$  is because in the original policy gradient update, successive rewards along the trajectory  $\boldsymbol{\tau}$  are discounted by  $\gamma$ , hence we do the same here. Meanwhile,  $\mathbf{w}_t$  is simultaneously updated using an online value function approximation learning algorithm such as one-step temporal differences using stepsize  $\alpha_{\mathbf{w}}$ .

## Chapter 17

# Quantitative Finance

### 17.1 Copulae

A copula is a multivariate probability distribution on support the unit hypercube, with uniform marginal distributions. That is, a copula can be treated as a cumulative distribution function  $C : [0, 1]^d \rightarrow [0, 1]$  satisfying the following properties.

- $C(0, \dots, 0) = 0$ .
- $C$  is  $d$ -non-decreasing. In the bivariate case, this means that for any two intervals  $[u_1, v_1] \subseteq [0, 1]$ ,  $[u_2, v_2] \subseteq [0, 1]$ , we have

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \quad (17.1.1)$$

This is the same as taking the 2-difference, analogous to the bivariate mass function derived from the cumulative distribution function. In the  $d$ -variate case, we take the  $d$ -difference.

- Since a copula has uniform marginals, this means

$$C(1, \dots, 1, u, 1, \dots, 1) = u \quad (17.1.2)$$

By the last property, this means  $C(1, \dots, 1, 0, 1, \dots, 1) = 0$ . But then due to the first and second properties, this implies

$$C(u_1, \dots, u_{i-1}, 0, u_{i+1}, \dots, u_d) = 0 \quad (17.1.3)$$

The utility of copulae are demonstrated in the following way. Suppose we have a continuous random vector  $(X_1, \dots, X_d)$  with marginal CDFs  $F_1, \dots, F_d$ . By the inverse transform sampling method, if we let

$$(U_1, \dots, U_d) = (F_1(X_1), \dots, F_d(X_d)) \quad (17.1.4)$$

then  $(U_1, \dots, U_d)$  will have Uniform  $(0, 1)$  marginals, and so will be a Copula distribution that has the same underlying dependence structure as  $(X_1, \dots, X_d)$ . Then this copula  $C$  is said to be the copula associated with the distribution of  $(X_1, \dots, X_d)$ . Conversely, if we have the quantile functions  $F_1^{-1}, \dots, F_d^{-1}$  and some way to generate a sample  $(U_1, \dots, U_d)$  from the copula of  $(X_1, \dots, X_d)$ , then we can generate a sample of  $(X_1, \dots, X_d)$  by

$$(X_1, \dots, X_d) = (F_1^{-1}(U_1), \dots, F_d^{-1}(U_d)) \quad (17.1.5)$$

From this, we can also see that the associated copula is invariant to monotonic transformations of the marginals (i.e. transforming each univariate variable with a strictly increasing function will not change the associated copula of a distribution).

### 17.1.1 Sklar's Theorem [102]

Sklar's theorem asserts that the copula associated with a continuous multivariate distribution with CDF  $F(x_1, \dots, x_d)$  and marginals  $F_1, \dots, F_d$  is unique, given by

$$C(u_1, \dots, u_d) = \Pr(U_1 \leq u_1, \dots, U_d \leq u_d) \quad (17.1.6)$$

$$= \Pr(F_1^{-1}(U_1) \leq F_1^{-1}(u_1), \dots, F_d^{-1}(U_d) \leq F_d^{-1}(u_d)) \quad (17.1.7)$$

$$= \Pr(X_1 \leq F_1^{-1}(u_1), \dots, X_d \leq F_d^{-1}(u_d)) \quad (17.1.8)$$

$$= F(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)) \quad (17.1.9)$$

This is because beginning from the CDF  $F(x_1, \dots, x_d)$ , we can perform all the same steps in reverse to obtain the same copula. However, if the distribution is discrete, then the associated copula is non-unique (i.e. we can find more than one copula from which we can use to generate a sample of  $X_1, \dots, X_d$ ). This is because  $F_i(U_i)$  will no longer be Uniform(0, 1) distributed, so we cannot perform the same steps in reverse to obtain a unique copula.

### 17.1.2 Fréchet–Hoeffding Bounds [102]

The Fréchet–Hoeffding bounds are upper and lower bounds on the copula distribution function such that

$$\underline{C}(u_1, \dots, u_d) \leq C(u_1, \dots, u_d) \leq \overline{C}(u_1, \dots, u_d) \quad (17.1.10)$$

for any  $(u_1, \dots, u_d) \in [0, 1]^d$ . To derive these bounds, first consider a bivariate copula  $C(u_1, u_2) = \Pr(U_1 \leq u_1, U_2 \leq u_2)$ . Then

$$\Pr(U_1 \leq u_1) + \Pr(U_2 \leq u_2) - 1 \leq \Pr(U_1 \leq u_1) + \Pr(U_2 \leq u_2) - \Pr(U_1 \leq u_1 \cup U_2 \leq u_2) \quad (17.1.11)$$

$$= \Pr(U_1 \leq u_1, U_2 \leq u_2) \quad (17.1.12)$$

$$\leq \min\{\Pr(U_1 \leq u_1), \Pr(U_2 \leq u_2)\} \quad (17.1.13)$$

Thus the bounds are

$$\max\{\Pr(U_1 \leq u_1) + \Pr(U_2 \leq u_2) - 1, 0\} \leq C(u_1, u_2) \leq \min\{\Pr(U_1 \leq u_1), \Pr(U_2 \leq u_2)\} \quad (17.1.14)$$

or alternatively, using the fact that a copula has uniform marginals,

$$\max\{u_1 + u_2 - 1, 0\} \leq C(u_1, u_2) \leq \min\{u_1, u_2\} \quad (17.1.15)$$

Using induction, this can be extended to multivariate copulae by

$$\max\{u_1 + \dots + u_d - d + 1, 0\} \leq C(u_1, \dots, u_d) \leq \min\{u_1, \dots, u_d\} \quad (17.1.16)$$

#### Comonotonicity

The Fréchet–Hoeffding upper bound is a valid copula satisfying the required properties, e.g. it satisfies the property

$$\overline{C}(1, \dots, 1, u, 1, \dots, 1) = \min\{1, \dots, 1, u, 1, \dots, 1\} \quad (17.1.17)$$

$$= u \quad (17.1.18)$$

Random variables with the Fréchet–Hoeffding upper bound as its associated copula are said to be comonotonic. This represents a form of perfect positive dependence. To generate a sample from this copula, we simply generate  $U \sim \text{Uniform}(0, 1)$ , and then take  $\mathbf{U} = (U, \dots, U)$  as

the sample. That is, all the variates are equal. To explain why, consider the bivariate case for simplicity, and then derive the conditional CDF as

$$\Pr(U_2 \leq u_2 | U_1 = u_1) = \frac{\frac{\partial}{\partial u_1} \min\{u_1, u_2\}}{1} \quad (17.1.19)$$

$$= \mathbb{I}_{\{u_2 > u_1\}} \quad (17.1.20)$$

Hence to generate a sample, we first generate  $U_1 \sim \text{Uniform}(0, 1)$ , and then generate  $U_2$  conditional on this  $U_1$  using the conditional CDF with the inverse transform sampling method. It follows that with  $\mathbb{I}_{\{u_2 > u_1\}}$  as the conditional CDF, we will have  $U_2 = U_1$ .

So generally, for an arbitrary random vector  $\mathbf{X}$ , an equivalent characterisation for comonotonicity is that we can generate a sample by taking  $(F_1^{-1}(U), \dots, F_d^{-1}(U))$ , where  $U \sim \text{Uniform}(0, 1)$ .

### Countermonotonicity

For a bivariate copula, the Fréchet–Hoeffding lower bound is a valid copula. To generate a sample from this copula, we derive the conditional CDF as

$$\Pr(U_2 \leq u_2 | U_1 = u_1) = \frac{\frac{\partial}{\partial u_1} \max\{u_1 + u_2 - 1, 0\}}{1} \quad (17.1.21)$$

$$= \mathbb{I}_{\{u_1 + u_2 > 1\}} \quad (17.1.22)$$

Hence to generate a sample, we first generate  $U_1 \sim \text{Uniform}(0, 1)$ , and then generate  $U_2$  conditional on this  $U_1$  using the conditional CDF with the inverse transform sampling method. It follows that with  $\mathbb{I}_{\{u_1 + u_2 > 1\}} = \mathbb{I}_{\{u_2 > 1 - u_1\}}$  as the conditional CDF, we will have  $U_2 = 1 - U_1$ . Thus,  $U_1$  and  $U_2$  move completely opposite to each other, representing a form of perfect negative dependence. Bivariate random variables are said to be countermonotonic if their associated copula is the Fréchet–Hoeffding lower bound. We can also generate a sample from a countermonotonic distribution by  $(F_1^{-1}(U), F_2^{-1}(1 - U))$ , where  $U \sim \text{Uniform}(0, 1)$ .

For the case of higher dimensions  $d \geq 3$  however, the Fréchet–Hoeffding lower bound is no longer a valid copula, because it does not satisfy the  $d$ -non-decreasing property. To demonstrate with  $d = 3$ , we use

$$\underline{C}(u_1, u_2, u_3) = \max\{u_1 + u_2 + u_3 - 2, 0\} \quad (17.1.23)$$

and the expression for the trivariate  $C$ -volume of the rectangle bounded between  $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$  and  $(1, 1, 1)$  to get

$$\begin{aligned} \underline{C}(1, 1, 1) - \underline{C}\left(1, 1, \frac{1}{2}\right) - \underline{C}\left(1, \frac{1}{2}, 1\right) - \underline{C}\left(\frac{1}{2}, 1, 1\right) \\ + \underline{C}\left(1, \frac{1}{2}, \frac{1}{2}\right) + \underline{C}\left(\frac{1}{2}, \frac{1}{2}, 1\right) + \underline{C}\left(\frac{1}{2}, 1, \frac{1}{2}\right) - \underline{C}\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) \\ = 1 - \frac{1}{2} - \frac{1}{2} - \frac{1}{2} + 0 + 0 + 0 - 0 \end{aligned} \quad (17.1.24)$$

which equals  $-\frac{1}{2}$ , and is thus negative.

### 17.1.3 Copula Density Functions

Consider a bivariate continuous distribution for  $(X, Y)$ , with joint CDF  $F_{XY}(x, y)$  and marginal CDFs  $F_X(x)$ ,  $F_Y(y)$ . The joint CDF can be written in terms of the copula  $C(u, v)$  as

$$F_{XY}(x, y) = \Pr(X \leq x, Y \leq y) \quad (17.1.25)$$

$$= \Pr(F_X(X) \leq F_X(x), F_Y(Y) \leq F_Y(y)) \quad (17.1.26)$$

$$= \Pr(U \leq F_X(x), V \leq F_Y(y)) \quad (17.1.27)$$

$$= C(F_X(x), F_Y(y)) \quad (17.1.28)$$

Hence we can obtain the joint PDF  $f_{XY}(x, y)$  via

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \quad (17.1.29)$$

$$= \frac{\partial^2 C(F_X(x), F_Y(y))}{\partial x \partial y} \quad (17.1.30)$$

$$= \frac{\partial}{\partial x} \left( \frac{\partial C(F_X(x), F_Y(y))}{\partial y} \right) \quad (17.1.31)$$

This can be differentiated using the chain rule:

$$f_{XY}(x, y) = \frac{\partial}{\partial x} \left( \frac{\partial C(F_X(x), v)}{\partial v} \cdot \frac{\partial F_Y(y)}{\partial y} \right) \quad (17.1.32)$$

$$= \frac{\partial}{\partial x} \left( \frac{\partial C(F_X(x), v)}{\partial v} \right) f_Y(y) \quad (17.1.33)$$

$$= \frac{\partial^2 C(u, v)}{\partial u \partial v} \frac{\partial F_X(x)}{\partial x} f_Y(y) \quad (17.1.34)$$

$$= c(u, v) f_X(x) f_Y(y) \quad (17.1.35)$$

where  $c(u, v) = \frac{\partial C(u, v)}{\partial u \partial v}$  is known as the copula density function. Thus

$$c(u, v) = \frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} \quad (17.1.36)$$

$$= \frac{f_{XY}(F_X^{-1}(u), F_Y^{-1}(v))}{f_X(F_X^{-1}(u)) f_Y(F_Y^{-1}(v))} \quad (17.1.37)$$

and

$$\frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} = c(u, v) \quad (17.1.38)$$

$$= c(F_X(x), F_Y(y)) \quad (17.1.39)$$

This can be generalised to a  $d$ -variate distribution:

$$\frac{f_{X_1, \dots, X_d}(x_1, \dots, x_d)}{f_{X_1}(x_1) \times \dots \times f_{X_d}(x_d)} = c(u_1, \dots, u_d) \quad (17.1.40)$$

$$= c(F_{X_1}(x_1), \dots, F_{X_d}(x_d)) \quad (17.1.41)$$

For example, if  $c$ ,  $f_1, \dots, f_d$ ,  $F_1, \dots, F_d$  are known, then the joint PDF can be worked out by

$$f_{X_1, \dots, X_d}(x_1, \dots, x_d) = f_{X_1}(x_1) \times \dots \times f_{X_d}(x_d) c(F_{X_1}(x_1), \dots, F_{X_d}(x_d)) \quad (17.1.42)$$

Or if  $f_1, \dots, f_d$ ,  $F_1^{-1}, \dots, F_d^{-1}$  and their joint PDF are known, then

$$c(u_1, \dots, u_d) = \frac{f_{X_1, \dots, X_d}(F_{X_1}^{-1}(u_1), \dots, F_{X_d}^{-1}(u_d))}{f_{X_1}(F_{X_1}^{-1}(u_1)) \times \dots \times f_{X_d}(F_{X_d}^{-1}(u_d))} \quad (17.1.43)$$

### 17.1.4 Maximum Likelihood Copulae Fitting

Suppose we are given some multivariate data  $\mathbf{X} \in \mathbb{R}^{n \times d}$  with  $n$  i.i.d. observations in  $d$  dimensions. It may be reasonably straightforward to fit parametric univariate distributions to each of the marginal datasets. However, if there is dependence between dimensions, it may be much more difficult to fit a parametric multivariate distribution to the whole dataset (as the univariate family may not have a natural multivariate extension). In this case, we can introduce a parametric copula  $C(u_1, \dots, u_d; \theta_C)$  parametrised in  $\theta_C$ , which may be fit against the dependence structure in the data. Denote each of the univariate likelihoods (for either discrete or continuous data) by

$$\mathcal{L}_1(\theta_1 | X_{1,1}, \dots, X_{n,1}) = p_1(X_{1,1}, \dots, X_{n,1} | \theta_1) \quad (17.1.44)$$

$$\vdots \quad (17.1.45)$$

$$\mathcal{L}_d(\theta_d | X_{1,d}, \dots, X_{n,d}) = p_d(X_{1,d}, \dots, X_{n,d} | \theta_d) \quad (17.1.46)$$

with corresponding parametric CDFs  $F_1(x_1; \theta_1), \dots, F_d(x_d; \theta_d)$ . We can obtain the parametric copula density by

$$c(u_1, \dots, u_d; \theta_C) = \frac{\partial^d C(u_1, \dots, u_d; \theta_C)}{\partial u_1 \dots \partial u_d} \quad (17.1.47)$$

The likelihood of the parameters  $(\theta_C, \theta_1, \dots, \theta_d)$  given data  $\mathbf{X}$  in terms of the copula density is then

$$\mathcal{L}(\theta_C, \theta_1, \dots, \theta_d | \mathbf{X}) = \prod_{i=1}^n c(F_1(X_{i,1}; \theta_1), \dots, F_d(X_{i,d}; \theta_d); \theta_C) \quad (17.1.48)$$

and the log-likelihood is

$$\log \mathcal{L}(\theta_C, \theta_1, \dots, \theta_d | \mathbf{X}) = \sum_{i=1}^n \log c(F_1(X_{i,1}; \theta_1), \dots, F_d(X_{i,d}; \theta_d); \theta_C) \quad (17.1.49)$$

The maximum likelihood estimate for the parameters  $(\theta_C, \theta_1, \dots, \theta_d)$  is

$$(\hat{\theta}_C, \hat{\theta}_1, \dots, \hat{\theta}_d) = \underset{\theta_C, \theta_1, \dots, \theta_d}{\operatorname{argmin}} \{-\log \mathcal{L}(\theta_C, \theta_1, \dots, \theta_d | \mathbf{X})\} \quad (17.1.50)$$

A computationally easier approach might be to first find the maximum likelihood parameters dimension-wise:

$$\hat{\theta}_1 = \underset{\theta_1}{\operatorname{argmin}} \{-\log \mathcal{L}_1(\theta_1 | X_{1,1}, \dots, X_{n,1})\} \quad (17.1.51)$$

$$\vdots \quad (17.1.52)$$

$$\hat{\theta}_d = \underset{\theta_d}{\operatorname{argmin}} \{-\log \mathcal{L}_d(\theta_d | X_{1,d}, \dots, X_{n,d})\} \quad (17.1.53)$$

and then fix these in the likelihood maximisation with respect to  $\theta_C$ :

$$\hat{\theta}_C = \underset{\theta_C}{\operatorname{argmin}} \left\{ -\log \mathcal{L}(\theta_C, \hat{\theta}_1, \dots, \hat{\theta}_d | \mathbf{X}) \right\} \quad (17.1.54)$$

or use  $\hat{\theta}_1, \dots, \hat{\theta}_d$  as initialisation points in an iterative optimisation algorithm.

## Semi-Parametric Copulae Fitting

If one does not wish to model the marginal distributions, and only the dependence structure of the copula, a semi-parametric maximum likelihood approach can be considered. The negative log-likelihood to minimise is now

$$\widehat{\theta}_C = \operatorname{argmin}_{\theta_C} \left\{ - \sum_{i=1}^n \log c \left( \widehat{F}_1(X_{i,1}), \dots, \widehat{F}_d(X_{i,d}); \theta_C \right) \right\} \quad (17.1.55)$$

where  $\widehat{F}_1, \dots, \widehat{F}_d$  are now non-parametric estimates for each of the marginal cumulative distributions (such as using the empirical distribution function).

