

# **Math Review Notes**

Gregory Faletto

## Contents

<b>1</b>	<b>Introduction</b>	<b>12</b>
<b>2</b>	<b>Linear Algebra</b>	<b>13</b>
2.1	Properties of Projection Matrices . . . . .	13
2.2	Eigenvalues, Eigenvectors, Diagonalization, Symmetric Matrices . . . . .	14
2.3	Positive Definite Matrices . . . . .	15
2.4	Matrix Decompositions . . . . .	16
2.4.1	Singular value decomposition and Pseudo-inverse . . . . .	16
2.4.2	Principal Components . . . . .	20
2.4.3	Low rank matrix approximation . . . . .	20
2.5	Inverting Matrices . . . . .	20
2.6	Other . . . . .	20
2.7	Practice Problems . . . . .	21
<b>3</b>	<b>Calculus</b>	<b>26</b>
3.1	Differentiation and common derivatives and integrals to know . . . . .	26
3.2	Matrix Differentiation . . . . .	26
3.3	Some theorems in higher dimensions . . . . .	28
3.4	Optimizing functions of several variables . . . . .	30
3.5	Lagrange Multipliers . . . . .	30
3.6	Line Integrals . . . . .	30
3.7	Miscellaneous . . . . .	31
3.8	Practice Problems . . . . .	32
<b>4</b>	<b>Differential Equations</b>	<b>38</b>
<b>5</b>	<b>Real Analysis</b>	<b>39</b>
5.1	Midterm 1 . . . . .	39
5.1.1	Homework 1 . . . . .	39

---

5.1.2	Homework 2 . . . . .	42
5.2	Midterm 2 . . . . .	45
5.2.1	Homework 3 . . . . .	45
5.2.2	Homework 4 . . . . .	46
5.3	Final . . . . .	47
5.3.1	Homework 5 . . . . .	47
5.3.2	Homework 6 . . . . .	50
5.4	More Theorems . . . . .	51
5.5	Problems from Practice Math GRE Subject Tests . . . . .	51
<b>6</b>	<b>Probability</b>	<b>54</b>
6.1	To Know for Math 505A Midterm 1 (Discrete Random Variables) . . . . .	54
6.1.1	Definitions . . . . .	54
6.1.2	Conditioning . . . . .	56
6.1.3	Convolution . . . . .	62
6.1.4	Compound Random Variables . . . . .	64
6.1.5	Odds and Ends . . . . .	65
6.1.6	Methods for Calculating Quantities . . . . .	69
6.1.7	Discrete Random Variable Distributions . . . . .	72
6.1.8	Indicator Method . . . . .	78
6.1.9	Linear transformations of random variables . . . . .	78
6.1.10	Poisson Paradigm (Poisson approximation for indicator method) . . . . .	78
6.1.11	Asymptotic Distributions . . . . .	80
6.2	Worked problems . . . . .	80
6.2.1	Example Problems That Will Likely Appear on Midterm (and Final) . . . . .	80
6.2.2	Problems we did in class that professor mentioned . . . . .	84
6.2.3	Problems we did on homework . . . . .	90
6.2.4	DSO Statistics Group Screening Exam Problems . . . . .	108
6.3	To Know for Math 505A Midterm 2 . . . . .	112

---

6.3.1	Definitions . . . . .	112
6.3.2	Probability-Generating Functions . . . . .	114
6.3.3	Moment-Generating Functions . . . . .	115
6.3.4	Characteristic Functions . . . . .	115
6.3.5	Continuous Random Variable Distributions . . . . .	116
6.3.6	More on Exponential Random Variables . . . . .	127
6.3.7	Multivariate Gaussian (Normal) Distributions . . . . .	131
6.4	Exponential Families . . . . .	133
6.4.1	Differential identities (Generalization of Moment-Generating Functions) . . . . .	137
6.5	KL Divergence (DSO 607) . . . . .	142
6.6	Worked problems . . . . .	142
6.6.1	Example Problems That Will Likely Appear on Midterm (and Final) . . . . .	142
6.6.2	More Problems From Homework . . . . .	148
6.7	Random Matrix Theory . . . . .	154
6.7.1	Large Deviation Theory . . . . .	155
6.7.2	Band Matrix . . . . .	160
6.7.3	Subgaussian Random Variable . . . . .	161
6.7.4	Topic 1: Invertibility of Matrices . . . . .	164
6.7.5	Random Graphs . . . . .	166
<b>7</b>	<b>Stochastic Processes</b>	<b>168</b>
7.1	Preliminaries . . . . .	168
7.2	Poisson Processes . . . . .	168
7.2.1	Poissonization Trick . . . . .	174
7.2.2	Time Sampling Poisson Processes . . . . .	179
7.2.3	Nonstationary Poisson Processes . . . . .	180
7.2.4	Queueing Systems . . . . .	185
7.3	Renewal Processes . . . . .	186
7.3.1	Stopping Times . . . . .	193

---

7.4 ISE 620 Midterm Solutions . . . . .	200
7.5 Renewal Reward Processes (Section 4.7 in <i>Introduction to Probability Models</i> , 3.6 <i>Stochastic Processes</i> ) . . . . .	203
7.5.1 Alternating Renewal Processes (Section 7.5.1 in <i>Introduction to Probability Models</i> . . .	212
7.5.2 Equilibrium renewal processes . . . . .	217
7.5.3 Regenerative Processes . . . . .	221
7.5.4 Example of Renewal Reward Processes to Patterns . . . . .	221
7.6 Markov Chains (Chapter 4 of <i>Stochastic Processes</i> ; Chapter 4 of <i>Introduction to Probability Models</i> ) . . . . .	222
7.6.1 Chapman-Kolmogorov Equations—section 4.2 of <i>Introduction to Probability Models</i> , section 4.2 of <i>Stochastic Processes</i> (p. 178 of pdf) . . . . .	225
7.6.2 Classification of States (Section 4.3 of <i>Introduction to Probability Models</i> ) . . . . .	226
7.6.3 Long-Run Proportions and Limiting Probabilities (Limit Theorems) (4.4 in <i>Introduction to Probability Models</i> , 4.3 in <i>Stochastic Processes</i> . . . . .	230
7.6.4 Branching Processes (Section 4.7 of <i>Introduction to Probability Models</i> , Section 4.5 of <i>Stochastic Processes</i> . . . . .	237
7.6.5 A Markov Chain Model of Algorithmic Efficiency (Section 4.6.1 of <i>Stochastic Processes</i> ) . . . . .	238
7.6.6 Time Reversible Markov Chains (Section 4.8 of <i>Introduction to Probability Models</i> , Section 4.7 of <i>Stochastic Processes</i> . . . . .	240
7.6.7 Semi-Markov Processes (Section 4.8 of <i>Stochastic Processes</i> (p. 224 of pdf), Section 7.6 of <i>Introduction to Probability Models</i> . . . . .	241
7.7 ISE 620 . . . . .	242
7.8 DSO Statistics Group Screening Exam Problems . . . . .	244
7.9 Simple Random Walk . . . . .	246
7.10 Martingales . . . . .	246
7.11 Brownian Motion . . . . .	247
<b>8 Asymptotics and Convergence</b>	<b>249</b>
8.1 Preliminaries (5.9 and 7.1, Grimmett and Stirzaker) . . . . .	249
8.2 Inequalities (8.6 of Pesaran) . . . . .	250
8.3 Modes of Convergence (7.2 of Grimmett and Stirzaker, 8.2 and 8.4 of Pesaran) . . . . .	255

---

8.4 More on convergence (7.2 of Grimmet and Stirzaker) . . . . .	259
8.4.1 Slutsky's Convergence Theorems (8.4.1 of Pesaran, 7.3 of Grimmett and Stirzaker) . .	263
8.5 Stochastic orders $\mathcal{O}_p(\cdot)$ and $o_p(\cdot)$ (Pesaran 8.5) . . . . .	264
8.6 Laws of Large Numbers and Central Limit Theorems (Pesaran 8.6; Grimmett and Stirzaker 7.4, 7.5) . . . . .	265
8.7 The case of dependent and heterogeneously distributed observations (Pesaran 8.8) . . . .	267
8.8 Worked Examples from Math 505A Midterm 2 . . . . .	268
8.9 Estimators and Central Limit Theorems (DSO 607) . . . . .	273
<b>9 Convex Optimization</b>	<b>274</b>
9.1 Convex Functions . . . . .	274
9.2 Schur Complement Trick . . . . .	285
9.2.1 Definition . . . . .	285
9.2.2 The Trick . . . . .	286
9.2.3 Example 1: Last Year's Final, Question 2(b) . . . . .	286
9.2.4 Example 2: Last Year's Final, Question 4(b) . . . . .	287
9.3 Duality . . . . .	287
9.4 MLE estimates . . . . .	288
9.5 Practice Final (2017 Final) . . . . .	288
<b>10 Mathematical Statistics</b>	<b>293</b>
10.1 Order Statistics . . . . .	293
10.2 Random Samples . . . . .	297
10.2.1 The Delta Method . . . . .	309
10.2.2 Simulation of Random Variables . . . . .	312
10.3 Data Reduction . . . . .	314
10.3.1 Sufficient Statistics . . . . .	314
10.3.2 Minimal Sufficient Statistics . . . . .	318
10.3.3 Ancillary Statistics . . . . .	324
10.3.4 Complete Statistics . . . . .	326

---

10.4 Point Estimation . . . . .	332
10.4.1 Heuristic Principles for Finding Good Estimators . . . . .	332
10.4.2 Evaluating Estimators . . . . .	333
10.4.3 Efficiency of an Estimator . . . . .	340
10.4.4 Bayes Estimation . . . . .	344
10.4.5 Method of Moments . . . . .	351
10.4.6 Maximum likelihood estimator . . . . .	352
10.4.7 Bayes estimator . . . . .	364
10.4.8 EM Algorithm . . . . .	364
10.4.9 Comparison of estimators . . . . .	364
10.5 Resampling and Bias Reduction . . . . .	364
10.5.1 Jackknife Resampling . . . . .	364
10.5.2 Bootstrapping . . . . .	365
10.6 Some Concentration of Measure . . . . .	365
10.6.1 Concentration for Independent Sums . . . . .	365
10.7 Math 541B . . . . .	366
10.8 Hypothesis Testing . . . . .	367
10.8.1 Neyman-Pearson Tests . . . . .	370
10.8.2 Consistency of Neyman-Pearson Tests . . . . .	376
10.8.3 Composite Hypothesis Testing . . . . .	378
<b>11 Linear Regression</b>	<b>392</b>
11.1 Chapter 1: Linear Regression . . . . .	392
11.1.1 Preliminaries . . . . .	392
11.1.2 Estimation . . . . .	392
11.2 Chapter 2: Multiple Regression . . . . .	398
11.3 Chapter 3: Hypothesis testing in regression . . . . .	403
11.3.1 ANOVA . . . . .	406
11.4 Chapter 4: Heteroskedasticity . . . . .	406

---

11.5 Chapter 5: Autocorrelated disturbances . . . . .	407
11.6 Quantile Regression . . . . .	408
11.6.1 Detecting outliers in multiple dimensions . . . . .	408
11.7 Transformed Linear Models . . . . .	409
11.7.1 Transformations of response . . . . .	409
11.7.2 Transforming predictor values . . . . .	410
11.8 Segmented regression, local regression, splines . . . . .	410
11.8.1 Local Regression . . . . .	411
11.8.2 Curse of Dimensionality (brief discussion) . . . . .	411
11.9 Dimension Reduction methods . . . . .	411
11.9.1 Principal components regression . . . . .	411
11.9.2 Partial least squares . . . . .	412
11.9.3 Dimension reduction by random matrix . . . . .	412
11.10 Goodness of fit, residuals, residual diagnostics, leverage . . . . .	413
11.10.1 Residual diagnostics . . . . .	413
11.11 DSO 607 . . . . .	414
11.11.1 Akaike Information Criterion (AIC) . . . . .	414
11.11.2 Bayesian Information Criterion (BIC) . . . . .	416
11.12 Ridge Regression . . . . .	419
11.13 Lasso . . . . .	422
11.13.1 Soft Thresholding . . . . .	424
11.13.2 Lasso theory . . . . .	425
11.13.3 Non-Negative Garotte . . . . .	431
11.13.4 LARS—Preliminaries and Intuition . . . . .	431
11.13.5 LARS . . . . .	433
11.14 Loss Functions . . . . .	435
11.14.1 Feature Selection properties . . . . .	436
11.15 Dantzig Selector . . . . .	440

---

11.16 Coordinate Descent . . . . .	441
11.17 Total Variational Distance . . . . .	441
11.18 Non-parametric regression . . . . .	441
11.18.1 Generalized additive models . . . . .	441
11.19 Mixture regression . . . . .	442
11.20 Missing observations . . . . .	442
11.21 Generalized linear models . . . . .	442
<b>12 Time Series</b>	<b>446</b>
12.1 Chapter 6: ARDL Models . . . . .	446
12.2 Chapters 12 and 13: Intro to Stochastic Processes and Spectral Analysis . . . . .	446
12.2.1 Worked Examples . . . . .	448
12.3 Some time series and their properties . . . . .	449
12.3.1 White noise process: . . . . .	449
12.3.2 MA(1) process: . . . . .	449
12.3.3 MA( $\infty$ ) process: . . . . .	450
12.3.4 AR(1) process: . . . . .	450
12.3.5 AR(2) process: . . . . .	451
12.3.6 AR( $p$ ) process: . . . . .	452
12.3.7 ARMA(1, 1) process: . . . . .	452
12.4 Chapter 14: Estimation of Stationary Time Series Processes . . . . .	454
12.4.1 Sufficient conditions for ergodicity of mean. (Book section 14.2.1) . . . . .	454
12.4.2 Estimation of autocovariances (Book section 14.2.2). . . . .	455
12.4.3 Worked examples . . . . .	456
12.5 Chapter 15: Unit Root Processes . . . . .	462
12.5.1 Worked Problems . . . . .	462
12.6 Chapter 17: Introduction to Forecasting . . . . .	469
12.6.1 17.7: Iterated and direct multi-step AR methods . . . . .	469
12.6.2 Worked Problems . . . . .	470

---

12.7 Chapter 18: Measurement and Modeling of Volatility . . . . .	473
12.7.1 Higher order GARCH models (Pesaran Section 18.4.2) . . . . .	473
12.7.2 Testing for GARCH effects (Pesaran Section 18.5.1) . . . . .	474
12.7.3 Worked Problems . . . . .	474
12.8 Chapter 21: Vector Autoregressive Models . . . . .	477
12.8.1 Worked Problems . . . . .	477
12.9 Chapter 22: Cointegration Analysis . . . . .	482
12.9.1 22.4 Cointegrating VAR: multiple cointegrating relations and 22.5: Identification of long-run effects . . . . .	483
12.9.2 Worked Problems . . . . .	484
12.10 Chapter 23: VARX Modeling . . . . .	484
12.10.1 Worked Problems . . . . .	484
12.11 Chapter 24: Impulse Response Analysis . . . . .	484
12.11.1 Worked Problems . . . . .	484
12.12 Chapter 33: Theory and Practice of GVAR Modeling . . . . .	487
<b>13 Statistical Learning</b>	<b>488</b>
13.1 Math 547 . . . . .	488
13.1.1 Perceptron Algorithm . . . . .	488
13.1.2 Mercer's Theorem . . . . .	488
13.2 Norms . . . . .	488
<b>14 Abstract Algebra</b>	<b>490</b>
14.1 Chapter 1: Binary Operations . . . . .	490
14.2 Chapter 2: Groups . . . . .	493
14.3 Chapter 3: The Symmetric Groups . . . . .	494
14.4 Chapter 4: Subgroups . . . . .	497
14.5 Chapter 5: The Group of Units of $\mathbb{Z}_n$ . . . . .	499
14.6 Chapter 6: Direct Products of Groups . . . . .	499
14.7 Chapter 7: Isomorphism of Groups . . . . .	500

---

14.8 Chapter 8: Cosets and Lagrange's Theorem . . . . .	502
14.9 Chapter 9: Introduction to Ring Theory . . . . .	504
<b>15 Miscellaneous</b>	<b>506</b>
15.1 Set Theory . . . . .	506
15.2 Other . . . . .	506

Last updated October 5, 2019

## 1 Introduction

These are notes I've collected on various math topics. I originally created this document to prepare for the GRE Math Subject test. Since then I've expanded it as I've reviewed concepts from past classes and reinforced concepts from new classes. This document is very much a work in progress, with many typos, omissions to be filled in, and probably errors. Nonetheless, I share this document in case it's useful to anyone else as a reference.

I use many sources throughout this document, which I either cite at the beginning of the section (for sources I use broadly) or as I use them (for sources I use for one or two isolated results).

## 2 Linear Algebra

These are my notes from taking EE 588 at USC, Math 541A at USC, and various other sources which I mostly cite within the text.

### 2.1 Properties of Projection Matrices

- i. Formula:

$$P = A(A^T A)^{-1} A^T$$

(Note that if  $A$  is an invertible (square) matrix, then  $P = A(A^T A)^{-1} A^T = AA^{-1}(A^T)^{-1} A^T = I$ .)

**The projection matrix projects any vector  $b$  into the column space of  $A$ .** In other words,  $p = Pb$  is the component of  $b$  in the column space, and the error  $e = b - Pb$  is the component in the orthogonal complement. ( $I - P$  is also a projection matrix. It projects  $b$  onto the orthogonal complement, and the projection is  $b - Pb = e$ .)

(Note that if  $A$  is an invertible (square) matrix, then its column space is all of  $\mathbb{R}^n$ , so  $b$  is already in the column space of  $A$ .)

- ii. The projection matrix is **idempotent**: it equals its square— $P^2 = P$ .
- iii. The projection matrix is **symmetric**: it equals its transpose— $P^T = P$ .
- iv. Conversely, **any symmetric idempotent matrix represents a projection**.  $P$  is unique for a given subspace.
- v. If  $A$  is an  $m \times n$  matrix with rank  $n$ , then  $\text{rank}(P) = n$ . The eigenvalues of  $P$  consist of  $n$  ones and  $m - n$  zeroes.  $P$  always contains  $n$  independent eigenvectors and is thus diagonalizable.

Suppose  $A$  is a square nonsingular matrix and  $\lambda$  is an eigenvalue of  $A$ . Then  $\lambda^{-1}$  is an eigenvalue of the matrix  $A^{-1}$ .

**Proposition 2.1.1.** The eigenvalues of an idempotent matrix equal either 0 or 1.

*Proof.* Let  $A$  be an idempotent matrix and let  $v$  be a unit-length eigenvector of  $H$  with corresponding eigenvalue  $\lambda$ . Then

$$Av = \lambda v \implies A(Av) = A(\lambda v).$$

Then

$$AAv = Av = \lambda v$$

and

$$\lambda Av = \lambda^2 v$$

so we have

$$\lambda v = \lambda^2 v \implies \lambda \in \{0, 1\}.$$

□

The trace of an idempotent matrix with rank  $r$  is  $r$ .

## 2.2 Eigenvalues, Eigenvectors, Diagonalization, Symmetric Matrices

### Notes on Diagonalization

Suppose the  $n \times n$  matrix  $A$  has  $n$  linearly independent eigenvectors. If these eigenvectors are the columns of a matrix  $S$ , then  $S^{-1}AS$  is a diagonal matrix  $\Lambda$ . The eigenvalues of  $A$  are on the diagonal of  $\Lambda$ :

$$S^{-1}AS = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

We call  $S$  the **eigenvector matrix** and  $\Lambda$  the **eigenvalue matrix**.

1. If the matrix  $A$  has no repeated eigenvalues, then its  $n$  eigenvectors are automatically independent. Therefore **any matrix with  $n$  distinct eigenvalues can be diagonalized**.
2. **The diagonalizing matrix  $S$  is not unique.** An eigenvector  $x$  can be multiplied by a constant and remains an eigenvector. We can multiply the columns of  $S$  by any nonzero constants and produce a new diagonalizing  $S$ . Repeated eigenvalues leave even more freedom in  $S$  (columns with identical eigenvalues can be interchanged).

(Note that for the trivial example  $A = I$ , any invertible  $S$  will do.  $S^{-1}IS$  is always diagonal, and  $\Lambda$  is just  $I$ . **All vectors are eigenvectors of the identity.**)

3. **Other matrices  $S$  will not produce a diagonal  $\Lambda$ .** Since  $\Lambda = S^{-1}AS$ ,  $S$  must satisfy  $S\Lambda = AS$ . Suppose the first column of  $S$  is  $y$ . Then the first column of  $S\Lambda$  is  $\lambda_1 y$ . If this is to agree with the first column of  $AS$ , which by matrix multiplication is  $Ay$ , then  $y$  must be an eigenvector:  $Ay = \lambda_1 y$ .

(Note that the *order* of the eigenvectors in  $S$  and the eigenvalues in  $\Lambda$  must match.)

4. Not all matrices possess  $n$  linearly independent eigenvectors, so **not all matrices are diagonalizable**.

**Diagonalizability of  $A$  depends on having enough ( $n$ ) independent eigenvectors. Invertibility of  $A$  depends on having nonzero eigenvalues.**

There is no connection between diagonalizability ( $n$  independent eigenvectors) and invertibility (no zero eigenvalues). The only indication given by the eigenvalues is that diagonalization can fail only if there are repeated eigenvalues. (But even then, it does not always fail—e.g.  $I$ .)

The test is to check, for an eigenvalue that is repeated  $p$  times, whether there are  $p$  independent eigenvectors—in other words, whether  $A - \lambda$  has rank  $n - p$ .

5. **Projection matrices always contain  $n$  independent eigenvectors and thus are always diagonalizable.**

**Eigenvalues of Symmetric Matrices:** If  $A$  is symmetric, then it has the following properties:

1.  $A$  has exactly  $n$  (not necessarily distinct) eigenvalues
2. There exists a set of  $n$  eigenvectors, one for each eigenvalue, that are mutually orthogonal (even if the eigenvalues are not distinct).

**Eigenvalues of the Inverse of a Matrix:** Suppose  $A$  is a square nonsingular matrix and  $\lambda$  is an eigenvalue of  $A$ . Then  $\lambda^{-1}$  is an eigenvalue of the matrix  $A^{-1}$ . Proof: Note that since  $A$  is nonsingular,  $A^{-1}$  exists and  $\lambda$  is nonnegative for all eigenvalues of  $A$ . Let  $\lambda$  be an eigenvalue of  $A$  and let  $x \neq 0$  be an eigenvector of  $A$  for  $\lambda$ . Suppose  $A$  is  $n$  by  $n$ . Then we have

$$A^{-1}x = A^{-1}\lambda^{-1}\lambda x = \lambda^{-1}A^{-1}\lambda x = \lambda^{-1}A^{-1}Ax = \lambda^{-1}x$$

**The inverse of a symmetric matrix is symmetric.** Proof: Let  $A$  be a symmetric matrix.

$$I = I'$$

$$AA^{-1} = (AA^{-1})'$$

$$A^{-1}A = (A^{-1})'A'$$

$$A^{-1}AA^{-1} = (A^{-1})'AA^{-1}$$

$$A^{-1} = (A^{-1})'$$

### 2.3 Positive Definite Matrices

**Proposition 2.3.1.** For any real matrix  $A \in \mathbb{R}^{m \times n}$  with  $m \geq n$ , the product  $A^T A$  is a positive semidefinite matrix.

*Proof.* Let  $z$  be a non-zero vector.  $A^T A$  is positive semidefinite if and only if  $z^T A^T A z \geq 0 \forall z \in \{\mathbb{R}^n \setminus \{0\}\}$ . Note that  $z^T A^T A z = (Az)^T (Az) = \|Az\|_2^2 \geq 0$ .

□

**Proposition 2.3.2.** For any real invertible matrix  $A \in \mathbb{R}^{n \times n}$ , the product  $A^T A$  is a positive definite matrix.

*Proof.* Let  $z$  be a non-zero vector.  $A'A$  is positive definite if and only if  $z'A'Az > 0 \forall z \in \{\mathbb{R}^n \setminus \{0\}\}$ . Note that  $z'A'Az = (Az)'(Az)$ . Because  $A$  is invertible and  $z \neq 0$ ,  $Az \neq 0$ , so  $(Az)'(Az) = \|Az\|_2^2 > 0$ .  $\square$

**Proposition 2.3.3 (Math 547 Homework Problem).** Let  $A$  be an  $m \times n$  real matrix with  $m \geq n$ . Then  $A$  has rank  $n$  if and only if  $A^T A$  is positive definite.

*Proof.* By Proposition 2.3.1,  $A^T A$  is positive semidefinite. Since  $A^T A$  is not invertible if one of its eigenvalues equals 0,  $A^T A$  is invertible (and therefore full rank) if and only if it is positive definite. But  $A^T A \in \mathbb{R}^{n \times n}$  is full rank if and only if the columns of  $A$  are linearly independent, for the following reason: multiplying  $A^T A$  means taking linear combinations of the rows of  $A^T$  (that is, the columns of  $A$ ) one at a time, with weightings according to the columns of  $A$ . These linear combinations will themselves be linearly independent if and only if the rows of  $A^T$  (the columns of  $A$ ) are linearly independent and for the basis of an  $n$ -dimensional subspace. And  $A^T A$  has linearly independent columns if and only if it is invertible.

So we have shown that  $A^T A$  is positive definite if and only if the columns of  $A$  are linearly independent. That itself is true if and only if  $A$  has rank  $n$ . Therefore  $A$  has rank  $n$  if and only if  $A^T A$  is positive definite.

$\square$

Every positive definite matrix is invertible and its inverse is also positive definite.

## 2.4 Matrix Decompositions

**Schur complement, Schur decomposition:**

**Proposition 2.4.1 (Schur complement formula).**

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

For information, see Section 9.2.

**QR decompositon**

**Orthogonal Decomposition**

**Spectral Decomposition (eigenvalue decomposition)**

**Generalized eigenvalue decomposition**

**Jordan decomposition**

**Cholesky decomposition**

### 2.4.1 Singular value decomposition and Pseudo-inverse

Suppose  $X \in \mathbb{R}^{n \times p}$  with  $\text{rank}(X) = r \leq \min\{n, p\}$ . Note that  $X^T X$  is positive semi-definite by Proposition 2.3.1. Therefore by the Spectral Decomposition Theorem, we can find a factorization

$$X^T X = V D V^T \quad (2.1)$$

where the columns of  $V$  are orthonormal eigenvectors of  $X^T X$  and  $D$  is a diagonal matrix containing the eigenvalues of  $X^T X$  on the diagonal. Since these eigenvalues are nonnegative, let  $D = \Sigma^2$  for the positive semi-definite diagonal matrix  $\Sigma := D^{1/2}$ . Note that  $\text{rank}(X^T X) = r \leq p$ , so  $D$  contains  $r$  nonzero entries.

If  $r < p$  then  $X^T X$  will only have  $r < p$  eigenvectors, so  $r$  columns of  $V$  will form an orthonormal basis for the  $r$ -dimensional column space of  $X^T X$ . But then there exist  $p - r$  vectors that are orthogonal to each other and the first  $r$  vectors of  $V$ , such that if these vectors form the last  $p - r$  columns of  $V$ , then  $V$  is still orthogonal and its columns still span  $\mathbb{R}^p$ . Further, the factorization (2.1) still holds since in this case these columns will be multiplied by the 0 entries in  $D$  anyway. Lastly, since these vectors are orthogonal to a basis for the column space of  $X^T X$ , these vectors lie in the null space of  $X^T X$ . That is, for any one of these columns  $v_j$ ,

$$X^T X v_j = 0_{p \times 1} \implies v_j^T X^T X v_j = 0 \iff (Xv_j)^T X v_j = 0 \iff X v_j = 0_{p \times 1}$$

so  $v_j$  also lies in the nullspace of  $X$ . Next, consider the matrix  $XV \in \mathbb{R}^{n \times p}$ . We know that

$$\left( \underbrace{X}_{n \times p} \underbrace{V}_{p \times p} \right)^T \underbrace{X}_{n \times p} \underbrace{V}_{p \times p} = \underbrace{V^T}_{p \times p} \underbrace{X^T}_{p \times n} \underbrace{X}_{n \times p} \underbrace{V}_{p \times p} = V^T V \Sigma^2 V^T V = \Sigma^2. \quad (2.2)$$

Let  $\tilde{D} \in \mathbb{R}^{r \times r}$  be a diagonal matrix containing the  $r$  nonzero entries of  $\Sigma$  so that we can write

$$\Sigma = \begin{bmatrix} \tilde{D} & 0_{r \times (p-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (p-r)} \end{bmatrix}$$

Then if we define

$$\tilde{\Sigma} := \begin{bmatrix} \tilde{D}^{-1} & 0_{r \times (p-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (p-r)} \end{bmatrix}$$

we can modify (2.2) as

$$(XV\tilde{\Sigma})^T XV\tilde{\Sigma} = \tilde{\Sigma} V^T X^T XV\tilde{\Sigma} = \tilde{\Sigma} V^T V \Sigma^2 V^T V \tilde{\Sigma} = \tilde{\Sigma} \Sigma^2 \tilde{\Sigma} = \begin{bmatrix} I_{r \times r} & 0_{r \times (p-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (p-r)} \end{bmatrix} \quad (2.3)$$

since

$$\tilde{\Sigma} \Sigma = \begin{bmatrix} \tilde{D}^{-1} & 0_{r \times (p-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (p-r)} \end{bmatrix} \begin{bmatrix} \tilde{D} & 0_{r \times (p-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (p-r)} \end{bmatrix} = \begin{bmatrix} I_{r \times r} & 0_{r \times (p-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (p-r)} \end{bmatrix} = \Sigma \tilde{\Sigma}.$$

That is,  $XV\tilde{\Sigma} \in \mathbb{R}^{n \times p}$  contains  $r$  orthogonal columns. By (2.2), the  $p - r$  remaining columns of  $XV$  equal 0. Let  $U := XV\tilde{\Sigma}$ . Then we have

$$U\Sigma = XV\tilde{\Sigma}\Sigma = XV \begin{bmatrix} I_{r \times r} & 0_{r \times (p-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (p-r)} \end{bmatrix} = XV \quad (2.4)$$

where the last step follows because the last  $p - r$  columns of  $V$  lie in the nullspace of  $X$ , so they already equaled 0 anyway. Finally, multiply both sides of (2.4) by  $V^T$  to obtain the **singular value decomposition**

$$X = U\Sigma V^T.$$

Now let's examine the properties of this decomposition. We argued earlier that the last  $p - r$  columns of  $V$  lie in the nullspace of  $X$ . It follows that since the first  $p$  columns of  $V$  are orthonormal and orthogonal to the nullspace of  $X$ , the first  $r$  columns of  $V$  are orthonormal basis vectors for the row space of  $X$ .

Next, we have already shown that  $U = XV\tilde{\Sigma}$  contains  $r$  orthogonal columns, and of course these columns lie in the column space of  $X$ . Therefore the first  $r$  columns of  $U$  form an orthonormal basis for the column space of  $X$ .

Now identity (2.4) yields a useful way of thinking about the singular value decomposition.  $V$  is a special orthonormal basis for the rowspace of  $X$  (perhaps with some extra orthogonal vectors spanning the nullspace of  $X$  if  $X$  does not have full column rank) with the property that when they are linearly transformed by  $X$ ,  $XV = U\Sigma$  forms  $r$  orthogonal vectors spanning the column space of  $X$  (plus  $p - r$  0 vectors if  $X$  doesn't have full column rank).

If we like, we can decompose  $U\Sigma$  into unit vectors  $U$  and a diagonal vector  $\Sigma$  containing the lengths of the orthogonal vectors  $U\Sigma$  on its diagonal. Then  $U$  is an orthonormal basis for the column space of  $X$ .

We call the diagonal entries of  $\Sigma$  the **singular values** of  $X$ :  $\sigma_i \in \mathbb{R}, i \in \{1, \dots, r\}$ .

The first  $r$  nonzero columns of  $U\Sigma$  are called the **principal components** of  $X$ . The columns of  $U \in \mathbb{R}^n$  are also called the **normalized principal components** of  $X$ . Sometimes we call the columns of  $V \in \mathbb{R}^p$  the **principal component directions** of  $X$ .

Lastly, although we have already examined the properties of  $X^T X$ , consider the decomposition

$$XX^T = \hat{U}\hat{D}\hat{U}^T \quad (2.5)$$

for some positive semidefinite diagonal matrix  $\hat{D} \in \mathbb{R}^{n \times n}$  and some orthonormal matrix  $\hat{U} \in \mathbb{R}^{n \times n}$ , which exists by the Spectral Theorem. Now consider expressing this in terms of the singular value decomposition:

$$XX^T = \underbrace{U}_{n \times p} \underbrace{\Sigma}_{p \times p} \underbrace{V^T}_{p \times p} (\underbrace{U}_{n \times p} \underbrace{\Sigma}_{p \times p} \underbrace{V^T}_{p \times p})^T = \underbrace{U}_{n \times p} \underbrace{\Sigma}_{p \times p} \underbrace{V^T}_{p \times p} \underbrace{V}_{p \times p} \underbrace{\Sigma}_{p \times p} \underbrace{U^T}_{p \times n} = \underbrace{U}_{n \times p} \underbrace{\Sigma^2}_{p \times p} \underbrace{U^T}_{p \times n}, \quad (2.6)$$

so we can form

$$\hat{U} = [U \quad u_{r+1} \quad \cdots \quad u_n]$$

(where  $u_{r+1}, \dots, u_n$  are orthonormal vectors completing an orthonormal basis for  $\mathbb{R}^n$  in the columns of  $\hat{U}$ ) and

$$\hat{D} = \begin{bmatrix} \Sigma^2 & 0_{p \times (n-p)} \\ 0_{(n-p) \times p} & 0_{(n-p) \times (n-p)} \end{bmatrix}$$

to yield (2.5) from (2.6).

⋮

**Math 547 explanation:** Suppose  $X \in \mathbb{R}^{n \times p}$ . Since  $X^T X \in \mathbb{R}^{p \times p}$ , by the Spectral Theorem, there exists an orthogonal matrix  $Q \in \mathbb{R}^{p \times p}$  and a real diagonal matrix  $D \in \mathbb{R}^{p \times p}$  such that

$$X^T X = Q^T D Q.$$

The entries of  $D$  (the eigenvalues of  $X^T X$ ) are nonnegative because if  $v \in \mathbb{R}^p$  is an eigenvector of  $X^T X$  with corresponding eigenvalue  $\lambda$ ,

$$\lambda \|v\|_2^2 = v^T (\lambda v) = v^T (X^T X v) = (Xv)^T (Xv) = \|Xv\|_2^2 \geq 0.$$

Similarly,  $XX^T \in \mathbb{R}^{n \times n}$  and there exists an orthogonal  $R \in \mathbb{R}^{n \times n}$  and a diagonal  $G \in \mathbb{R}^{n \times n}$  such that

$$XX^T = R^T G R.$$

Since  $D \succeq 0$ , we can write

$$X = \underbrace{R^T}_{n \times n} \underbrace{\sqrt{D}}_{p \times p} \underbrace{Q}_{p \times p}$$

⋮

$$X = \underbrace{U}_{N \times p} \underbrace{D}_{p \times p} \underbrace{V^T}_{p \times p}$$

since

$$(R^T \sqrt{D} Q)^T R^T \sqrt{D} Q = Q^T \sqrt{D} R R^T \sqrt{D} Q = Q^T D Q = X^T X$$

and

$$R^T \sqrt{D} Q (R^T \sqrt{D} Q)^T = R^T \sqrt{D} Q Q^T \sqrt{D} R = R^T D R = \dots = R^T G R = X X^T.$$

### 2.4.2 Principal Components

Again, the  $p$  columns of  $U\Sigma \in \mathbb{R}^n$  are the **principal components** of  $X$ . The first principal component direction  $v_1 \in \mathbb{R}^p$  satisfies the condition  $z_1 := Xv_1$  has the largest sample variance among all normalized linear combinations of the columns of  $X$ . That is,  $v_1$  satisfies

$$v_1 = \arg \max_{v \in \mathbb{R}^p, \|v\|_2=1} \text{Var}(Xv) = \arg \max_{v \in \mathbb{R}^p, \|v\|_2=1} v^T(X^T X)v$$

Note that  $z_1 = Xv_1 = \sigma_1 u_1$ , so  $z_1$  is the first principal component of  $X$  and  $u_1$  is the normalized first principal component.

### 2.4.3 Low rank matrix approximation

Note that using the singular value decomposition we can express  $X$  as a sum of rank 1 matrices:

$$X = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$$

An intuitive “best” way to approximate  $X$  as a rank  $k < r$  matrix is to take the first  $k$  of these terms:

$$X \approx X_k = \sum_{i=1}^k \sigma_i u_i v_i^T = U_k \Sigma_k V_k^T$$

where  $U_k \in \mathbb{R}^{n \times k}$  contains only the first  $k$  columns of  $U$  (corresponding to the  $k$  largest singular values),  $\Sigma_k \in \mathbb{R}^{k \times k}$  contains the  $k \times k$  top-left submatrix of  $\Sigma$ , and  $V_k \in \mathbb{R}^{p \times k}$  contains the first  $k$  columns of  $V$ . It turns out this is the best low-rank approximation of  $X$  in the sense that for any other rank- $k$  matrix  $B$  of size  $n \times p$ ,

$$\|X - X_k\|_F^2 \leq \|X - B\|_F^2.$$

## 2.5 Inverting Matrices

**Theorem 2.5.1 (Woodbury Matrix Identity (or Sherman-Morrison-Woodbury formula)).** For  $A \in \mathbb{R}^{n \times n}$ ,  $U \in \mathbb{R}^{n \times k}$ ,  $C \in \mathbb{R}^{k \times k}$ , and  $V \in \mathbb{R}^{v \times n}$ ,

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

**Theorem 2.5.2 (Binomial Inverse Thereom).**

## 2.6 Other

### Frobenius norm

From appendix of Time Series:

**Quadratic forms**

**Special matrices**

**Difference Equations**

## 2.7 Practice Problems

**[The Power Method]** This exercise gives an algorithm for finding the eigenvectors and eigenvalues of a symmetric matrix. In modern statistics, this is often a useful thing to do. The Power Method described below is not the best algorithm for this task, but it is perhaps the easiest to describe and analyze.

Let  $A$  be an  $n \times n$  real symmetric matrix. Let  $\lambda_1 \geq \dots \geq \lambda_n$  be the (unknown) eigenvalues of  $A$ , and let  $v_1, \dots, v_n \in \mathbb{R}^n$  be the corresponding (unknown) eigenvectors of  $A$  such that  $|v_i| = 1$  and such that  $Av_i = \lambda_i v_i$  for all  $1 \leq i \leq n$ .

Given  $A$ , our first goal is to find  $v_1$  and  $\lambda_1$ . For simplicity, assume that  $1/2 < \lambda_1 < 1$ , and  $0 \leq \lambda_n \leq \dots \leq \lambda_2 < 1/4$ . Suppose we have found a vector  $v \in \mathbb{R}^n$  such that  $|v| = 1$  and  $|\langle v, v_1 \rangle| > 1/n$ . Let  $k$  be a positive integer. Show that

$$A^k v$$

approximates  $v_1$  well as  $k$  becomes large. More specifically, show that for all  $k \geq 1$ ,

$$|A^k v - \langle v, v_1 \rangle \lambda_1^k v_1|^2 \leq \frac{n-1}{16^k}.$$

(Hint: use the spectral theorem for symmetric matrices.)

**Solution.** Since the eigenvectors for  $A$  are orthogonal, they form a basis for  $\mathbb{R}^n$ , so for any  $v \in \mathbb{R}^n$  we have  $v = \sum_{i=1}^n c_i v_i$  for some  $c = (c_1, \dots, c_n) \in \mathbb{R}^n$ . It also follows then that  $\langle v, v_1 \rangle = \langle \sum_{i=1}^n c_i v_i, v_1 \rangle = c_1 v'_1 v_1 = c_1$ . And finally, since  $\|v\| = 1$  and  $\|v_i\| = 1$  for all  $i$ , clearly we have  $-1 \leq c_i \leq 1$ . Using these facts, we have

$$\begin{aligned} \|A^k v - \langle v, v_1 \rangle \lambda_1^k v_1\|^2 &= \left\| \sum_{i=1}^n \lambda_i^k c_i v_i - \langle v, v_1 \rangle \lambda_1^k v_1 \right\|^2 = \left\| \sum_{i=1}^n \lambda_i^k c_i v_i - \lambda_1^k c_1 v_1 \right\|^2 = \left\| \sum_{i=2}^n \lambda_i^k c_i v_i \right\|^2 \\ &= \sum_{i=2}^n \lambda_i^{2k} c_i^2 v'_i v_i = \sum_{i=2}^n \lambda_i^{2k} c_i^2 \end{aligned}$$

Since by assumption  $0 \leq \lambda_n \leq \dots \leq \lambda_2 \leq 1/4$ ,  $\lambda_i^{2k} \leq 1/16^k$  for all  $i$ , so we have

$$\|A^k v - \langle v, v_1 \rangle \lambda_1^k v_1\|^2 \leq \frac{1}{16^k} \sum_{i=2}^n c_i^2$$

Since  $-1 \leq c_i \leq 1 \implies 0 \leq c_i^2 \leq 1$ , we have  $\sum_{i=2}^n c_i^2 \leq n-1$ , so this can be written as

$$\|A^k v - \langle v, v_1 \rangle \lambda_1^k v_1\|^2 \leq \frac{n-1}{16^k}$$

**Remark.** Since  $|\langle v, v_1 \rangle| \lambda_1^k > 2^{-k}/n$ , this inequality implies that  $A^k v$  is approximately an eigenvector of  $A$  with eigenvalue  $\lambda_1$ . That is, by the triangle inequality,

$$|A(A^k v) - \lambda_1(A^k v)| \leq |A^{k+1}v - \langle v, v_1 \rangle \lambda_1^{k+1} v_1| + \lambda_1 |\langle v, v_1 \rangle \lambda_1^k v_1 - A^k v| \leq 2 \frac{\sqrt{n-1}}{4^k}.$$

Moreover, by the reverse triangle inequality,

$$|A^k v| = |A^k v - \langle v, v_1 \rangle \lambda_1^k v_1 + \langle v, v_1 \rangle \lambda_1^k v_1| \geq \frac{1}{n} 2^{-k} - \frac{\sqrt{n-1}}{4^k}.$$

If we take  $k$  to be large (say  $k > 10 \log n$ ), and if we define  $z : equals A^k v$ , then  $z$  is approximately an eigenvector of  $A$ , that is

$$|A \frac{A^k v ||A^k v| - \lambda_1 \frac{A^k v}{|A^k v|}}{|A^k v|} 4n^{3/2} 2^{-k}| \leq 4n^{-4}.$$

And to approximately find the first eigenvalue  $\lambda_1$ , we simply compute

$$\frac{z^T A z}{z^T z}.$$

That is, we have approximately found the first eigenvector and eigenvalue of  $A$ .

To find the second eigenvector and eigenvalue, we can repeat the above procedure, where we start by choosing  $v$  such that  $\langle v, v_1 \rangle = 0$ ,  $|v| = 1$  and  $|\langle v, v_2 \rangle| > 1/(10\sqrt{n})$ . To find the third eigenvector and eigenvalue, we can repeat the above procedure, where we start by choosing  $v$  such that  $\langle v, v_1 \rangle = \langle v, v_2 \rangle = 0$ ,  $|v| = 1$  and  $|\langle v, v_3 \rangle| > 1/(10\sqrt{n})$ . And so on.

Google's PageRank algorithm uses the power method to rank websites very rapidly. In particular, they let  $n$  be the number of websites on the internet (so that  $n$  is roughly  $10^9$ ). They then define an  $n \times n$  matrix  $C$  where  $C_{ij} = 1$  if there is a hyperlink between websites  $i$  and  $j$ , and  $C_{ij} = 0$  otherwise. Then, they let  $B$  be an  $n \times n$  matrix such that  $B_{ij}$  is 1 divided by the number of 1's in the  $i^{th}$  row of  $C$ , if  $C_{ij} = 1$ , and  $B_{ij} = 0$  otherwise. Finally, they define

$$A = (.85)B + (.15)D/n$$

where  $D$  is an  $n \times n$  matrix all of whose entries are 1.

The power method finds the eigenvector  $v_1$  of  $A$ , and the size of the  $i^{th}$  entry of  $v_1$  is proportional to the "rank" of website  $i$ .

12. Let  $A$  be a  $2 \times 2$  matrix for which there is a constant  $k$  such that the sum of the entries in each row and each column is  $k$ . Which of the following must be an eigenvector of  $A$  ?

I.  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$

II.  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$

III.  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

- (A) I only      (B) II only      (C) III only      (D) I and II only      (E) I, II, and III

**Solution 12.** (C) This condition makes the matrix of the form

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

There is no reason that  $a = 0$  or  $b = 0$ , so there is no reason  $(1, 0)$  or  $(0, 1)$  should be eigenvectors. But it is easy to verify that  $(1, 1)$  must be.

24. Consider the system of linear equations

$$\begin{aligned} w + 3x + 2y + 2z &= 0 \\ w + 4x + y &= 0 \\ 3w + 5x + 10y + 14z &= 0 \\ 2w + 5x + 5y + 6z &= 0 \end{aligned}$$

with solutions of the form  $(w, x, y, z)$ , where  $w, x, y$ , and  $z$  are real. Which of the following statements is FALSE?

- (A) The system is consistent.
- (B) The system has infinitely many solutions.
- (C) The sum of any two solutions is a solution.
- (D)  $(-5, 1, 1, 0)$  is a solution.
- (E) Every solution is a scalar multiple of  $(-5, 1, 1, 0)$ .

**Solution 24.** (E) Looking at our answers, we can verify directly that  $(-5, 1, 1, 0)$  is a solution. Any multiple of  $(-5, 1, 1, 0)$  is also a solution, which shows that (A), (B), (C), and (D) are all true – leaving only (E). Another solution, for example, is  $(0, 2, -8, 5)$

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 2 & 3 & 4 & 5 \\ 0 & 0 & 3 & 4 & 5 \\ 0 & 0 & 0 & 4 & 5 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$$

34. Which of the following statements about the real matrix shown above is FALSE?

- (A)  $A$  is invertible.
- (B) If  $\mathbf{x} \in \mathbb{R}^5$  and  $A\mathbf{x} = \mathbf{x}$ , then  $\mathbf{x} = \mathbf{0}$ .
- (C) The last row of  $A^2$  is  $(0 \ 0 \ 0 \ 0 \ 25)$ .
- (D)  $A$  can be transformed into the  $5 \times 5$  identity matrix by a sequence of elementary row operations.
- (E)  $\det(A) = 120$

**Solution 34.** (B) An upper triangular matrix is easily verified to be invertible so long as its diagonal entries are all nonzero. Specifically,  $\det A$  is still the product of its diagonal entries, so (E) and (D) and (A) are all true. (C) can easily be verified to be true by computing that the bottom-right corner is 25 (the product of upper triangular matrices still being upper triangular). This leaves (B). (B) can be checked directly to be false: if we let  $x = (1, 0, 0, 0, 0)$ , then  $Ax = x$ .

37. Let  $V$  be a finite-dimensional real vector space and let  $P$  be a linear transformation of  $V$  such that  $P^2 = P$ . Which of the following must be true?

- I.  $P$  is invertible.
  - II.  $P$  is diagonalizable.
  - III.  $P$  is either the identity transformation or the zero transformation.
- (A) None      (B) I only      (C) II only      (D) III only      (E) II and III

**Solution 37.** (C)  $P^2 = P$  means that  $P$  is projection onto some subspace. There is no reason to believe that this should be invertible, but it should definitely be diagonalisable (with eigenbasis some basis of that subspace). III also need not be true if the subspace is anything proper or nontrivial.

50. Let  $A$  be a real  $2 \times 2$  matrix. Which of the following statements must be true?

- I. All of the entries of  $A^2$  are nonnegative.
  - II. The determinant of  $A^2$  is nonnegative.
  - III. If  $A$  has two distinct eigenvalues, then  $A^2$  has two distinct eigenvalues.
- (A) I only      (B) II only      (C) III only      (D) II and III only      (E) I, II, and III

**Solution 50.** (B) There is no reason that all the entries of  $A^2$  need to be nonnegative. Its determinant must be nonnegative though:  $\det(A^2) = (\det A)^2$ . For III, suppose  $A$  is the diagonal matrix with entries  $\pm\lambda$ . Then those are its eigenvalues, and they are distinct so long as  $\lambda \neq 0$ . But  $A^2$  has only one eigenvalue:  $\lambda^2$ .

51. Which of the following is an orthonormal basis for the column space of the real matrix  $\begin{pmatrix} 1 & -1 & 2 & -3 \\ -1 & 1 & -3 & 2 \\ 2 & -2 & 5 & -5 \end{pmatrix}$ ?

- (A)  $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$
- (B)  $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$
- (C)  $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ 0 \end{pmatrix} \right\}$
- (D)  $\left\{ \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ -3 \\ 5 \end{pmatrix} \right\}$
- (E)  $\left\{ \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} \right\}$

**Solution 51.** (E) The basis (C) is not orthogonal and (D) is not normal, so we can rule those out. We can throw out the first column, since it is the negation of the second. A little bit of math shows that the remaining  $3 \times 3$  matrix has determinant 0, so the rank of our column space is 2. That leaves only (A) and (E), but (A) cannot be correct. Our column space contains vectors that have nonzero third entry, so cannot lie in the span of that basis.

## 3 Calculus

These notes include some screenshots from Wikipedia as well as from *Calculus* by Gilbert Strang, available at <https://ocw.mit.edu/ans7870/resources/Strang/Edited/Calculus/Calculus.pdf>. I also used parts from some other resources which I mention when they arise.

### 3.1 Differentiation and common derivatives and integrals to know

**Theorem 3.1.1 (Clairaut's Theorem).** Let  $z = f(x, y)$  be a two variable real-valued function that is defined on a disk  $\mathcal{D}$  that contains the point  $(a, b)$ . Then if  $\frac{\partial^2 z}{\partial x \partial y}$  and  $\frac{\partial^2 z}{\partial y \partial x}$  are continuous on  $\mathcal{D}$ ,  $\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x}$ .

$$\begin{aligned}\frac{d}{dx}(\sin^{-1} x) &= \frac{1}{\sqrt{1-x^2}} & \frac{d}{dx}(\ln(x)) &= \frac{1}{x}, \quad x > 0 \\ \frac{d}{dx}(\cos^{-1} x) &= -\frac{1}{\sqrt{1-x^2}} & \frac{d}{dx}(\ln|x|) &= \frac{1}{x}, \quad x \neq 0 \\ \frac{d}{dx}(\tan^{-1} x) &= \frac{1}{1+x^2} & \frac{d}{dx}(\log_a(x)) &= \frac{1}{x \ln a}, \quad x > 0\end{aligned}$$

$$\begin{aligned}\int \tan u \, du &= \ln|\sec u| + C \\ \int \sec u \, du &= \ln|\sec u + \tan u| + C \\ \int \frac{1}{a^2+u^2} \, du &= \frac{1}{a} \tan^{-1}\left(\frac{u}{a}\right) + C \\ \int \frac{1}{\sqrt{a^2-u^2}} \, du &= \sin^{-1}\left(\frac{u}{a}\right) + C\end{aligned}$$

$$\int \ln u \, du = u \ln(u) - u + C$$

$$\int \sinh x \, dx = \cosh x + C \quad \int \cosh x \, dx = \sinh x + C$$

### 3.2 Matrix Differentiation

Recommended resource: “Matrix Differentiation ( and some other stuff )” by Randal J. Barnes (Department of Civil Engineering, University of Minnesota). Available for download at <https://atmos.washington.edu/~dennis/MatrixCalculus.pdf>.

More information not contained in that pdf (from the appendix of *Convex Optimization* by Stephen Boyd and Lieven Vandenberghe, available for free download at <https://web.stanford.edu/~boyd/cvxbook/>):

**Chain rule.** Suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is differentiable at  $x \in \text{int dom } f$  and  $g : \mathbb{R}^m \rightarrow \mathbb{R}^p$  is differentiable at  $f(x) \in \text{int dom } g$ . Define the composition  $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$  by  $h(z) = g(f(z))$ . Then

$$Dh(x) = Dg(f(x))Df(x)$$

In particular, if  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $g : \mathbb{R} \rightarrow \mathbb{R}$ ,

$$\nabla h(x) = g'(f(x))\nabla f(x)$$

**Example with an affine function.** Suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is differentiable,  $A \in \mathbb{R}^{n \times p}$ , and  $b \in \mathbb{R}^n$ . Define  $g : \mathbb{R}^p \rightarrow \mathbb{R}^m$  as  $g(x) = f(Ax + b)$  with  $\text{dom } g = \{x \mid Ax + b \in \text{dom } f\}$ . Then

$$\nabla g(x) = A^T \nabla f(Ax + b)$$

**Example 2.** Consider  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  where

$$f(x) = \log \sum_{i=1}^m \exp(a_i^T x + b_i) =$$

where  $a_1, \dots, a_m \in \mathbb{R}^n$  and  $b_1, \dots, b_m \in \mathbb{R}$ . Note that  $f(\cdot)$  can be expressed as a composition of  $Ax + b$  (where  $A \in \mathbb{R}^{m \times n}$  has rows  $a_1^T, \dots, a_m^T$ ) and the function  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  given by  $g(y) = \log(\sum_{i=1}^m \exp(y_i))$ . We have

$$\nabla g(y) = \left[ \sum_{i=1}^m e^{y_i} \right]^{-1} (\exp(y_1) \dots \exp(y_m))^T$$

so applying the chain rule yields

$$\nabla f(x) = \left[ \sum_{i=1}^m \exp(a_i^T x + b_i) \right]^{-1} A^T z$$

where  $z_i = \exp(a_i^T x + b_i)$ ,  $i = 1, \dots, m$ .

**Hessians.** The Hessian matrix of  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is denoted by  $\nabla^2 f(x)$  and is given by

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, n$$

The quadratic function

$$f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2} (x - x_0)^T \nabla^2 f(x_0) (x - x_0)$$

is called the **second-order approximation of  $f$  near  $x_0$** .

**Chain rule for second derivative.** A chain rule for the second derivative is difficult in general. Here are some special cases.

**Composition with scalar function.** Suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $g : \mathbb{R} \rightarrow \mathbb{R}$ , and  $h(x) = g(f(x))$ . We have

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

**Composition with affine function.** Suppose  $f : \mathbb{R} \rightarrow \mathbb{R}$ ,  $a \in \mathbb{R}^m$ ,  $b \in \mathbb{R}$ . Define  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  by  $g(x) = f(a^T x + b)$ . Then

$$\nabla^2 g(x) = a^T \nabla^2 f(a^T x + b) a$$

More generally, suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $A \in \mathbb{R}^{n \times m}$ , and  $b \in \mathbb{R}^n$ . Define  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  by  $g(x) = f(Ax + b)$ . Then

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b) A$$

### 3.3 Some theorems in higher dimensions

**Proposition 3.3.1 (Change of Variables).** If  $U$  is a “nice” subset of  $\mathbb{R}^2$  and  $\phi$  is an injective differentiable function on  $U$ , then

$$\int_{\phi(U)} f(u, v) dudv = \int_U f(\phi(x, y)) |J\phi(x, y)| dx dy$$

where  $J\phi(x, y)$  is the Jacobian of  $\phi$  at  $(x, y)$ .

**Taylor’s Theorem (first order).** (borrowed from <https://www.rose-hulman.edu/~bryan/lottamath/mTaylor.pdf>) Consider a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Let  $a \in \mathbb{R}^n$  be a fixed point. Then Taylor’s Theorem states:

If  $f(x)$  is differentiable on an open ball  $B$  around  $a$  and  $x \in B$  then

$$f(x) = f(a) + \nabla f(b)^T (x - a)$$

for some  $b$  on the line segment joining  $a$  and  $x$ .

This can also be expressed as follows. Let  $x, y \in \mathbb{R}^n$ . If  $f(x)$  is continuously differentiable, then

$$f(y) = f(x) + \nabla f(tx + (1-t)y)^T (y - x)$$

for some  $t \in [0, 1]$ .

*Proof.* Consider  $g(z) = f(zy + (1-z)x)$ . If  $f$  is differentiable then so is  $g$ . Then by the Mean Value Theorem, for some  $t \in (0, 1)$  we have  $g(1) - g(0) = g'(t)$ . By the chain rule,

$$g'(t) = \nabla f(x + t(y - x))^T (y - x)$$

Using  $g(1) = f(y)$  and  $g(0) = f(x)$ , we have

$$\iff \nabla f(tx + (1-t)y)^T (y - x) = g(1) - g(0) = f(y) - f(x)$$

□

**Taylor's Theorem (second order).** Consider a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Let  $a \in \mathbb{R}^n$  be a fixed point. Then Taylor's Theorem states:

If  $f(x)$  is twice differentiable on an open ball  $B$  around  $a$  and  $x \in B$  then

$$f(x) = f(a) + (x - a)^T \nabla f(a) + \frac{1}{2}(x - a)^T \nabla^2 f(b)(x - a)$$

for some  $b$  on the line segment joining  $a$  and  $x$ .

This can also be expressed as follows. Let  $x, y \in \mathbb{R}^n$ . If  $f(x)$  is twice continuously differentiable, then

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(ty + (1-t)x)(y - x)$$

for some  $t \in [0, 1]$ .

*Proof.* Consider  $g(z) = f(zy + (1-z)x)$ . If  $f$  is differentiable then so is  $g$ . Then by the second order case of Taylor's Theorem in one dimension, for some  $t \in (0, 1)$  we have  $g(1) = g(0) + g'(0) + (1/2)g''(t)$ . By the chain rule,

$$g''(t) = \frac{\partial}{\partial t} \nabla f(x + t(y - x))^T (y - x) = (y - x)^T \nabla^2 f(x + t(y - x))^T (y - x)$$

Using this result along with  $g(1) = f(y)$ ,  $g(0) = f(x)$ , and  $g'(0) = \nabla f(x)^T (y - x)$ , we have

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x + t(y - x))^T (y - x)$$

□

### 3.4 Optimizing functions of several variables

#### Functions of two variables [ edit ]

Suppose that  $f(x, y)$  is a differentiable [real function](#) of two variables whose second [partial derivatives](#) exist. The [Hessian matrix](#)  $H$  of  $f$  is the  $2 \times 2$  matrix of partial derivatives of  $f$ :

$$H(x, y) = \begin{pmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{pmatrix}.$$

Define  $D(x, y)$  to be the [determinant](#)

$$D(x, y) = \det(H(x, y)) = f_{xx}(x, y)f_{yy}(x, y) - (f_{xy}(x, y))^2,$$

of  $H$ . Finally, suppose that  $(a, b)$  is a critical point of  $f$  (that is,  $f_x(a, b) = f_y(a, b) = 0$ ). Then the second partial derivative test asserts the following:<sup>[1]</sup>

1. If  $D(a, b) > 0$  and  $f_{xx}(a, b) > 0$  then  $(a, b)$  is a local minimum of  $f$ .
2. If  $D(a, b) > 0$  and  $f_{xx}(a, b) < 0$  then  $(a, b)$  is a local maximum of  $f$ .
3. If  $D(a, b) < 0$  then  $(a, b)$  is a [saddle point](#) of  $f$ .
4. If  $D(a, b) = 0$  then the second derivative test is inconclusive, and the point  $(a, b)$  could be any of a minimum, maximum or saddle point.

#### Functions of many variables [ edit ]

For a function  $f$  of two or more variables, there is a generalization of the rule above. In this context, instead of examining the determinant of the Hessian matrix, one must look at the [eigenvalues](#) of the Hessian matrix at the critical point. The following test can be applied at any critical point  $(a, b, \dots)$  for which the Hessian matrix is [invertible](#):

1. If the Hessian is [positive definite](#) (equivalently, has all eigenvalues positive) at  $(a, b, \dots)$ , then  $f$  attains a local minimum at  $(a, b, \dots)$ .
2. If the Hessian is [negative definite](#) (equivalently, has all eigenvalues negative) at  $(a, b, \dots)$ , then  $f$  attains a local maximum at  $(a, b, \dots)$ .
3. If the Hessian has both positive and negative eigenvalues then  $(a, b, \dots)$  is a saddle point for  $f$  (and in fact this is true even if  $(a, b, \dots)$  is degenerate).

### 3.5 Lagrange Multipliers

: to flesh out! <http://tutorial.math.lamar.edu/Classes/CalcIII/LagrangeMultipliers.aspx>

### 3.6 Line Integrals

(p. 555 of Strang book)

Suppose a force in two-dimensional space is given by  $\mathbf{F} = M\mathbf{i} + N\mathbf{j}$ . Then the work done by this force on a particle moving along a curve  $C$  is given by

$$W = \int_C \mathbf{F} \cdot d\mathbf{R} = \int_C Mdx + Ndy$$

Along a curve in three-dimensional space the work done by a three-dimensional force  $\mathbf{F} = Mi + Nj + Pk$  is given by

$$W = \int_C \mathbf{F} \cdot \mathbf{T} ds = \int_C \mathbf{F} \cdot d\mathbf{R} = \int_C Mdx + Ndy + Pdz$$

where the tangent vector  $\mathbf{T}$  is given by

$$\mathbf{T} = \frac{d\mathbf{R}}{ds}$$

**Green's Theorem:** Suppose the region  $R$  is bounded by the simple closed piecewise smooth curve  $C$ . Then an integral over  $R$  equals a line integral around  $C$ :

$$\oint_C \mathbf{F} \cdot d\mathbf{R} = \oint_C Mdx + Ndy = \int \int_R \left( \frac{\partial N}{\partial x} - \frac{\partial M}{\partial y} \right) dx dy$$

**Line integrals chapter!** <http://tutorial.math.lamar.edu/Classes/CalcIII/LineIntegralsIntro.aspx>

**Surface integrals chapter!** <http://tutorial.math.lamar.edu/Classes/CalcIII/SurfaceIntegralsIntro.aspx>

### 3.7 Miscellaneous

**13A** The tangent plane at  $(x_0, y_0, z_0)$  has the same slopes as the surface  $z = f(x, y)$ . The equation of the tangent plane (a linear equation) is

$$z - z_0 = \left( \frac{\partial f}{\partial x} \right)_0 (x - x_0) + \left( \frac{\partial f}{\partial y} \right)_0 (y - y_0). \quad (1)$$

The normal vector  $\mathbf{N}$  to that plane has components  $(\partial f / \partial x)_0, (\partial f / \partial y)_0, -1$ .

**13B** The tangent plane to the surface  $F(x, y, z) = c$  has the linear equation

$$\left( \frac{\partial F}{\partial x} \right)_0 (x - x_0) + \left( \frac{\partial F}{\partial y} \right)_0 (y - y_0) + \left( \frac{\partial F}{\partial z} \right)_0 (z - z_0) = 0. \quad (7)$$

The normal vector is  $\mathbf{N} = \left( \frac{\partial F}{\partial x} \right)_0 \mathbf{i} + \left( \frac{\partial F}{\partial y} \right)_0 \mathbf{j} + \left( \frac{\partial F}{\partial z} \right)_0 \mathbf{k}$ .

$$dz = (\partial z / \partial x)_0 dx + (\partial z / \partial y)_0 dy \quad \text{or} \quad df = f_x dx + f_y dy. \quad (10)$$

This is the **total differential**. All letters  $dz$  and  $df$  and  $dw$  can be used, but  $\partial z$  and  $\partial f$  are not used. Differentials suggest small movements in  $x$  and  $y$ ; then  $dz$  is the resulting movement in  $z$ . On the tangent plane, equation (10) holds exactly.

The **directional derivative**, denoted  $D_v f(x, y)$ , is a derivative of a multivariable function in the direction of a vector  $\mathbf{v}$ . It is the scalar projection of the gradient onto  $\mathbf{v}$ .

$$D_v f(x, y) = \text{comp}_v \nabla f(x, y) = \frac{\nabla f(x, y) \cdot \mathbf{v}}{|\mathbf{v}|}$$

### 3.8 Practice Problems

**13F** The directional derivative is  $D_{\mathbf{u}} f = (\text{grad } f) \cdot \mathbf{u}$ . The level direction is perpendicular to  $\text{grad } f$ , since  $D_{\mathbf{u}} f = 0$ . **The slope  $D_{\mathbf{u}} f$  is largest when  $\mathbf{u}$  is parallel to  $\text{grad } f$** . That maximum slope is the length  $|\text{grad } f| = \sqrt{f_x^2 + f_y^2}$ :

$$\text{for } \mathbf{u} = \frac{\text{grad } f}{|\text{grad } f|} \text{ the slope is } (\text{grad } f) \cdot \mathbf{u} = \frac{|\text{grad } f|^2}{|\text{grad } f|} = |\text{grad } f|.$$

$$\int_C g(x, y) ds = \text{limit of } \sum_{i=1}^N g(x_i, y_i) \Delta s_i \text{ as } (\Delta s)_{\max} \rightarrow 0.$$

**The differential  $ds$  becomes  $(ds/dt)dt$ . Everything changes over to  $t$ :**

$$\int g(x, y) ds = \int_{t=a}^{t=b} g(x(t), y(t)) \sqrt{(dx/dt)^2 + (dy/dt)^2} dt.$$

19. Let  $f$  and  $g$  be twice-differentiable real-valued functions defined on  $\mathbb{R}$ . If  $f'(x) > g'(x)$  for all  $x > 0$ , which of the following inequalities must be true for all  $x > 0$ ?

- (A)  $f(x) > g(x)$
- (B)  $f''(x) > g''(x)$
- (C)  $f(x) - f(0) > g(x) - g(0)$
- (D)  $f'(x) - f'(0) > g'(x) - g'(0)$
- (E)  $f''(x) - f''(0) > g''(x) - g''(0)$

**Solution 19.** (C) There is no reason that  $f(x) > g(x)$ , or that  $f''(x) > g''(x)$ . But we do know that

$$\int_0^x f'(t) dt > \int_0^x g'(t) dt \implies f(x) - f(0) > g(x) - g(0).$$

This is precisely an answer.

22. What is the volume of the solid in  $xyz$ -space bounded by the surfaces  $y = x^2$ ,  $y = 2 - x^2$ ,  $z = 0$ , and  $z = y + 3$ ?

- (A)  $\frac{8}{3}$       (B)  $\frac{16}{3}$       (C)  $\frac{32}{3}$       (D)  $\frac{104}{105}$       (E)  $\frac{208}{105}$

**Solution 22.** (C) It looks like our  $x$ -coordinates are running over  $[-1, 1]$ , with  $y$  depending on  $x$  and  $z$  depending on  $y$ . To find the volume of the solid, we just need to integrate the constant function 1. We must therefore compute

$$\begin{aligned} \int_{-1}^1 \int_{x^2}^{2-x^2} \int_0^{y+3} 1 \, dz \, dy \, dx &= \int_{-1}^1 \int_{x^2}^{2-x^2} y + 3 \, dy \, dx \\ &= \int_{-1}^1 ((2 - x^2)^2/2 + 3(2 - x^2)) - ((x^2)^2/2 + 3(x^2)) \, dx \\ &= \int_{-1}^1 8 - 8x^2 \, dx \\ &= 8x - 8x^3/3 \Big|_{-1}^1 = (8 - 8/3) - (-8 + 8/3) = 32/3. \end{aligned}$$

24. Let  $h$  be the function defined by  $h(x) = \int_0^{x^2} e^{x+t} dt$  for all real numbers  $x$ . Then  $h'(1) =$

- (A)  $e - 1$       (B)  $e^2$       (C)  $e^2 - e$       (D)  $2e^2$       (E)  $3e^2 - e$

**Solution 24.** (E) We can actually just integrate this, and not worry about differentiation under the integral.

$$\int_0^{x^2} e^{x+t} dt = e^x \int_0^{x^2} e^t dt = e^x (e^{x^2} - 1) = e^{x^2+x} - e^x.$$

Then deriving that,

$$h'(x) = (2x + 1)e^{x^2+x} - e^x,$$

whence our result follows immediately.

26. Let  $f(x, y) = x^2 - 2xy + y^3$  for all real  $x$  and  $y$ . Which of the following is true?

- (A)  $f$  has all of its relative extrema on the line  $x = y$ .
- (B)  $f$  has all of its relative extrema on the parabola  $x = y^2$ .
- (C)  $f$  has a relative minimum at  $(0, 0)$ .
- (D)  $f$  has an absolute minimum at  $\left(\frac{2}{3}, \frac{2}{3}\right)$ .
- (E)  $f$  has an absolute minimum at  $(1, 1)$ .

**Solution 26.** (A) We are concerned about its extrema, we should find some partial derivatives.

$$f_x = 2x - 2y, \quad f_y = -2x + 3y^2.$$

We would like to know when they are both zero. The first equation gives us  $x = y$  and the second gives us  $2x = 3y^2$ , so that

$$2y = 3y^2 \implies (3y - 2)y = 0 \implies y = 0, 2/3.$$

Therefore our solutions are  $(0, 0)$  and  $(2/3, 2/3)$ . Indeed, our relative extrema are all on the line  $x = y$ . To do some more checking (which you should not do on the actual test),

$$f_{xx} = 2, \quad f_{yy} = 6y, \quad f_{xy} = f_{yx} = -2.$$

Then the determinant of the Hessian is  $12y - 4$ . This shows that  $(0, 0)$  is a saddle point. There is no reason that  $(2/3, 2/3)$  is an absolute minimum without further verification, and  $(1, 1)$  needn't be an extreme point.

27. Consider the two planes  $x + 3y - 2z = 7$  and  $2x + y - 3z = 0$  in  $\mathbb{R}^3$ . Which of the following sets is the intersection of these planes?

- (A)  $\emptyset$
- (B)  $\{(0, 3, 1)\}$
- (C)  $\{(x, y, z) : x = t, y = 3t, z = 7 - 2t, t \in \mathbb{R}\}$
- (D)  $\{(x, y, z) : x = 7t, y = 3 + t, z = 1 + 5t, t \in \mathbb{R}\}$
- (E)  $\{(x, y, z) : x - 2y - z = -7\}$

**Solution 27.** (D) First, we know that the intersection of two planes in  $\mathbb{R}^3$  should be either a plane or a line. In our case, the two planes are definitely not the same, so we will obtain a line. The slope of the line can be found by taking the cross product of the normal vectors of the two planes in question.

$$(1, 3, -2) \times (2, 1, -3) = \det \begin{bmatrix} i & j & k \\ 1 & 3 & -2 \\ 2 & 1 & -3 \end{bmatrix} = (-7, -1, -5).$$

The only solution corresponding to this slope is (D), as the coefficients of  $t$  in  $(x, y, z)$  are  $(7, 1, 5)$ .

32.  $\frac{d}{dx} \int_{x^3}^{x^4} e^{t^2} dt =$

- (A)  $e^{x^6} (e^{x^8-x^6} - 1)$
- (B)  $4x^3 e^{x^8}$
- (C)  $\frac{1}{\sqrt{1-e^{x^2}}}$
- (D)  $\frac{e^{x^2}}{x^2} - 1$
- (E)  $x^2 e^{x^6} (4x e^{x^8-x^6} - 3)$

**Solution 32.** (E) We can sort this out in two steps and apply the fundamental theorem to each.

$$\frac{d}{dx} \left( \int_{x^3}^0 e^{t^2} dt + \int_0^{x^4} e^{t^2} dt \right)$$

For the first,

$$\frac{d}{dx} \int_{x^3}^0 e^{t^2} dt = -\frac{d}{dx} \int_0^{x^3} e^{t^2} dt = -3x^2 e^{x^6}$$

For the second,

$$\frac{d}{dx} \int_0^{x^4} e^{t^2} dt = 4x^3 e^{x^8}.$$

All told, our integral is  $x^2 e^{x^6} (4x e^{x^8-x^6} - 3)$ .

41. Let  $\ell$  be the line that is the intersection of the planes  $x + y + z = 3$  and  $x - y + z = 5$  in  $\mathbb{R}^3$ . An equation of the plane that contains  $(0, 0, 0)$  and is perpendicular to  $\ell$  is

- (A)  $x - z = 0$
- (B)  $x + y + z = 0$
- (C)  $x - y - z = 0$
- (D)  $x + z = 0$
- (E)  $x + y - z = 0$

**Solution 41.** (A) The first plane is determined by the normal vector  $(1, 1, 1)$ , and the second determined by  $(1, -1, 1)$ . Therefore the slope of  $\ell$  is determined by a vector perpendicular to those, i.e. the cross product.

$$(1, 1, 1) \times (1, -1, 1) = \det \begin{bmatrix} i & j & k \\ 1 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix} = (2, 0, -2).$$

41. Let  $C$  be the circle  $x^2 + y^2 = 1$  oriented counterclockwise in the  $xy$ -plane. What is the value of the line integral  $\oint_C (2x - y) dx + (x + 3y) dy$ ?

- (A) 0
- (B) 1
- (C)  $\frac{\pi}{2}$
- (D)  $\pi$
- (E)  $2\pi$

**Solution 41.** (E) This is a classic Green's theorem problem.

$$\oint_{\partial D} L dx + M dy = \iint_D \left( \frac{\partial M}{\partial x} - \frac{\partial L}{\partial y} \right) dx dy.$$

In our case,

$$\oint_C (2x - y) dx + (x + 3y) dy = \iint_D (1 + 1) dx dy = 2A,$$

where  $A$  is the area of the unit circle, i.e.  $\pi$ .

So that is the slope of  $\ell$ . We need this to be the normal vector for the plane in question, so it seems that  $(1, 0, -1)$  is our best bet (out of the given options).

$$\begin{aligned}y' + xy &= x \\y(0) &= -1\end{aligned}$$

44. If  $y$  is a real-valued function defined on the real line and satisfying the initial value problem above, then  $\lim_{x \rightarrow -\infty} y(x) =$
- (A) 0      (B) 1      (C) -1      (D)  $\infty$       (E)  $-\infty$

**Solution 44.** (B) Putting it in simpler terms,

$$\frac{dy}{dx} + xy = x \implies \frac{dy}{dx} = x(1-y) \implies \frac{dy}{1-y} = x \, dx.$$

Integrating both sides, we obtain

$$-\log(1-y) = x^2/2 + C' \implies 1-y = Ce^{-x^2/2} \implies y = 1 - Ce^{-x^2/2}.$$

Solving the initial value problem gives  $C = 2$ . Furthermore, as  $x \rightarrow -\infty$ , the second term above vanishes so we get 1 in the limit.

48. Let  $g$  be the function defined by  $g(x, y, z) = 3x^2y + z$  for all real  $x$ ,  $y$ , and  $z$ . Which of the following is the best approximation of the directional derivative of  $g$  at the point  $(0, 0, \pi)$  in the direction of the vector  $\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$ ? (Note:  $\mathbf{i}$ ,  $\mathbf{j}$ , and  $\mathbf{k}$  are the standard basis vectors in  $\mathbb{R}^3$ .)
- (A) 0.2      (B) 0.8      (C) 1.4      (D) 2.0      (E) 2.6

**Solution 48.** (B) It would be good to recall the formula for the directional derivative. We take the gradient of the function then take its scalar product with the normalised vector in the direction we want. To begin,

$$\nabla g = (6xy, 3x^2, 1).$$

At the point  $(0, 0, \pi)$ , we have  $\nabla g = (0, 0, 1)$ . That works out pretty well for us. The normalised version of the vector  $(1, 2, 3)$  is  $(1/\sqrt{14}, 2/\sqrt{14}, 3/\sqrt{14})$ . Dotting this with  $(0, 0, 1)$  gives  $3/\sqrt{14}$ , and since  $\sqrt{14} = 3.5$  or so our answer should be closer to 0.8 than 0.2.

48. Consider the theorem: If  $f$  and  $f'$  are both strictly increasing real-valued functions on the interval  $(0, \infty)$ , then  $\lim_{x \rightarrow \infty} f(x) = \infty$ . The following argument is suggested as a proof of this theorem.

- (1) By the Mean Value Theorem, there is a  $c_1$  in the interval  $(1, 2)$  such that

$$f'(c_1) = \frac{f(2) - f(1)}{2 - 1} = f(2) - f(1) > 0.$$

- (2) For each  $x > 2$ , there is a  $c_x$  in  $(2, x)$  such that  $\frac{f(x) - f(2)}{x - 2} = f'(c_x)$ .

- (3) For each  $x > 2$ ,  $\frac{f(x) - f(2)}{x - 2} = f'(c_x) > f'(c_1)$  since  $f'$  is strictly increasing.

- (4) For each  $x > 2$ ,  $f(x) > f(2) + (x - 2)f'(c_1)$ .

- (5)  $\lim_{x \rightarrow \infty} f(x) = \infty$

Which of the following statements is true?

- (A) The argument is valid.
- (B) The argument is not valid since the hypotheses of the Mean Value Theorem are not satisfied in (1) and (2).
- (C) The argument is not valid since (3) is not valid.
- (D) The argument is not valid since (4) cannot be deduced from the previous steps.
- (E) The argument is not valid since (4) does not imply (5).

**Solution 48.** (A) The only issue here seems to be that (4) implies that  $f(x)$  gets very large so long as  $f'(c_1)$  is positive. But we know that it is, since  $f$  is a strictly increasing function. Therefore everything is satisfactory.

## 4 Differential Equations

61. A tank initially contains a salt solution of 3 grams of salt dissolved in 100 liters of water. A salt solution containing 0.02 grams of salt per liter of water is sprayed into the tank at a rate of 4 liters per minute. The sprayed solution is continually mixed with the salt solution in the tank, and the mixture flows out of the tank at a rate of 4 liters per minute. If the mixing is instantaneous, how many grams of salt are in the tank after 100 minutes have elapsed?

(A) 2      (B)  $2 - e^{-2}$       (C)  $2 + e^{-2}$       (D)  $2 - e^{-4}$       (E)  $2 + e^{-4}$

**Solution 61.** (E) We can set this up as a differential equation. Let  $s$  denote the amount of salt in the tank, and let  $t$  denote time. We have the initial condition of  $s(0) = 3$ .  $s'(t)$  depends on two factors: the salt flowing in and the salt flowing out. The salt flows in constantly at a rate of 0.08 grams per minute, and the salt flows out at a rate of  $4 \cdot (s/100) = s/25$  grams per minute. Therefore

$$s'(t) = \frac{ds}{dt} = 0.08 - s(t)/25 \implies \frac{ds}{dt} = 0.04(2 - s) \implies \frac{ds}{2 - s} = 0.04 dt.$$

Doing the usual calculus,

$$-\log(2 - s) = 0.04t + C' \implies 2 - s = Ce^{-0.04t} \implies s(t) = 2 - Ce^{-0.04t}.$$

The initial condition tells us that  $C = -1$ , so  $s(t) = 2 + e^{-0.04t}$ . Plugging in  $t = 100$  gives our answer.

## 5 Real Analysis

These are my notes from Math 4650: Analysis I at Cal State LA as well as Prof. Steven Heilman's notes from Math 541A at USC.

### 5.1 Midterm 1

#### 5.1.1 Homework 1

**Definition 5.1.** Let  $S \subseteq \mathbb{R}$ . We say that  $S$  is **bounded from above** if  $\exists b \in \mathbb{R}$  where

$$s \leq b \quad \forall s \in S$$

If this is the case, we call  $b$  an **upper bound** of  $S$ .

If  $b \leq c$  for all upper bounds  $c$  of  $S$ , we call  $b$  the **supremum** of  $S$ :  $b = \sup(S)$ .

**Definition 5.2.** We say that  $S$  is **bounded from below** if  $\exists a \in \mathbb{R}$  where

$$s \geq a \quad \forall s \in S$$

If this is the case, we call  $a$  a **lower bound** of  $S$ .

If  $a \geq d$  for all lower bounds  $d$  of  $S$ , we call  $a$  the **infimum** of  $S$ :  $a = \inf(S)$ .

**Proposition 5.1.1. Useful Sup/Inf Fact:** Let  $S \subseteq \mathbb{R}$ ,  $S \neq \emptyset$ .

(1) Suppose  $S$  is bounded from above by an element  $b$ . Then  $b = \sup(S) \iff \forall \epsilon > 0 \exists x \in S$  with

$$b - \epsilon < x \leq b$$

(2) Suppose  $S$  is bounded from below by an element  $a$ . Then  $a = \inf(S) \iff \forall \epsilon > 0 \exists x \in S$  with

$$a \leq x < a + \epsilon$$

**Completeness Axiom:** Let  $S$  be a nonempty subset of  $\mathbb{R}$ . If  $S$  is bounded from above, then  $\sup(S)$  exists. If  $S$  is bounded from below, then  $\inf(S)$  exists.

**Facts about absolute value:**

- 

**Proposition 5.1.2.**  $|x - y| < \epsilon \iff y - \epsilon < x < y + \epsilon$ .

*Proof.* In notes 08/23. □

- 

**Proposition 5.1.3.**  $|ab| = |a||b|$ .

*Proof.*

$$\begin{aligned}
 |ab| &= \begin{cases} ab & ab \geq 0 \\ -ab & ab < 0 \end{cases} = \begin{cases} ab & a \geq 0, b \geq 0 \\ -ab & a \geq 0, b < 0 \\ -ab & a < 0, b \geq 0 \\ ab & a < 0, b < 0 \end{cases} = \begin{cases} ab & a \geq 0, b \geq 0 \\ a(-b) & a \geq 0, b < 0 \\ (-a)b & a < 0, b \geq 0 \\ (-a)(-b) & a < 0, b < 0 \end{cases} \\
 &= \begin{cases} |a||b| & a \geq 0, b \geq 0 \\ |a||b| & a \geq 0, b < 0 \\ |a||b| & a < 0, b \geq 0 \\ |a||b| & a < 0, b < 0 \end{cases} \implies |ab| = |a||b|
 \end{aligned}$$

□

•

**Proposition 5.1.4.** Let  $\epsilon > 0$ . Then  $|a| < \epsilon \iff -\epsilon < a < \epsilon$ .

*Proof.* Follows from Proposition 5.1.2 if  $x = a, y = 0$ . □

•

**Proposition 5.1.5.**  $-|a| \leq a \leq |a|$

*Proof.* Follows from Proposition 5.1.2 if  $x = a, y = 0, \epsilon = |a|$ . □

•

**Theorem 5.1.6. Triangle Inequality:**  $|a + b| \leq |a| + |b|$ .

*Proof.* In notes 08/23. □

**Corollary 5.1.6.1. Triangle Inequality:**  $|a - b| \leq |a| + |b|$ .

*Proof.* Follows from Theorem 5.1.6, let  $b = -b$ . □

**Remark.** See also Theorem 8.2.11.

•

**Proposition 5.1.7.**  $| |a| - |b| | \leq |a - b|$ .

*Proof.* By Proposition 5.1.2,  $| |a| - |b| | \leq |a - b|$  if and only if

$$|b| - |a - b| \leq |a| \leq |b| + |a - b| \tag{5.1}$$

The left half of (5.1) is true by the Triangle Inequality (Theorem 5.1.6):

$$|b| = |a - (a - b)| \leq |a| + |a - b| \iff |b| \leq |a| + |a - b| \iff |b| - |a - b| \leq |a|$$

The right half of (5.1) is also true by the Triangle Inequality (Theorem 5.1.6):

$$|a| = |b + a - b| \leq |b| + |a - b|$$

Therefore

$$| |a| - |b| | \leq |a - b|.$$

□

*Proof.* (Alternative proof.) Note that by the Triangle Inequality (Theorem 5.1.6),

$$|a| = |a - b + b| \leq |a - b| + |b| \implies |a| - |b| \leq |a - b|$$

Also,

$$|b| = |b - a + a| \leq |b - a| + |a| \implies -|b - a| \leq |a| - |b| \implies -|a - b| \leq |a| - |b|$$

where the last step follows from Proposition 5.1.9. Therefore

$$-|a - b| \leq |a| - |b| \leq |a - b|$$

and by Proposition 5.1.2,

$$| |a| - |b| | \leq |a - b|.$$

□

•

**Proposition 5.1.8.** If  $a < x < b$  and  $a < y < b$  then  $|x - y| < b - a$ .

*Proof.*

$$y > a \implies -y < -a \implies b - y < b - a$$

$$b > y \implies b - y = |b - y| \implies |b - y| < b - a$$

By the Triangle Inequality (Theorem 5.1.6),

$$|x - y| = |x - b + b - y| \leq |x - b| + |b - y|$$

Since  $b < x$ ,  $|x - b| > 0$ . Therefore  $|x - y| < |b - y|$ .

$$\implies |x - y| < |b - y| < b - a$$

$$\implies |x - y| < b - a$$

□

*Proof.* (Alternative proof.) Break into two cases.

– **Case 1:**  $x \geq y$ . Then  $|x - y| = x - y$ . We know  $a < x < b \implies 0 < x - a < b - a$ .

$$a < y \implies -a > -y \implies x - a > x - y \implies x - y < x - a < b - a$$

$$\implies |x - y| < b - a$$

– **Case 2:**  $x < y$ . Then  $|x - y| = y - x$ . We know  $a < y < b \implies 0 < y - a < b - a$ .

$$a < x \implies -a > -x \implies y - a > y - x \implies y - x < y - a < b - a$$

$$\implies |x - y| < b - a$$

□

•

**Proposition 5.1.9.**  $|a - b| = |b - a|$

*Proof.*  $|a - b| = |(-1)(b - a)| = |-1||b - a| = |b - a|$ , where the second-to-last step follows from Proposition 5.1.2.

□

### 5.1.2 Homework 2

**Definition 5.3.** A sequence  $(a_n)$  of real numbers is said to **converge** to a **limit**  $L \in \mathbb{R}$  if  $\forall \epsilon > 0 \exists N > 0$  where

$$n \geq N \implies |a_n - L| < \epsilon$$

We say that  $(a_n)$  **diverges** if it does not converge.

**Definition 5.4.** A sequence  $(a_n)$  of real numbers is **bounded** if  $\exists M > 0$  where  $\forall n \in \mathbb{N}$

$$|a_n| \leq M.$$

**Theorem 5.1.10.** If  $(a_n)$  converges then  $(a_n)$  is bounded.

**Definition 5.5.** Let  $(a_n)$  be a sequence of real numbers. We say that  $(a_n)$  is a **Cauchy sequence** if  $\forall \epsilon > 0 \exists N$  where

$$n, m \geq N \implies |a_n - a_m| < \epsilon$$

**Theorem 5.1.11.**  $(a_n)$  is Cauchy if and only if  $(a_n)$  converges.

**Corollary 5.1.11.1.** If  $(a_n)$  is Cauchy then  $(a_n)$  is bounded.

*Proof.* Let  $\epsilon = 1$ . Since  $(a_n)$  is Cauchy,  $\exists N > 0 \mid n, m \geq N \implies$

$$|a_n - a_m| < 1$$

So,  $n \geq N \implies$

$$|a_n - a_N| < 1 \iff a_N - 1 < a_n < a_N + 1 \implies |a_n| < |a_N + 1| \leq |a_N| + 1$$

Let  $M = \max\{|a_1|, |a_2|, \dots, |a_{N-1}|, |a_N| + 1\}$ . Then  $|a_n| \leq M \forall n \geq 1$ . Therefore  $(a_n)$  is bounded.  $\square$

**Theorem 5.1.12. (Squeeze theorem.)** Suppose that  $\{a_n\}$ ,  $\{b_n\}$ , and  $\{c_n\}$  are sequences of real numbers such that  $a_n \leq b_n \leq c_n$  for all  $n$ . If both  $\{a_n\}$  and  $\{c_n\}$  converge to  $L$ , then  $\{b_n\}$  converges to  $L$ .

*Proof.* Let  $\epsilon > 0$ .  $(a_n)$  converges to  $L \implies$

$$\forall \epsilon > 0 \exists N_A \mid n \geq N_A \implies |a_n - L| < \epsilon$$

$(c_n)$  converges to  $L \implies$

$$\forall \epsilon > 0 \exists N_C \mid n \geq N_C \implies |c_n - L| < \epsilon$$

Let  $N = \max\{N_A, N_C\}$ . Then by one of our absolute values rules,  $n \geq N \implies$

$$|a_n - L| < \epsilon \iff L - \epsilon < a_n < L + \epsilon$$

$$|c_n - L| < \epsilon \iff L - \epsilon < c_n < L + \epsilon$$

Therefore since  $a_n \leq b_n \leq c_n$ ,

$$L - \epsilon < a_n \leq b_n \leq c_n < L + \epsilon \implies L - \epsilon < b_n < L + \epsilon \iff |b_n - L| < \epsilon$$

Therefore  $(b_n)$  converges to  $L$ .  $\square$

**Theorem 5.1.13.** Suppose that  $\{a_n\}$  and  $\{b_n\}$  are sequences of real numbers such that  $a_n \leq b_n$  for all  $n$ . If  $\{a_n\}$  and  $\{b_n\}$  converge to  $A$  and  $B$  respectively, then  $A \leq B$ .

*Proof.* Suppose  $A > B$ . Then let  $\epsilon = \frac{A-B}{4} > 0$ .  $(a_n)$  converges to  $A \implies$

$$\exists N_A \mid n \geq N_A \implies |a_n - A| < \epsilon \iff A - \epsilon < a_n < A + \epsilon$$

$(b_n)$  converges to  $B \implies$

$$\exists N_B \mid n \geq N_B \implies |b_n - B| < \epsilon \iff B - \epsilon < b_n < B + \epsilon$$

Then if  $n > \max\{N_A, N_B\}$ ,

$$A - \epsilon < a_n < A + \epsilon \iff A - \frac{A - B}{4} < a_n < A + \frac{A - B}{4} \iff \frac{3A}{4} + \frac{B}{4} < a_n < \frac{5A}{4} - \frac{B}{4}$$

$$B - \epsilon < b_n < B + \epsilon \iff B - \frac{A - B}{4} < b_n < B + \frac{A - B}{4} \iff \frac{5B}{4} - \frac{A}{4} < b_n < \frac{3B}{4} + \frac{A}{4}$$

This implies

$$b_n < \frac{3B}{4} + \frac{A}{4} = \frac{B}{4} + \frac{A}{4} + \frac{2B}{4} < \frac{B}{4} + \frac{A}{4} + \frac{2A}{4} = \frac{3A}{4} + \frac{B}{4} < a_n$$

Contradiction, since it is given that  $a_n \leq b_n \forall n$ . Therefore  $A \leq B$ .  $\square$

## 5.2 Midterm 2

### 5.2.1 Homework 3

**Definition 5.6. (Limits of functions at infinity.)** Let  $f$  be a real-valued function defined on some set  $D$  where  $D$  contains an interval of the form  $(a, \infty)$ . Let  $L \in \mathbb{R}$ . We say

$$\lim_{x \rightarrow \infty} f(x) = L$$

if  $\forall \epsilon > 0 \exists N \in \mathbb{R}$  where

$$x \geq N \implies |f(x) - L| < \epsilon.$$

**Definition 5.7.** Let  $D \subseteq \mathbb{R}$ . Let  $a \in \mathbb{R}$ . We say that  $a$  is a **limit point** (or “cluster point,” or “accumulation point”) of  $D$  if  $\forall \delta > 0 \exists x \in D$  where

$$x \neq a \text{ and } |x - a| < \delta$$

(Note that  $a$  may or may not be contained in  $D$ .)

**Definition 5.8. (Limit of a function at  $a$ .)** Let  $D \subseteq \mathbb{R}$  and  $f : D \rightarrow \mathbb{R}$ . Let  $a$  be a limit point of  $D$ . Let  $x \in D$ . We say that  $f$  has a *limit as  $x$  tends to  $a$*  if  $\exists L \in \mathbb{R}$  where  $\forall \epsilon > 0 \exists \delta > 0$  such that

$$0 < |x - a| < \delta \implies |f(x) - L| < \epsilon$$

and we write

$$\lim_{x \rightarrow a} f(x) = L$$

**Proposition 5.2.1. (Properties of Limits.)** Let  $D \subseteq \mathbb{R}$  and let  $a$  be a limit point of  $D$ . Suppose  $f : D \rightarrow \mathbb{R}$  and  $g : D \rightarrow \mathbb{R}$ . Let  $\alpha \in \mathbb{R}$ .

(1) If  $\lim_{x \rightarrow a} f(x) = L$  and  $\lim_{x \rightarrow a} g(x) = M$  then

(a)

$$\lim_{x \rightarrow a} \alpha = \alpha$$

(b)

$$\lim_{x \rightarrow a} [f(x) + g(x)] = L + M$$

(c)

$$\lim_{x \rightarrow a} [f(x) - g(x)] = L - M$$

(d)

$$\lim_{x \rightarrow a} [f(x) \cdot g(x)] = L \cdot M$$

(e)

$$\lim_{x \rightarrow a} [\alpha \cdot f(x)] = \alpha \cdot L$$

(2) If  $h : D \rightarrow \mathbb{R}$  and  $h(x) \neq 0 \forall x \in D$  and  $\lim_{x \rightarrow a} h(x) = H \neq 0$ , then

$$\lim_{x \rightarrow a} \frac{1}{h(x)} = \frac{1}{H}$$

Note that properties (2) and (1)(d) combined imply

$$\lim_{x \rightarrow a} \frac{f(x)}{h(x)} = \frac{L}{H}$$

### 5.2.2 Homework 4

**Definition 5.9. (Continuity.)** Let  $D \subseteq \mathbb{R}$  and  $f : D \rightarrow \mathbb{R}$  and  $a \in D$ . Then  $f$  is **continuous** at  $a$  if  $\lim_{x \rightarrow a} f(x)$  exists and

$$\lim_{x \rightarrow a} f(x) = f(a)$$

**Remark.** if  $f$  is continuous at  $a$ , then we can say  $\forall \epsilon > 0 \exists \delta > 0$  such that

$$|x - a| < \delta \implies |f(x) - L| < \epsilon$$

that is, we don't need to say  $0 < |x - a| < \delta$ .

**Definition 5.10.** If  $B \subseteq D$ , then  $f$  is **continuous on  $B$**  if  $f$  is continuous at every  $b \in B$ .

**Theorem 5.2.2. (Intermediate Value Theorem.)** Let  $f$  be continuous on  $[a, b]$  and suppose  $f(a) < f(b)$ .  $\forall d$  such that

$$f(a) < d < f(b)$$

$\exists c \in \mathbb{R}$  where

$$a < c < b, f(c) = d.$$

### 5.3 Final

#### 5.3.1 Homework 5

**Definition 5.11.** Let  $S \subseteq \mathbb{R}$ . We say  $x \in \mathbb{R}$  is an **interior point** of  $S$  if there exists an open interval  $(a, b)$  where

$$x \in (a, b) \text{ and } (a, b) \subseteq S.$$

**Definition 5.12. (Open sets.)** Let  $S \subseteq \mathbb{R}$ . We say  $S$  is **open** if every  $x \in S$  is an interior point of  $S$ .

**Definition 5.13. (Closed sets.)** Let  $S \subseteq \mathbb{R}$ . We say  $S$  is **closed** if  $\mathbb{R} \setminus S$  is open.

**Theorem 5.3.1.** A set is closed if and only if it contains all of its limit points.

(Facts about open and closed sets.) Suppose  $a, b \in \mathbb{R}$ . Then

•

**Proposition 5.3.2.**  $(a, \infty)$  is open.

*Proof.* Let  $x \in (a, \infty)$ . Since  $x > a$ ,  $\exists \epsilon > 0 \mid a + \epsilon = x$ . Then  $a = x - \epsilon < x - \frac{\epsilon}{2} < x < x + \frac{\epsilon}{2} < \infty$ . Therefore  $x \in (x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2}) \subseteq (a, \infty)$ , so  $(a, \infty)$  is open.  $\square$

•

**Proposition 5.3.3.**  $(-\infty, b)$  is open.

*Proof.* Let  $x \in (-\infty, b)$ . Since  $x < b$ ,  $\exists \epsilon > 0 \mid b - \epsilon = x$ . Then  $-\infty < x - \frac{\epsilon}{2} < x < x + \frac{\epsilon}{2} < x + \epsilon = b$ . Therefore  $x \in (x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2}) \subseteq (-\infty, b)$ , so  $(-\infty, b)$  is open.  $\square$

•

**Proposition 5.3.4.**  $(a, b)$  is open.

*Proof.* In class notes.  $\square$

•

**Proposition 5.3.5.** If  $a < b$ , then  $[a, b]$  is closed.

*Proof.* Consider  $\mathbb{R} \setminus [a, b] = (-\infty, a) \cup (b, \infty)$ . By Proposition 5.3.3,  $(-\infty, a)$  is open. By Proposition ra.hw5.5b,  $(b, \infty)$  is open. By Proposition 5.3.6, the union of two open sets is open. Therefore  $\mathbb{R} \setminus [a, b]$  is open, so  $[a, b]$  is closed.  $\square$

•

**Proposition 5.3.6.** If  $A$  and  $B$  are open, then  $A \cup B$  is open.

*Proof.* Since  $A$  is open,  $\forall x_A \in A \exists (a_A, b_A) \subseteq A \mid x_A \in (a_A, b_A)$ . Since  $B$  is open,  $\forall x_B \in B \exists (a_B, b_B) \subseteq B \mid x_B \in (a_B, b_B)$ .

Let  $x \in A \cup B$ . If  $x \in A$ , then per above  $\exists (a_A, b_A) \subseteq A \subseteq A \cup B \mid x_A \in (a_A, b_A)$ . If  $x \in B$ , then per above  $\exists (a_B, b_B) \subseteq B \subseteq A \cup B \mid x_B \in (a_B, b_B)$ . Therefore  $A \cup B$  is open.  $\square$

•

**Proposition 5.3.7.** If  $A$  and  $B$  are open, then  $A \cap B$  is open.

*Proof.* Since  $A$  is open,  $\forall x_A \in A \exists (a_A, b_A) \subseteq A \mid x_A \in (a_A, b_A)$ . Since  $B$  is open,  $\forall x_B \in B \exists (a_B, b_B) \subseteq B \mid x_B \in (a_B, b_B)$ .

Let  $x \in A \cap B$ . Then  $x \in A$  and  $x \in B$ , so  $\exists (a_A, b_A) \subseteq A \mid x_A \in (a_A, b_A)$ , and  $\exists (a_B, b_B) \subseteq B \mid x_B \in (a_B, b_B)$ . Let  $a = \max\{a_A, a_B\}$ , and  $b = \min\{b_A, b_B\}$ . Since  $x > a$  and  $x < b$ ,  $x \in (a, b)$ . Since  $(a, b) \subseteq (a_A, b_A) \subseteq A$  and  $(a, b) \subseteq (a_B, b_B) \subseteq B$ ,  $(a, b) \subseteq A \cap B$ . Therefore  $A \cap B$  is open.  $\square$

•

**Proposition 5.3.8.** If  $A$  and  $B$  are closed, then  $A \cup B$  is closed.

*Proof.* Since  $A$  is closed,  $\mathbb{R} \setminus A$  is open. Since  $B$  is closed,  $\mathbb{R} \setminus B$  is open.  $\mathbb{R} \setminus (A \cup B) = (\mathbb{R} \setminus A) \cap (\mathbb{R} \setminus B)$ . By Proposition 5.3.7 the intersection of two open sets is open. Therefore  $\mathbb{R} \setminus (A \cup B)$  is open, so  $A \cup B$  is closed.

 $\square$ 

•

**Proposition 5.3.9.** If  $A$  and  $B$  are closed, then  $A \cap B$  is closed.

*Proof.* Since  $A$  is closed,  $\mathbb{R} \setminus A$  is open. Since  $B$  is closed,  $\mathbb{R} \setminus B$  is open.  $\mathbb{R} \setminus (A \cap B) = (\mathbb{R} \setminus A) \cup (\mathbb{R} \setminus B)$ . By Proposition 5.3.6, the union of two open sets is open. Therefore  $\mathbb{R} \setminus (A \cap B)$  is open, so  $A \cap B$  is closed.

 $\square$ 

•

**Proposition 5.3.10.**  $\mathbb{R}$  is open and closed.

*Proof.* Let  $\epsilon > 0$ . Let  $x \in \mathbb{R}$ . Then  $x - \epsilon, x + \epsilon \in \mathbb{R}$ , and  $x \in (x - \epsilon, x + \epsilon)$ . Therefore  $\mathbb{R}$  is open.  $\mathbb{R}$  is closed because by Proposition 5.3.11,  $\mathbb{R} \setminus \mathbb{R} = \emptyset$  is open.

 $\square$ 

•

**Proposition 5.3.11.**  $\emptyset$  is open and closed.

*Proof.* To show that a set  $S$  is open, we must show that  $\forall x \in S \exists S' \subseteq S \mid x \in S'$  where  $S'$  is open. Since there are no  $x \in \emptyset$ , this condition is satisfied for  $\emptyset$ .  $\emptyset$  is closed because per Proposition 5.3.10,  $\mathbb{R} \setminus \emptyset = \mathbb{R}$  is open.

 $\square$ 

**Proposition 5.3.12.** Let  $x_1, x_2, \dots, x_n$  be real numbers. Let  $S$  be the finite set  $S = \{x_1, x_2, \dots, x_n\}$ . Then  $S$  is closed.

*Proof.* Consider  $\mathbb{R} \setminus S = (-\infty, x_1) \cup (x_1, x_2) \cup \dots \cup (x_{n-1}, x_n) \cup (x_n, \infty)$ .  $(-\infty, x_1)$  is open by Proposition 5.3.3.  $(x_n, \infty)$  is open by Proposition 5.3.2.

Consider  $(x_i, x_{i+1})$  where  $i \in \{1, 2, 3, \dots, n-1\}$ . Let  $x \in (x_i, x_{i+1})$ . Then since  $x > x_i$  and  $x < x_{i+1}$ ,  $\exists \epsilon > 0 \mid x_i + \epsilon = x$  and  $\exists \delta > 0 \mid x_{i+1} - \delta = x$ . Then  $x_i = x - \epsilon < x - \frac{\epsilon}{2} < x < x + \frac{\delta}{2} < x + \delta = x_{i+1}$ . Therefore  $x \in (x - \epsilon, x + \delta) \subseteq (x_i, x_{i+1})$ , so  $(x_i, x_{i+1})$  is open.

Finally, since  $\mathbb{R} \setminus S$ , by Proposition 5.3.2 (and induction)  $\mathbb{R} \setminus S$  is open. Therefore  $S$  is closed.  $\square$

**Proposition 5.3.13.** Let  $x_1, x_2, \dots, x_n$  be real numbers. Let  $S$  be the finite set  $S = \{x_1, x_2, \dots, x_n\}$ . Then  $S$  has no limit points.

*Proof.* Per Definition 5.7, we seek to show that (1)  $\forall x_i \in S \exists \delta_i$  such that  $\forall x_j \in D(x_j \neq x_i)$

$$|x_j - x_i| \geq \delta_i$$

and (2)  $\forall x \in \mathbb{R} \setminus S \exists \delta_x$  such that  $\forall x_i \in D$

$$|x_i - x| \geq \delta_x$$

(1) Let  $x_i \in S$ . Let  $\delta_i = \frac{1}{2} \min\{|x_i - x_k| \mid x_k \neq x_i\}$ . Then  $\forall x_j \neq x_i \in S$ ,

$$|x_i - x_j| \geq |x_i - x_k| > \delta_i$$

(2) Let  $x \in \mathbb{R} \setminus S$ . Let  $\delta_x = \frac{1}{2} \min\{|x - x_i| \mid x_i \in S\}$ . Then

$$|x_i - x| \geq \min\{|x - x_i| \mid x_i \in S\} > \delta_x$$

$\square$

**Definition 5.14.** Let  $S \subseteq \mathbb{R}$ . An **open cover** of  $S$  is a collection  $X = \{\mathcal{O}_\alpha \mid \alpha \in I\}$  where each set  $\mathcal{O}_\alpha$  is an open subset of  $\mathbb{R}$  such that

$$S \subseteq \bigcup_{\alpha \in I} \mathcal{O}_\alpha$$

(Here  $I$  is some set that indexes the  $\mathcal{O}_\alpha$ ).

**Definition 5.15.** If  $X' \subseteq X$  such that

$$S \subseteq \bigcup_{\mathcal{O}_\alpha \in X'} \mathcal{O}_\alpha$$

then  $X'$  is called a **subcover** of  $S$  contained in  $X$ . In addition, if  $X'$  is finite then we call  $X'$  a **finite subcover** of  $S$  contained in  $X$ .

**Definition 5.16. (Compactness.)** Let  $S \subseteq \mathbb{R}$ . We say that  $S$  is **compact** if every open cover of  $S$  contains a finite subcover.

**Definition 5.17.** Let  $S \subseteq \mathbb{R}$ . We say that  $S$  is **bounded** if  $\exists M > 0$  where  $S \subseteq [-M, M]$ .

**Remark.**  $S$  is bounded if and only if  $|s| \leq M \forall s \in S$ .

**Theorem 5.3.14. (Heine-Borel Theorem.)** Let  $S \subseteq \mathbb{R}$ .  $S$  is compact if and only if  $S$  is closed and bounded.

**Proposition 5.3.15.** Let  $x_1, x_2, \dots, x_n$  be real numbers. Let  $S$  be the finite set  $S = \{x_1, x_2, \dots, x_n\}$ . Then  $S$  is compact.

*Proof.* Let  $\{O_\alpha\}$  be an open cover of  $S$ . By definition of open cover,  $\forall i \exists O_{\alpha_i}$  such that  $x_i \in O_{\alpha_i}$ . Thus,  $\{O_{\alpha_1}, O_{\alpha_2}, \dots, O_{\alpha_n}\}$  is a finite subcover of  $S$ .  $\square$

**Proposition 5.3.16.** Let  $A$  and  $B$  be compact subsets of  $\mathbb{R}$ . Then  $A \cap B$  is compact.

*Proof.* Since  $A \cap B \subseteq A$ ,  $A \cap B \subseteq [-M_A, M_A]$ . Therefore  $A \cap B$  is bounded.

Since  $A \cap B$  is closed and bounded, by the Heine-Borel Theorem (Theorem ra.heine-borel.thm),  $A \cap B$  is compact.  $\square$

**Proposition 5.3.17.** Let  $A$  and  $B$  be compact subsets of  $\mathbb{R}$ . Then  $A \cup B$  is compact.

*Proof.* Let  $M = \max\{M_A, M_B\}$ . Note that  $[-M_A, M_A] \subseteq [M, M]$  and  $[-M_B, M_B] \subseteq [-M, M]$ . This implies  $A \subseteq [-M, M]$  and  $B \subseteq [-M, M]$ . Therefore  $A \cup B \subseteq [-M, M]$ .

Since  $A \cup B$  is closed and bounded, by the Heine-Borel Theorem (Theorem ra.heine-borel.thm),  $A \cup B$  is compact.  $\square$

**Theorem 5.3.18.** Let  $f : D \rightarrow \mathbb{R}$  be continuous on  $D$ . If  $X \subseteq D$  and  $X$  is compact (closed and bounded), then

$$f(\bar{x}) = \{f(x) \mid x \in X\}$$

is compact (closed and bounded).

**Corollary 5.3.18.1.** Suppose  $f : D \rightarrow \mathbb{R}$  where  $D$  is closed and bounded. Then there exists  $a, b \in D$  where  $f(a)$  is the min of  $f$  on  $D$  and  $f(b)$  is the max of  $f$  on  $D$ .

### 5.3.2 Homework 6

**Definition 5.18. (Uniform Continuity.)** Let  $D \subseteq \mathbb{R}$  and let  $f : D \rightarrow \mathbb{R}$ . We say that  $f$  is **uniformly continuous** on  $D$  if  $\forall \epsilon > 0 \exists \delta > 0$  where

$$x, y \in D \text{ and } 0 < |x - y| < \delta \implies |f(x) - f(y)| < \epsilon$$

**Theorem 5.3.19. (Uniform continuity implies continuity.)** Suppose  $f : D \rightarrow \mathbb{R}$  where  $D \subseteq \mathbb{R}$ . If  $f$  is uniformly continuous on  $D$ , then  $f$  is continuous at every  $a \in D$ .

## 5.4 More Theorems

**Theorem 5.4.1. Fubini's Theorem.** Let  $h : \mathbb{R}^2 \rightarrow \mathbb{R}$  be a continuous function such that  $\int \int_{\mathbb{R}^2} |h(x, y)| dx dy < \infty$ . Then

$$\int \int_{\mathbb{R}^2} h(x, y) dx dy = \int_{\mathbb{R}} \left( \int_{\mathbb{R}} h(x, y) dx \right) dy = \int_{\mathbb{R}} \left( \int_{\mathbb{R}} h(x, y) dy \right) dx$$

## 5.5 Problems from Practice Math GRE Subject Tests

38. Let  $A$  and  $B$  be nonempty subsets of  $\mathbb{R}$  and let  $f : A \rightarrow B$  be a function. If  $C \subseteq A$  and  $D \subseteq B$ , which of the following must be true?
- (A)  $C \subseteq f^{-1}(f(C))$
  - (B)  $D \subseteq f(f^{-1}(D))$
  - (C)  $f^{-1}(f(C)) \subseteq C$

**Solution 38.** (A) Neither of the equalities should hold – these are in fact nonsense statements, as one side lies in  $A$  and the other in  $B$ . To unravel the remaining two sets,

$$f^{-1}(f(C)) = \{x \in A : f(x) \in f(C)\}, \quad f(f^{-1}(D)) = f(\{y \in A : f(y) \in D\})$$

Clearly the second set must always be contained in  $D$ , but not the other way around. Similarly the first set certainly contains all  $c \in C$  (as  $f(c) \in f(C)$ ) but not the other way around.

47. The function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is defined as follows.

$$f(x) = \begin{cases} 3x^2 & \text{if } x \in \mathbb{Q} \\ -5x^2 & \text{if } x \notin \mathbb{Q} \end{cases}$$

Which of the following is true?

- (A)  $f$  is discontinuous at all  $x \in \mathbb{R}$ .
- (B)  $f$  is continuous only at  $x = 0$  and differentiable only at  $x = 0$ .
- (C)  $f$  is continuous only at  $x = 0$  and nondifferentiable at all  $x \in \mathbb{R}$ .
- (D)  $f$  is continuous at all  $x \in \mathbb{Q}$  and nondifferentiable at all  $x \in \mathbb{R}$ .
- (E)  $f$  is continuous at all  $x \notin \mathbb{Q}$  and nondifferentiable at all  $x \in \mathbb{R}$ .

**Solution 47.** (B) A classic kind of problem. We are clearly continuous and differentiable at 0. Anywhere else, near a rational number there is an irrational number and vice versa. Therefore there can be no continuity anywhere but at 0, and hence no differentiability either.

57. For each positive integer  $n$ , let  $x_n$  be a real number in the open interval  $\left(0, \frac{1}{n}\right)$ . Which of the following statements must be true?

I.  $\lim_{n \rightarrow \infty} x_n = 0$

II. If  $f$  is a continuous real-valued function defined on  $(0, 1)$ , then  $\{f(x_n)\}_{n=1}^{\infty}$  is a Cauchy sequence.

III. If  $g$  is a uniformly continuous real-valued function defined on  $(0, 1)$ , then  $\lim_{n \rightarrow \infty} g(x_n)$  exists.

- (A) I only      (B) I and II only      (C) I and III only      (D) II and III only      (E) I, II, and III

**Solution 57.** (C) I is true, since  $\lim_{n \rightarrow \infty} x_n$  must be bounded between 0 and  $\lim_{n \rightarrow \infty} 1/n = 0$ . Unfortunately,  $x_n$  does not converge inside  $(0, 1)$ . There is no reason therefore that  $f(x_n)$  should be a convergent sequence – suppose that  $f(x) = 1/x$ , so that  $f(x_n)$  is certainly not Cauchy. However, if  $g$  is uniformly continuous, then  $g$  extends to a continuous function on  $[0, 1]$ . Now  $x_n$  is a convergent sequence, so  $\lim_{n \rightarrow \infty} g(x_n) = g(\lim_{n \rightarrow \infty} x_n) = g(0)$  exists.

60. A real-valued function  $f$  defined on  $\mathbb{R}$  has the following property.

For every positive number  $\epsilon$ , there exists a positive number  $\delta$  such that

$$|f(x) - f(1)| \geq \epsilon \text{ whenever } |x - 1| \geq \delta.$$

This property is equivalent to which of the following statements about  $f$ ?

- (A)  $f$  is continuous at  $x = 1$ .
- (B)  $f$  is discontinuous at  $x = 1$ .
- (C)  $f$  is unbounded.
- (D)  $\lim_{|x| \rightarrow \infty} |f(x)| = \infty$
- (E)  $\int_0^{\infty} |f(x)| dx = \infty$

**Solution 60.** (D) While it looks like this is the opposite of continuity, that should read ‘there exists  $\epsilon > 0$ ’. What the statement says is that we not only get arbitrarily far away from  $f(1)$ , but we must for all  $x$  sufficiently far away from 1. So as  $|x|$  gets very large, so does  $|f(x)|$ .

63. For any nonempty sets  $A$  and  $B$  of real numbers, let  $A \cdot B$  be the set defined by

$$A \cdot B = \{xy : x \in A \text{ and } y \in B\}.$$

If  $A$  and  $B$  are nonempty bounded sets of real numbers and if  $\sup(A) > \sup(B)$ , then  $\sup(A \cdot B) =$

- (A)  $\sup(A) \sup(B)$
- (B)  $\sup(A) \inf(B)$
- (C)  $\max\{\sup(A) \sup(B), \inf(A) \inf(B)\}$
- (D)  $\max\{\sup(A) \sup(B), \sup(A) \inf(B)\}$
- (E)  $\max\{\sup(A) \sup(B), \inf(A) \sup(B), \inf(A) \inf(B)\}$

**Solution 63.** (E) The supremum is either going to be the product of the two largest positive numbers in  $A$  and  $B$  or the product of the two smallest negative numbers in  $A$  and  $B$ . That means we should look for  $\sup \cdot \sup$  or  $\inf \cdot \inf$ . However, it might be the case that  $B$  contains only negative numbers and  $A$  contains only positive numbers. Then the largest value in  $A \cdot B$  will be attained by the smallest positive element of  $A$  and the largest negative element of  $B$ , giving us our third option:  $\inf A \cdot \sup B$ .

## 6 Probability

These are my notes from taking Math 505A at USC taught by Sergey Lototsky, Math 541A at USC taught by Steven Heilman, ISE 620 at USC taught by Sheldon Ross (as well as the corresponding textbooks *Introduction to Probability Models* [Ross, 2014] and *Stochastic Processes* [Ross, 2008] by Sheldon Ross) and the textbook *Probability and Random Processes* (Grimmett and Stirzaker) 3rd edition [Grimmett and Stirzaker, 2001], Math 541B at USC taught by Stanislav Minsker, Statistics 100B at UCLA taught by Nicolas Christou, as well as a few other sources I cite within the text.

### 6.1 To Know for Math 505A Midterm 1 (Discrete Random Variables)

#### 6.1.1 Definitions

**Definition 6.1.** The **probability mass function** of a discrete random variable  $X$  is the function  $f : \mathbb{R} \rightarrow [0, 1]$  given by  $f(x) = \Pr(X = x)$ .

**Definition 6.2.** The **(cumulative) distribution function** of a discrete random variable  $F$  is given by

$$F(x) = \sum_{i:x_i \leq x} f(x_i)$$

**Definition 6.3.** The **joint probability mass function**  $f : \mathbb{R}^2 \rightarrow [0, 1]$  of two discrete random variables  $X$  and  $Y$  is given by

$$f(x, y) = \Pr(X = x \cap Y = y)$$

**Definition 6.4.** The **joint distribution function**  $F : \mathbb{R}^2 \rightarrow [0, 1]$  is given by

$$F(x, y) = \Pr(X \leq x \cap Y \leq y)$$

**Definition 6.5.** If  $\Pr(B) > 0$  then the **conditional probability** that  $A$  occurs given that  $B$  occurs is defined to be

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

**Definition 6.6. (Independent sets.)** Let  $A_1, A_2, \dots$  be subsets of a sample space  $\Omega$ , and let  $\mathbb{P}$  be a probability law on  $\Omega$ . We say that  $A_1, A_2, \dots$  are **independent** if for any finite subset  $S$  of  $\{1, 2, \dots\}$ , we have

$$\mathbb{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbb{P}(A_i)$$

**Definition 6.7. (Notation.)** Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable. Let  $B \subseteq \mathbb{R}$ . We define  $\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\}$ .

**Definition 6.8. (Independence of random variables.)** Random variables  $X_1, X_2, \dots$  are **independent** if for every  $B_1, B_2, \dots \subseteq \mathbb{R}$ , the events  $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \dots$  are independent; that is,

$$\mathbb{P}\left(\bigcap_{i=1}^n \{X_i \in B_i\}\right) = \prod_{i=1}^n \mathbb{P}(\{X_i \in B_i\})$$

**Remark.** A more informal definition is as follows: Two random variables  $X$  and  $Y$  are **independent** if and only if  $\Pr(X \cap Y) = \Pr(X)\Pr(Y)$ .

**Theorem 6.1.1. (Law of total probability).** If  $X$  is a random variable and  $Y$  is a discrete random variable taking on values  $y_1, y_2, \dots, y_n$ , then  $\Pr(X) = \sum_i \Pr(X | Y = y_i) \cdot \Pr(Y = y_i)$ . (Can be used to prove independence.)

**Lemma 6.1.2.** Let  $B_1, A_1, A_2, \dots, A_m$  be subsets of a sample space  $\Omega$  and let  $\mathbb{P}$  be a probability law on  $\Omega$ . Let  $X_1, X_2, \dots, X_m$ , and  $Z$  be random variables. If  $Z$  is independent of  $X_i$  for all  $i \in 1, \dots, m$ , then

$$\mathbb{P}\left(\bigcap_{i=1}^m \{X_i \in A_i\} \cap Z \in B_1\right) = \mathbb{P}\left(\bigcap_{i=1}^m \{X_i \in A_i\}\right) \mathbb{P}(Z \in B_1).$$

That is,  $\bigcap_{i=1}^m X_i$  is independent of  $Z$ .

*Proof.* Note that by the definition of conditional probability,

$$\begin{aligned} & \mathbb{P}\left(\bigcap_{i=1}^m X_i \in A_i \mid Z \in B_1\right) = \frac{\mathbb{P}(\bigcap_{i=1}^m \{X_i \in A_i\} \cap Z \in B_1)}{\mathbb{P}(Z \in B_1)} \\ &= \frac{\prod_{j=1}^{m-1} [\mathbb{P}(X_j \in A_j \mid \bigcap_{k=j+1}^m \{X_k \in A_k\} \cap Z \in B_1)] \cdot \mathbb{P}(X_m \in A_m \cap Z \in B_1)}{\mathbb{P}(Z \in B_1)} \\ &= \frac{\prod_{j=1}^{m-1} [\mathbb{P}(X_j \in A_j \mid \bigcap_{k=j+1}^m \{X_k \in A_k\}, Z \in B_1)] \cdot \mathbb{P}(X_m \in A_m \mid Z \in B_1) \mathbb{P}(Z \in B_1)}{\mathbb{P}(Z \in B_1)} \end{aligned}$$

Canceling and using the independence of  $X_i$  and  $Z$  for all  $i$ , we have

$$= \prod_{j=1}^{m-1} [\mathbb{P}(X_j \in A_j \mid \bigcap_{k=j+1}^m \{X_k \in A_k\})] \cdot \mathbb{P}(X_m \in A_m) = \mathbb{P}\left(\bigcap_{i=1}^m \{X_i \in A_i\}\right)$$

Then the result follows since

$$\mathbb{P}\left(\bigcap_{i=1}^m \{X_i \in A_i\} \cap Z \in B_1\right) = \mathbb{P}\left(\bigcap_{i=1}^m X_i \in A_i \mid Z \in B_1\right) \mathbb{P}(Z \in B_1) = \mathbb{P}\left(\bigcap_{i=1}^m \{X_i \in A_i\}\right) \mathbb{P}(Z \in B_1).$$

□

**Definition 6.9.** Two random variables  $X$  and  $Y$  are **uncorrelated** if  $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ .

**Proposition 6.1.3.** (a) Two random variables are uncorrelated if and only if their covariance  $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$  equals 0.

(b) If  $X$  and  $Y$  are independent then they are uncorrelated.

**Theorem 6.1.4.** If  $X$  and  $Y$  are independent and  $g, h : \mathbb{R} \rightarrow \mathbb{R}$ , then  $g(X)$  and  $h(Y)$  are also independent.

**Definition 6.10.** We say that a nonnegative random variable  $X$  is **lattice** with period  $d$  if

$$\sum_{n=0}^{\infty} \Pr(X = nd) = 1$$

and  $d$  is the largest number that satisfies this equation.

**Remark.** Most common discrete random variables are lattice, but not all discrete random variables are. For example, the discrete random variable  $X$  such that

$$\Pr(X = 1/i) = p_i > 0, \quad i \geq 1$$

such that  $\sum_{i=1}^{\infty} p_i = 1$  is not lattice. Another example is

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ \sqrt{2} & \text{with probability } 1/2 \end{cases}$$

because there is no  $d$  such that there are integers  $m$  and  $n$  so that

$$1 = md$$

$$\sqrt{2} = nd.$$

(If there were, then we would have

$$\frac{n}{m} = \frac{nd}{md} = \sqrt{2}$$

but there are no integers  $n$  and  $m$  such that this is true, since  $\sqrt{2}$  is irrational.)

### 6.1.2 Conditioning

**Definition 6.11.** The **conditional distribution function** of  $Y$  given  $X = x$ , written  $F_{Y|X}(\cdot \mid x)$ , is defined by

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

**Definition 6.12.** The **conditional probability mass function** of  $Y$  given  $X = x$ , written  $f_{Y|X}(\cdot | x)$ , is defined by

$$f_{Y|X}(y | x) = \Pr(Y = y | X = x)$$

**Definition 6.13 (Conditional Expectation (Math 541B definition)).** Let  $X, Y$  be random variables, and assume that  $\mathbb{E}|X| < \infty$ . Then the conditional expectation of  $X$  given  $Y$ , denoted  $\mathbb{E}(X | Y)$ , is a function of  $Y$ :  $g(Y) = \mathbb{E}(X | Y)$  such that for any bounded  $f$ ,

$$\mathbb{E}(Xf(Y)) = \mathbb{E}(g(Y)f(Y))$$

Now assume that  $\mathbb{E}X^2 < \infty$ ,  $\mathbb{E}Y^2 < \infty$ . Consider the Hilbert space  $(\mathbb{H}, \langle \cdot, \cdot \rangle)$  consisting of all random variables with finite second moment and  $\langle X, Y \rangle := \mathbb{E}(XY)$ . Consider  $L(Y) = \{\text{all random variables of the form } X = \phi(Y) \text{ where } \mathbb{E}Z^2 < \infty\}$ <sup>1</sup>. Then  $\mathbb{E}(X | Y)$  is the orthogonal projection of  $X$  onto the subspace  $L(Y)$ . Indeed,

$$\mathbb{E}(X \cdot f(Y)) = \langle X, f(Y) \rangle = \mathbb{E}(\mathbb{E}(X | Y)f(Y)) = \langle \mathbb{E}(X | Y), f(Y) \rangle$$

$$\iff \forall f, \langle X - \mathbb{E}(X | Y), f(Y) \rangle = 0.$$

(because  $\mathbb{E}[(X - \mathbb{E}(X | Y)) \cdot f(Y)] = \mathbb{E}(\mathbb{E}[Xf(Y) | Y] - \mathbb{E}[\mathbb{E}(X | Y)f(Y) | Y]) = \dots$ )

$(f(Y) \in L(Y))$ .

**Remark.** Think of expectation as the best summary of a random variable with one value, and conditional expectation as the best summary of one random variable as a function of another random variable.

**Remark.**

**Example 6.1.** Assume  $(X, Y)$  has joint pdf  $p(x, y)$ . Then  $\mathbb{E}(X | Y = y) = \int_{\mathbb{R}} x p_{X|Y}(x | y) dx$ , where  $p_{X|Y}(x | y)$  is the conditional density of  $X$  given  $Y$ ; that is,

$$p_{X|Y}(x | y) = p(x, y) / \int_{x \in \mathbb{R}} p(x, y) dx.$$

**Theorem 6.1.5. Iterated expectations:**

- (i)  $\mathbb{E}[\mathbb{E}(X | Y)] = \mathbb{E}(X)$  (**Law of Total Expectation**)
- (ii)  $\mathbb{E}[(X | Y) | Z] = \mathbb{E}(X | Y)$
- (iii)  $\mathbb{E}(E(XY | Y)) = \mathbb{E}(Y\mathbb{E}(X | Y))$

<sup>1</sup>clear that this subspace is linear: any linear combination of square integrable functions of  $Y$  is also square integrable. less clear that this is a closed space; that is, that any limit of functions in this space is also in this space. But it turns out that it is, so since this is a closed linear subspace.

*Proof.* (i) Discrete case:

$$\begin{aligned}\mathbb{E}[\mathbb{E}(X | Y)] &= \sum_y \mathbb{E}(X | Y = y) \Pr(Y = y) = \sum_y \sum_x x \Pr(X = x | Y = y) \Pr(Y = y) \\ &= \sum_y \sum_x x \Pr(X = x \cap Y = y) = \sum_x x \sum_y \Pr(X = x \cap Y = y) = \sum_x x \Pr(X = x) = \mathbb{E}(X)\end{aligned}$$

Continuous case:

$$\mathbb{E}[\mathbb{E}(X | Y)] = \int_{-\infty}^{\infty} \mathbb{E}(X | Y = y) f_Y(y) dy = \text{(by definition 1.75)} \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx \right) f_Y(y) dy$$

$$\text{(by Fubini's Theorem, Theorem 5.4.1)} \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} x f_{X|Y}(x | y) f_Y(y) dy \right) dx = \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}(X)$$

□

**Definition 6.14. Conditional Variance:**  $\text{Var}(X | Y) = \mathbb{E}[(X - \mathbb{E}(X | Y))^2 | Y]$

**Theorem 6.1.6 (Conditional Expectation as a Random Variable).** (i) Let  $X, Y$  be random variables such that  $(X, Y)$  is uniformly distributed on the triangle  $\{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0, x + y \leq 1\}$ . Then

$$\mathbb{E}(X|Y) = \frac{1}{2}(1 - Y).$$

(ii) Total Expectation Theorem:

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X).$$

- If  $X$  is a random variable, and if  $f(t) := \mathbb{E}(X - t)^2$ ,  $t \in \mathbb{R}$ , then the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is uniquely minimized when  $t = \mathbb{E}X$ . A similar minimizing property holds for conditional expectation. Let  $h : \mathbb{R} \rightarrow \mathbb{R}$ . Then the quantity  $\mathbb{E}(X - h(Y))^2$  is minimized among all functions  $h : \mathbb{R} \rightarrow \mathbb{R}$  when  $h(Y) = \mathbb{E}(X|Y)$ .

(iii)

$$\mathbb{E}(Xh(Y)|Y) = h(Y)\mathbb{E}(X|Y).$$

$$\mathbb{E}([\mathbb{E}(X|h(Y))|Y] = \mathbb{E}(X|h(Y)).$$

(iv)

$$\mathbb{E}(X|X) = X.$$

$$\mathbb{E}(X + Y|Z) = \mathbb{E}(X|Z) + \mathbb{E}(Y|Z).$$

(v) If  $Z$  is independent of  $X$  and  $Y$ , then

$$\mathbb{E}(X|Y, Z) = \mathbb{E}(X|Y).$$

(Here  $\mathbb{E}(X|Y, Z)$  is notation for  $\mathbb{E}(X|(Y, Z))$  where  $(Y, Z)$  is interpreted as a random vector, so that  $X$  is conditioned on the random vector  $(Y, Z)$ .)

*Proof.* (i) Note that since  $0 \leq x$  and  $x + y \leq 1$ , conditional on  $y = y \geq 0$   $x$  is uniformly distributed on  $[0, 1 - y]$ . That is,

$$\Pr(X \leq x \mid Y = y) = \begin{cases} 0 & x < 0 \\ x/(1-y) & 0 \leq x < 1-y \\ 1 & x \geq 1-y \end{cases}$$

Since the expected value of a random variable uniformly distributed on  $[a, b]$  is  $(a + b)/2$ , it follows that  $\mathbb{E}(X \mid Y = y) = (0 + 1 - y)/2 = (1/2)(1 - y)$ . Therefore  $\boxed{\mathbb{E}(X \mid Y) = \frac{1}{2}(1 - Y)}$ .

(ii) • Discrete case:

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X \mid Y)] &= \sum_y \mathbb{E}(X \mid Y = y) \Pr(Y = y) = \sum_y \sum_x x \Pr(X = x \mid Y = y) \Pr(Y = y) \\ &= \sum_y \sum_x x \Pr(X = x \cap Y = y) = \sum_x x \sum_y \Pr(X = x \cap Y = y) = \sum_x x \Pr(X = x) = \mathbb{E}(X) \end{aligned}$$

• Continuous case:

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X \mid Y)] &= \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y) dy = \text{(by definition 1.75)} \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx \right) f_Y(y) dy \\ &\quad \text{(by Fubini's Theorem)} \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y) dy \right) dx = \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}(X) \end{aligned}$$

• Next we will show that the quantity  $\mathbb{E}(X - h(Y))^2$  is minimized among all functions  $h : \mathbb{R} \rightarrow \mathbb{R}$  when  $h(Y) = \mathbb{E}(X \mid Y)$ . We seek

$$\arg \min_{\{h: \mathbb{R} \rightarrow \mathbb{R}\}} \mathbb{E}(X - h(Y))^2 = \arg \min_{\{h: \mathbb{R} \rightarrow \mathbb{R}\}} [\mathbb{E}(X^2) - 2\mathbb{E}[h(Y)]\mathbb{E}(X) + \mathbb{E}[h(Y)^2]].$$

This expression is quadratic in  $\mathbb{E}[h(Y)]$ . Differentiating with respect to  $\mathbb{E}[h(Y)]$  and setting equal to 0, we have

$$2\mathbb{E}[h(Y)] - 2\mathbb{E}(X) = 0 \iff \mathbb{E}[h(Y)] = \mathbb{E}(X) \implies \boxed{\arg \min_{\{h: \mathbb{R} \rightarrow \mathbb{R}\}} \mathbb{E}(X - h(Y))^2 = \mathbb{E}(X \mid Y)}$$

(iii) • Discrete case:

$$\mathbb{E}(Xh(Y) \mid Y) = \sum_{x \in \mathbb{R}} x \cdot h(Y) \cdot \Pr(X = x \mid Y) = h(Y) \sum_{x \in \mathbb{R}} x \cdot \Pr(X = x \mid Y) = h(Y)\mathbb{E}(X \mid Y).$$

Continuous case:

$$\mathbb{E}(Xh(Y) \mid Y) = \int_{x \in \mathbb{R}} x \cdot h(Y) \cdot f_{X|Y}(x) dx = h(Y) \int_{x \in \mathbb{R}} x \cdot f_{X|Y}(x) dx = h(Y)\mathbb{E}(X \mid Y).$$

- Discrete case: Note that

$$\mathbb{E}[X \mid h(Y)] = \sum_{x \in \mathbb{R}} x \Pr(X = x \mid h(Y)) = \sum_{x \in \mathbb{R}} x \mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)]$$

where  $\mathbf{1}_{\{X=x\}}$  is an indicator variable for  $X$  taking on the value  $x$ . Note that  $\mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)]$  is a function of  $Y$  (and a random variable). Then we have

$$\begin{aligned} \mathbb{E}(\mathbb{E}(X \mid h(Y)) \mid Y) &= \mathbb{E}\left(\sum_{x \in \mathbb{R}} x \mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)] \mid Y\right) = \sum_{x \in \mathbb{R}} \mathbb{E}[x \mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)] \mid Y] \\ &= (\text{by the previous result}) \sum_{x \in \mathbb{R}} \mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)] \mathbb{E}[x \mid Y] = \sum_{x \in \mathbb{R}} \Pr(X = x \mid h(Y)) \cdot x = \mathbb{E}(X \mid h(Y)). \end{aligned}$$

Continuous case: Note that

$$\mathbb{E}[X \mid h(Y)] = \int_{x \in \mathbb{R}} x f_{X|h(Y)}(x) dx.$$

Note that for a fixed  $x$ ,  $f_{X|h(Y)}(x)$  is a function of  $Y$  (and a random variable). Then we have

$$\begin{aligned} \mathbb{E}(\mathbb{E}(X \mid h(Y)) \mid Y) &= \mathbb{E}\left(\int_{x \in \mathbb{R}} x f_{X|h(Y)}(x) dx \mid Y\right) = \int_{x \in \mathbb{R}} \mathbb{E}[x \cdot f_{X|h(Y)}(x) \mid Y] dx \\ &= (\text{by the previous result}) \int_{x \in \mathbb{R}} f_{X|h(Y)}(x) \mathbb{E}[x \mid Y] dx = \int_{x \in \mathbb{R}} f_{X|h(Y)}(x) \cdot x = \mathbb{E}(X \mid h(Y)). \end{aligned}$$

- (iv) • Discrete case:

$$\mathbb{E}(X \mid X) = \sum_{x \in \mathbb{R}} x \Pr(X = x \mid X)$$

Note that

$$\Pr(X = x \mid X) = \begin{cases} 1 & x = X \\ 0 & \text{otherwise} \end{cases}$$

so we have

$$\mathbb{E}(X \mid X) = \sum_{x \in \mathbb{R}} x \Pr(X = x \mid X) = \dots + 0 + 0 + X + 0 + 0 + \dots = X.$$

Continuous case:

$$\mathbb{E}(X \mid X) = \int_{x \in \mathbb{R}} x \cdot dF_X$$

Note that

$$\Pr(X \leq x \mid X) = F_{X|X}(x) = \begin{cases} 0 & x < X \\ 1 & x \geq X \end{cases}$$

so we have

$$\mathbb{E}(X \mid X) = \int_{x \in \mathbb{R}} x \cdot dF_X = X.$$

- Discrete case:

$$\begin{aligned}
\mathbb{E}(X + Y \mid Z = z) &= \sum_{x \in \mathbb{Z}} \sum_{y \in \mathbb{Z}} (x + y) \Pr(X = x, Y = y \mid Z) \\
&= \sum_{x \in \mathbb{Z}} \sum_{y \in \mathbb{Z}} x \Pr(X = x, Y = y \mid Z) + \sum_{x \in \mathbb{Z}} \sum_{y \in \mathbb{Z}} y \Pr(X = x, Y = y \mid Z) \\
&= \sum_{x \in \mathbb{R}} x \Pr(X = x \mid Z) + \sum_{y \in \mathbb{R}} y \Pr(Y = y \mid Z) = \mathbb{E}(X \mid Z) + \mathbb{E}(Y \mid Z).
\end{aligned}$$

Continuous case:

$$\begin{aligned}
\mathbb{E}(X + Y \mid Z = z) &= \int_{x \in \mathbb{R}} \int_{y \in \mathbb{R}} (x + y) \Pr(X = x, Y = y \mid Z) dy dx \\
&= \int_{x \in \mathbb{R}} \int_{y \in \mathbb{R}} x \Pr(X = x, Y = y \mid Z) dy dx + \int_{x \in \mathbb{R}} \int_{y \in \mathbb{R}} y \Pr(X = x, Y = y \mid Z) dy dx \\
&= \int_{x \in \mathbb{R}} x \Pr(X = x \mid Z) dx + \int_{y \in \mathbb{R}} y \Pr(Y = y \mid Z) dy = \mathbb{E}(X \mid Z) + \mathbb{E}(Y \mid Z).
\end{aligned}$$

(v) Note by the definition of conditional probability that

$$\mathbb{P}(X = x \mid Y = y \cap Z = z) = \frac{\mathbb{P}(X = x \cap \{Y = y \cap Z = z\})}{\mathbb{P}(Y = y \cap Z = z)}$$

Using the independence of  $X$  and  $Y$  from  $Z$  and Lemma 6.1.2, we can express this as

$$= \frac{\mathbb{P}(X = x \cap Y = y) \mathbb{P}(Z = z)}{\mathbb{P}(Y = y) \mathbb{P}(Z = z)} = \frac{\mathbb{P}(X = x \cap Y = y)}{\mathbb{P}(Y = y)} = \mathbb{P}(X = x \mid Y = y)$$

by the definition of conditional probability. So  $\mathbb{P}(X = x \mid Y = y, Z) = \mathbb{P}(X = x, Y = y)$ . Therefore we have (in the discrete case)

$$\mathbb{E}(X \mid Y = y, Z) = \sum_{x \in \mathbb{R}} x \mathbb{P}(X = x \mid Y = y, Z) = \sum_{x \in \mathbb{R}} x \mathbb{P}(X = x, Y = y) = \mathbb{E}(X \mid Y = y) = g(y)$$

which implies that  $\mathbb{E}(X \mid Y, Z) = g(Y) = \mathbb{E}(X \mid Y)$ . In the continuous case, note that

$$F_{X|Y,Z}(x) = \mathbb{P}(X \leq x \mid Y, Z) = \mathbb{P}(X \leq x \mid Y) = F_{X|Y}(x)$$

by Lemma 6.1.2. Therefore we have

$$\mathbb{E}(X \mid Y, Z) = \int_{x \in \mathbb{R}} x dF_{X|Y,Z}(x) = \int_{x \in \mathbb{R}} x dF_{X|Y}(x) = \mathbb{E}(X \mid Y).$$

□

#### Corollary 6.1.6.1.

$$\mathbb{E}(h(Y) \mid Y) = h(Y).$$

*Proof.* Use the first result in part (iii) of Theorem 6.1.6 with  $X = 1$  and note that  $\mathbb{E}(1 | Y) = 1$ .

□

**Corollary 6.1.6.2.**

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

*Proof.* Use the second result in part (iii) of Theorem 6.1.6 with  $h(Y) = Y$ .

□

**Lemma 6.1.7.** If  $X \geq Z$  then  $\mathbb{E}(X|Y) \geq \mathbb{E}(Z|Y)$ .

*Proof.* Suppose that  $X$  and  $Z$  are nonnegative. Note that

$$\mathbb{E}(X | Y) = \int_0^\infty \Pr(X > t | Y) dt \geq \int_0^\infty \Pr(Z > t | Y) dt = \mathbb{E}(Z | Y)$$

where the third step follows since  $X \geq Z$ . If  $X$  and  $Z$  are not nonnegative, then

$$\begin{aligned} \mathbb{E}(X | Y) &= \mathbb{E}(\max\{X, 0\} | Y) - \mathbb{E}(\max\{-X, 0\} | Y) \\ &= \int_0^\infty \Pr(\max\{X, 0\} > t | Y) dt - \int_0^\infty \Pr(\max\{-X, 0\} > t | Y) dt \\ &\geq \int_0^\infty \Pr(\max\{Z, 0\} > t | Y) dt - \int_0^\infty \Pr(\max\{-Z, 0\} > t | Y) dt = \mathbb{E}(Z | Y) \end{aligned}$$

where the inequality follows since  $X \geq Z$ .

□

### 6.1.3 Convolution

**Theorem 6.1.8. Sums of random variables.** If  $X$  and  $Y$  are independent then

$$\Pr(X + Y = z) = f_{X+Y}(z) = \sum_x f_X(x) f_Y(z - x) = \sum_y f_X(z - y) f_Y(y)$$

**Remark. Convolution on the integers.** Let  $X, Y$  be independent integer-valued random variables. Let  $t \in \mathbb{Z}$ .

$$\begin{aligned} \Pr(X + Y = t) &= \sum_{j,k \in \mathbb{Z}: j+k=t} \Pr(X = j, Y = k) = \sum_{j \in F} \Pr(X = j, Y = t - j) = \sum_{j \in \mathbb{Z}} \Pr(X = j) \Pr(Y = t - j) \\ &= \sum_{j \in \mathbb{Z}} p_X(j) p_Y(t - j) \end{aligned}$$

**Definition 6.15.** Let  $g, h : \mathbb{Z} \rightarrow \mathbb{R}$  be functions. The **convolution** of  $g$  and  $h$ , denoted by  $g * h$ , is a function  $g * h : \mathbb{Z} \rightarrow \mathbb{R}$  defined by

$$(g * h)(t) = \sum_{j \in \mathbb{Z}} g(j)h(t - j) \quad \forall t \in \mathbb{Z}$$

**Definition 6.16. (Convolution on the real line.)** Let  $g, h : \mathbb{R} \rightarrow \mathbb{R}$  be functions. The **convolution** of  $g$  and  $h$ , denoted by  $g * h$ , is a function  $g * h : \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$(g * h)(t) = \int_{-\infty}^{\infty} g(x)h(t - x)dx, \quad \forall t \in \mathbb{R}$$

**Proposition 6.1.9.** Let  $X, Y$  be two continuous independent random variables such that  $\Pr(X + Y \leq t)$  is differentiable with respect to  $t \in \mathbb{R}$ . Then

$$f_{X+Y}(t) = (f_X * f_Y)(t), \quad \forall t \in \mathbb{R}$$

*Proof.*

$$\Pr(X + Y \leq t) = \int_{\{(x,y) \in \mathbb{R}^2 : x+y \leq t\}} f_{X,Y}(x,y)dxdy = \int_{x=-\infty}^{x=\infty} \int_{y=-\infty}^{y=t-x} f_X(x)f_Y(y)dydx$$

Then, since  $\Pr(X + Y \leq t)$  is differentiable with respect to  $t$ , we have by the Fundamental Theorem of Calculus

$$f_{X+Y}(t) = \frac{d}{dt} \Pr(X + Y \leq t) = \int_{x=-\infty}^{x=\infty} f_X(x) \frac{d}{dt} \int_{y=-\infty}^{y=t-x} f_Y(y)dydx = \int_{x=-\infty}^{x=\infty} f_X(x)f_Y(t-x)dx$$

□

**Example 6.2.** Let  $X, Y$  be independent standard Gaussian random variables. Then by Proposition 6.1.9,  $X + Y$  has density

$$\begin{aligned} f_{X+Y}(t) &= \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2}e^{-(t-x)^2/2}dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2+tx-t^2/2}dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-t/2)^2-t^2/4-t^2/2}dx \\ &= e^{-t^2/4} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-t/2)^2}dx = \frac{1}{2\pi} e^{-t^2/4} \int_{-\infty}^{\infty} e^{-x^2}dx \end{aligned}$$

Let  $x = y/\sqrt{2}, dx = dy/\sqrt{2}$ .

$$= \frac{1}{2\pi} e^{-t^2/4} \int_{-\infty}^{\infty} e^{-y^2/2} \cdot \frac{1}{\sqrt{2}} dy = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} e^{-t^2/4} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} e^{-t^2/4} \frac{1}{\sqrt{2\pi}}$$

which is the density for a Gaussian random variable distributed as  $\mathcal{N}(0, 1)$ .

**Definition 6.17. (Convolved cdfs; Ross's in-class definition from ISE 620.)** Suppose we have random variables  $X$  and  $Y$  with cdfs  $F_X$  and  $F_Y$  and pdfs  $f_X$  and  $f_Y$ . Then

$$\begin{aligned} (F_Y * F_X)(t) &= \Pr(X + Y \leq t) = \int_{-\infty}^{\infty} \Pr(X + Y \leq t \mid X = x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} F_Y(t - x) f_X(x) dx \end{aligned}$$

#### 6.1.4 Compound Random Variables

**Definition 6.18.** Let  $\{X_i\}$  be i.i.d. random variables. Let  $N$  be a random variable taking on positive integer values. Let

$$S = \sum_{i=1}^N X_i$$

Then  $S$  is a **compound random variable**.

**Proposition 6.1.10. (Wald's Equation.)** Let  $\{X_i\}$  be i.i.d. random variables with mean  $\mathbb{E}(X)$ . Let  $N$  be a random variable taking on positive integer values, and let  $S = \sum_{i=1}^N X_i$ . Then  $\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X)$ .

*Proof.*

$$\mathbb{E}(S \mid N) = \mathbb{E}\left(\sum_{i=1}^N X_i\right) = \sum_{i=1}^N \mathbb{E}(X_i) = N\mathbb{E}(X)$$

$$\mathbb{E}(S) = \mathbb{E}(\mathbb{E}[S \mid N]) = \mathbb{E}(N\mathbb{E}(X)) = \mathbb{E}(N)\mathbb{E}(X)$$

□

**Proposition 6.1.11.** Let  $\{X_i\}$  be i.i.d. random variables with mean  $\mathbb{E}(X)$ . Let  $N$  be a random variable taking on positive integer values, and let  $S = \sum_{i=1}^N X_i$ . Then  $\text{Cov}(N, S) = \mathbb{E}(X)\text{Var}(N)$ .

*Proof.* Let  $\mathbb{E}(X_i) = \mathbb{E}(X)$  for all  $i$ . We will use the result from Wald's Equation (Proposition 6.1.10:  $E(S) = \mathbb{E}(X)\mathbb{E}(N)$ ). We have

$$\text{Cov}(N, S) = \mathbb{E}(NS) - \mathbb{E}(N)\mathbb{E}(S) = \mathbb{E}[\mathbb{E}(NS \mid N)] - \mathbb{E}(N)\mathbb{E}(N)\mathbb{E}(X) = \mathbb{E}[N\mathbb{E}(S \mid N)] - \mathbb{E}(N)^2\mathbb{E}(X) \quad (6.1)$$

Note that

$$\mathbb{E}(S | N) = \mathbb{E}\left(\sum_{i=1}^N X_i\right) = \sum_{i=1}^N \mathbb{E}(X_i) = N\mathbb{E}(X) \implies \mathbb{E}[N\mathbb{E}(S | N)] = \mathbb{E}(N^2\mathbb{E}(X)) = \mathbb{E}(X)\mathbb{E}(N^2)$$

Plugging this into (6.1) yields

$$\text{Cov}(N, S) = \mathbb{E}(X)\mathbb{E}(N^2) - \mathbb{E}(N)^2\mathbb{E}(X) = \mathbb{E}(X)[\mathbb{E}(N^2) - \mathbb{E}(N)^2] = \mathbb{E}(X)\text{Var}(N)$$

□

**Definition 6.19 (Compound Poisson random variable).** Let  $N$  be a Poisson random variable. Let  $X_1, X_2, \dots$  be independent and identically distributed random variables that are also independent of  $N$ . Then

$$S := \sum_{i=1}^N X_i$$

is called a **compound Poisson random variable**.

**Proposition 6.1.12 (Variance of a compound Poisson random variable, from Ross *Introduction to Probability Models*).** For a compound Poisson random variable  $S = \sum_{i=1}^N X_i$  having  $\mathbb{E}(N) = \lambda$ ,  $\mathbb{E}(X_i) = \mu$ ,  $\text{Var}(X_i) = \sigma^2$ ,

$$\text{Var}(S) = \lambda\sigma^2 + \lambda\mu^2 = \lambda\mathbb{E}(X^2).$$

### 6.1.5 Odds and Ends

**Proposition 6.1.13. Inclusion-Exclusion Principle:**

(a)

$$\Pr\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq m} \Pr(A_{i1} \cap \dots \cap A_{ik}) \right)$$

(b)

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq m} |A_{i1} \cap \dots \cap A_{ik}| \right)$$

To prove this, we will first prove the Multi-Binomial Theorem.