### 17.1.5 Gaussian Copula

The Gaussian copula is a parametric family of copulae, with parameter being the correlation matrix  $R \in [-1, 1]^{d \times d}$ , with all diagonals equal to 1 and  $R \succeq 0$ . The copula distribution is defined by

$$C(u_1, \dots, u_d; R) = \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)) \quad (17.1.56)$$

where  $\Phi_R$  is the CDF of the zero-mean multivariate Gaussian with covariance matrix  $R$ , and  $\Phi^{-1}(\cdot)$  is the inverse univariate Gaussian CDF. Since  $\mathcal{N}(\mathbf{0}, R)$  has  $\mathcal{N}(0, 1)$  marginals, then  $\Phi^{-1}(U_j) \sim \mathcal{N}(0, 1)$  implies that  $U_j \sim \text{Uniform}(0, 1)$  for  $j = 1, \dots, d$ . This shows that the form of the Gaussian copula above is a valid copula distribution (in variables  $u_1, \dots, u_d$ ).

#### Gaussian Copula Density Function

To derive the form of the density function of the Gaussian copula, we first compute the  $d^{\text{th}}$  order derivative using the chain rule:

$$c(u_1, \dots, u_d; R) = \frac{\partial^d C(u_1, \dots, u_d; R)}{\partial u_1 \dots \partial u_d} \quad (17.1.57)$$

$$= \frac{\partial^d}{\partial u_1 \dots \partial u_d} \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)) \quad (17.1.58)$$

$$= \phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)) \cdot \frac{\partial \Phi^{-1}(u_1)}{\partial u_1} \times \dots \times \frac{\partial \Phi^{-1}(u_d)}{\partial u_d} \quad (17.1.59)$$

$$= \frac{\phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))}{\phi(\Phi^{-1}(u_1)) \dots \phi(\Phi^{-1}(u_d))} \quad (17.1.60)$$

where  $\phi_R$  and  $\phi(\cdot)$  denote Gaussian densities for the respective Gaussian CDFs, and each  $\frac{\partial \Phi^{-1}(u_j)}{\partial u_j} = \frac{1}{\phi(\Phi^{-1}(u_j))}$  comes by application of the inverse function theorem. This form shows that the Gaussian copula density is given by the ratio of a  $\mathcal{N}(\mathbf{0}, R)$  density over the product of  $\mathcal{N}(0, 1)$  densities (in variables  $\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)$ ). Substituting in the expressions for these densities,

$$\begin{aligned} c(u_1, \dots, u_d; R) &= \frac{1}{(2\pi)^{d/2} \det(R)^{1/2}} \exp \left( -\frac{1}{2} [\Phi^{-1}(u_1) \dots \Phi^{-1}(u_d)] R^{-1} \begin{bmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{bmatrix} \right) \\ &\div \frac{\exp(\Phi^{-1}(u_1)^2/2)}{\sqrt{2\pi}} \div \dots \div \frac{\exp(\Phi^{-1}(u_d)^2/2)}{\sqrt{2\pi}} \quad (17.1.61) \end{aligned}$$

Collecting exponents, we have

$$c(u_1, \dots, u_d; R) = \frac{\sqrt{2\pi} \times \dots \times \sqrt{2\pi}}{(2\pi)^{d/2} \det(R)^{1/2}} \\ \times \exp \left( -\frac{1}{2} [\Phi^{-1}(u_1) \ \dots \ \Phi^{-1}(u_d)] R^{-1} \begin{bmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{bmatrix} + \frac{\Phi^{-1}(u_1)^2}{2} + \dots + \frac{\Phi^{-1}(u_d)^2}{2} \right) \quad (17.1.62)$$

Then factorising the exponent yields:

$$c(u_1, \dots, u_d; R) = \frac{1}{\det(R)^{1/2}} \exp \left( -\frac{1}{2} [\Phi^{-1}(u_1) \ \dots \ \Phi^{-1}(u_d)] (R^{-1} - I) \begin{bmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{bmatrix} \right) \quad (17.1.63)$$

### Gaussian Copula Estimation

One approach to estimate the correlation parameters of a Gaussian copula is to take the Kendall correlation between each pair of variables, denoted  $\hat{\tau}_{jk}$  for variables  $j$  and  $k$ . Then using the relationship between the correlation and the Kendall correlation for a bivariate Gaussian, we estimate the Gaussian copula correlations as

$$\hat{\rho}_{jk} = \sin \left( \frac{\pi}{2} \hat{\tau}_{jk} \right) \quad (17.1.64)$$

#### 17.1.6 Archimedean Copulae

Archimedean copulae are a class of copula that are formed using a ‘generator function’  $\psi : [0, 1] \times \Theta \rightarrow [0, \infty)$  where  $\Theta$  is a parameter space for the generator function. Among some other regularity conditions (mostly related to smoothness and monotonicity of its inverse  $\psi^{-1}$ ), the generator function  $\psi(t; \theta)$  must satisfy the properties of being continuous and convex in  $t$  for all  $\theta \in \Theta$ . Moreover, it must strictly decreasing in  $t$  such that  $\psi(1; \theta) = 0$ . Then from the generator function, the copula is given by

$$C(u_1, \dots, u_d; \theta) = \psi^\dagger(\psi(u_1; \theta) + \dots + \psi(u_d; \theta); \theta) \quad (17.1.65)$$

where  $\psi^\dagger : [0, \infty) \times \Theta \rightarrow [0, 1]$  denotes a generalised inverse, defined by

$$\psi^\dagger(s; \theta) = \begin{cases} \psi^{-1}(s; \theta), & 0 \leq s \leq \psi(0; \theta) \\ 0, & \psi(0; \theta) < s < \infty \end{cases} \quad (17.1.66)$$

In other words, this generalised inverse restricts the range of  $\psi^{-1}$  so that it maps to  $[0, 1]$ . Due to the strictly decreasing property of the generator function, this generalised inverse can alternatively be written as  $\psi^\dagger(s; \theta) = \max\{0, \psi^{-1}(s; \theta)\}$ .

### Independence Copula

The independence copula is an Archimedean copula with the parameterless generator function  $\psi(t; \theta) = -\log t$ . Since  $\psi(0; \theta) = \infty$ , then the generalised inverse is the same as the inverse  $\psi^{-1}(s; \theta) = e^{-t}$ . The form of the Copula is then

$$C(u_1, \dots, u_d; \theta) = \exp(-(-\log u_1 - \dots - \log u_d)) \quad (17.1.67)$$

$$= u_1 \times \dots \times u_d \quad (17.1.68)$$

which is the CDF of  $d$  i.i.d. Uniform  $(0, 1)$  random variables. We can also say that that random variables are mutually independent if and only if their associated copula is the independence copula.

## Frank Copula

The Frank copula is an Archimedean copula with generator function

$$\psi(t, \theta) = -\log \left( \frac{e^{-\theta t} - 1}{e^{-\theta} - 1} \right) \quad (17.1.69)$$

and parameter space  $\theta \in \mathbb{R} \setminus \{0\}$  since the denominator inside the log is zero with  $\theta = 0$ . However, note that if we take the limit as  $\theta \rightarrow 0$  of the fraction inside the log using L'Hôpital's rule:

$$\lim_{\theta \rightarrow 0} \frac{e^{-\theta t} - 1}{e^{-\theta} - 1} = \lim_{\theta \rightarrow 0} \frac{-te^{-\theta t}}{-e^{-\theta}} \quad (17.1.70)$$

$$= t \quad (17.1.71)$$

Hence the limit of the generator function is

$$\lim_{\theta \rightarrow 0} \psi(t, \theta) = -\log t \quad (17.1.72)$$

which is the same as the generator function for the independence copula. So the generalised Frank copula allows for  $\theta \in \mathbb{R}$  and simply takes on the independence copula when  $\theta = 0$ . To determine the inverse generator function, we rearrange

$$s = -\log \left( \frac{e^{-\theta t} - 1}{e^{-\theta} - 1} \right) \quad (17.1.73)$$

$$e^{-s} (e^{-\theta} - 1) = e^{-\theta t} - 1 \quad (17.1.74)$$

$$\ln [e^{-s} (e^{-\theta} - 1) + 1] = -\theta t \quad (17.1.75)$$

$$t = -\frac{1}{\theta} \ln [e^{-s} (e^{-\theta} - 1) + 1] \quad (17.1.76)$$

Thus  $\psi^{-1}(s; \theta) = -\frac{1}{\theta} \ln [e^{-s} (e^{-\theta} - 1) + 1]$  and since  $\psi(0; \theta) = \infty$  like with the independence copula, the generalised inverse generator function is the same as the inverse generator function. The bivariate Frank copula has the form

$$C(u, v; \theta) = \psi^{-1}(\psi(u; \theta) + \psi(v; \theta); \theta) \quad (17.1.77)$$

$$= -\frac{1}{\theta} \ln \left[ 1 + \exp(\psi(u; \theta) + \psi(v; \theta)) (e^{-\theta} - 1) \right] \quad (17.1.78)$$

$$= -\frac{1}{\theta} \ln \left[ 1 + \exp \left[ \ln \left( \frac{e^{-\theta u} - 1}{e^{-\theta} - 1} \cdot \frac{e^{-\theta v} - 1}{e^{-\theta} - 1} \right) \right] (e^{-\theta} - 1) \right] \quad (17.1.79)$$

$$= -\frac{1}{\theta} \ln \left[ 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right] \quad (17.1.80)$$

Several properties of copulas can be verified for this form. Firstly,

$$C(0, 0; \theta) = -\frac{1}{\theta} \ln \left[ 1 + \frac{(1 - 1)(1 - 1)}{e^{-\theta} - 1} \right] \quad (17.1.81)$$

$$= 0 \quad (17.1.82)$$

Then,

$$C(u, 1; \theta) = -\frac{1}{\theta} \ln \left[ 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta} - 1)}{e^{-\theta} - 1} \right] \quad (17.1.83)$$

$$= -\frac{1}{\theta} \ln \left( 1 + e^{-\theta u} - 1 \right) \quad (17.1.84)$$

$$= -\frac{-\theta u}{\theta} \quad (17.1.85)$$

$$= u \quad (17.1.86)$$

and similarly,  $C(1, v; \theta) = v$ . This also confirms that  $C(1, 1; \theta) = 1$ .

## 17.2 Heavy-Tailed Distributions [61]

### 17.2.1 Long-Tailed Distributions

### 17.2.2 Subexponential Distributions

### 17.2.3 States of Randomness

## 17.3 Stochastic Orders

### 17.3.1 First-Order Stochastic Dominance

Stochastic dominance defines a way in which one random variable can be ‘bigger’ than another, and provides an ordering between random variables. For random variables  $X$  and  $Y$ , we say that  $Y$  first-order stochastically dominates  $X$ , and write  $X \preceq_{st} Y$  if

$$\Pr(Y > t) \geq \Pr(X > t) \quad (17.3.1)$$

for all  $t \in \mathbb{R}$ . Note that this can be alternatively written in terms of the CDFs by

$$1 - \Pr(Y \leq t) \geq 1 - \Pr(X \leq t) \quad (17.3.2)$$

$$\Pr(Y \leq t) \leq \Pr(X \leq t) \quad (17.3.3)$$

$$F_Y(t) \leq F_X(t) \quad (17.3.4)$$

Stochastic dominance is however only a *partial ordering* between random variables, i.e. it is possible for two arbitrary random variables  $X$  and  $Y$  that neither stochastically dominates the other. If both  $X \preceq_{st} Y$  and  $Y \preceq_{st} X$ , then thus would imply their CDFs are equal, and thus  $X$  and  $Y$  are equal in distribution, for which we can denote by  $X \stackrel{st}{=} Y$ .

Another way to interpret  $X \preceq_{st} Y$  is by saying that a decision maker with non-decreasing utility function  $u(\cdot)$  of wealth will prefer a gamble with payoff  $Y$  over a gamble with payoff  $X$ . We formalise this with the following theorem.

**Theorem 17.1.** *The following are each equivalent conditions for  $X \preceq_{st} Y$ :*

- $\Pr(Y > t) \geq \Pr(X > t)$  for all  $t \in \mathbb{R}$ .
- $\mathbb{E}[u(X)] \leq \mathbb{E}[u(Y)]$  for any weakly increasing (utility) function  $u(\cdot)$ , whenever the expectations exist.

*Proof.* We first begin with the CDF condition for stochastic dominance:

$$\Pr(Y > t) \geq \Pr(X > t) \quad (17.3.5)$$

$$\Pr(Y \leq t) \leq \Pr(X \leq t) \quad (17.3.6)$$

for all  $t \in \mathbb{R}$ . Since  $u(\cdot)$  is weakly increasing, then this implies

$$\Pr(u(Y) > t) \geq \Pr(u(X) > t) \quad (17.3.7)$$

$$\Pr(u(Y) \leq t) \leq \Pr(u(X) \leq t) \quad (17.3.8)$$

for all  $t \in \mathbb{R}$ . Denote for convenience  $U_Y := u(Y)$  and  $U_X := u(X)$ . The conditions above imply

$$\int_0^\infty \Pr(U_Y > t) dt - \int_{-\infty}^0 \Pr(U_Y \leq t) dt \geq \int_0^\infty \Pr(U_X > t) dt - \int_{-\infty}^0 \Pr(U_X \leq t) dt \quad (17.3.9)$$

$$\int_0^\infty (1 - F_{U_Y}(t)) dt - \int_{-\infty}^0 F_{U_Y}(t) dt \geq \int_0^\infty (1 - F_{U_X}(t)) dt - \int_{-\infty}^0 F_{U_X}(t) dt \quad (17.3.10)$$

which is a characterisation on each side of the expectations  $\mathbb{E}[U_Y]$  and  $\mathbb{E}[U_X]$  respectively using the CDFs. Hence

$$\mathbb{E}[u(Y)] \geq \mathbb{E}[u(X)] \quad (17.3.11)$$

To show the converse, suppose there exists some  $t^*$  such that  $\Pr(X > t^*) > \Pr(Y > t^*)$ . We demonstrate that we can find a weakly increasing utility function that results in a contradiction. Define a utility function as the indicator  $u(t) = \mathbb{I}_{\{t>t^*\}}$ , and denote for convenience  $U_Y^* := u(Y)$  and  $U_X^* := u(X)$ . Their expectations can be simply evaluated as

$$\mathbb{E}[U_X^*] = \Pr(X > t^*) \quad (17.3.12)$$

$$\mathbb{E}[U_Y^*] = \Pr(Y > t^*) \quad (17.3.13)$$

But this results in  $\mathbb{E}[U_X^*] > \mathbb{E}[U_Y^*]$ , which is a contradiction.  $\square$

Stochastic dominance can also be used to upper bound probabilities as follows.

**Corollary 17.1.** *Let  $Y_1, Y_2, Y', Y''$  be independent random variables. Suppose  $Y'$  is first-order stochastically dominated by  $Y_1$  and  $Y_2$  is first-order stochastically dominated by  $Y''$ . Then*

$$\Pr(Y_1 \leq Y_2) \leq \Pr(Y' \leq Y'') \quad (17.3.14)$$

*Proof.* By independence, we can write  $\Pr(Y_1 \leq Y_2)$  as

$$\Pr(Y_1 \leq Y_2) = \int_{-\infty}^\infty \Pr(Y_1 \leq y) dF_{Y_2}(y) \quad (17.3.15)$$

where we have written the probability differential as  $dF_{Y_2}(y)$  rather than  $f_{Y_2}(y) dy$  to explicitly allow for discrete random variables. From stochastic dominance,

$$\Pr(Y_1 \leq Y_2) \leq \int_{-\infty}^\infty \Pr(Y' \leq y) dF_{Y_2}(y) \quad (17.3.16)$$

$$= \int_{-\infty}^\infty F_{Y'}(y) dF_{Y_2}(y) \quad (17.3.17)$$

$$= \mathbb{E}_{Y_2}[F_{Y'}(Y_2)] \quad (17.3.18)$$

Since the CDF  $F_{Y'}(\cdot)$  is weakly increasing, we can treat it as though it were a utility function, which by stochastic dominance yields  $\mathbb{E}_{Y_2}[F_{Y'}(Y_2)] \leq \mathbb{E}_{Y''}[F_{Y'}(Y'')]$ . Hence

$$\Pr(Y_1 \leq Y_2) \leq \mathbb{E}_{Y''}[F_{Y'}(Y'')] \quad (17.3.19)$$

$$= \int_{-\infty}^\infty F_{Y'}(y) dF_{Y''}(y) \quad (17.3.20)$$

$$= \int_{-\infty}^\infty \Pr(Y' \leq y) dF_{Y''}(y) \quad (17.3.21)$$

$$= \Pr(Y' \leq Y'') \quad (17.3.22)$$

$\square$

## Statewise Dominance

Statewise dominance is a stronger notion than first-order stochastic dominance. If  $X(\omega)$  and  $Y(\omega)$  are random variables defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , we say that  $Y$  dominates  $X$  statewise everywhere if  $X(\omega) \leq Y(\omega)$  for all  $\omega \in \Omega$ . A more relaxed definition of statewise dominance is when  $X \leq Y$  holds almost surely, i.e.

$$\Pr(X \leq Y) = 1 \quad (17.3.23)$$

This relaxed definition then allows for the possibility that  $X(\omega) > Y(\omega)$  whenever  $\mathbb{P}(\omega) = 0$ . Statewise dominance then directly implies the definition of first-order stochastic dominance, because for all  $t \in \mathbb{R}$  we have

$$\Pr(Y > t) \geq \Pr(X > t) \quad (17.3.24)$$

Another characterisation of statewise dominance is given by the additive noise representation

$$X = Y + \eta \quad (17.3.25)$$

where the noise  $\eta$  is non-positive. These concepts are related to first-order stochastic dominance as follows.