**Lemma 6.1.14. (Multi-Binomial Theorem)**

$$\prod_{i=1}^d (x_i + y_i)^{n_i} = \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} \dots \sum_{k_d=0}^{n_d} \binom{n_1}{k_1} x_1^{k_1} y_1^{n_1-k_1} \binom{n_2}{k_2} x_2^{k_2} y_2^{n_2-k_2} \dots \binom{n_d}{k_d} x_d^{k_d} y_d^{n_d-k_d}$$

*Proof.*

□

We are now ready to prove the Inclusion-Exclusion Principle.

*Proof.* (Proof of (a).) Begin by noting that

$$\mathbf{1}_{\{\cup_{i=1}^n A_i\}} = 1 - \prod_{i=1}^n (1 - \mathbf{1}_{\{A_i\}}) \quad (6.2)$$

because the expression on the right will equal 1 if at least one term in the product equals 0 (that is, if  $\mathbf{1}_{\{A_i\}} = 1$  for some  $i \in 1, \dots, n$ ) and will equal 0 if every term in the product equals 1 (if  $\mathbf{1}_{\{A_i\}} = 0$  for every  $i \in 1, \dots, n$ ), which is exactly what we want. Expanding the right side of (6.2) using the Multi-Binomial Theorem (Lemma 6.1.14), we have

$$\begin{aligned} &= 1 - \prod_{i=1}^n (1 - \mathbf{1}_{\{A_i\}}) = 1 - (1 - \mathbf{1}_{\{A_1\}})(1 - \mathbf{1}_{\{A_2\}}) \cdots (1 - \mathbf{1}_{\{A_n\}}) \\ &= 1 - \left[ 1 + \sum_{k=1}^n (-1)^k \left( \sum_{1 \leq i_1 < \dots < i_k \leq k} \mathbf{1}_{\{A_{i_1}\}} \cdots \mathbf{1}_{\{A_{i_k}\}} \right) \right] = -1 \cdot \sum_{k=1}^n (-1)^k \left( \sum_{1 \leq i_1 < \dots < i_k \leq k} \mathbf{1}_{\{A_{i_1} \cap \dots \cap A_{i_k}\}} \right) \\ &= \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq k} \mathbf{1}_{\{A_{i_1} \cap \dots \cap A_{i_k}\}} \right) \\ \implies \mathbf{1}_{\{\cup_{i=1}^n A_i\}} &= \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq k} \mathbf{1}_{\{A_{i_1} \cap \dots \cap A_{i_k}\}} \right) \end{aligned} \quad (6.3)$$

Taking expectations of both sides of (6.3) yields

$$\begin{aligned} \Pr(\cup_{i=1}^n A_i) &= \mathbb{E} \left[ \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq k} \mathbf{1}_{\{A_{i_1} \cap \dots \cap A_{i_k}\}} \right) \right] \\ &= \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq k} \mathbb{E}[\mathbf{1}_{\{A_{i_1} \cap \dots \cap A_{i_k}\}}] \right) \\ &= \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq k} \Pr(A_{i_1} \cap \dots \cap A_{i_k}) \right) \end{aligned}$$

□

**Proposition 6.1.15 (a nice trick for upper bounding binomial sums, from Math 547).** For  $d < n$ ,

$$\sum_{i=0}^d \binom{n}{i} \leq \left( \frac{en}{d} \right)^d.$$

*Proof.*

$$\sum_{i=0}^d \binom{n}{i} (d/n)^i \leq \sum_{i=0}^d \binom{n}{i} (d/n)^i \leq \sum_{i=0}^n \binom{n}{i} (d/n)^i = [1 + (d/n)]^n \leq e^d \iff \sum_{i=0}^d \binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

□

**Proposition 6.1.16.** (**Proposition 1.6.1 in Sheldon Ross *A First Course in Probability*.**) There are  $\binom{n-1}{r-1}$  distinct positive integer-valued vectors  $(x_1, x_2, \dots, x_r)$ ,  $x_i > 0 \forall i$  satisfying the equation  $x_1 + x_2 + \dots + x_r = n$ .

*Proof.* (Not rigorous, but a justification.) Imagine we have  $n$  indistinguishable objects to allocate to  $r$  people. We lay out the  $n$  objects and take  $r - 1$  sticks to place in the  $n - 1$  spaces between them. The first person gets all the objects to the left of the leftmost stick, the second person gets the objects between the leftmost and second leftmost stick, and so on, until the last person gets all the objects to the right of the rightmost stick. The constraint that  $x_i$  be positive is equivalent to saying that each person must receive at least one object. Therefore we must place each stick in a different place. There are  $\binom{n-1}{r-1}$  ways to do this.

□

**Proposition 6.1.17.** (**Proposition 1.6.2 in Sheldon Ross *A First Course in Probability*.**) There are  $\binom{n+r-1}{r-1}$  distinct nonnegative integer-valued vectors  $(x_1, x_2, \dots, x_r)$ ,  $x_i \geq 0 \forall i$  satisfying the equation  $x_1 + x_2 + \dots + x_r = n$ .

*Proof.* We would like to solve the problem

$$x_1 + x_2 + \dots + x_r = n, x_i \geq 0 \forall i$$

Note that we can transform this problem in the following way:

$$x_1 + 1 + x_2 + 1 + \dots + x_r + 1 = n + 1 \cdot r, x_i + 1 \geq 1 \forall i$$

Letting  $y_i = x_i + 1$ , we have the equivalent system

$$y_1 + y_2 + \dots + y_r = n + r, y_i \geq 1 \forall i$$

Since  $y_i \geq 1 \iff y_i > 0$ , by Proposition 6.1.16, the number of distinct solutions to this equation is  $\binom{n+r-1}{r-1}$ .

□

**Proposition 6.1.18.** More generally, if we desire solutions to  $x_1 + x_2 + \dots + x_r = n$  such that  $x_i \geq k \in \mathbb{N}$ , then  $\binom{n+r \cdot (1-k)-1}{r-1} = \binom{n+r-1-rk}{r-1}$  solutions are possible.

*Proof.* We can construct a similar argument to that used in the proof of Proposition 6.1.17 by adding  $r \cdot (1 - k)$  to each side of  $x_1 + x_2 + \dots + x_r = n$ :

$$x_1 + 1 - k + x_2 + 1 - k + \dots + x_r + 1 - k = n + r \cdot (1 - k), \quad x_i \geq k \quad \forall i$$

Then substitute  $y_i = x_i + 1 - k$  to yield

$$y_1 + y_2 + \dots + y_r = n + r \cdot (1 - k), \quad y_i + k - 1 \geq k \quad \forall i$$

then apply Proposition 6.1.16 noting that  $y_i + k - 1 \geq k \iff y_i \geq 1 \iff x_i \geq k$  to yield the result.  $\square$

**Proposition 6.1.19.** Even more generally, suppose we desire solutions to  $x_1 + x_2 + \dots + x_r = n$  such that  $x_1 \geq k_1 \in \mathbb{N}, x_2 \geq k_2 \in \mathbb{N}, \dots, x_r \geq k_r \in \mathbb{N}$ . Then

$$\binom{n + \sum_{i=1}^r (1 - k_i) - 1}{r - 1} = \binom{n + r - 1 - \sum_{i=1}^r k_i}{r - 1}$$

solutions are possible.

*Proof.* Very similar to the proof of Proposition 6.1.18. Add  $\sum_{i=1}^r (1 - k_i)$  to each side of  $x_1 + x_2 + \dots + x_r = n$ :

$$x_1 + 1 - k_1 + x_2 + 1 - k_2 + \dots + x_r + 1 - k_r = n + \sum_{i=1}^r (1 - k_i), \quad x_i \geq k_i \quad \forall i$$

Then substitute  $y_i = x_i + 1 - k_i$  to yield

$$y_1 + y_2 + \dots + y_r = n + \sum_{i=1}^r (1 - k_i), \quad y_i + k_i - 1 \geq k_i \quad \forall i$$

Finally, apply Proposition 6.1.16 noting that  $y_i + k_i - 1 \geq k_i \iff y_i \geq 1 \iff x_i \geq k_i$  to yield the result.  $\square$

**Proposition 6.1.20.** Suppose we desire solutions to  $x_1 + x_2 + \dots + x_r = n$  such that  $\tau \leq r$  of the  $\{x_i\}$  exceed some threshold  $k$ . For example, if we have  $x_1 \geq k, x_2 \geq k, \dots, x_\tau \geq k$ , with  $x_{\tau+1}, \dots, x_r$  taking on arbitrary values, then the condition is satisfied. (The particular  $x_i$  that exceed  $k$  does not matter, as long as  $\tau$  of them exceed  $k$ ). Then

$$\binom{r}{\tau} \binom{n + r - 1 - k\tau}{r - 1}$$

solutions are possible.

*Proof.* By Proposition 6.1.19, the number of ways this condition can be met for a particular set of  $\tau$  variables  $x_i$  is  $\binom{n + r - 1 - k\tau}{r - 1}$ . Since there are  $\binom{r}{\tau}$  ways to choose which  $\tau$  variables will exceed  $k$ , the result follows.  $\square$

### 6.1.6 Methods for Calculating Quantities

- Expectation

—

**Definition 6.20.** (**Math 541A definition 1.37.**) Let  $\Omega$  be a sample space, let  $\mathbb{P}$  be a probability law on  $\Omega$ . Let  $X$  be a random variable on  $\Omega$ . Assume that  $X$  takes on nonnegative values; that is,  $X : \Omega \rightarrow [0, \infty)$ . We define the **expected value** of  $X$  by

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > t) dt$$

In analytic notation,  $\mathbb{E}(X) = \int_{\Omega} X(\omega) d\mathbb{P}(\omega)$ . More generally, if  $g : [0 \rightarrow \infty) \rightarrow [0 \rightarrow \infty)$  is a differentiable function such that  $g'$  is continuous and  $g(0) = 0$ , we define

$$\mathbb{E}(g(X)) = \int_0^\infty g'(t) \mathbb{P}(X > t) dt$$

For a general random variable  $X$ , if  $\mathbb{E}(\max\{X, 0\} < \infty)$  and if  $\mathbb{E}(\max\{-X, 0\} < \infty)$ , we then define  $\mathbb{E}(X) = \mathbb{E}(\max\{X, 0\}) - \mathbb{E}(\max\{-X, 0\})$ . Otherwise, we say that  $\mathbb{E}(X)$  is undefined.

**Definition 6.21.** In the above, taking  $g(t) = t^n$  for any positive integer  $n$ , for any  $t \geq 0$ , we have

$$\mathbb{E}(X^n) = \int_0^\infty n t^{n-1} \mathbb{P}(X > t) dt$$

**Remark.** If we assume that the expected value and the integral on  $\mathbb{R}$  can be commuted, then the following derivation of the formula for  $\mathbb{E}(g(X))$  can be given. From the Fundamental Theorem of Calculus, we have

$$g(X) = \int_0^X g'(t) dt = \int_0^\infty g'(t) \mathbf{1}_{\{X>t\}} dt$$

where  $\mathbf{1}_{\{X>t\}}$  is an indicator variable. Therefore

$$\mathbb{E}(g(X)) = \int_0^\infty g'(t) \mathbf{1}_{\{X>t\}} dt = \int_0^\infty g'(t) \mathbb{E}(\mathbf{1}_{\{X>t\}}) = \int_0^\infty g'(t) \mathbb{P}(X > t) dt.$$

—

**Definition 6.22.** (**Discrete random variables.**)  $\mathbb{E}(X) = \sum_x x \Pr(X = x)$

—

**Theorem 6.1.21.** (a)  $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$   
(b) If  $X \geq 0$  then  $\mathbb{E}(X) \geq 0$

—

**Theorem 6.1.22.** Expectation of sums is sum of expectations if sum is finite or if sum is infinite and all variables are positive, but not necessarily otherwise.

**Example 6.3** (**Example 4.18 in *Introduction to Probability Models*.**) Consider a sequence of discrete random variables  $X_1, X_2, \dots$  such that  $X_i = 1$  with probability  $1/2$  and  $X_i = -1$  with probability  $1/2$ . Let  $N$  be a stopping time:

$$N = \min\{n : X_1 + \dots + X_n = 1\} \implies 1 = X_1 + \dots + X_N.$$

Let  $I\{i \leq N\}$  be an indicator variable for  $i \leq N$ . So

$$1 = \sum_{i=1}^N X_i = \sum_{i=1}^{\infty} X_i I\{i \leq N\}.$$

Taking expectations we have

$$1 = \mathbb{E} \sum_{i=1}^{\infty} X_i I\{i \leq N\}.$$

Note that  $N \geq i$  if and only if you have not yet stopped after the first  $i - 1$  games, so it depends on the results of the previous games but not on  $X_i$ . So  $I\{i \leq N\}$  and  $X_i$  are independent. Therefore

$$\mathbb{E}[X_i I\{i \leq N\}] = \mathbb{E}(X_i) \mathbb{E}(I\{i \leq N\}) = 0$$

since  $\mathbb{E}(X_i) = 0$ . So

$$\sum_{i=1}^{\infty} \mathbb{E}[X_i I\{i \leq N\}] = \sum_{i=1}^{\infty} 0 = 0$$

which means that

$$\mathbb{E} \sum_{i=1}^{\infty} X_i I\{i \leq N\} \neq \sum_{i=1}^{\infty} \mathbb{E}[X_i I\{i \leq N\}].$$

See also Example 7.37 in the Stochastic Processes notes.

**Theorem 6.1.23. Law of the Unconscious Statistician:** If  $X$  has mass function  $f$ , and  $g : \mathbb{R} \rightarrow \mathbb{R}$ , then

$$\mathbb{E}(g(X)) = \sum_x g(x) f(x)$$

**Theorem 6.1.24.** Expectation is a linear operator:  $\mathbb{E}(\sum_i X_i) = \sum_i \mathbb{E}(X_i)$

**Theorem 6.1.25. (Layer cake formulation.)** If  $N$  is a discrete random variable taking on non-negative values, then  $\mathbb{E}(N) = \sum_{i=1}^{\infty} \Pr(N \geq i)$ .

*Proof.* Let  $\mathbf{1}_{\{i \leq N\}}$  be an indicator variable. Then

$$N = \sum_{i=1}^{\infty} \mathbf{1}_{\{i \leq N\}}.$$

Take expectations of both sides to yield the result. □

**Remark.** Also note that we can derive this result from Definition 6.20:

$$\mathbb{E}(X) = \int_0^{\infty} \Pr(X > t) dt = \sum_{k=1}^{\infty} \int_{k-1}^k \Pr(X > t) dt = \sum_{k=1}^{\infty} \Pr(X \geq k) = \sum_{k=0}^{\infty} \Pr(X > k).$$

Using Fubini's Theorem (Theorem 5.4.1) to rearrange the sum, we can arrive at

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{\infty} \Pr(X > k) = \sum_{k=0}^{\infty} \sum_{j=k+1}^{\infty} \Pr(X = j) = \sum_{0 \leq k < j < \text{infnty}} \Pr(X = j) \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^j \Pr(X = j) = \sum_{j=0}^{\infty} j \Pr(X = j)\end{aligned}$$

which is the usual definition for a discrete random variable.

- For the conditional expectation of one Gaussian random variable on another when the covariance or correlation between them is known, see Proposition 6.3.29. For the conditional expectation of a set of Gaussian random variables on another set when the covariance matrix is known, see Proposition 6.3.30.

- Variance

–

**Definition 6.23.**  $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$

–

**Proposition 6.1.26. (Useful reformulation:)**  $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

–

**Theorem 6.1.27. (Some useful results):**

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

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

(c)  $\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \pm 2ab\text{Cov}(X, Y)$

(d) **Variance-Covariance Expansion.**  $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$

–

**Definition 6.24.** Conditional variance:

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

–

**Theorem 6.1.28. Law of Total Variance:**  $\text{Var}(X) = \text{Var}(\mathbb{E}(X | Y)) + \mathbb{E}(\text{Var}(X | Y))$

*Proof.*

$$\text{Var}(\mathbb{E}(X | Y)) = \mathbb{E}[(\mathbb{E}(X | Y))^2] - [\mathbb{E}(\mathbb{E}(X | Y))]^2 = \mathbb{E}[(\mathbb{E}(X | Y))^2] - \mathbb{E}(X)^2 \quad (6.4)$$

$$\text{Var}(X | Y) = \mathbb{E}(X^2 | Y) - (\mathbb{E}(X | Y))^2 \implies \mathbb{E}[\text{Var}(X | Y)] = \mathbb{E}(X^2) - \mathbb{E}[(\mathbb{E}(X | Y))^2] \quad (6.5)$$

Adding together (6.4) and (6.5) yields

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

$$= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \text{Var}(X)$$

□

**Corollary 6.1.28.1. (Rao-Blackwell Theorem.)**  $\text{Var}(X) \geq \text{Var}(\mathbb{E}(X | Y))$

*Proof.* Follows immediately from Theorem 6.1.28 by noting that since the variance is nonnegative,  $\mathbb{E}(\text{Var}(X | Y)) \geq 0$ .

□

—

**Proposition 6.1.29.** If  $c \in \mathbb{R}$ , then  $\text{Var}(c) = 0$ .

- For the conditional variance of one Gaussian random variable on another when the covariance or correlation between them is known, see Proposition 6.3.29. For the conditional variance of a set of Gaussian random variables on another set when the covariance matrix is known, see Proposition 6.3.30.

- Covariance

—

**Definition 6.25.**  $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$

—

**Proposition 6.1.30. (Useful reformulation):**  $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

—

**Theorem 6.1.31. (Some useful results):**

- (a)  $\text{Cov}(aX, BY) = ab\text{Cov}(X, Y)$
- (b)  $\text{Cov}(X, X) = \text{Var}(X)$
- (c)  $\text{Cov}(aX + bY) = ac\text{Var}(X) + bd\text{Var}(Y) + (ad + bc)\text{Cov}(X, Y)$
- (d)  $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$
- (e)  $\text{Cov}\left(\sum_{i=1}^n X_i, Y\right) = \sum_{i=1}^n \text{Cov}(X_i, Y)$
- (f)  $\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$

—

**Definition 6.26.** Conditional covariance:

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

—

**Theorem 6.1.32. Law of Total Covariance:**

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

### 6.1.7 Discrete Random Variable Distributions

**Binomial:** Binomial( $n, p$ ) (sum of  $n$  Bernoulli random variables)

- Mass function:  $\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

- Distribution:  $\Pr(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}$
- Expectation:  $\mathbb{E}(X) = np$
- Variance:  $\text{Var}(X) = np(1-p)$

**Multinomial:**

**Definition 6.27** (Multinomial( $n, p_1, p_2, \dots, p_{r-1}$ ) distribution). Suppose that  $n$  independent trials, each of which results in either outcome  $1, 2, \dots, r$  with respective probabilities  $p_1, p_2, \dots, p_r$  (with  $\sum_i p_i = 1$ ), are performed. Let  $N_i$  denote the number of trials resulting in outcome  $i$ . Then the joint distribution of  $N_1, \dots, N_r$  is called the **multinomial distribution**.

- Mass function:

$$\Pr(\mathbf{N} = \mathbf{N}) = \binom{n}{N_1, N_2, \dots, N_r} \prod_{i=1}^r p_i^{N_i} = \frac{n!}{N_1! N_2! \dots N_r!} \prod_{i=1}^r p_i^{N_i}$$

*Proof.* For  $r = 2$ , we have the binomial distribution:  $\Pr(N_1 = x_1) = \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1}$ ;  $\Pr(N_2 = x_2) = \binom{n}{x_2} p_2^{x_2} (1-p_2)^{n-x_2}$ . But in the general case, it is still true that  $N_i$  is a binomial random variable for all  $i$ . So in general we have  $\Pr(N_i = x_i) = \binom{n}{x_i} p_i^{x_i} (1-p_i)^{n_i-x_i}$ ,  $i \in \{1, 2, \dots, r\}$ .

We would like the distribution of the random vector  $\mathbf{N} = (N_1, \dots, N_r)$ . First consider the case where  $n = 1$ ; that is, the Multinouli distribution. Note that in this case, we have  $\Pr(\mathbf{N} = \mathbf{N}) = \prod_{i=1}^r p_i^{N_i}$ , where  $N_i$  are the entries of the observed vector  $\mathbf{N}$  ( $N_i \in \{0, 1\}$ ).

Now consider an arbitrary  $n \in \mathbb{N}$ . Then  $\mathbf{N}$  is the sum of  $n$  Multinouli random variables. Just as before, the probability of a particular  $\mathbf{N}$  obtained in a particular ordering is  $\Pr(\mathbf{N} = \mathbf{N}) = \prod_{i=1}^r p_i^{N_i}$ . However, we must also consider the number of possible orderings in which these successes could have occurred. This number of orderings is exactly equal to the multinomial coefficient,

$$\binom{n}{N_1, N_2, \dots, N_r} = \frac{n!}{N_1! N_2! \dots N_r!}$$

Therefore the joint probability mass function for  $\mathbf{N}$  is

$$\Pr(\mathbf{N} = \mathbf{N}) = \binom{n}{N_1, N_2, \dots, N_r} \prod_{i=1}^r p_i^{N_i} = \frac{n!}{N_1! N_2! \dots N_r!} \prod_{i=1}^r p_i^{N_i}$$

□

- Distribution:  $\Pr(X \leq k) =$
- Expectation:  $\mathbb{E}(X) =$
- Variance:  $\text{Var}(X) =$

**Proposition 6.1.33.**  $\text{Cov}(N_i, N_j) = -np_i p_j$ .

*Proof.* For a Multinoulli random variable, we have

$$\text{Cov}(N_i, N_j) = \mathbb{E}(N_i N_j) - \mathbb{E}(N_i)\mathbb{E}(N_j) = 0 - p_i p_j = \dots = -p_i p_j$$

(where  $\mathbb{E}(N_i)\mathbb{E}(N_j) = 0$  because at least one of them must equal 0). Since a multinomial random variable is the sum of  $n$  independent Multinoulli random variables, in the general case we have

$$\text{Cov}(N_i, N_j) = -np_i p_j$$

□

**Poisson:** Poisson( $\lambda$ ): an approximation of the binomial distribution for  $n$  very large,  $p$  very small,  $np \rightarrow \lambda \in (0, \infty)$ .

- Mass function:

$$\Pr(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

- Distribution:  $\Pr(X \leq k) = \sum_{i=0}^k \frac{e^{-\lambda} \lambda^i}{i!}$
- Expectation:  $\mathbb{E}(X) = \lambda$  (derive from basic definitions)
- Variance:  $\text{Var}(X) = \lambda$
- Moment-generating function:  $M_X(t) = e^{\lambda(e^t - 1)}$

**Proposition 6.1.34.** Let  $X \sim \text{Binomial}(n, p)$ . Then

$$\lim_{n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda} X \sim \text{Poisson}(np).$$

*Proof.*

$$\begin{aligned} & \lim_{n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda} \frac{n!}{(n-k)! k!} p^k (1-p)^{n-k} \\ &= \frac{1}{k!} \cdot \lim_{n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda} (1-p)^{n-k} p^k \prod_{i=0}^{k-1} (n-i) = \frac{1}{k!} \cdot \lim_{n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda} \left(1 - \frac{np}{n}\right)^{n-k} p^k \prod_{i=0}^{k-1} (n-i) \end{aligned}$$

Using  $\lim_{n \rightarrow \infty} (1 - \lambda/n)^n = \exp(\lambda)$ , and letting  $\lambda = np$ , we have

$$= \frac{\exp(-np)(np)^k}{k!} = \boxed{\frac{\exp(-\lambda)\lambda^k}{k!}} \sim \text{Poisson}(np)$$

□

**Proposition 6.1.35.** Let  $X \sim \text{Poisson}(\lambda)$ . If  $\lambda$  is sufficiently large (say  $\lambda > 20$ , then we can use the approximation

$$X \sim \mathcal{N}(\lambda, \lambda)$$

*Proof.* (Informal justification.) By Proposition 6.1.34, a Poisson distribution can be thought of as a close approximation of a binomial distribution. Since a binomial distribution can be approximated by a normal distribution for  $n$  large and  $np$  not too large, the same is true of a Poisson distribution.  $\square$

**Proposition 6.1.36.** Suppose  $X \sim \text{Poisson}(\mu)$ ,  $Y \sim \text{Poisson}(\nu)$ , with  $X \perp\!\!\!\perp Y$ . Then  $\{X + Y\} \sim \text{Poisson}(\mu + \nu)$ .

*Proof.* (exercise; use characteristic functions.)

$\square$

**Proposition 6.1.37.** Suppose  $X \sim \text{Poisson}(\mu)$ ,  $Y \sim \text{Poisson}(\nu)$ , with  $X \perp\!\!\!\perp Y$ . Then  $\{X \mid X + Y = t\} \sim \text{Bin}(t, \mu/(\mu + \nu))$

*Proof.*

$$\mathbb{P}_{\mu, \nu}(X = x \mid X + Y = t) = \frac{\mathbb{P}_{\mu, \nu}(X = x, Y = t - x)}{\mathbb{P}_{\mu, \nu}(X + Y = t)} = \frac{\mu^x}{x!} e^{-\mu} \cdot \frac{\nu^{t-x}}{(t-x)!} e^{-\nu} \cdot \frac{t! e^{\mu+\nu}}{(\mu+\nu)^t}$$

where we used Proposition 6.1.36 to get  $\{X + Y\} \sim \text{Poisson}(\mu + \nu)$ .

$$= \binom{t}{x} \left( \frac{\mu}{\mu + \nu} \right)^x \left( \frac{\nu}{\mu + \nu} \right)^{t-x}$$

so  $\{X \mid T = t\} \sim \text{Bin}(t, \mu/(\mu + \nu))$ .

$\square$

**Geometric:**  $G_1(p)$ : the number of Bernoulli trials before the first success.

- Mass function:  $\Pr(X = k) = p(1-p)^{k-1}$
- Distribution:  $\Pr(X \leq k) = \sum_{i=1}^k p(1-p)^{i-1}$
- Expectation:  $\mathbb{E}(X) = 1/p$
- Variance:  $\text{Var}(X) = (1-p)/p^2$

**Negative binomial:**  $\text{NB}(r, p)$ : The number of Bernoulli trials required for  $r$  successes. (Can be derived as the sum of  $r$  identically distributed geometric random variables.)

- Mass function:  $\Pr(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$

- Distribution:  $\Pr(X \leq k) = \sum_{i=r}^k \binom{i-1}{r-1} p^r (1-p)^{i-r}$
- Expectation:  $\mathbb{E}(X) = \frac{r}{p}$
- Variance:  $\text{Var}(X) = \frac{r(1-p)}{p^2}$
- Moment-generating function:

$$M_X(t) = \frac{(pe^t)^r}{[1 - (1-p)e^t]^r}$$

**Hypergeometric:** Hypergeometric( $N, M, K$ ): When drawing a sample of size  $K$  from a group of  $N$  items,  $M$  of which are special,  $X$  is the number of special items retrieved.

- Mass function:

$$\Pr(X = k) = \frac{\binom{M}{k} \binom{N-M}{K-k}}{\binom{N}{K}}$$

- Distribution:

$$\Pr(X \leq k) = \sum_{i=0}^k \frac{\binom{M}{i} \binom{N-M}{K-i}}{\binom{N}{K}}$$

- Expectation:  $\mathbb{E}(X) = \frac{MK}{N}$

*Proof.* Let  $Y_j$  be an indicator variable for special item  $j$  being selected. Note that  $X = \sum_{j=1}^M Y_j$  and that

$$\mathbb{E}(Y_j) = \frac{\binom{1}{1} \binom{N-1}{K-1}}{\binom{N}{K}} = \frac{1 \cdot \frac{(N-1)!}{(N-K)!(K-1)!}}{\frac{N!}{(N-K)!K!}} = \frac{K}{N},$$

so we have

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{j=1}^M Y_j\right) = \sum_{j=1}^M \mathbb{E}(Y_j) = \frac{MK}{N}.$$

Alternative proof:

Let  $Z_i$  be an indicator variable for the  $i$ th selected item to be special. Then  $X = \sum_{i=1}^K Z_i$  and  $\mathbb{E}(Z_i) = \frac{M}{N}$ , so we have

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^K Z_i\right) = \sum_{i=1}^K \mathbb{E}(Z_i) = \frac{MK}{N}.$$

Variance:  $\text{Var}(X) =$  (find by indicator method. Proof from notes, I think this may have errors:)

$\Sigma_{\text{random}}$

$\underline{\text{Ex}} \quad X \sim \mathcal{H}(M, m, n)$

$X = \sum_{k=1}^n X_k$        $X_k = \begin{cases} 1, & \text{if element is "special"} \\ 0, & \text{if not.} \end{cases}$

$E(X_k) = \frac{m}{M}$        $\Rightarrow E(X) = n \cdot \frac{m}{M}$

$V_{\text{var}}(X) = \sum_{k=1}^n V_{\text{var}}(X_k) + \sum_{k \neq m} \sum_{m} \text{cov}(X_k, X_m)$

$= n p(1-p) - 2 \binom{n}{2} \left[ \frac{m(m-1)}{M(M-1)} - \left(\frac{m}{M}\right)^2 \right]$

□

**Proposition 6.1.38.** Let  $X \sim \text{Hypergeometric}(N, M, K)$ . Then

$$\lim_{M, N \rightarrow \infty, M/N \rightarrow p} \Pr(X = k) \sim \text{Binomial}(K, p)$$

*Proof.*

$$\begin{aligned}
 \lim_{M, N \rightarrow \infty, M/N \rightarrow p} p_k(M, N, K) &= \lim_{M, N \rightarrow \infty, M/N \rightarrow p} \frac{\binom{M}{k} \binom{N-M}{K-k}}{\binom{N}{K}} \\
 &= \lim_{M, N \rightarrow \infty, M/N \rightarrow p} \frac{M!(N-M)!/[k!(M-k)!(K-k)!(N-M-K+k)!]}{N!/[K!(N-K)!]} \\
 &= \lim_{M, N \rightarrow \infty, M/N \rightarrow p} \frac{K!}{(K-k)!k!} \cdot \frac{M!(N-M)!(N-K)!}{N!(M-k)!(N-M-K+k)!} \\
 &= \lim_{M, N \rightarrow \infty, M/N \rightarrow p} \binom{K}{k} \cdot \frac{M!/(M-k)!}{N!/(N-k)!} \cdot \frac{(N-M)!(N-K)!}{(N-k)!(N-M-(K-k))!} \\
 &= \binom{K}{k} \lim_{M, N \rightarrow \infty, M/N \rightarrow p} \frac{M!/(M-k)!}{N!/(N-k)!} \cdot \frac{(N-M)!(N-M-(K-k))!}{(N-K+(K-k))!/(N-K)!} \\
 &= \binom{K}{k} \lim_{M, N \rightarrow \infty, M/N \rightarrow p} \prod_{i=0}^{k-1} \frac{M-i}{N-i} \cdot \prod_{j=0}^{K-k-1} \frac{N-M-j}{N-K+1+j} \\
 &= \binom{K}{k} \left(\frac{M}{N}\right)^k \left(\frac{N-M}{N}\right)^{K-k} \\
 &= \binom{K}{k} \left(\frac{M}{N}\right)^k \left(1 - \frac{M}{N}\right)^{K-k} = \binom{K}{k} p^k (1-p)^{K-k}
 \end{aligned}$$

□

### 6.1.8 Indicator Method

**Proposition 6.1.39.** If  $\mathbf{1}_{A_k}$  is an indicator then

(a)

$$\text{Cov}(\mathbf{1}_{A_k}, \mathbf{1}_{A_m}) = \mathbb{E}(\mathbf{1}_{A_k} \mathbf{1}_{A_m}) - \mathbb{E}(\mathbf{1}_{A_k})\mathbb{E}(\mathbf{1}_{A_m}) = \Pr(A_k \cap A_m) - \Pr(A_k)\Pr(A_m)$$

(b)

$$\text{Var}(\mathbf{1}_{A_k}) = \mathbb{E}(\mathbf{1}_{A_k}^2) = \mathbb{E}(\mathbf{1}_{A_k})^2 = \Pr(A_k) - (\Pr(A_k))^2$$

**Theorem 6.1.40.**  $X$  is independent of  $Y$  if and only if  $X$  is independent of  $\mathbf{1}_A$ ,  $A \in Y$ .

Example problems: 505A Homework 3 problem 9(a)

Worked examples in p. 56 - 59 of Grimmett and Stirzaker 3rd edition.

### 6.1.9 Linear transformations of random variables

#### 6.1.10 Poisson Paradigm (Poisson approximation for indicator method)

**Theorem 6.1.41.** (Theorem 4.12.9, p. 129 of Grimmett and Stirzaker.) Let  $A_i$  be an event. If  $X = \sum_{i=1}^m \mathbf{1}_{A_i}$  where  $\mathbf{1}_{A_i}$  is an indicator variable for  $A_i$ , and the  $A_i$  are only weakly dependent on each other, then

$$\text{As } m \rightarrow \infty, \quad X \sim \text{Poisson}(\mathbb{E}(X))$$

More specifically, let  $B_i$  be  $n$  independent Bernoulli random variables with probabilities  $p_i$ . If  $Y = \sum_{i=1}^n B_i$  then

$$\text{As } n \rightarrow \infty, \quad Y \sim \text{Poisson} \left( \mathbb{E} \left( \sum_i B_i \right) \right) = \text{Poisson} \left( \sum_i \mathbb{E} B_i \right) = \text{Poisson} \left( \sum_i p_i \right)$$

*Proof.* Full proof available in Grimmett and Stirzaker, section 4.12, page 129. A justification of the first claim is as follows: if the  $A_i$  are independent and  $\Pr(A_i) = p \forall i$ , then  $X \sim \text{Binomial}(m, p)$ . Then by Proposition 6.1.34, the result follows. It turns out that this result holds up if the probabilities are not necessarily identical (but all small) and the variables are not necessarily independent (but only weakly dependent).  $\square$

#### Solution. Alternative Solution to Exercise 1 (Matching Problem):

Let  $X$  be the number of matches. Let  $P_n = \Pr(X = 0)$  given that there are  $n$  people. Let  $Y$  be an indicator variable for the first person who receives their sandwich receiving the correct one. Note that

$$P_n = \Pr(X = 0) = \Pr(X = 0 \mid Y = 1)(1/n) + \Pr(X = 0 \mid Y = 0)(n - 1)/n.$$

Note that if the first person didn't match ( $Y = 0$ ), we have  $n - 2$  people with their hats left, but one of the remaining  $n - 1$  people can't get their sandwich because it was taken by the first person. Therefore

$$\Pr(X = 0 \mid Y = 0) = \Pr(\{\text{the extra person selects the first person's hat, and the rest of the people pick the wrong hat}\})$$

$$+ \Pr(\{\text{the extra person does not select the first person's hat, and the rest of the people don't have any matches}\})$$

$$= \frac{1}{n-1} \cdot P_{n-2} + P_{n-1}$$

This yields

$$P_n = \frac{1}{n} P_{n-2} + \frac{n-1}{n} P_{n-1} \iff P_n - P_{n-1} = -\frac{1}{n} (P_{n-1} - P_{n-2})$$

which is a recursive formula. Now we seek to find a closed form solution. We have

$$P_3 - P_2 = -\frac{1}{3} (P_2 - P_1) = -\frac{1}{3!} \iff P_3 = P_2 - \frac{1}{3!} = \frac{1}{2!} - \frac{1}{3!}$$

$$P_4 - P_3 = -\frac{1}{4} (P_3 - P_2) = \frac{1}{4!} \iff P_4 = P_3 + \frac{1}{4!} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!}$$

⋮

$$P_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots + \frac{(-1)^n}{n!} = \sum_{k=0}^n \frac{(-1)^k}{k!} \rightarrow e^{-1} \text{ as } n \rightarrow \infty$$

so we see that for large  $n$  we approximately have  $P_n \sim \text{Poisson}(1)$ .

Now we consider  $\Pr(X = k)$ . Consider a set of  $k$  people who all have matches and no one else matches. The probability that everyone in a set of  $k$  people match is  $(n - k)!/n!$ . The probability that none of the other  $n - k$  people match is  $P_{n-k}$  per above. Since there are  $\binom{n}{k}$  ways to chose groups of  $k$  people, we have

$$\Pr(X = k) = \binom{n}{k} \frac{(n - k)!}{n!} \cdot P_{n-k} = \frac{P_{n-k}}{k!} \xrightarrow{n \rightarrow \infty} \frac{e^{-1} 1^k}{k!}$$

so again we see that for large  $n$  we approximately have  $P_n \sim \text{Poisson}(1)$ .

### 6.1.11 Asymptotic Distributions

**Proposition 6.1.42.**

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$$

**Theorem 6.1.43. Stirling's Formula:**

$$n! \sim n^n e^{-n} \sqrt{2\pi n}$$

That is,

$$\lim_{n \rightarrow \infty} \frac{n^n e^{-n} \sqrt{2\pi n}}{n!} = 1.$$

## 6.2 Worked problems

### 6.2.1 Example Problems That Will Likely Appear on Midterm (and Final)

- (1) (a) **Fall 2011 Problem 1 (same as HW1 problem 5; similar to HW3 problem 2(5).)** Let  $A$  and  $B$  be events such that  $0 < \Pr(A) < 1$ . Show that if  $\Pr(B | A) = \Pr(B | A^c)$ , then  $A$  and  $B$  are independent.
- (b) Let  $X$  and  $Y$  be two discrete random variables, each taking only two possible values. Show that if  $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$  then  $X$  and  $Y$  are independent.

**Solution.**

- (a)  $A$  and  $B$  are independent if and only if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

We know that

$$\Pr(B) = \Pr(B|A) \cdot \Pr(A) + \Pr(B|A^c) \cdot \Pr(A^c)$$

$$= \Pr(B|A) \cdot \Pr(A) + \Pr(B|A) \cdot (1 - \Pr(A)) = \Pr(B|A) \cdot \Pr(A) + \Pr(B|A) - \Pr(B|A) \cdot \Pr(A)$$

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

Also, we know that since  $\Pr(A) \neq 0$ ,

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

Per above  $\Pr(B|A) = \Pr(B)$ , so we have

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

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

which is what we were trying to prove. So the answer is true.

(b) Without loss of generality, let  $X$  and  $Y$  have mass functions

$$X = \begin{cases} x_1 & \text{with probability } \Pr(A) \\ x_2 & \text{with probability } \Pr(A^c) \end{cases}$$

$$Y = \begin{cases} y_1 & \text{with probability } \Pr(B) \\ y_2 & \text{with probability } \Pr(B^c) \end{cases}$$

Then  $X \perp\!\!\!\perp Y \iff \Pr(A \cap B) = \Pr(A)\Pr(B)$ . Let  $\alpha = X - x_2$ ,  $\beta = Y - y_2$ ; that is,

$$\alpha = \begin{cases} x_1 - x_2 & \text{with probability } \Pr(A) \\ 0 & \text{with probability } \Pr(A^c) \end{cases}$$

$$\beta = \begin{cases} y_1 - y_2 & \text{with probability } \Pr(B) \\ 0 & \text{with probability } \Pr(B^c) \end{cases}$$

Then we have

- $\mathbb{E}(\alpha) = (x_1 - x_2)\Pr(A)$
- $\mathbb{E}(\beta) = (y_1 - y_2)\Pr(B)$
- $\mathbb{E}(\alpha\beta) = (x_1 - x_2)(y_1 - y_2)\Pr(A \cap B)$

which we can use to obtain

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}[(\alpha + x_2)(\beta + y_2)] = \mathbb{E}(\alpha\beta) + y_2\mathbb{E}(\alpha) + x_2\mathbb{E}(\beta) + x_2y_2 \\ &= (x_1 - x_2)(y_1 - y_2)\Pr(A \cap B) + y_2(x_1 - x_2)\Pr(A) + x_2(y_1 - y_2)\Pr(B) + x_2y_2 \end{aligned} \quad (6.6)$$

$$\begin{aligned} \mathbb{E}(X)\mathbb{E}(Y) &= \mathbb{E}(\alpha + x_2)\mathbb{E}(\beta + y_2) = [x_2 + (x_1 - x_2)\Pr(A)][y_2 + (y_1 - y_2)\Pr(B)] \\ &= x_2y_2 + x_2(y_1 - y_2)\Pr(B) + y_2(x_1 - x_2)\Pr(A) + (x_1 - x_2)(y_1 - y_2)\Pr(A)\Pr(B) \end{aligned} \quad (6.7)$$

Using  $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ , we set (6.6) and (6.7) equal to each other. Canceling terms appearing in both yields

$$(x_1 - x_2)(y_1 - y_2)\Pr(A \cap B) = (x_1 - x_2)(y_1 - y_2)\Pr(A)\Pr(B) \iff \Pr(A \cap B) = \Pr(A)\Pr(B)$$

which proves the independence of  $X$  and  $Y$  if  $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ .

- (2) **Fall 2013 Qual Problem 1.** Consider a sequence of independent tosses of a pair of fair dice. Compute the probability that the sum 4 will occur before the sum 5.

**Solution.** Let  $Y_k$  be the outcome of the  $k$ th toss. Let  $X_{k1}$  be the number of the first die on the  $k$ th toss and  $X_{k2}$  be the outcome of the second die. Note that

$$\Pr(Y_k = 4) = \Pr(X_{k1} = 1 \cap X_{k2} = 3) + \Pr(X_{k1} = 2 \cap X_{k2} = 2) + \Pr(X_{k1} = 3 \cap X_{k2} = 1) = 3 \cdot \frac{1}{36} = \frac{1}{12}$$

$$\Pr(Y_k = 5) = \Pr(X_{k1} = 1 \cap X_{k2} = 4) + \Pr(X_{k1} = 2 \cap X_{k2} = 3) + \Pr(X_{k1} = 3 \cap X_{k2} = 2)$$

$$+ \Pr(X_{k1} = 4 \cap X_{k2} = 1) = 4 \cdot \frac{1}{36} = \frac{1}{9}$$

Let  $A_k$  be the event that  $Y_k = 4$  and  $Y_j \neq 4$  or  $5$ ,  $j = 1, \dots, k-1$ . Note that all  $A_k$  are mutually exclusive and

$$\Pr(A_k) = \frac{1}{12} \cdot \left(1 - \frac{3+4}{36}\right)^{k-1} = \frac{1}{12} \cdot \left(\frac{29}{36}\right)^{k-1}.$$

Then

$$\begin{aligned} \Pr(\{\text{roll a 4 before a 5}\}) &= \Pr\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \Pr(A_k) = \frac{1}{12} \sum_{k=1}^{\infty} \left(\frac{29}{36}\right)^{k-1} = \frac{1/12}{1 - 29/36} \\ &= \frac{3}{36-29} = \boxed{\frac{3}{7}} \end{aligned}$$

- (3) **Spring 2013 Problem 1, in notes from 09/21.** Let  $X$  and  $Y$  be random variables such that  $\mathbb{E}(X | Y) = Y, \mathbb{E}(Y | X) = X, \mathbb{E}(X^2) < \infty, \mathbb{E}(Y^2) < \infty$ . Show that  $\mathbb{E}(X - Y)^2 = 0$  (or equivalently, show  $\Pr(X = Y) = 1$ ).

**Solution.**

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

$$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY | Y)) = \mathbb{E}(Y\mathbb{E}(X | Y)) = \mathbb{E}(Y \cdot Y) = \mathbb{E}(Y^2)$$

Also,

$$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY | X)) = \mathbb{E}(X\mathbb{E}(Y | X)) = \mathbb{E}(X \cdot X) = \mathbb{E}(X^2)$$

Therefore

$$\mathbb{E}(X - Y)^2 = 0$$

- (4) **Fall 2016 Problem 4.** Consider a group of  $n \geq 4$  people, among whom are Alice, Bob, Charles, and Diana, standing in a row. Assume that all possible orderings of the  $n$  people are equally likely.
- (a) Compute the probability that Charles stands somewhere between Alice and Bob. (Note: this does not mean that the three are necessarily adjacent; there can be other people between Alice and Bob.)
  - (b) Compute the probability that Diana stands somewhere between Alice and Bob given that Charles stands somewhere between Alice and Bob.
  - (c) Let  $X$  be the number of people who stand between Alice and Bob. Compute the expected value and the variance of  $X$ . (Note: Alice and Bob themselves are not counted in this number.)

**Solution.**

- (a) For this part,  $n$  does not make a difference; all we need to know is the ordering of  $A$ ,  $B$ , and  $C$ . This is because conditional on a specific ordering of  $A$ ,  $B$ , and  $C$ , all arrangements of everyone else are equally likely; and conversely, the a particular ordering of  $A$ ,  $B$ , and  $C$  is independent of which three particular slots are made available to them. Examining the permutations of  $A$ ,  $B$ , and  $C$ , two of them have  $C$  in the middle, so the answer is  $2/6 = \boxed{1/3}$ .
- (b) Similarly, the answer is independent of  $n$ , so we work with  $n = 4$ . All possible orderings with Charles between Alice and Bob are as follows:

$$ACDB, ADCB, BCDA, BDCA, ACBD, BCAD, DACB, DBCA$$

The first four of these have Diana between Alice and Bob, so the answer is  $4/8 = \boxed{1/2}$ .

- (c) Let  $I_k$  be an indicator variable for the event that person  $k$  is between  $A$  and  $B$ . By the result from part (a),  $\mathbb{E}(I_k) = 1/3$ . Then we have

$$\text{Var}(I_k) = \mathbb{E}(I_k^2) - \mathbb{E}(I_k)^2 = 1^2 \cdot \Pr(I_k = 1) - \frac{1}{9} = \frac{1}{3} - \frac{1}{9} = \frac{2}{9}$$

Noting that the four arrangements above with Charles and Diana in between Alice and Bob are the only ones where this will be the case of the  $4! = 24$  possible orderings, we have

$$\mathbb{E}(I_k I_j) = \frac{4}{24} = \frac{1}{6}$$

so

$$\text{Cov}(I_j, I_k) = \mathbb{E}(I_k I_j) - \mathbb{E}(I_k)\mathbb{E}(I_j) = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}$$

Therefore

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^{n-2} I_k\right) = \sum_{i=1}^{n-2} \frac{1}{3} = \frac{n-2}{3} \\ \text{Var}(X) &= \text{Var}\left(\sum_{i=1}^{n-2} I_k\right) = \sum_{i=1}^{n-2} \text{Var}(I_k) + 2 \sum_{1 \leq j < k \leq n-2} \text{Cov}(I_j, I_k) = \frac{2(n-2)}{9} + [(n-2)^2 - (n-2)] \cdot \frac{1}{18} \\ &= \frac{4(n-2)}{18} + \frac{(n-2)(n-3)}{18} = \boxed{\frac{(n-2)(n+1)}{18}} \end{aligned}$$

- (5) **Spring 2015 Problem 2.** A deck of 52 cards is shuffled thoroughly. Someone goes through all 52 cards, scoring 1 point each time 2 cards of the same value are consecutive (that is, when two consecutive cards have the same rank but different suits). For example, the sequence 9H, 8H, 7D, 6C, 7S, 7H, 7C scores 2 points, one for the 7H following the 7S, one for the 7C following the 7H. Let  $X$  be the total score.

- (a) Compute  $\mathbb{E}(X)$ .
- (b) Compute  $\Pr(X = 39)$ . (Note that there are 13 different ranks and you cannot score more than 3 per rank.)
- (c) In the line below, circle the number that you think is the closest to the value  $\Pr(X = 0)$  and briefly explain your choice:

$$\frac{1}{1000}, \quad \frac{1}{500}, \quad \frac{1}{100}, \quad \frac{1}{50}, \quad \frac{1}{20}, \quad \frac{1}{10}, \quad \frac{1}{5}, \quad \frac{1}{2}$$

**Solution.**

- (a) Start by assuming the permutation is cyclic (that is, after the last card you go back to the beginning). Let  $Y$  be the number of matches in this situation. Let  $A_i$  be the event that the  $i$ th card is followed by a match. Then  $\Pr(A_i) = 3/51 = 1/17$ , so

$$Y = \sum_{k=1}^{52} \mathbf{1}_{\{A_k\}} \implies \mathbb{E}(Y) = \sum_{i=1}^{52} \mathbb{E}(\mathbf{1}_{\{A_i\}}) = \sum_{i=1}^{52} \Pr(A_i) = 52 \cdot 1/17 = 52/17$$

Note that  $\mathbb{E}(Y) = \mathbb{E}(X) + \mathbb{E}(\mathbf{1}_{\{A_i\}})$  because if the permutation is cyclic then you have one extra opportunity to match at the end.

$$\implies \mathbb{E}(X) = \frac{52}{17} - \frac{1}{17} = \frac{51}{17} = \boxed{3}$$

- (b)  $\Pr(X = 39) = \Pr(\{\text{all possible matches occur}\})$ , so in this event all cards of the same rank are clustered together. There are 13 clusters of 4 cards, so there are  $13!$  ways to order the clusters and  $4!$  ways to order the cards within each cluster. Therefore

$$\boxed{\Pr(X = 39) = \frac{13!(4!)^{13}}{52!}}$$

- (c) Because  $A_i$  are only weakly dependent and  $\Pr(A_i)$  is small for all  $A_i$ , we can use the Poisson approximation (see Section 6.1.10); that is,  $X \sim \text{Poisson}(\mathbb{E}(X)) = \text{Poisson}(3)$ . Therefore

$$\Pr(X = 0) \approx \frac{e^{-3} \cdot 3^0}{0!} = \frac{1}{e^3} \approx \frac{1}{2.8^3} \approx \boxed{\frac{1}{20}}$$

### 6.2.2 Problems we did in class that professor mentioned

**(Fall 2014 Problem 1) (Variation of Midterm problem 1 above)** Let  $A$  and  $B$  be two events with  $0 < \Pr(A) < 1$ ,  $0 < \Pr(B) < 1$ . Define the random variables  $\xi = \xi(\omega)$  and  $\eta = \eta(\omega)$  by

$$\xi(\omega) = \begin{cases} 5 & \text{if } \omega \in A \\ -7 & \text{if } \omega \notin A \end{cases}, \quad \eta(\omega) = \begin{cases} 2 & \text{if } \omega \in B \\ 3 & \text{if } \omega \notin B \end{cases}$$

True or false: the events  $A$  and  $B$  are independent if and only if the random variables  $\xi$  and  $\eta$  are uncorrelated?

**Solution.** ( $\implies$ ) Suppose  $A$  and  $B$  are independent. Then  $\xi$  and  $\eta$  are uncorrelated if and only if  $\mathbb{E}(\xi\eta) = \mathbb{E}(\xi)\mathbb{E}(\eta)$ . We can write  $\xi = 5 \cdot \mathbf{1}_A - 7 \cdot \mathbf{1}_{A^c}$  and  $\eta = 2 \cdot \mathbf{1}_B + 3 \cdot \mathbf{1}_{B^c}$ . So we have

$$\xi\eta = (5 \cdot \mathbf{1}_A - 7 \cdot \mathbf{1}_{A^c})(2 \cdot \mathbf{1}_B + 3 \cdot \mathbf{1}_{B^c}) = 10 \cdot \mathbf{1}_{A \cap B} + 15 \cdot \mathbf{1}_{A \cap B^c} - 14 \cdot \mathbf{1}_{A^c \cap B} - 21 \cdot \mathbf{1}_{A^c \cap B^c}$$

$$\implies \mathbb{E}(\xi\eta) = 10 \Pr(A \cap B) + 15 \Pr(A \cap B^c) - 14 \Pr(A^c \cap B) - 21 \Pr(A^c \cap B^c)$$

Then

$$\mathbb{E}(\xi)\mathbb{E}(\eta) = (5\Pr(A) - 7\Pr(A^c))(2\Pr(B) + 3\Pr(B^c))$$

$$= 10\Pr(A \cap B) + 15\Pr(A \cap B^c) - 14\Pr(A^c \cap B) - 21\Pr(A^c \cap B^c) = \mathbb{E}(\xi\eta)$$

where the second-to-last step follows from the independence of  $A$  and  $B$ . Therefore  $\eta$  and  $\xi$  are uncorrelated.

( $\Leftarrow$ ) Now suppose  $\eta$  and  $\xi$  are uncorrelated. Then  $\xi$  and  $\eta$  are independent if and only if  $\Pr(\xi \cap \eta) = \Pr(\xi)\Pr(\eta)$ . Define

$$\alpha(\omega) = \xi(\omega) + 7 = \begin{cases} 12 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}, \quad \beta(\omega) = \eta(\omega) - 3 = \begin{cases} -1 & \text{if } \omega \in B \\ 0 & \text{if } \omega \notin B \end{cases}$$

Then we have

$$(\alpha\beta)(\omega) = \begin{cases} -12 & \text{if } \omega \in A \cap B \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathbb{E}(\xi\eta) = \mathbb{E}[(\alpha - 7)(\beta + 3)] = \mathbb{E}(\alpha\beta) + 3\mathbb{E}(\alpha) - 7\mathbb{E}(\beta) - 21$$

$$\mathbb{E}(\xi)\mathbb{E}(\eta) = (\mathbb{E}(\alpha) - 7)(\mathbb{E}(\beta) + 3) = \mathbb{E}(\alpha)\mathbb{E}(\beta) - 7\mathbb{E}(\beta) + 3\mathbb{E}(\alpha) - 21$$

Since by assumption  $\mathbb{E}(\xi\eta) = \mathbb{E}(\xi)\mathbb{E}(\eta)$ , this yields  $\mathbb{E}(\alpha\beta) = \mathbb{E}(\alpha)\mathbb{E}(\beta)$ . But

$$\mathbb{E}(\alpha\beta) = -12\Pr(A \cap B), \quad \mathbb{E}(\alpha)\mathbb{E}(\beta) = 12\Pr(A)(-1)\Pr(B) = -12\Pr(A)\Pr(B)$$

Therefore  $\Pr(\xi \cap \eta) = \Pr(\xi)\Pr(\eta)$  and  $\xi$  and  $\eta$  are independent.

**Exercise 1. Example (Letter/envelope matching problem; sometimes referred to as Montmort's matching problem).** An assistant brings  $n$  sandwiches for  $n$  employees at a company. Each employee ordered a unique sandwich, but unfortunately the assistant forgot to ask that the sandwiches be labeled, so they are all indistinguishable, wrapped in the same paper. The assistant plans to distribute one sandwich to each employee and hope for the best. Let  $X$  be the number of sandwiches that are delivered to the correct person.

(a) What is the probability of at least one match; that is,  $\Pr(X \geq 1)$ ?

(b) What is the probability of  $r$  correct matches?

(c) What  $\mathbb{E}(X)$ ?

(d) What is  $\text{Var}(X)$ ?

**Solution.**

(a) Let  $A_k$  be an indicator variable for the event that sandwich  $k$  is matched to the correct employee.

Then

$$\Pr(X \geq 1) = \Pr\left(\bigcup_{k=1}^n A_k\right)$$

Consider that if there are  $k$  correct matches, there are  $\binom{n}{k}$  sets of  $k$  sandwiches that could be correctly distributed. Also, the probability of a particular set of  $k$  sandwiches being correctly distributed is  $(n - k)!/n!$ . So we have

$$\Pr(X = k) = \binom{n}{k} \frac{(n - k)!}{n!}$$

Therefore by the Inclusion-Exclusion Principle (Proposition 6.1.13),

$$\begin{aligned} \Pr\left(\bigcup_{k=1}^n A_k\right) &= \sum_{k=1}^n (-1)^{k-1} \Pr(X = k) = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} \frac{(n - k)!}{n!} = \sum_{k=1}^n (-1)^{k-1} \frac{n!}{(n - k)!k!} \frac{(n - k)!}{n!} \\ &= \sum_{k=1}^n \frac{(-1)^{k-1}}{k!} = \frac{(-1)^0}{0!} - \sum_{k=0}^n \frac{(-1)^k}{k!} = \boxed{1 - \sum_{k=0}^n \frac{(-1)^k}{k!}} \end{aligned}$$

As  $n \rightarrow \infty$ , we have

$$1 - \sum_{k=0}^n \frac{(-1)^k}{k!} \rightarrow 1 - e^{-1} = \boxed{1 - \frac{1}{e}}$$

(b) Clearly there is only one way to match all the sandwiches correctly, so  $\Pr(X = r \mid r = n) = 1/n!$ . Also, note that it is impossible to match all but one sandwich, so  $\Pr(X = r \mid r = n - 1) = 0$ . Only the cases for  $r \leq n - 2$  are nontrivial. Using a similar argument as part (a), we see that for any set of  $m$  sandwiches, the probability that at least one was correctly distributed is

$$\Pr\left(\bigcup_{k=1}^m A_k\right) = \sum_{k=1}^m (-1)^{k-1} \frac{(n - k)!}{n!}$$

and that the probability that *any* set of  $m$  sandwiches contained at least one correct match is

$$\begin{aligned} \sum_{k=1}^m (-1)^{k-1} \binom{n}{k} \frac{(n - k)!}{n!} &= \sum_{k=1}^m (-1)^{k-1} \frac{n!}{(n - k)!k!} \frac{(n - k)!}{n!} \\ &= \sum_{k=1}^m \frac{(-1)^{k-1}}{k!} = 1 + \sum_{k=2}^m \frac{(-1)^{k-1}}{k!} = 1 - \sum_{k=2}^m \frac{(-1)^k}{k!} \end{aligned}$$

So for  $m \geq 2$ , the probability of *no* correct matches is  $\sum_{k=2}^m \frac{(-1)^k}{k!}$  if  $n \geq 2$ , and of course 0 if  $n = 1$ . Therefore the probability of  $r$  matches is the probability of any one set of  $r$  sandwiches all matching

and none of the remaining  $n - r$  sandwiches matching times the number of sets of  $r$  sandwiches; that is,

$$\begin{aligned}\Pr(X = r \mid r \leq n - 2) &= \binom{n}{r} \cdot \frac{(n-r)!}{n!} \cdot \left( \sum_{k=2}^{n-r} \frac{(-1)^k}{k!} \right) = \frac{r!}{(n-r)!r!} \cdot \frac{(n-r)!}{n!} \sum_{k=2}^{n-r} \frac{(-1)^k}{k!} \\ &= \frac{1}{r!} \sum_{k=2}^{n-r} \frac{(-1)^k}{k!}\end{aligned}$$

Therefore we have

$$\boxed{\Pr(X = r) = \begin{cases} \frac{1}{r!} \sum_{k=2}^{n-r} \frac{(-1)^k}{k!} & r \leq n - 2 \\ 0 & r = n - 1 \\ \frac{1}{r!} & r = n \end{cases}}$$

(c)

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{k=1}^n A_k\right) = \sum_{k=1}^n \mathbb{E}(A_k) = \sum_{k=1}^n \Pr(A_k = 1) = n \cdot \frac{1}{n} = \boxed{1}$$

(d)

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

$$\mathbb{E}(X^2) = \mathbb{E}\left(\sum_{k=1}^n A_k\right)^2 = \mathbb{E}\left(\sum_{k=1}^n A_k^2 + 2 \sum_{1 \leq i < j \leq n} A_i A_j\right) = \sum_{k=1}^n \mathbb{E}(A_k^2) + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}(A_i A_j)$$

Because

$$\mathbb{E}(A_k^2) = 1^2 \cdot \Pr(A_k = 1) = \frac{1}{n}$$

$$\mathbb{E}(A_i A_j) = \Pr(A_i = 1 \cap A_j = 1) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)} = \frac{1}{2} \cdot \binom{n}{2}^{-1}$$

we have

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{k=1}^n \frac{1}{n} + 2 \sum_{1 \leq i < j \leq n} \frac{1}{n(n-1)} = 1 + 2 \cdot \binom{n}{2} \cdot \frac{1}{2} \cdot \binom{n}{2}^{-1} = 2 \\ \implies \text{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 2 - 1 = \boxed{1}\end{aligned}$$

**HW3 Problem 2(5).** Verify:  $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$  if  $X$  and  $Y$  are independent.

**Solution.**  $X$  and  $Y$  are independent if and only if

$$\Pr(X \cap Y) = \Pr(X) \cdot \Pr(Y) \iff \Pr(X = x \cap Y = y) = \Pr(X = x) \Pr(Y = y)$$

$$\iff \Pr(X = x | Y = y) \cdot \Pr(Y = y) = \Pr(X = x) \Pr(Y = y) \iff \Pr(X = x | Y = y) = \Pr(X = x)$$

$$\implies E(X | Y) = \sum_x x \cdot \Pr(X = x | Y = y) = \sum_x x \cdot \Pr(X = x) = \mathbb{E}(X)$$

**HW3 Problem 2 (parts 1 - 4).** Verify:

- (1)  $\mathbb{E}(\mathbb{E}(X | Y)) = \mathbb{E}(X)$
- (2)  $\mathbb{E}(g(Y)X | Y) = g(Y)\mathbb{E}(X | Y)$
- (3)  $\text{Cov}(\mathbb{E}(X | Y), Y) = \text{Cov}(X, Y)$
- (4)  $Y$  and  $X - \mathbb{E}(X | Y)$  are uncorrelated.

**Solution.**

(1)

$$\begin{aligned} \mathbb{E}(\mathbb{E}(X | Y)) &= \sum_y \mathbb{E}(X | Y) \Pr(Y = y) = \sum_y \left[ \sum_x x \cdot \Pr(X = x | Y = y) \Pr(Y = y) \right] \\ &= \sum_y \left[ \sum_x x \cdot \Pr(X = x \cap Y = y) \right] = \sum_y \left[ \sum_x x \cdot \Pr(Y = y | X = x) \cdot \Pr(X = x) \right] \\ &= \sum_x \left[ x \cdot \Pr(X = x) \cdot \sum_y (\Pr(Y = y | X = x)) \right] = \sum_x \left[ x \cdot \Pr(X = x) \cdot 1 \right] \\ &= \mathbb{E}(X) \end{aligned}$$

(2) 2

(3)

$$\begin{aligned} \text{Cov}(\mathbb{E}(X | Y), Y) &= \mathbb{E}\left(\left[\mathbb{E}(X | Y) - \mathbb{E}(\mathbb{E}(X | Y))\right]\left[Y - \mathbb{E}(Y)\right]\right) \\ &= \mathbb{E}\left(\left[\mathbb{E}(X | Y) - \mathbb{E}(X)\right]\left[Y - \mathbb{E}(Y)\right]\right) = \mathbb{E}\left(\mathbb{E}(X | Y)Y - \mathbb{E}(X)Y - \mathbb{E}(X | Y)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)\right) \\ &= \mathbb{E}(\mathbb{E}(X | Y)Y) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)\mathbb{E}(\mathbb{E}(X | Y)) + \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(X | Y)Y) - \mathbb{E}(Y)\mathbb{E}(X) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \text{Cov}(X, Y) \end{aligned}$$

- (4)  $Y$  and  $X - \mathbb{E}(X | Y)$  are uncorrelated if and only if  $\text{Cov}(Y, X - \mathbb{E}(X | Y)) = 0 \iff \mathbb{E}(Y \cdot [X - \mathbb{E}(X | Y)]) - \mathbb{E}(Y)\mathbb{E}(X - \mathbb{E}(X | Y)) = 0$ .

$$\mathbb{E}(Y \cdot [X - \mathbb{E}(X | Y)]) - \mathbb{E}(Y)\mathbb{E}(X - \mathbb{E}(X | Y)) = \mathbb{E}(YX - Y\mathbb{E}(X | Y)) - \mathbb{E}(Y)\mathbb{E}(X) + \mathbb{E}(Y)\mathbb{E}(\mathbb{E}(X | Y))$$

$$= \mathbb{E}(YX) - \mathbb{E}(Y\mathbb{E}(X | Y)) - \mathbb{E}(Y)\mathbb{E}(X) + \mathbb{E}(Y)\mathbb{E}(X) = \mathbb{E}(YX) - \mathbb{E}(YX) = 0$$

### Spring 2018 Problem 2 (did not complete)

**2.** Consider positions 1 to  $n$  arranged in a circle, so that 2 comes after 1, 3 comes after 2, ...,  $n$  comes after  $n - 1$ , and 1 comes after  $n$ . Similarly, take 1 to  $n$  as values, with cyclic order, and consider all  $n!$  ways to assign values to positions, bijectively, with all  $n!$  possibilities equally likely. For  $i = 1$  to  $n$ , let  $X_i$  be the indicator that position  $i$  and the one following are filled in with two consecutive values in increasing order, and define

$$S_n = \sum_{i=1}^n X_i, \quad T_n = \sum_{i=1}^n iX_i$$

For example, with  $n = 6$  and the circular arrangement 314562, we get  $X_3 = 1$  since 45 are consecutive in increasing order, and similarly  $X_4 = X_6 = 1$ , so that  $S_6 = 3, T_6 = 13$ .

- a) Compute the mean and the variance of  $S_n$ .
- b) Compute the mean and the variance of  $T_n$ .

**Fall 2008 Problem 2 (HW1 Problem 10).** Consider a lottery with  $n^2$  tickets, of which only  $n$  tickets win prizes. Let  $p_n$  be the probability that, out of  $n$  randomly selected tickets, at least one wins a prize. Compute  $\lim_{n \rightarrow \infty} p_n$ .

**Solution.** There are  $\binom{n^2}{n}$  possible sets of  $n$  tickets. The number of these sets that do not contain at least one winner (that is, they only contain members of the  $n^2 - n$  losing tickets) is  $\binom{n^2 - n}{n}$ . Therefore the probability of selecting a set of  $n$  tickets that contains at least one winner is

$$\begin{aligned} p_n &= 1 - \binom{n^2 - n}{n} / \binom{n^2}{n} = 1 - \frac{(n^2 - n)!}{n!(n^2 - n - n)!} / \frac{(n^2)!}{(n^2 - n)!n!} = 1 - \frac{(n^2 - n)!}{n!(n^2 - 2n)!} \cdot \frac{(n^2 - n)!n!}{(n^2)!} \\ &= 1 - \frac{(n^2 - n)!}{(n^2 - 2n)!} \cdot \frac{(n^2 - n)!}{(n^2)!} = 1 - \prod_{i=0}^{n-1} (n^2 - n - i) / \prod_{i=0}^{n-1} (n^2 - i) = 1 - \prod_{i=0}^{n-1} \frac{n^2 - n - i}{n^2 - i} \\ &= 1 - \prod_{i=0}^{n-1} \left( \frac{n^2 - i}{n^2 - i} - \frac{n}{n^2 - i} \right) = 1 - \prod_{i=0}^{n-1} \left( 1 - \frac{n}{n^2 - i} \right) \end{aligned}$$

Therefore

$$\lim_{n \rightarrow \infty} p_n = \lim_{n \rightarrow \infty} \left[ 1 - \prod_{i=0}^{n-1} \left( 1 - \frac{n}{n^2 - i} \right) \right] = 1 - \lim_{n \rightarrow \infty} \prod_{i=0}^n \left( 1 - \frac{n}{n^2 - i} \right) = 1 - \lim_{n \rightarrow \infty} \prod_{i=0}^n \left( 1 - \frac{n \cdot \frac{1}{n}}{\frac{n^2}{n} - \frac{i}{n}} \right)$$

$$= 1 - \lim_{n \rightarrow \infty} \prod_{i=0}^n \left(1 - \frac{1}{n - \frac{i}{n}}\right) = 1 - \lim_{n \rightarrow \infty} \prod_{i=0}^n \left(1 - \frac{1}{n}\right) = 1 - \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = \boxed{1 - \exp(-1)}$$

### 6.2.3 Problems we did on homework

**Fall 2017 Problem 2 (Homework 3 Problem 6).** An urn contains  $2n$  balls, coming in pairs: two balls are labeled “1”, two balls are labeled “2”, ..., two balls are labeled “ $n$ ”. A sample of size  $n$  is taken without replacement. Denote by  $N$  the number of pairs in the sample. Compute the expected value and the variance of  $N$ . **You do not need to simplify the expression for the variance.**

**Solution.** Let  $X_k$  be an indicator variable for both balls labeled  $k$  being in the sample. Note that

$$\mathbb{E}(X_k) = \Pr(X_k = 1) = \frac{\binom{2n-2}{n-2}}{\binom{2n}{n}} = \frac{(2n-2)!}{(n-2)!n!} / \frac{(2n)!}{n!n!} = \frac{(2n-2)!n!}{(2n)!(n-2)!} = \frac{n(n-1)}{2n(2n-1)} = \frac{n-1}{2(2n-1)}$$

Now since  $N = \sum_{k=1}^n X_k$ , we have

$$\mathbb{E}(N) = \mathbb{E}\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n \mathbb{E}(X_k) = \boxed{\frac{n(n-1)}{2(2n-1)}}$$

To obtain the variance, note that

$$\mathbb{E}(N^2) = \mathbb{E}\left(\sum_{k=1}^n X_k\right)^2 = \mathbb{E}\left(\sum_{k=1}^n X_k^2 + 2 \sum_{1 \leq i < j \leq n} X_i X_j\right) = \sum_{k=1}^n \mathbb{E}(X_k^2) + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}(X_i X_j)$$

Because

$$\mathbb{E}(X_k^2) = 1^2 \cdot \Pr(X_k = 1) = \mathbb{E}(X_k) = \frac{n-1}{2(2n-1)}$$

$$\mathbb{E}(X_i X_j) = \Pr(X_i = 1 \cap X_j = 1) = \frac{\binom{2n-4}{n-4}}{\binom{2n}{n}} = \frac{(2n-4)!}{(n-4)!n!} / \frac{(2n)!}{n!n!} = \frac{(2n-4)!n!}{(2n)!(n-4)!}$$

$$= \frac{n(n-1)(n-2)(n-3)}{2n(2n-1)(2n-2)(2n-3)} = \frac{(n-1)(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)}$$

we have

$$\mathbb{E}(N^2) = \sum_{k=1}^n \frac{n-1}{2(2n-1)} + 2 \sum_{1 \leq i < j \leq n} \frac{(n-1)(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)} = \frac{n(n-1)}{2(2n-1)} + 2 \binom{n}{2} \frac{(2n-4)!n!}{(2n)!(n-4)!}$$

$$\begin{aligned}
&= \frac{n(n-1)}{2(2n-1)} + \frac{n!}{(n-2)!} \cdot \frac{(2n-4)!n!}{(2n)!(n-4)!} = \frac{n(n-1)}{2(2n-1)} + n(n-1) \cdot \frac{(n-1)(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)} \\
&= \frac{n(n-1)}{2(2n-1)} + \frac{n(n-1)^2(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)} \\
\implies \text{Var}(N) &= \mathbb{E}(N^2) - \mathbb{E}(N)^2 = \boxed{\frac{n(n-1)}{2(2n-1)} + \frac{n(n-1)^2(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)} - \frac{n^2(n-1)^2}{4(2n-1)^2}}
\end{aligned}$$

**Fall 2017 Problem 3 (HW3 Problem 8—almost full solution)**

Let  $U_1, U_2, \dots$  be iid random variables, uniformly distributed on  $[0, 1]$ , and let  $N$  be a Poisson random variable with mean value equal to 1. Assume that  $N$  is independent of  $U_1, U_2, \dots$  and define

$$Y = \begin{cases} 0 & \text{if } N = 0 \\ \max_{1 \leq i \leq N} U_i & \text{if } N > 0 \end{cases}$$

Compute the expected value of  $Y$ .

**Solution.** Since  $Y$  is a function of  $N$ , let  $Y = y(N)$ . By the Law of the Unconscious Statistician,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y | N)) = \mathbb{E}(\mathbb{E}(\max_{1 \leq i \leq N} U_i | N = n))$$

Let  $Z_n = \max_{1 \leq i \leq n} U_i$ . The cdf of  $Z_n$  can be calculated as follows:

$$\Pr(Z_n \leq x) = \Pr(\max_{1 \leq i \leq n} U_i \leq x) = \Pr(U_1 \leq x \cap U_2 \leq x \cap \dots \cap U_n \leq x) = x^n$$

for  $x \in [0, 1]$ . Therefore the pdf of  $Z_n$  is its derivative,  $nx^{n-1}$ . So we have

$$\mathbb{E}(\max_{1 \leq i \leq N} U_i | N = n) = \mathbb{E}(Z_n) = \int_0^1 x n x^{n-1} dx = n \int_0^1 x^n dx = n \frac{x^{n+1}}{n+1} \Big|_0^1 = \frac{n}{n+1}$$

Plugging this into the expression for  $\mathbb{E}(Y)$  yields

$$\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(\mathbb{E}(Y | N)) = \sum_{n=0}^{\infty} \frac{n}{n+1} \Pr(N = n) = \sum_{n=1}^{\infty} \frac{n}{n+1} \frac{\exp(-1) 1^n}{n!} \\
&= \frac{1}{e} \sum_{n=1}^{\infty} \frac{n+1-1}{(n+1)!} = \frac{1}{e} \left( \sum_{n=1}^{\infty} \frac{n+1}{(n+1)!} - \sum_{n=1}^{\infty} \frac{1}{(n+1)!} \right) = \frac{1}{e} \left( \sum_{n=1}^{\infty} \frac{1}{n!} - \sum_{m=2}^{\infty} \frac{1}{m!} \right)
\end{aligned}$$

$$= \frac{1}{e} [e - 1 - (e - 1 - 1)] = \boxed{\frac{1}{e}}$$

**Fall 2013 Problem 3/Spring 2011 Problem 2 (HW3 Problem 9; coupon collector problem)**

Only parts I didn't do: Let  $D$  be the event that no box receives more than 1 ball. Fix  $a \in (0, 1)$ . If both  $n, d \rightarrow \infty$  together, what relation must they satisfy in order to have  $\Pr(D) \rightarrow a$ ?

**HW3 Problem 9.** Consider  $n$  (different) balls placed at random in  $m$  boxes so that each of  $m^n$  configurations is equally likely.

- (a) Compute the expected value and the variance of the number of empty boxes.
  - (b) Show that if  $\lim_{m,n \rightarrow \infty} m \exp(-n/m) = \lambda \in (0, \infty)$ , then, in the same limit, the number of empty boxes has Poisson distribution with parameter  $\lambda$ .
  - (c) For  $k \geq 1$  such that  $k+3 \leq m$ , define the event  $A_k$  that the boxes  $k, k+1, k+2, k+3$  are empty. Assuming that  $m > 8$ , compute  $\Pr(A_1 \cup A_3 \cup A_5)$ . How will the answer change if  $m = 8$ ?
  - (d) Now imagine that the balls are dropped one-by-one (with each ball equally likely to go into any of the  $m$  boxes, independent of all other balls), and denote by  $N_m$  the minimal number of balls required to fill all the boxes. Compute  $\mathbb{E}(N_m)$ ,  $\text{Var}(N_m)$  and
- $$\lim_{m \rightarrow \infty} \Pr\left(\frac{N_m - m \log m}{m} \leq x\right)$$
- (e) Suppose we instead place an unlimited number of balls into the  $m$  boxes until we have  $k$  consecutive balls land in the same box (it doesn't matter which box). What is the expected number of balls we will drop until this happens?

**Solution.**

- (a) Let  $A_i$  be the event that the  $i$ th box is empty. Let  $\mathbf{1}_{A_i}$  be the indicator for  $A_i$ . Then  $X = \sum_{i=1}^m \mathbf{1}_{A_i}$ .

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^m \mathbf{1}_{A_i}\right) = \sum_{i=1}^m (\mathbb{E} \mathbf{1}_{A_i}) = \sum_{i=1}^m \Pr(A_i) = \sum_{i=1}^m \left(\frac{m-1}{m}\right)^n = \boxed{\frac{(m-1)^n}{m^{n-1}}}$$

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^m \mathbf{1}_{A_i}\right) = \sum_{i=1}^m \text{Var}(\mathbf{1}_{A_i}) + 2 \sum_{1 \leq i < j \leq m} \text{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j})$$

$$\text{Var}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) = \mathbb{E}(\mathbf{1}_{A_i} \mathbf{1}_{A_j}) - \mathbb{E}(\mathbf{1}_{A_i})^2 = \Pr(A_i \cap A_j) - \Pr(A_i)^2 = \left(\frac{m-1}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}$$

$$\text{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) = \mathbb{E}(\mathbf{1}_{A_i} \mathbf{1}_{A_j}) - \mathbb{E}(\mathbf{1}_{A_i}) \mathbb{E}(\mathbf{1}_{A_j}) = \Pr(A_i \cap A_j) - \Pr(A_i) \Pr(A_j) = \left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}$$

$$\implies \text{Var}(X) = m \cdot \left[ \left(\frac{m-1}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n} \right] + \frac{m!}{(m-2)!} \left[ \left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n} \right]$$

$$= \frac{(m-1)^n}{m^{n-1}} - \frac{(m-1)^{2n}}{m^{2n-1}} + (m^2 - m) \left[ \left( \frac{m-2}{m} \right)^n - \left( \frac{m-1}{m} \right)^{2n} \right]$$

$$\boxed{\text{Var}(X) = \frac{(m-1)^n}{m^{n-1}} - \frac{(m-1)^{2n}}{m^{2n-1}} + (m-1) \left[ \frac{(m-2)^n}{m^{n-1}} - \frac{(m-1)^{2n}}{m^{2n-1}} \right]}$$

(b) Note that

$$X = \sum_{i=1}^m \mathbf{1}_{A_i}$$

and that the  $A_i$  are only weakly dependent on each other, especially as  $m$  and  $n$  increase. Therefore as  $m, n \rightarrow \infty$ , the Poisson paradigm (see Section 6.1.10) suggests  $X \sim \text{Poisson}(\mathbb{E}(X))$ . We have

$$\mathbb{E}(X) = \frac{(m-1)^n}{m^{n-1}}$$

so

$$\begin{aligned} \lim_{n,m \rightarrow \infty} \mathbb{E}(X) &= \lim_{n,m \rightarrow \infty} m \cdot \left( \frac{m-1}{m} \right)^n = \lim_{n,m \rightarrow \infty} m \cdot \left( 1 - \frac{1}{m} \right)^n = \lim_{n,m \rightarrow \infty} m \cdot \left[ \left( 1 - \frac{1}{m} \right)^m \right]^{n/m} \\ &\approx \lim_{n,m \rightarrow \infty} m \cdot [e^{-1}]^{n/m} = \lim_{n,m \rightarrow \infty} m e^{-n/m} \end{aligned}$$

Using

$$\lim_{m,n \rightarrow \infty} m \exp(-n/m) = \lambda \in (0, \infty)$$

we have  $\boxed{X \sim \text{Poisson}(\lambda) \text{ as } m, n \rightarrow \infty}$ .

(c)

$$\Pr(A_1 \cup A_3 \cup A_5) = \Pr(A_1) + \Pr(A_3) + \Pr(A_5) - \Pr(A_1 \cap A_3) - \Pr(A_1 \cap A_5) - \Pr(A_3 \cap A_5) + \Pr(A_1 \cap A_3 \cap A_5)$$

We have

$$\Pr(A_1) = \Pr(A_3) = \Pr(A_5) = \left( \frac{m-4}{m} \right)^n$$

$$\Pr(A_1 \cap A_3) = \Pr(A_3 \cap A_5) = \left( \frac{m-6}{m} \right)^n$$

$$\Pr(A_1 \cap A_5) = \Pr(A_1 \cap A_3 \cap A_5) = \left( \frac{m-8}{m} \right)^n$$

Therefore

$$\Pr(A_1 \cup A_3 \cup A_5) = 3 \left( \frac{m-4}{m} \right)^n - 2 \left( \frac{m-6}{m} \right)^n = \boxed{\frac{3(m-4)^n - 2(m-6)^n}{m^n}}$$

- (d)  $N_m$  is the minimal number of balls required to fill all the boxes. Let  $T_i$  be the number of balls that have to be dropped to fill the  $i$ th box after  $i-1$  boxes have been filled. The probability of filling a new box after  $i-1$  boxes have been filled is  $\frac{m-(i-1)}{m}$ . (Note that  $T_1$  should be identically 1 regardless of  $m \geq 1$ ; this checks out using this expression.) Therefore  $T_i$  has a geometric distribution with  $E(T_i) = \frac{m}{m-(i-1)}$ . Since  $N_m = \sum_{i=1}^m T_i$ , we have

$$\mathbb{E}(N_m) = \mathbb{E}\left(\sum_{i=1}^m T_i\right) = \sum_{i=1}^m \mathbb{E}(T_i) = \sum_{i=1}^m \frac{m}{m-(i-1)} = \boxed{m \sum_{i=1}^m \frac{1}{i}}$$

Because the  $T_i$  are independent, we have

$$\begin{aligned} \text{Var}(N_m) &= \text{Var}\left(\sum_{i=1}^m T_i\right) = \sum_{i=1}^m \text{Var}(T_i) = \sum_{i=1}^m \left(1 - \frac{m-(i-1)}{m}\right) \left/\left(\frac{m-(i-1)}{m}\right)^2\right. \\ &= \sum_{i=1}^m \frac{i-1}{m} \cdot \left(\frac{m}{m-(i-1)}\right)^2 = \boxed{m \sum_{i=1}^m \frac{i-1}{[m-(i-1)]^2}} \end{aligned}$$

Finally, to find

$$\lim_{m \rightarrow \infty} \Pr\left(\frac{N_m - m \log m}{m} \leq x\right)$$

begin by noting that we can also express  $N_m$  as

$$\Pr(N_m \leq k) = \Pr(X_{m,k} = 0)$$

where  $X_{m,k}$  is defined as  $X$  is in part (b) with  $k$  being the number of balls that have been dropped so far,  $k \in \mathbb{N} \geq m$ . (For  $k < m$ ,  $\Pr(N_m \leq k) = 0$ .)

Again, let  $A_{i,k}$  be the event that the  $i$ th box is empty after dropping  $k$  balls. Then because  $X_{m,k} = \sum_{i=1}^m \mathbf{1}_{A_{i,k}}$  and the  $A_{i,k}$  are only weakly dependent on each other (especially as  $m$  becomes large), the Poisson paradigm (see Section 6.1.10) again suggests that as  $m \rightarrow \infty$ ,  $X_{m,k} \sim \text{Poisson}(\lambda_k)$  where  $\lambda_k = \mathbb{E}(X_{m,k})$  is defined as above. Therefore we have

$$\begin{aligned} \lim_{m \rightarrow \infty} \Pr\left(\frac{N_m - m \log m}{m} \leq x\right) &= \lim_{m \rightarrow \infty} \Pr(N_m \leq xm + m \log m) = \lim_{m \rightarrow \infty} \Pr(X_{m,xm+m \log m} \\ &= 0) \approx \frac{\exp(-\lambda_{xm+m \log m}) \cdot \lambda_{xm+m \log m}^0}{0!} = \exp(-\lambda_{xm+m \log m}) \end{aligned}$$

And we have

$$\begin{aligned} \lambda_{xm+m \log m} &= \lim_{m \rightarrow \infty} m \exp\left(-\frac{xm + m \log m}{m}\right) = \lim_{m \rightarrow \infty} m \exp(-x - \log m) = \lim_{m \rightarrow \infty} m/m \exp(-x) \\ &= \exp(-x) \end{aligned}$$

which yields

$$\lim_{m \rightarrow \infty} \Pr \left( \frac{N_m - m \log m}{m} \leq x \right) = \exp(\exp(-x))$$

- (e) Let  $N = N_k$  be the number of balls that are dropped until  $k$  consecutive balls land in the same box, and likewise for  $N_{k-1}$ . Suppose we have already observed  $k-1$  consecutive outcomes (of any kind) in  $N_{k-1}$  trials. Then we finish on the next term (by having another consecutive outcome) with probability  $1/m$ . Otherwise we have a different outcome and then repeat the same process again. So we have

$$\mathbb{E}(N_k | N_{k-1}) = N_{k-1} + 1 \cdot \frac{1}{m} + \mathbb{E}(N_k) \cdot \left(1 - \frac{1}{m}\right)$$

Therefore

$$\begin{aligned} \mathbb{E}(N) &= \mathbb{E}(N_k) = \mathbb{E}[\mathbb{E}(N_k | N_{k-1})] = \mathbb{E}(N_{k-1}) + \frac{1}{m} + \left(1 - \frac{1}{m}\right)\mathbb{E}(N_k) \\ &\iff \frac{1}{m}\mathbb{E}(N_k) = \mathbb{E}(N_{k-1}) + \frac{1}{m} \iff \mathbb{E}(N_k) = m\mathbb{E}(N_{k-1}) + 1 \end{aligned}$$

We have a recursive formula. Note that  $\mathbb{E}(N_1) = 1$  because the number of trials until there is 1 consecutive outcome of any kind is simply 1. We can then calculate as follows:

$$\mathbb{E}(N_2) = m\mathbb{E}(N_{2-1}) + 1 = m + 1$$

$$\mathbb{E}(N_3) = m\mathbb{E}(N_{3-1}) + 1 = m(m + 1) + 1 = 1 + m + m^2$$

$$\mathbb{E}(N_4) = m\mathbb{E}(N_{4-1}) + 1 = m(1 + m + m^2) + 1 = 1 + m + m^2 + m^3$$

⋮

$$\mathbb{E}(N_k) = \sum_{i=0}^{k-1} m^i = \frac{1 \cdot (1 - m^k)}{1 - m} = \boxed{\frac{m^k - 1}{m - 1}}$$

**Fall 2012 Problem 1 (HW2 Problem 10/HW 1 Problem 9)** Only part I didn't do: Find the mean and variance of  $S_n = X_1 + \dots + X_n$ , the total number of white balls added to the urn up to time  $n$ .

**HW1 Problem 9.** An urn contains  $b$  black and  $w$  white balls. At each step, a ball is removed from the urn at random and then put back together with one more ball of the same color. Compute the probability  $p_n$  to get a black ball on step  $n$ ,  $n \geq 1$ .

**Solution. Step 1:**

$$p_1 = \frac{b}{b+w}$$

**Step 2:** We need to separately consider the cases where a black ball was selected on step 1 (with probability  $p_1$ ) or a white ball (with probability  $1 - p_1$ ).

$$\begin{aligned}
 p_2 &= p_1 \cdot \frac{b+1}{b+w+1} + (1-p_1) \cdot \frac{b}{b+w+1} = p_1 \left( \frac{b+1}{b+w+1} - \frac{b}{b+w+1} \right) + \frac{b}{b+w+1} \\
 &= p_1 \left( \frac{1}{b+w+1} + \frac{1}{p_1} \frac{b}{b+w+1} \right) = p_1 \left( \frac{1}{b+w+1} + \frac{b+w}{b} \frac{b}{b+w+1} \right) \\
 &= p_1 \left( \frac{b+w+1}{b+w+1} \right) = p_1 \\
 \implies p_2 &= p_1 = \frac{b}{b+w}
 \end{aligned}$$

**Step 3:** Regardless of the previous steps, there are now  $b+w+2$  balls in the urn. Since we know that  $p_1 = p_2$ , the probability that we have selected  $k$  black balls so far (and thus, the probability that there are currently  $b+k$  black balls in the urn) is given by

$$\begin{aligned}
 \Pr(k \text{ balls chosen in first 2 rounds}) &= \binom{2}{k} p_1^k (1-p_1)^{2-k} = \binom{2}{k} \left( \frac{b}{b+w} \right)^k \left( \frac{w}{b+w} \right)^{2-k} \\
 &= \binom{2}{k} \frac{b^k w^{2-k}}{(b+w)^2}
 \end{aligned}$$

for  $k \in \{0, 1, 2\}$ . Given that we have selected  $k$  black balls so far, the probability of selecting a black ball this time is  $\frac{b+k}{b+w+2}$ . Therefore the probability of selecting a black ball this round is

$$\begin{aligned}
 p_3 &= \sum_{k=0}^2 \binom{2}{k} \frac{b^k w^{2-k}}{(b+w)^2} \frac{b+k}{b+w+2} = \frac{1}{(b+w+2)(b+w)^2} \sum_{k=0}^2 \binom{2}{k} (b+k) b^k w^{2-k} \\
 &= \frac{1}{(b+w+2)(b+w)^2} \left( \binom{2}{0} bw^2 + \binom{2}{1} (b+1)bw + \binom{2}{2} (b+2)b^2 \right) \\
 &= \frac{bw^2 + 2(b+1)bw + (b+2)b^2}{(b+w+2)(b+w)^2} = \frac{b}{b+w} \left( \frac{w^2 + 2bw + 2w + b^2 + 2b}{b^2 + bw + 2b + wb + w^2 + 2w} \right) \\
 &= \frac{b}{b+w} \left( \frac{w^2 + 2bw + 2w + b^2 + 2b}{b^2 + 2bw + 2b + w^2 + 2w} \right) = \frac{b}{b+w} = p_1
 \end{aligned}$$

There seems to be a clear pattern here. Let's find the general formula by induction.

**Step  $n + 1$ :** Assume that the probability of choosing a black ball on steps  $1, 2, \dots, n$  was  $\frac{b}{b+w}$  each time.  
(a bunch of boring stuff, then it worked.)

**HW2 Problem 10.** Random variables  $(X_1, \dots, X_n)$  are called *exchangeable* if  $\Pr(X_1 = x_1, \dots, X_n = x_n) = \Pr(X_{\tau(1)} = x_1, \dots, X_{\tau(n)} = x_n)$  for all real numbers  $x_1, \dots, x_n$  and every permutation  $\tau$  of the set  $\{1, \dots, n\}$ . In the setting of Problem 9 from Homework 1, let  $X_k = 1$  if a white ball is drawn on step  $k$ , and  $X_k = 0$  otherwise. Show that the random variables  $X_1, \dots, X_n$  are exchangeable for every  $n \geq 2$ .

**Solution.** For  $n = 2$ : There are two cases which we must show are equal to show exchangeability:

$$\Pr(X_1 = 0, X_2 = 1) = \Pr(X_1 = 1, X_2 = 0)$$

First,

$$\begin{aligned} \Pr(X_1 = 0, X_2 = 1) &= \Pr(\text{black first}) \Pr(\text{white second} \mid \text{black first}) = \left(\frac{b}{b+w}\right) \left(\frac{w}{b+w+1}\right) \\ &\quad \left(\frac{w}{b+w}\right) \left(\frac{b}{b+w+1}\right) = \Pr(X_1 = 1, X_2 = 0) \end{aligned}$$

which proves exchangeability for  $n = 2$ . In the general case, we seek to show that  $X_1, \dots, X_n$  are exchangeable. That is, in all  $n+1$  unordered sets  $\mathbb{X}_k = \{x_{1k}, x_{2k}, \dots, x_{nk} \mid x_{ik} \in \{0, 1\}, \sum_i x_{ik} = k\}$ , in all  $\binom{n}{k}$  permutations of  $\mathbb{X}_k$ ,

$$\Pr(\mathbb{X}_{kj} = \Pr(\mathbb{X}_{kj'})$$

where  $j$  and  $j'$  denote different permutations of  $\mathbb{X}_k$ . That is,

$$\Pr(X_1 = x_{1k}, X_2 = x_{2k}, \dots, X_n = x_{nk}) = \Pr(X_{j_1} = x_{1k}, X_{j_2} = x_{2k}, \dots, X_{j_n} = x_{nk})$$

where  $j_1, j_2, \dots, j_n$  index the permuted variables. Consider  $\mathbb{X}_{kj^*}$  where all  $k$  white balls are chosen first and all  $n - k$  black balls are chosen last. We have

$$\begin{aligned} \Pr(\mathbb{X}_{kj^*}) &= \prod_{i=1}^k \left(\frac{w+i-1}{b+w+i-1}\right) \cdot \prod_{i=k+1}^n \left(\frac{b+i-k-1}{b+w+i-1}\right) \\ &= \prod_{i=1}^n \left(\frac{1}{b+w+i-1}\right) \cdot \left[ \prod_{i=1}^k (w+i-1) \prod_{i=k+1}^n (b+i-k-1) \right] = \prod_{i=1}^n \left(\frac{1}{b+w+i-1}\right) \cdot \left[ \prod_{i=1}^k (w+i-1) \prod_{i'=1}^{n-k} (b+i'-1) \right] \end{aligned}$$

It is easy to see that the leftmost product will always equal the product of the denominators, regardless of the permutation, since one ball is added to the urn after every draw. Similarly, regardless of permutation, the numerator of the probability of drawing the  $i$ th white ball will always equal  $w + i - 1$ , the number of white balls already in the urn. Likewise, the numerator of the probability of drawing the  $i'$ th black ball is

always  $b + i' - 1$ . Because multiplication is commutative, all permutations of these numbers will have equal products. Therefore  $\Pr(\mathbb{X}_{kj^*}) = \Pr(\mathbb{X}_{kj})$  for all  $k$ . That is,

$$\Pr(X_1 = x_1, \dots, X_n = x_n) = \Pr(X_{\tau(1)} = x_1, \dots, X_{\tau(n)} = x_n)$$

for all  $(x_1, \dots, x_n) \in \mathbb{R}^n$ , all  $n \in \mathbb{Z}$  such that  $n \geq 2$ , all permutations  $\tau$ .

**Homework 2 Problem 2.** Consider the function

$$f(x) = \begin{cases} C(2x - x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Could  $f$  be a distribution function? If so, determine  $C$ .
- (b) Could  $f$  be a probability density function? If so, determine  $C$ .

**Solution.**

- (a) If  $f$  is a distribution function,  $\lim_{x \rightarrow -\infty} f(x) = 0$ ,  $\lim_{x \rightarrow \infty} f(x) = 1$ , and  $f'(x) \geq 0 \forall x \in \mathbb{R}$ .  $f$  clearly does not meet the second or third conditions and is therefore not a distribution function.
- (b) If  $f$  is a density function then  $\int_{-\infty}^{\infty} f(x)dx = 1$  and  $f(x) \geq 0 \forall x \in \mathbb{R}$ .

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_0^2 C(2x - x^2)dx = C \left[ x^2 - \frac{x^3}{3} \right]_0^2 = C \left( 4 - \frac{8}{3} - 0 \right) = C \cdot \frac{4}{3} \\ &= 1 \iff C = \frac{3}{4} \end{aligned}$$

Next we check that  $f$  is always nonnegative. It equals zero except on  $(0, 2)$ .

$$\frac{3}{4}(2x - x^2) \geq 0 \iff x(2 - x) \geq 0 \iff x \in (0, 2)$$

Therefore  $f$  is nonnegative  $\forall x \in \mathbb{R}$ , so  $f$  is a probability density function if  $C = \frac{3}{4}$ .

**HW1 Problem 8.** Two people,  $A$  and  $B$ , are involved in a duel. The rules are simple: shoot at each other once; if at least one is hit, the duel is over, if both miss, repeat (go to the next round), and so on. Denote by  $p_A$  and  $p_B$  the probabilities that  $A$  hits  $B$  and  $B$  hits  $A$  with one shot, and assume that hitting/missing is independent from round to round. Compute the probabilities of the following events:

- (a) the duel ends and  $A$  is not hit;
- (b) the duel ends and both are hit;
- (c) the duel ends after round number  $n$ ;
- (d) the duel ends after round number  $n$  GIVEN that  $A$  is not hit;
- (e) the duel ends after  $n$  rounds GIVEN that both are hit;
- (f) the duel goes on forever.

**Solution.**

- (a) Let  $A_k$  denote the event that the duel is ended by  $A$  shooting  $B$  in the  $k$ th round (with neither person being shot in the first  $k - 1$  rounds). Note that  $\{A_k|k = 1, 2, \dots\}$  are all mutually exclusive. Therefore the probability of the duel ending without  $A$  being hit is  $\sum_{k=1}^{\infty} A_k$ . Because the probabilities in each round are constant and independent,

$$A_k = (1 - p_A)^{k-1} p_A (1 - p_B)^k$$

So the probability that the duel ends and  $A$  is not hit is

$$\sum_{k=1}^{\infty} A_k = \sum_{k=1}^{\infty} (1 - p_A)^{k-1} p_A (1 - p_B)^k = p_A (1 - p_B) \sum_{k=1}^{\infty} (1 - p_A)^{k-1} (1 - p_B)^{k-1}$$

This is an infinite geometric series. Since the ratio  $(1 - p_A)(1 - p_B)$  has absolute value less than 1, the sum can be calculated.

$$\sum_{k=1}^{\infty} A_k = p_A (1 - p_B) \cdot \frac{1}{1 - (1 - p_A)(1 - p_B)} = \frac{p_A (1 - p_B)}{p_A + p_B - p_A p_B} = \boxed{\frac{p_A (1 - p_B)}{p_A (1 - p_B) + p_B}}$$

- (b) Similar to part (a). Let  $C_k$  denote the event that the duel is ended with both players being shot in the  $k$ th round (with neither person being shot in the first  $k - 1$  rounds). Again,  $\{C_k|k = 1, 2, \dots\}$  are all mutually exclusive, so the probability of the duel ending in these circumstances is  $\sum_{k=1}^{\infty} C_k$ . We have

$$C_k = (1 - p_A)^{k-1} p_A (1 - p_B)^{k-1} p_B$$

$$\begin{aligned} \sum_{k=1}^{\infty} C_k &= \sum_{k=1}^{\infty} (1 - p_A)^{k-1} p_A (1 - p_B)^{k-1} p_B = p_A p_B \sum_{k=1}^{\infty} (1 - p_A)^{k-1} (1 - p_B)^{k-1} \\ &= p_A p_B \cdot \frac{1}{1 - (1 - p_A)(1 - p_B)} = \boxed{\frac{p_A p_B}{p_A + p_B - p_A p_B}} \end{aligned}$$

Note that this value is less than the answer from part (a) if  $p_B < \frac{1}{2}$  and greater if  $p_B > \frac{1}{2}$

- (c) Let  $B_k$  denote the event that the duel is ended by  $B$  shooting  $A$  in the  $k$ th round (with neither person being shot in the first  $k - 1$  rounds), with

$$B_k = (1 - p_A)^k p_B (1 - p_B)^{k-1}$$

Let  $A_k$  and  $C_k$  be defined as above. Note that  $\{A_k|k = 1, 2, \dots\}$ ,  $\{B_k|k = 1, 2, \dots\}$ ,  $\{C_k|k = 1, 2, \dots\}$  are all mutually exclusive, and that the event that the duel ends in round  $n$  is  $\{A_n \cup B_n \cup C_n\}$ . So the probability of the duel ending in round  $n$  is

$$\begin{aligned} \Pr(A_n \cup B_n \cup C_n) &= \Pr(A_n) + \Pr(B_n) + \Pr(C_n) \\ &= (1 - p_A)^{n-1} p_A (1 - p_B)^n + (1 - p_A)^n p_B (1 - p_B)^{n-1} + (1 - p_A)^{n-1} p_A (1 - p_B)^{n-1} p_B \\ &= (1 - p_A)^{n-1} (1 - p_B)^{n-1} [p_A (1 - p_B) + (1 - p_A) p_B + p_A p_B] \\ &= \boxed{(1 - p_A)^{n-1} (1 - p_B)^{n-1} (p_A + p_B - p_A p_B)} \end{aligned}$$

- (d) Let  $A_k, B_k, C_k$  be defined as above. The event that the duel ends at round  $n$  without  $A$  being hit is given by  $\{A_n\}$ .

$$\Pr(A_n) = \boxed{(1 - p_A)^{n-1} p_A (1 - p_B)^n}$$

- (e) Let  $A_k, B_k, C_k$  be defined as above. The event that the duel ends at round  $n$  with both players being hit is given by  $\{C_n\}$ .

$$\Pr(C_n) = \boxed{(1 - p_A)^{n-1} p_A (1 - p_B)^{n-1} p_B}$$

- (f) Let  $A_k, B_k, C_k$  be defined as above. The probability that the duel never ends is equal to 1 - the probability that the duel ends at some point, which is  $\{A_k | k = 1, 2, \dots\} \cup \{B_k | k = 1, 2, \dots\} \cup \{C_k | k = 1, 2, \dots\}$ . Since all of these events are mutually exclusive, we have

$$\begin{aligned} 1 - \Pr(\{A_k | k = 1, 2, \dots\} \cup \{B_k | k = 1, 2, \dots\} \cup \{C_k | k = 1, 2, \dots\}) &= 1 - \sum_{k=1}^{\infty} (A_k + B_k + C_k) \\ &= 1 - \sum_{k=1}^{\infty} ((1 - p_A)^{k-1} p_A (1 - p_B)^k + (1 - p_A)^k p_B (1 - p_B)^{k-1} + (1 - p_A)^{k-1} p_A (1 - p_B)^{k-1} p_B) \\ &= 1 - [p_A(1 - p_B) + (1 - p_A)p_B + p_A p_B] \sum_{k=1}^{\infty} (1 - p_A)^{k-1} (1 - p_B)^{k-1} \\ &= 1 - [p_A(1 - p_A p_B) + p_B(1 - p_A) p_B + p_A p_B] \cdot \frac{1}{1 - (1 - p_A)(1 - p_B)} \\ &= 1 - \frac{p_A - p_A p_B + p_B - p_A p_B + p_A p_B}{p_A + p_B - p_A p_B} = 1 - \frac{p_A + p_B - p_A p_B}{p_A + p_B - p_A p_B} = \boxed{0} \end{aligned}$$

### Homework 1 Problem 1.

- (I) Seven different gifts are distributed among 10 children. How many different outcomes are possible if every child can receive (a) at most one gift, (b) at most two gifts, (c) any number of gifts?  
 (II) Answer the same questions if the gifts are identical (but the children are still different).

### Solution.

(I) (a)  $\binom{10}{7} 7! = \boxed{604,800}$

(b) Clearly all outcomes that satisfy part (I)(a) also satisfy these conditions, so we start with  $\binom{10}{7} 7! = 604,800$  possible outcomes. In addition, the following outcomes are possible:

- (i) **A set of 6 children receive gifts; one child receives two gifts.** There are  $\binom{10}{6}$  ways to pick a group of 6 children to receive the gifts. Next, there are  $\binom{6}{1} = 6$  ways to choose which child receives two gifts. Finally, there are  $7!/2!$  unique ways to distribute the gifts among the children once a particular partition is chosen (since order matters for all of the gifts except for the two that are received by the same child).

(ii) **A set of 5 children receive gifts; two children receive two gifts.** There are  $\binom{10}{5}$  ways to pick a group of 5 children to receive the gifts. Next, there are  $\binom{5}{2}$  ways to choose which of these children receive one gift and which receive two. Finally, there are  $7!/(2!2!)$  unique ways to distribute the gifts among the children once a particular partition is chosen (since order matters for all of the gifts except for the two batches of two gifts that are received by the same child).

(Note that without the restriction that a child can receive at most two gifts, another possibility is that 1 child could receive 3 gifts, but that wouldn't work in this case.)

(iii) **A set of 4 children receive gifts; three children each receive two gifts.** There are  $\binom{10}{4}$  ways to pick a group of 4 children to receive the gifts. Next, there are  $\binom{4}{3} = 4$  ways to choose which of these children receive one gift and which receive two. Finally, there are  $7!/(2!2!2!)$  unique ways to distribute the gifts among the children once a particular partition is chosen (since order matters for all of the gifts except for the three batches of two gifts that are received by the same child).

(Again, there are other possibilities for 4 children to receive 7 gifts, but none that satisfy the condition that no child receives more than 2 gifts.)

Clearly each of these outcomes are mutually exclusive. Therefore the answer is

$$\begin{aligned} & \binom{10}{7} 7! + \binom{10}{6} \cdot \binom{6}{1} \cdot \frac{7!}{2!} + \binom{10}{5} \cdot \binom{5}{2} \cdot \frac{7!}{2!2!} + \binom{10}{4} \cdot \binom{4}{3} \cdot \frac{7!}{2!2!2!} \\ &= 7! \cdot \left( \frac{10!}{3!} + \frac{10!}{6!4!} \cdot 6 \cdot \frac{1}{2} + \frac{10!}{5!5!} \cdot \frac{5!}{3!2!} \cdot \frac{1}{4} + \frac{10!}{4!6!} \cdot \frac{4!}{3!} \cdot \frac{1}{8} \right) \\ &= 7!10! \cdot \left( \frac{1}{3!} + \frac{1}{6!4!} \cdot \frac{6}{2} + \frac{1}{5!} \cdot \frac{1}{3!2!} \cdot \frac{1}{4} + \frac{1}{6!3!} \cdot \frac{1}{8} \right) \\ &= \boxed{7,484,400} \end{aligned}$$

(c)  $10^7 = \boxed{10,000,000}$

(II) (a)  $\binom{10}{7} = \boxed{120}$

(b) Clearly all outcomes that satisfy part (I)(a) also satisfy these conditions, so we start with  $\binom{10}{7} = 120$  possible outcomes. In addition, the following outcomes are possible:

- (i) A set of 6 children receive gifts; one child receives two gifts (6 distinct ways this could happen for each set of 6 children).
- (ii) A set of 5 children receive gifts; two children receive two gifts ( $\binom{5}{2}$  distinct ways this could happen for each set of 5 children).
- (iii) A set of 4 children receive gifts; three children each receive two gifts (4 distinct ways this could happen for each set of 4 children).

Clearly each of these outcomes are mutually exclusive. Therefore the answer is

$$\binom{10}{7} + \binom{10}{6} \cdot \binom{6}{7-6} + \binom{10}{5} \cdot \binom{5}{7-5} + \binom{10}{4} \cdot \binom{4}{7-4} = \boxed{4,740}$$

(c) By Proposition 6.1.17, the number of nonnegative integer-valued vectors  $(x_1, x_2, \dots, x_r)$  satisfying the equation

$$x_1 + x_2 + \dots + x_r = n$$

is equal to  $\binom{n+r-1}{r-1} = \binom{7+10-1}{10-1} = \boxed{11,440}$ .

**Homework 1 Problem 2.**

- (I) 20 different gifts are distributed among seven children. How many different outcomes are possible if every child can receive (a) at least one gift, (b) at least two gifts, (c) any number of gifts?
- (II) Answer the same questions if the gifts are identical (but the children are still different).
- (III) Now try to generalize problems (1) and (2).

**Solution.**

- (I) (a) There are  $7^{20}$  possible allocations of gifts if we have no restrictions. If one child doesn't get a gift, there are  $\binom{7}{1}$  ways to choose which child that is and  $6^{20}$  subsequent allocations of gifts. Likewise, there are  $\binom{7}{2} \cdot (7 - 2)^{20}$  ways to allocate the gifts if two children don't receive gifts,  $\binom{7}{3} \cdot (7 - 3)^{20}$  ways if three children don't receive gifts,  $\binom{7}{4} \cdot (7 - 4)^{20}$  ways if four children don't receive gifts,  $\binom{7}{5} \cdot (7 - 5)^{20}$  ways if five children don't receive gifts, and  $\binom{7}{6} \cdot (7 - 6)^{20}$  ways if six children don't receive gifts.

Let  $A_i$  denote the number of allocations in which  $i$  children do not receive gifts. In order to make the calculation, we must use the Inclusion-Exclusion principle (Proposition 6.1.13) (because, for example, some of the allocations in which three children don't receive gifts include allocations where four or more children don't receive gifts, and we don't want to double-count). Therefore the number of ways that at least one child can not receive a gift (i.e. the complement of every child receiving at least one gift) is

$$\begin{aligned} \left| \bigcup_{i=1}^6 A_i \right| &= \sum_{i=1}^6 |A_i| - \sum_{1 \leq i < j \leq 6} |A_i \cap A_j| + \sum_{1 \leq i < j < k \leq 6} |A_i \cap A_j \cap A_k| - \dots \\ &\quad + (-1)^{6-1} |A_1 \cap A_2 \cap A_3 \cap \dots \cap A_6| \end{aligned}$$

Fortunately, these allocations are nested in the sense that all the allocations where e.g. 5 children do not receive gifts are a subset of all the allocations where 4 children do not receive gifts; that is

$$A_6 \subset A_5 \subset A_4 \subset A_3 \subset A_2 \subset A_1$$

which implies e.g.

$$A_1 \cap A_2 \cap A_3 \cap \dots \cap A_6 = A_6,$$

$$\sum_{1 \leq i < j \leq 6} |A_i \cap A_j| = 5|A_6| + 4|A_5| + 3|A_4| + 2|A_3| + |A_2|$$

So we have

$$\begin{aligned} \left| \bigcup_{i=1}^6 A_i \right| &= |A_6| + |A_5| + |A_4| + |A_3| + |A_2| + |A_1| - (5|A_6| + 4|A_5| + 3|A_4| + 2|A_3| + |A_2|) \\ &\quad + (4|A_6| + 3|A_5| + 2|A_4| + |A_3|) - (3|A_6| + 2|A_5| + |A_4|) + \dots - |A_6| \end{aligned}$$

$$= |A_1| - |A_2| + |A_3| - |A_4| + |A_5| - |A_6|$$

$$\begin{aligned} &= \binom{7}{1} \cdot 6^{20} - \binom{7}{2} \cdot (7-2)^{20} + \binom{7}{3} \cdot (7-3)^{20} - \binom{7}{4} \cdot (7-4)^{20} \\ &\quad + \binom{7}{5} \cdot (7-5)^{20} - \binom{7}{6} \cdot (7-6)^{20} \end{aligned}$$

The final answer is

$$\begin{aligned} 7^{20} - \left| \bigcup_{i=1}^6 A_i \right| &= 7^{20} - \binom{7}{1} \cdot 6^{20} + \binom{7}{2} \cdot (7-2)^{20} - \binom{7}{3} \cdot (7-3)^{20} + \binom{7}{4} \cdot (7-4)^{20} \\ &\quad - \binom{7}{5} \cdot (7-5)^{20} + \binom{7}{6} \cdot (7-6)^{20} \approx \boxed{5.616 \cdot 10^{16}} \end{aligned}$$

- (b) Similar to above, but more complicated. The complement of every child receiving at least two gifts is that at least one child doesn't receive a gift (same as above) or at least one child only receives one gift. So we start from the baseline answer above, and subtract out all the possible allocations in which at least one child receives one gift.

If one child only receives one gift (and the rest receive more than one), there are  $\binom{7}{1}$  ways to choose which child that is,  $\binom{20}{1}$  ways to choose which gift that child receives, and  $6^{20-1}$  allocations of the remaining gifts. If two children receive only one gift, there are  $\binom{7}{2}$  ways to choose which children those are,  $\binom{20}{2} \cdot 2!$  ways to choose which gifts those children get and distribute them among those children, and  $(7-2)^{20-2}$  ways to allocate the remaining gifts. Likewise, if three children receive only one gift there are  $\binom{7}{3} \binom{20}{3} \cdot 3! \cdot (7-3)^{20-3}$  ways to allocate the gifts,  $\binom{7}{4} \binom{20}{4} \cdot 4! \cdot (7-4)^{20-4}$  ways if four children receive only one gift,  $\binom{7}{5} \binom{20}{5} \cdot 5! \cdot (7-5)^{20-5}$  ways if five children receive only one gift, and  $\binom{7}{6} \binom{20}{6} \cdot 6! \cdot (7-6)^{20-6}$  ways if six children don't receive gifts.

Let  $B_j$  be the event that  $j$  children receive only one gift. Note that  $B_1 \cap A_i$  is nonempty  $\forall i < 7-1$ ,  $B_2 \cap A_i$  is nonempty  $\forall i < 7-2$ , and in general,  $B_j \cap A_i$  is nonempty  $\forall i < 7-j$ ,  $j \in \{1, 2, \dots, 6\}$ . Applying the Inclusion-Exclusion Principle (Proposition 6.1.13) in a similar way as in part (I)(a), the answer is

$$\boxed{7^{20} - \left| \bigcup_{i=1}^6 A_i \right| - \left| \bigcup_{j=1}^6 B_j \right| + \sum_{i \in \{1, \dots, 6\}, j \in \{1, \dots, 6\}} \left| A_i \cap B_j \right|}$$

Per part (I)(a), the first two terms approximately equal  $5.616 \cdot 10^{16}$ . Clearly

$$\bigcup_{i \in \{1, \dots, 6\}, j \in \{1, \dots, 6\}} \left( A_i \cap B_j \right) \subset \bigcup_{j=1}^6 B_j$$

which implies

$$-\left| \bigcup_{j=1}^6 B_j \right| + \left| A_i \cap \bigcap_{i \in \{1, \dots, 6\}, j \in \{1, \dots, 6\}} B_j \right| < 0$$

so the answer to this part will be less than  $5.616 \cdot 10^{16}$ , which makes sense.

Calculating  $\left| \bigcup_{j=1}^6 B_j \right|$  is not too difficult using Inclusion-Exclusion:

$$\left| \bigcup_{j=1}^6 B_j \right| = \sum_{j=1}^6 |B_j| - \sum_{1 \leq j < k \leq 6} |B_j \cap B_k| + \sum_{1 \leq j < k < \ell \leq 6} |B_j \cap B_k \cap B_\ell| - \dots$$

$$+(-1)^{6-1} |B_1 \cap B_2 \cap B_3 \cap \dots \cap B_6|$$

where since

$$B_6 \subset B_5 \subset B_4 \subset B_3 \subset B_2 \subset B_1$$

which implies e.g.

$$B_1 \cap B_2 \cap B_3 \cap \dots \cap B_6 = B_6,$$

$$\sum_{1 \leq j < k \leq 6} |B_j \cap B_k| = 5|B_6| + 4|B_5| + 3|B_4| + 2|B_3| + |B_2|$$

we have

$$\begin{aligned} \left| \bigcup_{j=1}^6 B_j \right| &= |B_6| + |B_5| + |B_4| + |B_3| + |B_2| + |B_1| - (5|B_6| + 4|B_5| + 3|B_4| + 2|B_3| + |B_2|) \\ &\quad + (4|B_6| + 3|B_5| + 2|B_4| + |B_3|) - (3|B_6| + 2|B_5| + |B_4|) + \dots - |B_6| \\ &= |B_1| - |B_2| + |B_3| - |B_4| + |B_5| - |B_6| \\ &= \binom{7}{1} \binom{20}{1} \cdot (7-1)^{20-1} - \binom{7}{2} \binom{20}{2} \cdot 2! \cdot (7-2)^{20-2} + \binom{7}{3} \binom{20}{3} \cdot 3! \cdot (7-3)^{20-3} - \binom{7}{4} \binom{20}{4} \cdot 4! \cdot (7-4)^{20-4} \\ &\quad + \binom{7}{5} \binom{20}{5} \cdot 5! \cdot (7-5)^{20-5} - \binom{7}{6} \binom{20}{6} \cdot 6! \cdot (7-6)^{20-6} \\ &\approx 5.846 \cdot 10^{16} \end{aligned}$$

However, calculating

$$\sum_{i \in \{1, \dots, 6\}, j \in \{1, \dots, 6\}} \left| A_i \bigcap B_j \right|$$

is very difficult because, for example,  $B_2 \cap A_3$  is nonempty but  $B_2 \not\subset A_3$  and  $A_3 \not\subset B_2$ .

(c)  $7^{20} \approx 7.979 \cdot 10^{16}$

(II) (a) By Proposition 6.1.16, there are  $\binom{19}{6} = 27,132$  ways to do this.

- (b) Similar to Problem 1 part (II)(c), if the vector  $(x_1, x_2, \dots, x_7)$  represents the number of gifts given to each child, we would like a solution such that

$$x_1 + x_2 + \dots + x_7 = 20, x_i \geq 2 \forall i$$

By Proposition 6.1.18, the number of possible allocations under these conditions, is  $\binom{20+7\cdot(1-2)-1}{7-1} = \binom{12}{6} = \boxed{924}$ .

- (c) By Proposition 6.1.17, the number of nonnegative integer-valued vectors  $(x_1, x_2, \dots, x_r)$  satisfying the equation

$$x_1 + x_2 + \dots + x_r = n$$

is equal to  $\binom{n+r-1}{r-1}$ . In distributing 20 identical gifts to 7 different children, we can imagine the vector  $(x_1, x_2, \dots, x_{10})$  represents the number of gifts given to each child (where  $x_i$  is a nonnegative integer for all  $i$ ). So we have  $n = 20$  and  $r = 7$ . Therefore the number of possible allocations is

$$\binom{20+7-1}{7-1} = \boxed{165,765,600}$$

- (III) Generalization of 1(I): If there are  $g$  distinguishable gifts and  $c \geq g$  children, the number of distinct allocations if each child can receive

(a) at most one gift is  $\binom{c}{g}g!$ .

(b) at most two gifts is

$$\sum_{i=c-g+1}^g \binom{c}{i} \cdot \binom{i}{g-i} \cdot \frac{g!}{(2!)^{g-i}}$$

(c) any number of gifts is  $c^g$ .

Generalization of 1(II): If there are  $g$  identical gifts and  $c \geq g$  children, the number of distinct allocations if each child can receive

(a) at most one gift is  $\binom{c}{g}$ .

(b) at most two gifts is

$$\sum_{i=c-g+1}^g \binom{c}{i} \cdot \binom{i}{g-i}$$

(c) any number of gifts is  $\binom{g+c-1}{c-1}$

Generalization of 2(I): If there are  $g$  distinguishable gifts and  $c \leq g$  children, the number of distinct allocations if each child must receive

(a) at least one gift is

$$c^g - \sum_{i=1}^{c-1} (-1)^{i+1} \binom{c}{i} \cdot (c-i)^g$$

(b) at least two gifts is

$$c^g - \sum_{i=1}^{c-1} (-1)^{i+1} \binom{c}{i} \cdot (c-i)^g - \sum_{i=1}^{c-1} (-1)^{i+1} \binom{c}{i} \binom{g}{i} \cdot i! \cdot (c-i)^{g-i}$$

(c) any number of gifts is  $c^g$

Generalization of 2(II): If there are  $g$  identical gifts and  $c \leq g$  children, the number of distinct allocations if each child must receive

(a) at least one gift is

$$\binom{g-1}{c-1}$$

(b) at least two gifts is

$$\binom{g-c-1}{c-1}$$

(c) any number of gifts is

$$\binom{g+c-1}{c-1}$$

**Homework 1 Problem 4.** You have \$20K to invest, and have a choice of stocks, bonds, mutual funds, or a CD. Investments must be made in multiples of \$1K, and there are minimal amounts to be invested: \$2K in stocks, \$2K in bonds, \$3K in mutual funds, and \$4K in the CD. Count the number of choices in each situation: (a) You want to invest in all four, (b) you want to invest in at least three out of four.

**Solution.**

(a) If the vector  $(x_S, x_B, x_{MF}, x_{CD})$  represents the amount of money (in thousands of dollars) invested in each instrument, we would like a solution such that

$$x_S + x_B + x_{MF} + x_{CD} = 20$$

where

$$x_S \geq 2, x_B \geq 2, x_{MF} \geq 3, x_{CD} \geq 4$$

In a way similar to the proof for Proposition 6.1.18, note that we can transform this problem in the following way:

$$x_S - 1 + x_B - 1 + x_{MF} - 2 + x_{CD} - 3 = 20 - (1 + 1 + 2 + 3)$$

where

$$x_S - 1 \geq 1, x_B - 1 \geq 1, x_{MF} - 2 \geq 1, x_{CD} - 3 \geq 1$$

Letting  $y_S = x_S - 1, y_B = x_B - 1, y_{MF} = x_{MF} - 2, y_{CD} = x_{CD} - 3$ , we have the equivalent system

$$y_S + y_B + y_{MF} + y_{CD} = 13, y \geq 1 \quad \forall y$$

By Proposition 6.1.16, the number of distinct solutions to this equation, and therefore the number of possible allocations under these conditions, is  $\binom{13-1}{4-1} = \boxed{220}$ .

(b) Enumerate the  $\binom{4}{3} = 4$  possibilities.

**(i) Invest in stocks, bonds, and mutual funds.**

$$x_S + x_B + x_{MF} = 202$$

where

$$x_S \geq 2, x_B \geq 2, x_{MF} \geq 3$$

Note that we can transform this problem in the following way:

$$x_S - 1 + x_B - 1 + x_{MF} - 2 = 20 - (1 + 1 + 2)$$

where

$$x_S - 1 \geq 1, x_B - 1 \geq 1, x_{MF} - 2 \geq 1$$

Letting  $y_S = x_S - 1, y_B = x_B - 1, y_{MF} = x_{MF} - 2$ , we have the equivalent system

$$y_S + y_B + y_{MF} = 16, y \geq 1 \forall y$$

Therefore the number of possible allocations under these conditions is  $\binom{16-1}{3-1} = [105]$ .

**(ii) Invest in stocks, bonds, and CDs.**

$$x_S + x_B + x_{CD} = 20$$

where

$$x_S \geq 2, x_B \geq 2, x_{CD} \geq 4$$

Note that we can transform this problem in the following way:

$$x_S - 1 + x_B - 1 + x_{CD} - 3 = 20 - (1 + 1 + 3)$$

where

$$x_S - 1 \geq 1, x_B - 1 \geq 1, x_{CD} - 3 \geq 1$$

Letting  $y_S = x_S - 1, y_B = x_B - 1, y_{CD} = x_{CD} - 3$ , we have the equivalent system

$$y_S + y_B + y_{CD} = 15, y \geq 1 \forall y$$

Therefore the number of possible allocations under these conditions is  $\binom{15-1}{3-1} = [91]$ .

**(iii) Invest in stocks, mutual funds, and CDs.**

$$x_S + x_{MF} + x_{CD} = 2$$

where

$$x_S \geq 2, x_{MF} \geq 3, x_{CD} \geq 4$$

Note that we can transform this problem in the following way:

$$x_S - 1 + x_{MF} - 2 + x_{CD} - 3 = 20 - (1 + 2 + 3)$$

where

$$x_S - 1 \geq 1, x_{MF} - 2 \geq 1, x_{CD} - 3 \geq 1$$

Letting  $y_S = x_S - 1, y_{MF} = x_{MF} - 2, y_{CD} = x_{CD} - 3$ , we have the equivalent system

$$y_S + y_{MF} + y_{CD} = 14, y \geq 1 \forall y$$

Therefore the number of possible allocations under these conditions is  $\binom{14-1}{3-1} = \boxed{78}$ .

- (iv) **Invest in bonds, mutual funds, and CDs.**

$$x_B + x_{MF} + x_{CD} = 2$$

where

$$x_B \geq 2, x_{MF} \geq 3, x_{CD} \geq 4$$

Note that we can transform this problem in the following way:

$$x_B - 1 + x_{MF} - 2 + x_{CD} - 3 = 20 - (1 + 2 + 3)$$

where

$$x_B - 1 \geq 1, x_{MF} - 2 \geq 1, x_{CD} - 3 \geq 1$$

Letting  $y_B = x_B - 1, y_{MF} = x_{MF} - 2, y_{CD} = x_{CD} - 3$ , we have the equivalent system

$$y_B + y_{MF} + y_{CD} = 14, y \geq 1 \forall y$$

therefore the number of possible allocations under these conditions is  $\binom{14-1}{3-1} = \boxed{78}$ .

- (v) **Invest in all four:** per part 4(a), there are  $\boxed{220}$  ways to do this.

Note that all of these possibilities are mutually exclusive. Therefore the total number is

$$\binom{16-1}{3-1} + \binom{15-1}{3-1} + \binom{14-1}{3-1} + \binom{14-1}{3-1} + \binom{13-1}{4-1} = 105 + 91 + 78 + 78 + 220 = \boxed{572}$$

#### 6.2.4 DSO Statistics Group Screening Exam Problems

**Exercise 2 (2017 DSO Statistics Group In-Class Screening Exam, Question 1).** Let  $X_1, X_2, \dots, X_k$  be independent standard normal random variables and  $\gamma_1(t), \dots, \gamma_k(t)$  infinitely differentiable functions of a real variable defined on a closed, bounded interval, such that  $\sum_{i=1}^k \gamma_i^2(t) = 1$  for all  $t$ . Let  $Z(t) = \sum_{i=1}^k \gamma_i(t)X_i$ . Let  $\dot{Z}(t), \ddot{Z}(t)$ , etc. denote first, second, etc. derivatives of  $Z(t)$  with respect to  $t$ .

- (a) Show that  $\text{Cov}(Z(t), \dot{Z}(t)) = 0$ .
- (b) Evaluate  $\mathbb{E}(Z(t) | \ddot{Z}(t))$  in terms of  $\ddot{Z}(t)$  and expressions of the form

$$\sum_{i=1}^k (\gamma_i(t))^a (\delta^m \gamma_i(t) / \delta t^m)^b,$$

for some  $a, b, m$  values.

**Solution.**

(a)

$$\begin{aligned}
\dot{Z}(t) &= \frac{\partial}{\partial t} \sum_{i=1}^k \gamma_i(t) X_i = \sum_{i=1}^k \dot{\gamma}_i(t) X_i \\
\implies \mathbb{E}(\dot{Z}(t)) &= \sum_{i=1}^k \mathbb{E}(\dot{\gamma}_i(t) X_i) = 0 \\
\implies \text{Cov}(Z(t), \dot{Z}(t)) &= \mathbb{E}[(Z(t) - \mathbb{E}[Z(t)])(\dot{Z}(t) - \mathbb{E}[\dot{Z}(t)])] = \mathbb{E}[Z(t)\dot{Z}(t)] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^k \gamma_i(t) X_i\right) \left(\sum_{i=1}^k \dot{\gamma}_i(t) X_i\right)\right] = \sum_{i=1}^k \mathbb{E}(\gamma_i(t) \dot{\gamma}_i(t) X_i^2) + 0 = \sum_{i=1}^k \gamma_i(t) \dot{\gamma}_i(t)
\end{aligned}$$

since  $\mathbb{E}(X_i^2) = 1$ . But

$$\sum_{i=1}^k \gamma_i \dot{\gamma}_i(t) = 0 \quad (6.8)$$

because

$$\sum_{i=1}^k \gamma_i^2(t) = 1 \iff \frac{\partial}{\partial t} \left( \sum_{i=1}^k \gamma_i^2(t) \right) = 0 \iff 2 \sum_{i=1}^k \gamma_i(t) \dot{\gamma}_i(t) = 0,$$

so the conclusion follows.

(b)

$$\begin{aligned}
Z(t) = \sum_{i=1}^k \gamma_i(t) X_i &\implies Z(t) \sim \mathcal{N}\left(0, \sum_{i=1}^k \gamma_i^2(t)\right) = \mathcal{N}(0, 1) \\
\ddot{Z}(t) = \sum_{i=1}^k \ddot{\gamma}_i(t) X_i &\implies \ddot{Z}(t) \sim \mathcal{N}\left(0, \sum_{i=1}^k \ddot{\gamma}_i^2(t)\right)
\end{aligned}$$

Also, we have

$$\begin{aligned}
\text{Cov}(Z(t), \ddot{Z}(t)) &= \mathbb{E}[(Z(t) - \mathbb{E}[Z(t)])(\ddot{Z}(t) - \mathbb{E}[\ddot{Z}(t)])] = \mathbb{E}[Z(t)\ddot{Z}(t)] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^k \gamma_i(t) X_i\right) \left(\sum_{i=1}^k \ddot{\gamma}_i(t) X_i\right)\right] = \sum_{i=1}^k \mathbb{E}(\gamma_i(t) \ddot{\gamma}_i(t) X_i^2) + 0 = \sum_{i=1}^k \gamma_i(t) \ddot{\gamma}_i(t) = -\sum_{i=1}^k \dot{\gamma}_i^2(t)
\end{aligned}$$

because differentiating (6.8) yields

$$\frac{\partial}{\partial t} \sum_{i=1}^k \gamma_i \dot{\gamma}_i(t) = 0 \iff \sum_{i=1}^k (\dot{\gamma}_i^2(t) + \gamma_i(t) \ddot{\gamma}_i(t)) = 0 \iff \sum_{i=1}^k \gamma_i(t) \ddot{\gamma}_i(t) = -\sum_{i=1}^k \dot{\gamma}_i^2(t).$$

Since both distributions are Gaussian, we have

$$\mathbb{E}(Z(t) | \ddot{Z}(t)) = \mathbb{E}[Z(t)] + \frac{\text{Cov}[Z(t), \ddot{Z}(t)]}{\text{Var}[\ddot{Z}(t)]} (\ddot{Z}(t) - \mathbb{E}[\ddot{Z}(t)]) = \left[ \sum_{i=1}^k \dot{\gamma}_i^2(t) \right]^{-1} \cdot \left[ -\sum_{i=1}^k \dot{\gamma}_i^2(t) \right] \ddot{Z}(t).$$

**Exercise 3 (2018 DSO Statistics Group In-Class Screening Exam, Question 1).** (a) For each  $x > 0$ , let  $M(x)$  be a real-valued random variable and set  $M(0) = 0$ . Assume that the random function  $M(x)$  is monotone non-decreasing on  $[0, \infty)$ . Define  $T(y) := \inf\{x \geq 0 : M(x) \geq y\}$ . Suppose that  $e^{-y}T(y)$  converges in distribution to an  $\text{Exponential}(\lambda)$  random variable when  $y \rightarrow \infty$ .

- (i) Find non-random  $a(x)$  and  $b(x) > 0$  such that  $(M(x) - a(x))/b(x)$  converges in distribution for  $x \rightarrow \infty$  to a non-degenerate random variable.
  - (ii) Provide the distribution function of the limit random variable. What is the name of this distribution?
- (b) Let  $X \sim \text{Bin}(n, p)$ . Find  $\mathbb{E}(1/(i + X))$ , for  $i = 1, 2$ . Hint: Recall that  $\int x^\alpha dx = x^{\alpha+1}/(\alpha+1) + C$  for  $\alpha \neq -1$ .

**Solution.**

- (a) (i) We have

$$T(y) = \inf\{x \geq 0 : M(x) \geq y\}; \quad \lim_{y \rightarrow \infty} \Pr(e^{-y}T(y) \leq a) = 1 - e^{-\lambda a} \quad \forall a \geq 0.$$

Note that

$$\lim_{y \rightarrow \infty} \Pr(e^{-y}T(y) \leq a) = \lim_{y \rightarrow \infty} \Pr(T(y) \leq ae^y)$$

Because  $T(y) = \inf\{x \geq 0 : M(x) \geq y\}$  and  $M(\cdot)$  is monotonically increasing, we have the inequality  $M(z) \geq y$  for all  $z \geq x$ . Therefore  $T(y) \leq ae^y \iff M(ae^y) \geq y$ . Let  $z = ae^y \implies y = \log(z/a)$ ; then we have

$$\lim_{y \rightarrow \infty} \Pr(T(y) \leq ae^y) = \lim_{y \rightarrow \infty} \Pr(M(ae^y) \geq y) = \lim_{z \rightarrow \infty} \Pr(M(z) \geq \log(z) - \log(a))$$

Let  $b = -\log a$  to get

$$\begin{aligned} &= \lim_{z \rightarrow \infty} \Pr(M(z) - \log(z) \geq b) = 1 - e^{-\lambda a} \implies \lim_{z \rightarrow \infty} \Pr(M(z) - \log(z) \geq b) = 1 - e^{-\lambda e^{-b}} \\ &\iff \lim_{z \rightarrow \infty} \Pr(M(z) - \log(z) \leq b) = e^{-\lambda e^{-b}} \end{aligned}$$

- (ii) The distribution function is  $F(x) = e^{-\lambda e^{-x}}$ , a Gumbel distribution with parameter  $\lambda$ . This is a proper cdf because  $\lim_{x \rightarrow \infty} F(x) = 1$  and  $\lim_{x \rightarrow -\infty} F(x) = 0$ .

- (b) Using hint:

$$\begin{aligned} &\int_0^1 t^x dt = \left[ \frac{t^{x+1}}{x+1} \right]_0^1 = \frac{1}{x+1} \\ &\iff \mathbb{E} \left[ \int_0^1 t^X dt \right] = \mathbb{E} \left( \frac{1}{X+1} \right) \iff \int_0^1 \mathbb{E}(t^X) dt = \mathbb{E} \left( \frac{1}{X+1} \right) \\ &\iff \int_0^1 \sum_{i=0}^n t^i \Pr(X=i) dt = \mathbb{E} \left( \frac{1}{X+1} \right) \iff \int_0^1 \sum_{i=0}^n t^i \binom{n}{i} p^i (1-p)^{n-i} dt = \mathbb{E} \left( \frac{1}{X+1} \right) \end{aligned}$$

$$\begin{aligned}
&\iff (1-p)^n \int_0^1 \sum_{i=0}^n \binom{n}{i} \left(\frac{tp}{1-p}\right)^i dt = \mathbb{E}\left(\frac{1}{X+1}\right) \\
&\iff (1-p)^n \int_0^1 \left(1 + \frac{tp}{1-p}\right)^n dt = \mathbb{E}\left(\frac{1}{X+1}\right) \\
&\iff \int_0^1 (1-p+tp)^n dt = \mathbb{E}\left(\frac{1}{X+1}\right)
\end{aligned}$$

Let  $u = 1 - p + tp \implies du = p dt$ . Then we can write

$$\frac{1}{p} \int_{1-p}^1 u^n du = \mathbb{E}\left(\frac{1}{X+1}\right) \iff \frac{1}{p} \left[ \frac{u^{n+1}}{n+1} \right]_{1-p}^1 = \mathbb{E}\left(\frac{1}{X+1}\right) \iff \mathbb{E}\left(\frac{1}{X+1}\right) = \frac{1}{p} \left[ \frac{1 - (1-p)^{n+1}}{n+1} \right]$$

Now we consider  $\mathbb{E}\left[\frac{1}{X+2}\right]$ .

$$\begin{aligned}
&\int_0^1 t^{x+1} dt = \left[ \frac{t^{x+2}}{x+2} \right]_0^1 = \frac{1}{x+2} \\
&\iff \mathbb{E}\left[\int_0^1 t^{X+1} dt\right] = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \int_0^1 \mathbb{E}(t^{X+1}) dt = \mathbb{E}\left(\frac{1}{X+2}\right) \\
&\iff \int_0^1 \sum_{i=0}^n t^{i+1} \Pr(X=i) dt = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \int_0^1 \sum_{i=0}^n t^{i+1} \binom{n}{i} p^i (1-p)^{n-i} dt = \mathbb{E}\left(\frac{1}{X+2}\right) \\
&\iff (1-p)^n \int_0^1 t \sum_{i=0}^n \binom{n}{i} \left(\frac{tp}{1-p}\right)^i dt = \mathbb{E}\left(\frac{1}{X+2}\right) \\
&\iff (1-p)^n \int_0^1 t \left(1 + \frac{tp}{1-p}\right)^n dt = \mathbb{E}\left(\frac{1}{X+2}\right) \\
&\iff \int_0^1 t (1-p+tp)^n dt = \mathbb{E}\left(\frac{1}{X+2}\right)
\end{aligned}$$

Let  $u = 1 - p + tp \implies du = p dt$  and  $t = (u + p - 1)/p$ . Then we can write

$$\begin{aligned}
&\frac{1}{p^2} \int_{1-p}^1 u^n (u+p-1) du = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \frac{1}{p^2} \int_{1-p}^1 [u^{n+1} + (p-1)u^n] du = \mathbb{E}\left(\frac{1}{X+2}\right) \\
&\iff \frac{1}{p^2} \left[ \frac{u^{n+2}}{n+2} + (p-1) \frac{u^{n+1}}{n+1} \right]_{1-p}^1 = \mathbb{E}\left(\frac{1}{X+2}\right) \\
&\iff \mathbb{E}\left(\frac{1}{X+2}\right) = \frac{1}{p^2} \left[ \frac{1 - (1-p)^{n+2}}{n+2} - (1-p) \frac{1 - (1-p)^{n+1}}{n+1} \right]
\end{aligned}$$

### 6.3 To Know for Math 505A Midterm 2

#### 6.3.1 Definitions

**Definition 6.28.** A random variable  $X$  is **continuous** if its distribution function  $F(x) = \Pr(X \leq x)$  can be written as

$$F(x) = \int_{-\infty}^x f(u)du$$

for some integrable  $f : \mathbb{R} \rightarrow [0, \infty)$ .

**Definition 6.29.** The function  $f$  is called the **(probability) density function** of the continuous random variable  $X$ .

**Proposition 6.3.1.** If  $X$  has pdf  $f_X(x)$ , then for  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ ,

$$h(x) = \frac{1}{\sigma} f_X\left(\frac{x - \mu}{\sigma}\right)$$

is a pdf. In this setting  $\mu$  is sometimes called a “location parameter” and  $\sigma$  is called a “scale parameter.”

**Definition 6.30.** The **joint distribution function** of  $X$  and  $Y$  is the function  $F : \mathbb{R}^2 \rightarrow [0, 1]$  given by

$$F(x, y) = \Pr(X \leq x \cap Y \leq y)$$

**Definition 6.31.** The random variables  $X$  and  $Y$  are **jointly continuous** with **joint (probability) density function**  $f : \mathbb{R}^2 \rightarrow [0, \infty)$  if

$$F(x, y) = \int_{v=-\infty}^y \int_{u=-\infty}^x f(u, v)dudv \text{ for each } x, y \in \mathbb{R}$$

**Definition 6.32.** Two continuous random variables are **independent** if and only if  $\{X \leq x\}$  and  $\{Y \leq y\}$  are independent events for all  $x, y \in \mathbb{R}$ .

Ways to show independence:

- Use Definition 6.32: show that  $\Pr(X \leq x \cap Y \leq y) = \Pr(X \leq x)\Pr(Y \leq y)$  for all  $x, y \in \mathbb{R}$ .

•

**Theorem 6.3.2.** The random variables  $X$  and  $Y$  are independent if and only if  $F(x, y) = F_X(x)F_Y(y)$  for all  $x, y \in \mathbb{R}$ .

•

**Proposition 6.3.3.** For continuous random variables, the previous condition is equivalent to requiring  $f(x, y) = f_X(x)f_Y(y)$ .

•

**Theorem 6.3.4.** If two variables are bivariate normal, they are independent if and only if their covariance

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dxdy$$

is equal to 0.

**Theorem 6.3.5. (Change of variables.)** Let  $X_1, Y_1$  be random variables with joint PDF  $f_{X_1, Y_1}$ . Let  $X_2, Y_2$  be random variables with joint PDF  $f_{X_2, Y_2}$ . Let  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  and let  $S : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  so that  $ST(x, y) = (x, y)$  and  $TS(x, y) = (x, y)$  for every  $(x, y) \in \mathbb{R}^2$ . Let  $J(x, y)$  denote the determinant of the Jacobian of  $S$  at  $(x, y)$ . Then

$$f_{X_2, Y_2}(x, y) = f_{X_1, Y_1}(S(x, y)) |J(x, y)|.$$

*Proof.* If the transformation from  $X_1, Y_1$  to  $X_2, Y_2$  is given by  $S(X_1, Y_1) = (X_2, Y_2)$ , then the change of variables formula from calculus is as follows:

$$\int \int_A f_{X_1, Y_1}(x, y)dxdy = \int \int_B f_{X_1, Y_1}(S(x, y)) |J(x, y)| dxdy$$

where  $A \subseteq \text{domain}(f_{X_1, Y_1}(\cdot))$ ,  $B$  is the transformation of the region  $A$  under  $S$ , and  $|J(x, y)|$  is the Jacobian of  $S$  at  $(x, y)$ . It follows from the definition of joint pdfs that the integrand on the right is the joint pdf of  $(X_2, Y_2)$ ; that is,

$$f_{X_2, Y_2}(x, y) = f_{X_1, Y_1}(S(x, y)) |J(x, y)|.$$

□

- Characteristic functions:

**Theorem 6.3.6.**  $X$  and  $Y$  are independent if and only if  $\phi_{X,Y}(s, t) = \phi_X(s)\phi_Y(t)$ .

**Theorem 6.3.7. (Theorem 4.2.3, Grimmett and Stirzaker.)** Let  $X$  and  $Y$  be random variables, and let  $g, h : \mathbb{R} \rightarrow \mathbb{R}$ . If  $X$  and  $Y$  are independent, then so are  $g(X)$  and  $h(Y)$ .

**Definition 6.33.** The **correlation coefficient** between random variables  $X$  and  $Y$  is given by

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

**Theorem 6.3.8.** The correlation coefficient satisfies  $|\rho| \leq 1$ .

*Proof.* Apply the Cauchy-Schwarz Inequality (Theorem 8.2.4) to  $X - \mathbb{E}(X)$  and  $Y - \mathbb{E}(Y)$ :

$$\text{Cov}(X, Y)^2 = (\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]^2)^2 \leq \mathbb{E}[(X - \mathbb{E}(X))^2]\mathbb{E}[(Y - \mathbb{E}(Y))^2] = \text{Var}(X)\text{Var}(Y)$$

$$\iff \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)\text{Var}(Y)} \leq 1 \iff \rho^2 \leq 1 \iff |\rho| \leq 1$$

□

**Theorem 6.3.9. (Stein identity.)** Let  $X$  be a standard Gaussian random variable, so that  $X$  has density  $x \rightarrow e^{-x^2/2}/\sqrt{2\pi}$ ,  $\forall x \in \mathbb{R}$ . Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a continuously differentiable function such that  $g$  and  $g'$  have polynomial volume growth. That is,  $\exists a, b > 0$  such that  $|g(x)|, |g'(x)| \leq a(1 + |x|)^b$ ,  $\forall x \in \mathbb{R}$ . Then

$$\mathbb{E}Xg(X) = \mathbb{E}g'(X).$$

*Proof.* Examining the left side, we have

$$\mathbb{E}(Xg(X)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xg(x)e^{-x^2/2} dx$$

We will use integration by parts with  $u = g(x) \implies du = g'(x)dx$ ,  $dv = xe^{-x^2/2}dx \implies v = -e^{-x^2/2}$  to yield the result:

$$\mathbb{E}(Xg(X)) = -g(x)e^{-x^2/2} + \int_{-\infty}^{\infty} g'(x)e^{-x^2/2} dx = \mathbb{E}(g'(X)).$$

□

**Example.** Using Theorem 6.3.9, recursively compute  $\mathbb{E}X^k$  for any positive integer  $k$ . Alternatively, for any  $t > 0$ , show that  $\mathbb{E}e^{tX} = e^{t^2/2}$ , i.e. compute the **moment generating function** of  $X$ . Then, using  $\frac{d^k}{dt^k}|_{t=0} \mathbb{E}e^{tX} = \mathbb{E}X^k$  and using the power series expansion of the exponential, compute  $\mathbb{E}X^k$  directly from the identity  $\mathbb{E}e^{tX} = e^{t^2/2}$ .

**Solution.** We can use this result to recursively calculate  $\mathbb{E}(X^k)$  for any positive integer  $k$ . Suppose we have  $\mathbb{E}(X^{k-1})$ . Letting  $g(X) = X^{k-1} \implies g'(X) = (k-1)X^{k-2}$ , we have

$$\mathbb{E}(X^k) = \mathbb{E}(X \cdot X^{k-1}) = \mathbb{E}(Xg(X)) = \mathbb{E}(g'(X)) \iff \boxed{\mathbb{E}(X^k) = (k-1)\mathbb{E}(X^{k-2})}.$$

Since  $X \sim \mathcal{N}(0, 1)$ , we have  $\mathbb{E}(X) = 0$ ,  $\mathbb{E}(X^2) = \text{Var}(X) + \mathbb{E}(X)^2 = 1 + 0 = 1$ . Therefore we have

$$\mathbb{E}(X^k) = \begin{cases} \prod_{i=1}^{(k-1)/2} (k - (2i-1)) \mathbb{E}(X) & k \text{ is odd} \\ \prod_{i=1}^{k/2} (k - (2i-1)) \mathbb{E}(X^2) & k \text{ is even} \end{cases}$$

$$\boxed{\mathbb{E}(X^k) = \begin{cases} 0 & k \text{ is odd} \\ \prod_{i=1}^{k/2} (k - (2i-1)) & k \text{ is even} \end{cases}}$$

### 6.3.2 Probability-Generating Functions

**Definition 6.34.**

$$G_X(s) = \mathbb{E}(s^X)$$

**Theorem 6.3.10. Some useful properties:**

- (a)  $\mathbb{E}(X) = G'_X(1)$ ,  $\mathbb{E}[X(X-1)\cdots(X-k+1)] = G^{(k)}(1)$
- (b) If  $X$  and  $Y$  are independent then  $G_{X+Y}(s) = G_X(s)G_Y(s)$ .

### 6.3.3 Moment-Generating Functions

**Definition 6.35.**

$$M_X(t) = \mathbb{E}(e^{tX})$$

**Theorem 6.3.11 (Some useful properties).** (a)  $\mathbb{E}(X) = M'_X(0)$ ,  $\mathbb{E}(X^k) = M^{(k)}(0)$

(b) If  $X_1, X_2, \dots, X_n$  are independent then  $M_{X_1+X_2+\dots+X_n}(t) = \prod_{i=1}^n M_{X_i}(t)$ .

*Proof.* (a)

(b)

$$M_{X_1+X_2+\dots+X_n}(t) = \mathbb{E} \exp \left( t \sum_{i=1}^n X_i \right) = \mathbb{E} \prod_{i=1}^n e^{tX_i} = \prod_{i=1}^n \mathbb{E} e^{tX_i} = \prod_{i=1}^n M_{X_i}(t).$$

□

### 6.3.4 Characteristic Functions

**Definition 6.36.**

$$\phi_X(t) = \mathbb{E}(e^{itX})$$

**Proposition 6.3.12.** Necessary and sufficient conditions for a function to be a characteristic function:

- (a)  $\phi_X(0) = 1$
- (b)  $|\phi(t)| \leq 1 \forall t$
- (c)  $\phi$  is uniformly continuous on  $\mathbb{R}$
- (d)  $\phi$  is positive semidefinite; that is,

$$\sum_{i,j} \phi(t_j - t_k) z_j \bar{z}_k \geq 0 \text{ for all real } t_1, t_2, \dots, t_n \text{ and complex } z_1, z_2, \dots, z_n$$

Or, equivalently, or every set of real numbers  $t_1, t_2, \dots, t_n$ , the matrix  $\phi(t_i - t_j), i, j \in \{1, 2, \dots, n\}$  is Hermitian and nonnegative definite.