**Theorem 17.2.** *The following are each equivalent conditions for  $X \preceq_{\text{st}} Y$ :*

- There exist random variables  $\widehat{X}$  and  $\widehat{Y}$  defined on the same probability space, such that  $\widehat{X} \stackrel{\text{st}}{=} X$ ,  $\widehat{Y} \stackrel{\text{st}}{=} Y$ , and

$$\Pr(\widehat{X} \leq \widehat{Y}) = 1 \quad (17.3.26)$$

- There exists a random variable  $\eta$  possibly dependent with  $Y$ , with  $\Pr(\eta \leq 0) = 1$ , such that  $X \stackrel{\text{st}}{=} Y + \eta$ .

*Proof.* As shown above, we readily have that  $\Pr(\widehat{X} \leq \widehat{Y}) = 1$  implies  $\Pr(Y > t) \geq \Pr(X > t)$  for all  $t$ , because of equality in the distributions. For the converse, we show existence of some  $\widehat{X}, \widehat{Y}$  via the following construction. Supposing  $X \stackrel{\text{st}}{\leq} Y$ , then  $F_Y(t) \leq F_X(t)$ , so that

$$F_Y^{-1}(u) \geq F_X^{-1}(u) \quad (17.3.27)$$

for all  $u \in (0, 1)$ . Let  $U \sim \text{Uniform}(0, 1)$ , and make  $\widehat{X} = F_X^{-1}(U)$  and  $\widehat{Y} = F_Y^{-1}(U)$  over the same probability space. Then due to the inverse probability integral transform, we have

$$\widehat{X} \stackrel{\text{st}}{=} X \quad (17.3.28)$$

$$\widehat{Y} \stackrel{\text{st}}{=} Y \quad (17.3.29)$$

Moreover,

$$\Pr(\widehat{X} \leq \widehat{Y}) = \Pr(F_X^{-1}(U) \leq F_Y^{-1}(u)) \quad (17.3.30)$$

$$= 1 \quad (17.3.31)$$

as required.

As for the additive noise condition, existence of the additive noise representation directly implies stochastic dominance because

$$\Pr(X > t) = \Pr(Y + \eta > t) \quad (17.3.32)$$

$$= \Pr(Y > t - \eta) \quad (17.3.33)$$

$$\leq \Pr(Y > t) \quad (17.3.34)$$

since  $\Pr(Y > t)$  is non-increasing in  $t$ , and  $\eta \leq 0$ . For the converse, we show the existence via the following construction. Let  $U \sim \text{Uniform}(0, 1)$  as before, and make  $Y = F_Y^{-1}(U)$  and  $\eta = F_X^{-1}(U) - F_Y^{-1}(U) \leq 0$  over the same probability space. Then it is shown that

$$Y + \eta = F_Y^{-1}(U) + F_X^{-1}(U) - F_Y^{-1}(U) \quad (17.3.35)$$

$$= F_X^{-1}(U) \quad (17.3.36)$$

$$\stackrel{\text{st}}{=} X \quad (17.3.37)$$

□

### 17.3.2 Second-Order Stochastic Dominance

If two random variables  $X$  and  $Y$  cannot be ranked with first-order stochastic dominance, they may still be able to be ranked using second-order stochastic dominance (and hence can be used to ‘break ties’). Loosely speaking, a random variable  $Y$  second-order stochastically dominates another random variable  $X$  if a gamble with payoff  $Y$  is less ‘risky’ than a gamble with payoff  $X$ . To explain further, let the second-order CDF of  $X$  be defined as

$$F_X^{(2)}(x) = \int_{-\infty}^x F_X(t) dt \quad (17.3.38)$$

i.e. in the same way that we integrate the PDF to obtain the CDF, we integrate the CDF to obtain the second-order CDF. Observe a general property of the second-order CDF  $F_X^{(2)}(x)$  is that it is non-decreasing convex, since its derivative is the CDF which is non-negative and non-decreasing. Also,  $F_X^{(2)}(x)$  will be non-negative and non-decreasing just like the CDF. Using integration by parts, the second-order CDF can be shown to be equal to

$$F_X^{(2)}(x) = [F_X(t)t]_{-\infty}^x - \int_{-\infty}^x f_X(t) t dt \quad (17.3.39)$$

$$= x F_X(x) - \int_{-\infty}^{\infty} f_X(t) t \mathbb{I}_{\{t \leq x\}} dt \quad (17.3.40)$$

$$= x \Pr(X \leq x) - \mathbb{E}[X \mathbb{I}_{\{X \leq x\}}] \quad (17.3.41)$$

$$= x \mathbb{E}[\mathbb{I}_{\{X \leq x\}}] - \mathbb{E}[X \mathbb{I}_{\{X \leq x\}}] \quad (17.3.42)$$

$$= \mathbb{E}[\mathbb{I}_{\{X \leq x\}}(x - X)] \quad (17.3.43)$$

$$= \mathbb{E}[\max\{x - X, 0\}] \quad (17.3.44)$$

The quantity  $\max\{x - X, 0\}$  can be thought of as a *shortfall* variable, which gives the difference if  $X$  falls short of  $x$ , and zero otherwise. Hence the second-order CDF  $F_X^{(2)}(x)$  quantifies the average shortfall for a given  $x$ , which is a notion of risk. Then, we say that  $Y$  second-order stochastically dominates  $X$  and write  $X \stackrel{(2)}{\preceq} Y$  if

$$F_Y^{(2)}(t) \leq F_X^{(2)}(t) \quad (17.3.45)$$

for all  $t \in \mathbb{R}$ . Equivalently, we have

$$\mathbb{E}[\max\{t - Y, 0\}] \leq \mathbb{E}[\max\{t - X, 0\}] \quad (17.3.46)$$

for all  $t \in \mathbb{R}$ . Note that if  $X \stackrel{\text{st}}{\preceq} Y$ , then this is a sufficient condition for  $X \stackrel{(2)}{\preceq} Y$ , because

$$F_X^{(2)}(t) - F_Y^{(2)}(t) = \int_{-\infty}^t F_X(s) ds - \int_{-\infty}^t F_Y(s) ds \quad (17.3.47)$$

$$= \int_{-\infty}^t (F_X(s) - F_Y(s)) ds \quad (17.3.48)$$

$$\geq 0 \quad (17.3.49)$$

Another way to equivalently characterise second-order stochastic dominance is in terms of risk aversion. A risk-averse decision maker will have a non-decreasing concave utility function  $u(\cdot)$ , and  $X \stackrel{(2)}{\preceq} Y$  means that  $Y$  has more expected utility than  $X$ , demonstrated as follows.

**Theorem 17.3.** *The following are each equivalent conditions for  $X \stackrel{(2)}{\preceq} Y$ :*

- $F_Y^{(2)}(t) \leq F_X^{(2)}(t)$  for all  $t \in \mathbb{R}$ .
- $\mathbb{E}[u(X)] \leq \mathbb{E}[u(Y)]$  for any non-decreasing concave (utility) function  $u(\cdot)$ , whenever the expectations exist.

*Proof.* Suppose the condition  $\mathbb{E}[u(X)] \leq \mathbb{E}[u(Y)]$  holds. A sketch of the function will reveal that  $-\max\{t - X, 0\}$  is non-decreasing concave in  $x$  for all  $t$ . Hence

$$\mathbb{E}[-\max\{t - X, 0\}] \leq \mathbb{E}[-\max\{t - Y, 0\}] \quad (17.3.50)$$

for all  $t$ , which is the same as  $F_Y^{(2)}(t) \leq F_X^{(2)}(t)$  for all  $t$ , so the second-order CDF condition is implied. For the converse, suppose the condition  $F_Y^{(2)}(t) \leq F_X^{(2)}(t)$  holds, which is the same as  $\mathbb{E}[-\max\{t - X, 0\}] \leq \mathbb{E}[-\max\{t - Y, 0\}]$ . The function  $-\max\{t - x, 0\}$  is kinked, with an upwards linear slope to the left of  $t$ , and horizontal to the right of  $t$ . Given this shape, we reason that any non-decreasing concave function  $u(\cdot)$  can be built up (in the limit, if required) of positive combinations of shifted  $-\max\{t - x, 0\}$  as follows (in the countable case, for simplicity):

$$u(x) = - \sum_i a_i \max\{t_i - x, 0\} \quad (17.3.51)$$

where the  $a_i > 0$ . Therefore the second-order CDF condition implies

$$\mathbb{E}[u(X)] = \mathbb{E}\left[- \sum_i a_i \max\{t_i - X, 0\}\right] \quad (17.3.52)$$

$$= \sum_i a_i \mathbb{E}[-\max\{t_i - X, 0\}] \quad (17.3.53)$$

$$\leq \sum_i a_i \mathbb{E}[-\max\{t_i - Y, 0\}] \quad (17.3.54)$$

$$= \mathbb{E}[u(Y)] \quad (17.3.55)$$

□

By letting  $u(x) = x$ , we see that a necessary condition for  $X \stackrel{(2)}{\preceq} Y$  is that  $\mathbb{E}[X] \leq \mathbb{E}[Y]$ .

### Concave Stochastic Order [178]

The concave stochastic order is closely related to second-order stochastic dominance. We say that  $Y$  dominates  $X$  in concave stochastic order and denote  $X \stackrel{\text{cv}}{\preceq} Y$  if

$$\mathbb{E}[u(X)] \leq \mathbb{E}[u(Y)] \quad (17.3.56)$$

for all concave  $u(\cdot)$ . Note the difference here is that the definition applies to all concave functions, rather than just non-decreasing concave functions. Hence concave stochastic order is

stronger than second-order stochastic dominance, in that  $X \preceq_{\text{cv}} Y$  implies  $X \preceq_{(2)} Y$ . Also, if we defined a notion of increasing concave stochastic order (where  $u(\cdot)$  must be weakly increasing), then this becomes equivalent to second-order stochastic dominance.

Typical concave functions are arch-shaped, so if  $\mathbb{E}[u(Y)]$  is larger, this intuitively suggests that  $Y$  is less likely to take on extreme values as  $X$ . This carries the same notion as  $Y$  being riskier than  $X$ . Also since  $u(x) = x$  and  $u(x) = -x$  are both concave, then we have  $\mathbb{E}[X] \leq \mathbb{E}[Y]$  and  $-\mathbb{E}[X] \leq -\mathbb{E}[Y]$  which must mean that

$$\mathbb{E}[X] = \mathbb{E}[Y] \quad (17.3.57)$$

Therefore, equality in expectation is a necessary condition for concave stochastic order. Since the expectations are equal, then in order for  $Y$  to have a smaller variance than  $X$ , we must require  $\mathbb{E}[X^2] \geq \mathbb{E}[Y^2]$ . Because  $u(x) = -x^2$  is concave, then this is indeed true, and therefore

$$\text{Var}(X) \geq \text{Var}(Y) \quad (17.3.58)$$

The condition  $\mathbb{E}[X] = \mathbb{E}[Y]$  is important in the distinction between concave stochastic order and second-order stochastic dominance, as the following result illustrates.

**Theorem 17.4.**  $X \preceq_{\text{cv}} Y$  if and only if  $X \preceq_{(2)} Y$  and  $\mathbb{E}[X] = \mathbb{E}[Y]$ .

*Proof.* As we have already shown above that  $X \preceq_{\text{cv}} Y$  implies both  $X \preceq_{(2)} Y$  and  $\mathbb{E}[X] = \mathbb{E}[Y]$ , it suffices to show the converse. From second-order stochastic dominance, we know that any non-decreasing concave function can be built from positive combinations of shifted  $-\max\{t - x, 0\}$ . Then to build any general concave function, we need to use negative slopes, i.e. write  $u(\cdot)$  as

$$u(x) = - \sum_i a_i \max\{t_i - x, 0\} - \sum_j b_j (x - c_j) \quad (17.3.59)$$

where the  $a_i > 0$ ,  $b_j > 0$  and the  $t_i$ ,  $c_j$  are at appropriate locations. Then because  $\mathbb{E}[X] = \mathbb{E}[Y]$  by hypothesis, this implies

$$\mathbb{E}[u(X)] = \mathbb{E}\left[- \sum_i a_i \max\{t_i - X, 0\} - \sum_j b_j (X - c_j)\right] \quad (17.3.60)$$

$$= \sum_i a_i \mathbb{E}[-\max\{t_i - X, 0\}] - \sum_j b_j (\mathbb{E}[X] - c_j) \quad (17.3.61)$$

$$\leq \sum_i a_i \mathbb{E}[-\max\{t_i - Y, 0\}] - \sum_j b_j (\mathbb{E}[X] - c_j) \quad (17.3.62)$$

$$= \sum_i a_i \mathbb{E}[-\max\{t_i - Y, 0\}] - \sum_j b_j (\mathbb{E}[Y] - c_j) \quad (17.3.63)$$

$$= \mathbb{E}[u(Y)] \quad (17.3.64)$$

□

## Convex Stochastic Order

We say that  $Y$  dominates  $X$  in convex stochastic order and denote  $X \preceq_{\text{cx}} Y$  if

$$\mathbb{E}[u(X)] \leq \mathbb{E}[u(Y)] \quad (17.3.65)$$

for all convex  $u(\cdot)$ . Since if  $u(\cdot)$  is convex then  $-u(\cdot)$  is concave, then convex stochastic order is equivalent to concave stochastic order in the sense that  $X \preceq_{\text{cv}} Y$  if and only if  $Y \preceq_{\text{cv}} X$ . Analogous properties can then be stated for the convex stochastic order.

Typical convex functions are U-shaped, so if  $\mathbb{E}[u(Y)]$  is larger, then this intuitively suggests that  $Y$  is more likely to take on extreme values, carrying the notion that  $Y$  is riskier.

### Mean-Preserving Spreads [128, 133]

The idea of a mean-preserving spread can equivalently characterise the concave stochastic order. Loosely speaking, a mean-preserving spread of a distribution or random variable ‘spreads’ the probability towards the tails, while preserving the mean. To make this more precise, define two random variables  $X, Y$  with probability densities  $f_X(\cdot), f_Y(\cdot)$  respectively (analogous definitions apply if dealing with probability masses). Random variable  $X$  is said to be a mean-preserving spread of  $Y$  if  $\mathbb{E}[X] = \mathbb{E}[Y]$ , and there exists two points  $\underline{t}, \bar{t} \in \mathbb{R}$  such that:

- $f_X(\cdot)$  assigns no more probability than  $f_Y(\cdot)$  to every subinterval in  $(\underline{t}, \bar{t})$ .
- $f_X(\cdot)$  assigns at least as much probability as  $f_Y(\cdot)$  to every subinterval in  $(-\infty, \underline{t}) \cup (\bar{t}, \infty)$ .

Another related notion is via additive noise, which also has the effect of spreading out the distribution. Suppose the relation between  $X$  and  $Y$  can be expressed as

$$X =_{\text{st}} Y + \varepsilon \quad (17.3.66)$$

where  $\varepsilon$  is a random variable such that  $\mathbb{E}[\varepsilon|Y] = 0$ . Note that this implicitly requires  $\mathbb{E}[Y] = \mathbb{E}[X]$  because

$$\mathbb{E}[Y] = \mathbb{E}[X] + \mathbb{E}[\varepsilon] \quad (17.3.67)$$

$$= \mathbb{E}[X] + \mathbb{E}[\mathbb{E}[\varepsilon|X]] \quad (17.3.68)$$

$$= \mathbb{E}[X] \quad (17.3.69)$$

Equipped with these concepts, we can formally establish the following equivalent characterisations of the concave stochastic order.

**Theorem 17.5.** *The following are each equivalent conditions for  $X \preceq_{\text{cv}} Y$ :*

- $X$  is a mean-preserving spread of  $Y$ .
- There exists a random variable  $\varepsilon$  with  $\mathbb{E}[\varepsilon|Y] = 0$  such that  $X =_{\text{st}} Y + \varepsilon$ .

*Proof.* Firstly supposing additive noise, then

$$\mathbb{E}[u(Y)] = \mathbb{E}[u(X + \varepsilon)] \quad (17.3.70)$$

$$= \mathbb{E}[\mathbb{E}[u(X + \varepsilon)|X]] \quad (17.3.71)$$

When  $u(\cdot)$  is concave, applying Jensen’s inequality on the inner expectation above yields

$$\mathbb{E}[u(Y)] \leq \mathbb{E}[u(\mathbb{E}[X + \varepsilon|X])] \quad (17.3.72)$$

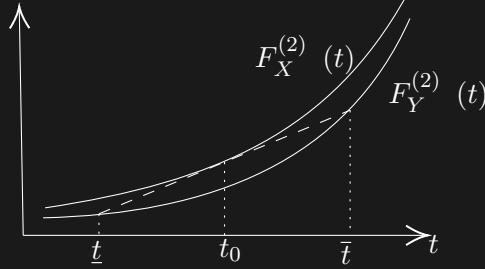
$$= \mathbb{E}[u(X + \mathbb{E}[\varepsilon|X])] \quad (17.3.73)$$

$$= \mathbb{E}[u(X)] \quad (17.3.74)$$

which is the definition of  $X \preceq_{\text{cv}} Y$ .

We now show that  $X \preceq_{\text{cv}} Y$  implies a mean-preserving spread, via the following construction.

Because  $X \preceq_{\text{cv}} Y$ , the second-order CDFs satisfy  $F_Y^{(2)}(t) \leq F_X^{(2)}(t)$ . Also recall that the second-order CDFs are non-decreasing convex. Pick an arbitrary location  $t_0 \in \mathbb{R}$ , and draw the line tangent to  $F_X^{(2)}(t_0)$  until it intersects  $F_Y^{(2)}(\cdot)$  at the locations  $t$  and  $t'$ , with  $t \leq t'$ .