**Remark.** Relationship between characteristic functions and probability and moment-generating functions:

$$\phi_X(t) = M_X(it) = G_X(e^{it})$$

**Theorem 6.3.13.** Some useful properties:

- (a)  $X \perp\!\!\!\perp Y \implies \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$
- (b)  $Y = aX + b \implies \phi_Y(t) = e^{itb}\phi_X(at)$
- (c)  $\phi_X^{(k)}(0) = i^k \mathbb{E}(X^k)$
- (d)  $\phi_{X,Y}(s,t) = \mathbb{E}(e^{isX}e^{itY})$
- (e)  $X \perp\!\!\!\perp Y \iff \phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$

**Theorem 6.3.14.** Other facts from notes on course website

- (a) If  $\phi(t)$  is even,  $\phi(0) = 1$ ,  $\phi$  is convex for  $t > 0$ , and  $\lim_{t \rightarrow \infty} \phi(t) = 0$ , then  $\phi$  is a characteristic function of an absolutely continuous random variable.
- (b) If  $\phi$  is a characteristic function and  $\phi(t) = 1 + o(t^2), t \rightarrow 0$ , then  $\phi(t) = 1$  for all  $t$ . The random variable with such a characteristic function must have zero mean and zero variance. In particular, if  $r > 2$ , then  $\exp(-|t|^r)$  is not a characteristic function.
- (c) If  $\phi(t) = e^{p(t)}$  is a characteristic function and  $p = p(t)$  is a polynomial, then the degree of  $p$  is at most 2. For example,  $e^{t^2-t^4}$  is not a characteristic function.
- (d) If  $\xi$  is absolutely continuous, then  $\lim_{|t| \rightarrow \infty} |\phi_\xi(t)| = 0$  (Riemann-Lebesgue).
- (e) If  $\int_{-\infty}^{\infty} |\phi_\xi(t)| dt < \infty$ , then  $\xi$  is absolutely continuous with pdf

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \phi(t) dt$$

### 6.3.5 Continuous Random Variable Distributions

**Uniform:**  $U(a, b)$

- Probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- Cumulative distribution function:

$$F(x) = \Pr(X \leq x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x \leq b \\ 1 & x > b \end{cases}$$

- Probability-generating function:
- Moment-generating function:

$$M_X(t) = \frac{1}{(b-a)t} [\exp(bt) - \exp(at)]$$

*Proof.*

$$M_X(t) = \mathbb{E}(\exp(tx)) = \int_a^b \frac{1}{b-a} \cdot \exp(tx) dx = \frac{1}{b-a} \left[ \frac{1}{t} \exp(tx) \right]_a^b = \frac{1}{(b-a)t} [\exp(bt) - \exp(at)]$$

□

- Characteristic function:

$$\frac{2}{(b-a)t} \sin\left(\frac{1}{2}(b-a)t\right) \exp\left(i(a+b)\frac{t}{2}\right)$$

- Expectation:  $\mathbb{E}(X) = (b-a)/2$
- Variance:  $\text{Var}(X) = (b-a)^2/12$

**Proposition 6.3.15.** If  $X \sim U(0, 1)$ , then  $Y = -\log(X) \sim \text{Exponential}(1)$ .

*Proof.*

$$\begin{aligned} \Pr(Y \leq y) &= \Pr(-\log(X) \leq y) = \Pr(\log(X) \geq -y) = \Pr(X \geq e^{-y}) = \int_{\exp(-y)}^{\infty} f_X(t) dt = \int_{\exp(-y)}^1 dt \\ &= 1 - e^{-y} \end{aligned}$$

which is the cdf for an exponential distribution with mean 1. □

**Normal (or Gaussian):**  $\mathcal{N}(\mu, \sigma^2)$

- Probability density function:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Cumulative distribution function:  $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:  $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$
- Characteristic function:  $\phi(t) = \exp(i\mu t - (1/2)\sigma^2 t^2)$ . Standard normal:  $\phi(t) = \exp((-1/2)t^2)$ .
- Expectation:  $\mathbb{E}(X) = \mu$
- Variance:  $\text{Var}(X) = \sigma^2$

**Theorem 6.3.16.** Let  $X := \Omega \rightarrow \mathbb{R}^n$  be a random Gaussian variable with the **standard Gaussian distribution**:

$$\Pr(X \in A) := \int_A e^{-(x_1^2 + \dots + x_n^2)/2} dx (2\pi)^{-n/2}, \quad \forall A \subset \mathbb{R}^n \text{ measurable.}$$

Let  $v_1, \dots, v_m$  be vectors in  $\mathbb{R}^n$ . Let  $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  be the standard inner product on  $\mathbb{R}^n$ , so that  $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$  for any  $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n) \in \mathbb{R}^n$ .

Let  $v \in \mathbb{R}^n$ . Then  $\langle X, v \rangle$  is a mean zero Gaussian with variance  $\langle v, v \rangle$ .

**Remark.** Similar to Exercise 6.2. You can then use an induction argument to prove the case for the sum of arbitrarily many Gaussian random variables.

*Proof.* Let  $X = (X_1, X_2, \dots, X_n)$  and  $v = (v_1, v_2, \dots, v_n)$ . Then  $\langle X, v \rangle = \sum_{i=1}^n X_i v_i$ . Let  $Y_i = v_i X_i$ , so  $\langle X, v \rangle = \sum_{i=1}^n Y_i$ . Note that  $Y_i \sim \mathcal{N}(0, v_i^2)$ . Recall that the moment-generating function of a Gaussian random variable is  $\mathbb{E}e^{tX} = e^{\mu t + \sigma^2 t^2/2}$ , so we have  $M_{Y_i}(t) = \exp(v_i^2 t^2/2)$ . Next we seek the distribution of  $\sum_{i=1}^n v_i X_i = \sum_{i=1}^n Y_i$ . From Proposition 1.67, we have

$$M_{Y_1+Y_2+\dots+Y_n}(t) = \mathbb{E} \exp \left( t \sum_{i=1}^n Y_i \right) = \mathbb{E} \prod_{i=1}^n e^{t Y_i} = \prod_{i=1}^n \mathbb{E} e^{t Y_i} = \prod_{i=1}^n M_{Y_i}(t).$$

Then

$$\prod_{i=1}^n M_{Y_i}(t) = \prod_{i=1}^n \exp(v_i^2 t^2/2) = \exp \left( \frac{t^2}{2} \sum_{i=1}^n v_i^2 \right) = \exp \left( \frac{t^2}{2} \cdot \langle v, v \rangle \right)$$

which is the same as the moment-generating function for a mean zero Gaussian random variable with variance  $\langle v, v \rangle$ . By the provided uniqueness result from Problem 6 (“If  $Y$  and  $Z$  are two random variables whose MGFs coincide in a neighborhood of 0 ( $\exists \delta > 0$  for which  $M_Y(u) = M_Z(u) < \infty$  for all  $u \in [-\delta, \delta]$ ), then  $Y$  and  $Z$  have the same distribution.”), the result follows.  $\square$

**Proposition 6.3.17.** Let  $X_1, X_2, \dots, X_n$  be i.i.d. random sample from  $\mathcal{N}(\mu, \sigma)$ . Then the sum of these  $n$  observations  $T = \sum_{i=1}^n X_i$  also follows the normal distribution.

*Proof.*

$$T = \sum_{i=1}^n X_i = X_1 + X_2 + \dots + X_n$$

Since  $X_i$  is  $\mathcal{N}(\mu, \sigma)$ , we have

$$M_{X_i}(t) = \exp \left( \mu t + \frac{t^2 \sigma^2}{2} \right)$$

Since all observations are independent, we have

$$M_T(t) = \prod_{i=1}^n M_{X_i}(t) = \left( \exp \left( \mu t + \frac{t^2 \sigma^2}{2} \right) \right)^n = \boxed{\exp \left( n \mu t + \frac{t^2 n \sigma^2}{2} \right)}$$

which is the moment generating function for a normal distribution with mean  $n\mu$  and standard deviation  $\sigma\sqrt{n}$ .  $\square$

**Lognormal:** If the random variable  $X$  follows the normal distribution with  $\mu = 0$ ,  $\sigma^2 = 1$  and  $Y = e^X$ , then  $Y$  has a lognormal distribution.

- Probability density function:

$$f_X(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(\frac{-1}{2}(\log(x))^2\right)$$

*Proof.*

$$F_Y(y) = \Pr(Y \leq y) = \Pr(e^X \leq y) = \Pr(X \leq \log(y)) \implies F_Y(y) = F_X(\log(y))$$

$$\implies f_Y(y) = \frac{d}{dy} F_X(\log(y)) = f_X(\log(y)) \frac{1}{y}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}x^2} \implies f_Y(y) = \frac{1}{y\sqrt{2\pi}} \exp\left(\frac{-1}{2}(\log(y))^2\right)$$

□

- Cumulative distribution function:  $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:  $M_X(t) =$
- Characteristic function:  $\phi(t) =$
- Expectation:  $\mathbb{E}(X) =$
- Variance:  $\text{Var}(X) =$

**Gamma:**  $\Gamma(\alpha, \beta)$

- Probability density function:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} = \frac{1}{\Gamma(\alpha, \beta)} x^{\alpha-1} e^{-x/\beta}$$

- Cumulative distribution function:

$$F(x) = \Pr(X \leq x) =$$

- Probability-generating function:
- Moment-generating function:

$$\left(\frac{1/\beta}{1/\beta - t}\right)^\alpha = \left(\frac{1}{1 - \beta t}\right)^\alpha$$

*Proof.*

$$\mathbb{E}(e^{tX}) = \int_{\mathbb{R}} e^{tx} f(x) dx = \int_0^\infty e^{tx} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(1/\beta-t)} dx$$

Using integration by parts, one can find the identity

$$\int_0^\infty x^a e^{-bx} dx = \frac{\Gamma(a+1)}{b^{a+1}}.$$

Using this yields

$$\mathbb{E}(e^{tX}) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \frac{\Gamma(\alpha)}{(1/\beta - t)^\alpha} = \left( \frac{1/\beta}{1/\beta - t} \right)^\alpha = \left( \frac{1}{1 - \beta t} \right)^\alpha$$

□

- Characteristic function:
- Expectation:  $\mathbb{E}(X) = \alpha\beta$
- Variance:  $\text{Var}(X) = \alpha\beta^2$

**Proposition 6.3.18.** Let  $X_i \sim \text{Gamma}(\alpha_i, \beta)$  for  $i = 1, 2, \dots, n$ . Then

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right).$$

*Proof.* Using the moment-generating function for a Gamma distribution as well as Theorem 6.3.11(b), we have

$$M_{X_1+\dots+X_n}(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n \left( \frac{1}{1 - \beta t} \right)^{\alpha_i} = \left( \frac{1}{1 - \beta t} \right)^{\sum_{i=1}^n \alpha_i}$$

which is the same as the moment-generating function for a  $\Gamma(\sum_{i=1}^n \alpha_i, \beta)$  distribution. By the provided uniqueness result (“If  $Y$  and  $Z$  are two random variables whose MGFs coincide in a neighborhood of 0 ( $\exists \delta > 0$  for which  $M_Y(u) = M_Z(u) < \infty$  for all  $u \in [-\delta, \delta]$ ), then  $Y$  and  $Z$  have the same distribution.”), the result follows. □

**Proposition 6.3.19.** Let  $X \sim \text{Gamma}(\alpha, \beta)$ . Then as  $\beta \rightarrow \infty$ ,  $X \xrightarrow{d} \mathcal{N}(\alpha\beta, \alpha\beta^2)$ .

*Proof.* See <http://www.math.wm.edu/~leemis/chart/UDR/PDFs/GammaNormal1.pdf>. □

**Proposition 6.3.20 (Stats 100B homework problem).** The method of moments estimators for  $\alpha$  and  $\beta$  are

$$\hat{\alpha} = \frac{n\bar{x}^2}{\sum_{i=1}^n (X_i^2 - \bar{x}^2)}$$

$$\hat{\beta} = \frac{1}{n\bar{x}} \sum_{i=1}^n (X_i^2 - \bar{x}^2)$$

*Proof.*

$$\mathbb{E}(X) = \alpha\beta \implies \hat{\alpha}\hat{\beta} = \bar{x}$$

$$\hat{\alpha} = \frac{\bar{x}}{\hat{\beta}}$$

$$\text{Var}(X) = \alpha\beta^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

$$\implies \hat{\alpha}\hat{\beta}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{x}^2$$

$$\frac{\bar{x}}{\hat{\beta}}\hat{\beta}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{x}^2$$

$$\bar{x}\hat{\beta} = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{x}^2$$

$$\hat{\beta} = \frac{1}{n\bar{x}} \sum_{i=1}^n X_i^2 - \bar{x} = \frac{1}{n\bar{x}} \sum_{i=1}^n X_i^2 - \frac{n\bar{x}^2}{n\bar{x}}$$

$$\hat{\beta} = \frac{1}{n\bar{x}} \sum_{i=1}^n (X_i^2 - \bar{x}^2)$$

$$\implies \hat{\alpha} = \frac{\bar{x}}{\hat{\beta}} = \frac{n\bar{x}^2}{\sum_{i=1}^n (X_i^2 - \bar{x}^2)}$$

□

**Proposition 6.3.21** (Some useful formulae and integrals related to the gamma function).

- Definition:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

- Beta function (see also the information on the Beta distribution):

$$B(z_1, z_2) := \int_0^1 t^{z_1-1} (1-t)^{z_2-1} dt = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$$

- Generalization of the factorial function:

$$\Gamma(z+1) = z\Gamma(z)$$

(in particular, for any integer  $n \geq 1$ ,  $\Gamma(n) = (n-1)!$ .)

- $\Gamma(1/2) = \sqrt{\pi}$ .
- Using integration by parts, one can find the identity

$$\int_0^\infty x^a e^{-bx} dx = \frac{\Gamma(a+1)}{b^{a+1}}.$$

- **Euler's Reflection Formula:** For  $z \notin \mathbb{Z}$ ,

$$\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}$$

- **Legendre Duplication Formula:**

$$\Gamma(z)\Gamma\left(z + \frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$$

- For all  $z \in \mathbb{C}$ ,

$$\lim_{n \rightarrow \infty} \frac{\Gamma(n+z)}{\Gamma(n)n^z} = 1.$$

- From 2018 DSO Statistics Screening Exam:

$$\lim_{x \rightarrow 0} x\Gamma(x) = 1$$

- See Theorem 6.1.43 (Stirling's Formula)

$$\Gamma(n+1) \sim n^n e^{-n} \sqrt{2\pi n}$$

That is,

$$\lim_{n \rightarrow \infty} \frac{n^n e^{-n} \sqrt{2\pi n}}{\Gamma(n+1)} = 1.$$

$\chi_k^2$  (**chi-squared**): special case of a gamma distribution:  $\Gamma(k/2, 2)$ . Also the sum of  $k$  independent standard normally distributed variables.

- Probability density function:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} = \frac{1}{\Gamma(k/2, 2)} x^{k/2-1} e^{-x/2}$$

For  $\chi_1^2$ :  $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x/2} x^{-1/2}$ . For  $\chi_2^2$ :  $f(x) = \frac{1}{2} e^{-x/2}, x > 0$ .

*Proof.* See notes from Math 541A. □

- Cumulative distribution function:  $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:  $(1 - 2t)^{-k/2}$  for  $t < 1/2$

- Characteristic function:
- Expectation:  $\mathbb{E}(X) = k/2 \cdot 2 = k$
- Variance:  $\text{Var}(X) = k/2 \cdot 2^2 = 2k$

**Exponential:** (special case of a gamma distribution:  $\Gamma(1, \beta)$ . Also a special case of a Weibull distribution with  $\beta = 1$ .)

- Probability density function:  $f(x) = \frac{1}{\beta} \exp(-x/\beta) = \lambda e^{-\lambda x}$
- Cumulative distribution function:  $F(x) = \Pr(X \leq x) = 1 - e^{-\lambda x}$
- Probability-generating function:
- Moment-generating function:  $\frac{\lambda}{\lambda-t}$
- Characteristic function:
- Expectation:  $\mathbb{E}(X) = \beta = \lambda^{-1}$

*Proof.*

$$\mathbb{E}(X) = \int_0^\infty \bar{F}(t)dt = \int_0^\infty e^{-\lambda t}dt = \frac{1}{\lambda}$$

□

- Variance:  $\text{Var}(X) = \beta^2 = \lambda^{-2}$

*Proof.* See Definition 6.21 above for the definition of  $E(X^n)$ . Then using that:

$$\mathbb{E}(X^2) = 2 \int_0^\infty te^{-\lambda t}dt = \frac{2}{\lambda} \int_0^\infty \lambda te^{-\lambda t}dt = \frac{2}{\lambda} \mathbb{E}(X) = \frac{2}{\lambda^2}$$

Then use  $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$  to yield the result. □

**Remark.** Note also the general case:

$$\mathbb{E}(X^n) = n \int_0^\infty t^{n-1} \bar{F}(t)dt = n \int_0^\infty t^{n-1} e^{-\lambda t}dt = \frac{n}{\lambda} \mathbb{E}(X^{n-1})$$

For much more on exponential random variables, see Section 6.3.6.

**Cauchy:**

- Probability density function:

$$f(x) = \frac{1}{\pi(1+x^2)} \text{ (standard Cauchy)} , f(x) = \frac{1}{\pi\sigma(1+(x-\mu)^2/\sigma^2)} \text{ (general)}$$

- Cumulative distribution function:  $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:
- Characteristic function:

- Expectation: does not exist
- Variance: does not exist (Cauchy distribution has no moments.)

**Beta:** Recall:

$$B(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

$$\implies \frac{B(\alpha + 1, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + 1)\Gamma(\beta)}{\Gamma(\alpha + 1 + \beta)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha}{\alpha + \beta}$$

- Probability density function:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} (x)^{\alpha-1} (1-x)^{\beta-1}$$

- Cumulative distribution function:  $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:

$$1 + \sum_{k=1}^{\infty} \left( \prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r} \right) \frac{t^k}{k!}$$

- Characteristic function:
- Expectation:  $\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta}$
- Variance:

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

**Student's  $t_n$ :** can be defined as

$$T = \frac{Z}{\sqrt{V/v}} \implies T \sim t_v$$

where  $Z \sim \mathcal{N}(0, 1)$ ,  $V \sim \chi_v^2$ , and  $Z \perp\!\!\!\perp V$ .

- Probability density function:

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \cdot \Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

- Cumulative distribution function:  $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:
- Characteristic function:
- Expectation:  $\mathbb{E}(X) = 0$
- Variance:  $\text{Var}(X) = n/(n - 2)$

**Snedecor's  $F$ -distribution:** can be defined as

$$X = \frac{U/d_1}{V/d_2} \implies X \sim F_{d_1, d_2}$$

where  $U \sim \chi^2_{d_1}$ ,  $V \sim \chi^2_{d_2}$ , and  $U \perp\!\!\!\perp V$ .

– Probability density function:

$$\begin{aligned} f(x) &= \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \left(1+t(p/q)\right)^{-(p+q)/2}, \quad \forall t > 0 \\ &= p^{p/2}q^{q/2} \cdot \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1}}{(pt+q)^{(p+q)/2}} \end{aligned}$$

– Cumulative distribution function:  $F(x) =$

– Probability-generating function:

– Moment-generating function:

– Characteristic function:

– Expectation:  $\mathbb{E}(X) =$

– Variance:  $\text{Var}(X) =$

**Proposition 6.3.22.** Let  $X$  be a chi squared random variables with  $p$  degrees of freedom. Let  $Y$  be a chi squared random variable with  $q$  degrees of freedom, with  $X$  and  $Y$  independent. Then  $(X/p)/(Y/q)$  is an  $F$ -distributed random variable with  $p$  and  $q$  degrees of freedom.

*Proof.* Let  $c = p/2$ ,  $d = q/2$ . Note that

$$f_X(x) = \frac{1}{\Gamma(c)2^c}x^{c-1}e^{-x/2}, \quad x > 0$$

$$f_Y(y) = \frac{1}{\Gamma(d)2^d}y^{d-1}e^{-y/2}, \quad y > 0$$

We first seek  $F_{X/Y}(t)$ . Note that  $F_{X/Y}(t) = \Pr(X/Y \leq t) = \Pr(X \leq tY)$ . Thinking about this graphically, we can calculate this as

$$\Pr(X \leq tY) = \int_0^\infty \int_0^{ty} f_{X,Y}(x,y) dx dy$$

By assumption,  $X$  and  $Y$  are independent, so

$$f_{X,Y}(x,y) = \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}}x^{c-1}y^{d-1}e^{-x/2}e^{-y/2}$$

Plugging this in we have

$$\begin{aligned} \Pr(X \leq tY) &= \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty \int_0^{ty} x^{c-1}y^{d-1}e^{-x/2}e^{-y/2} dx dy \\ &= \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty \int_0^{ty} x^{c-1}e^{-x/2} dx y^{d-1}e^{-y/2} dy \end{aligned}$$

Rather than solving this integral, we next differentiate with respect to  $t$ :

$$\begin{aligned} f_{X/Y}(t) &= \frac{d}{dt} F_{X/Y}(t) = \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty y(ty)^{c-1} e^{-ty/2} y^{d-1} e^{-y/2} dy \\ &= \frac{t^{c-1}}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty y^{c+d-1} e^{-y(t+1)/2} dy \end{aligned} \quad (6.9)$$

Compare the integrand of (6.9) to the pdf of a Gamma distributed random variable

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

to see that with some manipulations we can express the integrand of (6.9) as the pdf of a Gamma distributed random variable with parameters  $\alpha = c + d$ ,  $\beta = 2/(t + 1)$ :

$$f_{X/Y}(t) = \frac{t^{c-1} \Gamma(c+d)}{\Gamma(c)\Gamma(d)(t+1)^{c+d}} \int_0^\infty \frac{1}{[2/(t+1)]^{c+d} \Gamma(c+d)} y^{c+d-1} e^{-y(t+1)/2} dy$$

Then the integral becomes 1, so we have

$$f_{X/Y}(t) = \frac{t^{c-1} \Gamma(c+d)}{\Gamma(c)\Gamma(d)(t+1)^{c+d}}$$

Substitute back in  $c = p/2$ ,  $d = q/2$ :

$$f_{X/Y}(t) = \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1}}{(t+1)^{(p+q)/2}}$$

Now take into account the constants:

$$\begin{aligned} f_{(X/p)/(Y/q)}(t) &= \frac{p}{q} f_{X/Y}\left(\frac{p}{q}t\right) = \frac{p}{q} \cdot \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{(pt/q)^{p/2-1}}{(pt/q+1)^{(p+q)/2}} \\ &= \frac{t^{(p/2)-1} (p/q)^{p/2} \Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \left(1 + t(p/q)\right)^{-(p+q)/2} \end{aligned}$$

(or this can be expressed as)

$$= \left(\frac{p}{q}\right)^{p/2} \cdot \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1} q^{(p+q)/2}}{(pt+q)^{(p+q)/2}} = p^{p/2} q^{q/2} \cdot \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1}}{(pt+q)^{(p+q)/2}}$$

□

### Weibull:

- Probability density function:  $f(x) = \alpha \beta x^{\beta-1} \exp(-\alpha x^\beta)$
- Cumulative distribution function:  $F(x) = \Pr(X \leq x) = 1 - \exp(-\alpha x^\beta)$
- Probability-generating function:
- Moment-generating function:
- Characteristic function:
- Expectation:  $\mathbb{E}(X) =$

- Variance:  $\text{Var}(X) =$

**Pareto:** (parameters  $\alpha, x_m$ )

- Probability density function:  $f(x) = \alpha x_m^\alpha / x^{\alpha+1}$  for  $x \geq 1, 0$  otherwise.
- Cumulative distribution function:  $F(x) = 1 - (x_m/x)^\alpha$  for  $x \geq 1, 0$  otherwise.
- Probability-generating function:
- Moment-generating function:
- Characteristic function:
- Expectation:  $\mathbb{E}(X) = \alpha x_m / (\alpha - 1)$  for  $\alpha > 1, \infty$  otherwise.
- Variance:  $\text{Var}(X) = \frac{x_m^2 \alpha}{(\alpha-1)^2(\alpha-2)}$  for  $\alpha > 2, \infty$  otherwise.

### 6.3.6 More on Exponential Random Variables

**Remark.** Recall Proposition 6.3.15: If  $X \sim U(0, 1)$ , then  $Y = -\log(X) \sim \text{Exponential}(1)$ .

**Proposition 6.3.23.** Let  $X$  be a random variable. Then  $X$  is exponentially distributed if and only if  $X$  has the **memoryless** property; that is,

$$\Pr(X > s + t \mid X > t) = \Pr(X > s) \quad \forall s, t \geq 0$$

or, equivalently,

$$\Pr(X > s + t) = \Pr(X > s) \Pr(X > t) \quad \forall s, t \geq 0$$

*Proof.* See Ross *Introduction to Probability Models* section 5.2.2. □

**Remark.** Exponential distributions can be derived as follows: Let  $X$  be a nonnegative random variable with cdf  $F$  and pdf  $f$ . Define the **failure** or **hazard rate**

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

The intuition for the hazard rate is as follows: think of  $X$  as the lifetime and consider the probability that a unit with age  $t$  fails within some timespan  $(t, t+h)$  with  $h$  small; that is,  $\Pr(t < X < t+h \mid X > t)$ . Letting  $1 - F(t) = \bar{F}(t)$ , we have

$$\Pr(t < X < t+h \mid X > t) = \frac{\Pr(t < X < t+h)}{\Pr(X > t)} = \frac{\int_t^{t+h} f(s)ds}{\bar{F}(t)} \approx \frac{f(t)}{\bar{F}(t)} \text{ for small } h$$

If this quantity is constant at  $\lambda$  (i.e. the process is “memoryless,” see Proposition 6.3.23), we have an exponential distribution:

$$\lambda = \frac{f(t)}{\bar{F}(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}$$

Indeed, a process is memoryless if and only if it is exponential. To see this, note that

$$\int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1 - F(s)}ds$$

Letting  $u = 1 - F(s) \implies du = -f(s)ds$ , we have

$$\begin{aligned} &= \int_{u=1}^{u=\bar{F}(t)} -\frac{du}{u} = \int_{\bar{F}(t)}^1 \frac{du}{u} = \log(1) - \log(\bar{F}(t)) = -\log \bar{F}(t) \\ &\implies \int_0^t \lambda(s)ds = -\log \bar{F}(t) \implies \boxed{\bar{F}(t) = \exp\left(-\int_0^t \lambda(s)ds\right)} \end{aligned}$$

If the process is memoryless, we must have  $\lambda(s) = \lambda$  (constant). Then  $\bar{F}(t) = \exp(-\lambda t)$  and we have the exponential distribution. See the Weibull distribution below for a more general case.

**Proposition 6.3.24.** Let  $X$  be an exponential random variable. Then

- (a)  $\mathbb{E}(X - s \mid X > s) = \mathbb{E}(X)$
- (b)  $\mathbb{E}(X - s) = \Pr(X > s)\mathbb{E}(X)$  and

*Proof.* (a) By the memoryless property (Proposition 6.3.23),  $\Pr(X > t + s \mid X > s) = \Pr(X - s > t \mid X > s) = \Pr(X > t)$ . Then we have

$$\mathbb{E}(X - s \mid X > s) = \int_0^\infty \Pr(X - s > t \mid X > s)dt = \int_0^\infty \Pr(X > t)dt = \mathbb{E}(X)$$

- (b) By the memoryless property (Proposition 6.3.23),  $\Pr(X > t + s) = \Pr(X - s > t) = \Pr(X > t)\Pr(X > s)$ . Then we have

$$\mathbb{E}(X - s) = \int_0^\infty \Pr(X > t)\Pr(X > s)dt = \Pr(X > s) \int_0^\infty \Pr(X > t)dt = \Pr(X > s)\mathbb{E}(X)$$

□

**Proposition 6.3.25.** Let  $X$  be an exponential random variable. Then  $X - t \mid X > t$  is identically distributed as  $X$ .

*Proof.* Because  $X$  is memoryless, we know by Proposition 6.3.23 that  $\Pr(X > s + t \mid X > t) = \Pr(X > s)$ , which is to say  $\Pr(X \leq s + t \mid X > t) = \Pr(X \leq s) \iff \Pr(X - t \leq s \mid X > t) = \Pr(X \leq s)$ ; that is,  $X - t \mid X > t$  and  $X$  have identical distributions.

□

**Proposition 6.3.26.** Let  $X$  be an exponential random variable. Then  $\mathbb{E}[X^2 \mid X > t] = \mathbb{E}[(X + t)^2]$ .

*Proof.* By Proposition 6.3.25,  $X$  and  $X - t \mid X > t$  have identical distributions. That means they have identical variances, so

$$\text{Var}(X - t \mid X > t) = \text{Var}(X) \iff \mathbb{E}[(X - t)^2 \mid X > t] - [\mathbb{E}(X - t \mid X > t)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$

Using  $\mathbb{E}(X - t \mid X > t) = \mathbb{E}(X)$  we have

$$\mathbb{E}[(X-t)^2 | X > t] - [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \iff \mathbb{E}[(X-t)^2 | X > t] = \mathbb{E}(X^2)$$

$$\mathbb{E}[(X-t)^2 | X > t] - [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \iff \mathbb{E}[X^2 - 2tX + t^2 | X > t] = \mathbb{E}(X^2)$$

$$\iff \mathbb{E}[X^2 | X > t] - 2t\mathbb{E}(X | X > t) + t^2 = \mathbb{E}(X^2) \iff \mathbb{E}[X^2 | X > t] = \mathbb{E}(X^2) + 2t\mathbb{E}(X | X > t) - t^2$$

Using  $\mathbb{E}(X | X > t) = t + \mathbb{E}(X)$ , we have

$$\mathbb{E}[X^2 | X > t] = \mathbb{E}(X^2) + 2t(t + \mathbb{E}(X)) - t^2 = \mathbb{E}(X^2) + 2t\mathbb{E}(X) + t^2 = \mathbb{E}[(X+t)^2]$$

□

**Example 6.4. (ISE 620):** Enter a bank with one teller, 5 present upon your arrival, all service times are exponential with parameter 1.  $T$  is time you spend in system.

$$\mathbb{E}(T) = R + \sum_{i=1}^t S_i \implies \mathbb{E}(T) = \mathbb{E}(R) + \sum_{i=1}^5 \mathbb{E}(S_i)$$

$$\mathbb{E}(T) = \mathbb{E}(R) + \frac{5}{\lambda} = \frac{6}{\lambda}$$

**Example 6.5. (ISE 620):** Suppose  $X \sim \text{Exponential}(\lambda)$  and  $Y \sim \text{Exponential}(\mu)$ . Then

$$\Pr(X < Y) = \int_0^\infty \Pr(Y > X | X = x) \cdot \lambda e^{-\lambda x} dx = \int_0^\infty e^{-\mu x} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{-(\lambda+\mu)x} dx = \boxed{\frac{\lambda}{\lambda + \mu}}$$

**Remark. (ISE 620).** The following is an intuitive explanation for why  $\min\{X, Y\}$  is distributed exponentially if  $X$  and  $Y$  are exponential: Suppose we require two machines with failure times  $X$  and  $Y$  distributed exponentially, and we need both to work. If at time  $t$  they are both still working, the remaining expected time is the same as initially because of the memoryless property (Proposition 6.3.23). Thus  $\min\{X, Y\}$  is also memoryless, and exponential. Do the math:

$$\Pr(\min\{X, Y\} > t) = \Pr(X > t, Y > t) = e^{-\lambda t} e^{-\mu t} = e^{-(\lambda+\mu)t}$$

The same is not true for maximum:

$$\Pr(\max\{X, Y\} > t) = (1 - e^{-\lambda t})(1 - e^{-\mu t})$$

which is not exponential. Intuitively, if we only need one machine to be working, just because we are working at time  $t$  does not mean we are in the same position as we were initially; one machine could have failed. Now the general case:  $X_1, \dots, X_n \sim \text{Exponential}(\lambda_i)$ .

$$\Pr(\min\{X_i\} > t) = \Pr(X_1 > t \cap \dots \cap X_n > t) = \prod_{i=1}^n \Pr(X_i > t) = \exp\left(t \sum_{i=1}^n \lambda_i\right)$$

so the minimum of an arbitrary number of exponential random variables is distributed exponentially. Now

$$\Pr(X_i < \min_{j \neq i} \{X_j\}) = \Pr(X_i < Z)$$

where  $Z \sim \text{Exponential}(\sum_{j \neq i} \lambda_j)$ . So by the other example, this probability is  $\lambda_i / \sum_{j=1}^n \lambda_j$ .

**Proposition 6.3.27. (ISE 620).** If  $X$  and  $Y$  are independent exponential random variables,  $\min\{X, Y\}$  is independent of whether  $X$  or  $Y$  is smaller. That is

$$\Pr(\min\{X, Y\} > t \mid X < Y) = \Pr(\min\{X, Y\} > t).$$

*Proof.*

$$\begin{aligned} \Pr(\min\{X, Y\} > t \mid X < Y) &= \frac{\Pr(Y > X > t)}{\Pr(X < Y)} = \frac{1}{\lambda/(\lambda + \mu)} \cdot \int_0^\infty \Pr(Y > X > t \mid X = s) \lambda e^{-\lambda s} ds \\ &= \frac{\lambda + \mu}{\lambda} \int_t^\infty e^{-\mu s} \lambda e^{-\lambda s} ds = \int_t^\infty (\lambda + \mu) e^{-(\lambda + \mu)s} ds = e^{-(\lambda + \mu)t} = \Pr(\min\{X, Y\} > t) \end{aligned}$$

□

**Example 6.6. (ISE 620).** You have two servers with exponential service times with parameters  $\mu_1$  and  $\mu_2$ . You wait until the first person is done serving a customer, then you are served by that server.  $T$  is the time you spend in the system. What is  $E(T)$ ?

$$\begin{aligned} \mathbb{E}(T) &= \mathbb{E}(T \mid \mu_1) \Pr(\mu_1) + \mathbb{E}(T \mid \mu_2) \Pr(\mu_2) = \mathbb{E}(T \mid \mu_1) \frac{\mu_1}{\mu_1 + \mu_2} + \mathbb{E}(T \mid \mu_2) \frac{\mu_2}{\mu_1 + \mu_2} \\ &= \left( \frac{1}{\mu_1} + \mathbb{E}(X_1 \mid X_1 < X_2) \right) \frac{\mu_1}{\mu_1 + \mu_2} + \left( \frac{1}{\mu_2} + \mathbb{E}(X_2 \mid X_2 < X_1) \right) \frac{\mu_2}{\mu_1 + \mu_2} \\ &= \left( \frac{1}{\mu_1} + \mathbb{E}(\min\{X_1, X_2\}) \right) \frac{\mu_1}{\mu_1 + \mu_2} + \left( \frac{1}{\mu_2} + \mathbb{E}(\min\{X_1, X_2\}) \right) \frac{\mu_2}{\mu_1 + \mu_2} \\ &= \left( \frac{1}{\mu_1} + \frac{1}{\mu_1 + \mu_2} \right) \frac{\mu_1}{\mu_1 + \mu_2} + \left( \frac{1}{\mu_2} + \frac{1}{\mu_1 + \mu_2} \right) \frac{\mu_2}{\mu_1 + \mu_2} \end{aligned}$$

**Proposition 6.3.28. (Equation (5.5) in Sheldon Ross *Introduction to Probability Models*.)** For independent exponential variables  $T_1$  and  $T_2$  with means  $1/\lambda_1$  and  $1/\lambda_2$ ,  $\Pr(T_2 < T_1) = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ .

*Proof.*

$$\begin{aligned} \Pr(T_2 < T_1) &= \int_0^\infty \Pr(T_2 < T_1 \mid T_2 = t) \lambda_2 e^{-\lambda_2 t} dt = \int_0^\infty \Pr(t < T_1) \lambda_2 e^{-\lambda_2 t} dt = \int_0^\infty e^{-\lambda_1 t} \lambda_2 e^{-\lambda_2 t} dt \\ &= \int_0^\infty \lambda_2 e^{-(\lambda_1 + \lambda_2)t} dt = \frac{\lambda_2}{\lambda_1 + \lambda_2} \end{aligned}$$

□

### 6.3.7 Multivariate Gaussian (Normal) Distributions

**Definition 6.37.** From <http://pluto.huji.ac.il/~pchiga/teaching/MathStat/SIAnotes2013.pdf> (definition 2b6): A random vector  $X = (X_1, X_2)$  is Gaussian with mean  $\mu = (\mu_1, \mu_2)$  and the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

if it has a joint pdf of the form

$$f_X(x) = \frac{1}{2\pi\sigma_2\sigma_2\sqrt{1-\rho^2}} \exp \left[ -\frac{1}{2} \frac{1}{1-\rho^2} \left( \frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right) \right]$$

for  $x \in \mathbb{R}^2$ .

**Definition 6.38. (From Heilman Math 541A; Vector-valued Gaussian random variables.)**  $Z := (Z_1, \dots, Z_d) \in \mathbb{R}^d$  is a **Gaussian random vector** if for all  $v \in \mathbb{R}^d$ ,  $\langle v, Z \rangle$  is a Gaussian random variable (in the usual sense). Equivalently, any linear combination of  $Z_1, \dots, Z_d$  is a Gaussian random variable. Also,  $\text{t}Z$  has covariance matrix  $a_{ij} = \mathbb{E}[(Z_i - \mathbb{E}(Z_i))(Z_j - \mathbb{E}(Z_j))]$ ,  $1 \leq i < j \leq d$ .

**Definition 6.39.** A random vector  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  is Gaussian with mean  $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)$  and covariance matrix  $\boldsymbol{\Sigma} = [\sigma_{ij}]$  if it has a joint pdf of the form

$$f_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{x}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \cdot \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

**Definition 6.40 (Multivariate Normal distribution, Math 541B definition).**  $Y \in \mathbb{R}^d$  is multivariate normal if and only if there exist  $B \in \mathbb{R}^{d \times k}$  and  $C \in \mathbb{R}^d$  such that  $Y = BX + C$  where  $X = (X_1, \dots, X_k)$  with  $X_j \sim \mathcal{N}(0, 1)$  for all  $j \in [k]$  and  $X_j \perp\!\!\!\perp X_i$  for all  $i, j \in [k], i \neq j$ .

**Proposition 6.3.29. [Conditional distribution of one Gaussian random variable on another.]**

From <http://pluto.huji.ac.il/~pchiga/teaching/MathStat/SIAnotes2013.pdf> (Proposition 3c1)]

Let  $X$  be a Gaussian random variable in  $\mathbb{R}^2$  such that

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right) = \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix} \right)$$

Then  $f_{X_1|X_2}(x_1; x_2)$  is Gaussian with the (conditional) mean

$$\mathbb{E}(X_1 | X_2 = x_2) = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2) = \mu_1 + \frac{\sigma_{12}^2}{\sigma_2^2}(x_2 - \mu_2)$$

and the (conditional) variance

$$\text{Var}(X_1 | X_2 = x_2) = \sigma_1^2(1 - \rho^2) = \sigma_1^2 - \frac{\sigma_{12}^4}{\sigma_2^2}$$

That is, the conditional distribution of  $X_1$  given  $X_2 = x_2$  is

$$X_1 \mid X_2 = x_2 \sim \mathcal{N} \left( \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2), (1 - \rho^2) \sigma_1^2 \right) = \mathcal{N} \left( \mu_1 + \frac{\sigma_{12}^2}{\sigma_2^2} (x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^4}{\sigma_2^2} \right)$$

**Proposition 6.3.30 (Generalization of Proposition 6.3.29; from [https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution#Conditional\\_distributions](https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions))**  
Suppose

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

where  $\boldsymbol{\mu}_1 \in \mathbb{R}^q$ ,  $\boldsymbol{\mu}_2 \in \mathbb{R}^{p-q}$ ,  $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{q \times q}$ ,  $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{(p-q) \times q}$ , and  $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{(p-q) \times (p-q)}$ . Then

$$\{\mathbf{X}_1 \mid \mathbf{X}_2\} \sim \mathcal{N} \left( \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{12}^T \right) \quad (6.10)$$

**Remark.** Note that the conditional covariance matrix in (6.10) is the Schur complement (see Sections 9.2 and 2.4) of  $\boldsymbol{\Sigma}_{22}$  in  $\boldsymbol{\Sigma}$ .

**Remark.** A similar Berry-Esseen theorem (Theorem 8.6.8) exists for multivariate Gaussian distributions.

**Remark.** Note that this matches the OLS coefficients in the univariate case. In other words, the univariate OLS formula can be derived using only this fact.

Recall Theorem 6.3.4: if two variables are bivariate normal, they are independent if and only if their covariance

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dxdy$$

equals 0.

**Proposition 6.3.31 (Stats 100B homework problem).** Let  $(X_i, Y_i), i = 1, 2, \dots, n$ , be a random sample from a bivariate normal distribution, where  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  are independent.. Then the joint moment generating function of  $(\bar{X}, \bar{Y})$  is

$$\exp \left( \mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \frac{1}{n} \boldsymbol{\Sigma} \mathbf{t} \right) \sim \boxed{\mathcal{N} \left( \boldsymbol{\mu}, \frac{1}{n} \boldsymbol{\Sigma} \right)}$$

and  $(\bar{X}, \bar{Y})$  is bivariate normal with mean  $\boldsymbol{\mu}$  and variance-covariance matrix  $\frac{1}{n} \boldsymbol{\Sigma}$ .

*Proof.* Let  $\mathbf{W}_i = (X_i, Y_i)$ . Then the moment-generating function of  $\mathbf{W}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  is given by

$$M_{\mathbf{W}_i}(\mathbf{t}) = \exp(t' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Sigma} \mathbf{t})$$

where  $\boldsymbol{\mu}$  is a two-dimensional column vector and  $\boldsymbol{\Sigma}$  is a two-by-two matrix. Then

$$(\bar{X}, \bar{Y}) = \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i(\mathbf{t}) = \sum_{i=1}^n \frac{1}{n} \mathbf{W}_i(\mathbf{t})$$

Since  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  are independent,

$$\begin{aligned} M_{(\bar{X}, \bar{Y})}(\mathbf{t}) &= \prod_{i=1}^n M_{\mathbf{W}_i}\left(\frac{1}{n} \mathbf{t}\right) \\ &= \left[ \exp\left(\frac{1}{n} \mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \frac{1}{n} \mathbf{t}' \boldsymbol{\Sigma} \frac{1}{n} \mathbf{t}\right) \right]^n = \exp\left(n \frac{1}{n} \mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} n \frac{1}{n^2} \mathbf{t}' \boldsymbol{\Sigma} \mathbf{t}\right) \\ &= \exp\left(\mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \frac{1}{n} \boldsymbol{\Sigma} \mathbf{t}\right) \boxed{\mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{n} \boldsymbol{\Sigma}\right)} \end{aligned}$$

Thus  $(\bar{X}, \bar{Y})$  is bivariate normal with mean  $\boldsymbol{\mu}$  and variance-covariance matrix  $\frac{1}{n} \boldsymbol{\Sigma}$ .

□

**Example 6.7 (Example from 541B).** Let  $X \sim \mathcal{N}(0, 1)$ . Define

$$Y := \begin{cases} X & |X| \leq a \\ -X & |X| > a \end{cases}$$

Then (a) for any  $a > 0$ ,  $Y \sim \mathcal{N}(0, 1)$ , (b)  $\exists a > 0$  such that  $\mathbb{E}(XY) = 0$ . However,  $X$  and  $Y$  are clearly not independent. In particular the joint distribution of  $X$  and  $Y$  is not multivariate normal.

## 6.4 Exponential Families

(For more notes, see Section 11.21 on generalized linear models.)

**Definition 6.41.** Informally, an **exponential family** is a family of probability distributions that depends on a parameter  $w \in \mathbb{R}^k$ .

Formally, let  $n, k$  be positive integers and let  $\mu$  be a *measure* on  $\mathbb{R}^n$  (that is, a probability law that does not necessarily sum to 1). Let  $t_1, \dots, t_k : \mathbb{R}^n \rightarrow \mathbb{R}$ . Let  $h : \mathbb{R}^n \rightarrow [0, \infty]$ , and assume  $h$  is not identically zero. For any  $w = (w_1, \dots, w_k) \in \mathbb{R}^k$ , define

$$a(w) := \log \left[ \int_{\mathbb{R}^n} h(x) \exp \left( \sum_{i=1}^k w_i t_i(x) \right) d\mu(x) \right], \quad \forall x \in \mathbb{R}^n$$

The set  $\{w \in \mathbb{R}^k\}$  is called the **natural parameter space**. On this set, the function

$$f_w(x) := h(x) \exp \left( \sum_{i=1}^k w_i t_i(x) - a(w) \right), \quad \forall x \in \mathbb{R}^n$$

satisfies  $\int_{\mathbb{R}^n} f_w(x) d\mu(s) = 1$  (by the definition of  $a(w)$ ). (Why?)

$$\int_{\mathbb{R}^n} f_w(x) d\mu(s) = \frac{\int_{\mathbb{R}^n} h(x) \exp \left( \sum_{i=1}^k w_i t_i(x) \right) d\mu(x)}{\int_{\mathbb{R}^n} h(x) \exp \left( \sum_{i=1}^k w_i t_i(x) \right) d\mu(x)} = 1.$$

So, the set of functions (which can be interpreted as probability density functions, or as probability mass functions according to  $\mu$ )  $\{f_w : \theta \in \Theta : a(w(\theta)) < \infty\}$  is called a  **$k$ -parameter exponential family in canonical form**.

More generally, let  $\Theta \in \mathbb{R}^k$  be any set and let  $w : \Theta \rightarrow \mathbb{R}^k$ . We define a  **$k$ -parameter exponential family** to be a set of functions  $\{f_\theta : \theta \in \Theta\}$ , where

$$f_\theta(x) := h(x) \exp \left( \sum_{i=1}^k w_i(\theta) t_i(x) - a(w(\theta)) \right), \quad \forall x \in \mathbb{R}^n$$

satisfies

Also, an exponential family is called **curved** if the dimension of  $\Theta$  is less than  $k$ .

**Example 6.8.** A Gaussian random variable has mean  $\mu$  and standard deviation  $\sigma$ , and is in an exponential family:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \mu \in \mathbb{R}, \sigma > 0$$

In this case,  $k = 2$ .

⋮

**(Continuation of example 3.2.)** If we instead write this in canonical form, we get

$$a(w) = \frac{\mu^2}{2\sigma^2} + \log(\sigma) = \left( \frac{\mu}{\sigma^2} \right)^2 \left[ (-2) \frac{(-1)}{\sigma^2} \right]^{-1} - \frac{1}{2} \log \left( (-2) \frac{(-1)}{2\sigma^2} \right) = \frac{w_1^2}{4w_2} - \frac{1}{2} \log(-2w_2)$$

That is, in this case you can get rid of the thetas and write this in canonical form. Define

$$f_w(x) = h(x) \exp \left( \sum_{i=1}^2 w_i t_i(x) - a(w) \right), \quad \forall x \in \mathbb{R}$$

where  $w$  ranges in  $\{(w_1, w_2) \in \mathbb{R}^2 : w_2 > 0\}$ .

**Remark.** (notes remark 3.4) (**location family.**) Let  $X$  be a random variable with density  $f : \mathbb{R} \rightarrow \mathbb{R}$ . Let  $\mu \in \mathbb{R}$ . Then the densities  $\{f(x + \mu)\}_{\mu \in \mathbb{R}}$  are called a **location family** of  $X$ . This family may or may not be an exponential family.

**Remark.** (notes remark 3.6) (**Scale family.**) Let  $X$  be a random variable with density  $f : \mathbb{R} \rightarrow \mathbb{R}$ . Let  $\sigma > 0$ . The family of densities  $\{\sigma^{-1}f(x/\sigma)\}_{\sigma>0}$  is called a **scale family**. Note

$$\int_{\mathbb{R}} \sigma^{-1} f(x/\sigma) dx = \int_{\mathbb{R}} f(y) dy = 1$$

(substituting  $y = x/\sigma$ ,  $dy = dx/\sigma$ ).

**Remark.** (notes remark 3.7) (**Location and scale family.**) The family of densities  $\{\sigma^{-1}f((x+\mu)/\sigma)\}_{\sigma>0, \mu \in \mathbb{R}}$  is called a **location and scale family**. This family may or may not be an exponential family (although Gaussian random variables are one example where it is.)

**Exercise 4.** Try to write a binomial random variable with parameters  $n$  and  $p$  as a two-parameter exponential family. (Note: it's impossible to do, but instructive to try.)

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!} \exp(x \log(p)) (1-p)^{n-x} = \dots$$

It turns out you can do it with one parameter with  $p$ , but not with two parameters with  $n$ .

**Example 6.9 (Example 3.15 in 541A notes).** Write a binomial random variable with parameters  $n$  and  $p$  as an exponential family (when  $n$  is fixed), then take derivatives in  $p$ .

Recall

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 \text{ for any other } x.$$

Keep  $n$  fixed and look at  $p$ .

$$\binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \exp(x \log p + (n-x) \log(1-p)) = \binom{n}{x} \exp[x \log(p/(1-p)) - (-1)n \log(1-p)]$$

Define  $h : \mathbb{R} \rightarrow \mathbb{R}$  so that

$$h(x) = \begin{cases} \binom{n}{x} & 0 \leq x \leq n \text{ is an integer} \\ 0 & \text{otherwise} \end{cases}$$

Let  $\theta := p, \Theta := (0, 1), t(x) := x, w(\theta) := \log(\theta/(1-\theta)), a(w(\theta)) = -n \log(1-\theta)$ .

So,  $f_\theta(x) := h(x) \exp [w(\theta)t(x) - a(w(\theta))], \forall x \in \mathbb{R}$ .

Now we will take derivatives. As in previous example (last class),

$$e^{-a(w(\theta))} \frac{\partial}{\partial \theta_1} e^{a(w(\theta))} = \mathbb{E}_\theta \left( \sum_{i=1}^k \frac{\partial w_i}{\partial \theta_1} t_i \right)$$

So in this case

$$e^{-a(w(\theta))} \frac{\partial}{\partial \theta_1} e^{a(w(\theta))} = \mathbb{E}_\theta \left( \frac{d}{d\theta} w(\theta) t \right).$$

Plugging in yields

$$(1-\theta)^n \frac{d}{d\theta} (1-\theta)^{-n} = \frac{n}{1-\theta}$$

on the left side and

$$= \left( \frac{1}{\theta} + \frac{1}{1-\theta} \right) \mathbb{E}_\theta(X) = \mathbb{E}_\theta(X) \left( \frac{1}{\theta(1-\theta)} \right)$$

on the right side which yields

$$\mathbb{E}_\theta(X) = \frac{n}{1-\theta} (\theta(1-\theta)) = n\theta \implies \mathbb{E}(X) = pn$$

**Remark.** This is a probability mass function since it is defined only on integers.

**Example 6.10 (GSBA 604).** Suppose  $Y \sim \text{Poisson}(\mu)$ . Express it as an exponential family.

**Solution.**

$$\begin{aligned} \mathbb{P}_\mu(Y=y) &= \frac{e^{-\mu} \mu^y}{y!}, \quad y \in \{0, 1, \dots\} \\ &= \frac{e^{y \log \mu - \mu}}{y!} \end{aligned}$$

So the natural parameter  $\eta$  is  $\log \mu$ , the normalizing function  $\psi(\eta) = \mu$ , and the carrier density  $c(y) = 1/y!$ .

#### 6.4.1 Differential identities (Generalization of Moment-Generating Functions)

Recall that the moment-generating function for a Gaussian random variable is

$$\mathbb{E}(e^{tX}) = e^{t^2/2} \quad \forall t \in \mathbb{R}$$

Consequently,

$$\left. \frac{d^m}{dt^m} \right|_{t=0} \mathbb{E}(e^{tX}) = \mathbb{E}(X^m)$$

for any integer  $m > 0$ . We can do a similar thing for an exponential family—we can differentiate the parameters of exponential families and find out information about the exponential family. As in Definition 6.41, let

$$a(w) := \log \int_{\mathbb{R}^n} h(x) \exp \left( \sum_{i=1}^k w_i t_i(x) \right) d\mu(x)$$

Define

$$W := \{w \in \mathbb{R}^k : a(w) < \infty\}.$$

**Question:** Is  $a(w)$  differentiable?

**Lemma 6.4.1** (Lemma 3.8 in 541A notes). The function  $a(w)$  is continuous and has continuous partial derivatives of all orders on the interior of  $W$ . Moreover, we can compute these derivatives by differentiating under the integral sign.

*Proof.* We prove only the case of a first order partial derivative. Consider the case of the partial derivative with respect to  $w_1$  at  $w$  in the interior of  $W$ . Let  $e_1 = (1, 0 \dots, 0) \in \mathbb{R}^k$ . Since the exponential function is analytic, it suffices to show that the partial derivative of  $e^{a(w)}$  exists in the direction  $e_1$ . We form the difference quotient for  $e^{a(w)}$  as follows:

$$\begin{aligned} \frac{\exp[a(w + \epsilon e_1)] - \exp[a(w)]}{\epsilon} &= \frac{1}{\epsilon} \int_{\mathbb{R}^n} h(x) \left[ \exp \left( \epsilon t_1(x) + \sum_{i=1}^k w_i t_i(x) \right) - \exp \left( \sum_{i=1}^k w_i t_i(x) \right) \right] d\mu(x) \\ &= \int_{\mathbb{R}^n} h(x) \frac{\exp[\epsilon t_1(x)] - 1}{\epsilon} \exp \left( \sum_{i=1}^k w_i t_i(x) \right) d\mu(x). \end{aligned} \tag{6.11}$$

By the Mean Value Theorem,

$$\frac{|e^a - 1|}{a} = \frac{|e^a - e^0|}{a - 0} \leq e^a \text{ if } a > 0, \text{ or } 1 \text{ if } a < 0$$

so we have

$$|e^a - 1| \leq |a| \max\{1, e^a\} \leq |a|e^{|a|} \leq e^{2|a|} \leq e^{2a} + e^{-2a} \quad \forall a \in \mathbb{R}.$$

In particular, if  $a = \epsilon t_1(x)$ , we have

$$|e^{\epsilon t_1(x)} - 1| \leq e^{2\epsilon t_1(x)} + e^{-2\epsilon t_1(x)} \quad (6.12)$$

Therefore, examining the integrand of (6.11), we have

$$\begin{aligned} & \left| h(x) \frac{\exp[\epsilon t_1(x)] - 1}{\epsilon} \exp\left(\sum_{i=1}^k w_i t_i(x)\right) \right| \leq h(x) \left| \frac{\exp[\epsilon t_1(x)] - 1}{\epsilon} \right| \exp\left(\sum_{i=1}^k w_i t_i(x)\right) \\ & \leq (\text{by (6.12)}) h(x) (e^{2\epsilon t_1(x)} + e^{-2\epsilon t_1(x)}) \exp\left(\sum_{i=1}^k w_i t_i(x)\right) \\ & \quad \vdots \\ & = h(x) |t_1(x)| \exp\left(\epsilon t_1(x) + \sum_{i=1}^k w_i t_i(x)\right) + h(x) |t_1(x)| \exp\left(-\epsilon t_1(x) + \sum_{i=1}^k w_i t_i(x)\right) \end{aligned} \quad (6.13)$$

If both of the expressions in (6.13) always have finite expected value uniformly for all  $\epsilon > 0$ , we will be done by the Dominated Convergence Theorem (Theorem 6.4.2). (Assuming those things are bounded uniformly in expectation then the limit of the integrals is the integral of the limits.)

□

**Remark. Notes from proof of Lemma 3.8.**

$\left|\frac{e^a - 1}{a}\right| \leq e^a + e^{-a}$ . Then

$$\left| h(x) \exp\left(\epsilon t_1(x) + \sum_{i=1}^k w_i t_i(x)\right) - \exp\left(\sum_{i=1}^k w_i t_i(x)\right) \right| \leq h(x) \exp\left(\sum_{i=1}^k w_i t_i(x)\right) |t_1(x)| \left( \exp(\epsilon t_1(x)) + \exp(-\epsilon t_1(x)) \right)$$

**Theorem 6.4.2** (Dominated convergence theorem (Math 541A)). Let  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  and  $Y : \Omega \rightarrow [0, \infty)$  such that  $|X_i| \leq Y$  for all  $i \geq 1$  and  $\mathbb{E}(Y) < \infty$ . Assume  $X_1, X_2, \dots$  converges almost surely to  $X : \Omega \rightarrow \mathbb{R}$ . Then

$$\lim_{i \rightarrow \infty} \mathbb{E}(x_i) = \mathbb{E}\left(\lim_{i \rightarrow \infty} X_i\right) = \mathbb{E}(X)$$

**Corollary 6.4.2.1** (Corollary 3.11 in 541A notes.). Let  $\epsilon > 0$ . Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable such that  $\mathbb{E}(e^{wX}) < \infty$  for all  $w \in (-\epsilon, \epsilon)$ . Then for any integer  $n \geq 1$ ,  $\mathbb{E}(X^n)$  exists and

$$\left. \frac{d^n}{dw^n} \right|_{w=0} e^{wX} = \mathbb{E}(X^n).$$

*Proof.* Apply Lemma 6.4.1 when  $\mu = \mathbb{P}, h = 1, k = 1, t(x) = x$ : we see that

$$a(w) = \log \int_{\mathbb{R}^n} e^{wx} d\mathbb{P}(x)$$

□

*Proof (GSBA 604 proof, just for first two moments).* Notation: The density is

$$g_\eta(y) = c(y) \exp \{ \eta y - \psi(\eta) \}$$

where  $c(y)$  is the **carrier density**,  $\eta$  is the **natural parameter**, and  $\psi(\cdot)$  is the **normalizing function**, chosen so that it is a proper density: that is, choose  $\psi(\eta)$  such that

$$\int_{\mathbb{R}} g_\eta(y) dy = 1,$$

so

$$e^{-\psi(y)} = \left( \int_{\mathbb{R}} c(y) e^{\eta y} dy \right)^{-1}$$

Now

$$\int g_\eta(y) dy = 1$$

differentiate with respect to  $\eta$ :

$$\begin{aligned} \int e^{\eta y - \psi(\eta)} [y - \frac{d\psi(\eta)}{d\eta}] c(y) dy &= 0 \\ \implies \int y g_\eta(y) dy - \frac{d\psi(\eta)}{d\eta} \int e^{\eta y - \psi(\eta)} c(y) dy &= 0 \quad (6.14) \end{aligned}$$

so then  $\mathbb{E}(y) - \frac{d\psi(\eta)}{d\eta} = 0 \implies \mathbb{E}(y) = \frac{d\psi(\eta)}{d\eta}$ .

Next,

$$\begin{aligned}
& \int y \frac{dg_\eta(y)}{d\eta} dy - \frac{d\psi^2(\eta)}{d\eta^2} = 0 \\
& \implies \int y \cdot e^{\eta y - \psi(\eta)} \left( y - \frac{d\psi(\eta)}{d\eta} \right) c(y) dy - \frac{d\psi^2(\eta)}{d\eta^2} = 0 \\
& \implies \int y^2 g_\eta(y) dy - \underbrace{\int y e^{\eta y - \psi(\eta)} c(y) dy}_{\frac{d\psi(\eta)}{d\eta}} - \frac{d\psi^2(\eta)}{d\eta^2} = 0 \\
& \implies \mathbb{E}(Y^2) - \left[ \frac{d\psi(\eta)}{d\eta} \right]^2 - \frac{d\psi^2(\eta)}{d\eta^2} = 0 \\
& \implies \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \frac{d\psi^2(\eta)}{d\eta^2}.
\end{aligned}$$

□

**Remark.** Notes from example 3.13.

$$\begin{aligned}
e^{-a(w)} \frac{\partial}{\partial w_2} e^{a(w)} &= \int_{\mathbb{R}} t_1(x) h(x) \exp \left( \sum_{i=1}^k w_i t_i(x) - a(w) \right) dx \\
&= \int_{\mathbb{R}} t_1(x) f_x(x) dx
\end{aligned} \tag{6.15}$$

So by chain rule

$$\begin{aligned}
e^{-a(w(\theta))} \frac{\partial}{\partial \theta_1} e^{a(w(\theta))} &= e^{-a(w(\theta))} \sum_i \frac{\partial e^{a(w)}}{\partial w_i} \frac{\partial w_i}{\partial \theta_1} \\
&= (\text{by (6.15)}) \sum_{i=1}^k \frac{\partial w_i}{\partial \theta_1} \mathbb{E}_\theta t_i = \mathbb{E}_\theta \left( \sum_{i=1}^k t_i \frac{\partial w_i}{\partial \theta_1} \right)
\end{aligned} \tag{6.16}$$

**Example 6.11.** Recall Example 3.3. Gaussian, mean  $\mu$ , variance  $\sigma^2$ .  $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$ ,  $t_1(x) = x$ ,  $t_2(x) = x^2$ ,  $w_1(\theta) = \theta_1/\theta_2 = \mu/\sigma^2$ ,  $w_2(\theta) = -1/(2\sigma^2) = -1/(2\theta_2)$ ,  $a(w(\theta)) = \theta_1^2/(2\theta_2) + (1/2)\log(\theta_2) = \mu^2/(2\sigma^2) + \log(\sigma^2)$ .

Complete both sides of (6.16).

$$e^{-a(w(\theta))} \frac{\partial}{\partial \theta_1} e^{a(w(\theta))} = \frac{\theta_1}{\theta_2} = \frac{\mu}{\sigma^2}$$

$$\frac{\partial}{\partial \theta_1} e^{a(w(\theta))} = \frac{d}{d\theta_1} \sqrt{\theta_2} \exp \left( \frac{\theta_1^2}{2\theta_2} \right) = \frac{\theta_1}{\theta_2} \exp(a(w(\theta)))$$

Right side:

$$\frac{dw_1}{d\theta_2} = \frac{1}{\theta_2}, \quad \frac{dw_2}{d\theta_1} = 0$$

$$\implies \mathbb{E}_\theta \left( \sum_{i=1}^k t_i \frac{dw_i}{d\theta_1} \right) = \mathbb{E}_\theta (t_1 \frac{dw_1}{d\theta_1} + t_2 \frac{dw_2}{d\theta_1})$$

$$= \mathbb{E}_\theta (x/\theta_2) \implies \mathbb{E}_\theta (x) = \theta_1 = \mu$$

**Theorem 6.4.3 (Theorem 3.4.2 from Casella and Berger).** If  $X$  is a random variable in an exponential family, then

$$\mathbb{E} \left( \sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X) \right) = \frac{\partial}{\partial \theta_j} a(w(\theta)). \quad (6.17)$$

**Example 6.12.** Recall Example 3.3 again: Gaussian, mean  $\mu$ , variance  $\sigma^2$ .  $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$ ,  $t_1(x) = x$ ,  $t_2(x) = x^2$  k  $w_1(\theta) = \theta_1/\theta_2 = \mu/\sigma^2$ ,  $w_2(\theta) = -1/(2\sigma^2) = -1/(2\theta_2)$ ,  $a(w(\theta)) = \theta_1^2/(2\theta_2) + (1/2)\log(\theta_2) = \mu^2/(2\sigma^2) + \log(\sigma^2)$ .

Complete both sides of (6.17).

$$\frac{\partial}{\partial \theta_1} a(w(\theta)) = \frac{\partial}{\partial \theta_1} \left( \theta_1^2/(2\theta_2) + (1/2)\log(\theta_2) \right) = \frac{\theta_1}{\theta_2}.$$

Left side:

$$\begin{aligned} \mathbb{E} \left( \sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_1} t_i(X) \right) &= \mathbb{E} \left( \frac{\partial}{\partial \theta_1} w_1(\theta) t_1(X) + \frac{\partial}{\partial \theta_1} w_2(\theta) t_2(X) \right) = \mathbb{E} \left( \frac{\partial}{\partial \theta_1} \frac{\theta_1}{\theta_2} \cdot x - \frac{1}{2} \frac{\partial}{\partial \theta_1} \frac{1}{\theta_2} x^2 \right) \\ &= \frac{1}{\theta_2} \mathbb{E}(x) \\ \implies \frac{1}{\theta_2} \mathbb{E}(x) &= \frac{\theta_1}{\theta_2} \implies \mathbb{E}(X) = \theta_1 = \mu. \end{aligned}$$

Repeating by taking the partial derivatives with respect to  $\theta_2$  instead:

$$\frac{\partial}{\partial \theta_2} a(w(\theta)) = \frac{\partial}{\partial \theta_2} \left( \theta_1^2/(2\theta_2) + (1/2)\log(\theta_2) \right) = \frac{\theta_1^2}{2} \cdot \frac{-1}{\theta_2^2} + \frac{1}{2\theta_2} = \frac{\theta_2 - \theta_1^2}{2\theta_2^2}$$

Left side:

$$\begin{aligned} \mathbb{E} \left( \sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_2} t_i(X) \right) &= \mathbb{E} \left( \frac{\partial}{\partial \theta_2} w_1(\theta) t_1(X) + \frac{\partial}{\partial \theta_2} w_2(\theta) t_2(X) \right) = \mathbb{E} \left( \frac{\partial}{\partial \theta_2} \frac{\theta_1}{\theta_2} \cdot x - \frac{1}{2} \frac{\partial}{\partial \theta_2} \frac{1}{\theta_2} x^2 \right) \\ &= \mathbb{E} \left( -\frac{\theta_1}{\theta_2^2} \cdot x + \frac{1}{2\theta_2^2} x^2 \right) = -\frac{\theta_1}{\theta_2^2} \mathbb{E}(X) + \frac{1}{2\theta_2^2} \mathbb{E}(X^2) = -\frac{\theta_1^2}{\theta_2^2} + \frac{1}{2\theta_2^2} \mathbb{E}(X^2) \\ \implies -\frac{\theta_1^2}{\theta_2^2} + \frac{1}{2\theta_2^2} \mathbb{E}(X^2) &= \frac{\theta_2 - \theta_1^2}{2\theta_2^2} \iff -2\theta_1^2 + \mathbb{E}(X^2) = \theta_2 - \theta_1^2 \iff \mathbb{E}(X)^2 = \theta_2 + \theta_1^2 = \sigma^2 + \mu^2. \end{aligned}$$

## 6.5 KL Divergence (DSO 607)

Let  $f(\cdot | \theta) : \mathbb{R}^k \rightarrow \mathbb{R}^n$  be a family of parameterized probability densities with  $\theta \in \Theta$ . Suppose the true model is parameterized by  $\theta_0 \in \mathbb{R}^s$ , and we are interested in comparing a different model parameterized by  $\theta \in \mathbb{R}^k$  to this model. The likelihood ratio  $\tau : \mathbb{R}^n \rightarrow \mathbb{R}$  defined by  $\tau(x; \theta, \theta_0) := f(x | \theta)/f(x | \theta_0)$  is a good tool for this comparison. Define the **discrimination** between  $\theta$  and  $\theta_0$  at  $x$  to be  $\Phi(\tau(x; \theta, \theta_0))$  for some function  $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ , and define the **mean discrimination**  $D(\theta; \theta_0) : \mathbb{R}^k \rightarrow \mathbb{R}$  between  $\theta$  and  $\theta_0$  as

$$D(\theta; \theta_0) := \int_{\mathbb{R}^n} \Phi(\tau(x; \theta, \theta_0)) \cdot f(x | \theta_0) dx = \mathbb{E}_{\theta_0} [\Phi(\tau(X; \theta, \theta_0))].$$

To choose  $\Phi$  so that  $D(\theta; \theta_0)$  behaves like a distance, we would like  $\Phi(1) = 0$  since this covers the case where  $\theta = \theta_0$ , because under the assumption that  $f(x | \theta) = f(x | \theta_0) \forall x \in \mathbb{R}^n \iff \theta = \theta_0$ , we have  $\tau(x; \theta, \theta_0) = 1 \forall x \in \mathbb{R}^n \iff \theta = \theta_0$ . In particular, we would like this to be the minimum of the function; that is,  $\tau''(1; \theta, \theta_0) > 0$ . Lastly, we would also like  $\tau'(x; \theta, \theta_0) > 0 \forall x \in \mathbb{R}^n$ . One such choice is  $\Phi(t) := -2 \log(t)$ . This yields the **Kullback-Leibler (KL) divergence**  $I(\theta; \theta_0) : \mathbb{R}^k \rightarrow \mathbb{R}$

$$\begin{aligned} I(\theta; \theta_0) &:= -2 \int_{\mathbb{R}^n} \log \left( \frac{f(x | \theta)}{f(x | \theta_0)} \right) \cdot f(x | \theta_0) dx = 2 \int_{\mathbb{R}^n} [\log(f(x | \theta_0)) - \log(f(x | \theta))] \cdot f(x | \theta_0) dx \\ &= 2\mathbb{E}_{\theta_0} [\log(f(X | \theta_0))] - 2\mathbb{E}_{\theta_0} [\log(f(X | \theta))]. \end{aligned}$$

Of course, in practice we will estimate  $\theta_0$  as best as we can, by maximizing the **probabilistic negentropy**

$$\mathbb{E}_Z I(\theta; \hat{\theta}_0(Z)) := 2\mathbb{E}_{\theta_0} [\log(f(X | \theta_0))] - 2\mathbb{E}_{\theta_0, Z} \left[ \log \left( f \left( X | \hat{\theta}_0(Z) \right) \right) \right]$$

where  $Z$  describes the probability distribution of the sample data and  $\hat{\theta}_0(Z)$  is our estimate of  $\theta_0$  from the data.

Way we wrote this in DSO 607: KL Divergence of density  $f$  (estimated) from density  $g$  (true model/distribution):

$$I(g; f) = \int [\log g(z)] g(z) dz - \int [\log f(z)] g(z) dz = \mathbb{E}_g \log(g(z)) - \mathbb{E}_g (\log(f(z)))$$

One application of KL Divergence is AIC for model selection; see Section 11.11.1. See also Section 10.4.6.

## 6.6 Worked problems

### 6.6.1 Example Problems That Will Likely Appear on Midterm (and Final)

- (1) Let  $X$  be uniform on  $[1, 5]$ , let  $Y$  be uniform on  $[0, 1]$ , and assume that  $X$  and  $Y$  are independent.

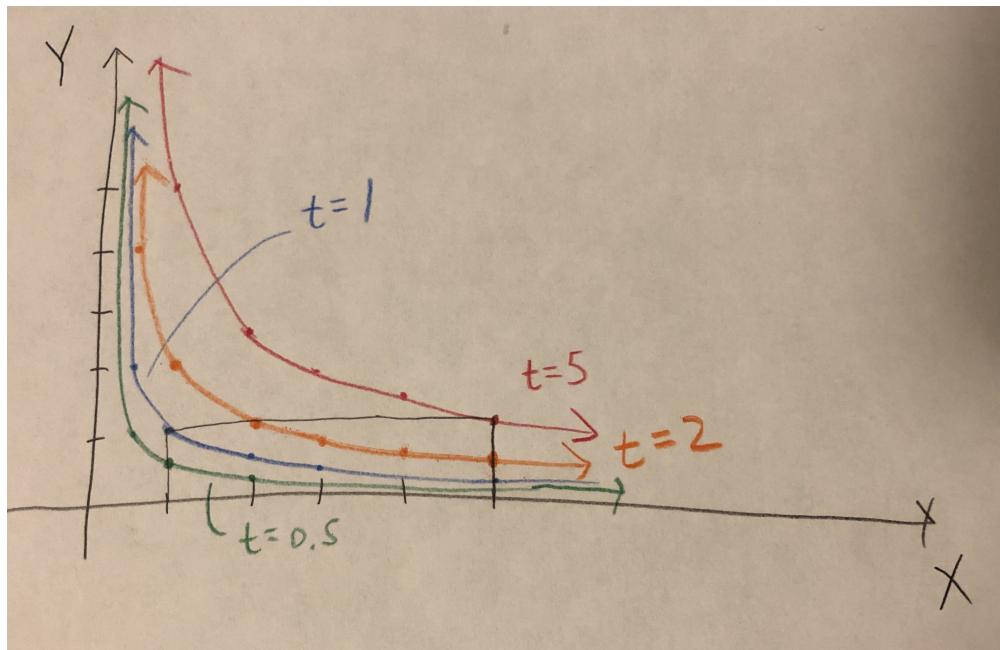
- (a) Compute the probability density function of the product  $XY$ .
- (b) **Only part included on midterm.** Compute the cumulative distribution function of the ratio  $X/Y$ .
- (c) Compute the characteristic function of the sum  $X + Y$ .
- (d) Compute the moment-generating function of the random variable  $X - \ln(Y)$ .

**Solution.**

- (a) We will find the cdf and then differentiate to yield the pdf. Observe that

$$F_{XY}(t) = \Pr(XY \leq t) = \Pr(Y \leq tX^{-1})$$

Plotting the density function of  $XY$  along with plots of  $F_{XY}(t)$  as a function of  $X$  for various values of  $t$ , we have the following:



Now since both  $X$  and  $Y$  are distributed uniformly, for a given  $t$ ,  $\Pr(Y \leq tX^{-1})$  is the area under the curve and in the rectangle, weighted by  $1/4$  since the rectangle has total area 4 but total probability 1. It is clear that the four regimes we need to consider are (1)  $t < 0$ , (2)  $0 \leq t < 1$ , (3)  $1 \leq t < 5$ , and (4)  $t \geq 5$ .

- (1)  $t < 0$ : The curve lies below the rectangle, so there is no area below the curve and in the rectangle. Therefore  $\boxed{\Pr(Y \leq tX^{-1} \mid t < 0) = 0}$ . (This is also clear since  $tX^{-1}$  would be a negative number and  $Y$  is nonnegative.)
- (2)  $0 \leq t < 1$ : Integrating the relevant area, we have

$$\Pr(Y \leq tX^{-1} \mid 0 \leq t < 1) = \frac{1}{4} \int_1^5 \frac{t}{x} dx = \frac{t}{4} [\log(x)]_1^5 = \boxed{\frac{t}{4} \log(5)}$$

- (3)  $1 \leq t < 5$ : In this case, the area is a rectangle of height 1 and width  $t - 1$  plus the area under the curve from  $t$  to 5.

$$\Pr(Y \leq tX^{-1} \mid 1 \leq t < 5) = \frac{1}{4} \left( 1 \cdot (t-1) + \int_t^5 \frac{t}{x} dx \right) = \frac{1}{4} \left( t-1 + t [\log(x)]_t^5 \right) = \boxed{\frac{1}{4} [t(1 + \log(5/t)) - 1]}$$

(4)  $t \geq 5$ : In this case, the entire rectangle lies below the curve. Therefore  $\boxed{\Pr(Y \leq tX^{-1} \mid t \geq 5) = 1}$ .