Call  $H(t)$  the function that is  $F_Y^{(2)}(t)$ , except linearly interpolated between  $\underline{t}$  and  $\bar{t}$ . This is formally specified as follows:

$$H(t) = \begin{cases} F_Y^{(2)}(t), & t \leq \underline{t} \\ F_Y^{(2)}(\underline{t}) + \frac{F_Y^{(2)}(\bar{t}) - F_Y^{(2)}(\underline{t})}{\bar{t} - \underline{t}}(t - \underline{t}), & t \in (\underline{t}, \bar{t}) \\ F_Y^{(2)}(\bar{t}), & t \geq \bar{t} \end{cases} \quad (17.3.75)$$

Due to convexity, we have  $F_Y^{(2)}(t) \leq H(t) \leq F_X^{(2)}(t)$  for all  $t$ . The derivative  $H'(t)$  is also a valid CDF, since it is non-decreasing, and satisfies  $\lim_{t \rightarrow -\infty} H'(t) = \lim_{t \rightarrow -\infty} F_Y(t) = 0$  and  $\lim_{t \rightarrow \infty} H'(t) = \lim_{t \rightarrow \infty} F_Y(t) = 1$ . Let  $Z$  be a random variable which has CDF  $H'(t)$ . Since  $H(t)$  is linear over  $(\underline{t}, \bar{t})$ , then  $H'(t)$  will be flat over the same interval. Hence by construction  $Z$  is a ‘spread’ of  $Y$ , since it has redistributed the probability to the locations  $\underline{t}, \bar{t}$ . We are left to show is that mean-preserving. Using the expression for the expectation in terms of the CDF, if  $\mathbb{E}[Z] = \mathbb{E}[Y]$ , then this is the same as

$$\int_0^\infty (1 - H'(t)) dt - \int_{-\infty}^0 H(t) dt = \int_0^\infty (1 - F_Y(t)) dt - \int_{-\infty}^0 F_Y(t) dt \quad (17.3.76)$$

Rearranging gives

$$\begin{aligned} \int_0^\infty (F_Y(t) - H'(t)) dt + \int_{-\infty}^0 (F_Y(t) - H(t)) dt &= \int_{-\infty}^\infty (F_Y(t) - H'(t)) dt \\ &= 0 \end{aligned} \quad (17.3.77)$$

Hence their expectations are equal if  $\lim_{t \rightarrow \infty} (F_Y^{(2)}(t) - H(t)) = 0$ , which is true by construction. Now consider constructing a successive sequence of mean-preserving spreads  $H_1(t), H_2(t)$ , etc., which defines a sequence of random variables  $Z_1, Z_2$ , etc. If this is done appropriately, then in the limit  $H_\infty(t) \equiv F_X^{(2)}(t)$ . Their derivatives are also identical, meaning  $Z_\infty \stackrel{\text{st}}{=} X$ . Observe that if  $Z_{n+1}$  is a mean-preserving spread of  $Z_n$  and moreover  $Z_n$  is mean-preserving spread of  $Y$ , then  $Z_{n+1}$  is also a mean-preserving spread of  $Y$ . Therefore,  $X$  is a mean-preserving spread of  $Y$ .

Finally, we show that our construction of a mean-preserving spread implies an additive noise representation. We seek a random variable  $\varepsilon_1$  with  $\mathbb{E}[\varepsilon_1|Y] = 0$  such that  $Z \stackrel{\text{st}}{=} Y + \varepsilon_1$ . Define  $\varepsilon_1$  as being dependent on  $Y$  with the specification

$$\varepsilon_1(y) = \begin{cases} (\underline{t} - y) \mathbb{I}_{\{y \in (\underline{t}, \bar{t})\}}, & \text{w.p. } \frac{\bar{t} - y}{\bar{t} - \underline{t}} \\ (\bar{t} - y) \mathbb{I}_{\{y \in (\underline{t}, \bar{t})\}}, & \text{w.p. } \frac{y - \underline{t}}{\bar{t} - \underline{t}} \end{cases} \quad (17.3.79)$$

Thus  $\varepsilon_1(y) = 0$  when  $y \notin (\underline{t}, \bar{t})$ , and when  $y \in (\underline{t}, \bar{t})$ :

$$\mathbb{E} [\varepsilon_1(Y) | Y = y, Y \in (\underline{t}, \bar{t})] = (\underline{t} - y) \frac{\bar{t} - y}{\bar{t} - \underline{t}} + (\bar{t} - y) \frac{y - \underline{t}}{\bar{t} - \underline{t}} \quad (17.3.80)$$

$$= \frac{(\underline{t} - y)(\bar{t} - y)}{\bar{t} - \underline{t}} - \frac{(\bar{t} - y)(\bar{t} - y)}{\bar{t} - \underline{t}} \quad (17.3.81)$$

$$= 0 \quad (17.3.82)$$

Hence  $\mathbb{E}[\varepsilon_1|Y] = 0$ . Furthermore,  $Y + \varepsilon_1$  takes on a value of either  $\underline{t}$  or  $\bar{t}$  when  $Y \in (\underline{t}, \bar{t})$ , which fits the same characterisation as  $Z$ , since all the probability in  $(\underline{t}, \bar{t})$  has been redistributed to the locations  $\underline{t}, \bar{t}$ . This has also been done in such a way that

$$\mathbb{E}[Y + \varepsilon_1] = \mathbb{E}[\mathbb{E}[Y + \varepsilon_1|Y]] \quad (17.3.83)$$

$$= \mathbb{E}[Y + \mathbb{E}[\varepsilon_1|Y]] \quad (17.3.84)$$

$$= \mathbb{E}[Y] \quad (17.3.85)$$

$$= \mathbb{E}[Z] \quad (17.3.86)$$

which is together enough to show that  $Z = \underset{\text{st}}{\mathbb{E}}[Y + \varepsilon_1]$ . Now consider an appropriate sequence of constructions (following that used in the mean-preserving spread):

$$Z_1 = \underset{\text{st}}{\mathbb{E}}[Y + \varepsilon_1] \quad (17.3.87)$$

$$Z_2 = \underset{\text{st}}{\mathbb{E}}[Y + \varepsilon_2] \quad (17.3.88)$$

$$\vdots \quad (17.3.89)$$

with  $\mathbb{E}[\varepsilon_1|Y] = 0$ ,  $\mathbb{E}[\varepsilon_2|Z_1] = 0$ , etc. in the same manner. Then  $Z_\infty = \underset{\text{st}}{\mathbb{E}}[X]$  and we can express

$$X = \underset{\text{st}}{\mathbb{E}}[Y + \sum_{i=1}^{\infty} \varepsilon_i] \quad (17.3.90)$$

Note that

$$\mathbb{E}[\varepsilon_1 + \varepsilon_2|Y] = \mathbb{E}[\varepsilon_1|Y] + \mathbb{E}[\varepsilon_2|Y] \quad (17.3.91)$$

$$= \mathbb{E}[\mathbb{E}[\varepsilon_2|Z_1, Y]|Y] \quad (17.3.92)$$

$$= \mathbb{E}[\mathbb{E}[\varepsilon_2|Z_1]|Y] \quad (17.3.93)$$

$$= \mathbb{E}[0|Y] \quad (17.3.94)$$

$$= 0 \quad (17.3.95)$$

because  $\varepsilon_2$  and  $Y$  are conditionally independent given  $Z_1$ . Then by induction we can show  $\mathbb{E}[\sum_{i=1}^{\infty} \varepsilon_i|X] = 0$ , so take  $\varepsilon = \sum_{i=1}^{\infty} \varepsilon_i$  and therefore this gives an additive noise representation from a mean-preserving spread.  $\square$

### 17.3.3 Higher-Order Stochastic Dominance [180]

Extending the second-order CDF, we can define the  $k^{\text{th}}$ -order CDF of a random variable  $X$  by

$$F_X^{(k)}(x) = \int_{-\infty}^x F_X^{(k-1)}(t) dt \quad (17.3.96)$$

for  $k = 2, 3, 4, \dots$ , with  $F_X^{(1)}(x) = F_X(x)$  as the usual CDF. This is not to be confused with higher-order distribution functions of a stochastic process. Although  $F_X^{(1)}(x)$  could have jumps

in the function,  $F_X^{(2)}(x)$  will be continuous once we integrate. Also recall that  $F_X^{(2)}(x)$  is non-negative and non-decreasing convex. By recycling the same arguments, we have that  $F_X^{(2)}(x)$  is continuous, non-negative and non-decreasing convex for all  $k \geq 2$ .

A natural definition for higher-order stochastic dominance follows from the  $k^{\text{th}}$ -order CDF. We say that  $Y$  stochastically dominates  $X$  in  $k^{\text{th}}$ -order and denote  $X \preceq_{(k)} Y$  if

$$F_Y^{(k)}(t) \leq F_X^{(k)}(t) \quad (17.3.97)$$

for all  $t \in \mathbb{R}$ . In the same way that first-order stochastic dominance implies second-order stochastic dominance, we have that  $k^{\text{th}}$ -order stochastic dominance implies  $m^{\text{th}}$ -order stochastic dominance for all  $m > k$ .

#### 17.3.4 Multivariate Stochastic Dominance [178]

Notions of multivariate stochastic dominance can extend univariate stochastic dominance. The analogue to first-order stochastic dominance, sometimes referred to the *usual multivariate stochastic order*, is defined as follows. A random vector  $\mathbf{Y}$  is said to stochastically dominate  $\mathbf{X}$  in the usual stochastic order, and denoted  $\mathbf{X} \preceq_{\text{st}} \mathbf{Y}$ , if

$$\Pr(\mathbf{X} \in \mathbb{U}) \leq \Pr(\mathbf{Y} \in \mathbb{U}) \quad (17.3.98)$$

for all upper sets  $\mathbb{U} \subseteq \mathbb{R}^n$ . An upper set can be defined as a set such that if  $t \in \mathbb{U}$ , then  $t' \in \mathbb{U}$  for all  $t' \geq t$  (interpreted as a component-wise inequality). Intuitively speaking,  $\mathbf{Y}$  is at least as likely as  $\mathbf{Y}$  to take on large values (formally defined in terms of upper sets). Analogous to univariate stochastic dominance, multivariate stochastic dominance can be equivalently characterised in terms of expectations of non-decreasing functions.

**Theorem 17.6.** *The following are each equivalent conditions for  $\mathbf{X} \preceq_{\text{st}} \mathbf{Y}$ :*

- $\Pr(\mathbf{X} \in \mathbb{U}) \leq \Pr(\mathbf{Y} \in \mathbb{U})$  upper sets  $\mathbb{U} \subseteq \mathbb{R}^n$ .
- $\mathbb{E}[u(\mathbf{X})] \leq \mathbb{E}[u(\mathbf{Y})]$  for any weakly increasing function  $u(\cdot)$  (interpreted as component-wise increasing), whenever the expectations exist.

*Proof.* Suppose  $\Pr(\mathbf{X} \in \mathbb{U}) \leq \Pr(\mathbf{Y} \in \mathbb{U})$  for all upper sets. Introducing the indicator  $\mathbb{I}_{\{t \in \mathbb{U}\}}$ , we have

$$\mathbb{E}[\mathbb{I}_{\{\mathbf{X} \in \mathbb{U}\}}] \leq \mathbb{E}[\mathbb{I}_{\{\mathbf{Y} \in \mathbb{U}\}}] \quad (17.3.99)$$

for all upper sets. Now any weakly increasing  $u(\cdot)$  can be expressed (in the limit, if need be) as a linear combination of indicators of a sequence of upper sets:

$$u(t) = \sum_i a_i \mathbb{I}_{\{t \in \mathbb{U}_i\}} + b \quad (17.3.100)$$

Therefore it follows that

$$\mathbb{E}[u(\mathbf{X})] = \sum_i a_i \mathbb{E}[\mathbb{I}_{\{\mathbf{X} \in \mathbb{U}_i\}}] + b \quad (17.3.101)$$

$$\leq \sum_i a_i \mathbb{E}[\mathbb{I}_{\{\mathbf{Y} \in \mathbb{U}_i\}}] + b \quad (17.3.102)$$

$$= \mathbb{E}[u(\mathbf{Y})] \quad (17.3.103)$$

For the converse, suppose that  $\mathbb{E}[u(\mathbf{X})] \leq \mathbb{E}[u(\mathbf{Y})]$  for all weakly increasing  $u(\cdot)$ . Let  $\mathbb{U}^*$  be a potential upper set such that  $\Pr(\mathbf{X} \in \mathbb{U}^*) > \Pr(\mathbf{Y} \in \mathbb{U}^*)$ . But then consider the function  $u(t) = \mathbb{I}_{\{t \in \mathbb{U}^*\}}$  which is weakly increasing. Then by hypothesis,

$$\mathbb{E}[\mathbb{I}_{\{\mathbf{X} \in \mathbb{U}^*\}}] \leq \mathbb{E}[\mathbb{I}_{\{\mathbf{Y} \in \mathbb{U}^*\}}] \quad (17.3.104)$$

which is equivalent to  $\Pr(\mathbf{X} \in \mathbb{U}^*) \leq \Pr(\mathbf{Y} \in \mathbb{U}^*)$ . This contradicts the existence of such a  $\mathbb{U}^*$ , therefore it must be that  $\Pr(\mathbf{X} \in \mathbb{U}) \leq \Pr(\mathbf{Y} \in \mathbb{U})$  for all upper sets.  $\square$

### Upper Orthant Dominance [178]

A weaker notion of multivariate stochastic dominance compared to (and is implied by) the usual multivariate stochastic dominance is upper orthant dominance. However it is still a natural extension of first-order stochastic dominance to multivariate distributions. Let  $F_{\mathbf{X}}(\mathbf{x})$  and  $F_{\mathbf{Y}}(\mathbf{y})$  be the CDFs of random vectors  $\mathbf{X}$  and  $\mathbf{Y}$  respectively. Let  $\bar{F}_{\mathbf{X}}(\mathbf{x})$  and  $\bar{F}_{\mathbf{Y}}(\mathbf{y})$  denote the respective multivariate survival functions, i.e.

$$\bar{F}_{\mathbf{X}}(\mathbf{x}) = \Pr(X_1 > x_1, \dots, X_n > x_n) \quad (17.3.105)$$

$$\bar{F}_{\mathbf{Y}}(\mathbf{y}) = \Pr(Y_1 > y_1, \dots, Y_n > y_n) \quad (17.3.106)$$

We say that  $\mathbf{Y}$  stochastically dominates  $\mathbf{X}$  in upper orthant order, and denote  $\mathbf{X} \xrightarrow[\text{uo}]{} \mathbf{Y}$ , if

$$\bar{F}_{\mathbf{X}}(\mathbf{t}) \leq \bar{F}_{\mathbf{Y}}(\mathbf{t}) \quad (17.3.107)$$

for all  $\mathbf{t} \in \mathbb{R}^n$ . This has the same intuitive meaning as the usual stochastic order, that  $\mathbf{Y}$  is as likely as  $\mathbf{X}$  to take on large values (but formally defined this time in terms of upper orthants).

A concept of *lower orthant dominance* also follows. We say that  $\mathbf{Y}$  stochastically dominates  $\mathbf{X}$  in lower orthant order, and denote  $\mathbf{X} \xrightarrow[\text{lo}]{} \mathbf{Y}$ , if

$$F_{\mathbf{Y}}(\mathbf{t}) \leq F_{\mathbf{X}}(\mathbf{t}) \quad (17.3.108)$$

for all  $\mathbf{t} \in \mathbb{R}^n$ . Intuitively,  $\mathbf{Y}$  is no more likely than  $\mathbf{X}$  to take on small values (formally defined in terms of lower orthants).

## 17.4 Portfolio Optimisation

### 17.4.1 Kelly Criterion

Suppose an opportunity to place a bet yields  $b$  net odds (i.e. for every 1 unit of currency wagered, a win results in  $1+b$  gross return, while a loss results in no return). The probability of winning the bet is given by  $p$ . The Kelly criterion investigates the optimal betting amount for this bet. Assume the bettor has a log utility function of wealth, given by

$$u(W) = \log W \quad (17.4.1)$$

The Kelly criterion maximises the expected utility (hence expected log wealth) from an initial wealth  $W_0$  and a wager  $w$ :

$$\mathbb{E}[u(W)] = p \log(W_0 + wb) + (1-p) \log(W_0 - w) \quad (17.4.2)$$

For further simplicity, denote the betting fraction  $f = w/W_0$  so that

$$\mathbb{E}[u(W)] = \frac{p \log(1+fb) + (1-p) \log(1-f)}{W_0} \quad (17.4.3)$$

We find the optimal betting fraction  $f^*$  to maximise expected utility. Note that the expected utility is a concave function. Taking the derivative yields

$$\frac{\partial \mathbb{E}[u(W)]}{\partial f} = \frac{1}{W_0} \left( \frac{pb}{1+fb} + \frac{p-1}{1-f} \right) \quad (17.4.4)$$

Setting this to zero gives

$$\frac{1}{W_0} \left( \frac{pb}{1+f^*b} + \frac{p-1}{1-f^*} \right) = 0 \quad (17.4.5)$$

$$pb(1-f^*) = (1+f^*b)(1-p) \quad (17.4.6)$$

$$f^* = \frac{pb+p-1}{b} \quad (17.4.7)$$

Letting  $q := 1 - p$ , we rewrite this as

$$f^* = \frac{pb-q}{b} \quad (17.4.8)$$

which gives the optimal betting fraction in terms of the probability of winning/losing and the odds.

### 17.4.2 Modern Portfolio Theory

We describe the *Markowitz model* in portfolio optimisation. Let  $r \in \mathbb{R}^n$  be a random vector for the rate of returns for  $n$  assets over a fixed period of time. Let  $x_0 \in \mathbb{R}^n$  denote the vector containing an investor's initial wealth allocated to each of the  $n$  assets. Then let vector  $u \in \mathbb{R}^n$  denote the transacted amounts in each of the assets in the fixed period, so the investor's updated wealth is given by

$$x = x_0 + u \quad (17.4.9)$$

Assume that total wealth is conserved, meaning  $\mathbf{1}^\top x = \mathbf{1}^\top x_0$ , or

$$\mathbf{1}^\top u = 0 \quad (17.4.10)$$

Thus the goal is to choose the optimal  $u$  with regard to several objectives.

#### Mean-Variance Portfolio Optimisation

Suppose the expected returns  $\bar{r} := \mathbb{E}[r]$  as well as covariance of returns  $\Sigma := \text{Cov}(r)$  are known. Then the investor's return at the end of the fixed period is a random variable  $R = r^\top x$  which has mean and variance

$$\mathbb{E}[R] = \bar{r}^\top x \quad (17.4.11)$$

$$\text{Var}(R) = x^\top \Sigma x \quad (17.4.12)$$

respectively. Maximising the return  $\mathbb{E}[R]$  and minimising the risk  $\text{Var}(R)$  are both desirable, assuming investors are risk-averse for the latter. To formulate a single objective problem, we could choose  $u$  to maximise the return under a prescribed maximum admissible level  $\sigma^2$  of risk, under the budgetary constraint:

$$\begin{aligned} \bar{R}^*(\sigma) = \max_u & \quad \bar{r}^\top (x_0 + u) \\ \text{s.t.} & \quad x = x_0 + u \\ & \quad x^\top \Sigma x \leq \sigma^2 \\ & \quad \mathbf{1}^\top u = 0 \\ & \quad \mathbb{I}_{ss} x \geq \mathbf{0} \end{aligned} \quad (17.4.13)$$

where  $\mathbb{I}_{ss}$  is a flag that denotes whether short-selling is allowed (and hence, whether the investor is allowed to hold a negative amount of wealth in an asset). This optimisation problem has a linear objective and quadratic constraints, thus it can be cast as a second-order cone program, which is convex.

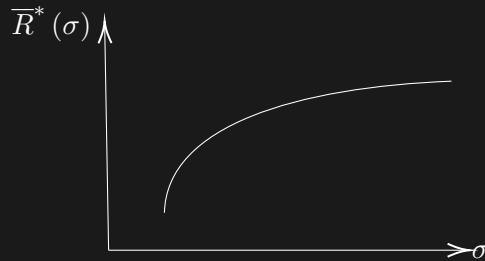
An alternative formulation is to minimise the risk, subject to a lower bound  $\mu$  on the expected return:

$$\begin{aligned} \sigma^*(\mu) = \min_u & (x_0 + u)^\top \Sigma (x_0 + u) \\ \text{s.t.} & x = x_0 + u \\ & \bar{r}^\top x \geq \mu \\ & \mathbf{1}^\top u = 0 \\ & \mathbb{I}_{ss} x \geq \mathbf{0} \end{aligned} \tag{17.4.14}$$

This is a quadratic program, therefore it is convex.

### Efficient Frontier

If values of the maximum return  $\bar{R}^*(\sigma)$  for prescribed admissible risk  $\sigma$  is plotted against  $\sigma$ , this is known as the efficient frontier. This is because any portfolio which yields return less than  $\bar{R}^*(\sigma)$  with risk  $\sigma$  is less efficient than any portfolio on the efficient frontier. Also, since increasing  $\sigma$  relaxes the optimisation problem (increasing the size of the feasible set), it follows that  $\bar{R}^*(\sigma)$  is non-decreasing in  $\sigma$  (over the domain where the problem is feasible with respect to  $\sigma$ ).



Alternatively, an efficient frontier could be drawn by plotting the minimum risk  $\sigma^*(\mu)$  subject to a lower bound  $\mu$  on the expected return, against  $\mu$ . Since decreasing  $\mu$  relaxes the optimisation problem (by increasing the size of the feasible set), it follows that  $\sigma^*(\mu)$  is non-decreasing in  $\mu$  (over the domain where the problem is feasible with respect to  $\mu$ ).

Under some regularity conditions, we can show that the efficient frontiers computed either way coincide. Assume that  $\bar{R}^*(\cdot)$  and  $\sigma^*(\cdot)$  are both strictly increasing in their respective arguments. For some nominal  $\sigma_0$ , let  $R_1^* := \bar{R}^*(\sigma_0)$  be a point on the efficient frontier computed via  $\bar{R}^*(\cdot)$ . Since there exists a possible portfolio with  $(\sigma_0, R_1^*)$ , then

$$\sigma^*(R_1^*) \leq \sigma_0 \tag{17.4.15}$$

by definition of  $\sigma^*(\cdot)$ . Suppose we have strict inequality, i.e.  $\sigma_1^* := \sigma^*(R_1^*) < \sigma_0$ . By the fact that  $\bar{R}^*(\cdot)$  is strictly increasing, we would have  $\bar{R}^*(\sigma_1^*) < \bar{R}^*(\sigma_0) = R_1^*$ . However, this implies the existence of a portfolio with  $(\sigma_1^*, R_1^*)$ , that has expected return greater than  $\bar{R}^*(\sigma_1^*)$ . This is contradictory, therefore it must be that

$$\sigma^*(\bar{R}^*(\sigma_0)) = \sigma_0 \tag{17.4.16}$$

In the same way, for some nominal  $\mu_0$  we will have

$$\bar{R}^*(\sigma^*(\mu_0)) = \mu_0 \tag{17.4.17}$$

## Sharpe Ratio Portfolio Optimisation

Let  $r_0$  denote the risk-free rate of return, i.e. there always exists an asset (excluded from the  $n$  assets under consideration) which yields return  $r_0$  with probability one. The Sharpe ratio of a portfolio  $x$  quantifies the expected return in excess of the risk-free rate, per unit risk:

$$\text{SR}(x) = \frac{\bar{r}^\top x - r_0}{\sqrt{x^\top \Sigma x}} \quad (17.4.18)$$

Assume that:

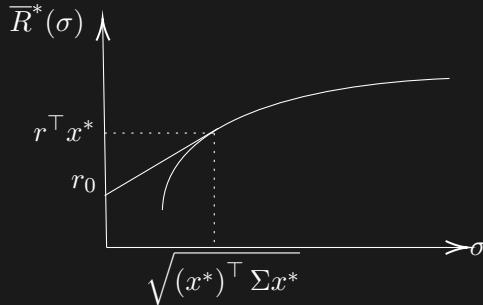
- $r_0 < \bar{r}^\top x$  for any portfolio  $x$  on the efficient frontier.
- $\Sigma \succ \mathbf{0}$ , so that  $x^\top \Sigma x$  is always positive.

Then the Sharpe-optimal portfolio may be found by solving

$$\begin{aligned} \max_u \quad & \text{SR}(x_0 + u) \\ \text{s.t.} \quad & x = x_0 + u \\ & \bar{r}^\top x > r_0 \\ & \mathbf{1}^\top u = 0 \\ & \mathbb{I}_{ss} x \geq \mathbf{0} \end{aligned} \quad (17.4.19)$$

Upon finding the Sharpe-optimal portfolio  $x^*$ , the point  $\left(\sqrt{(x^*)^\top \Sigma x^*}, \bar{r}^\top x^*\right)$  will lie on the efficient frontier. To see why, if  $x^*$  maximises the Sharpe ratio, then there cannot be another portfolio with risk  $\sqrt{(x^*)^\top \Sigma x^*}$  that has return greater than  $\bar{r}^\top x^*$  (otherwise this would have a larger Sharpe ratio). Hence  $\bar{R}^* \left( \sqrt{(x^*)^\top \Sigma x^*} \right) = \bar{r}^\top x^*$ .

Consider the straight line that connects  $(0, r_0)$  to  $\left(\sqrt{(x^*)^\top \Sigma x^*}, \bar{r}^\top x^*\right)$ . This will have a slope of  $\frac{\bar{r}^\top x^* - r_0}{\sqrt{(x^*)^\top \Sigma x^*}}$ , which is actually the optimal Sharpe ratio. This line is called the *capital allocation line*. If the investor is allowed to additionally invest in the risk-free asset, then they may achieve a portfolio with any point on this line, via a combination of the risk-free asset and the Sharpe-optimal portfolio. This is visualised as follows.



The form of Sharp ratio optimisation problem above is non-convex, but it can be cast into a convex form as follows. Introduce the slack variable  $\gamma > 0$ , and multiply the numerator and denominator of the Sharpe ratio by  $\gamma$ , so it becomes

$$\text{SR}(x) = \frac{\bar{r}^\top \tilde{x} - \gamma r_0}{\sqrt{\tilde{x}^\top \Sigma \tilde{x}}} \quad (17.4.20)$$

where  $\tilde{x} = \gamma x_0 + \tilde{u}$  and  $\tilde{u} = \gamma u$ . Also since the Sharpe ratio is positive, maximising  $\text{SR}(x)$  is equivalent to minimising  $1/\text{SR}(x)$ . The problem can be rewritten as

$$\begin{aligned} \min_{\tilde{u} \in \mathbb{R}^n, \gamma > 0} \quad & \frac{\sqrt{(\gamma x_0 + \tilde{u})^\top \Sigma (\gamma x_0 + \tilde{u})}}{\bar{r}^\top (\gamma x_0 + \tilde{u}) - \gamma r_0} \\ \text{s.t.} \quad & \tilde{x} = \gamma x_0 + \tilde{u} \\ & \bar{r}^\top \tilde{x} > \gamma r_0 \\ & \mathbf{1}^\top \tilde{u} = 0 \\ & \mathbb{I}_{ss} \tilde{x} \geq \mathbf{0} \end{aligned} \tag{17.4.21}$$

Note that there are infinitely many solutions to this problem, since  $u = \tilde{u}/\gamma$ , so if  $(\tilde{u}^*, \gamma^*)$  is a solution, then so is  $(c\tilde{u}^*, c\gamma^*)$ , for any  $c > 0$ . To resolve this, we can fix the denominator to  $\bar{r}^\top \tilde{x} - \gamma r_0 = 1$ , so we end up finding the solution that makes the denominator one. The problem now becomes

$$\begin{aligned} \min_{\tilde{u} \in \mathbb{R}^n, \gamma > 0} \quad & \sqrt{(\gamma x_0 + \tilde{u})^\top \Sigma (\gamma x_0 + \tilde{u})} \\ \text{s.t.} \quad & \tilde{x} = \gamma x_0 + \tilde{u} \\ & \bar{r}^\top \tilde{x} - \gamma r_0 = 1 \\ & \mathbf{1}^\top \tilde{u} = 0 \\ & \mathbb{I}_{ss} \tilde{x} \geq \mathbf{0} \end{aligned} \tag{17.4.22}$$

Lastly, note that  $\sqrt{(\gamma x_0 + \tilde{u})^\top \Sigma (\gamma x_0 + \tilde{u})} = \|\Sigma^{1/2} \tilde{x}\|_2$ , so introduce another slack variable  $t \geq 0$  and pose the equivalent problem

$$\begin{aligned} \min_{\tilde{u} \in \mathbb{R}^n, \gamma > 0, t \geq 0} \quad & t \\ \text{s.t.} \quad & \left\| \Sigma^{1/2} \tilde{x} \right\|_2 \leq t \\ & \tilde{x} = \gamma x_0 + \tilde{u} \\ & \bar{r}^\top \tilde{x} - \gamma r_0 = 1 \\ & \mathbf{1}^\top \tilde{u} = 0 \\ & \mathbb{I}_{ss} \tilde{x} \geq \mathbf{0} \end{aligned} \tag{17.4.23}$$

This is equivalent because a solution will satisfy  $\|\Sigma^{1/2} \tilde{x}\|_2 = t$ . To see why this is the case, suppose the solution satisfied  $\|\Sigma^{1/2} \tilde{x}\|_2 < t$ . Then this is a contradiction, because this implies we could lower the objective by making  $t = \|\Sigma^{1/2} \tilde{x}\|_2$ . This problem is a second-order cone program (which is convex), due to the presence of the  $\|\Sigma^{1/2} \tilde{x}\|_2 \leq t$  constraint, and the linear equality constraints.

### Value-at-Risk Portfolio Optimisation

If  $R = r^\top x$  is the return of a portfolio  $x$ , then the loss of the portfolio is simply the negative return, i.e.  $L = -R$ . The value-at-risk (VaR) of a portfolio  $x$  at level  $\alpha \in (0, 1)$  is denoted  $\text{VaR}_\alpha(x)$ , and satisfies the following equivalent characterisations:

- The probability of a loss  $\text{VaR}_\alpha(x)$  or greater is no more than  $\alpha$ .
- The probability of a return  $-\text{VaR}_\alpha(x)$  or less is no more than  $\alpha$ .
- The probability of a loss less than  $\text{VaR}_\alpha(x)$  is more than  $1 - \alpha$ .

- The probability of a return more than  $-\text{VaR}_\alpha(x)$  is more than  $1 - \alpha$ .

The value-at-risk is one way to measure risk, for which a larger value (with fixed  $\alpha$ ) indicates a larger risk. The value of  $\alpha$  is usually chosen to be small, so that the value-at-risk represents a possible amount of loss that is ‘rare’. Of course, there may be more than one value which satisfies the characterisations above. Formally, the value-at-risk is defined as

$$\text{VaR}_\alpha(x) = \inf \left\{ \zeta \in \mathbb{R} : \Pr(-r^\top x \geq \zeta) \leq \alpha \right\} \quad (17.4.24)$$

where the infimum is taken because the distribution of the loss in general could be a discrete distribution, so that  $\Pr(-r^\top x \geq \zeta)$  is left-continuous and non-increasing in  $\zeta$ . This means that  $\Pr(-r^\top x \geq \zeta) \geq \Pr(-r^\top x > \zeta)$  for any  $\zeta$ , hence if we ever had  $\Pr(-r^\top x \geq \zeta) > \alpha > \Pr(-r^\top x > \zeta)$ , then the infimum would always exist, whereas the minimum would not.

The value-at-risk can also be equivalently defined as

$$\text{VaR}_\alpha(x) = \inf \left\{ \zeta \in \mathbb{R} : \Pr(r^\top x \leq -\zeta) \leq \alpha \right\} \quad (17.4.25)$$

$$= -\sup \left\{ \xi \in \mathbb{R} : \Pr(r^\top x \leq \xi) \leq \alpha \right\} \quad (17.4.26)$$

If the CDF of the return distribution is continuous and strictly increasing, then the supremum is equal to

$$\sup \left\{ \xi \in \mathbb{R} : \Pr(r^\top x \leq \xi) \leq \alpha \right\} = \inf \left\{ \xi \in \mathbb{R} : \Pr(r^\top x \leq \xi) \geq \alpha \right\} \quad (17.4.27)$$

$$= F_{r^\top x}^{-1}(\alpha) \quad (17.4.28)$$

where  $F_{r^\top x}^{-1}(\alpha)$  is the quantile function of the return distribution. Thus in this case, the value-at-risk is

$$\text{VaR}_\alpha(x) = -F_{r^\top x}^{-1}(\alpha) \quad (17.4.29)$$

i.e. the top  $100\alpha$  percentile of losses.

Suppose the rate of return  $r$  is normally distributed with  $r \sim \mathcal{N}(\bar{r}, \Sigma)$ . Then the return is also normally distributed with

$$r^\top x \sim \mathcal{N}(\bar{r}^\top x, x^\top \Sigma x) \quad (17.4.30)$$

and the value-at-risk is

$$\text{VaR}_\alpha(x) = -\Phi_{\bar{r}^\top x, x^\top \Sigma x}^{-1}(\alpha) \quad (17.4.31)$$

where  $\Phi_{\bar{r}^\top x, x^\top \Sigma x}^{-1}(\cdot)$  is the quantile function of the  $\mathcal{N}(\bar{r}^\top x, x^\top \Sigma x)$  distribution. In portfolio optimisation, we can include a maximum value-at-risk as a constraint, and then maximise the expected return subject to this. Write this constraint as

$$\text{VaR}_\alpha(x) \leq v \quad (17.4.32)$$

Under the normally distributed returns assumption, this can be rearranged as follows:

$$-\Phi_{\bar{r}^\top x, x^\top \Sigma x}^{-1}(\alpha) \leq v \quad (17.4.33)$$

$$\Phi_{\bar{r}^\top x, x^\top \Sigma x}^{-1}(\alpha) \geq -v \quad (17.4.34)$$

$$\alpha \geq \Phi_{\bar{r}^\top x, x^\top \Sigma x}^{-1}(-v) \quad (17.4.35)$$

$$\Pr(r^\top x \leq -v) \leq \alpha \quad (17.4.36)$$

From here, we have

$$\Pr(r^\top x \leq -v) = \Pr\left(\frac{r^\top x - \bar{r}^\top x}{\sqrt{x^\top \Sigma x}} \leq \frac{-v - \bar{r}^\top x}{\sqrt{x^\top \Sigma x}}\right) \quad (17.4.37)$$

$$= \Phi\left(\frac{-v - \bar{r}^\top x}{\sqrt{x^\top \Sigma x}}\right) \quad (17.4.38)$$

$$\leq \alpha \quad (17.4.39)$$

since  $\frac{r^\top x - \bar{r}^\top x}{\sqrt{x^\top \Sigma x}}$  is standard normal. Then

$$1 - \Phi\left(\frac{-v - \bar{r}^\top x}{\sqrt{x^\top \Sigma x}}\right) \geq 1 - \alpha \quad (17.4.40)$$

Due to the symmetry about zero of the standard Gaussian,

$$\Phi\left(\frac{\bar{r}^\top x + v}{\sqrt{x^\top \Sigma x}}\right) \geq 1 - \alpha \quad (17.4.41)$$

$$\frac{\bar{r}^\top x + v}{\sqrt{x^\top \Sigma x}} \geq \Phi^{-1}(1 - \alpha) \quad (17.4.42)$$

If  $\alpha < 1/2$  (which should naturally be satisfied, since  $\alpha$  is usually taken to be small), then  $\Phi^{-1}(1 - \alpha) > 0$  and we have

$$\left\| \Sigma^{1/2} x \right\|_2 = \sqrt{x^\top \Sigma x} \quad (17.4.43)$$

$$\leq \frac{\bar{r}^\top x + v}{\Phi^{-1}(1 - \alpha)} \quad (17.4.44)$$

which is a second-order cone inequality in  $x$ . Thus the following second-order cone program can be formulated:

$$\begin{aligned} \max_u \quad & \bar{r}^\top (x_0 + u) \\ \text{s.t.} \quad & x = x_0 + u \\ & \left\| \Sigma^{1/2} x \right\|_2 \leq \frac{\bar{r}^\top x + v}{\Phi^{-1}(1 - \alpha)} \\ & \mathbf{1}^\top u = 0 \\ & \mathbb{I}_{\text{ss}} x \geq \mathbf{0} \end{aligned} \quad (17.4.45)$$

to maximise the expected return subject to an upper bound  $v$  on the value-at-risk, assuming normally distributed returns and  $\alpha < 1/2$ .