So we have

$$F_{XY}(t) = \begin{cases} 0 & t < 0 \\ \frac{t}{4} \log(5) & 0 \leq t < 1 \\ \frac{1}{4} [t(1 + \log(5/t)) - 1] & 1 \leq t < 5 \\ 1 & t \geq 5 \end{cases}$$

Finally, differentiating yields

$$f_{XY}(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{4} \log(5) & 0 \leq t < 1 \\ \frac{1}{4} \log\left(\frac{5}{t}\right) & 1 \leq t < 5 \\ 0 & t \geq 5 \end{cases}$$

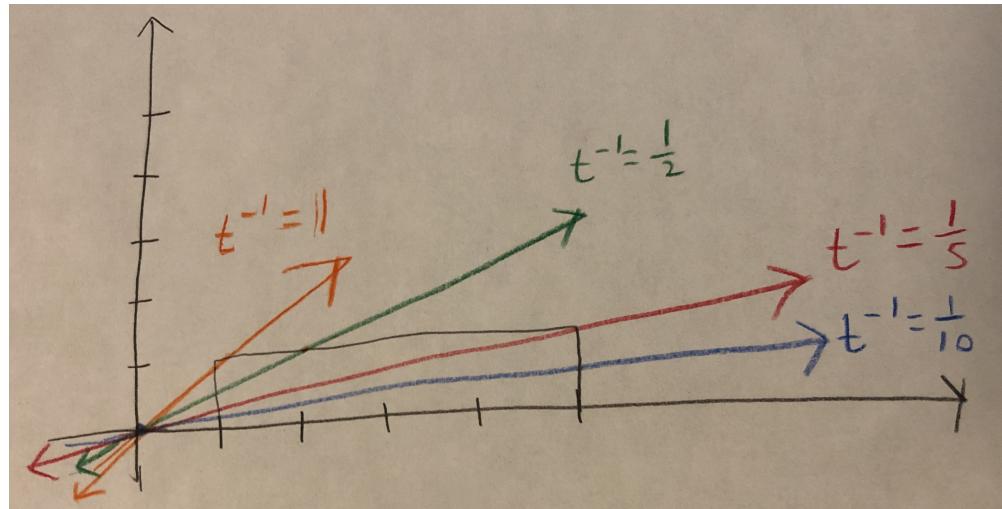
since

$$\begin{aligned} \frac{d}{dt} \left( \frac{1}{4} [t(1 + \log(5/t)) - 1] \right) &= \frac{1}{4} \left[ 1 + \log(5/t) + t \left( \frac{1}{5/t} \cdot -5 \cdot t^{-2} \right) \right] = \frac{1}{4} \left[ 1 + \log(5/t) + t(-1 \cdot t^{-2}) \right] \\ &= \frac{1}{4} \log\left(\frac{5}{t}\right) \end{aligned}$$

(b) **Only part included on midterm.** We will proceed in a similar way as part (a). Observe that

$$F_{X/Y}(t) = \Pr\left(\frac{X}{Y} \leq t\right) = \Pr(Y \geq X/t)$$

Plotting the density function of  $X/Y$  along with plots of  $F_{X/Y}(t)$  as a function of  $X$  for various values of  $t$ , we have the following:



Now since both  $X$  and  $Y$  are distributed uniformly, for a given  $t$ ,  $\Pr(Y \geq X/t)$  is the area above the curve and in the rectangle, weighted by  $1/4$  since the rectangle has total area 4 but total probability 1. It is clear that the three regimes we need to consider are (1)  $t^{-1} \geq 1 \iff t \leq 1$ , (2)  $1/5 \leq t^{-1} < 1 \iff 1 < t \leq 5$ , and (3)  $0 < t^{-1} < 1/5 \iff t > 5$ .

- (1)  $t \leq 1$ : The curve lies above the rectangle, so there is no area above the curve and in the rectangle. Therefore  $\Pr(Y \geq X/t \mid t \leq 1) = 0$ . (This is also clear since  $X/t$  would have to be greater than 1 and  $Y$  is less than or equal to 1.)
- (2)  $1 < t \leq 5$ : The relevant area is the triangle above the green line in the rectangle. Note that it intersects the vertical line at  $Y = 1/t$  and the horizontal line at  $X = t$ .

$$\Pr(Y \geq X/t \mid 1 < t \leq 5) = \frac{1}{4} \cdot \frac{1}{2} \left(1 - \frac{1}{t}\right)(t-1) = \frac{1}{8} \left(t - 1 - 1 + \frac{1}{t}\right) = \frac{1}{8} \left(t - 2 + \frac{1}{t}\right)$$

- (3)  $t > 5$ : In this case, the area is a trapezoid above the blue line and in the rectangle. Note that the blue line intersects the left vertical line at  $Y = 1/t$  and the right vertical line at  $Y = 5/t$ .

$$\Pr(Y \geq X/t \mid t > 5) = \frac{1}{4} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{t} + 1 - \frac{5}{t}\right) \cdot 4 = \frac{1}{2} \cdot \left(2 - \frac{6}{t}\right) = 1 - \frac{3}{t}$$

So we have

$$F_{XY}(t) = \begin{cases} 0 & t \leq 1 \\ \frac{1}{8} \left(t - 2 + \frac{1}{t}\right) & 1 < t \leq 5 \\ 1 - \frac{3}{t} & t > 5 \end{cases}$$

- (c) The characteristic function for a uniform distribution on  $[a, b]$  is

$$\frac{2}{(b-a)t} \sin\left(\frac{1}{2}(b-a)t\right) \exp\left(i(a+b)\frac{t}{2}\right).$$

Using the fact that  $X \perp\!\!\!\perp Y \implies \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$ , we have

$$\begin{aligned} \phi_{X+Y}(t) &= \phi_X(t)\phi_Y(t) = \frac{2}{(5-1)t} \sin\left(\frac{1}{2}(5-1)t\right) \exp\left(i(5+1)\frac{t}{2}\right) \cdot \frac{2}{t} \sin\left(\frac{1}{2}t\right) \exp\left(i\frac{t}{2}\right) \\ &= \frac{1}{2t} \sin(2t) \exp(3it) \cdot \frac{2}{t} \sin\left(\frac{1}{2}t\right) \exp\left(i\frac{t}{2}\right) = \boxed{\frac{1}{t^2} \exp\left(\frac{7}{2}it\right) \cdot \sin(2t) \sin\left(\frac{1}{2}t\right)} \end{aligned}$$

- (d) The moment-generating function for a uniform distribution on  $[a, b]$  is

$$M_X(t) = \mathbb{E}(\exp(tx)) = \int_a^b \frac{1}{b-a} \cdot \exp(tx) dx = \frac{1}{b-a} \left[ \frac{1}{t} \exp(tx) \right]_a^b = \frac{1}{(b-a)t} [\exp(bt) - \exp(at)]$$

Therefore the moment-generating function for  $X$  is  $t^{-1}[\exp(t) - 1]$ . Note that

$$\Pr(Y \leq y) = \Pr(-\log(X) \leq y) = \Pr(\log(X) \geq -y) = \Pr(X \geq e^{-y}) = \int_{\exp(-y)}^{\infty} dt = \int_{\exp(-y)}^1 dt$$

Substituting  $t = e^{-u}$  (so that we have  $u = -\log(t)$ ,  $dt = -e^{-u}du$ , we have

$$\Pr(Y \leq y) = - \int_y^0 e^{-u} du = [e^{-u}]_y^0 = 1 - e^{-y}$$

which is the cdf for an exponential distribution with mean 1. Therefore  $Y = -\log(X) \sim \text{Exponential}(1)$ , so

$$M_Y(t) = \frac{1}{1-t}$$

Using the fact that if  $X$  and  $Y$  are independent then  $M_{X+Y}(t) = M_X(t)M_Y(t)$ , we have

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \frac{\exp(t) - 1}{t} \cdot \frac{1}{1-t} = \boxed{\frac{\exp(t) - 1}{t - t^2}}$$

- (2) **Fall 2016 Problem 2.** Let  $X$  and  $Y$  be i.i.d. exponential with mean 1. Show that for every  $t > 0$  the events  $\{\omega : \min\{X, Y\} > t\}$  and  $\{\omega : X < Y\}$  are independent.

**Solution.** Note that  $\min\{X, Y\} > t \iff X > t \cap Y > t$ .

- $\Pr(\min\{X, Y\} > t) = \Pr(X > t \cap Y > t) = \Pr(X > t) \Pr(Y > t)$

$$= \int_t^\infty e^{-x} dx \int_t^\infty e^{-y} dy = -e^{-x}|_t^\infty - e^{-y}|_t^\infty = \boxed{e^{-2t}}$$

- $\Pr(X < Y)$ : Note that in Figure 2, the region that satisfies this condition is region  $G_1$  plus  $G_2$  (note that  $X$  and  $Y$  are nonnegative). Therefore we can find this probability by integrating the joint pdf over that region.

$$\Pr(X < Y) = \iint_{\{G_1+G_2\}} f_{X,Y}(x, y) dxdy$$

Note that the joint pdf is the probability of the marginal pdfs since  $X$  and  $Y$  are independent.

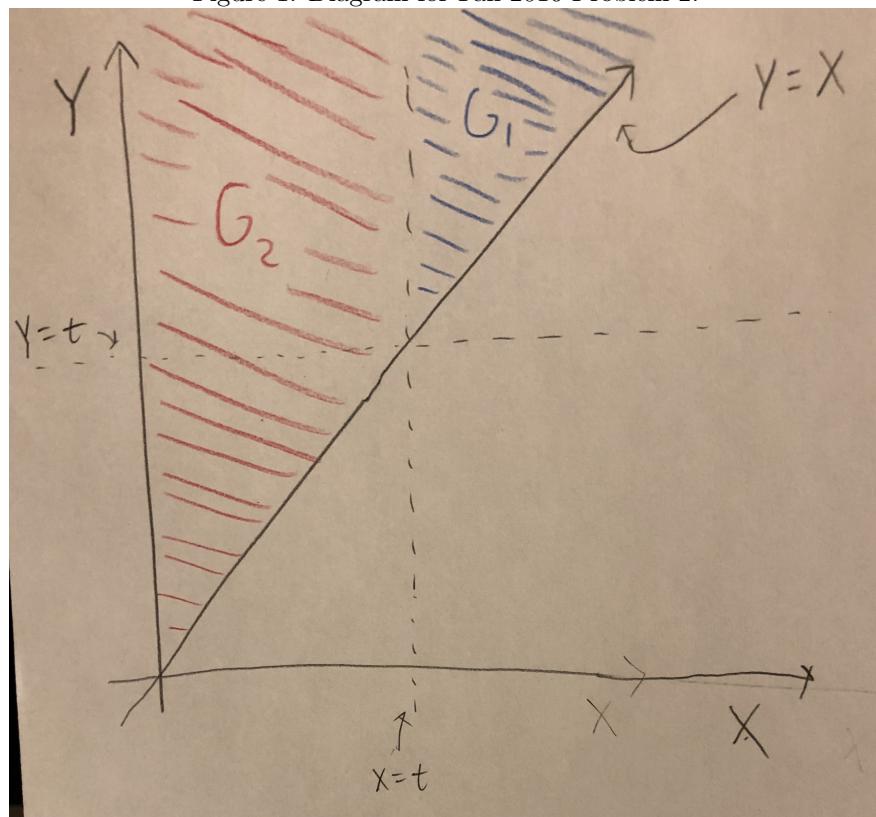
$$\begin{aligned} &= \int_0^\infty \int_x^\infty e^{-x-y} dy dx = \int_0^\infty e^{-x} \int_x^\infty e^{-y} dy dx \\ &\quad \int_x^\infty e^{-y} dy = -e^{-y}|_x^\infty = e^{-x} \end{aligned}$$

$$\implies \Pr(X < Y) = \int_0^\infty e^{-2x} dx = -\frac{1}{2}e^{-2x}|_0^\infty = \boxed{\frac{1}{2}}$$

- $\Pr(X < Y \cap \min\{X, Y\} > t)$ : Note that in Figure 2, the region that satisfies this condition is region  $G_1$ . Therefore we can find this probability by integrating the joint pdf over that region.

$$\begin{aligned} \Pr(X < Y \cap \min\{X, Y\} > t) &= \iint_{G_1} f_{X,Y}(x, y) dxdy = \int_t^\infty \int_t^y e^{-x-y} dx dy = \int_t^\infty e^{-y} \int_t^y e^{-x} dx dy \\ &\quad \int_t^y e^{-x} dx = -e^{-x}|_t^y = -e^{-y} + e^{-t} \\ \implies \Pr(X < Y \cap \min\{X, Y\} > t) &= \int_t^\infty e^{-y} (-e^{-y} + e^{-t}) dy = \int_t^\infty (e^{-t-y} - e^{-2y}) dy \\ &= \frac{1}{2}e^{-2y} - e^{-t}e^{-y}|_t^\infty = -\frac{1}{2}e^{-2t} - e^{-2t} = \boxed{\frac{1}{2}e^{-2t}} \end{aligned}$$

Figure 1: Diagram for Fall 2016 Problem 2.



Note that

$$\Pr(X < Y \cap \min\{X, Y\} > t) = \frac{1}{2} \cdot e^{-2t} = \Pr(X < Y) \Pr(\min\{X, Y\} > t)$$

Therefore the events  $\{\omega : \min\{X, Y\} > t\}$  and  $\{\omega : X < Y\}$  are independent for every  $t > 0$ .

- (3) In a certain area, earthquakes happen at a frequency of one every four days. What is the probability that more than 100 earthquakes will occur in this area in one year (365 days)?

**Solution.** We can think of this as a Poisson process (see section 7.2) with  $\lambda = 1/4$ . Then there are two ways to obtain the answer: we can either examine the number of earthquakes in a 365 day period  $N(365)$  and find the probability that  $N(365) > 100$ , or we can examine the number of days until the 101st earthquake  $T_{101}$  and find the probability that  $T_{101} < 365$ .

- (i) **Number of earthquakes in 365 days:** Let  $N(t)$  be the number of earthquakes that occur in  $t$  days after the start of this process. By Theorem 7.2.3,  $N(t) \sim \text{Poisson}(t \cdot 1/4)$ . Then

$$\Pr(N(t) > 100) = \sum_{j=101}^{\infty} \frac{(365 \cdot 1/4)^j \exp(-365 \cdot 1/4)}{j!}$$

To obtain an answer for this, we can use the normal approximation to a Poisson distribution (Proposition 6.1.35):

$$\begin{aligned} N(t) \sim \mathcal{N}(t/4, t/4) \implies \Pr(N(365) > 100) &\approx \Pr\left(\mathcal{N}(0, 1) > \frac{100.5 - 365/4}{\sqrt{365/4}}\right) \\ &= \Pr\left(\mathcal{N}(0, 1) > \frac{100.5 - 91.25}{\sqrt{91.25}}\right) \approx \Pr\left(\mathcal{N}(0, 1) > \frac{9.25}{9.1}\right) \approx \boxed{0.1664} \end{aligned}$$

- (ii) **Number of days before 100th earthquake:** Let  $T_n$  be the number of days until the  $n$ th earthquake happens. By Corollary 7.2.4.1,  $T_n \sim \text{Gamma}(n, 4)$ . Then

$$\Pr(T_{101} < 365) = \int_0^{365} \frac{1}{\Gamma(101, 4)} x^{101-1} e^{-x/4} dx$$

To obtain an answer for this, we can use the normal approximation to a Gamma distribution (Proposition 6.3.19):

$$\begin{aligned} T_n \sim \mathcal{N}(404, 1616) \implies \Pr(T_{101} < 365) &\approx \Pr\left(\mathcal{N}(0, 1) < \frac{365 - 404}{\sqrt{1616}}\right) \approx \Pr\left(\mathcal{N}(0, 1) < \frac{-40}{40}\right) \\ &= \Pr(\mathcal{N}(0, 1) < -1) \approx \boxed{0.1660} \end{aligned}$$

## 6.6.2 More Problems From Homework

### Homework 5 Problem 4.

Let  $X_1, X_2, \dots$  be i.i.d. having moment-generating functions  $M_X = M_X(t), t \in (-\infty, \infty)$ . Let  $N$  be an integer-valued random variable with moment-generating function  $M_N = M_N(t), t \in (-\infty, \infty)$ . Assume that  $N$  is independent of all  $X_k$  and define  $S = \sum_{k=1}^N X_k$ . Confirm that the random variable  $S$  has the moment-generating function  $M_S = M_S(t)$  defined for all  $t \in (-\infty, \infty)$  and

$$M_S(t) = M_N(M_X(t))$$

Then use the result to derive the formulae

$$\mathbb{E}(S) = \mu_N \mu_X, \text{Var}(S) = (\sigma_N^2 - \mu_N) \mu_X^2 + \mu_N \sigma_X^2$$

where  $\mu_N = \mathbb{E}(N)$ ,  $\mu_X = \mathbb{E}(X_1)$ ,  $\sigma_N^2 = \text{Var}(N)$ , and  $\sigma_X^2 = \text{Var}(X_1)$ . How will the above computations change if we use the characteristic function  $\phi_X$  instead of the moment-generating function  $M_X$ ?

**Solution.**

$$M_S(t) = \mathbb{E}(e^{tS}) = \mathbb{E}[\mathbb{E}(e^{tS} \mid N)] = \sum_{n=0}^{\infty} \mathbb{E}(e^{tS} \mid N = n) \Pr(N = n) = \sum_{n=0}^{\infty} \mathbb{E}(e^{t(X_1+X_2+\dots+X_n)} \mid N = n) \Pr(N = n)$$

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{tX_1} e^{tX_2} \cdots e^{tX_n}) \Pr(N = n)$$

By independence of the  $X_i$  we have

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{tX_1}) \mathbb{E}(e^{tX_2}) \cdots \mathbb{E}(e^{tX_n}) \Pr(N = n)$$

which, since the  $X_i$  are i.i.d., can be written as

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{tX_1})^n \Pr(N = n) = \sum_{n=0}^{\infty} (M_X(t))^n \Pr(N = n)$$

But since  $G_N(s) = \mathbb{E}(s^N) = \sum_{n=0}^{\infty} s^n \Pr(N = n)$ , this can be written as

$$M_S(t) = G_N(M_X(t))$$

as desired. Note that

$$M'_S(t) = G'_N(M_X(t)) M'_X(t)$$

$$M''_S(t) = G''_N(M_X(t))(M'_X(t))^2 + G'_N(M_X(t))M''_X(t)$$

So we have

- $\mathbb{E}(S) = M'_S(0) = G'_N(M_X(0)) M'_X(0) = G'_N(1) \mathbb{E}(X_1) = \mathbb{E}(N) \mathbb{E}(X_1) = \mu_N \mu_X$

- $\text{Var}(S) = \mathbb{E}(S^2) - \mathbb{E}(S)^2 = M_S''(0) - (M_S'(0))^2$

$$= G_N''(M_X(0))(M_X'(0))^2 + G_N'(M_X(0))M_X''(0) - \mu_N^2\mu_X^2 = G_N''(1)\mathbb{E}(X_1)^2 + G_N'(1)\text{Var}(X_1) - \mu_N^2\mu_X^2$$

$$= \mathbb{E}[N(N-1)]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2 = \mathbb{E}[N^2 - N]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2$$

$$= [\mathbb{E}(N^2) - \mathbb{E}(N)^2 + \mathbb{E}(N)^2 - \mathbb{E}(N)]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2 =$$

$$= [\text{Var}(N) + \mathbb{E}(N)^2 - \mathbb{E}(N)]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2 = (\sigma_N^2 + \mu_N^2 - \mu_N)\mu_X^2 + \mu_N\sigma_X^2 - \mu_N^2\mu_X^2$$

$$= \boxed{(\sigma_N^2 - \mu_N)\mu_X^2 + \mu_N\sigma_X^2}$$

To use the characteristic function  $\phi_X$  instead of the moment generating function  $M_X$ , we would do the following:

$$\begin{aligned} \phi_S(t) &= \mathbb{E}(e^{itS}) = \mathbb{E}[\mathbb{E}(e^{itS} \mid N)] = \sum_{n=0}^{\infty} \mathbb{E}(e^{itS} \mid N = n) \Pr(N = n) = \sum_{n=0}^{\infty} \mathbb{E}(e^{it(X_1 + X_2 + \dots + X_n)} \mid N = n) \Pr(N = n) \\ &= \sum_{n=0}^{\infty} \mathbb{E}(e^{itX_1} e^{tX_2} \cdots e^{itX_n}) \Pr(N = n) \end{aligned}$$

By independence of the  $X_i$  we have

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{itX_1}) \mathbb{E}(e^{itX_2}) \cdots \mathbb{E}(e^{itX_n}) \Pr(N = n)$$

which, since the  $X_i$  are i.i.d., can be written as

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{itX_1})^n \Pr(N = n) = \sum_{n=0}^{\infty} (\phi_X(t))^n \Pr(N = n)$$

But since  $G_N(s) = \mathbb{E}(s^N) = \sum_{n=0}^{\infty} s^n \Pr(N = n)$ , this can be written as

$$\phi_S(t) = G_N(\phi_X(t))$$

### Homework 5 Problem 7.

- (a) Let  $X_1, X_2, \dots, X_n$  be independent with mean zero and finite third moment. Prove that

$$\mathbb{E}(X_1 + \dots + X_n)^3 = \mathbb{E}X_1^3 + \dots + \mathbb{E}X_n^3$$

**Solution.**

- (a) Let  $\mathbb{E}(\exp(it_i X_i)) = \phi_{X_i}(t_i)$ . Let  $S_n = \sum_{i=1}^n X_i$ . Then by independence the characteristic function for  $S_n$  is

$$\mathbb{E}(\exp(itS_n)) = \phi_{S_n}(t) = \prod_{i=1}^n \phi_{X_i}(t)$$

Then

$$\mathbb{E}(X_1 + X_2 + \dots + X_n)^3 = \mathbb{E}(S_n^3) = \phi_{S_n}^{(3)}(0)$$

$$= \sum_{i=1}^n \phi_{X_i}^{(3)}(0) \cdot \left( \prod_{j \in \{1, \dots, n\}, j \neq i} \phi_{X_j}(0) \right) + C \left[ \sum_{i=1}^n \left( \sum_{j \in \{1, \dots, n\}, j \neq i} \phi_{X_i}^{(2)}(0) \phi_{X_j}^{(1)}(0) \right) \cdot \left( \prod_{k \in \{1, \dots, n\}, k \neq i, j} \phi_{X_k}(0) \right) \right]$$

where  $C$  is some coefficient resulting from the multinomial expansion of  $S_n$  after repeated differentiation product rules. But because  $\mathbb{E}(X_i) = 0$ ,  $\phi_{X_i}^{(1)}(0) = 0 \forall i$ , so the second term goes to 0. Therefore we have

$$\mathbb{E}(X_1 + X_2 + \dots + X_n)^3 = \sum_{i=1}^n \phi_{X_i}^{(3)}(0) \cdot \left( \prod_{j \in \{1, \dots, n\}, j \neq i} \phi_{X_j}(0) \right) = \sum_{i=1}^n \mathbb{E}(X_i^3) \cdot 1^{n-1} = \sum_{i=1}^n \mathbb{E}(X_i^3)$$

as desired.

### Homework 6 Problem 10.

- (a) For  $p \in (0, 1)$ , let  $x(p)$  be the smallest number of people so that there is a better than  $100 \cdot p\%$  chance to have at least two born on the same day. Find an approximate expression for  $x(p)$ , and sketch the graph of the function  $x = x(p)$ .
- (b) Repeat part (a) when you want at least three people to share a birthday.

**Solution.**

- (a) Let  $f(x)$  be the probability of no matches in birthdays in a group of  $x$  people; that is,

$$f(x) = \frac{365 \cdot 364 \cdot 363 \cdots (365 - x + 1)}{365^x} = \frac{1}{365^x} \cdot \frac{365!}{(365 - x)!} = \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{x-1}{365}\right)$$

Using the first order Taylor approximation  $\exp(-k/x) \approx 1 - k/x$ , we have

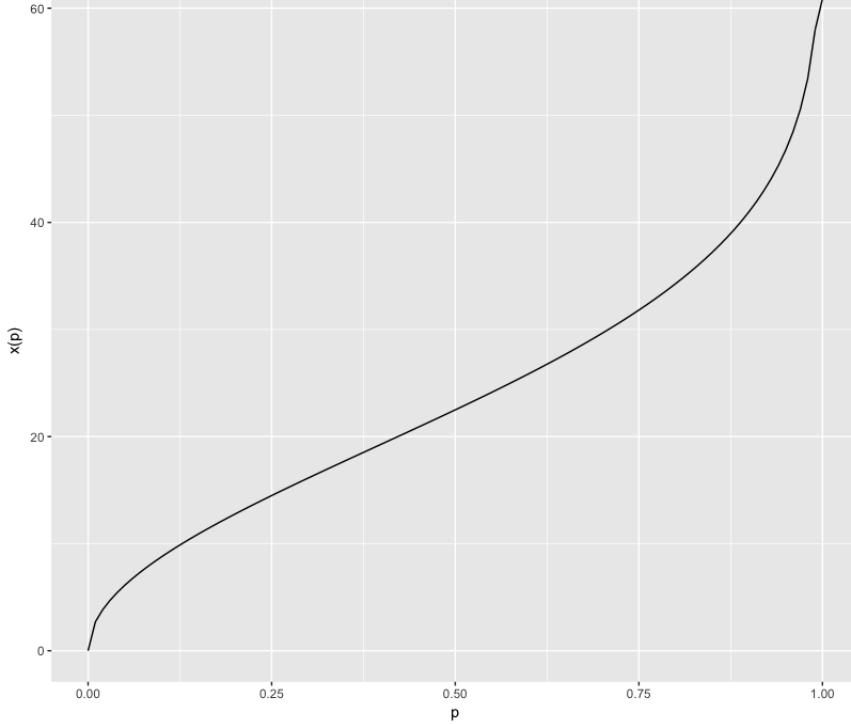
$$\begin{aligned} f(x) &= \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{x-1}{365}\right) \approx \exp(-1/365) \exp(-2/365) \cdots \exp(-(x-1)/365) \\ &= e^{-(x^2-x)/(2 \cdot 365)} \end{aligned}$$

We want the probability of a match to be at least  $p$ ; that is,  $f(x) \leq 1 - p$ . Setting this equal to  $q = 1 - p$ , we have

$$e^{-(x^2-x)/(2 \cdot 365)} = q \iff -\frac{x^2-x}{730} = \log(q) \iff x^2 - x + 730 \log(q) = 0$$

$$\implies x = 0.5 + \sqrt{1/4 + 730 \log(1/q)} \approx \boxed{\sqrt{2 \cdot 365 \log(1/q)}}$$

where we discard the negative root because we have to have a nonnegative number of people, and we don't worry about the decimals since this is an approximation and we have to round up to the nearest whole person anyway.



- (b) For a group of three people, the Poisson approximation (see Section 6.1.10) is more convenient. The number of groups of 3 people in a room of  $x$  people is  $\binom{x}{3}$ . For a group of three people, the probability that all three have the same birthday is  $1 \cdot 1/365 \cdot 1/365 = 365^{-2}$ . Therefore we can think of the number of matches of three people as distributed Poisson with expectation  $\binom{x}{3} \cdot 365^{-2}$ . Then we have the probability of at least one “success” (triplet with three matched birthdays) is

$$1 - \frac{\exp(-\lambda)\lambda^0}{0!} = 1 - \exp\left(-\binom{x}{3} \cdot 365^{-2}\right)$$

We set this equal to  $p$  and solve:

$$\begin{aligned} p = 1 - \exp\left(-\binom{x}{3} \cdot 365^{-2}\right) &\iff -\binom{x}{3} \cdot 365^{-2} = \log(1-p) \iff \frac{x!}{(x-3)!3!} = 365^2 \cdot \log\left(\frac{1}{1-p}\right) \\ &\iff x(x-1)(x-2) = 6 \cdot 365^2 \cdot \log\left(\frac{1}{1-p}\right) \iff (x^2-x)(x-2) = x^3 - 3x^2 + 2x = 6 \cdot 365^2 \cdot \log\left(\frac{1}{1-p}\right) \end{aligned}$$

This has a unique real solution, but it is hard to find.

**Exercise 5.** Let  $X, Y, Z$  be independent uniform on  $(0, 1)$ . Compute the cdfs of  $XY$ ,  $X/Y$ , and  $XY/Z$ .

**Solution.**

Using the information from part (a), and the fact that  $f_X(x) = 1$  (for  $x \in [0, 1]$ ) and likewise for  $f_Y(y)$ :

- $XY$ :

$$\begin{aligned} F_{XY}(z) &= \int_0^\infty f_X(x) \int_{-\infty}^{z/x} f_Y(y) dy dx - \int_{-\infty}^0 f_X(x) \int_\infty^{z/x} f_Y(y) dy dx \\ &= \int_0^1 [(z/x)\mathbf{1}_{\{0 < z/x \leq 1\}} + \mathbf{1}_{\{z/x > 1\}}] dx = \int_0^1 [(z/x)\mathbf{1}_{\{z \leq x\}} + \mathbf{1}_{\{z > x\}}] dx = \int_0^z dx + \int_z^1 (z/x) dx \\ &= z + z \log(x) \Big|_z^1 = z + z \log(1) - z \log(z) = z(1 - \log(z)) \end{aligned}$$

$$\Rightarrow F_{XY}(z) = \begin{cases} 0 & z \leq 0 \\ z(1 - \log(z)) & 0 < z \leq 1 \\ 1 & z > 1 \end{cases}$$

- $X/Y$ :

$$\begin{aligned} F_{X/Y}(z) &= \int_0^\infty f_Y(y) \int_{-\infty}^{zy} f_X(x) dx dy - \int_{-\infty}^0 f_Y(y) \int_\infty^{zy} f_X(x) dx dy \\ &= \int_0^1 [zy\mathbf{1}_{\{0 < zy \leq 1\}} + \mathbf{1}_{\{zy > 1\}}] dy = \int_0^1 [zy\mathbf{1}_{\{y > 0 \cap y \leq 1/z\}} + \mathbf{1}_{\{y > 1/z\}}] dy = \int_0^{1/z} zy \cdot dy + \int_{1/z}^1 dy \\ &= \frac{zy^2}{2} \Big|_0^{1/z} + (1 - 1/z) = \frac{z}{2z^2} + 1 - \frac{2}{2z} = 1 - \frac{1}{2z} \\ \Rightarrow F_{X/Y}(z) &= \begin{cases} 0 & z \leq 0 \\ 1 - \frac{1}{2z} & 0 < z \leq 1 \\ 1 & z > 1 \end{cases} \\ &= \begin{cases} 0 & z \leq 0 \\ z/2 & 0 < z \leq 1 \\ 1 - \frac{1}{2z} & z > 1 \end{cases} \end{aligned}$$

- $XY/Z$ : Consider this the cdf of the quotient of  $W = XY$  and  $Z$ .

$$\begin{aligned}
F_U(u) &= \int_0^\infty f_Z(z) \int_{-\infty}^{uz} f_W(w) dw dz - \int_{-\infty}^0 f_Z(z) \int_\infty^{uz} f_W(w) dw dz \\
&= \int_0^1 \int_0^{uz} -\log(w) \mathbf{1}_{\{0 < uz \leq 1\}} dw dz = \int_0^1 -[w \log(w) - w]_0^{uz} \mathbf{1}_{\{0 < z \leq 1/u\}} dz \\
&= \int_0^{1/u} uz [1 - \log(uz)] dz = \frac{u}{4} z^2 (3 - 2 \log(uz)) \Big|_0^{1/u} = \frac{u}{4u^2} (3 - 2 \log(1)) - 0 = \frac{3}{4u} \\
\implies F_{XY/Z}(u) &= \begin{cases} 0 & u \leq 0 \\ \frac{3}{4u} & 0 < u \leq 3/4 \\ 1 & u > 3/4 \end{cases}
\end{aligned}$$

## 6.7 Random Matrix Theory

**Definition 6.42 (Wigner-type random matrix).**  $H \in \mathbb{R}^{N \times N}$  is a **Wigner-type random matrix** if each entry  $H_{ij}$  equals either 1 or  $-1$  with equal probability and  $H$  is symmetric ( $H = H^+$ ). Other than the symmetry restriction, the entries are independent.

**Remark.** Because a Wigner-type random matrix is symmetric, the eigenvalues are real.

**Example 6.13.** 10 i.i.d. random variables  $X_1, \dots, X_{10}$ , where  $\Pr(X_1 = m) = 1/3$  for  $m \in \{2, 3, 4\}$ . What is the distribution of  $S = \sum_k X_k$ ?

This is not an interesting problem since a computer can easily calculate the answer, even if the number of random variables is in the thousands.

**Example 6.14.**  $n$  i.i.d. random variables  $X_1, \dots, X_n$ , where  $n$  is large. Find  $f(n)$  and  $g(n)$  such that

$$Z_n = \frac{S_n - f(n)}{g(n)}$$

has a nontrivial limiting distribution.

### Solution.

We know if  $\mathbb{E}X_k^2 < \infty$ , choosing  $f(n) = n\mathbb{E}X_1$  and  $g(n) = \sqrt{n\text{Var}(X_1)}$  allows convergence to a standard Gaussian random variable under the Central Limit Theorem. Also note that if  $n$  is large,  $S_n \approx n\mathbb{E}X_1$  and  $(S_n - n\mathbb{E}X_1) = \mathcal{O}(\sqrt{n\text{Var}(X_1)})$ .

Let  $\hat{S}_n$  be the largest eigenvalue of a Wigner random matrix  $H_n$ . It turns out that  $\hat{S}_n \approx 2\sqrt{n}$  and  $\hat{S} - 2\sqrt{n} = \mathcal{O}(n^{1/6})$ . Lastly,

$$\frac{\hat{S} - 2\sqrt{n}}{n^{1/6}}$$

converges in distribution to a Tracy-Widom distribution [Johnstone, 2001].

**Exercise 6.** Suppose  $X_1, \dots, X_n$  are i.i.d. uniform on  $[0, 1]$ . Let  $Y_1, \dots, Y_n$  be the order statistics  $X_{(1)}, \dots, X_{(n)}$ . Let  $Z_n = Y_{n/2+1} - Y_{n/2}$ . Find  $f(n), g(n)$  such that

$$\frac{Z_n - f(n)}{g(n)} \xrightarrow{d} \text{non-trivial distribution.}$$

**Example 6.15.** Let  $X_n = \sum_{k=1}^n X_k$  where  $X_k \sim \text{Ber}(1/2)$ . We know that

$$\frac{S_n - n/2}{1/2\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

However, what is  $\mathbb{P}(S_n > 0.6n)$ ? It turns out that  $\mathcal{P}(S_n > 0.6n) \sim e^{-cn}$ .

For a Gaussian Orthogonal Ensemble, eigenvalues tend to be between -2 and 2. Difference between adjacent eigenvalues  $\lambda_i - \lambda_{i+1}$  is of interest. Tends to be of order  $1/n$ , so after multiplying by  $n$ , tends to be of order 1.

$$n(\lambda_i - \lambda_{i+1}) \xrightarrow{d} \text{Gaudin-Mehta distribution.}$$

Approximately given by Wigner's

$$\mathbb{P}(s) = \frac{\pi s}{2} e^{-\frac{\pi}{4}s^2}.$$

Bulk Universality Conjecture: Gap distribution for the bulk eigenvalues for a large generic  $d$ -regular graph converges to this distribution.

**Universality Conjecture (Wigner-Dyson-Mehta):** bulk eigenvalue statistics are universal, depending only on the symmetry class of the matrix ensemble, regardless of the distribution of individual matrix entries. Proven by Erdos, Yau, and Tau and Vu (under 4th moment condition)

### 6.7.1 Large Deviation Theory

Wigner-type matrix  $H$ . Let the eigenvalues be  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ . Suppose we normalize  $H$  by dividing it by the square root of  $n$ ; that is, let  $H_1 = H/\sqrt{n}$ . Define a measure

$$\mu_n := \frac{1}{n} \sum_{k=1}^n \delta_{\lambda_k}$$

so for  $I \subset \mathbb{R}$ ,

$$\mu_n(I) = \#\{i : \lambda_i \in I\}.$$

Note that  $\mu_n$  is random because  $H$  is random. We will show that

$$\mu_n \xrightarrow{a.s.} \mu.$$

**Proposition 6.7.1.** For any continuous smooth function  $g$ ,

$$\int g(t) d\mu_n(t) \xrightarrow{a.s.} \int g(t) d\mu(t).$$

We can also write this as

$$\langle \mu_n, g \rangle \xrightarrow{a.s.} \langle \mu, g \rangle.$$

In other words, for any interval  $[a, b] \subset \mathbb{R}$ ,

$$\mu_n([a, b]) \xrightarrow{a.s.} \mu([a, b]).$$

where

$$d\mu = \frac{1}{2\pi} \sqrt{(4 - x^2)_+} dx$$

where

$$(4 - x^2)_+ = \begin{cases} 4 - x^2 & 4 - x^2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Roughly speaking, for a Wigner random matrix where  $\mathbb{E} H_{ij} = 0$  and  $\mathbb{E} H_{ij}^2 = 1$  and  $\mathbb{E} H_{ij}^p$  exists and is finite for all  $p$ , if  $n$  is large then the number of eigenvalues of  $H_n$  in the interval  $[a, b]$  divided by  $n$  is close to the number

$$\int_a^b \frac{1}{2\pi} \sqrt{(4 - x^2)_+} dx.$$

Let  $\lambda_n^{(n)}$  be the largest eigenvalue of  $H_n$ . We have that

$$\frac{\lambda_n^{(n)} - 2}{n^{2/3}} \xrightarrow{d} \text{Tracy-Widom distribution.}$$

**Proposition 6.7.2.** Suppose  $k = \lfloor n/2 \rfloor$ . Let  $\lambda_k^{(n)}$  be the  $k$ -th smallest eigenvalue of  $H_n$ .  $\gamma_k$  is defined as

$$\int_{-2}^{\gamma_k} \frac{1}{2\pi} \sqrt{(4 - x^2)_+} = \frac{k}{n}$$

Then

- (1)  $\lambda_k^{(n)} \approx \gamma_k$
- (2)  $\lambda_k^{(n)} - \gamma_k = \mathcal{O}(\frac{1}{2}\sqrt{\log n})$ .
- (3)

$$\frac{\lambda_k^{(n)} - \gamma_k}{n^{-1}\sqrt{\log n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

These results hold regardless of the exact distribution of  $H_{ij}$ . This property is referred to as **universality**. (Like the Central Limit Theorem.)

**Proposition 6.7.3.** Fix  $a, b \in \mathbb{R}$ ,  $E_0 \in \mathbb{R}$ .

$$I_{E_0, a, b} = [E_0 + a/n, E_0 + b/n].$$

Then the number of eigenvalues of  $H$  in  $I_{E_0, a, b}$  is  $\mathcal{O}(1)$ . (Regardless of the exact distribution of  $H_{ij}$ .)

**Proposition 6.7.4.** Let  $\lambda_k^{(n)}$  be an eigenvalue of  $H_n$ . Let  $u_k^{(n)}$  be the corresponding unit-length eigenvector; that is,

$$Hu_k^{(n)} = \lambda_k^{(n)} u_k^{(n)}, \quad \|u_k^{(n)}\|_2 = 1.$$

Then

1. For all fixed  $w \in \mathbb{R}^n$  such that  $\|wrVert_2 = 1$ ,

$$\frac{|\langle w, u_k^{(n)} \rangle|^2}{1/n} \xrightarrow{d} (\mathcal{N}(0, 1))^2.$$

2. Let  $g_k^w := |\langle w, u_k^{(n)} \rangle|^2$ . Then for  $k \neq \ell$ ,

$$(g_k^w, g_\ell^w) \xrightarrow{d} ([\mathcal{N}(0, 1)]^2, [\mathcal{N}(0, 1)]^2)$$

and the two vectors are independent.

3. Let  $w, w'$  be two orthogonal unit vectors ( $\langle w, w' \rangle = 0$ ). Then

$$(g_k^w, g_k^{w'}) \xrightarrow{d} ([\mathcal{N}(0, 1)]^2, [\mathcal{N}(0, 1)]^2)$$

and the two vectors are independent.

Let  $X$  have i.i.d. entries but it is not necessarily symmetric. We will investigate the eigenvalues of  $X^T X$ , which converge to a Marchenko–Pastur distribution.

Let  $X$  have eigenvalues  $\lambda_k \in \mathbb{C}$ . Let  $\mu_n = n^{-1} \sum_{k=1}^n \delta_{\lambda_k}$ . Then  $n\mu_n(A)$  is the number of eigenvalues of  $X$  in  $A$  for  $A \subset \mathbb{C}$ .

$$\mu(A) = \frac{\text{Area}(A \cap D)}{\pi}, \quad D := \{z : |z| \leq 1\}.$$

This is called a **circular law** (density  $1/\pi$  in  $D$ , 0 outside of  $D$ ). Results exist if  $\mathbb{E}X_i = 0$ ,  $\mathbb{E}X_{ij}^2 = 1$ .

**Proposition 6.7.5.**  $z \in \mathbb{C}$ ,  $B_n = z^\ell$ ,  $\ell = N^{-1/2+\delta}$ , then the number of eigenvalues of  $X$  in  $B_n$  is roughly  $\pi^{-1}\ell^2 n$ . where  $B_n$  is a box with width  $\ell$ .

**Proposition 6.7.6.** Suppose  $u_k$  is an eigenvector of  $X$  with  $\|u_k\|_2 = 1$ . Let  $u_k(m)$  be the  $m$ th component of  $X$ . Note  $\|u_k\|_2^2 = 1$ ,  $u_k$  has  $n$  components. With high probability,

$$\max_m |u_k(m)|^2 \leq n^{-1+\epsilon} \quad \forall \epsilon > 0.$$

**Delocalization:** eigenvectors are spread out fairly uniformly over a range of possible values.

Opposite case: Suppose we have a diagonal random matrix  $\tilde{X}_{ij}$ . Then the eigenvectors are the unit vectors  $e_i$ , so the eigenvectors are very localized.

Some other Wigner-type random matrices:

- (1)  $H$ ,  $\mathbb{E}H_{ij} = 0$ ,  $\mathbb{E}H_{ij}^2 = \mathcal{O}(1)$ ,  $\sum \mathbb{E}(H_{ij})^2 = N$ . This is a generalized Wigner random matrix.
  - (2) Remove the condition that  $\sum \mathbb{E}(H_{ij})^2 = N$ . This is a **universal Wigner matrix**. In this case, the eigenvectors may have some favored direction.
  - (3)  $\mathbb{E}H_{ij}^2 = 1$  but mean is not 0. Then similar results.
  - (4) Suppose  $X$  is Wigner and  $T$  is a fixed matrix. Then there are results for  $TX$ .
  - (5)  $H_{ij}$  does not have order 1 all the time; that is, some entries are on different scales than others (some much larger). Examples:
    - (a) Sparse matrix, like Erdos-Renyi matrix. E-R graph: line connecting  $V_i$ ,  $V_j$  if and only iff  $A_{ij} = 1$ .  $\mathbb{P}(A_{ij} = 1) = p(n)$ , some function of  $n$ . All  $A_{ij}$  are independent of each other besides the fact that  $A$  is symmetric.
- $A = H + A_0$ , where  $A_{ij}$  is defined as above,  $\mathbb{E}H_{ij} = 0$ ,  $\mathbb{E}H_{ij}^2 = p_n(1 - p_n)$ . Note that if  $p_n \ll 1$ ,

⋮

we have

$$\text{Var} \left( \frac{1}{\sqrt{p_n}} A_{ij} \right) \approx 1, \quad \mathbb{E} \left( \frac{1}{\sqrt{p_n}} A_{ij} \right) \approx 0,$$

but  $\hat{A} = p_n^{-1/2} A$ ,  $\hat{A}_{ij} \gg 1$ ; that is, this is a sparse matrix with mostly 0 entries, and the entries that are nonzero are large.

- (b) **Band matrix.**  $H_{ij} \cdot f(|i - j|)$ , where

$$f(m) = \begin{cases} 1 & m \leq W \\ 0 & m > W. \end{cases}$$

The interesting case is when  $W \ll n$ .

Topics:

- **Topic 1:** For a matrix  $A$ , find optimal  $m$  and  $M$  such that for all  $x$ ,  $m\|x\| < \|Ax\| < M\|x\|$  (i.e.,  $m$  is the smallest singular value and  $M$  is the largest). Then the condition number is  $\kappa(A) = M/m$ .

Suppose  $A \in \mathbb{R}^{n \times m}$ ,  $a_{ij} = g_{ij}$ ,  $g_{ij} \sim \mathcal{N}(0, 1)$ . Marchenko–Pastur Law:

$$\|A\| \approx \sqrt{n} + \sqrt{m}$$

$$\sigma_{\min}(A) \approx \sqrt{n} - \sqrt{m}$$

If  $A$  is a tall matrix,  $n \gg m$ , then  $\kappa(A)$  is constant. In square case,  $\sigma_{\min}(A) < 0 \iff A$  is invertible.

- **Least Singular Value of square matrices:** Intuition: eigenvector should not have a preference—“uniform on the sphere.” True for Gaussian matrices.

Typical  $X \sim \text{Uniform}(S^{n-1})$ :

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

$$\mathbb{E}|X_i|^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E}|X_j|^2 = \frac{1}{n}$$

where we used the fact that  $X$  is a unit vector. Therefore  $|X_i| \sim n^{-1/2}$ .

- $\|X\|_\infty = \max_j |X_j|$ :  
 $\mathbb{P}(|X_i| > t/\sqrt{n}) \leq \exp(-C_1 t^2)$  for  $t > C_1$ .

$$\mathbb{P}\left(|X_i| > \frac{R\sqrt{\log n}}{\sqrt{n}}\right) \leq \exp(-CR^2 \log n)$$

$$\mathbb{P}\left(\exists j \text{ such that } |X_j| > \frac{R\sqrt{\log n}}{\sqrt{n}}\right) < n \exp(-CR^2 \log n)$$

$$\|X\|_\infty \leq R\sqrt{\frac{\log n}{n}} \text{ with probability } \geq 1 - n^{1- CR^2}.$$

- **Topic 2:**  $\ell_\infty$  delocalization. Suppose we have  $A$  with  $\mathbb{E}a_{ij} = 0$ ,  $\mathbb{E}A_{ij}^2 = 1$ . Eigenvector  $v$ ,  $\|v\|_2 = 1 \implies \|v\|_\infty < \frac{\log C}{\sqrt{n}}$ .
- **Topic 3:** No-gap delocalization of eigenvectors.

$$X = \begin{pmatrix} \sqrt{2/n} \\ \vdots \\ \sqrt{2/n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

(first  $n/2$  entries are  $\sqrt{2/n}$ , rest are 0.)

For every  $I \subset \{1, \dots, n\}$  with  $|I| = \epsilon n$ , let  $I$  index over the nonzero entries of an eigenvector. Want to show that  $\|V_i\|_2 \geq \epsilon^c$ .

- **Topic 0:** Subgaussian random matrices

Linear algebra:  $A \in \mathbb{R}^{n \times m}$ . Singular value decomposition:

$$A = U\Sigma V^T$$

where  $UU^T = I_n$ ,  $VV^T = I_m$ , and  $\Sigma$  is a diagonal matrix with diagonal entries equal to the singular values.

### 6.7.2 Band Matrix

$n \times n$  random matrix with bandwidth  $w$ .  $H_{ij} \neq 0 \implies |i - j| \leq w$ . Let  $w = n^a$  for  $0 < a < 1$ . We have for the nonzero entries,  $H_{ij} = \pm 1$ ,  $\mathbb{E}H_{ij} = 0$ ,  $\mathbb{E}H_{ij}^2 = 1$ .

Let  $H$  be a Wigner band matrix. Since  $H$  is symmetric, decompose  $H$  as  $H = UDU^+$ , where  $D$  is diagonal. Then  $H^m = U D^m U^+$ . When  $H$  is a band matrix with reasonably small  $w$ , then for any eigenvector  $u$  of  $H$ , usually  $|u(k)|^2$  is contained within a narrow range.

Roughly speaking, for each eigenvector  $u_k$  of  $H$  there exists an integer  $a_k$ , where  $1 \leq a_k \leq n$ , such that  $u(k) \approx 0$  if  $|k \cdot a| \gg w^2$ .

From linear algebra: if  $\lambda_\ell$  and  $u_\ell$  are corresponding eigenvalues and eigenvectors of  $H$ ,

$$[H^m]_{bc} = \sum_\ell \lambda_\ell^m u_\ell(b) u_\ell^+(c).$$

As you raise the matrix to higher and higher powers, the bandwidth will initially increase but eventually won't.

$$u_\ell(b) \neq 0, u_\ell(c) \neq 0 \implies |b - c| = o(w^2).$$

This is so-called **localization** of a band matrix. **Conjecture:** this occurs for  $w = n^a$  when  $a < 1/2$ . When  $a > 1/2$  we have delocalization.

For  $d = 3$ , a band matrix is delocalized as long as  $a > 0$ .

1. **GOE/GUE:** Wigner-type random matrix (e.g.,  $\mathbb{E}H_{ij} = 0$ ,  $\mathbb{E}H_{ij}^2 = 1$ ). If  $H_{ij} \sim \mathcal{N}(0, 1)$ , then it's called GOE (Gaussian orthogonal \*). If  $H$  is GOE, then for any orthogonal matrix  $O$  ( $OO^+ = I$ ), then  $OHO^+$  has the same distribution as  $H$ . Also, the probability density of  $\lambda_1, \dots, \lambda_n$  of  $H$  is

$$\prod e^{-\sum_k 1/4\lambda_k^2} \prod_{i < j} |\lambda_i - \lambda_j|.$$

this is an **orthogonal polynomial**.

Recall important properties of Wigner matrix: independent entries, orthogonal invariance. For a symmetric matrix, the only solution is a GOE matrix. For Hermitian, must be GUE.

2. **Comparison Method:** unknown  $X$ , known  $\tilde{X}$ , for Wigner,  $\mathbb{E}\lambda_n^{(n)} \approx \mathbb{E}_n^{(n)}$ .

Look at Central Limit Theorem again:  $X_i$  i.i.d.,  $\mathbb{E}X_i = 0$ ,  $\mathbb{E}X_i^2 = 1$ , then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \xrightarrow{d} \mathcal{N}(0, 1).$$

Special case: if  $X_i \sim \mathcal{N}(0, 1)$ , result holds exactly even in finite samples.

Need to prove:

$$\mu_n \rightarrow \mathcal{N}(0, 1) = \mu$$

Need to prove: for all  $f \in C_{b,c}^\infty$  (smooth functions),  $\langle f, \mu_n \rangle \rightarrow \langle f, \mu \rangle$ . For all  $f$ ,

$$\mathbb{E}f\left(\frac{\sum_i X_i}{\sqrt{n}}\right) \rightarrow \mathbb{E}f(Z),$$

where  $Z \sim \mathcal{N}(0, 1)$ . So need to prove

$$\mathbb{E}f\left(\frac{\sum_i X_i}{\sqrt{n}}\right) - \mathbb{E}f\left(\frac{\sum_i X_i^G}{\sqrt{n}}\right) = o(1) \quad (6.18)$$

where  $X_i^G$  is Gaussian. (show the thing we want to study and something we know about (Gaussian R.V.s) are close together)

Let

$$S_n = \sum_{i=1}^n X_i, \quad S_n^G = \sum_{i=1}^n X_i^G, \quad \tilde{S}_n^k = \sum_{i=1}^k X_i + \sum_{i=k+1}^n X_i^G.$$

Note that  $\tilde{S}_n^n = S_n$  and  $\tilde{S}_n^0 = S_n^G$ . We can express the left side of (6.18) as

$$\sum_k \left[ \mathbb{E}f\left(\frac{\tilde{S}_n^{k+1}}{\sqrt{n}}\right) - \mathbb{E}f\left(\frac{\tilde{S}_n^k}{\sqrt{n}}\right) \right]$$

### 6.7.3 Subgaussian Random Variable

$g \sim \mathcal{N}(0, 1)$ ,  $t \rightarrow \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$ . Fact

1.

$$\mathbb{P}(|g| > t) \leq 2 \exp(-Ct^2)$$

2.

$$G = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}$$

$g_i \sim \mathcal{N}(0, 1)$ , with density function  $x \rightarrow (2\pi)^{-n/2} \exp(-\sum_i X_i^2/2)$ .

So the density only depends on  $\|X\|_2$ , density is invariant under rotation, which means if  $U$  is an orthogonal matrix, then  $UG \sim G$ .

**Definition 6.43 (Subgaussian random variable).** A random variable  $X$  is called **subgaussian** if

$$\|X\|_\Psi = \inf \left\{ s > 0 : \mathbb{E}F\left(\frac{|X|}{s}\right) \leq 1 \right\} < \infty$$

(“find  $s$  such that  $\mathbb{E}F(|X|/s) = 1$ .”)

**Remark.**  $\|\cdot\|_\Psi$  is a norm (the **subgaussian norm**); one can check that it satisfies the triangle inequality, homogeneity, and it is nonnegative.

**Proposition 6.7.7.** The following are equivalent:

1. There exists a  $\kappa$  such that  $\|X\|_\Psi \leq \kappa$ .
2. There exists a  $\kappa$  such that for all  $t > 0$ ,  $\mathbb{P}(|X| > t) \leq 2 \exp(-t^2/\kappa^2)$ .
3. There exists a  $\kappa$  such that  $\sup_{p>1} \left\{ \frac{\mathbb{E}(|X|^p)^{1/p}}{\sqrt{p}} \right\} \leq \kappa$ .
4. There exists a  $\kappa$  such that  $\mathbb{E}\exp(\lambda X) \leq 2 \exp(\lambda^2 \kappa^2)$ .
5. (only when  $\mathbb{E}X = 0$ )  $\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2 \kappa^2)$ .

*Proof.* First we show that (1) implies (2). We will use exponential Markov's inequality: if  $Z$  is nonnegative,

$$\mathbb{P}(Z < t) \leq \frac{\mathbb{E}Z}{t}$$

$$\begin{aligned} \mathbb{P}(|X| > t) &= \mathbb{P}(\exp(|X|^2/\kappa^2) > \exp(t^2/\kappa^2)) \leq \frac{\mathbb{E}\exp(|X|^2/\kappa^2)}{\exp(t^2/\kappa^2)} = \frac{\mathbb{E}[f(|X|/\kappa) + 1]}{\exp(t^2/\kappa^2)} \\ &\leq \frac{2}{\exp(t^2/\kappa^2)} = 2 \exp(-t^2/\kappa^2). \end{aligned}$$

which implies  $g$  is subgaussian. Now we show that 2 implies 3.

$$\mathbb{E}|X|^p = \mathbb{E} \left( \int_0^{|X|^p} 1 \, dt \right) = \int_0^\infty \mathbf{1}_{[t,\infty)}(|X|^p) dt = \int_0^\infty \mathbb{P}(|X|^p < Mt) \, dt$$

Let  $u^p = t$ . We will use integration by parts

$$\begin{aligned} &= \int_0^\infty pu^{p-1}\mathbb{P}(|X| > u) du \leq \int_0^\infty pu^{p-1}2\exp(-u^2/\kappa^2) du = \int_0^\infty \underbrace{pu^{p-2}}_w \cdot \underbrace{u2\exp(-u^2/\kappa^2)}_{v'} du \\ &= [wv]_0^\infty - \int_0^\infty w'v du = pu^{p-2}(-\kappa^2 \exp(-u^2/\kappa^2))_0^\infty + \int_0^\infty p(p-2)u^{p-3}\kappa^2 \exp(-u^2/\kappa^2) du \\ &= p(p-2)(p-4)C^p\kappa^p \leq C^p 9p^{p/2}\kappa^p \end{aligned}$$

$p$ th root  $\implies \sqrt{p}C\kappa$ .

Next we will show 3 implies 1.

⋮

□

**Proposition 6.7.8.** Consider  $a \in S^{n-1}$ . Suppose  $X_1, \dots, X_n$  are mean 0 with  $\|X_i\|_\Psi \leq \kappa$  for all  $i$ . Then  $\sum_i a_i X_i$  is subgaussian (for all  $n \in \mathbb{Z}_{++}$ ).

*Proof.*

$$\begin{aligned} \mathbb{E} \exp \left( \lambda \sum_i a_i X_i \right) &= \mathbb{E} \prod_i \exp(\lambda a_i X_i) = \prod_i \mathbb{E} \exp(\lambda a_i X_i) \leq \prod_i \mathbb{E} \exp(C\lambda^2 a_i^2 \kappa^2) \\ &= \mathbb{E} \exp \left( C\lambda^2 \left( \sum_i a_i^2 \right) \kappa^2 \right) = \mathbb{E} \exp(C_1 \lambda^2 \kappa^2). \end{aligned}$$

Therefore  $\sum_i a_i X_i$  is subgaussian (and  $\|\sum_i a_i X_i\|_\Psi \leq C_1 \kappa$ ).

□

**Examples of subgaussian random variables:** Gaussian, Bernoulli, bounded.

**Definition 6.44 (Hilbert-Schmidt Norm).**

$$\|A\|_{HS} = \sqrt{\sum_{i,j} a_{ij}^2}.$$

**Remark.** Note

$$\mathrm{Tr}(A^* A) = \|A\|_{HS}^2 = \sum_i s_i^2(A)$$

(the sum of the squared singular values of  $A$ ).

**Proposition 6.7.9 (Hanson-Wright Inequality).**

*Proof.*

□

#### 6.7.4 Topic 1: Invertibility of Matrices

**Very tall case:** Suppose  $A \in \mathbb{R}^{n \times m}$ , with  $n \gg m$ . Entries  $a_{ij}$  are independent with  $\mathbb{E}a_{ij} = 0$ ,  $\mathbb{E}a_{ij}^2 = 1$ , and  $\|a_{ij}\|_\Psi \leq \kappa$ . A left inverse of  $A$ ,  $A^+$  exists if the smallest singular value of  $A$  is positive; that is,  $\delta_m(A) > 0$ . Then the columns of  $A$  span  $\mathbb{R}^m$ , and there exists a matrix  $A^+$  such that  $A^+A = I_m$ .

**Proposition 6.7.10 (Operator norm of  $A$ ).**

$$\mathbb{P}(\|A\| \geq R(\sqrt{n} + \sqrt{m})) \leq \exp(-C_2 R^2(\sqrt{n} + \sqrt{m})^2)$$

for  $R > C_1$ .

*Proof.*

•

$$\|A\| = \max_{x \in S^{n-1}} \|Ax\|_2 = \max_{x \in S^{n-1}} \max_{y \in S^{n-1}} \langle y, Ax \rangle$$

- For fixed  $y \in S^{n-1}$ ,  $x \in S^{n-1}$ ,

$$\langle y, .Ax \rangle = \sum_{i,j} y_i a_{ij} x_j$$

Using the subgaussianity of  $a_{ij}$ ,

$$\left\| \sum_{i,j} x_i x_j a_{ij} \right\|_\Psi \leq C \sqrt{\sum_{i,j} (y_i x_j)^2 \kappa} = CK$$

Using the second definition of the subgaussian norm,

$$\implies \mathbb{P}(|\langle y, Ax \rangle| > t) \leq 2 \exp(-Ct^2/\kappa^2)$$

We need an upper bound for all  $x \in S^{m-1}$ ,  $y \in S^{n-1}$ .

**Definition 6.45 ( $\epsilon$ -net).**  $\mathcal{M} \subseteq S^{m-1}$  is an  $\epsilon$ -net of  $S^{m-1}$ , if for all  $x \in S^{m-1}$ , there exists an  $x' \in \mathcal{M}$  such that  $\|x - x'\|_2 \leq \epsilon$ .

Take a  $1/4$ -net in  $S^{m-1}$ , called  $\mathcal{M}$ . Likewise, take a  $1/4$ -net in  $S^{n-1}$ , called  $\mathcal{N}$ .

$$\mathbb{P}(\exists x \in \mathcal{M}, y \in \mathcal{N} \text{ s.t. } |\langle y, Ax \rangle| > t) \leq |\mathcal{M}| |\mathcal{N}| \exp(-Ct^2).$$

Suppose for all  $x \in \mathcal{M}$ ,  $y \in \mathcal{N}$ , we have  $|\langle y, Ax \rangle| \leq t$ . Let  $x_0 \in S^{m-1}$ ,  $y_0 \in S^{n-1}$  such that  $\|A\| = \langle y_0, Ax_0 \rangle$ . Take  $y \in \mathcal{N}$ ,  $x \in \mathcal{M}$  such that  $\|y - y_0\| \leq 1/4$ ,  $\|x - x_0\| \leq 1/4$ . Then

$$\|A\| = \langle y, Ax_0 \rangle + \langle y_0 - y, Ax_0 \rangle \leq \langle y, Ax_0 \rangle + \|y_0 - y\| \|A\|$$

$$\begin{aligned} &= \langle y, Ax \rangle + \langle y, A(x_0 - x) \rangle + \frac{1}{4} \|A\| \leq \langle y, Ax \rangle + \|y\| \|A\| \|x_0 - x\| + \frac{1}{4} \|A\| \\ &\implies \|A\| \leq \langle y, Ax \rangle + \frac{1}{2} \|A\| \\ &\iff \|A\| \leq 2\langle y, Ax \rangle \leq 2t \end{aligned}$$

since we already have  $|\langle y, Ax \rangle| \leq t$ .

Then

$$\mathbb{P}(\forall x \in \mathcal{M}, \forall y \in \mathcal{N}, |\langle y, Ax \rangle| \leq t) \geq 1 - |\mathcal{M}||\mathcal{N}| \exp(-Ct^2/\kappa^2)$$

therefore with probability at least  $1 - |\mathcal{M}||\mathcal{N}| \exp(-Ct^2/\kappa^2)$ , we have  $\|A\| \leq 2t$ .

□

Next, adjust  $t$  according to  $|\mathcal{M}||\mathcal{N}|$  to maximize the probability.

**Proposition 6.7.11.** For  $\epsilon < 1$ , let  $\mathcal{N}$  be an  $\epsilon$ -net of  $S^{n-1}$ . Then  $\mathcal{N}$  satisfies  $|\mathcal{N}| < (3/\epsilon)^n$ .

*Proof.* Let  $B_2^n$  denote the Euclidean unit ball in  $\mathbb{R}^n$ . Then  $x + \epsilon B_2^n = \{x + \epsilon y : y \in B_2^n\}$ . We will show that for all  $x, y \in \mathcal{N}$ ,

$$(x + \epsilon/2 B_2^n) \cap (y + \epsilon/2 B_2^n) = \emptyset.$$

Construct  $\mathcal{N}$  by picking a point  $x_1 \in S^{n-1}$ ,  $x_2$  so that  $|x_1 - x_2| > \epsilon$ , keep choosing points until there isn't any more empty space. Once we stop, there are no more points in  $S^{n-1}$  such that its epsilon distance away from all points in  $\mathcal{N}$  which means  $\mathcal{N}$  is an epsilon net. Then

$$\bigcup_{x \in \mathcal{N}} (x + \epsilon/2 B_2^n) \subset (1 + \epsilon/2) B_2^n,$$

so

$$\left| \bigcup_{x \in \mathcal{N}} (x + \epsilon/2 B_2^n) \right| < |(1 + \epsilon/2) B_2^n| \implies |\mathcal{N}| \cdot |\epsilon/2 B_2^n| < |1 + \epsilon/2 B_2^n|$$

$$\implies |\mathcal{N}|(\epsilon/2)^n < (1 + \epsilon/2)^n \implies |\mathcal{N}| < (2/\epsilon + 1)^n < (3/\epsilon)^n.$$

□

**Proposition 6.7.12 (Paley-Zygmund Inequality).** If  $Z$  is a nonnegative random variable with  $\mathbb{E}Z = 1, \mathbb{E}Z^2 < K$ , then for  $\lambda \in (0, 1)$ , there exists  $p > 0$  such that  $\mathbb{P}(Z > \lambda) > p$ .

*Proof.*

$$\begin{aligned} 1 = \mathbb{E}Z &= \mathbb{E}Z\mathbf{1}_{\{Z \in [0, \lambda]\}} + \mathbb{E}Z\mathbf{1}_{\{Z \in (\lambda, \infty)\}} \leq \lambda + \sqrt{\mathbb{E}Z^2 \mathbb{E}(\mathbf{1}_{\{Z \in (\lambda, \infty)\}})} \leq \lambda + \sqrt{K \mathbb{P}(Z > \lambda)} \\ &\iff \frac{(1 - \lambda)^2}{K} \leq \mathbb{P}(Z > \lambda), \end{aligned}$$

which proves the statement for  $p = \frac{(1 - \lambda)^2}{K}$ .

□

### 6.7.5 Random Graphs

1. **Erdos-Renyi graph:**  $\mathcal{G}(N, p)$  line connecting  $V_i, V_j$  if and only iff  $A_{ij} = 1$ .  $\mathbb{P}(A_{ij} = 1) = p(n)$ , some function of  $n$ . All  $A_{ij}$  are independent of each other besides the fact that  $A$  is symmetric. connected with probability  $p$ .
2. **Random  $d$ -regular graph model:**  $\mathcal{G}_{n,d}$ : set of  $d$ -regular graphs with  $n$  vertices.

These are used for statistical inference, particularly community detection.

$$|G_{nd}| = \Pr(G \text{ simple}) \cdot \frac{(nd - 1)!!}{(d!)^n}$$

**Theorem 6.7.13.**  $\mathbb{P}(G \text{ is simple}) = e^{-(d^2 - 1)/4}$ .

**Remark.** It doesn't go to 0 as  $n \rightarrow \infty$ .

1.

$$|G_{nd}| = (1 + o(1))e^{-(d^2 - 1)/2} \cdot \frac{(nd - 1)!!}{(d!)^n}$$

2. efficient way to sample a  $d$ -regular graph. Expected waiting time is  $e^{(d^2 - 1)/4}$ .

3. If you can probe property A is true for the configuration model with probability  $1 - o(1)$ , then the same thing is true for the uniform model.

$$o(1) = \mathbb{P}(A \text{ is not true}) \geq \mathbb{P}(\mathcal{G} \text{ is simple}, A \text{ is not true for } \mathcal{G})$$

$$= \mathbb{P}(A \text{ is not true} \mid \mathcal{G} \text{ is simple}) \mathbb{P}(\mathcal{G} \text{ simple})$$

$$\approx \mathbb{P}_{\text{uniform}}(A \text{ is not true}) e^{-(d^2 - 1)/4}$$

$$\mathbb{P}_{\text{uniform}}(A \text{ is not true}) = o(1) \text{ as } n \rightarrow \infty.$$

## 7 Stochastic Processes

These notes are based on my notes from ISE 620 at USC taught by Sheldon Ross (along with the textbooks *Stochastic Processes* [Ross, 2008] and *Introduction to Probability Models* [Ross, 2014] by Sheldon Ross) *Time Series and Panel Data Econometrics* (1st edition) by M. Hashem Pesaran [Pesaran, 2015] and coursework for Economics 613: Economic and Financial Time Series I at USC, as well as notes from *Probability and Random Processes* by Grimmett and Stirzaker [Grimmett and Stirzaker, 2001].

### 7.1 Preliminaries

**Definition 7.1.** A **stochastic process** is a collection of random variables  $X(t), t \geq 0$  (in the continuous case) or  $X_1, X_2, \dots$  in the discrete case such that ...

**Definition 7.2.** A stochastic process  $\{N(t), t \geq 0\}$  is said to be a **counting process** if  $N(t)$  represents the total number of events that have occurred up to time  $t$ . Hence, a counting process  $N(t)$  must satisfy

- (i)  $N(t) \geq 0$
- (ii)  $N(t)$  is integer-valued.
- (iii) If  $s < t$  then  $N(s) \leq N(t)$ .
- (iv) For  $s < t$ ,  $N(s) - N(t)$  equals the number of events that have occurred in the interval  $(s, t]$ .

**Definition 7.3.** We say a counting process  $\{N(t), t \geq 0\}$  has **independent increments** if the numbers of events that occur in disjoint time intervals are independent; that is, for all  $t_0 < t_1 < \dots < t_n$ ,  $N(t_1) - N(t_0), \dots, N(t_n) - N(t_{n-1})$  are independent.

**Definition 7.4.** A counting process  $\{N(t), t \geq 0\}$  has **stationary** increments if the distribution of the number of events that occur in any interval of time depends only on the length of the time interval. That is, if  $N(t+s) - N(s)$  has a distribution that does not depend on  $s$  (is the same for all  $s$ ; only depends on  $t$ ).

### 7.2 Poisson Processes

**Definition 7.5 (Poisson Process, Grimmet and Stirzaker definition).** A **Poisson process with intensity**  $\lambda$  is a process  $N = \{N(t) : t \geq 0\}$  taking values in  $S = \{0, 1, 2, \dots\}$  such that

- (a)  $N(0) = 0$ ; if  $s < t$  then  $N(s) \leq N(t)$ .

$$(b) \Pr(N(t+h) = n+m \mid N(t) = n) = \begin{cases} \lambda h + o(h) & \text{if } m = 1, \\ o(h) & \text{if } m > 1; \\ 1 - \lambda h + o(h) & \text{if } m = 0 \end{cases}$$

- (c) If  $s < t$ , the number  $N(t) - N(s)$  of emissions in the interval  $(s, t]$  is independent of the times of emissions during  $[0, s]$ .

**Remark.**  $\lambda$  can be interpreted as the average or long-run frequency of the Poisson process.

**Definition 7.6 (Poisson Process, Ross definition, 2.1.2 in Stochastic Processes).** The counting process  $\{N(t) : t \geq 0\}$  is said to be a **Poisson process with rate  $\lambda$**  if

- (a)  $N(0) = 0$
- (b)  $\{N(t) : t \geq 0\}$  has independent increments
- (c)  $\Pr(N(t+h) - N(t) = 1) = \lambda h + o(h)$
- (d)  $\Pr(N(t+h) - N(t) \geq 2) = o(h)$

**Remark.** Note that a Poisson process has stationary increments.

**Lemma 7.2.1 (ISE 620 Ross Lemma 1).** Let  $N(t)$  be a Poisson process. For a fixed time  $s$ , let  $N_s(t) = N(s+t) - N(s)$  be the difference between two Poisson processes. Then  $\{N_s(t), t \geq 0\}$  is a Poisson process.

*Proof.* We verify the conditions in Definition 7.6.

- (a)  $N_s(0) = N(s) - N(s) = 0$ , QED.
- (b)  $\{N_s(t) : t \geq 0\}$  has independent increments, yes.
- (c)  $\Pr(N_s(t+h) - N_s(t) = 1) = \Pr(N(s+t+h) - N_s(s+t) = 1) = \lambda h + o(h)$ , QED.
- (d)  $\Pr(N_s(t+h) - N_s(t) \geq 2) = (\Pr(N(s+t+h) - N_s(s+t) \geq 2) = o(h)$ , QED.

□

**Lemma 7.2.2 (ISE 620 Ross Lemma 2).** Let  $N(t)$  be a Poisson process. Then  $P_0(t) = \Pr(N(t) = 0) = e^{-\lambda t}$ .

*Proof.*

$$P_0(t+h) = \Pr(N(t+h) = 0) = \Pr(N(t+h) = 0, N(t) = 0)$$

$$= \Pr(N(t) = 0) \Pr(N(t+h) - N(t) = 0) \text{ (by independent increments)}$$

Using conditions (iii) and (iv) of Definition 7.6,

$$P_0(t+h) = P_0(t)(1 - \lambda h - o(h)) = P_0(t)(1 - \lambda h) + o(h)$$

$$\iff \frac{P_0(t+h) - P_0(t)}{h} = \frac{-\lambda h P_0(t)}{h} + \frac{o(h)}{h}$$

Taking the limit as  $h \rightarrow 0$ , we have

$$P'_0(t) = -\lambda P_0(t) \iff \frac{P'_0(t)}{P_0(t)} = -\lambda \iff \log(P_0(t)) = -\lambda t + C$$

$$\iff P_0(t) = \lambda e^{-\lambda t} \iff P_0(0) = 1$$

□

**Theorem 7.2.3.** [Grimmett and Stirzaker theorem 6.8.2] Let  $N(t)$  be a Poisson process with intensity  $\lambda$ . Then  $N(t)$  has the Poisson distribution with parameter  $\lambda t$ ; that is,

$$\Pr(N(t) = j) = \frac{(\lambda t)^j \exp(-\lambda t)}{j!}, \quad j = 0, 1, 2, \dots$$

*Proof.* See Grimmett and Stirzaker section 6.8.2, page 247. Ross proof:

$$\begin{aligned} \Pr(N(t) = n) &= \frac{1}{\Pr(X_{n+1} = t - s \mid X_n = s)} \int_0^t \Pr(N(t) = n \mid S_n = s) \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} ds = 1 \\ &= \Pr(X_{n+1} > t - s) = e^{-\lambda(t-s)} \\ \Pr(N(t) = n) &= \int_0^t e^{-\lambda(t-s)} \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} ds \\ &= \frac{e^{-\lambda t} \lambda^n}{(n-1)!} \int_0^t s^{n-1} ds = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \end{aligned}$$

□

**Corollary 7.2.3.1.**  $N(t+s) - N(s) \sim \text{Poisson}(\lambda t)$

*Proof.* By Lemma 7.2.1, ?? is a Poisson process. . .

□

**Definition 7.7 (Poisson Process, Ross definition, 2.1.1 in Stochastic Processes).** The counting process  $\{N(t) : t \geq 0\}$  is said to be a **Poisson process with rate  $\lambda$**  if

- (a)  $N(0) = 0$
- (b)  $\{N(t) : t \geq 0\}$  has independent increments
- (c) The number of events in any interval of length  $t$  is Poisson distributed with mean  $\lambda t$ . That is, for all  $s, t \geq 0$ ,

$$\Pr(N(t+s) - N(s) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad n = 0, 1, \dots$$

**Remark.** Note that it follows from condition (iii) in Definition 7.7 that a Poisson process has stationary increments and also that  $\mathbb{E}(N(t)) = \lambda t$ .