---

### 17.4.3 Capital Asset Pricing Model [118]

## 17.5 Discrete-Time Derivatives Pricing [27, 182, 202]

### 17.5.1 Binomial Trees

## 17.6 Continuous-Time Derivatives Pricing [93, 144, 217]

### 17.6.1 Black-Scholes Model

## 17.7 Optimal Stopping

### 17.7.1 Odds Algorithm [36]

#### Secretary Problem

### 17.7.2 Changepoint Detection [114]

### 17.7.3 Optional Stopping Theorem

## 17.8 Ruin Theory

## 17.9 Financial Econometrics

### 17.9.1 Beta Regression

### 17.9.2 Autoregressive Conditional Heteroskedasticity (ARCH) Models [172]

Consider a time-series  $X_t$  of the form

$$X_t = \mathbb{E}[X_t | \mathcal{F}_{t-1}] + U_t \quad (17.9.1)$$

where  $\mathcal{F}_{t-1} = (X_{t-1}, X_{t-2}, \dots)$  denotes the information up until time  $t - 1$ , and  $U_t$  is an error term. The form of  $\mathbb{E}[X_t | \mathcal{F}_{t-1}]$  is arbitrary but treated as known, i.e. it could be from an AR specification, or it could even be for a nonlinear model. If there is *conditional homoskedasticity*, then the conditional variance of this process is a constant that does not depend on  $\mathcal{F}_{t-1}$ :

$$\text{Var}(X_t | \mathcal{F}_{t-1}) = \text{Var}(U_t | \mathcal{F}_{t-1}) \quad (17.9.2)$$

$$= \text{Var}(U_t) \quad (17.9.3)$$

Note that an ARMA model with lagged error terms would still be considered conditionally homoskedastic, because those lagged error terms have zero variance when conditioned on  $\mathcal{F}_{t-1}$  since they are each computed by

$$U_{t-j} = X_{t-j} - \mathbb{E}[X_{t-j} | \mathcal{F}_{t-j-1}] \quad (17.9.4)$$

When there is conditional heteroskedasticity, the conditional variance is no longer constant, and depends on  $\mathcal{F}_{t-1}$ . An ARCH model is a model for the error terms  $U_t$ , which also specifies the conditional variance of the error terms. The model defines the error term to be given by

$$U_t = \sigma_t \varepsilon_t \quad (17.9.5)$$

where  $\varepsilon_t$  is a sequence of zero-mean white noise independent of the sequence  $U_t$ , with  $\text{Var}(\varepsilon_t) = 1$ , and  $\sigma_t$  is the conditional standard deviation. Thus, the conditional variance is

$$\text{Var}(U_t | \mathcal{F}_{t-1}) = \sigma_t^2 \quad (17.9.6)$$

and  $\sigma_t^2$  is modelled by

$$\sigma_t^2 = \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \quad (17.9.7)$$

where  $\omega > 0$  is a constant and  $\alpha_1, \dots, \alpha_q \geq 0$  are parameters, which are constrained to ensure that the conditional variance is always positive. We denote this model by  $\text{ARCH}(q)$ . Note that the conditional variance does indeed depend on the past values  $\mathcal{F}_{t-1} = (X_{t-1}, X_{t-2}, \dots)$ , because we can write each of the past error terms as  $U_{t-j} = X_{t-j} - \mathbb{E}[X_{t-j} | \mathcal{F}_{t-j-1}]$ . Also, each of the squared error terms can be written as

$$U_t^2 = \varepsilon_t^2 \left( \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \right) \quad (17.9.8)$$

so we can see that  $U_t^2$  evolves in an autoregressive fashion, except we have multiplicative noise (rather than additive noise) with mean  $\mathbb{E}[\varepsilon_t^2] = 1$ . Then the model for the series  $X_t$  is given by

$$X_t = \mathbb{E}[X_t | \mathcal{F}_{t-1}] + \varepsilon_t \sqrt{\omega + \sum_{j=1}^q \alpha_j U_{t-j}^2} \quad (17.9.9)$$

The intuition behind an ARCH process is that the variance of the error term is dependent on the magnitudes of the previous errors. Thus, periods of large errors are likely to be clustered together, as are periods of small errors.

### AR Representation of ARCH Models

Note that while  $U_t^2 = \varepsilon_t^2 \left( \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \right)$  appears like an AR process with independent multiplicative noise, it can actually be represented as an AR process with additive noise. Define the *innovation* term for this model by

$$\eta_t := U_t^2 - \sigma_t^2 \quad (17.9.10)$$

$$= U_t^2 - \omega - \sum_{j=1}^q \alpha_j U_{t-j}^2 \quad (17.9.11)$$

Then

$$U_t^2 = \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 + \eta_t \quad (17.9.12)$$

This looks like an AR process in  $U_t^2$ , but we investigate the properties on the innovation sequence  $\eta_t$ . Note that

$$\eta_t = (\sigma_t \varepsilon_t)^2 - \sigma_t^2 \quad (17.9.13)$$

$$= \sigma_t^2 (\varepsilon_t^2 - 1) \quad (17.9.14)$$

So

$$\mathbb{E}[\eta_t] = \mathbb{E}[\sigma_t^2] \mathbb{E}[\varepsilon_t^2 - 1] \quad (17.9.15)$$

$$= 0 \quad (17.9.16)$$

since  $\mathbb{E}[\varepsilon_t^2] = 1$ . Next,  $\eta_t$  can be shown to be white. For  $t \neq s$ , we have

$$\text{Cov}(\eta_t, \eta_s) = \mathbb{E}[\sigma_t (\varepsilon_t^2 - 1) \sigma_s (\varepsilon_s^2 - 1)] \quad (17.9.17)$$

$$= \mathbb{E} [\sigma_t \sigma_s (\varepsilon_t^2 - 1)] \mathbb{E} [(\varepsilon_s^2 - 1)] \quad (17.9.18)$$

$$= 0 \quad (17.9.19)$$

Lastly, the unconditional variance of  $\eta_t$  is given by

$$\text{Var}(\eta_t) = \mathbb{E} [\eta_t^2] \quad (17.9.20)$$

$$= \mathbb{E} [\sigma_t^4] \mathbb{E} [(\varepsilon_t^2 - 1)^2] \quad (17.9.21)$$

$$= \mathbb{E} [\sigma_t^4] (\mathbb{E} [\varepsilon_t^4] - 2\mathbb{E} [\varepsilon_t^2] + 1) \quad (17.9.22)$$

$$= \mathbb{E} [\sigma_t^4] (\mathbb{E} [\varepsilon_t^4] - 1) \quad (17.9.23)$$

The term  $\varepsilon_t$  is assumed to have finite fourth moment, e.g. if  $\varepsilon_t$  is assumed to be standard normal, then  $\mathbb{E} [\varepsilon_t^4] = 3$ . Taking this value for simplicity, then

$$\text{Var}(\eta_t) = 2\mathbb{E} [\sigma_t^4] \quad (17.9.24)$$

$$= 2\mathbb{E} \left[ \left( \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \right)^2 \right] \quad (17.9.25)$$

This will involve the autocorrelations of  $U_t^2$ . So  $\text{Var}(\eta_t)$  can be a constant that does not depend on time, if  $U_t^2$  is a weakly stationary AR process (which is itself determined by the coefficients  $\alpha_1, \dots, \alpha_q$ ). Therefore,  $U_t^2$  can be represented by an AR process with constant variance but conditionally heteroskedastic white noise  $\eta_t$ , with conditional variance (assuming standard normal  $\varepsilon_t$ ):

$$\text{Var}(\eta_t | \mathcal{F}_{t-1}) = \mathbb{E} [\sigma_t^4 (\varepsilon_t^2 - 1)^2 | \mathcal{F}_{t-1}] \quad (17.9.26)$$

$$= 2 \left( \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \right)^2 \quad (17.9.27)$$

### Moments of ARCH Processes [43]

The unconditional expectation of the error terms can be computed to be zero:

$$\mathbb{E}[U_t] = \mathbb{E}[\mathbb{E}[U_t | \mathcal{F}_{t-1}]] \quad (17.9.28)$$

$$= \mathbb{E}[\mathbb{E}[\sigma_t \varepsilon_t | \mathcal{F}_{t-1}]] \quad (17.9.29)$$

$$= \mathbb{E}[\sigma_t \mathbb{E}[\varepsilon_t | \mathcal{F}_{t-1}]] \quad (17.9.30)$$

$$= 0 \quad (17.9.31)$$

since  $\mathbb{E}[\varepsilon_t | \mathcal{F}_{t-1}] = 0$  by assumption. Hence the unconditional variance is equal to the second moment:

$$\text{Var}(U_t) = \mathbb{E}[U_t^2] \quad (17.9.32)$$

$$= \mathbb{E} \left[ \varepsilon_t^2 \left( \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \right) \right] \quad (17.9.33)$$

$$= \mathbb{E}[\varepsilon_t^2] \mathbb{E} \left[ \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \right] \quad (17.9.34)$$

$$= \omega + \sum_{j=1}^q \alpha_j \mathbb{E}[U_{t-j}^2] \quad (17.9.35)$$

$$= \omega + \sum_{j=1}^q \alpha_j \text{Var}(U_{t-j}) \quad (17.9.36)$$

Note that this is a linear difference equation in  $\text{Var}(U_t)$ , so we can analyse its stability properties in the same way as was established for AR models. It follows that if the roots of

$$1 - \alpha_1 L - \cdots - \alpha_q L^q = 0 \quad (17.9.37)$$

lie outside the unit circle, then the difference equation is stable, and the solution is given by the stationary unconditional variance, by making  $\text{Var}(U_t) = \text{Var}(U_{t-1}) = \cdots = \text{Var}(U_{t-q})$ :

$$\text{Var}(U_t) = \frac{\omega}{1 - \alpha_1 - \cdots - \alpha_q} \quad (17.9.38)$$

From this, we can see that we also require  $\alpha_1 + \cdots + \alpha_q < 1$  in order for this unconditional variance make sense. But since  $\alpha_1, \dots, \alpha_q \geq 0$ , then this is already implied by the stability condition (to see why, if we had  $\alpha_1 + \cdots + \alpha_q \geq 1$  this would suggest  $\omega + \sum_{j=1}^q \alpha_j \text{Var}(U_t) > \text{Var}(U_t)$  which is contradictory). It is also instructive to consider the fourth moments of  $U_t$ , which we will compute only for the ARCH(1), for simplicity.

$$\mathbb{E}[U_t^4] = \mathbb{E}[\varepsilon_t^4 (\omega + \alpha_1 U_{t-1}^2)^2] \quad (17.9.39)$$

$$= \mathbb{E}[\varepsilon_t^4] (\omega^2 + 2\alpha_1 \omega \mathbb{E}[U_{t-1}^2] + \alpha_1^2 \mathbb{E}[U_{t-1}^4]) \quad (17.9.40)$$

If we assume that  $\varepsilon_t$  is normally distributed, then its fourth moment is given by  $\mathbb{E}[\varepsilon_t^4] = 3$ , and substituting the expression for the stationary second moment  $\mathbb{E}[U_t^2] = \mathbb{E}[U_{t-1}^2]$  from above for the ARCH(1) specification gives

$$\mathbb{E}[U_t^4] = 3 \left( \omega^2 + \frac{2\alpha_1 \omega^2}{1 - \alpha_1} + \alpha_1^2 \mathbb{E}[U_{t-1}^4] \right) \quad (17.9.41)$$

This yields a first-order linear difference equation for  $\mathbb{E}[U_t^4]$ . We now assume that  $\alpha_1 < 1/\sqrt{3}$ , which will also play a role later. Then this difference equation is stable, and the stationary fourth moment can be found by rearranging:

$$\mathbb{E}[U_t^4] (1 - 3\alpha_1^2) = 3 \left( \omega^2 + \frac{2\alpha_1 \omega^2}{1 - \alpha_1} \right) \quad (17.9.42)$$

and then solving for  $\mathbb{E}[U_t^4]$ :

$$\mathbb{E}[U_t^4] = \frac{3}{1 - 3\alpha_1^2} \left( \omega^2 + \frac{2\alpha_1 \omega^2}{1 - \alpha_1} \right) \quad (17.9.43)$$

$$= \frac{3\omega^2}{1 - 3\alpha_1^2} \left( 1 + \frac{2\alpha_1}{1 - \alpha_1} \right) \quad (17.9.44)$$

$$= \frac{3\omega^2}{1 - 3\alpha_1^2} \left( \frac{1 - \alpha_1 + 2\alpha_1}{1 - \alpha_1} \right) \quad (17.9.45)$$

$$= \frac{3\omega^2(1 + \alpha_1)}{(1 - 3\alpha_1^2)(1 - \alpha_1)} \quad (17.9.46)$$

which is guaranteed to have a positive denominator via the assumption  $\alpha_1 < 1/\sqrt{3}$ . The unconditional kurtosis  $\kappa_U$  can then be calculated as

$$\kappa_U = \frac{\mathbb{E}[U_t^4]}{\text{Var}(U_t)^2} \quad (17.9.47)$$

$$= \frac{3\omega^2(1+\alpha_1)}{(1-3\alpha_1^2)(1-\alpha_1)} \cdot \frac{(1-\alpha_1)^2}{\omega^2} \quad (17.9.48)$$

$$= \frac{3(1+\alpha_1)(1-\alpha_1)}{1-3\alpha_1^2} \quad (17.9.49)$$

$$= \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2} \quad (17.9.50)$$

Therefore  $\kappa_U > 3$  since  $1 - \alpha_1^2 > 1 - 3\alpha_1^2$ . So the error terms in an ARCH model can be characterised as being leptokurtic.

### Estimation of ARCH Models

Suppose we have gathered some data for a series  $X_t$  and observed the error terms  $U_1, \dots, U_T$ , which can be obtained from  $U_t = X_t - \mathbb{E}[X_t | \mathcal{F}_{t-1}]$ . This assumes that we know  $\mathbb{E}[X_t | \mathcal{F}_{t-1}]$ . If it is not known however, it can still be replaced by some estimate  $\widehat{\mathbb{E}}[X_t | \mathcal{F}_{t-1}]$  of the conditional expectation (e.g. through another fitted time-series model for  $X_t$ ). Then we can treat

$$\widehat{U}_t = X_t - \widehat{\mathbb{E}}[X_t | \mathcal{F}_{t-1}] \quad (17.9.51)$$

as the observed series. From this, the parameters  $\omega, \alpha_1, \dots, \alpha_q$  in an ARCH( $q$ ) model can be fitted. One way to do this is to use the AR representation for  $U_t^2$ , and then use standard approaches for fitting AR models, such as least squares or the Yule-Walker equations.

### Maximum Likelihood Estimation of ARCH Models [63]

To estimate the parameters  $\omega, \alpha_1, \dots, \alpha_q$  in an ARCH( $q$ ) model via a maximum likelihood approach, we can assume that the terms  $\varepsilon_t$  are standard normal, so that the conditional distribution of  $U_t$  given  $\mathcal{F}_{t-1}$  is also normally distributed. The conditional mean and variance are given by

$$\mathbb{E}[U_t | \mathcal{F}_{t-1}] = \mathbb{E}[\sigma_t \mathbb{E}[\varepsilon_t | \mathcal{F}_{t-1}]] \quad (17.9.52)$$

$$= 0 \quad (17.9.53)$$

and

$$\text{Var}(U_t | \mathcal{F}_{t-1}) = \sigma_t^2 \quad (17.9.54)$$

$$= \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 \quad (17.9.55)$$

Hence the conditional normal density, and thus the conditional likelihood is expressed as

$$f(u_t | \mathcal{F}_{t-1}) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp\left(-\frac{u_t^2}{2\sigma_t^2}\right) \quad (17.9.56)$$

$$= \frac{1}{\sqrt{\omega + \sum_{j=1}^q \alpha_j U_{t-j}^2} \sqrt{2\pi}} \exp\left(-\frac{u_t^2}{2(\omega + \sum_{j=1}^q \alpha_j U_{t-j}^2)^2}\right) \quad (17.9.57)$$

$$= \mathcal{L}(\omega, \alpha_1, \dots, \alpha_q; u_t | \mathcal{F}_{t-1}) \quad (17.9.58)$$

Then the full likelihood of the ARCH parameters given  $U_1, \dots, U_T$  is factorised as

$$\mathcal{L}(\omega, \alpha_1, \dots, \alpha_q; u_1, \dots, u_T) = f(u_1, \dots, u_T) \quad (17.9.59)$$

$$= f(u_T | \mathcal{F}_{T-1}) f(u_{T-1} | \mathcal{F}_{T-2}) \dots f(u_{q+1} | \mathcal{F}_q) f(u_q, \dots, u_1) \quad (17.9.60)$$

Then the log likelihood is

$$\log \mathcal{L}(\omega, \alpha_1, \dots, \alpha_q; u_1, \dots, u_T) = \log f(u_1, \dots, u_T; \omega, \alpha_1, \dots, \alpha_q) \quad (17.9.61)$$

$$= \sum_{t=q+1}^T \log f(u_t | \mathcal{F}_{t-1}) + \log f(u_q, \dots, u_1) \quad (17.9.62)$$

$$\begin{aligned} &= -\frac{T-q}{2} \log(2\pi) - \frac{1}{2} \sum_{t=q+1}^T \log \left( \omega + \sum_{j=1}^q \alpha_j u_{t-j}^2 \right) \\ &\quad - \frac{1}{2} \sum_{t=q+1}^T \frac{u_t^2}{\left( \omega + \sum_{j=1}^q \alpha_j u_{t-j}^2 \right)^2} + \log f(u_q, \dots, u_1) \end{aligned} \quad (17.9.63)$$

All the terms in the log likelihood can be evaluated given  $U_1, \dots, U_T$ , except for the initial log density  $\log f(u_q, \dots, u_1)$ . However, this does not matter too much if we assume  $f(u_q, \dots, u_1)$  does not depend on the parameters, or if we condition the likelihood on  $U_1, \dots, U_q$ . The contribution of  $f(u_q, \dots, u_1)$  to the likelihood will be negligible for large  $T$ . Then, the maximum likelihood estimate can be obtained by solving the problem

$$\begin{aligned} (\hat{\omega}, \hat{\alpha}_1, \dots, \hat{\alpha}_q) &= \underset{\omega, \alpha_1, \dots, \alpha_q}{\operatorname{argmax}} \log \mathcal{L}(\omega, \alpha_1, \dots, \alpha_q; u_1, \dots, u_T) \\ \text{s.t. } &\omega > 0 \\ &\alpha_j \geq 0, \quad j = 1, \dots, q \\ &\alpha_1 + \dots + \alpha_q < 1 \end{aligned} \quad (17.9.64)$$

where the constraint  $\alpha_1 + \dots + \alpha_q < 1$  is imposed with the aim of fitting a stationary ARCH model.

### 17.9.3 Generalised Autoregressive Conditional Heteroskedasticity (GARCH) Models

A GARCH model is a generalisation of the ARCH model, and is denoted by GARCH  $(p, q)$ . We have

$$U_t = \sigma_t \varepsilon_t \quad (17.9.65)$$

as before, and the term  $\sigma_t^2$  is now modelled with an autoregressive component:

$$\sigma_t^2 = \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2 \quad (17.9.66)$$

where  $\beta_1, \dots, \beta_p \geq 0$  are parameters for the  $p^{\text{th}}$  order AR component. So when  $p = 0$ , this just becomes an ARCH  $(q)$  model.

#### ARMA Representation of GARCH Models

A GARCH model can be represented as an ARMA model in the series  $U_t^2$ . In the same way as representing an ARCH model with an AR model, define the innovation sequence  $\eta_t := U_t^2 - \sigma_t^2$  so that  $\sigma_t^2 = U_t^2 - \eta_t$  and  $\eta_t = \sigma_t^2 (\varepsilon_t^2 - 1)$ . Substituting these relations, we have

$$U_t^2 = \sigma_t^2 + \eta_t \quad (17.9.67)$$

$$= \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2 + \eta_t \quad (17.9.68)$$

$$= \omega + \sum_{j=1}^q \alpha_j U_{t-j}^2 + \sum_{j=1}^p \beta_j (U_{t-j}^2 - \eta_{t-j}) + \eta_t \quad (17.9.69)$$

$$= \omega + \sum_{j=1}^{\max\{q,p\}} (\alpha_j + \beta_j) U_{t-j}^2 + \eta_t - \sum_{j=1}^p \beta_j \eta_{t-j} \quad (17.9.70)$$

where it is understood that  $\alpha_j = 0$  whenever  $j > q$  and  $\beta_j = 0$  whenever  $j > p$ . This now appears like an ARMA  $(\max\{q,p\}, p)$  model in series  $U_t^2$  and with noise sequence  $\eta_t$ . We may show that  $\eta_t$  is a martingale difference sequence, i.e.  $\mathbb{E}[\eta_t | \mathcal{F}_t] = 0$  as follows.

$$\mathbb{E}[\eta_t | \mathcal{F}_t] = \mathbb{E}[U_t^2 - \sigma_t^2 | \mathcal{F}_t] \quad (17.9.71)$$

$$= \mathbb{E}[\sigma_t^2 \varepsilon_t^2 | \mathcal{F}_t] - \mathbb{E}[\sigma_t^2 | \mathcal{F}_t] \quad (17.9.72)$$

$$= \mathbb{E}[\varepsilon_t^2] \mathbb{E}[\sigma_t^2 | \mathcal{F}_t] - \mathbb{E}[\sigma_t^2 | \mathcal{F}_t] \quad (17.9.73)$$

$$= 0 \quad (17.9.74)$$

under the usual assumption  $\mathbb{E}[\varepsilon_t^2] = 1$ . Then the unconditional mean is

$$\mathbb{E}[\eta_t] = \mathbb{E}[\mathbb{E}[\eta_t | \mathcal{F}_t]] \quad (17.9.75)$$

$$= 0 \quad (17.9.76)$$

and since  $\eta_t$  is mean independent given  $\eta_{t-1}$ ,  $\eta_{t-2}$ , etc. then this also implies

$$\text{Cov}(\eta_t, \eta_s) = 0 \quad (17.9.77)$$

for all  $t \neq s$ . However the unconditional variance  $\text{Var}(\eta_t)$  may in general not be a constant in time, meaning that  $\eta_t$  can be considered non-stationary white noise.

### ARCH Representation of GARCH Models [8]

A GARCH model can be represented as an ARCH  $(\infty)$  model. Consider a GARCH  $(1, 1)$ , which has

$$\sigma_t^2 = \omega + \alpha_1 U_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (17.9.78)$$

Then via successive substitution of  $\sigma_t^2$ , we get

$$\sigma_t^2 = \omega + \alpha_1 U_{t-1}^2 + \beta_1 (\omega + \alpha_1 U_{t-2}^2 + \beta_1 \sigma_{t-2}^2) \quad (17.9.79)$$

$$= \omega + \alpha_1 U_{t-1}^2 + \beta_1 [\omega + \alpha_1 U_{t-2}^2 + \beta_1 (\omega + \alpha_1 U_{t-3}^2 + \dots)] \quad (17.9.80)$$

$$= \sum_{j=0}^{\infty} \beta_1^j \omega + \sum_{j=0}^{\infty} \beta_1^j \alpha_1 U_{t-1}^2 \quad (17.9.81)$$

$$= \sum_{j=0}^{\infty} \beta_1^j \omega + \sum_{j=1}^{\infty} \beta_1^{j-1} \alpha_1 U_{t-1}^2 \quad (17.9.82)$$

If  $0 \leq \beta_1 < 1$ , then

$$\sigma_t^2 = \frac{\omega}{1 - \beta_1} + \sum_{j=1}^{\infty} \beta_1^{j-1} \alpha_1 U_{t-1}^2 \quad (17.9.83)$$

which is an ARCH  $(\infty)$  model. The same principle can be applied to show that higher order GARCH models are also ARCH  $(\infty)$  models. The practicality of this result is that a GARCH  $(1, 1)$  model can be fitted instead of a high order ARCH  $(q)$  model, the former which has fewer parameters than the latter.

## Moments of GARCH Processes

Like with ARCH models, the unconditional expectation of  $U_t$  is zero:

$$\mathbb{E}[U_t] = \mathbb{E}[\sigma_t \varepsilon_t] \quad (17.9.84)$$

$$= \mathbb{E}[\sigma_t] \mathbb{E}[\varepsilon_t] \quad (17.9.85)$$

$$= 0 \quad (17.9.86)$$

Thus the unconditional variance is the second moment. Using the ARMA representation of GARCH models, we have

$$\text{Var}(U_t) = \mathbb{E}[U_t^2] \quad (17.9.87)$$

$$= \omega + \sum_{j=1}^{\max\{q,p\}} (\alpha_j + \beta_j) \mathbb{E}[U_{t-j}^2] + \mathbb{E}[\eta_t] - \sum_{j=1}^p \beta_j \mathbb{E}[\eta_{t-j}] \quad (17.9.88)$$

$$= \omega + \sum_{j=1}^{\max\{q,p\}} (\alpha_j + \beta_j) \mathbb{E}[U_{t-j}^2] \quad (17.9.89)$$

Comparing to the ARCH case, the conditions for  $U_t$  to have a stationary variance are evident, which now depend on the roots of the characteristic polynomial with coefficients  $\alpha_j + \beta_j$ . The stationary variance will be given by

$$\text{Var}(U_t) = \frac{\omega}{1 - \sum_{j=1}^{\max\{q,p\}} (\alpha_j + \beta_j)} \quad (17.9.90)$$

where

$$\sum_{j=1}^{\max\{q,p\}} (\alpha_j + \beta_j) = \sum_{j=1}^q \alpha_j + \sum_{j=1}^p \beta_j \quad (17.9.91)$$

$$< 1 \quad (17.9.92)$$

The GARCH process also exhibits similar heavy-tailed characteristics to the ARCH process. Assuming that  $\varepsilon_t$  is standard normal so that  $\mathbb{E}[\varepsilon_t^4] = 3$ , we have

$$\mathbb{E}[U_t^4] = \mathbb{E}[\sigma_t^4 \varepsilon_t^4] \quad (17.9.93)$$

$$= \mathbb{E}[\varepsilon_t^4] \mathbb{E}[\sigma_t^4] \quad (17.9.94)$$

$$= 3\mathbb{E}[\sigma_t^4] \quad (17.9.95)$$

We focus on computing  $\mathbb{E}[\sigma_t^4]$ :

$$\mathbb{E}[\sigma_t^4] = \mathbb{E}[(\omega + \alpha_1 U_{t-1}^2 + \beta_1 \sigma_{t-1}^2)] \quad (17.9.96)$$

$$= \omega^2 + 2\alpha_1 \omega \mathbb{E}[U_{t-1}^2] + 2\beta_1 \omega \mathbb{E}[\sigma_{t-1}^2] + \alpha_1^2 \mathbb{E}[U_{t-1}^4] + 2\alpha_1 \beta_1 \mathbb{E}[U_{t-1}^2 \sigma_{t-1}^2] + \beta_1^2 \mathbb{E}[\sigma_{t-1}^4] \quad (17.9.97)$$

The stationary second moment of  $U_t$  is  $\mathbb{E}[U_t^2] = \frac{\omega}{1 - \alpha_1 - \beta_1}$  while for  $\sigma_t$  is can be found from

$$\mathbb{E}[\sigma_t^2] = \omega + \alpha_1 \mathbb{E}[U_{t-1}^2] + \beta_1 \mathbb{E}[\sigma_{t-1}^2] \quad (17.9.98)$$

Thus

$$\mathbb{E}[\sigma_t^2] = \frac{\omega}{1 - \alpha_1 - \beta_1} \quad (17.9.99)$$

Moreover, we have  $\mathbb{E} [U_{t-1}^4] = 3\mathbb{E} [\sigma_{t-1}^4]$  and

$$\begin{aligned}\mathbb{E} [U_{t-1}^2 \sigma_{t-1}^2] &= \mathbb{E} [\varepsilon_{t-1}^2] \mathbb{E} [\sigma_{t-1}^4] \\ &= \mathbb{E} [\sigma_{t-1}^4]\end{aligned}\quad (17.9.100)$$

$$(17.9.101)$$

Substituting all these yields

$$\mathbb{E} [\sigma_t^4] = \omega^2 + \frac{2\alpha_1\omega^2}{1 - \alpha_1 - \beta_1} + \frac{2\beta_1\omega^2}{1 - \alpha_1 - \beta_1} + 3\alpha_1^2 \mathbb{E} [\sigma_{t-1}^4] + 2\alpha_1\beta_1 \mathbb{E} [\sigma_{t-1}^4] + \beta_1^2 \mathbb{E} [\sigma_{t-1}^4] \quad (17.9.102)$$

The stability condition for this difference equation is  $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$ , so the stationary fourth moment of  $\sigma_t$  is given by

$$\mathbb{E} [\sigma_t^4] (1 - 3\alpha_1^2 - 2\alpha_1\beta_1 - \beta_1^2) = \omega^2 \left( \frac{1 - \alpha_1 - \beta_1 + 2\alpha_1 + 2\beta_1}{1 - \alpha_1 - \beta_1} \right) \quad (17.9.103)$$

which becomes

$$\mathbb{E} [\sigma_t^4] = \omega^2 \frac{(1 + \alpha_1 + \beta_1) / (1 - \alpha_1 - \beta_1)}{1 - 3\alpha_1^2 - 2\alpha_1\beta_1 - \beta_1^2} \quad (17.9.104)$$

Now the kurtosis of  $U_t$  is

$$\kappa_U = \frac{\mathbb{E} [U_t^4]}{\text{Var} (U_t)^2} \quad (17.9.105)$$

$$= \frac{3\mathbb{E} [\sigma_t^4]}{\mathbb{E} [U_t^2]^2} \quad (17.9.106)$$

$$= 3\omega^2 \frac{(1 + \alpha_1 + \beta_1) / (1 - \alpha_1 - \beta_1)}{1 - 3\alpha_1^2 - 2\alpha_1\beta_1 - \beta_1^2} \cdot \frac{(1 - \alpha_1 - \beta_1)^2}{\omega^2} \quad (17.9.107)$$

$$= 3 \frac{(1 + \alpha_1 + \beta_1)(1 - \alpha_1 - \beta_1)}{1 - 3\alpha_1^2 - 2\alpha_1\beta_1 - \beta_1^2} \quad (17.9.108)$$

$$= 3 \frac{1 - (\alpha_1 + \beta_1)^2}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} \quad (17.9.109)$$

$$> 3 \quad (17.9.110)$$

since  $1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2 < 1 - (\alpha_1 + \beta_1)^2$ .

### Maximum Likelihood Estimation of GARCH Models

Maximum likelihood to estimate parameters  $(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta})$  with  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)$  and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$  in a GARCH  $(q, p)$  just extends the way it is done in the ARCH case. Although we now need to leave out or condition on the first  $\max \{q, p\}$  terms. The log likelihood is now written as

$$\begin{aligned}\log \mathcal{L} (\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}; u_1, \dots, u_T) &= -\frac{T - \max \{q, p\}}{2} \log (2\pi) \\ &\quad - \frac{1}{2} \sum_{t=\max\{q,p\}+1}^T \log \left( \omega + \sum_{j=1}^q \alpha_j u_{t-j}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2 \right) \\ &\quad - \frac{1}{2} \sum_{t=q+1}^T \frac{u_t^2}{\left( \omega + \sum_{j=1}^q \alpha_j u_{t-j}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2 \right)^2} + \log f(u_{\max\{q,p\}}, \dots, u_1)\end{aligned}\quad (17.9.111)$$

where the  $\sigma_t^2$  can be computed recursively through the GARCH specification. However, we do not have values for  $\sigma_1, \dots, \sigma_{\max\{q,p\}}$ . In practice, these can be replaced by the unconditional sample variance of the  $U_t$  values. The maximum likelihood estimate is found by solving

$$\begin{aligned} (\hat{\omega}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \underset{\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}}{\operatorname{argmax}} \quad & \log \mathcal{L}(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}; u_1, \dots, u_T) \\ \text{s.t.} \quad & \omega > 0 \\ & \alpha_j \geq 0, \quad j = 1, \dots, q \\ & \beta_j \geq 0, \quad j = 1, \dots, p \\ & \sum_{j=1}^q \alpha_j + \sum_{j=1}^p \beta_j < 1 \end{aligned} \tag{17.9.112}$$



## Chapter 18

# Physics

### 18.1 Hamiltonian Monte-Carlo [34]

### 18.2 Statistical Mechanics

#### 18.2.1 Maxwell-Boltzmann Distribution

#### 18.2.2 Gibbs Distribution

#### 18.2.3 Brownian Motion

#### 18.2.4 Fokker-Planck Equations

#### 18.2.5 Statistical Thermodynamics [112]

#### Second Law of Thermodynamics [47]

### 18.3 Statistical Physics

#### 18.3.1 Mean Sojourn Time

### 18.4 Langevin Dynamics

#### 18.4.1 Langevin Monte-Carlo

### 18.5 Mean Field Theory

### 18.6 Quantum Mechanics [78, 214, 216]

#### 18.6.1 Quantum Probability

#### Quantum Probability Spaces

#### 18.6.2 Quantum Computing

#### Grover's Algorithm

#### 18.6.3 Quantum Stochastic Calculus

#### Quantum Filtering

#### Quantum Control

### 18.7 Econophysics [135]

#### 18.7.1 Statistical Finance [208]

---

#### 18.7.2 Quantum Finance

# Bibliography

- [1] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.
- [2] T. W. Anderson. *The Statistical Analysis of Time Series*. John Wiley & Sons, Hoboken, 1971.
- [3] William J. Anderson. *Continuous-Time Markov Chains: An Applications-Oriented Approach*. Springer, 1991.
- [4] Barry C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. SIAM, Philadelphia, PA, 2008.
- [5] L. Arnold. *Stochastic Differential Equations: Theory and Applications*. Wiley, New York, 1974.
- [6] Robert B. Ash. *Basic Probability Theory*. DOVER PUBN INC, 2008.
- [7] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.
- [8] Dimitrios Asteriou and Stephen G. Hall. *Applied Econometrics*. Palgrave Macmillan, Basingstoke England New York, 2nd edition, 2011.
- [9] Karl J. Astrom. *Introduction to Stochastic Control Theory*. Academic Press, 1970.
- [10] Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, 2007.
- [11] Adelchi Azzalini. *The Skew-Normal and Related Families*. Cambridge University Press, Cambridge, 2014.
- [12] Paolo Baldi. *Stochastic Calculus: An Introduction through Theory and Exercises*. Springer, Cham, Switzerland, 2017.
- [13] Pierre Baldi and Søren Brunak. *Bioinformatics: The Machine Learning Approach*. A Bradford Book, 2nd edition, 2001.
- [14] Badi Baltagi. *Econometric Analysis of Panel Data*. J. Wiley & Sons, Chichester Hoboken, NJ, 3rd edition, 2005.
- [15] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.
- [16] Heinz Bauer. *Probability Theory*. de Gruyter, 1996.
- [17] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer New York, 2 edition, 1985.