**Definition 7.8 (Ross Stochastic Processes definition, Section 2.2).** Let  $X_n$  denote the time between the  $n - 1$ st and  $n$ th event in a Poisson process. The sequence  $\{X_n, n \geq 1\}$  is called the **sequence of interarrival times**.

**Definition 7.9 (Grimmett and Stirzaker definition).** Let  $N(t)$  be a Poisson process with intensity  $\lambda$ . Let  $T_0, T_1, \dots$  be given by

$$T_0 = 0, T_n = \inf\{t : N(t) = n\} \quad (7.1)$$

so that  $T_n$  is the time of the  $n$ th arrival. The **interarrival times** are the random variables  $X_1, X_2, \dots$  given by

$$X_n = T_n - T_{n-1}. \quad (7.2)$$

**Remark.** From knowledge of  $N$ , we can find the values of  $X_1, X_2, \dots$  by (7.1) and (7.2). Conversely, we can construct  $N$  from knowledge of the  $X_i$  by

$$T_n = \sum_{i=1}^n X_i, \quad N(t) = \max\{n : T_n \leq t\} \quad (7.3)$$

**Theorem 7.2.4. (Grimmett and Stirzaker theorem 6.8.10.)** Let  $N(t)$  be a Poisson process with intensity  $\lambda$ . Let  $T_0, T_1, \dots$  be given by (7.1) and let  $X_n$  be given by (7.2). Then the random variables  $\{X_n\}$  are independent, each having an exponential distribution with parameter  $\lambda$ .

*Proof.* See Grimmett and Stirzaker section 6.8.2, page 249.  $\square$

**Corollary 7.2.4.1.** Let  $N(t)$  be a Poisson process with intensity  $\lambda$ . Let  $T_0, T_1, \dots$  be given by (7.1). Then  $T_n \sim \text{Gamma}(n, \lambda^{-1})$ .

*Proof.* By (7.3),  $T_n = \sum_{i=1}^n X_i$ .  $X_i \sim \text{Exponential}(\lambda)$  by Theorem 7.2.4, which means  $X_i \sim \text{Gamma}(1, \lambda^{-1})$ . Then by Proposition 6.3.18,  $T_n \sim \text{Gamma}(n, \lambda^{-1})$ .  $\square$

**Lemma 7.2.5 (ISE 620 Ross in-class Lemma 3, Proposition 2.2.1 in Stochastic Processes).** Let  $N(t)$  be a Poisson process. Let  $X_1, X_2, \dots$  be the interarrival times. Then  $X_1, X_2, \dots$  are independently and identically distributed as  $\text{Exponential}(\lambda)$ . Since the sum of exponential distributions is Gamma, we have  $S_n = \sum_{i=1}^n X_i \sim \text{Gamma}(n, 1/\lambda)$ .

*Proof.* See proof on page 64 (section 2.2) of *Stochastic Processes*. Follows from Lemmas 7.2.1 and 7.2.2.  $\square$

**Remark.** If the arrival times are IID exponential, then the process is Poisson. (Will come later in notes.)

**Remark.**

$$X_1 < t \iff N(t) = 1$$

$$\Pr(X_2 > t \mid X_a = s) = \Pr(N(S+t) - N(s) = 0 \mid X_1 = s) = \Pr(N(s+t) - N(s) = 0) = \Pr(N_s(t) = 0) = e^{-\lambda t}$$

**Theorem 7.2.6.** Suppose events occur in a Poisson process with parameter lambda,  $\text{PP}(\lambda)$ . Let  $N(t)$  be the number of events that occur by time  $t$ . Suppose each event is independent of all that has previously occurred; that is, each event is type  $i = 1, \dots, r$  with probability  $p_i$ , and  $\sum_{i=1}^r p_i = 1$ . Let  $N_i(t)$  be the number of type  $i$  events that occur by time  $t$ .  $\{N_i(t)\} \sim \text{PP}(\lambda p_i)$ . Moreover, these processes are independent for different  $i$ .

*Proof.* Verify axioms of Poisson process:

1.  $N_i(0) = 0$
2. independent increments? Yes, the number of type  $i$  events in each interval is totally independent of what happened in previous intervals.
3.  $\Pr(N_i(t+h) - N_i(t) = 1) = \Pr(N(t+h) - N(t) = 1, \text{ event is type } i) + \dots$   
probability of two or more events (and then one of them is type  $i$ ) is  $o(h)$ .

$$= \lambda h p_i + o(h)$$

4.  $\Pr(N_i(t+h) - N_i(t) \geq 2) \leq \Pr(N(t_h) - N(t) \geq 2) = o(h)$

Therefore  $N_i(t)$  is a Poisson process with rate  $\lambda p_i$ . We could have also proven this by looking at the inter-arrival times and showing that they are i.i.d. exponential.

□

*Proof: Alternative, stronger statement.*

$$\Pr(N_1(t) = n_1, N_2(t) = n_2, \dots, N_r(t) = n_r)$$

we want to show that these random variables are independent. Let  $n = \sum_{i=1}^r n_i$ . Then

$$= \Pr(N_1(t) = n_1, \dots, N_r(t) = n_r \mid N(t) = n) \Pr(N(t) = n)$$

But this is a Multinomial distribution. So we have

$$\begin{aligned} &= \frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i^{n_i} \cdot e^{-\lambda t} \frac{(\lambda t)^{\sum_i n_i}}{n!} \\ &\quad \prod_{i=1}^r e^{-\lambda t p_i} \frac{(\lambda t p_i)^{n_i}}{n_i!} \end{aligned}$$

And these are just Poisson probabilities.

Then using Proposition 7.2.7, we have independence (???)

□

**Proposition 7.2.7** (offhand claim Ross made). If  $\Pr(X = n, Y = m) = g(n)h(m)$  (that is, if probability can be broken into products of functions) then variables are independent.

*Proof.*

$$\Pr(X = n) = \sum_m g(n)h(m) = g(n)C_h$$

$$\Pr(Y = m) = h(m) \sum_n g(n) = h(m)C_g$$

$$1 = \sum_n \sum_m g(n)h(m) = C_g C_h$$

So this is true as long as  $C_g C_h = 1$  which it does.

□

**Example 7.1.** Suppose in particular  $\lambda = 10$  so we expect 10 people to arrive every hour, either men or women. What is the expected number of women to arrive given that 7 men arrived? (5—women and men's arrivals are independent.)

**Example 7.2 (Coupon collecting problem, see also probability notes.).**  $r$  types of coupons collected with probability  $p_1, \dots, p_r$ . Let  $N$  be the number of coupons you collect until you have a complete set. What is  $\mathbb{E}(N)$ ?

**Solution.** Let  $m(\mathcal{S})$  be the mean number of draws required to obtain at least one coupon of type  $i$  for each  $i \in \mathcal{S}$ . Note that  $m(\emptyset) = 0$ , and

$$\begin{aligned} m(\mathcal{S}) &= 1 + \sum_{i \in \mathcal{S}} p_i m(\mathcal{S} \setminus i) + \sum_{i \notin \mathcal{S}} p_i m(\mathcal{S}) \\ \implies m(\mathcal{S}) &= \frac{1 + \sum_{i \in \mathcal{S}} p_i m(\mathcal{S} \setminus i)}{\sum_{i \notin \mathcal{S}} p_i} \end{aligned}$$

But unless  $r$  is small this is going to be hard to compute, so this doesn't really work.

**Solution.** New idea: Let  $N_i$  be the number of draws required to obtain types  $1, \dots, i$ . We want  $N_r$ . Note that  $\mathbb{E}(N_1) = 1/p_1$ . We have  $N_{i+1} = N_i + A$  where  $A$  is the additional time required.

$$\mathbb{E}(N_{i+1}) = \mathbb{E}(N_i) + \mathbb{E}(A)$$

if there is a type  $i+1$  in the original group, then  $\mathbb{E}(A) = 0$ . If not, note that

$$\mathbb{E}(A) = \begin{cases} 1/(p_i + 1) & \Pr(\{\text{a type } i+1 \text{ coupon has not already been collected}\}) \\ 0 & \Pr(\{\text{a type } i+1 \text{ coupon has already been collected}\}) \end{cases}$$

$$\implies \mathbb{E}(A) = \frac{1}{p_i + 1} \cdot \Pr(\{\text{a type } i+1 \text{ coupon has not already been collected}\})$$

$$\implies \mathbb{E}(A) = \frac{1}{p_i + 1} \cdot \Pr(\{\text{type } i+1 \text{ is the last of types } 1, \dots, i+1 \text{ to be collected}\})$$

$$= \frac{1}{p_i + 1} \cdot \frac{\sum_{j=1}^{i+1} p_j}{p_{i+1}}$$

this didn't work either.

### Solution.

Let  $N_i$  be the number to get type  $i$ . So  $N_i \sim \text{Geometric}(p_i)$ . Then  $N = \max\{N_i\}$ . So

$$\Pr(N \leq k) = \Pr(N_1 \leq k, \dots, N_r \leq k)$$

but you can't multiply out probabilities because  $N_i$  is not independent.

#### 7.2.1 Poissonization Trick

Suppose we collect coupons at times distributed according to a Poisson process with rate  $\lambda = 1$ . Let event type  $i$  be the event of collecting a type  $i$  coupon. Let  $T_i$  be the time until collecting a type  $i$  coupon. Note that by Theorem 7.2.6 and Lemma 7.2.5,  $T_i \sim \text{Exponential}(\lambda \cdot p_i) = \text{Exponential}(p_i)$ . Then the time to collect at least one of every type is  $T = \max_i\{T_i\}$ . And by Theorem 7.2.6, the  $T_i$  are all independent. Return to the coupon collecting problem:

**Solution.** Recall the layer cake formulation for expected value:

$$\mathbb{E}(T) = \int_0^\infty \Pr(T > t) dt$$

We have (using independence of the  $T_i$ )

$$\Pr(T > t) = 1 - \Pr(T \leq t) = 1 - \Pr(T_1 \leq t, \dots, T_r \leq t) = 1 - \prod_{i=1}^r \Pr(T_i \leq t) = 1 - \prod_{i=1}^r (1 - e^{-p_i t})$$

$$\implies \boxed{\mathbb{E}(T) = \int_0^\infty \left(1 - \prod_{i=1}^r (1 - e^{-p_i t})\right) dt}$$

$$= \sum_i \frac{1}{p_i} - \sum_{i < j} \frac{1}{p_i + p_j} + \sum_{i,j,k} \frac{1}{[p_i + p_j + p_k]} - \dots + \frac{(-1)^{r+1}}{p_1 + \dots + p_r}$$

But we want  $\mathbb{E}(N)$ . Recall that  $T$  is the time of the  $N$ th event. So

$$T = \sum_{i=1}^N X_i$$

where  $X_1, X_2, \dots$  are the interarrival times of a Poisson process with rate  $\lambda = 1$ . But  $N$  is independent of  $X_1, X_2, \dots$  (the type of coupons you get has nothing to do with the time between coupons). So  $T$  is a compound random variable. Therefore

$$\mathbb{E}(T) = \mathbb{E}(N)\mathbb{E}(X) = \mathbb{E}(N) \cdot \frac{1}{\lambda} = \mathbb{E}(N)$$

Therefore we have  $\boxed{\mathbb{E}(N) = \int_0^\infty \left(1 - \prod_{i=1}^r (1 - e^{-p_i t})\right) dt}.$

**Example 7.3.** A type is a *singleton* if after you get a complete set, you have only one object of that type. Suppose we collect coupons until we have at least one of every type, and that all coupons are equally likely to be collected on each draw. What is the expected number of singletons when you stop?

**Solution.** Let  $X$  be the number of singletons when you stop. Let  $I_j$  be an indicator variable for type  $j$  being a singleton. Then

$$\mathbb{E}(X) = \mathbb{E}\left[\sum_{j=1}^r I_j\right] = \sum_{j=1}^r \mathbb{E}(I_j) = \sum_{j=1}^r \Pr(\{j \text{ is a singleton}\})$$

⋮

$$= \frac{1}{r} \sum_{i=1}^r \frac{1}{r-i+1}$$

⋮

Let  $T_i$  be the time you get your first type  $i$  coupon. The probability we want is that at the time when you have at least one of every type except  $j$ , you have either 0 or 1 coupons of type  $j$ . That is, if  $S_2^{(j)}$  is the time the second card of type  $j$  is selected, we must have  $S_2^{(j)} > \max_{i \neq j} \{T_i\}$ . So we seek  $\Pr(S_2^{(j)} > \max_{i \neq j} \{T_i\})$ . Note that  $S_2^{(j)} \sim \text{Gamma}(2, p_j)$  by Corollary 7.2.4.1, so

$$f_{S_2^{(j)}}(s) = \frac{1}{p_j^2} s^{2-1} e^{-p_j s} = p_j e^{-p_j s} p_j \cdot s.$$

So we have

$$\Pr(I_j = 1) = \Pr(S_2^{(j)} > \max_{i \neq j} \{T_i\}) = \int_0^\infty \Pr(\max_{i \neq j} \{T_i\} < S_2^{(j)} \mid S_2^{(j)} = s) \cdot f_{S_2^{(j)}}(s) ds$$

$$= \int_0^\infty \prod_{i \neq j} (1 - e^{-p_i s}) \cdot p_j e^{-p_j s} p_j \cdot s \cdot ds$$

which yields

$$\boxed{\mathbb{E}(X) = \sum_{j=1}^r p_j^2 \int_0^\infty e^{-p_j s} \cdot s \cdot \prod_{i \neq j} (1 - e^{-p_i s}) \cdot ds}$$

**Remark.** Per Example 7.4, suppose people arrive at a bus stop according to a Poisson process with rate  $\lambda$ . A bus arrives at (fixed) time  $T$ . Let  $W$  be the sum of the waiting times for everyone at the bus stop. Let  $S_j$  be the arrival time of the  $j$ th person. (Note that if  $X_i$  are the interarrival times then  $S_j = \sum_{i=1}^j X_i$ .) The number of people who get on the bus is  $N(t)$ , the Poisson counting process. So

$$W = \sum_{i=1}^{N(T)} (T - S_i) = N(T)T - \sum_{i=1}^{N(T)} S_i$$

**Proposition 7.2.8.** In a Poisson process, if we know that one event has occurred by time  $t$ , then the distribution of times of the event is uniform between 0 and  $t$ . That is,

$$S_1 < x \mid N(t) = 1 \sim U(0, t)$$

*Proof.*

$$\begin{aligned} \Pr(S_1 < x \mid N(t) = 1) &= \frac{\Pr(S_1 < x \cap N(t) = 1)}{\Pr(N(t) = 1)} = \frac{\Pr(N(x) = 1 \cap N(t) - N(x) = 0)}{\exp(-\lambda t) \lambda t} \\ &= \frac{\Pr(N(x) = 1) \Pr(N(t) - N(x) = 0)}{\exp(-\lambda t) \lambda t} = \frac{\lambda x e^{-\lambda x} \cdot e^{-\lambda(t-x)}}{\exp(-\lambda t) \lambda t} = \frac{x}{t} \end{aligned}$$

which is the cdf for a random variable distributed as  $U(0, t)$ .

□

**Remark.** If  $X_i \sim U(0, t)$ , then  $f(x) = 1/t$ ,  $0 < x < t$ , so per Proposition 10.1.2,

$$f_{X_{(1)}, \dots, X_{(n)}}(x_1, \dots, x_n) = \frac{n!}{t^n}, \quad 0 < x < t$$

**Theorem 7.2.9.** In a Poisson process, if we know that  $n$  events have occurred by time  $t$  ( $N(t) = n$ ), then the set of event times  $\{S_1, \dots, S_n\}$  are distributed as a set of  $n$  i.i.d. uniform random variables between 0 and  $t$ . That is, the unordered times are distributed uniformly on the interval.

More precisely: Given  $N(t) = n$ ,  $S_1, \dots, S_n$  are distributed as order statistics of  $n$  i.i.d.  $U(0, t)$  random variables.

*Proof.* We will examine the density function

$$f_{S_1, \dots, S_n, |N(t)=n}(t_1, \dots, t_n), \quad 0 < t_1 < \dots < t_n < t$$

Per the above remark, we want to show that this is  $n!/t^n$ . Note that by Bayes' Theorem

$$f_{S_1, \dots, S_n, |N(t)=n}(t_1, \dots, t_n) = \frac{f_{S_1, \dots, S_n}(t_1, \dots, t_n) \cdot \Pr(N(t) = n | S_1 = t_1, \dots, S_n = t_n)}{\Pr(N(t) = n)}. \quad (7.4)$$

Let  $X_1, X_2, \dots$  be the interarrival times. So the condition above is equivalent to  $X_1 = t_1, X_2 = t_2 - t_1, \dots, X_n = t_n - t_{n-1}$ . Recall that all the  $\{X_i\}$  are independent. So we have

$$f_{S_1, \dots, S_n}(t_1, \dots, t_n) = f_{X_1, \dots, X_n}(t_1, t_2 - t_1, \dots, t_n - t_{n-1})$$

Also, we can interpret  $\Pr(N(t) = n | S_1 = t_1, \dots, S_n = t_n)$  as the probability that there are 0 arrivals between times  $t_n$  and  $t$ ; that is,

$$\Pr(N(t) = n | S_1 = t_1, \dots, S_n = t_n) = e^{-\lambda(t-t_n)}$$

So we can write (7.4) as

$$\begin{aligned} f_{S_1, \dots, S_n, |N(t)=n}(t_1, \dots, t_n) &= \frac{f_{X_1, \dots, X_n}(t_2 - t_1, \dots, t_n - t_{n-1}) \cdot e^{-\lambda(t-t_n)}}{e^{-\lambda t}(\lambda t)^n/n!} \\ &= \frac{\lambda e^{-\lambda t_1} \lambda e^{-\lambda(t_2-t_1)} \dots \lambda e^{-\lambda(t_n-t_{n-1})} \cdot e^{-\lambda(t-t_n)}}{e^{-\lambda t}(\lambda t)^n/n!} \\ &= \frac{e^{-\lambda t} \lambda^n}{e^{-\lambda t}(\lambda t)^n/n!} = \frac{n!}{t^n} \end{aligned}$$

□

**Example 7.4.** Suppose people arrive at a bus stop according to a Poisson process with rate  $\lambda$ . A bus arrives at (fixed) time  $T$ . What is the expected value of  $W$ , the sum of the waiting times for everyone at the bus stop?

**Solution.** Let  $S_j$  be the arrival time of the  $j$ th person. (Note that if  $X_i$  are the interarrival times then  $S_j = \sum_{i=1}^j X_i$ .) The number of people who get on the bus is  $N(T)$ , the Poisson counting process. So

$$W = \sum_{j=1}^{N(T)} (T - S_j) = N(T)T - \sum_{j=1}^{N(T)} S_j$$

$$\mathbb{E}(W) = \mathbb{E}(N(T)T) - \mathbb{E}\left(\sum_{j=1}^{N(T)} S_j\right) = \lambda T^2 - \mathbb{E}\left(\sum_{j=1}^{N(T)} S_j\right)$$

Note that by Theorem 7.2.9, if  $\{U_i\}$  are i.i.d. uniform random variables on  $[0, t]$ ,

$$\mathbb{E}\left(\sum_{j=1}^{N(T)} S_j \mid N(T) = n\right) = \mathbb{E}\left(\sum_{j=1}^n S_j \mid N(T) = n\right) = \mathbb{E}\left(\sum_{j=1}^n U_i\right);$$

that is, if we know that  $n$  arrivals have occurred by time  $T$ , then the (unordered) arrival times are distributed uniformly on  $[0, T]$ . But

$$\mathbb{E}\left(\sum_{j=1}^n U_i\right) = \sum_{j=1}^n \mathbb{E}U_i = \frac{nT}{2}$$

Also note that because the  $U_i$  is independent from  $N(T)$  (the arrival time of the  $j$ th person has nothing to do with how many people arrive by time  $T$ ),  $\sum_{j=1}^{N(T)} S_j$  is a compound random variable, which means

$$\mathbb{E}\left(\sum_{j=1}^{N(T)} S_j\right) = \mathbb{E}\left[\mathbb{E}\left(\sum_{j=1}^{N(T)} S_j \mid N(T) = n\right)\right] = \mathbb{E}\left(\frac{N(T)T}{2}\right) = \boxed{\frac{\lambda T^2}{2}}$$

⋮

Note that the sum of the ordered values is equal to the sum of the unordered values, so we have

$$\mathbb{E}\left(\sum_{i=1}^n X_{(i)}\right) = \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) = n\mu$$

**Example 7.5.** Another thing: Let  $I_j$  be an indicator variable for the event  $X_{(j)} < t$ . We are interested in the distribution of  $\sum_{j=1}^n I_j$ .

**Solution.** Note that the  $I_j$  is not independent because if  $I_1 = 0$  (so the smallest one is greater than  $t$ ) then it must be the case that  $I_2 = 0$  too. Also note that the sum of the ordered values less than  $t$  is the same as the sum of the unordered values less than  $t$ . So this is distributed as binomial with parameters  $n$  and  $p = F(t)$ .

**Example 7.6.** Suppose that shocks occur according to a Poisson process with rate  $\lambda$ . Let  $D_i$  be the damage caused by shock  $i$ , where  $\{D_i\}$  are i.i.d. and independent of  $\{N(t)\}$ . The damages dissipate at an exponential rate  $\alpha$ ; that is, damage of value  $d$  has value  $de^{-as}$  after a time  $s$ . Damages are cumulative. What is the total damage by time  $t$ ?

**Solution.** We have

$$D(t) = \sum_{i=1}^{N(t)} D_i e^{-a(t-S_i)}$$

so

$$\begin{aligned}\mathbb{E}(D(t) \mid N(t) = n) &= \mathbb{E}\left(\sum_{i=1}^{N(t)} D_i e^{-a(t-S_i)} \mid N(t) = n\right) = \sum_{i=1}^{N(t)} \mathbb{E}(D_i e^{-a(t-S_i)} \mid N(t) = n) \\ &= \sum_{i=1}^{N(t)} \mathbb{E}(D_i) \mathbb{E}(e^{-a(t-S_i)} \mid N(t) = n) = \mathbb{E}(D_i) e^{-at} \sum_{i=1}^n \mathbb{E}(e^{aS_i} \mid N(t) = n)\end{aligned}$$

Let  $\{U_i\}$  be i.i.d. random variables on  $[0, t]$ . Then by Theorem 7.2.9, we can write this as

$$\begin{aligned}&= \mathbb{E}(D_i) e^{-at} \sum_{i=1}^n \mathbb{E}(e^{aU_i} \mid N(t) = n) = \mathbb{E}(D_i) e^{-at} \cdot n \int_0^t \frac{e^{ax}}{t} dx = \mathbb{E}(D_i) e^{-at} \cdot \frac{n}{at} (e^{at} - 1) \\ &= \mathbb{E}(D_i) \cdot \frac{n}{at} (1 - e^{-at})\end{aligned}$$

So finally we have

$$\mathbb{E}(D(t)) = \mathbb{E}[\mathbb{E}(D(t) \mid N(t) = n)] = \mathbb{E}\left[\mathbb{E}(D_i) \cdot \frac{N(t)}{at} (1 - e^{-at})\right] = \mathbb{E}(D_i) \cdot \frac{\mathbb{E}(N(t))}{at} (1 - e^{-at}) = \boxed{\mathbb{E}(D_i) \cdot \frac{\lambda}{a} (1 - e^{-at})}$$

### 7.2.2 Time Sampling Poisson Processes

**Proposition 7.2.10 (Time Sampling (Proposition 5.3 in Introduction to Probability Models)).** Suppose we have events that happen as a Poisson process with rate  $\lambda$ . Each event is of type  $1, \dots, r$  independently of what has come before. An event at time  $s$  is type  $i$  with probability  $P_i(s)$ . Let  $N_i(t)$  be the number of type  $i$  events by time  $t$ . Then  $N_1(t), \dots, N_r(t)$  are independent Poisson random variables with mean  $\mathbb{E}[N_i(t)] = \lambda \int_0^t P_i(s) ds$ .

*Proof.* I'm not sure this proof makes sense? Note that

$$\Pr(N_1(t) = n_1, \dots, N_r(t) = n_r) = \Pr(N_1(t) = n_1, \dots, N_r(t) = n_r \mid N(t) = \sum_{i=1}^r n_i) \cdot \Pr\left(N(t) = \sum_{i=1}^r n_i\right).$$

Given  $N(t) = n$ , the  $n$  event times are i.i.d.  $U(0, t)$  by Theorem 7.2.9. Let  $P_i$  be the probability that a particular event is of type  $i$  given a time  $t$  and the fact that  $N(t) = n$ . Then

$$P_i = \int_0^t P_i(s) \cdot f_U(s) ds = \int_0^t \frac{P_i(s)}{t} ds$$

Since  $\Pr\left(N(t) = \sum_{i=1}^r n_i\right) = e^{-\lambda t}$ , we have

$$\Pr(N_1(t) = n_1, \dots, N_r(t) = n_r) = \frac{n!}{n_1! \cdots n_r!} P_1^{n_1} \cdots P_r^{n_r} e^{-\lambda t}$$

□

**Remark.** Notation for Queueing: in many cases we suppose times between successive arrivals are independent with a common distribution  $F$ . Then it is a renewal process. (If they are exponential, we get a Poisson process.) We have the following notation for a renewal process if  $k$  number of servers and  $G$  is the distribution of the service time:

$$F/G/k$$

If the distribution is exponential, we use  $M$  (for “memoryless” or “Markovian”). ( $E$  stands for Erlang, sum of i.i.d. exponentials).

**Example 7.7.** We have an  $M/G/\infty$  queuing system. Arrivals are a Poisson process with rate  $\lambda$ . Fix  $t$  and let  $X(t)$  be the number of people in the system at time  $t$ . Let  $Y(t)$  be the number of people who have completed service at time  $t$ . Let the events be arrivals of customers, and call it a Type 1 event if the customer is still in the system at time  $t$ , and Type 2 if the customer completes service by time  $t$  (and Type 3 otherwise).

Suppose a customer arrives at time  $s < t$ . Then the customer is Type 1 if their service time exceeds  $t - s$ , which happens with probability  $P_1(s) = \bar{G}(t - s)$ . Similarly  $P_2(s) = G(t - s)$ . Let  $N_1(t)$  be the number of Type I events that happen by time  $t$  and similarly for  $N_2(t)$ . Then by Proposition 7.2.10 we have

$$X(t) = N_1(t) \sim \text{Poisson} \left( \lambda \int_0^t \bar{G}(t-s) ds \right) = \text{Poisson} \left( \lambda \int_0^t \bar{G}(y) dy \right)$$

and

$$Y(t) = N_2(t) \sim \text{Poisson} \left( \lambda \int_0^t G(t-s) ds \right) = \text{Poisson} \left( \lambda \int_0^t G(y) dy \right)$$

which implies the independence of  $X(t)$  and  $Y(t)$ .

### 7.2.3 Nonstationary Poisson Processes

**Definition 7.10 (Nonstationary Poisson process.).** The counting process  $\{N(t), t \geq 0\}$  is said to be a **nonstationary Poisson process** (or **nonhomogeneous Poisson process**) with intensity function  $\lambda(t), t \geq 0$  if

- (i)  $N(0) = 0$ .
- (ii)  $\{N(t)\}$  has independent increments.

- (iii)  $\Pr(N(t+h) - N(t) = 1) = \lambda(t)h + o(h).$
- (iv)  $\Pr(N(t+h) - N(t) \geq 2) = o(h).$

**Remark.** Note the similarities to Definition 7.6.

**Lemma 7.2.11.** Let  $\{N(t), t \geq 0\}$  be a nonstationary Poisson process. Let  $N_s(t) = N(s+t) - N(s)$ . Then  $\{N_s(t), t \geq 0\}$  is a nonstationary Poisson Process with intensity  $\lambda_s(t) = \lambda(s+t)$ .

**Remark.** Note the similarity to Lemma 7.2.1.

*Proof.* Note the parts of Definition 7.10 above:

- (i)  $N_s(0) = N(s) - N(s) = 0.$
- (ii)  $\{N_s(t)\} = \{N(s+t) - N(s)\}$  has independent increments since  $N(t)$  has independent increments.
- (iii) **Not sure about this part?**  $\Pr(N_s(t+h) - N_s(t) = 1) = \Pr(N(s+t+h) - N(s) - [N(s+t) - N(s)] = 1) = \Pr(N(s+t+h) - N(s+t) = 1) = \lambda(t)h + o(h).$
- (iv)  $\Pr(N_s(t+h) - N_s(t) \geq 2) = o(h)$  by similar argument to (iii).

□

**Definition 7.11.** Let  $m(t) = \int_0^t \lambda(s)ds$ . Note that  $m'(t) = \lambda(t)$ . We call  $m(t)$  the **mean value function**.

**Remark.** Note that the mean value function for  $N_s(t)$  is

$$m_s(t) = \int_0^t \lambda_s(y)dy = \int_0^t \lambda(s+y)dy = \int_s^{s+t} \lambda(x)dx = m(t+s) - m(s).$$

**Lemma 7.2.12.** Let  $N(t)$  be a nonstationary Poisson process. Then  $\Pr(N(t) = 0) = e^{-m(t)}$  for any  $t \geq 0$ .

**Remark.** Note the similarity to Lemma 7.2.2.

*Proof.* Let  $P(t) = \Pr(N(t) = 0)$ . Then by independent increments

$$\Pr(N(t+h) = 0) = \Pr(N(t) = 0 \cap N(t+h) - N(t) = 0) = \Pr(N(t) = 0) \cdot \Pr(N(t+h) - N(t) = 0) \quad (7.5)$$

Let  $P_0(t+h) = \Pr(N(t+h) = 0)$ . Then using Definition 7.10 we have

$$\Pr(N(t+h) - N(t) = 0) = 1 - \Pr(N(t+h) - N(t) = 1) - \Pr(N(t+h) - N(t) \geq 2) = 1 - \lambda(t+h) + o(h)$$

so we can write (7.5) as

$$P_0(t+h) = P_0(t)[1 - \lambda(t+h) + o(h)]$$

$$\iff \frac{P_0(t+h) - P_0(t)}{h} = -\lambda(t+h) \frac{P_0(t)}{h} + \frac{o(h)}{h}$$

Taking the limit as  $h \rightarrow 0$  yields

$$P'_0(t) = -\lambda(t)P_0(t) \iff \int_0^s \frac{P'_0(t)}{P_0(t)} ds = \int_0^s -\lambda(t) dt \iff \log P_0(t) = \int_0^s -\lambda(t) dt \Big|_0^s$$

$$\iff P_0(s) = e^{-m(s)}$$

□

**Remark (Interarrival times).** Let  $T_1$  be the time of the first event. Note that  $T_1 > t \iff N(t) = 0$ . So

$$\bar{F}_{T_1}(t) = \Pr(T_1 > t) = \Pr(N(t) = 0) = e^{-m(t)}$$

which means

$$f_{T_1}(t) = \frac{d}{dt}(-\bar{F}_{T_1}(t)) = m'(t)e^{-m(t)} = \lambda(t)e^{-m(t)}.$$

Note that because  $\lambda(t)$  is not constant, the interarrival times are not i.i.d.

**Theorem 7.2.13.** Let  $N(t)$  be the counting process for a nonstationary Poisson process. Then  $N(t) \sim \text{Poisson}(m(t))$ .

*Proof.* We must show that

$$\Pr(N(t) = n) = e^{-m(t)} \frac{(m(t))^n}{n!}, \quad n = 0, 1, \dots$$

We have already shown that  $\Pr(N(t) = 0) = e^{-m(t)} \frac{(m(t))^0}{0!} = e^{-m(t)}$  this is true when  $n = 0$  in Lemma 7.2.12. We will show that this expression holds for  $n = 1, 2, \dots$  by induction. Assume for a fixed  $n$

$$\Pr(N(t) = n) = e^{-m(t)} \frac{(m(t))^n}{n!}.$$

We seek

$$\Pr(N(t) = n+1) = \int_0^t \Pr(N(t) = n+1 \mid T_1 = s) f_{T_1}(s) ds \tag{7.6}$$

Note that (using the property of independent increments)

$$\Pr(N(t) = n+1 \mid T_1 = s) = \Pr(N(t) - N(s) = n \mid T_1 = s) = \Pr(N(t) - N(s) = n)$$

$$= \Pr(N_s(t-s) = n)$$

By Lemma 7.2.11 above,  $N_s(\cdot)$  is a Poisson process. So using that and the induction hypothesis, we have

$$\Pr(N(t) = n+1 \mid T_1 = s) = e^{-m_s(t-s)} \frac{(m_s(t-s))^n}{n!} = e^{-[m(t)-m(s)]} \frac{[m(t)-m(s)]^n}{n!}.$$

Substituting this expression back in to (7.6) and using  $f_{T_1}(t) = \lambda(t)e^{-m(t)}$ , we have

$$\begin{aligned} \Pr(N(t) = n+1) &= \int_0^t e^{-[m(t)-m(s)]} \frac{[m(t)-m(s)]^n}{n!} \cdot \lambda(s)e^{-m(s)} ds \\ &= \frac{e^{-m(t)}}{n!} \int_0^t [m(t)-m(s)]^n \cdot \lambda(s) ds \end{aligned}$$

Substituting  $y = m(t) - m(s) \implies dy = -\lambda(s)ds$ , we have

$$= \frac{e^{-m(t)}}{n!} \int_{m(t)}^0 -y^n dy = \frac{e^{-m(t)}}{n!} \int_0^{m(t)} y^n dy = e^{-m(t)} \frac{(m(t))^{n+1}}{(n+1)!}.$$

Therefore the result follows by induction.  $\square$

**Corollary 7.2.13.1.**  $N(t+s) - N(s) \sim \text{Poisson}(m(t+s) - m(s)) = \text{Poisson}(\int_s^{s+t} \lambda(y) dy)$ .

*Proof.* Follows almost immediately from Theorem 7.2.13, since if  $N(t) \sim \text{Poisson}(m(t))$ ,

$\square$

**Proposition 7.2.14.** Suppose events occur according to a Poisson process with rate  $\lambda$ .  $\{N(t)\}$  is the counting process. An event at time  $s$  is type 1 with probability  $p(s)$ . Let  $N_1(t)$  be the number of type 1 events by time  $t$ . Then  $\{N_1(t)\}$  is a nonhomogeneous Poisson Process with intensity  $\lambda(t) = \lambda p(t)$ .

**Remark.** Note the similarity to Proposition 7.2.10.

*Proof.* Note the parts of Definition 7.10 above:

- (i)  $N_1(0) = 0$ : yes.
- (ii)  $\{N_1(t)\}$  has independent increments: yes, follows pretty much immediately.
- (iii) Since the probability of exactly one event in the interval that happens to be type 1 is  $\lambda h P_1(t)$ , we have  $\Pr(N_1(t+h) - N_1(t) = 1) = \lambda h P_1(t) = \lambda(t)h + o(h)$ .
- (iv)  $\Pr(N_1(t+h) - N_1(t) \geq 2) = o(h)$  by similar argument to (iii).

$\square$

So now we have that  $N_1(t) \sim \text{Poisson}(\int_0^t \lambda p_1(s)ds)$ . So if every time an event happens it is of type  $i$  with a certain probability that varies over time, the counting process is for each  $i$  is a nonstationary Poisson process, and all the processes are independent. (Another way to understand the time sampling result from before.)

What if the arrival process is nonstationary and the probability of a type  $i$  event is also time varying?  $\lambda(t), p(t)$ . It's a nonstationary PP with intensity  $\lambda(t)p(t)$ . (Probability in an interval is  $\lambda(t)hp(t) + o(h)$ .)

One more result about  $M/G/\infty$  processes:

**Theorem 7.2.15 (Example 5.25 in *Introduction to Probability Models*).** The **departure process** from an  $M/G/\infty$  queue is a nonstationary Poisson process with intensity function  $\lambda(t) = \lambda G(t)$ .

*Proof.* Let  $D(t)$  be the number of departures by time  $t$ . Examine the axioms of Definition 7.10 above:

- (i)  $D(0) = 0$ : yes.
- (ii)  $\{D(t)\}$  has independent increments: think of each arrival as an event. Types: type 1 if they depart in interval 1, 2 if they depart in 2, 3 if they depart in 3, 4 if they depart elsewhere, where 1, 2, and 3 are sequential nonoverlapping intervals. Note that for a given arrival time, the type of the event is dependent only on the service time. Since the arrival times are independent by assumption and the service times are also independent, the increments are independent.
- (iii)  $\Pr(D(t+h) - D(t) = 1)$ : Call an event type 1 if they depart in the interval  $(t, t+h)$ . Then  $P_1(s) = G(t+h+s) - G(t+s) = g(t-s) \cdot h + o(h)$ . So

$$D(t+h) - D(t) \sim \text{Poisson}(\lambda h \int_0^t g(t-s)ds + o(h)) = \text{Poisson}(\lambda h \int_0^t g(y)dy + o(h)) \text{Poisson}(\lambda h G(t) + o(h))$$

which means  $\Pr(D(t+h) - D(t) = 1) = \lambda G(t)h e^{-\lambda G(t)h} =$  (by Taylor expansion of exp)  $\lambda G(t)h + o(h)$  which is what we wanted to show.

- (iv)  $\Pr(D(t+h) - D(t) \geq 2) = 1 - (\Pr(D(t+h) - D(t) = 0) - (\Pr(D(t+h) - D(t) = 1) = 1 - e^{-\lambda G(t)h} - \lambda G(t)h + o(h) =$  (by Taylor expansion of exp)  $1 + \lambda G(t)h - \lambda G(t)h + o(h) = o(h)$ .

□

**Remark.** Notice in the limit as  $t \rightarrow \infty$  it converges to a Poisson process with rate  $\lambda$ .

**Example 7.8.** Poisson process, event  $i$  is associated with reward  $V_i$ .  $\{X(t), t \geq 0\}$  is amount of money you get at time  $t$ .

$$X(t) = \sum_{i=1}^r V_i N_i(t) \tag{7.7}$$

or

$$\mathbb{E}(X(t)) = \sum V_i \mathbb{E}(V_i(t)) = \sum V_i \lambda \alpha_i t = \lambda t \sum \alpha_i v_i = \lambda t \mathbb{E}(X)$$

nice thing about this representation is  $V_i N_i(t)$  are independent, so variance is sum of variances. Using  $\text{Var}(N_i(t)) = \lambda \alpha_i t$ ,

$$\text{Var}(X(t)) = \sum_{i=1}^r V_i^2 \lambda \alpha_i t = \lambda t \sum V_i^2 \alpha_i = \lambda t \mathbb{E}(X^2)$$

so we verified the formulas we derived a different way.

Note that  $N_i(t)$  is approximately Gaussian for large  $t$  (because Poisson goes to normal for large  $t$ ). Therefore when  $t$  is large  $X(t)$  is also approximately Gaussian.

#### 7.2.4 Queueing Systems

Say we have a  $M/G/1$  queuing process. That is, arrivals are  $PP(\lambda)$ , the service time has distribution  $G$ , and there is one server. Note that if no one is in line, the waiting time (length of an “idle period”) until the next arrival is exponential (since it’s a Poisson process). Consider the “busy periods,” that is, the time from when a customer arrives until the time the next idle period starts.

Let  $B$  be the length of a busy period. Let  $S$  be the service time of the initial customer. Note that  $B$  is independent from what happened before. Let  $A$  be the additional time after the first customer is served until the next idle period so that  $B = S + A$ . Let  $N(S)$  be the number of arrivals during  $S$ . Note that  $A$  depends on  $N(S)$ .

Suppose  $N(S) = n$ . If  $n = 0$ , then  $A = 0$ . If  $n = 1$ ,  $A$  has the same distribution as  $B_1$ . If  $n = 2$ ,  $A$  has the same distribution as the sum of two  $B$ s. And so on. This tells us we can say

$$B = S + \sum_{i=1}^{N(S)} B_i$$

If we condition on  $S = s$ ,

$$\{B \mid S = s\} = s + \sum_{i=1}^{N(s)} B_i$$

Note that  $\sum_{i=1}^{N(s)} B_i$  is a compound Poisson random variable. So

$$\mathbb{E}(B \mid S = s) = s + \mathbb{E}(N(s))\mathbb{E}(B_i) = s + \mathbb{E}(B)\lambda s$$

$$\text{Var}(B \mid S = s) = \lambda s \mathbb{E}(B^2)$$

or

$$\mathbb{E}(B | S) = S + \mathbb{E}(N(S))\mathbb{E}(B_i) = S + \mathbb{E}(B)\lambda S$$

$$\text{Var}(B | S) = \lambda S \mathbb{E}(B^2)$$

So

$$\mathbb{E}(B) = \mathbb{E}[\mathbb{E}(B | S)] = \mathbb{E}(S) + \lambda \mathbb{E}(S)\mathbb{E}(B) \implies \boxed{\mathbb{E}(B) = \frac{\mathbb{E}(S)}{1 - \lambda \mathbb{E}(S)}}$$

Note the similarity to the sum of an infinite geometric series. Similarly, if  $\lambda \mathbb{E}(S) \geq 1 \iff \mathbb{E}(S) \geq 1/\lambda$  then the expected waiting time is infinite because the arrival times under the Poisson process ( $1/\lambda$  in expectation) are faster in expectation than the service times. Also,

$$\text{Var}(B) = \lambda \mathbb{E}(S)\mathbb{E}(B^2) + (1 + \lambda \mathbb{E}(B))^2 \text{Var}(S)$$

Note that  $\text{Var}(B) = \mathbb{E}(B)^2$ , etc., then you work out the answer (see textbook for complete answer.)

### End of Poisson processes

## 7.3 Renewal Processes

**Definition 7.12 (Stieltjes Integral).**

$$\int_a^b h(x)dF(x) = \lim_{n \rightarrow \infty} \sum_{i=1}^n h(x_i)(F(x_i) - F(x_{i-1}))$$

In particular, we often suppose that  $F(x)$  is the distribution function for a random variable  $X$ . Then we have (if  $X$  is continuous)

$$\int_a^b h(x)dF(x) = \lim_{n \rightarrow \infty} \sum_{i=1}^n h(x_i)(\Pr(x_{i-1} < X \leq x_i)) = \int_a^b h(x)f_X(x)dx = \mathbb{E}(h(x))$$

or if  $X$  is discrete

$$\int_a^b h(x)dF(x) = \lim_{n \rightarrow \infty} \sum_{i=1}^n h(x_i)(\Pr(x_{i-1} < X \leq x_i)) = \mathbb{E}(h(x)).$$

Generalization of Poisson processes. A counting process where the interarrival times are i.i.d. Let  $X_1, X_2, \dots$  be i.i.d. non-negative random variables with distribution function  $F$ . We require  $F(0) = \Pr(X \leq 0) = \Pr(X = 0) < 1$ . Let

$$\mathbb{E}(X_1) = \mu = \int_0^\infty x dF(x) = \int_0^\infty \bar{F}(t) dt$$

where  $0 < \mu \leq \infty$ . Let  $N(t)$  be the counting process (this is the largest value of  $n$  for which the  $n$ th event has occurred at time  $t$ , so  $N(t) = \max\{n : S_n \leq t\}$ ). Let  $S_n$  be the arrival time for the  $n$ th event. Define  $S_0 = 0$ ,  $S_n = \sum_{i=1}^n X_i$ .

**Definition 7.13 (Renewal process; Definition 3.1.1 in *Stochastic Processes*).** Let  $\{X_n, n = 1, 2, \dots\}$  be a sequence of nonnegative independent random variables with a common distribution  $F$ . To avoid trivialities, suppose that  $F(0) = \Pr(\{X_n = 0\}) < 1$ . We shall interpret  $X_n$  as the time between the  $(n - 1)$ st and  $n$ th event. Let

$$S_0 = 0, \quad S_n = \sum_{i=1}^n X_i, n \geq 1$$

so that  $S_n$  is the time of the  $n$ th event. As the number of events by time  $t$  will equal the largest value of  $n$  for which the  $n$ th event occurs before or at time  $t$ , we have that  $N(t)$ , the number of events by time  $t$ , is given by

$$N(t) = \sup\{n : S_n \leq t\}.$$

The counting process  $\{N(t), t \geq 0\}$  is called a **renewal process**. We can write this as  $\{N(t), t \geq 0\}$  as a renewal process with intensity distribution  $F$ .

**Remark.** Let  $\mu = \mathbb{E}(X_n) = \int_0^\infty x dF(x)$  denote the mean time between successive events and note that from the assumptions  $X_n \geq 0$  and  $F(0) < 1$  it follows that  $0 < \mu \leq \infty$ .

Every time an event occurs is a “renewal:” from that time on, all the arrivals are i.i.d. (Between events it is unclear what is going on until the next event happens; depends on the nature of  $F$ . If we have a Poisson process, then it is memoryless, but otherwise it is not.)

**Example 7.9 (St. Petersburg Paradox).** Idea: you play a game, you get the amount of money  $X$ . In order to play the game you have to pay, so what’s a fair amount to pay? Belief:  $\mathbb{E}(X)$  (that’s a rational price).

Game: fair coin, flip until heads occurs. If it occurs on trial  $n$ , you win  $2^n$  dollars. Note that

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} 2^i \cdot \left(\frac{1}{2}\right)^i = \infty.$$

**Proposition 7.3.1.** With probability 1,  $N(t) < \infty$  for all  $t$ .

*Proof.*

$$\frac{S_n}{n} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{a.s.} \mu > 0 \text{ as } n \rightarrow \infty$$

Because the denominator goes to  $\infty$  and the ratio doesn't go to 0, we must have  $S_n \xrightarrow{a.s.} \infty$ . Therefore  $\Pr(N(t) < \infty) = 1$  for all  $t$ .

□

**Proposition 7.3.2.** With probability 1,  $N(\infty) = \lim_{t \rightarrow \infty} N(t) = \infty$ . (There's never a last renewal—there will always be another.)

*Proof.* Recall **Boole's Inequality** (which can be proven in a similar manner to Theorem 7.6.17):

$$\Pr\left(\bigcup_{n=1}^{\infty}\{A_n\}\right) \leq \sum_{n=1}^{\infty}\Pr(A_n)$$

(with equality if the events are disjoint). Then

$$\Pr(N(\infty) < \infty) = \Pr(X_n = \infty) \text{ (for some } n) = \Pr\left(\bigcup_{n=1}^{\infty}\{X_n = \infty\}\right) \leq \sum_{n=1}^{\infty}\Pr(X_n = \infty) = 0$$

□

**Definition 7.14 (Notation for  $n$ -fold self-convolution).** Recall the notation for convolution: if  $X \sim F \perp\!\!\!\perp Y \sim G$ , we have  $X + Y \sim F * G$ . Then

$$S_n \sim F * F * \dots * F(t).$$

Let

$$F_n(t) := F * F * \dots * F(t)$$

the  $n$ -fold convolution of  $F$  with itself. (In general, this is very hard to get a closed-form notation for except in some simple cases. More of a theoretical construct than something practical to compute.)

**Proposition 7.3.3.**  $\Pr(N(t) = n) = F_n(t) - F_{n+1}(t)$ .

*Proof.*

$$N(t) \geq n \iff S_n \leq t.$$

$$\implies \Pr(N(t) \geq n) = \Pr(S_n \leq t)$$

Then we have

$$\Pr(N(t) = n) = \Pr(N(t) \geq n) - \Pr(N(t) \geq n+1) = F_n(t) - F_{n+1}(t).$$

(Again, this is more of a theoretical construct than something practical to compute.)

□

**Definition 7.15 (Renewal function).** Let  $m(t) = \mathbb{E}(N(t))$ . We call  $m(t)$  the **renewal function**.

**Remark.**

$$m(t) = \sum_{n=1}^{\infty} \Pr(N(t) \geq n) = \sum_{n=1}^{\infty} F_n(t).$$

**Lemma 7.3.4.**  $m(t) < \infty$  for all  $t$ .

*Proof.* Skipped in class, not very enlightening.

□

**Proposition 7.3.5 (Renewal Equation).**

$$m(t) = F(t) + \int_0^t m(t-s)dF(s).$$

*Proof.*

$$m(t) = \mathbb{E}(N(t)) = \int_0^{\infty} \mathbb{E}(N(t) \mid X_1 = s)dF(s)$$

Note that

$$\mathbb{E}(N(t) \mid X_1 = s) = \begin{cases} 1 + m(t-s) & s \leq t \\ 0 & s > t \end{cases}$$

so we have

$$m(t) = \int_0^t (1 + m(t-s))dF(s) = F(t) + \int_0^t m(t-s)dF(s)$$

□

**Remark.** Remember we did a problem where  $X_i \sim \text{Unif}(0, 1)$  and asked how many you have to sum until the number of greater than 1. We ended up with  $1/e$ . This is the smallest value of  $n$  for which the event occurred after time  $n$  (or something?)

$$N = \min\{X_1 + \dots + X_n > 1\} = N(1) + 1$$

(the first event that occurred after time 1. We showed that this was equal to  $e$  basically by solving the renewal equation. Only feasible because of uniform distribution between 0 and 1.)

**Remark.** Note that

$$S_{N(t)} := (\text{time of the most recent event before or at time } t),$$

$$S_{N(t+1)} := (\text{time of the first event after time } t).$$

At what rate do renewals occur?

**Theorem 7.3.6 (Strong Law for Renewal Processes).**

$$\frac{N(t)}{t} \xrightarrow{\text{a.s.}} \frac{1}{\mu}$$

(Intuitively: since  $X_i$  are the time between renewals, the average time between renewals is  $\mathbb{E}(X_i) = \mu$ . So the average rate of renewals is equal to one over the average time between events.)

*Proof.* Note that

$$S_{N(t)} \leq t < S_{N(t)+1}$$

$$\iff \frac{S_{N(t)}}{N(t)} \leq \frac{t}{N(t)} < \frac{S_{N(t)+1}}{N(t)}$$

As  $t \rightarrow \infty$ , the quantity on the left converges to the mean by the Strong Law of Large Numbers (see the proof of Proposition 7.3.1 above). For the quantity on the right, we have

$$\frac{S_{N(t)+1}}{N(t)} = \frac{S_{N(t)+1}}{N(t) + 1} \cdot \frac{N(t) + 1}{N(t)} \xrightarrow{\text{a.s.}} \mu \cdot 1$$

which means

$$\frac{t}{N(t)} \xrightarrow{\text{a.s.}} \mu \iff \frac{N(t)}{t} \xrightarrow{\text{a.s.}} \frac{1}{\mu}.$$

□

**Remark.** In many applications:  $X_1, X_2, \dots$  are independent non-negative random variables.  $X_1 \sim G$ ,  $X_i \sim F, i \geq 1$ . (That is, once the first event occurs we have a renewal process, but before then we don't.)

**Definition 7.16 (Delayed renewal process).** Define

$$N_d(t) := \max\{n : X_1 + \dots + X_n \leq t\}.$$

We call  $\{N_d(t), t \geq 0\}$  a **delayed renewal process**.

**Remark.** Almost all limiting results for renewal processes apply for delayed renewal processes as well. (In the long run, that first waiting period doesn't really make a difference.) For example, consider the Strong Law for Renewal Processes (Theorem 7.3.6) (quick note: let  $N(s) = -1, s < 0$ ):

$$N_d(t) = 1 + N(t - X_1) \iff \frac{N_d(t)}{t} = \frac{1}{t} + \frac{N(t - X_1)}{t - X_1} \frac{t - X_1}{t}$$

By the Strong Law for Renewal processes,

$$\frac{N(t - X_1)}{t - X_1} \xrightarrow{\text{a.s.}} \frac{1}{\mu_F}$$

where  $\mu_F = \int_0^\infty \overline{F}(t)dt$ . Since  $\lim_{t \rightarrow \infty} \frac{t - X_1}{t} = 1$ , we have the same result as for the Strong Law for Renewal Processes:

$$\frac{N_d(t)}{t} \xrightarrow{\text{a.s.}} \frac{1}{\mu_F}$$

**Example 7.10.** Arrivals to a single server queue according to a Poisson process with rate  $\lambda$ . However, they will only enter if the server is free when they arrive. Service time has distribution  $G$ . (This is often called a **loss model** or **Erlang loss model**, more generally with  $k$  servers.)

- (a) At what rate do customers enter the system?
- (b) What proportion of arrivals enter the system?

**Solution.**

- (a) Note that the events of customers entering are a renewal process because everything probabilistically starts all over again when a customer arrived. Specifically it's a delayed renewal process because the time of the first arrival is exponential with rate  $\lambda$ , but the rest of them are more complicated because there must be a service and an arrival (the sum of two random variables). So the rate of arrivals is 1 over the expected time between events, or

$$\frac{1}{\mathbb{E}(S + I)}$$

where  $S$  is the service time and  $I$  is the interarrival time for the next customer. Let  $\mathbb{E}(X) = \mu_G = \int_0^\infty \overline{G}(t)dt$ . Then

$$\mathbb{E}(S + I) = \mu_G + \frac{1}{\lambda}$$

so the answer is

$$\frac{1}{\mu_G + \frac{1}{\lambda}} = \boxed{\frac{\lambda}{1 + \lambda\mu_G}}.$$

- (b) Rate of service arrivals divided by overall rates of arrivals:

$$\frac{\lambda}{1 + \lambda\mu_G} / \lambda = \boxed{\frac{1}{1 + \lambda\mu_G}}.$$

**Example 7.11.** Suppose we have a bin with an infinite number of coins. Each coin has value  $p$  of landing on heads, and suppose the value of  $p$  for a randomly chosen coin is distributed as Uniform(0, 1). We draw coins and flip them. Every turn we can either draw a new coin or flip one of the coins we already have. If we want to maximize the proportion of coins that flip heads, what is the optimal strategy?

**Solution.** Consider a strategy of giving up on a coin if it comes up tails once. Every time you flip tails, the process renews. Let  $N(n)$  be the number of tails in the first  $n$  flips. Then  $N(n)$  is a renewal process. So we know that

$$\frac{N(n)}{n} \xrightarrow{a.s.} \frac{1}{\mathbb{E}(T)}$$

where  $T$  is the time between tails. Let  $P$  be a random variable for the probability of a coin flipping heads and note that

$$\mathbb{E}(T) = \int_0^1 \mathbb{E}(T | P = p) dp = \int_0^1 \frac{1}{1-p} dp = -\log(1-p) \Big|_0^1 = \infty$$

so the long-run proportion of coins that land heads is 1.

**Proposition 7.3.7 (Homework problem; *Introduction to Probability Models* Ch. 7 Problem 30).** For a renewal process, let  $A(t) = t - S_{N(t)}$  be the age at time  $t$ . If  $\mu < \infty$ , then

$$\frac{A(t)}{t} \xrightarrow{w.p.1} 0.$$

*Proof.* We hope to show that

$$\frac{t - S_{N(t)}}{t} \xrightarrow{w.p.1} 0.$$

Note that

$$\frac{t - S_{N(t)}}{t} = \frac{t - S_{N(t)}}{N(t)} \cdot \frac{N(t)}{t} = \left( \frac{t}{N(t)} - \frac{S_{N(t)}}{N(t)} \right) \frac{N(t)}{t}$$

By the Strong Law for Renewal Processes (Theorem 7.3.6),  $\frac{N(t)}{t} \xrightarrow{w.p.1} \mu^{-1}$ , where  $\mu = \mathbb{E}(X)$  is the expected interarrival time. Since the expected interarrival time is finite,  $N(t) \rightarrow \infty$  as  $t \rightarrow \infty$ . Therefore

$$\frac{S_{N(t)}}{N(t)} = \frac{X_1 + \dots + X_{N(t)}}{N(t)} \xrightarrow{w.p.1} \mu.$$

Lastly, since  $N(t) > 0$  with probability 1 as  $t \rightarrow \infty$  and  $t > 0$ , by the Continuous Mapping Theorem (Theorem asym.contmappthm),

$$\frac{t}{N(t)} = \left( \frac{N(t)}{t} \right)^{-1} \xrightarrow{w.p.1} (\mu^{-1})^{-1} = \mu.$$

Therefore by another application of the Continuous Mapping Theorem (Theorem asym.contmappthm),

$$\frac{t - S_{N(t)}}{t} = \left( \frac{t}{N(t)} - \frac{S_{N(t)}}{N(t)} \right) \frac{N(t)}{t} \xrightarrow{w.p.1} (\mu - \mu) \cdot \mu^{-1} = 0.$$

□

### 7.3.1 Stopping Times

**Definition 7.17 (Stopping times).** Let  $X_1, X_2, \dots$  be independent random variables. We say the non-negative integer valued random variable  $N$  is a **stopping time** for  $X_1, X_2, \dots$  if the event  $\{N = n\}$  is independent of  $X_{n+1}, X_{n+2}, \dots$ . (This definition is more general than saying it must depend on previous times because it could be random and not depend on anything (except the current one) or you could have a finite memory, etc.)

**Theorem 7.3.8 (Wald's Equation).** Suppose  $X_1, X_2, \dots$  are i.i.d. with  $\mathbb{E}(X) < \infty$ . Let  $N$  be a stopping time for  $X_1, X_2, \dots$  such that  $\mathbb{E}(N) < \infty$ . Then

$$\mathbb{E}\left(\sum_{i=1}^N X_i\right) = \mathbb{E}(N)\mathbb{E}(X).$$

**Remark.** Note the similarity to compound random variables. But the proof differs because  $N$  is not independent of the  $\{X_i\}$ .

*Proof.* Let  $I_i$  be an indicator variable for  $I \leq N$ . Then

$$\sum_{i=1}^N X_i = \sum_{i=1}^{\infty} X_i I_i \implies \mathbb{E}\left(\sum_{i=1}^N X_i\right) = \mathbb{E}\left(\sum_{i=1}^{\infty} X_i I_i\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i I_i)$$

Note that  $I_i$  depends on  $X_1, \dots, X_{i-1}$  but not on  $X_i$  (it only depends on whether you play the  $i$ th game, not the outcome of that game. You only decide whether to play the  $i$ th game based on the outcomes of the previous games.). Therefore  $X_i \perp\!\!\!\perp I_i$ . So we have

$$= \sum_{i=1}^{\infty} \mathbb{E}(X_i) \mathbb{E}(I_i) = \mathbb{E}(X) \sum_{i=1}^{\infty} \mathbb{E}(I_i) = \mathbb{E}(X) \sum_{i=1}^{\infty} \Pr(N \geq i) = \mathbb{E}(X) \mathbb{E}(N)$$

or we can write

$$\mathbb{E}(X) \sum_{i=1}^{\infty} \mathbb{E}(I_i) = \mathbb{E}(X) \mathbb{E}\left(\sum_{i=1}^{\infty} I_i\right) = \mathbb{E}(X) \mathbb{E}(N).$$

□

**Remark.** How do we justify

$$\mathbb{E}\left(\sum_{i=1}^{\infty} X_i I_i\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i I_i)?$$

Do the same thing again but replace all the  $X_i$  with  $|X_i|$ . Then the proof works, and Wald's Equation follows by Lebesgue's Dominated Convergence Theorem.

**Example 7.12.** Let

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{otherwise} \end{cases}$$

One stopping time could be  $N_1 = \min\{n; X_1 + \dots + X_n = k\}$ . Then by Wald's Equation (Theorem 7.3.8),

$$\mathbb{E}\left(\sum_{i=1}^N X_i\right) = \mathbb{E}(N)\mathbb{E}(X) \iff k = \mathbb{E}(N)(p) \implies \mathbb{E}(N) = \frac{k}{p}.$$

Another could be  $N_2 = \min\{n : X_{n-1} = X_n = 1\}$ .

**Example 7.13.** Let

$$X_i = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases}$$

with  $p > 1/2$ . Note that  $\mathbb{E}(X_i) = 2p - 1 > 0$ . One stopping time could be  $N = \min\{n : X_1 + \dots + X_n = 10\}$ . (Note that by the Strong Law of Large Numbers, with probability 1 this will eventually happen.) Then by Wald's Equation (Theorem 7.3.8),

$$\mathbb{E}\left(\sum_{i=1}^N X_i\right) = \mathbb{E}(N)\mathbb{E}(X) \iff 10 = \mathbb{E}(N)(2p - 1) \implies \mathbb{E}(N) = \frac{10}{2p - 1}.$$

What's the mean amount of time until you're up by one dollar?

$$m = 1 + (1-p)\mathbb{E}(\text{up } 2) = 1 + (1-p)2m$$

Possible stopping rule: stop when you're winning money.

$$1 = \sum_{i=1}^N X_i = \mathbb{E}(N)\mathbb{E}(X) = 0$$

but it turns out Wald's equation doesn't apply because  $\mathbb{E}(N) = \infty$  (the mean number of plays to get ahead is infinite).

**Example 7.14.** Let

$$X_i = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

with  $p > 1/2$ . Note that  $\mathbb{E}(X_i) = 2p - 1 > 0$ . One stopping time could be  $N = \min\{n : X_1 + \dots + X_n = 1\}$ . It turns out from Markov chain theory that this will eventually happen with probability 1. We can't apply Wald's Equation (Theorem 7.3.8) because  $\mathbb{E}(N)$  is not finite. Note that if we try, a contradiction results:

$$1 = \mathbb{E}(N) \cdot 0 = 0$$

**Example 7.15.** Let  $U_i \sim \text{Uniform}(0, 1)$ ,  $U_i$  i.i.d. One stopping time is  $N = \min\{n : U_n > 0.8\}$ . Then by Wald's Equation (Theorem 7.3.8),

$$\mathbb{E}\left(\sum_{i=1}^N X_i\right) = \mathbb{E}(N)\mathbb{E}(X) = 10 \cdot 0 = 0$$

so the expected winnings when you stop is 0 (regardless of your stopping time, since it's a fair game).

Back to renewal theory: Let  $X_1, X_2, \dots$  be a sequence of random variables with interarrival times distributed as  $F$ . Note that  $N(t) = 5 \iff X_1 + \dots + X_5 \leq t, X_1 + \dots + X_6 > t$  so this is not a stopping time. But we could stop at the first event that occurs after time  $t$ , so  $N(t) + 1$  is a stopping time. Note that

$$N(t) + 1 = n \iff N(t) = n - 1, X_1 + \dots + X_{n-1} \leq t, X_1 + \dots + X_n > t.$$

(You can't say "I'll stop at the last event before  $t$ " because at the time of that event you wouldn't have been able to know it was the last event before  $t$ .)

**Corollary 7.3.8.1 (Corollary to Wald's Equation).**

$$\mathbb{E}\left(\sum_{i=1}^{N(t)+1} X_i\right) = \mu(m(t) + 1)$$

**Theorem 7.3.9 (Elementary Renewal Theorem).**

$$\frac{m(t)}{t} \xrightarrow{\text{a.s.}} \frac{1}{\mu}.$$

*Proof.* Note that

$$S_{N(t)+1} > t$$

We have

$$\mathbb{E}(S_{N(t)+1}) > t$$

Using Corollary 7.3.8.1,

$$\mu(m(t) + 1) > t \iff m(t) + 1 > \frac{t}{\mu} \iff \frac{m(t)}{t} > \frac{1}{\mu} - \frac{1}{t}$$

Then as  $t \rightarrow \infty$

$$\liminf \frac{m(t)}{t} \geq \frac{1}{\mu}.$$

So if we can show  $\liminf \frac{m(t)}{t} \leq \frac{1}{\mu}$ , we're done. Assume  $\Pr(X_i \leq M) = 1$  for some  $M \in \mathbb{R}$ . Then

$$\begin{aligned} S_{N(t)+1} < t + M &\implies \mu(m(t) + 1) < t + M \implies m(t) < \frac{t}{\mu} + \frac{M}{\mu} - 1 \\ &\implies \frac{m(t)}{t} < \frac{1}{\mu} + \frac{M}{t\mu} - \frac{1}{t} \\ &\implies \limsup \frac{m(t)}{t} \leq \frac{1}{\mu}. \end{aligned}$$

Now consider the general case. Consider  $X_1, X_2, \dots$ . Let

$$X_i^* = \begin{cases} X_i & \text{if } X_i \leq M \\ M & \text{if } X_i > M \end{cases}$$

and let  $N^*(t) = \max\{n : X_1^* + \dots + X_n^* \leq t\}$ . Then since  $N^*(t)$  has smaller interarrival times than  $N(t)$ ,

$$\begin{aligned} N(T) \leq N^*(t) &\implies \frac{\mathbb{E}(N(t))}{t} \leq \frac{\mathbb{E}(N^*(t))}{t} \\ &\implies \lim \frac{m(t)}{t} \leq \lim \frac{\mathbb{E}(N^*(t))}{t} = \frac{1}{\mathbb{E}(X_i^*)} \end{aligned}$$

Note that (since  $X_I^* = X_I$  if  $X_i \leq M$ )

$$\mathbb{E}(X_i^*) = \int_0^M x dF(x) + M \bar{F}(M) \rightarrow \mu + 0 = \mu \text{ as } M \rightarrow \infty$$

so

$$\lim \frac{m(t)}{t} \leq \frac{1}{\mathbb{E}(\min\{X, M\})}$$

which is true for every  $M$ . Letting  $M \rightarrow \infty$ , we have

$$\mathbb{E}(\min\{X, M\}) \rightarrow \mathbb{E}(X) \text{ as } M \rightarrow \infty$$

by Lebesgue's Monotone convergence theorem. ("if you have a sequence of variables  $X_n \leq X_{n+1} \leq \dots$  then  $\mathbb{E}(\lim_{n \rightarrow \infty} X_n) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n)$ ".)

□

**Remark.** This technique is called “coupling;” relating something you don’t know to something you do.

**Remark.** Why doesn’t the Elementary Renewal Theorem follow from the Strong Law for Renewal Processes (Theorem 7.3.6)? Consider this counterexample: Let  $U \sim \text{Uniform}(0, 1)$  and let  $X_n$  be a random variable such that

$$X_n = \begin{cases} n & U < 1/n \\ 0 & U > 1/n \end{cases}$$

Of course with probability 1,  $U > 0$ ; that is,  $U = \epsilon > 0$ . Note that  $X_n = 0$  for all  $n$  sufficiently large ( $1/n < \epsilon$ ). So we see that  $X_n \xrightarrow{a.s.} 0$ . But for all  $n$  sufficiently large

$$\mathbb{E}(X_n) = n \cdot \frac{1}{n} = 1.$$

So we have

$$\lim_{n \rightarrow \infty} X_n = 0 = \mathbb{E}(\lim_{n \rightarrow \infty} X_n) \neq \lim_{n \rightarrow \infty} \mathbb{E}(X_n) = 1$$

In summary, just because  $\frac{N(t)}{t} \xrightarrow{a.s.} \frac{1}{\mu}$  doesn’t mean  $\frac{m(t)}{t} \xrightarrow{a.s.} \frac{1}{\mu}$ .

**Proposition 7.3.10.** Consider a delayed renewal process  $N_d(t)$  and let  $m_d(t) = \mathbb{E}(N_d(t))$ . Recall that the first arrival has distribution  $G$  and the rest have distribution  $F$ . Then  $m_d(t) \xrightarrow{a.s.} 1/\mu$ .

*Proof.*

$$m_d(t) = \int_0^t \mathbb{E}(N_d(t) | X_1 = s) dG(s)$$

Note that

$$\mathbb{E}(N_d(t) | X_1 = s) = 1 + m(t - s)$$

so we have

$$\begin{aligned} m_d(t) &= \int_0^t (1 + m(t - s)) dG(s) = G(t) + \int_0^t m(t - s) dG(s) \\ &\iff \frac{m_d(t)}{t} = \frac{G(t)}{t} + \frac{1}{t} \int_0^t m(t - s) dG(s) \xrightarrow{a.s.} \frac{1}{\mu} \end{aligned}$$

by using  $m(y)/y \xrightarrow{a.s.} 1/\mu$ . (Details can be fleshed out.)

□

**Proposition 7.3.11.** Suppose  $X_1, X_2, \dots$  are integer valued, and there is only one renewal at a time. Suppose  $\Pr(X_i = 0) = 0$ . Then

*Proof.* Note that

$$\frac{\mathbb{E}(N_d(t))}{n} \rightarrow \frac{1}{\mu}$$

by the Strong Law for Renewal Processes (Theorem 7.3.6). Note that

$$N_d(n) = \sum_{j=1}^n I_j$$

where  $I_j$  is an indicator variable for whether there is a renewal at time  $j$ . Then

$$\mathbb{E}(N_d(n)) = \sum_{j=1}^n \mathbb{E}(I_j) = \sum_{j=1}^n \Pr(I_j = 1)$$

Then by the Elementary Renewal Theorem (Theorem 7.3.9),

$$\frac{1}{n} \sum_{j=1}^n \Pr(I_j = 1) \xrightarrow{a.s.} \frac{1}{\mathbb{E}(X_i)}$$

□

Recall we say that  $a_n \rightarrow a$  pointwise if  $\lim_{n \rightarrow \infty} a_n = a$ .

**Definition 7.18 (Caesaro Convergence).** We say  $a_n$  Caesaro converges to  $a$  if

$$\frac{a_1 + \dots + a_n}{n} \rightarrow a \text{ as } n \rightarrow \infty.$$

**Remark.** Pointwise convergence implies Caesaro convergence, but not the other way around.

When does  $\Pr(\{\text{renewal at } n\}) \rightarrow 1/\mu$ ? Let  $\Pr(\{\text{renewal at } j\}) = P_j$  and suppose  $P_n \rightarrow P^*$ . This implies Caesaro convergence

$$\frac{P_1 + \dots + P_n}{n} \rightarrow p^*$$

and  $p^* = 1/\mu$  where  $\mu$  is the expected time between renewals.

These examples are very confusing—look for explanations in textbook? Seems to be same as Section 7.9.2 in *Introduction to Probability Models*.

**Example 7.16.** Let  $Y_1, Y_2, \dots$  i.i.d. Suppose  $p_j = \Pr(Y_i = j)$ . We will keep doing draws until we observe the pattern 1, 2, 1, 3, 4, 1, 2, 1. Let  $N$  be the number of trials until this occurs. What is  $\mathbb{E}(N)$ ?

**Solution.** Suppose we consider the process as continuing to happen forever even after a pattern appears, and treat a pattern arriving as an event. let  $N(n)$  be the number of events by  $n$ . Is  $\{N(n)\}$  a renewal process? That is, once an event occurs does everything start all over? Yes, so it's either an ordinary or delayed renewal process. Note that it's delayed because the last three digits of the pattern match the first three, so once a pattern has just happened, another one could happen in 5 draws. Note that

$$\Pr(\text{pattern appears at time } n) = P_1 P_2 \dots = P_1^4 P_2^2 P_3 P_4$$

so the expected time between renewals is  $1/(P_1^4 P_2^2 P_3 P_4)$ , except that we need to take into account the fact that we already have the first three digits of the pattern if we just got one. Note that

$$\Pr(\text{pattern appears at time } n \mid \text{just had pattern}) = P_1^2 P_2 P_3 P_4$$

We can look at this a different way. The time until we get a pattern  $T_p$  is the time until we get 1,2,1  $T_{121}$  plus the time until we get the rest of the pattern  $T_{rest}$ .

$$T_p = T_{121} + T_{rest} \implies \mathbb{E}(T_p) = \mathbb{E}(T_{121}) + 1/(P_1^4 P_2^2 P_3 P_4)$$

Then  $\Pr(\text{renewal at } n) = P_1^2 P_2$  so the expected time between renewals is  $1/(P_1^2 P_2)$ . Then the time between renewals is the additional time between renewals given that you currently have 1, so

$$\mathbb{E}(\text{time between renewals}) = \mathbb{E}(\text{additional time when you have a 1}) = 1/(P_1^2 P_2)$$

Do it again: time to get to 1,2,1 is time to get to 1 plus additional:

$$T_{121} = T_1 + T_{21} \implies \mathbb{E}(T_{121}) = \mathbb{E}(T_1) + 1/(P_1^2 P_2) = 1/P_1 + 1/(P_1^2 P_2)$$

putting this all together, we have

$$\mathbb{E}(T_p) = 1/P_1 + 1/(P_1^2 P_2) + 1/(P_1^4 P_2^2 P_3 P_4).$$

**Example 7.17.** Flips coins, get heads with probability  $P$ . Then  $T_k$  is the time to land  $k$  heads in a row. What is the expected time until  $k$  heads in a row?

**Solution.** Note that the probability of a renewal at  $n$  is  $p^k$ . So the expected time between renewals is equal to  $1/p^k$  not taking into account that the next time we could use the heads we already have. Note that

$$T_k = T_{k-1} + T_{rest}$$

so

$$\mathbb{E}(T_k) = \mathbb{E}(T_{k=1} + 1/p^k = 1/p^k + \mathbb{E}(T_{k-2}) + 1/(p^{k-1}) + \dots$$

## 7.4 ISE 620 Midterm Solutions

**Exercise 7.**

$$\text{Var}\left(\sum_{i=1}^n X_i\right)$$

**Solution.**

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \text{Var}\left(\sum_{i=1}^n (n+1-i)X_i\right) = \sum_{i=1}^n \frac{(n+1-i)^2}{\lambda^2} = \sum_{j=1}^n j^2/\lambda^2$$

or

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(S_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(S_i, S_j)$$

Note that

$$\text{Cov}(S_i, S_j) = \text{Cov}\left(S_i, S_i + \sum_{k=i+1}^j X_k\right) = \text{Var}(S_i) + 0 = i/\lambda^2$$

**Exercise 8.**

$$\text{Var}\left(\sum_{i=1}^{N(t)} S_i \mid N(t)\right)$$

**Solution.**

$$\text{Var}\left(\sum_{i=1}^{N(t)} S_i \mid N(t) = n\right) = \text{Var}\left(\sum_{i=1}^n U_i\right)$$

where  $U_1, \dots, U_n \sim$  i.i.d.  $U(0, t)$  and they are ordered. Since the sum of the ordered values is the same as the sum of the unordered values and they are i.i.d., we have

$$\text{Var}\left(\sum_{i=1}^n U_i\right) = \sum_{i=1}^n \text{Var}(U_i) = \frac{nt^2}{12}.$$

**Exercise 9.** A wins point with probability  $p_1$  if A is server, A wins with probability  $p_2$  if B serves. (a) serve next if you just won. (b) alternate serve.