- [18] James O. Berger and Robert L. Wolpert. *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*. Institute of Mathematical Statistics, Hayward, Calif, 2nd edition, 1988.
  - [19] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, 1994.
  - [20] Dimitri P. Bertsekas. *Dynamic Programming and Stochastic Control*. Academic Press, 1976.
  - [21] Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2nd edition, 2008.
  - [22] S. Bhatnagar, H. L. Prasad, and L. A. Prashanth. *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*. Springer, London New York, 2013.
  - [23] Herman Bierens. *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge University Press, Cambridge, UK New York, 2005.
  - [24] Patrick Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
  - [25] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc., 2006.
  - [26] Gunnar Blom. *Probability and Statistics: Theory and Applications*. Springer, 1989.
  - [27] Stephen Blyth. *Introduction to Quantitative Finance*. Oxford University Press, USA, City, 2013.
  - [28] Serguei Bobkov and Michel Ledoux. *One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*. American Mathematical Society, Providence, RI, 2019.
  - [29] Denis Bosq. *Mathematical Statistics and Stochastic Processes*. Wiley, 2012.
  - [30] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
  - [31] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5th edition, 2015.
  - [32] Ulisses Braga-Neto. *Fundamentals of Pattern Recognition and Machine Learning*. Springer, Cham, Switzerland, 2020.
  - [33] Peter Brockwell. *Time Series: Theory and Methods*. Springer-Verlag, New York, 2nd edition, 1991.
  - [34] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.
  - [35] Robert Brown. *Introduction to Random Signals and Applied Kalman Filtering: With MATLAB Exercises*. John Wiley, Hoboken, NJ, 2012.
  - [36] F. Thomas Bruss. Sum the odds to one and stop. *The Annals of Probability*, 28(3):1384–1391, jun 2000.
  - [37] Sébastien Bubeck and Nicolo Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems*. Now Publishers Inc, 2012.
-

- [38] Ovidiu Calin. *An Informal Introduction to Stochastic Calculus with Applications*. WORLD SCIENTIFIC PUB CO INC, 2015.
- [39] Ovidiu Calin. *Deep Learning Architectures: A Mathematical Approach*. Springer, Cham, 2020.
- [40] Ovidiu Calin and Constantin Udriște. *Geometric Modeling in Probability and Statistics*. Springer, Cham, 2014.
- [41] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2001.
- [42] Christopher Chatfield. *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC, Boca Raton, FL, 6th edition, 2003.
- [43] Christopher Chatfield and Haipeng Xing. *The Analysis of Time Series: An Introduction with R*. CRC Press, Boca Raton, Florida, 7th edition, 2019.
- [44] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.
- [45] Pierre Collet, Servet Martínez, and Jaime San Martín. *Quasi-Stationary Distributions: Markov Chains, Diffusions and Dynamical Systems*. Springer, 2012.
- [46] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, 1999.
- [47] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.
- [48] Trevor F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall/CRC, 2nd edition, 2000.
- [49] Imre Csiszar and Paul Shields. *Information Theory and Statistics: A Tutorial*. Now Publishers Inc, 2004.
- [50] H. A. David and H. N. Nagaraja. *Order Statistics*. John Wiley, Hoboken, N.J, 2005.
- [51] Russell Davidson and James G. MacKinnon. *Econometric Theory and Methods*. Oxford University Press, New York, 2004.
- [52] Jan G. De Gooijer. *Elements of Nonlinear Time Series Analysis and Forecasting*. Springer, Cham, Switzerland, 2017.
- [53] Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.
- [54] Arnaud Doucet and Adam M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.
- [55] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.
- [56] Bradley Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.
- [57] Shlomo Engelberg. *Random Signals and Noise: A Mathematical Introduction*. London CRC Press, Boca Raton, 2007.

- [58] Michael D. Ernst. Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):676–685, nov 2004.
- [59] N. Etemadi. An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 55(1):119–122, feb 1981.
- [60] Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. Wiley, Hoboken, N.J, 4th edition, 2011.
- [61] Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer, 2nd edition, 2013.
- [62] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.
- [63] Jürgen Franke, Wolfgang Karl Härdle, and Christian Matthias Hafner. *Statistics of Financial Markets: An Introduction*. Springer, Cham, Switzerland, 5th edition, 2019.
- [64] Andrzej Galecki and Tomasz Burzykowski. *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer, New York, 2013.
- [65] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, 3rd edition, 2014.
- [66] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric Statistical Inference*. Chapman and Hall/CRC, 2011.
- [67] John Gittins. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2011.
- [68] Phillip Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer, New York, 3rd edition, 2005.
- [69] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.
- [70] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012.
- [71] Robert M. Gray. *Entropy and Information Theory*. Springer, 2011.
- [72] William H. Greene. *Econometric Analysis*. Pearson, 7th edition, 2012.
- [73] Priscilla E. Greenwood and Michael S. Nikulin. *A Guide to Chi-Squared Testing*. Wiley, New York, 1996.
- [74] Charles M. Grinstead and J. Laurie Snell. *Grinstead and Snell's Introduction to Probability*. ORANGE GROVE TEXTS, 2009.
- [75] Peter D. Grunwald. *The Minimum Description Length Principle*. The MIT Press, 2007.
- [76] Maya R. Gupta and Yihua Chen. *Theory and Use of the EM Algorithm*. Now Publishers, 2011.
- [77] James Douglas Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- [78] Richard W. Hamming. *The Art of Probability: For Scientists and Engineers*. Basic Books, 1991.

- [79] Wolfgang Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer, Berlin, 4th edition, 2015.
- [80] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.
- [81] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press LLC, Boca Raton, 2015.
- [82] Michio Hatanaka. *Time-Series-Based Econometrics: Unit Roots and Co-Integrations*. Oxford University Press, Oxford New York, 1996.
- [83] Melvin Hausner. *Elementary Probability Theory*. Springer, 1977.
- [84] Fumio Hayashi. *Econometrics*. Princeton University Press, Princeton, 2000.
- [85] M. H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, New York, 1996.
- [86] Ralf Herbrich, Thore Graepel, Peter Bollmann-Sdorra, and Klaus Obermayer. Learning preference relations for information retrieval. In *AAAI Workshop Text Categorization and Machine Learning*, 1998.
- [87] Robert V. Hogg, Joseph W. McKean, and Allen T. Craig. *Introduction to Mathematical Statistics*. Pearson, 7th edition, 2013.
- [88] Milan Holický. *Introduction to Probability and Statistics for Engineers*. Springer-Verlag GmbH, 2013.
- [89] Myles Hollander and Douglas A. Wolfe. *Nonparametric Statistical Methods*. Wiley-Interscience, 2nd edition, 1999.
- [90] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. Wiley, 3rd edition, 2014.
- [91] Ivana Horova, Jan Koláček, and Jiří Zelinka. *Kernel Smoothing in MATLAB: Theory and Practice of Kernel Smoothing*. World Scientific Publishing Co, 2012.
- [92] Stefan Höst. *Information and Communication Theory*. IEEE Press, Piscataway, NJ, 2019.
- [93] John C. Hull. *Options, Futures, and Other Derivatives*. Pearson, 10th edition, 2018.
- [94] Jorge Hurtado. *Structural Reliability: Statistical Learning Perspectives*. Springer, Berlin New York, 2004.
- [95] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018.
- [96] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.
- [97] Shun ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.
- [98] Rolf Isermann. *Identification of Dynamical Systems: An Introduction with Applications*. Springer, Berlin London, 2011.
- [99] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

- [100] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK New York, NY, 2003.
  - [101] Tony Jebara. *Machine Learning: Discriminative and Generative*. Kluwer Academic Publishers, Boston, 2004.
  - [102] Harry Joe. *Dependence Modeling with Copulas*. CRC Press, Boca Raton, 2014.
  - [103] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.
  - [104] Simon J. Julier and Jeffrey K. Uhlmann. A general method for approximating nonlinear transformations of probability distributions. Technical report, University of Oxford, 1996.
  - [105] Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear Estimation*. Prentice Hall, Upper Saddle River, N.J, 2000.
  - [106] Olav Kallenberg. *Foundations of Modern Probability*. Springer, 1997.
  - [107] K. J. Keesman. *System Identification: An Introduction*. Springer, London New York, 2011.
  - [108] Oscar Kempthorne and Leroy Folks. *Probability, Statistics, and Data Analysis*. Iowa State University Press, Ames, 1971.
  - [109] Maurice G. Kendall. *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. Charles Griffin & Company, 3rd edition, 1951.
  - [110] Rafail. Z. Khasminskii. *Stochastic Stability of Differential Equations*. Springer, Berlin Heidelberg, 2nd edition, 2012.
  - [111] J. F. Kiviet. *Monte Carlo Simulation for Econometricians*. Now, Boston, 2012.
  - [112] Oliver Knill. *Probability Theory and Stochastic Processes with Applications*. Overseas Press, New Delhi, India, 2009.
  - [113] Michael Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, N.Y, 2008.
  - [114] Vikram Krishnamurthy. *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge University Press, Cambridge, 2016.
  - [115] Dirk P. Kroese, Thomas Taimre, and Zdravko I. Botev. *Handbook of Monte Carlo Methods*. Wiley, 2011.
  - [116] P.R. Kumar and Pravin Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice Hall, 1986.
  - [117] Harold J. Kushner. *Stochastic Stability and Control*. Academic Press, 1967.
  - [118] Tze Leung Lai and Haipeng Xing. *Statistical Models and Methods for Financial Markets*. Springer, New York, 2008.
  - [119] Kenneth Lange. *Applied Probability*. Springer-Verlag GmbH, 2nd edition, 2010.
  - [120] Richard J. Larsen and Morris L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Pearson, 6th edition, 2018.
  - [121] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge New York, NY, 2020.
-

- [122] A. J. Lee. *U-Statistics: Theory and Practice*. Marcel Dekker, 1990.
  - [123] Lawrence M Leemis and Jacquelyn T McQueston. Univariate distribution relationships. *The American Statistician*, 62(1):45–53, feb 2008.
  - [124] E. L. Lehmann. *Testing Statistical Hypotheses*. Springer, 2nd edition, 1986.
  - [125] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, 1999.
  - [126] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.
  - [127] Alberto Leon-Garcia. *Probability, Statistics, and Random Processes For Electrical Engineering*. Pearson, 3rd edition, 2008.
  - [128] Haim Levy. *Stochastic Dominance: Investment Decision Making under Uncertainty*. Springer, 3rd edition, 2016.
  - [129] Ming Li and Paul M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2008.
  - [130] Lennart Ljung. *System Identification*. Prentice Hall, 2nd edition, 1999.
  - [131] Eugene Lukacs. *Characteristic Functions*. Charles Griffin & Co., Ltd., 2nd edition, 1970.
  - [132] Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. New York Springer, Berlin, 2005.
  - [133] Mark Machina and John Pratt. Increasing risk: Some direct constructions. *Journal of Risk and Uncertainty*, 14(2):103–127, 1997.
  - [134] S. G. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier/Academic Press, Amsterdam Boston, 2009.
  - [135] Rosario N. Mantegna and H. Eugene Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, Cambridge, UK New York, 1999.
  - [136] Vance Martin, Stan Hurn, and David Harris. *Econometric Modelling with Time Series: Specification, Estimation and Testing*. Cambridge University Press, Cambridge, 2013.
  - [137] William Q. Meeker, Gerald J. Hahn, and Luis A. Escobar. *Statistical Intervals: A Guide for Practitioners and Researchers*. Wiley, 2nd edition, 2017.
  - [138] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.
  - [139] Michael B. Miller. *Mathematics and Statistics for Financial Risk Management*. Wiley, 2013.
  - [140] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018.
  - [141] Stanley Mulaik. *Foundations of Factor Analysis*. Taylor and Francis, Boca Raton, FL, 2nd edition, 2009.
  - [142] Rémi Munos. *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*. Now Publishers, Hanover, Massachusetts, 2014.
-

- [143] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press Ltd, 2012.
- [144] Salih N. Neftci. *An Introduction to the Mathematics of Financial Derivatives*. Academic Press, 2000.
- [145] Roger B. Nelsen. *An Introduction to Copulas*. Springer, 1st edition, 1999.
- [146] Abraham Neyman and Sylvain Sorin, editors. *Stochastic Games and Applications*. Springer, 2003.
- [147] V. V. Nguyen and E. F. Wood. Review and unification of linear identifiability concepts. *SIAM Review*, 24(1):34–51, jan 1982.
- [148] Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.
- [149] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.
- [150] Daniel P. Palomar and Sergio Verdu. Lautum information. *IEEE Transactions on Information Theory*, 54(3):964–975, mar 2008.
- [151] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Boston, 4th edition, 2002.
- [152] Yudi Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford New York, 2001.
- [153] Georg Pflug. *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Kluwer Academic, Boston, Mass, 1996.
- [154] P. C. B. Phillips. Exact small sample theory in the simultaneous equations model. In *Handbook of Econometrics: Volume 1*, pages 449–516. Elsevier, 1983.
- [155] Rik Pintelon and Johan Schoukens. *System Identification: A Frequency Domain Approach*. Wiley IEEE Press, Hoboken, N.J. Piscataway, NJ, 2nd edition, 2012.
- [156] Warren Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, Hoboken, N.J, 2nd edition, 2011.
- [157] Warren B. Powell and Ilya O. Ryzhov. *Optimal Learning*. Wiley, Hoboken, New Jersey, 2012.
- [158] Friedrich Pukelsheim. The three sigma rule. *The American Statistician*, 48(2):88, may 1994.
- [159] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.
- [160] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT University Press Group Ltd, 2006.
- [161] Sidney Resnick. *A Probability Path*. Birkhäuser, New York, 1999.
- [162] Brian Ripley. *Stochastic Simulation*. Wiley, New York, 1987.
- [163] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Radar Library, 2004.

- 
- [164] Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Verlag, New York, 2nd edition, 2001.
- [165] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 1st edition, 1999.
- [166] Christian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Springer Verlag, 2010.
- [167] Michael Rockinger and Eric Jondeau. Entropy densities with an application to autoregressive conditional skewness and kurtosis. *Journal of Econometrics*, 106(1):119–142, jan 2002.
- [168] Sheldon Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, 1983.
- [169] Sheldon M. Ross. *Introductory Statistics*. Academic Press, 4th edition, 2017.
- [170] Sheldon M. Ross. *Introduction to Probability Models*. Academic Press, 12th edition, 2019.
- [171] Y. A. Rozanov. *Probability Theory: A Concise Course*. Dover Publications, 1977.
- [172] David Ruppert. *Statistics and Finance: An Introduction*. Springer, 2004.
- [173] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 3rd global edition, 2016.
- [174] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. *A Tutorial on Thompson Sampling*. now publishers inc., Boston Delft, 2018.
- [175] Neil Salkind, editor. *Encyclopedia of Measurement and Statistics*. SAGE Publications, Thousand Oaks, Calif, 2007.
- [176] Mark Schervish. *Theory of Statistics*. Springer New York, New York, NY, 1995.
- [177] E Seneta. *Non-negative Matrices and Markov Chains*. Springer, New York, 2006.
- [178] Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer, 2007.
- [179] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- [180] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics Mathematical Optimization Society, Philadelphia, Pennsylvania, 2nd edition, 2014.
- [181] M. Sharafi and J. Behboodian. The balakrishnan skew-normal density. *Statistical Papers*, 49(4):769–778, jan 2007.
- [182] Steven Shreve. *Stochastic Calculus for Finance I: The Binomial Asset Pricing Model*. Springer, 2004.
- [183] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, 4th edition, 2017.
- [184] Dan Simon. *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley-Interscience, 2006.
- [185] Aleksandrs Slivkins. *Introduction to Multi-Armed Bandits*. Now Publishers, 2019.

- [186] Christopher G. Small. *Expansions and Asymptotics for Statistics*. Chapman and Hall/CRC, 2010.
  - [187] Torsten Soderstrom and Petre Stoica. *System Identification*. Prentice Hall, 1989.
  - [188] T. T. Soong. *Random Differential Equations in Science and Engineering*. Academic Press, New York, 1973.
  - [189] James C. Spall. *Introduction to Stochastic Search and Optimization*. Wiley-Interscience, 2003.
  - [190] Henry Stark and John W. Woods. *Probability and Random Processes with Applications to Signal Processing*. Prentice Hall, 2001.
  - [191] Robert Stengel. *Optimal Control and Estimation*. Dover Publications, New York, 1994.
  - [192] James H. Stock and Mark W. Watson. *Introduction to Econometrics*. Pearson, Boston, updated 3rd edition, 2015.
  - [193] James H. Stock and Mark W. Watson. *Introduction to Econometrics*. Pearson, New York, NY, 4th edition, 2020.
  - [194] Petre Stoica. *Spectral Analysis of Signals*. Prentice Hall, 2005.
  - [195] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2nd edition, 2018.
  - [196] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Elsevier Academic Press, London San Diego, 1st edition, 2015.
  - [197] Charles W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1992.
  - [198] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. The MIT Press, 2005.
  - [199] Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.
  - [200] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley-Interscience, 1st edition, 2002.
  - [201] Ferdinand van der Heijden, Robert P. Duin, Dick de Ridder, and David M. J. Tax. *Classification, Parameter Estimation, and State Estimation : An Engineering Approach Using MATLAB*. Wiley, Chichester, West Sussex, Eng. Hoboken, NJ, 1st edition, 2004.
  - [202] John van der Hoek and Robert J. Elliott. *Binomial Models in Finance*. Springer, New York, NY, 2006.
  - [203] John van der Hoek and Robert J. Elliott. *Introduction to Hidden Semi-Markov Models*. Cambridge University Press, 2018.
  - [204] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
  - [205] Marno Verbeek. *A Guide to Modern Econometrics*. Wiley, Hoboken, NJ, 5th edition, 2017.
  - [206] Michel Verhaegen and Vincent Verdult. *Filtering and System Identification: A Least Squares Approach*. Cambridge University Press, 2007.
-

- [207] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.
- [208] Johannes Voit. *The Statistical Mechanics of Financial Markets*. Springer, Berlin New York, 2005.
- [209] Martin Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, United Kingdom New York, NY, USA, 2019.
- [210] Ruye Wang. *Introduction to Orthogonal Transforms: With Applications in Data Processing and Analysis*. Cambridge University Press, Cambridge, 2012.
- [211] Larry Wasserman. *All of Nonparametric Statistics*. Springer-Verlag GmbH, 2006.
- [212] Larry Wasserman. *All of Statistics*. Springer New York, 2013.
- [213] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1, jan 1982.
- [214] Peter Whittle. *Probability via Expectation*. Springer, 2000.
- [215] Samuel S. Wilks. *Mathematical Statistics*. Nabu Press, 1962.
- [216] David Williams. *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press, Cambridge New York, 2001.
- [217] Paul Wilmott. *Paul Wilmott Introduces Quantitative Finance*. Wiley, 2nd edition, 2007.
- [218] Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 2013.
- [219] Roy D. Yates and David J. Goodman. *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*. JOHN WILEY & SONS INC, 2nd edition, 2005.
- [220] Roy D. Yates and David J. Goodman. *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*. Wiley, 3rd edition, 2014.
- [221] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 2012.
- [222] Eric Zivot and Jiahui Wang. *Modeling Financial Time Series with S-PLUS®*. Springer, New York, NY, 2nd edition, 2006.