**Solution.**

- (a) A serving is a renewal. Suppose  $A$  just won a point. Let  $X$  be the number of games until  $A$  wins again. Frequency with which  $A$  wins converges to  $1/\mathbb{E}(X)$  by the Strong Law for Renewal Processes (Theorem 7.3.6). Condition on who wins next game. Let  $Y$  be an indicator variable for  $A$  winning the next game. Then

$$\mathbb{E}(X) = \mathbb{E}(X | Y = 1)p_1 + \mathbb{E}(X | Y = 0)(1 - p_1)$$

Note that if  $Y = 1$ ,  $X = 1$ . For  $\mathbb{E}(X | Y = 0)$ ,  $B$  keeps serving until  $A$  wins, so the number of games until  $A$  wins is a geometric random variable with probability  $p_2$ , so the mean is  $1/p_2$ . So we have

$$\mathbb{E}(X) = 1 \cdot p_1 + (1 + 1/p_2)(1 - p_1) = 1 + \frac{1 - p_1}{p_2} = \frac{p_2 + 1 - p_1}{p_2}$$

and the rate is

$$\frac{p_2}{p_2 + 1 - p_1}$$

- (b) Not a renewal—everything doesn't start all over again when  $A$  wins a point, because just because  $A$  just won doesn't tell us anything about who wins. A renewal would happen every time  $A$  wins on her serve. Let  $X$  be the number of games until the next renewal. Then the answer we seek is  $1/\mathbb{E}(X)$ . Let  $Y$  be an indicator variable for  $A$  winning on her next serve. Note that

$$\mathbb{E}(X) = \mathbb{E}(X | Y = 1)p_1 + \mathbb{E}(X | Y = 0)(1 - p_1)$$

Note that  $\mathbb{E}(X | Y = 1) = 2$ ,  $\mathbb{E}(X | Y = 0) = 2 + \mathbb{E}(X)$ . So we have

$$\mathbb{E}(X) = 2p_1 + (2 + \mathbb{E}(X))(1 - p_1) \iff p_1\mathbb{E}(X) = 2p_1 + 2 - 2p_1 \iff \mathbb{E}(X) = 2/p_1$$

Do the same thing again for when  $A$  wins on  $B$ 's serves. Then add them together.

A renewal would happen every time  $A$  wins on  $B$ 's serve. Let  $Z$  be the number of games until the next renewal. Then the answer we seek is  $1/\mathbb{E}(Z)$ . Let  $Y$  be an indicator variable for  $A$  winning on  $B$ 's next serve. Note that

$$\mathbb{E}(Z) = \mathbb{E}(Z | Y = 1)p_2 + \mathbb{E}(Z | Y = 0)(1 - p_2)$$

Note that  $\mathbb{E}(Z | Y = 1) = 2$ ,  $\mathbb{E}(Z | Y = 0) = 2 + \mathbb{E}(Z)$ . So we have

$$\mathbb{E}(Z) = 2p_2 + (2 + \mathbb{E}(Z))(1 - p_2) \iff p_2\mathbb{E}(Z) = 2p_2 + 2 - 2p_2 \iff \mathbb{E}(Z) = 2/p_2$$

So the long-run proportion of games won by  $A$  is

$$\frac{1}{\mathbb{E}(X)} + \frac{1}{\mathbb{E}(Z)} = \frac{p_1 + p_2}{2}.$$

**Exercise 10.**  $M/G/\infty$ ,  $\Pr(X(t) = j | N(t) = n)$ .

### Solution.

The arrival times are i.i.d. uniform on  $(0, t)$ . So each one has the same probability of being in the system at time  $t$ , and the number who are in the system is a binomial random variable.

$$\Pr(X(t) = j \mid N(t) = n) = \binom{n}{j} p^j (1-p)^{n-j}$$

where  $p$  is the probability of being in the system at time  $t$ :

$$p = \int_0^t \Pr(\text{in system} \mid \text{arrived at } s) \cdot (1/t) ds = \frac{1}{t} \int_0^t \bar{G}(t-s) ds.$$

**Exercise 11.** number of claims  $N$  poisson with mean lambda, claim amounts ‘are uniformly distributed on  $(0, 1)$ , we want  $\mathbb{E}(\max\{X_1, \dots, X_N\})$ .

**Solution.** Note that

$$\Pr(\max\{X_1, \dots, X_n\} \leq t) = t^n \implies \Pr(\max\{X_1, \dots, X_n\} > t) = 1 - t^n$$

$$\implies \mathbb{E}(\max\{X_1, \dots, X_n\}) = \int_0^1 (1 - t^n) dt = \frac{n}{n+1} = 1 - \frac{1}{n+1}$$

so this is the expected value if we condition on  $n$ . Therefore

$$\mathbb{E}(\max\{X_1, \dots, X_N\}) = \mathbb{E}(\max\{X_1, \dots, X_N\} \mid N = n) \Pr(N = n)$$

$$= \sum_{n=0}^{\infty} \left(1 - \frac{1}{n+1}\right) \cdot e^{-\lambda} \frac{\lambda^n}{n!} = \dots$$

**Exercise 12.**

$$\mathbb{E}(X(t) \mid N(s) = n)$$

where

$$X(t) = \sum_{i=1}^{N(t)} X_i$$

**Solution.**

We have

$$\begin{aligned} X(t) \mid N(s) = n &= \sum_{i=1}^{N(s)} X_i + \sum_{i=N(s)+1}^{N(t)} X_i = \sum_{i=1}^n X_i + \sum_{i=N(s)+1}^{N(t)} X_i \\ \implies \mathbb{E}(X(t) \mid N(s)) &= \mathbb{E}\left(\sum_{i=1}^n X_i\right) + \mathbb{E}\left(\sum_{i=N(s)+1}^{N(t)} X_i\right) = n\mathbb{E}(X) + \mathbb{E}(X)\lambda(t-s) \end{aligned}$$

where  $\lambda(t-s)$  is the expected number of events between  $s$  and  $t$ .

## 7.5 Renewal Reward Processes (Section 4.7 in *Introduction to Probability Models*, 3.6 *Stochastic Processes*)

**Definition 7.19 (Renewal reward process).** Suppose we have a renewal process with interarrival times  $X_1, X_2, \dots$ . Suppose we receive a reward  $R_i$  when renewal  $i$  occurs (after time  $X_i$  from the previous renewal). The reward is allowed to depend on what happens during the previous period  $X_i$ , but every time renewal happens, process starts again. So we assume the vectors  $(X_i, R_i), i \geq 1$  are i.i.d. Let  $R(t)$  be the total reward by time  $t$ ; that is,

$$R(t) = \sum_{i=1}^{N(t)} R_i.$$

Then  $\{R(t), t \geq 0\}$  is a **renewal reward process**.

**Proposition 7.5.1 (Proposition 7.3 in *Introduction to Probability Models*; sometimes called Renewal Reward Theorem).**

(a)

$$\frac{R(t)}{t} \xrightarrow{\text{a.s.}} \frac{\mathbb{E}(R)}{\mathbb{E}(X)}$$

(generalization of strong law)

(b)

$$\frac{\mathbb{E}(R(t))}{t} \xrightarrow{\text{a.s.}} \frac{\mathbb{E}(R)}{\mathbb{E}(X)}$$

*Proof.* (a)

$$\frac{R(t)}{t} = \frac{1}{t} \sum_{i=1}^{N(t)} R_i = \frac{1}{N(t)} \sum_{i=1}^{N(t)} R_i \cdot \frac{N(t)}{t}$$

Since  $N(t) \rightarrow \infty$  as  $t \rightarrow \infty$ , by Strong Law of Large Numbers (Theorem 8.6.4)

$$\frac{1}{N(t)} \sum_{i=1}^{N(t)} R_i \xrightarrow{\text{a.s.}} \mathbb{E}(R)$$

By Strong Law for Renewal Processes (Theorem 7.3.6),

$$\frac{N(t)}{t} \xrightarrow{\text{a.s.}} 1/\mathbb{E}(X).$$

Then by the Continuous Mapping Theorem (Theorem 8.4.13), the result follows.

(b)

□

**Example 7.18 (Same as exam question).** A wins point with probability  $p_1$  if A is server, A wins with probability  $p_2$  if B serves. alternate serve.

**Solution.** Every time  $A$  serves is a renewal. 1 reward each time  $A$  wins a point. Expected earning per cycle is  $p_1 + p_2$ . Cycles are length 2. Works out to  $(p_1 + p_2)/2$ .

**Proposition 7.5.2.** We've assumed that the renewal is earned at the end of the cycle. But the result holds even if the reward is earned during the interval gradually, not all at once at the end of the renewal interval (assuming all rewards are nonnegative). So

$$\frac{R(t)}{t} \xrightarrow{a.s.} \frac{\mathbb{E}(R)}{\mathbb{E}(X)}.$$

Also,

$$\mathbb{E}\left[\frac{R(t)}{t}\right] \xrightarrow{a.s.} \frac{\mathbb{E}(R)}{\mathbb{E}(X)}.$$

*Proof.*  $R(t)$  is the total reward you get by time  $t$ . So  $\sum_{i=1}^{N(t)} R_i \leq R(t)$  (the inequality holds if you've earned a bit since the last renewal). We have

$$\sum_{i=1}^{N(t)} R_i \leq R(t) \leq \sum_{i=1}^{N(t)+1} R_i$$

Dividing through by  $t$ :

$$\frac{1}{t} \sum_{i=1}^{N(t)} R_i \leq \frac{R(t)}{t} \leq \frac{1}{N(t)+1} \sum_{i=1}^{N(t)+1} R_i \cdot \frac{N(t)+1}{t}$$

But by Proposition 7.5.1

$$\frac{1}{t} \sum_{i=1}^{N(t)} R_i \xrightarrow{a.s.} \mathbb{E}(R)/\mathbb{E}(X),$$

$$\frac{1}{N(t)+1} \sum_{i=1}^{N(t)+1} R_i \xrightarrow{a.s.} \mathbb{E}(R)$$

$$\frac{N(t)+1}{t} \xrightarrow{a.s.} 1/\mathbb{E}(X)$$

So by the Continuous Mapping Theorem,

$$\frac{R(t)}{t} \xrightarrow{a.s.} \frac{\mathbb{E}(R)}{\mathbb{E}(X)}.$$

□

**Example 7.19 (Similar to Example 7.15 in *Introduction to Probability Models*).** People arrive to train station in a Poisson process with rate  $\lambda$ .  $K$  is the cost to dispatch the train. Suppose we incur a cost  $c$  per unit of time waiting per customer. What dispatching policy minimizes the long-run average cost per unit time?

**Solution.**

The structure of the optimal policy is to wait until a certain number of people arrive and then dispatch a train ( $n$  policy). So we want to determine this number. (Another policy is a  $t$  policy: dispatch a train every  $t$  units of time. But it turns out the best  $n$  policy is better than the best  $t$  policy. We may be interested in the difference in minimal costs because  $t$  policies are easier to employ and more convenient for customers.)

- (a)  **$n$ -policy:** Note that every time we dispatch a train, there is a renewal. At every renewal time there is a cost incurred. The long run average cost per unit time is the expected cost during a cycle divided by the expected time of a cycle. Note that the expected time between cycles is the expected amount of time for  $n$  people to arrive, which is simply  $n/\lambda$ . The cost of a cycle is

$$cX_2 + 2cX_3 + 3cX_4 + \dots + (n-1)cX_n + K$$

so the expected cost is

$$\mathbb{E}(X)(c + 2c + 3c + \dots + (n-1)c) + K = \frac{c}{\lambda} \frac{(n-1)n}{2} + K$$

Therefore taking the ratio, the average cost per cycle is

$$\left( \frac{1}{\lambda} \frac{c(n-1)n}{2} + K \right) / \left( \frac{n}{\lambda} \right) = \frac{c(n-1)}{2} + \frac{\lambda K}{n}.$$

Now we want to pick  $n$  to minimize this quantity. Treat  $n$  as continuous and take the derivative.

$$\frac{c}{2} - \frac{\lambda K}{n^2} = 0 \iff n = \sqrt{\frac{2\lambda K}{c}}$$

which makes the minimum average cost

$$\frac{1}{2} \left( c \sqrt{\frac{2\lambda K}{c}} - c \right) + K \sqrt{\frac{c}{2\lambda K}} = \sqrt{2\lambda c K} - \frac{c}{2}.$$

- (b)  **$t$ -policy:** Dispatch a train every  $t$  time units. Note that we have a renewal every time you dispatch a train. The average cost per unit time is the expected cost per cycle  $\mathbb{E}(C)$  divided by  $t$ . Note that

$$\begin{aligned} C &= c \sum_{i=1}^{N(t)} (t - S_i) = c \left( tN(t) - \sum_{i=1}^{N(t)} S_i \right) + K \\ \implies \mathbb{E}(C) &= c \mathbb{E} \left( tN(t) - \sum_{i=1}^{N(t)} S_i \right) + K = c \left[ \lambda t^2 - \mathbb{E} \left( \sum_{i=1}^{N(t)} S_i \right) \right] + K \end{aligned}$$

Note that

$$\mathbb{E}\left(\sum_{i=1}^{N(t)} S_i \mid N(t) = n\right) = \mathbb{E}\left(\sum_{i=1}^n S_i\right) = \mathbb{E}\left(\sum_{i=1}^n U_i\right) = \frac{nt}{2} \implies \mathbb{E}\left(\sum_{i=1}^{N(t)} S_i \mid N(t)\right) = \frac{tN(t)}{2}$$

where  $U_i \sim \text{U}(0, t)$ .

$$\begin{aligned} \implies \mathbb{E}(C) &= c\lambda t^2 - c\mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^{N(t)} S_i \mid N(t)\right)\right] + K = c\lambda t^2 - c\mathbb{E}\left[\frac{tN(t)}{2}\right] + K = c\lambda t^2 - c\frac{\lambda t^2}{2} + K \\ \implies \mathbb{E}(C) &= \frac{c\lambda t^2}{2} + K \end{aligned}$$

Therefore the average cost per cycle is

$$\frac{1}{t} \cdot \left(\frac{c\lambda t^2}{2} + K\right) = \frac{\lambda ct}{2} + \frac{K}{t}$$

We seek the best  $t$  to minimize this cost. Take the derivative with respect to  $t$ .

$$\frac{\lambda c}{2} - \frac{K}{t^2} = 0 \implies t = \sqrt{\frac{2K}{\lambda c}}$$

So the minimum average cost is

$$\sqrt{\frac{2K}{\lambda c}} \frac{\lambda c}{2} + \frac{K}{t} \sqrt{\frac{\lambda c}{2K}} = K \sqrt{\frac{\lambda c}{2K}}.$$

**Example 7.20.** It costs  $C$  to buy a new car. The car lasts for a time  $L$  having distribution  $F$  before it fails. If the car fails, you incur cost  $K$ . Suppose you have a policy of buying a new car when the old car either fails or reaches a certain age  $t$ .

**Solution.** Note that a renewal occurs when you get a new car or when it fails. The cost per cycle is  $C$  if  $L > t$ ,  $C + K$  if  $L \leq t$ . The time for a cycle is  $\min\{L, t\}$ . Therefore the average cost per unit time is the expected cost of a cycle  $C + K \Pr(L < t) = C + KF(t)$ . The average lifetime is

$$\int_0^t x dF(x) + \int_t^\infty t dF(x) = \int_0^t x dF(x) + t \bar{F}(t).$$

Therefore the average cost per cycle is

$$\left(c + KF(t)\right) \Big/ \left(\int_0^t x dF(x) + t \bar{F}(t)\right).$$

**Definition 7.20 (Age of a renewal process).** We call  $A(t) = t - S_{N(t)}$  the **age of the renewal process at time  $t$** .

**Definition 7.21 (Excess (or residual) lifetime at time  $t$ ).** We call  $Y(t) = S_{N(t)+1} - t$  be the **excess (or residual) lifetime of a renewal process at time  $t$** .

**Proposition 7.5.3.** Let  $X$  be the interarrival time for a renewal reward process. Then

(a) with probability 1,

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t A(s)ds = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

(b) With probability 1,

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t Y(s)ds = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

(c) With probability 1,

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbb{E}[A(s)]ds = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

(d) With probability 1,

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbb{E}[Y(s)]ds = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

*Proof.* (a) Imagine we earn a reward at a rate equal to the age of the renewal process. Then the amount earned by time  $t$  is

$$\int_0^t A(s)ds.$$

When there is a new renewal, the process starts over again, so this is a renewal reward process. The reward earned during each cycle is

$$\int_0^X tdt = \frac{X^2}{2}.$$

Taking expectations and taking the ratio of these in the limit, we have

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t A(s)ds = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

For more detail, see Example 7.18 in *Introduction to Probability Models*.

(b) Imagine we earn a reward at a rate equal to the residual lifetime of the renewal process. Then the amount earned by time  $t$  is

$$\int_0^t Y(s)ds.$$

When there is a new renewal, the process starts over again, so this is a renewal reward process. The reward earned during each cycle is

$$\int_0^X (X - t)dt = X^2 - \frac{X^2}{2} = \frac{X^2}{2}.$$

Taking expectations and taking the ratio of these in the limit, we have

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t A(s) ds = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

For more detail, see Example 7.19 in *Introduction to Probability Models*.

- (c) Similar to (a).
- (d) Similar to (c).

□

**Remark.** This makes sense because if the process is infinitely long, if you look at it backwards in time the distribution of renewals is the same as if you look forward in time. But when you switch direction, the residual lifetimes and ages reverse meaning. Therefore since the distribution doesn't change depending on the direction you look at it from, the long-term expected residual lifetime and age ought to be equal.

**Proposition 7.5.4 (Homework 7, *Introduction to Probability Models* Ch. 7 Problem 30).**

$$\frac{A(t)}{t} \xrightarrow{a.s.} 0.$$

*Proof.*  $A(t)$  is the time since the last renewal; that is,  $A(t) = t - S_{N(t)}$ . So we hope to show that with probability one,

$$\frac{t - S_{N(t)}}{t} \rightarrow 0 \text{ as } t \rightarrow \infty$$

(or, equivalently)

$$\frac{t - S_{N(t)}}{t} \xrightarrow{w.p.1} 0.$$

Note that

$$\frac{t - S_{N(t)}}{t} = \frac{t - S_{N(t)}}{N(t)} \cdot \frac{N(t)}{t} = \left( \frac{t}{N(t)} - \frac{S_{N(t)}}{N(t)} \right) \frac{N(t)}{t}$$

By the Strong Law for Renewal Processes (Theorem 7.3.6),  $\frac{N(t)}{t} \xrightarrow{w.p.1} \mu^{-1}$ , where  $\mu = \mathbb{E}(X)$  is the expected interarrival time. Since the expected interarrival time is finite,  $N(t) \rightarrow \infty$  as  $t \rightarrow \infty$ . Therefore

$$\frac{S_{N(t)}}{N(t)} = \frac{X_1 + \dots + X_{N(t)}}{N(t)} \xrightarrow{w.p.1} \mu.$$

Lastly, since  $N(t) > 0$  with probability 1 as  $t \rightarrow \infty$  and  $t > 0$ , by the Continuous Mapping Theorem

$$\frac{t}{N(t)} = \left( \frac{N(t)}{t} \right)^{-1} \xrightarrow{w.p.1} (\mu^{-1})^{-1} = \mu.$$

Therefore by another application of the Continuous Mapping Theorem,

$$\frac{t - S_{N(t)}}{t} = \left( \frac{t}{N(t)} - \frac{S_{N(t)}}{N(t)} \right) \frac{N(t)}{t} \xrightarrow{w.p.1} (\mu - \mu) \cdot \mu^{-1} = 0.$$

□

**Example 7.21 (Similar to Example 7.20 in *Introduction to Probability Models*).** People arrive to a bus stop in a Poisson process with rate  $\lambda$ . Buses arrive in an independent renewal process after time  $T$  with distribution  $F$  where  $\mathbb{E}(T) = \mu$ .

- (a) What is the long-run average number of people picked up by a bus when it arrives?
- (b) Suppose busses arrive according to a Poisson process with rate  $\lambda$ ; that is,  $F(t) = 1 - e^{-\alpha t}$ . What is the long-run average number of people waiting at any given time (averaged over all time)?
- (c) What is the average amount of time a person waits for a bus (averaged over all people)? (Let  $W_i$  be the waiting time of person  $i$ . We seek  $\lim_{n \rightarrow \infty} \sum_{i=1}^n W_i / n$ .)

**Solution.**

- (a) Every time a bus arrives, we have a renewal. The number of people picked up by a bus is  $N(T)$ . Note that  $\{N(t)\} \sim PP(\lambda)$ . Note that

$$\begin{aligned} \mathbb{E}(N(T) | T = t) &= \mathbb{E}(N(t) | T = t) = (\text{by independence}) \mathbb{E}(N(t)) = \lambda t \implies \mathbb{E}(N(T) | T) = \lambda T \\ &\implies \mathbb{E}(N(T)) = \mathbb{E}(\mathbb{E}(N(T) | T)) = \lambda \mathbb{E}(T) = \lambda \mu. \end{aligned}$$

- (b) Let  $N_s(s)$  be the number of people waiting at time  $s$ . We seek

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t N_w(s) ds.$$

Imagine that at time  $s$  we earn at a rate  $N_w(s)$ . Then we want the average reward per unit time. Then this is a renewal reward process, where renewals happen when a bus arrives. The expected reward earned per cycle is

$$\frac{1}{\mathbb{E}(T)} \cdot \mathbb{E}\left(\int_0^T N(s) ds\right) \tag{7.8}$$

Note that

$$\begin{aligned} \mathbb{E}\left(\int_0^T N(s) ds | T = t\right) &= \mathbb{E}\left(\int_0^t N(s) ds\right) = (\text{by Fubini's Theorem}) \int_0^t \mathbb{E}(N(s)) ds = \int_0^t \lambda s ds = \frac{\lambda t^2}{2} \\ &\implies \mathbb{E}\left(\int_0^T N(s) ds | T\right) = \frac{\lambda T^2}{2} \end{aligned}$$

Therefore we have that the expected reward earned per cycle is (plugging into (7.8))

$$\frac{\lambda \mathbb{E}(T^2)}{2\mathbb{E}(T)}.$$

Because arrivals occur in a Poisson process ( $F(t) = 1 - e^{-\alpha t}$ ), we have  $\mathbb{E}(T) = \alpha^{-1}$ ,  $\mathbb{E}(T^2) = 2\alpha^{-2}$ . Therefore the expected reward earned per cycle is

$$\frac{\lambda 2\alpha^{-2}}{2\alpha^{-1}} = \frac{\lambda}{\alpha}.$$

(c)  $\frac{\lambda}{\alpha}$ .

**Remark.** Note that the average number of people picked up by a bus now matches the average number of people waiting at any given time. This is counterintuitive because you would think at the time the bus comes, the most people are there, so the average number of people picked up should be larger than the average number of people waiting in general.

This illustrates the **PASTA principle** and the **inspection paradox**. The paradox is resolved by the fact that the average in (b) is averaged over all time, while the average in (c) is averaged over all people.

**Theorem 7.5.5.** Queue, customers arrive in a renewal process with distribution  $F$ , so

$$\lambda^{-1} = \int_0^\infty \bar{F}(t)dt.$$

Each customer eventually leaves. Let  $L$  be the average number of customers in the system, averaged over all time. Let  $X(s)$  be the number of customers in the system at time  $s$ . Note that

$$L = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t X(s)ds.$$

Let  $W$  be the average time a customer spends in the system. Let  $W_i$  be the amount of time customer  $i$  spends in the system. Note that

$$W = \lim_{n \rightarrow \infty} \frac{W_1 + \dots + W_n}{n}.$$

Then  $L = \lambda W$ .

*Proof.*

$$L = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t X(s)ds$$

Suppose we are paid at any time at a rate equal to the number of customers in the system. Then we can interpret  $L$  as the average reward per unit time. We can consider a renewal as occurring upon the arrival of the first customer when the system was previously empty. Then we have a renewal reward process. Let  $T$  be the time between renewals. Assume the first arrival occurs at time  $t = 0$  (so this is not a delayed renewal reward process). Note that

$$L = \frac{1}{\mathbb{E}(T)} \mathbb{E} \left( \int_0^T X(s) ds \right) \iff \mathbb{E} \left( \int_0^T X(s) ds \right) = L \mathbb{E}(T) \quad (7.9)$$

Note that

$$W = \lim_{n \rightarrow \infty} \frac{W_1 + \dots + W_n}{n}$$

Let  $N$  be the number of customers served in a busy period. Then when customer  $N+1$  arrives, the system is empty and there is a renewal, so  $T$  is the time of the  $N+1$ st arrival. So

$$T = \sum_{i=1}^N X_i$$

where  $X_i$  are the interarrival times ( $X_i$  is the time between the  $i$ th and  $i+1$ st arrival). Note that  $N$  depends on the interarrival times; if the interarrival times are small,  $N$  will be larger. So this is not a compound random variable, but we can use Wald's Equation (Theorem 7.3.8) if  $N$  is a stopping time for  $X_1, \dots, X_n$ . Note that  $N = 1$  if and only if  $X_1$  is greater than the service time of the initial customer. And in general,  $N = k$  depends only on the arrival times up to  $X_k$ . So  $N$  is a stopping time and we can use Wald's Equation. Therefore

$$\mathbb{E}(T) = \frac{\mathbb{E}(N)}{\lambda}. \quad (7.10)$$

Suppose each customer pays us 1 dollar per unit time. Then at time  $s$  we are earning at rate  $X(s)$ , so

$$\int_0^T X(s) ds$$

is the total amount you earn by time  $T$ . But we can also calculate this amount by adding up the amount of time each customer spends in the system:

$$\int_0^T X(s) ds = \sum_{i=1}^N W_i.$$

So we have (using (7.9) and (7.10))

$$\begin{aligned} \mathbb{E} \left( \int_0^T X(s) ds \right) &= \mathbb{E} \sum_{i=1}^N W_i \iff L \mathbb{E}(T) = \mathbb{E} \sum_{i=1}^N W_i \iff L = \frac{\mathbb{E}(\sum_{i=1}^N W_i)}{\mathbb{E}(N)/\lambda} \\ &= (\text{by Wald's Equation}) \frac{\lambda \mathbb{E}(N) \mathbb{E}(W_i)}{\mathbb{E}(N)} = \lambda W. \end{aligned}$$

□

**Exercise 13.** Queue, customers arrive in a renewal process with distribution  $F$ , so

$$\lambda^{-1} = \int_0^\infty \bar{F}(t)dt.$$

Each customer eventually leaves. Let  $L$  be the average number of customers in the system, averaged over all time. Let  $X(s)$  be the number of customers in the system at time  $s$ . Note that

$$L = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t X(s)ds.$$

Let  $W$  be the average time a customer spends in the system. Let  $W_i$  be the amount of time customer  $i$  spends in the system. Note that

$$W = \lim_{n \rightarrow \infty} \frac{W_1 + \dots + W_n}{n}.$$

**Solution.**

### 7.5.1 Alternating Renewal Processes (Section 7.5.1 in *Introduction to Probability Models*)

**Definition 7.22 (Alternating renewal process).** Suppose we have a process that is on for time  $Y_1$ , off for time  $Z_1$ , on for time  $Y_2$ , off for time  $Z_2$ , etc if  $(X_i, Y_i)$ ,  $i \geq 1$  are i.i.d. Call  $\{X(t), t \geq 0\}$  an **alternating renewal process**, where  $X(t)$  is an indicator variable for the process being on at time  $t$ .

**Proposition 7.5.6 (Proposition 7.4 in *Introduction to Probability Models*; similar to Theorem 7.5.7 (Theorem 3.4.4 in *Stochastic Processes*)).**

(a) The long-run proportion of time that the system is on is

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t I\{X(s) = 1\}ds = \frac{\mathbb{E}(Y)}{\mathbb{E}(Y) + \mathbb{E}(Z)}.$$

(b)

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \Pr(X(s) = 1)ds = \frac{\mathbb{E}(Y)}{\mathbb{E}(Y) + \mathbb{E}(Z)}.$$

*Proof.* (a) Imagine we earn 1 dollar per unit time when the system is on and nothing when the system is off. Note that we have a renewal when the system turns back on after having been turned off. So this is a renewal reward process.

(b) Something about

$$\frac{R(t)}{t} \rightarrow \frac{\mathbb{E}(R)}{\mathbb{E}(T)}$$

and

$$\frac{\mathbb{E}R(t)}{t} \rightarrow \frac{\mathbb{E}(R)}{\mathbb{E}(T)}$$

□

**Proposition 7.5.7** (Theorem 3.4.4 in *Stochastic Processes*; Similar to Proposition 7.5.6 (Proposition 7.4 in *Introduction to Probability Models*)). If the cycle distribution is not lattice (see Definition 6.10), then

$$\Pr(\text{on at } t) \rightarrow \frac{\mathbb{E}(Y)}{\mathbb{E}(Y) + \mathbb{E}(Z)} \text{ as } t \rightarrow \infty$$

where  $\mathbb{E}(Y)$  is the expected length of time intervals when the system is on and  $\mathbb{E}(Z)$  is the expected length of time intervals when the system is off.

**Example 7.22.** Suppose you have insurance and you pay at a rate  $r_1$  until you have an accident. Then the rate is  $r_2$  until  $s$  units pass without an accident, in which case you go back to paying rate  $r_1$ . Suppose accidents occur according to a Poisson process with rate  $\lambda$ . Then the probability you have to pay at rate  $r_1$  is

$$\frac{\mathbb{E}(\text{on})}{\mathbb{E}(\text{on}) + \mathbb{E}(\text{off})}.$$

where we define the process to be “on” when your pay rate is  $r_1$  and “off” when your pay rate is  $r_2$ . Note that

$$\mathbb{E}(\text{on}) = 1/\lambda$$

$$\mathbb{E}(\text{off}) = \int_0^\infty \mathbb{E}(\text{off} \mid T = x) \lambda e^{-\lambda x} dx$$

We have

$$\mathbb{E}(\text{off} \mid T = x) = \begin{cases} x + \mathbb{E}(\text{off}) & x < s \\ s & x > s \end{cases}$$

Substituting this in yields

$$\mathbb{E}(\text{off}) = \int_0^s (x + \mathbb{E}(\text{off})) \lambda e^{-\lambda x} dx + s \int_s^\infty \lambda e^{-\lambda x} dx = \int_0^s x \lambda e^{-\lambda x} dx + \mathbb{E}(\text{off})(1 - e^{-\lambda s}) + s e^{-\lambda s}.$$

Recall that the age of the renewal process at time  $t$  is  $A(t) - t - S_{N(t)}$  and the excess time at  $t$  is  $Y(t) = S_{N(t)+1} - t$ . We have already shown (in Proposition 7.5.3) that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t A(s)ds = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t Y(s)ds = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

**Example 7.23 (Example 7.28 in *Introduction to Probability Models*).**  $M/G/\infty$  queue, (arrivals are a Poisson Process). Consider the system to be “on” if the system is empty and “off” when the system is not empty (busy period). Note that  $\mathbb{E}(\text{on}) = \lambda^{-1}$ . Let  $\mathbb{E}(B)$  be the expected length of the “off” time (busy period). Find  $\mathbb{E}(B)$ .

**Solution.** A busy time starts when a first customer arrives. The long-run proportion of time the system is on is

$$\frac{\mathbb{E}(\text{on})}{\mathbb{E}(\text{on}) + \mathbb{E}(\text{off})}.$$

By Proposition 7.5.7, since this is nonlattice, this equals the limiting probability that the system is on at time  $t$ . That is,

$$\lim_{t \rightarrow \infty} \Pr(\text{on at } t) = \frac{\mathbb{E}(\text{on})}{\mathbb{E}(\text{on}) + \mathbb{E}(\text{off})}.$$

Let  $X(t)$  be the number remaining in the system at time  $t$ . By Proposition 7.2.10,

$$X(T) \sim \text{Poisson} \left( \lambda \int_0^t \bar{G}(s)ds \right).$$

So

$$\lim_{t \rightarrow \infty} \Pr(\text{on at } t) = \Pr(X(t) = 0) = \exp \left( -\lambda \int_0^t \bar{G}(s)ds \right)$$

Plugging in we have **might have made a mistake here:**  $S \sim G$ , then  $\lim_{t \rightarrow \infty} \Pr(\text{on at } t) = \mathbb{E}(S)$

$$\exp \left( -\lambda \int_0^t \bar{G}(s)ds \right) = \frac{\lambda^{-1}}{\lambda^{-1} + \mathbb{E}(\text{off})}.$$

**think this bottom one below might be the right one**

$$\exp \left( -\mathbb{E}(S) \right) = \frac{\lambda^{-1}}{\lambda^{-1} + \mathbb{E}(\text{off})}.$$

**Definition 7.23 (Equilibrium distribution of a renewal process).** We call

$$F_e(x) = \frac{1}{\mu} \int_0^x \bar{F}(t) dt$$

the **equilibrium distribution** of a renewal process with interarrival distribution  $F$ , where  $\mu$  is the expected length of a cycle in the renewal process.

**Proposition 7.5.8 (Age of equilibrium distribution; restatement of Theorem 7.5.7 (Theorem 3.4.4 in *Stochastic Processes*); similar to Examples 7.26 and 7.27 in *Introduction to Probability Models*; Similar to Proposition 7.5.6 (Proposition 7.4 in *Introduction to Probability Models*)).** With probability 1, the long-run proportion of time that the age is less than  $x$  is

$$\lim_{s \rightarrow \infty} \frac{1}{s} \int_0^s I\{A(t) < x\} dt = F_e(x)$$

and the long-run proportion of the time that the excess is less than  $x$

$$\lim_{s \rightarrow \infty} \frac{1}{s} \int_0^s I\{Y(t) < x\} dt = F_e(x)$$

both equal  $F_e(x) = \frac{1}{\mu} \int_0^x \bar{F}(t) dt$ . These are also equal to

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \Pr(A(s) < x) ds = F_e(x)$$

and

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \Pr(Y(s) < x) ds = F_e(x)$$

If  $F$  is not lattice, then

$$\lim \Pr(A(t) < a) = \lim \Pr(Y(t) < a) = F_e(t).$$

*Proof.* Say the system is “on” at time  $t$  if  $A(t) < x$ , “off” otherwise. A cycle is a renewal. So the system is “on” for the first  $x$  units of a renewal cycle, and then it is “off.” Once a renewal occurs, the age goes back to 0, so this is an alternating renewal process. The “on” time in a cycle is  $\min\{x, X\}$  because it’s on for the first  $x$  units of time the item is in use of a cycle, unless the cycle is less than  $x$ , in which case it’s on for the whole length of the cycle. Therefore we have that the expected “on” time is

$$\frac{\mathbb{E}(\text{on time in cycle})}{\mu} = \frac{\mathbb{E}(\min\{x, X\})}{\mu} = F_e(x)$$

where  $\mu$  is the average length of a cycle. Note that

$$\mathbb{E}(\min\{x, X\}) = \int_0^\infty \Pr(\min\{x, X\} > t) dt = \int_0^x \bar{F}(t) dt$$

so we have that the expected “on” time is

$$\frac{1}{\mu} \int_0^x \bar{F}(t) dt = F_e(x)$$

Now we will show that the expected excess is the same as the expected age. Say that the system is “on” at  $t$  if  $Y(t) > x$  and “off” otherwise. A cycle happens every time a renewal occurs. So if a renewal cycle is longer than  $x$ , the system will be initially on and then become off for the last  $x$  time units of a renewal. So the off time is  $\min\{x, X\}$ . Again, by a similar argument as above we get the same result. (Intuitively this makes sense because the distribution of interarrival times is the same if you look backward in time as it is if you look forward in time.)

□

Recall Proposition 7.5.3: The average age and average excess both equal  $\mathbb{E}(X^2)/(2\mathbb{E}(X))$ . Note that

$$A(t) + Y(t) = S_{N(t)+1} - S_{N(t)} = X_{N(t)+1}.$$

**Proposition 7.5.9 ([Proposition 3.4.6 in *Stochastic Processes*]).** With probability 1, the average value of  $X_{N(t)+1}$  is

$$\frac{\mathbb{E}(X^2)}{\mathbb{E}(x)} > \mathbb{E}(X)$$

*Proof.* Suppose  $X_e \sim F_e$  where

$$F_e(x) = \frac{1}{\mu} \int_0^x \bar{F}(y) dy$$

Note that

$$\frac{dF_e(x)}{dx} = \frac{\bar{F}(x)}{\mu}$$

We have

$$\begin{aligned} \mathbb{E}(X_e) &= \int_0^\infty x dF_e(x) = \int_0^\infty x \frac{\bar{F}(x)}{\mu} dx = \int_0^\infty \frac{x}{\mu} \int_x^\infty dF(y) dx = (\text{by Fubini's Theorem}) \int_0^\infty \int_0^y \frac{x}{\mu} dx dF(y) \\ &= \int_0^\infty \frac{y^2}{2\mu} dF(y) = \frac{\mathbb{E}(X^2)}{2\mu} \end{aligned}$$

□

**Proposition 7.5.10 (Problem on homework 7; inspection paradox).**

$$\Pr(X_{N(t)+1} > x) > \bar{F}(x).$$

### 7.5.2 Equilibrium renewal processes

**Definition 7.24 (Equilibrium renewal process).** Suppose we have a delayed renewal process where  $X_1 \sim F_e$ ,  $X_i \sim F \forall i > 1$ . Then  $\{N_e(t), t \geq 0\}$  is called an **equilibrium renewal process**. (the time until the first event has the equilibrium distribution.) We let  $m_e(t) = \mathbb{E}(N_e(t))$ .

**Example 7.24.** Suppose you arrive at some time  $t$  during a renewal process with finite expected renewal time  $\mu$ . Then you will have to wait  $Y(t)$  time until the next renewal. Let  $F_t$  be the distribution function of  $Y(t)$ . After that first arrival, all of the other arrival times will have distribution  $F$ . If the time  $t$  at which you start observing is very large, you observe an equilibrium renewal process (because  $F_t \rightarrow F_e$  as  $t \rightarrow \infty$ ).

**Proposition 7.5.11 (Theorem 3.5.2 from *Stochastic Processes*).** Let  $\{N_e(t), t \geq 0\}$  be an equilibrium renewal process. Let the excess at time  $t$  of  $\{N_e(t)\}$  be  $Y_e(t)$ .

- (1)  $Y_e(t) \sim F_e$  for all  $t$ .
- (2)  $N_e(t+s) - N_e(t)$  has the same distribution for all  $t$  (this counting process has stationary increments—see Definition 7.4).
- (3)  $m_e(t) = t/\mu$ .

*Proof (More of an intuitive argument than a complete proof).* (1) Consider an ordinary renewal process with interarrival times distributed as  $F$ . Suppose we start observing at a very large time  $t'$ , so we are observing an equilibrium renewal process. That is, the time until the first event is distributed as  $F_e$ . We'd like to know the time until the next event after you've waited a time  $t$ . We have

$$Y_e(t) = Y(t+t') \sim F_e$$

since  $t'$  is very very large.

- (2) Similar argument—after arriving after a very large amount of time, the distribution is in equilibrium, so it makes no difference when you start watching, just matters how large  $s$  is.
- (3)

$$N_e(t+s) = N_e(t+s) - N_e(t) + N_e(t)$$

$$\iff \mathbb{E}(N_e(t+s)) = \mathbb{E}[N_e(t+s) - N_e(t)] + \mathbb{E}[N_e(t)] \iff m_e(t+s) = m_e(s) + m_e(t)$$

where  $\mathbb{E}(N_e(t+s)) = m_e(t+s)$  by definition,  $\mathbb{E}[N_e(t+s) - N_e(t)] = m_e(s)$  by the result from part (2), and  $\mathbb{E}[N_e(t)] = m_e(t)$  by definition. It is the case that  $f(x+y) = f(x) + f(y) \implies f(x) = cx$  for some constant  $c$  if  $f$  is measurable. Because of that, we have that

$$m_e(t) = ct$$

where  $c$  is some constant. By the Elementary Renewal Theorem (Theorem 7.3.9),

$$\frac{m_e(t)}{t} \rightarrow \frac{1}{\mu}$$

which implies that  $c$  is  $\mu^{-1}$ .

□

Proving that the previous statement is true for any measurable function requires measure theory, but it is easy to prove it is true if  $f$  is differentiable.

**Lemma 7.5.12.** If  $f$  is differentiable,  $f(x + h) = f(x) + f(h)$ .

*Proof.* Note that

$$\lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h} = f'(x)$$

Note that

$$\lim_{h \rightarrow 0} \frac{f(h)}{h} = f'(0)$$

Then  $f'(x) = c$ ,  $f(x) = cx + k$ , so  $f(x) = cx$ .

□

**Theorem 7.5.13 (Blackwell's Theorem (Theorem 3.4.1 in *Stochastic Processes*, p.110)).** (i)

If  $F$  is not lattice, then

$$m(t + a) - m(t) \rightarrow \frac{a}{\mu} \text{ as } t \rightarrow \infty$$

(ii) If  $F$  is lattice with period  $d$ , then

$$\mathbb{E}(\text{number of renewals at } nd) \rightarrow \frac{d}{\mu} \text{ as } n \rightarrow \infty.$$

**Remark.** If  $\Pr(X_i = 0) = 0$  then this means that  $\Pr(\text{renewal at } nd) \rightarrow d/\mu$  as  $n \rightarrow \infty$ .

*Proof (intuitive; rigorous proof is quite technical).* (i) If you arrive at time  $t$ , then  $\{N(s), s \geq 0\}$  is a delayed renewal process. That is,  $X_1$  has the distribution of  $Y(t) \sim F_t$ , and the remaining  $X_i \sim F$ . So

$$m(t + a) - m(t) = \mathbb{E}[N_t(a)].$$

We already know that as  $t \rightarrow \infty$ ,  $F_t \rightarrow F_e$ . So as  $t \rightarrow \infty$ ,  $m(t + a) - m(t) = \mathbb{E}[N_t(a)] \rightarrow \mathbb{E}[N_e(a)]$ . But by Proposition 7.5.11, **the result follows** (??).

(ii) Suppose we have a renewal process where

$$\sum_{i=0}^{\infty} \Pr(X = i) = 1.$$

Then  $m(n)$  is the expected number of renewals by time  $n$ . That is,

$$m(n) = \mathbb{E} \sum_{i=0}^n (\text{number of renewals at } i) = \sum_{i=0}^n \mathbb{E}(\text{number of renewals at } i)$$

Then by the Elementary Renewal Theorem (Theorem 7.3.9),

$$\frac{m(n)}{n} = \frac{1}{n} \sum_{i=0}^n \mathbb{E}(\text{number of renewals at } i) \rightarrow \frac{1}{\mu}$$

In general, this does not imply that

$$\mathbb{E}(\text{number of renewals at } nd) \rightarrow \frac{d}{\mu} \text{ as } n \rightarrow \infty.$$

For example, consider the discrete random variable

$$X_i = \begin{cases} 2 & \text{with probability } 1/2 \\ 4 & \text{with probability } 1/2 \end{cases}$$

Then

$$\mathbb{E}(\text{number of renewals at } n) = m(n) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\text{number of renewals at } i) \rightarrow \frac{1}{\mu}$$

But of course all of these sums are 0 when  $i$  is odd. So we could also write this as

$$\begin{aligned} m(n) &= \sum_{i=1}^{\lfloor n/2 \rfloor} \frac{1}{n} \mathbb{E}(\text{number of renewals at } 2i) \rightarrow \frac{1}{\mu} \\ \iff m(n) &= \frac{1}{2} \sum_{i=1}^{\lfloor n/2 \rfloor} \frac{1}{n/2} \mathbb{E}(\text{number of renewals at } 2i) \rightarrow \frac{1}{\mu} \\ \iff m(n) &= \sum_{i=1}^{\lfloor n/2 \rfloor} \frac{1}{n/2} \mathbb{E}(\text{number of renewals at } 2i) \rightarrow \frac{2}{\mu} \\ &\vdots \end{aligned}$$

You get pointwise convergence in the lattice case with period  $d$  when  $d$  is the period.

□

Recall Proposition 7.5.3: The average age and average excess both equal  $\mathbb{E}(X^2)/(2\mathbb{E}(X_e))$ , where  $X_e \sim F_e$ .

**Theorem 7.5.14 (Proposition 3.4.6 in *Stochastic Processes*).** If  $F$  is not lattice,

$$\lim_{t \rightarrow \infty} \mathbb{E}(A(t)) = \lim_{t \rightarrow \infty} \mathbb{E}(Y(t)) = \frac{\mathbb{E}(X^2)}{2\mathbb{E}(X)}.$$

*Proof.* Won't prove; requires Key Renewal Theorem, which is beyond scope of this course. Some notes: we know that  $F_{Y(t)} \rightarrow F_e$ , or  $Y(t) \rightarrow Y(\infty)$  where  $Y(\infty) \sim F_e$ . Then

$$\lim \mathbb{E}(Y(t)) = (? \text{ can't justify now}) \mathbb{E}(\lim(Y(t))) = \mathbb{E}(X_\infty)$$

□

**Corollary 7.5.14.1 (Corollary 3.4.7 in *Stochastic Processes*, p. 121).** If  $\mathbb{E}(X^2) < \infty$  and  $F$  is nonlattice, then

$$m(t) - \frac{t}{\mu} \rightarrow \frac{\mathbb{E}(X^2)}{2\mu^2} - 1 \text{ as } t \rightarrow \infty$$

*Proof.* Recall that by Wald's Equation (Theorem 7.3.8),

$$\mathbb{E}\left(\sum_{i=1}^{N(t)+1} X_i\right) = \mu(m(t) + 1)$$

So

$$\begin{aligned} \mu(m(t) + 1) = \mathbb{E}(t + Y(t)) = t + \mathbb{E}(Y(t)) &\iff m(t) + 1 = \frac{t}{\mu} + \frac{\mathbb{E}(Y(t))}{\mu} \iff m(t) - \frac{t}{\mu} = \frac{\mathbb{E}(Y(t))}{\mu} - 1 \\ &\implies m(t) - \frac{t}{\mu} \rightarrow \frac{\mathbb{E}(X^2)}{2\mu^2} - 1 \text{ as } t \rightarrow \infty \\ &\implies m(t) \approx \frac{t}{\mu} + \frac{\mathbb{E}(X^2)}{2\mu^2} - 1 \end{aligned}$$

□

**Theorem 7.5.15 (Central Limit Theorem for Renewal Processes; Theorem 7.3 in *Introduction to Probability Models*).** Let  $N(t)$  be a renewal process. Then

$$N(t) \xrightarrow{d} \mathcal{N}\left(\frac{t}{\mu}, \frac{t\sigma^2}{\mu^3}\right)$$

where  $\mu = \mathbb{E}(X_i)$  and  $\sigma^2 = \text{Var}(X_i)$ .

**Remark.** If the renewal process is a Poisson process with rate  $\lambda$ , then  $F(x) = 1 - e^{-\lambda x}$ . Then  $\mu = \mathbb{E}(X) = \lambda^{-1}$ ,  $\sigma^2 = \text{Var}(X) = \lambda^{-2}$ ,  $t/\mu = \lambda t$ ,  $t\sigma^2/\mu^2 = t\lambda^3/\lambda^2 = \lambda t$ , which implies the limiting distribution is  $\mathcal{N}(\lambda t, \lambda t)$ , as expected.

*Proof.* Note that  $N(t) < n$  if and only if the  $n$ th event occurred after  $t$ , or  $X_1 + \dots + X_n > t$ . So

$$\Pr(N(t) < n) = \Pr(X_1 + \dots + X_n > t)$$

Then by the Central Limit Theorem,

$$X_1 + \dots + X_n \sim \mathcal{N}(n\mu, n\sigma^2) \text{ (approximately)}$$

⋮

□

### 7.5.3 Regenerative Processes

**Theorem 7.5.16 (Theorem 3.7.1 in *Stochastic Processes*).** Let  $\{X(t), t \geq 0\}$  be a stochastic process with state space  $\{0, 1, 2, \dots\}$  having the property that there exist time points at which the state restarts itself (with probability 1). Let  $S_1, S_2, \dots$  constitute the event times of a renewal process. We call  $X(t)$  a **regenerative process**. So  $\{S_1, S_2, \dots\}$  constitute the event times of a renewal process. We say a cycle is completed every time a renewal occurs. Let  $N(t) = \max\{n : S_n \leq t\}$  denote the number of cycles by time  $t$ .

If  $F$ , the distribution of a cycle, has a density over some interval, and if  $\mathbb{E}(S_1) < \infty$ , then

$$P_j = \lim_{t \rightarrow \infty} \Pr(X(t) = j) = \frac{\mathbb{E}(\text{amount of time in state } j \text{ in a cycle})}{\mathbb{E}(\text{time of a cycle})}.$$

### 7.5.4 Example of Renewal Reward Processes to Patterns

**Example 7.25 (Seems to be similar to Section 7.9.2 in *Introduction to Probability Models*).** Suppose we have i.i.d. data. Number  $i$  appears with probability  $p_i$ . We want to see the mean time until we see the pattern 1213121. An event is when the pattern appears. The next event is the next time the pattern appears, but we can't use data from the last event. Earn \$1 whenever the last 7 values are 1313121 (so you are allowed to use previous values for the reward process). That is, if the cycle is

$$\dots 1213121 | 3121 \dots 1213121 |$$

the events happen at the | marks, but in between there were two rewards (one at the second event, one at the bold 1). So this is a delayed renewal reward process  $(X_i, R_i), i \geq 2$ . Note that the limiting results will be the same for the delayed renewal process as for the normal one. So the expected average reward per unit time is the expected reward during the cycle divided by the expected time of a cycle;

$$\mathbb{E}(\text{average reward per unit time}) = \frac{\mathbb{E}(\text{reward during the cycle})}{\mathbb{E}(T)}$$

We want  $\mathbb{E}(T)$ , the expected length of a cycle. We know  $\mathbb{E}(\text{average reward per unit time})$  is

$$\mathbb{E}(\text{average reward per unit time}) = p_1^4 p_2^2 p_3$$

Now we want  $\mathbb{E}(\text{reward during the cycle})$ . Suppose a cycle just ended (the pattern just appeared). Then we could earn a reward after 4 values (if the last 4 terms of the pattern appear again) after 6 values (if the

last 6 terms of the pattern appear again). And we will earn a reward when the cycle ends (after the full pattern appears again). So if  $Y_i$  is the reward earned during after  $i$  new values,

$$\text{reward during the cycle} = Y_4 + Y_6 + 1 \implies \mathbb{E}(\text{reward during the cycle}) = \mathbb{E}(Y_4) + \mathbb{E}(Y_6) + 1 = p_1^2 p_2 p_3 + p_1^3 p_2^2 p_3 + 1$$

$$\implies \mathbb{E}(T) = \frac{p_1^2 p_2 p_3 + p_1^3 p_2^2 p_3 + 1}{p_1^4 p_2^2 p_3} = \frac{1}{p_1^2 p_2} + \frac{1}{p_1} + \frac{1}{p_1^4 p_2^2 p_3}$$

Note that the mean time between renewals is  $\frac{1}{p_1^4 p_2^2 p_3}$ , the mean time to get 121 is  $\frac{1}{p_1^2 p_2}$ , and the mean time to get 1 is  $\frac{1}{p_1}$ .

**Example 7.26 (Similar to example 7.38 in *Introduction to Probability Models*).** Suppose we are flipping coins, probability of flipping heads is  $p$ . The pattern is  $n$  heads in a row. Cycle is. a pattern without using data from last cycle. Get a reward of 1 each time the last  $n$  values were all heads. Then

$$p^n = \mathbb{E}(\text{average reward per unit time}) = \frac{\mathbb{E}(\text{reward during the cycle})}{\mathbb{E}(T)}$$

Note that

$$\begin{aligned} \text{reward during the cycle} &= Y_1 + \dots + Y_{n-1} + 1 \implies \mathbb{E}(\text{reward during the cycle}) = \sum_{i=1}^{n-1} \mathbb{E}(Y_i) + 1 \\ &= \sum_{i=1}^{n-1} p^i + 1 \\ \implies p^n &= \frac{1}{\mathbb{E}(T)} \left( \sum_{i=1}^{n-1} p^i + 1 \right) \implies \mathbb{E}(T) = p^{-n} \left( \sum_{i=1}^{n-1} p^i + 1 \right) = \left( \frac{1}{p} \right)^n + \left( \frac{1}{p} \right)^{n-1} + \dots + \frac{1}{p} \end{aligned}$$

## 7.6 Markov Chains (Chapter 4 of *Stochastic Processes*; Chapter 4 of *Introduction to Probability Models*)

Suppose we have  $X_0, X_1, X_2, \dots$ , where  $X_n$  is the state of the system at time  $n$ . The set of possible values of states are the nonnegative integers (the number of possible states will be either finite or countable). In a Markov chain, the probability of reaching state  $j$  given that you are currently in state  $i$  is  $P_{ij}$ ; that is,

$$\Pr(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P_{ij}$$

That is, the next state depends only on the current state, not any of the previous states (all the information is contained in the current state). In other words, every starts over again when you reach state  $i$ ; so there is a renewal every time you reach state  $i$ . We say a system has the **Markovian property** if given the present state  $X_n$ , the next state  $X_{n+1}$  is independent of the past  $X_{n-1}, X_{n-2}, \dots$

**Definition 7.25 (Markov chain).** Let  $S$  be the set of possible values of a system.  $S$  must be either finite or countably infinite (in general we will take  $S = \mathbb{N}$ ). Then  $\{X_0, X_1, \dots\}$  is said to be a **Markov chain** with transition probabilities  $\Pr(X_{n+1} = j \mid X_n = i)P_{i,j}$ ,  $i, j \in S$  if

$$\Pr(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \Pr(X_{n+1} = j \mid X_n = i) = P_{ij}.$$

We call the matrix

$$P = [P_{ij}]$$

the **transition probability matrix**. Note that  $\sum_j P_{ij} = 1$ , so the rows of the transition probability matrix all add to 1.

**Remark.** Given the initial state  $\Pr(X_0 = i_0)$  and the probability transition matrix, all probabilities about  $X_0, X_1, \dots$  can (in theory) be determined. That is,

Note that

$$\Pr(X_n = i_n) = \sum_{i_0, i_1, \dots, i_{n-1}} \Pr(X_0 = i_0)P_{i_0, i_1}P_{i_1, i_2} \dots, P_{i_{n-1}, i_n}.$$

**Example 7.27 (Examples 4.1 and 4.4 in *Introduction to Probability Models*).** Suppose we have a state based on weather today and yesterday. Then probability of rain tomorrow is the following:

$$\begin{cases} dd \rightarrow r & \text{with probability 0.3} \\ rd \rightarrow r & \text{with probability 0.4} \\ dr \rightarrow r & \text{with probability 0.6} \\ rr \rightarrow r & \text{with probability 0.7} \end{cases}$$

The transition matrix looks as follows:

$$P = \begin{bmatrix} & dd & rd & dr & rr \\ dd & 0.7 & 0 & 0.3 & 0 \\ rd & 0.6 & 0 & 0.4 & 0 \\ dr & 0 & 0.4 & 0 & 0.6 \\ rr & 0 & 0.3 & 0 & 0.7 \end{bmatrix}$$

**Example 7.28 (Random walk; examples 4.5 and 4.6 in *Introduction to Probability Models*).**

$$P_{i,i+1} = p, P_{i,i-1} = 1 - p, \quad S = \mathbb{Z}$$

One example is Gambler's Ruin:  $X_0 = i$ , stop when fortune is either 0 or  $N$ . (See Example 7.41 and several solutions in Grimmett and Stirzaker.)

**Example 7.29 (Similar to example 4.10 in *Introduction to Probability Models*, similar to homework problem from Math 505A).** Urn with 2 balls, red and blue. Each period, choose ball from urn at random. If red then replace with blue. If blue then replace with red with probability 0.7, with blue with probability 0.3. What is the long-run proportion of time the chosen ball is red?

**Solution.** Let  $X_n$  be an indicator variable for the  $n$ th ball chosen to be red. Note that if you just chose a red ball, it is not clear what the probability of choosing a red ball next is, because it depends on what color the other ball in the urn is. So consider the state of the Markov chain to be the number of red balls in the urn, and let  $X_n$  be the number of red balls in the urn before the  $n$ th withdrawal. Then we have the following transition matrix:

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0.5 & 0.5 \cdot 0.3 & 0.5 \cdot 0.7 \\ 0 & 1 & 0 \end{bmatrix}$$

Let  $\pi_j$  be the long-run proportion of the time the system is in state  $j$ . Note that every time we reach state  $j$  everything starts all over again, so that's a renewal (ordinary renewal process) if you are certain to eventually return to state  $j$  (we will assume this is true for now). We have

$$\pi_0 \cdot 0 + \pi_1 \cdot \frac{1}{2} + \pi_2 = \pi_2 + \frac{1}{2}\pi_1$$

is the long-run probability of choosing a red ball.

**Example 7.30 (Embedded Markov chain; example 4.1(a) in *Stochastic Processes*).**  $M/G/1$  queueing process. The number of people currently in the system is not Markovian, but it would be if the service time were exponential (since then it wouldn't matter how long the customer had been in service). But the number of people in the system immediately after the  $n$ th service completion is Markovian. (Called an **embedded Markov chain** because it is a Markov chain if you only look at it at certain times.)

Note that

$$\begin{aligned} P_{0j} = a_j &= \Pr(j \text{ arrivals during service time}) = \int_0^\infty \Pr(j \text{ arrivals during service time } t) dG(t) \\ &= \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dG(t) \end{aligned}$$

More generally, note that for  $i > 0, j \geq i - 1$

$$\begin{aligned} P_{ij} = a_{j-i+1} &= \Pr(j-i+1 \text{ arrivals during service time}) = \int_0^\infty \Pr(j-i+1 \text{ arrivals during service time } t) dG(t) \\ &= \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{(j-i+1)}}{(j-i+1)!} dG(t). \end{aligned}$$

**Example 7.31 (Example 4.1(b) in *Stochastic Processes*).** if  $X_n$  is the number of people in the system as seen by the  $n$ th arrival, then this is an (embedded) Markov chain.

⋮

We want  $P_{ij}, j > 0$ . If  $n$  is the number of people served, then

$$i + 1 - n = j \iff n = i + 1 - j$$

Then

$$P_{ij} = \int \Pr(i+1-j \text{ services in time } T \mid T=s) dF(s) = \int e^{-\mu s} \frac{(\mu s)^{i+1-j}}{(1+1-j)!} dF(s)$$

$P_{i0}$  is different because if we end with 0 people, all of the previous people had to depart. But they could have departed at any time before the next person arrived. Imagine that when all the customers leave, the server keeps serving “imaginary” customers. Then  $P_{i0}$  is the probability of at least  $i+1$  customers (real or imaginary) being served.

**Definition 7.26** ( $n$ -step transition probabilities).

$$P_{ij}^n = \Pr(X_n = j \mid X_0 = i)$$

### 7.6.1 Chapman-Kolmogorov Equations—section 4.2 of *Introduction to Probability Models*, section 4.2 of *Stochastic Processes* (p. 178 of pdf)

**Proposition 7.6.1** (Chapman-Kolmogorov Equations).

**Definition 7.27** (Matrix of  $n$ -step transition probabilities).

$$\mathbf{P}^{(n)} = [P_{ij}^n]$$

**Example 7.32** (Similar to example 4.10 in *Introduction to Probability Models*). Urn initially has 1 red ball 1 blue. At each stage choose a ball at random. If red, replace with blue. If blue, replace with blue with probability 1/3 or red with probability 2/3. What is the probability that the 5th ball selected is red?

**Solution.** Note that per the example from last time, knowing the ball you just drew is not sufficient to have a Markov chain (need to know color of other ball in urn to get probabilities of next state). So let  $X_n$  be the number of red balls in the urn after you draw the  $n$ th ball, and note that  $X_0 = 1$ . Note that there are three states. The probability transition matrix is

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 \end{bmatrix}$$

Note that the probability the 5th ball selected is red is equal to

$$\sum_{i=1}^2 \Pr(5th \text{ ball is red} \mid X_4 = i) P_{ii}^4 = 0 \cdot P_{10}^4 + \frac{1}{2} P_{11}^4 + 1 \cdot P_{12}^4$$

Now we calculate  $P^4$ .

$$P^2 = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{4}{9} & \frac{3}{9} & \frac{2}{9} \\ \frac{3}{12} & \frac{25}{36} & \frac{2}{36} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

$$P^4 = P^2 P^2 = \begin{bmatrix} \frac{4}{9} & \frac{3}{9} & \frac{2}{9} \\ \frac{3}{12} & \frac{25}{36} & \frac{2}{36} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{4}{9} & \frac{3}{9} & \frac{2}{9} \\ \frac{3}{12} & \frac{25}{36} & \frac{2}{36} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{bmatrix} = \dots$$

**Example 7.33 (Example 4.12 from *Introduction to Probability Models*).** Part (c) (not in book):  $N$  is the number of flips until either a run of 3 heads or 3 tails.  $\Pr(N = 8)$ ?

**Solution.** Say start at state 0, state  $i$  if you are on a run of  $i$  heads for  $i = 1, 2, 3$ , and state  $i$  if you are on a state of  $i - 3$  tails,  $i = 4, 5, 6$ . Then we have

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & P_{03} & P_{04} & P_{05} & P_{06} \\ P_{10} & P_{11} & P_{12} & P_{13} & P_{14} & P_{15} & P_{16} \\ P_{20} & P_{21} & P_{22} & P_{23} & P_{24} & P_{25} & P_{26} \\ P_{30} & P_{31} & P_{32} & P_{33} & P_{34} & P_{35} & P_{36} \\ P_{40} & P_{41} & P_{42} & P_{43} & P_{44} & P_{45} & P_{46} \\ P_{50} & P_{51} & P_{52} & P_{53} & P_{54} & P_{55} & P_{56} \\ P_{60} & P_{61} & P_{62} & P_{63} & P_{64} & P_{65} & P_{66} \end{bmatrix} = \begin{bmatrix} 0 & p & 0 & 0 & 1-p & 0 & 0 \\ 0 & 0 & p & 0 & 1-p & 0 & 0 \\ 0 & 0 & 0 & p & 1-p & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 1-p & 0 \\ 0 & p & 0 & 0 & 0 & 0 & 1-p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

In general, if  $T_i$  is the number of transitions to enter state  $i$  even state 0 is 0, we can change state  $i$  into an absorbing state.

### 7.6.2 Classification of States (Section 4.3 of *Introduction to Probability Models*)

We say state  $j$  is *accessible* from state  $i$  if for some  $n \geq 0$   $P_{ij}^n > 0$ . Note that  $P_{ij}^n = \Pr(X_n = j \mid X_0 = i)$ . So  $j$  is accessible from  $i$  if and only if starting in  $i$  it is possible that the Markov chain is every in  $j$ . So if  $j$  is not accessible from  $i$

$$\Pr(\text{ever in } j \mid X_0 = i) = \Pr\left(\bigcup_n \{X_n = j\} \mid X_0 = i\right) \leq \sum_n \Pr(\{X_n = j\} \mid X_0 = i) = 0$$

Two states  $i$  and  $j$  that are accessible to each other are said to *communicate*, and we write  $i \leftrightarrow j$ . Two states that communicate are said to be in the same *class*.

A Markov chain is **irreducible** if there is only one class.

#### Example 7.34.

$$P = \begin{bmatrix} 0.2 & 0.8 & 0 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0 & 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

Then the classes are  $\{0, 1\}$ ,  $\{2\}$ , and  $\{3, 4\}$ . Note that for the classes  $\{0, 1\}$  and  $\{3, 4\}$  you will stay there forever—they are essentially Markov chains themselves. But if you start in state 2 you will eventually go to one of the other states and never go back. In contrast consider

$$P = \begin{bmatrix} 0.2 & 0.8 & 0 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

where all three states have this property. We can merge classes  $\{2\}$  and  $\{3, 4\}$  like this:

$$P = \begin{bmatrix} 0.2 & 0.8 & 0 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0 & 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0.1 & 0.5 & 0.4 \end{bmatrix}$$

Let  $N_k$  be the number of transitions until you first enter state  $k$ . (discussion on p. 195 - 196 of *Introduction to Probability Models*.)

**Example 7.35 (Example 4.12 in *Introduction to Probability Models*).**

$$\Pr(N = 8) = P_{02}^7 p = P_{00}^5 p^3$$

**Definition 7.28.** For any states  $i$  and  $j$  in a Markov process, let  $f_{ij}^n$  be the probability that starting in  $i$  the first transition into  $j$  occurs at time  $n$ . Formally,

$$f_{ij}^0 = 0, \quad f_{ij}^n = \Pr(X_n = j, X_k \neq j, k = 1, \dots, n-1 \mid X_0 = i).$$

Let

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^n.$$

Then  $f_{ij}$  denotes the probability of ever making a transition into state  $j$  given that the process starts in  $i$ .

**Proposition 7.6.2 (Homework problem; Problem 4.4 in *Stochastic Processes*).**

$$P_{ij}^n = \sum_{k=0}^n f_{ij}^k P_{jj}^{n-k}.$$

*Proof.* Recall that  $P_{ij}^n$  is the probability that a process in state  $i$  will be in state  $j$  after  $n$  additional transitions. Also,  $f_{ij}^n$  is the probability that starting in  $i$  the first transition into  $j$  occurs at time  $n$ . Note that

$$P_{ij}^n = \sum_{k=0}^n \Pr(\text{first reach } j \text{ at time } k) \cdot \Pr(\text{end at } j \text{ starting from } j \text{ in } n-j \text{ steps}) = \sum_{k=0}^n f_{ij}^k P_{jj}^{n-k}.$$

□

**Proposition 7.6.3 (Homework problem; Problem 4.5 in *Stochastic Processes*).** Let

$$P_{ij/k} = \Pr(X_n = j, X_\ell \neq k, \ell = 1, \dots, n-1 \mid X_0 = i).$$

(This is the probability that at time  $n$  we are in state  $j$  and we never reached state  $k$  between time 0 and  $n$ , given that we started in state  $i$ .) Then for  $i \neq j$ ,

$$P_{ij}^n = \sum_{k=0}^n P_{ii}^k P_{ij/i}^{n-k}.$$

*Proof.* There are two classes of ways we can reach  $j$  starting from  $i$ : we either go back to  $i$  at some point and then end up at  $j$ , or we go straight to  $j$  without ever going back to  $i$ . If we go back to  $i$  for the last time at time  $k \in \{1, \dots, n-1\}$ , we can end up at  $j$  at time  $n$  with probability  $P_{ij/i}^{n-k}$ . We end up back at state  $i$  for the last time at time  $k$  before time  $n$  (regardless of what happened before time  $k$ ) with probability  $P_{ii}^k$ . Therefore,

$$\text{for } i \neq j, P_{ij}^n = \sum_{k=0}^n P_{ii}^k P_{ij/i}^{n-k}.$$

□

**Definition 7.29 (Recurrent state).** We call state  $i$  **recurrent** if starting in  $i$  the Markov Chain returns to  $i$  with probability 1.

**Definition 7.30 (Transient state).** We call state  $i$  **transient** if starting in  $i$  the probability the Markov Chain returns to  $i$  is less than 1.

**Proposition 7.6.4 (Proposition 4.1 in *Introduction to Probability Models*, Proposition 4.2.3 in *Stochastic Processes*).** In a Markov chain, state  $i$  is recurrent if

$$\sum_{n=1}^{\infty} P_{ii}^n = \infty$$

and transient if

$$\sum_{n=1}^{\infty} P_{ii}^n < \infty.$$

*Proof.* State  $i$  is recurrent if with probability 1 a process starting at  $i$  will eventually return to  $i$ . However, by the Markovian property it follows that this process will return again to  $i$  with probability 1, and so on. There is no last term, so with probability 1, the number of visits to  $i$  will be infinite.

On the other hand, suppose  $i$  is transient. Then each time the process returns to  $i$  there is a positive probability  $1 - f_{ii}$  that it will never again return. That is, the probability the total time in  $i$  is  $n$  periods given you start in  $i$  is

$$f_{ii}^{n_i-1}(1 - f_{ii}).$$

Therefore the number of visits is geometric with finite mean  $(1 - f_{ii})^{-1}$ . Let  $I_n$  be an indicator variable for  $X_n = i$ . Then the total number of visits to state  $i$  is  $\sum_{n=0}^{\infty} I_n$  and the expected number of visits if we start at  $i$  is

$$\mathbb{E} \sum_{n=0}^{\infty} \{I_n = 1 \mid X_0 = i\} = \sum_{n=0}^{\infty} \mathbb{E}\{I_n = 1 \mid X_0 = i\} = \sum_{n=0}^{\infty} \Pr\{X_n = i \mid X_0 = i\}$$

□

**Corollary 7.6.4.1 (Corollary 4.2 in *Introduction to Probability Models*, Corollary 4.2.4 in *Stochastic Processes*).** If state  $i$  is recurrent and state  $i$  communicates with state  $j$ , then state  $j$  is recurrent.

*Proof.* Let  $m$  and  $n$  be such that  $P_{ij}^n < 0$ ,  $P_{ji}^m > 0$  (these values exist since  $i$  and  $j$  communicate). Then for any  $s \geq 0$ ,

$$P_{jj}^{m+n+s} \geq P_{ji}^m P_{ii}^s P_{ij}^n.$$

Therefore

$$\sum_{s=0}^{\infty} P_{jj}^{m+n+s} \geq P_{ji}^m P_{ij}^n \sum_{s=0}^{\infty} P_{ii}^s = \infty \implies \sum_{n=0}^{\infty} P_{jj}^n = \infty$$

where  $\sum_{s=0}^{\infty} P_{ii}^s = \infty$  by definition of  $i$  being recurrent.

□

**Corollary 7.6.4.2 (Corollary 4.2.5 in *Stochastic Processes*).** If state  $i$  is recurrent and state  $i$  communicates with state  $j$ , then

$$f_{ij} = \Pr(\text{ever enter } j \mid X_0 = i) = 1$$

*Proof.* Suppose  $X_0 = i$ , and let  $n$  be such that  $P_{ij}^n > 0$ . Say that we miss opportunity 1 if  $X_n \neq j$ . If we miss opportunity 1, then let  $T_1$  denote the next time we enter  $i$ . Note that  $T_1$  is finite with probability 1 by Corollary 7.6.4.1. Say we miss opportunity 2 if  $X_{T_1+n} \neq j$ . If opportunity 2 is missed, let  $T_2$  denote the next time we enter  $i$  and say we miss opportunity 3 if  $X_{T_2+n} \neq j$ , and so on. It is easy to see that the opportunity number of the first success is a geometric random variable with mean  $1/P_{ij}^n$ , and is thus finite with probability 1. The result follows since  $i$  being recurrent implies that the number of potential opportunities is infinite.

□

So we have the recurrence and transience are class properties (if one element in the class has them, they all do).

**Example 7.36** (Simple random walk; example 4.18 in *Introduction to Probability Models*).

$$\frac{(2n)!}{n!n!} \frac{(2n)^{2n+1/2} e^{-2n} \sqrt{2\pi}}{n^{2n+1} e^{-2n} 2\pi} = \frac{4^n}{\sqrt{n\pi}}$$

Homework problem: If  $R$  is a recurrent class,  $i \in R$ ,  $j \notin R$ , then  $P_{ij} = 0$ . Why? Suppose  $P_{ij} > 0$ . But since  $i$  doesn't communicate with  $j$ ,  $P_{ij}^n = 0$  for all  $n$ . Contradiction.

### 7.6.3 Long-Run Proportions and Limiting Probabilities (Limit Theorems) (4.4 in *Introduction to Probability Models*, 4.3 in *Stochastic Processes*)

Let  $\pi_j$  be the long-run proportion of the time you are in state  $j$ . That is,

$$\pi_j = \lim_{n \rightarrow \infty} \frac{N_n(j)}{n}$$

where  $N_n(j)$  is the time in  $j$  during the first  $n$  periods. Then this equals  $1/m_j$  where  $m_j$  is the expected time to re-enter  $j$  given that we started in  $j$  if state  $j$  is recurrent.

Bad approach to find:

$$m_j = \sum_k \mathbb{E}[\text{return to } j \mid X_1 = k] P_{jk} = P_{jj} + \sum_{k \neq j} (1 + m_{kj}) P_{jk} = 1 + \sum_{k \neq j} m_{kj} P_{jk}$$

$$m_{ij} = 1 + \sum_{k \neq j} P_{ik} m_{kj}$$

for  $|S|$  states.

**Proposition 7.6.5** (Proposition 4.4 in *Introduction to Probability Models*). If the Markov chain is irreducible and recurrent, then for any initial state,

$$\pi_j = \frac{1}{m_j}.$$

**Proposition 7.6.6** (Proposition 4.5 in *Introduction to Probability Models*).

**Theorem 7.6.7** (Theorem 4.1 in *Introduction to Probability Models*). In an irreducible Markov chain is positive recurrent, then the long-run proportions are the unique solution of the equations

$$\pi_j = \sum_i \pi_i P_{ij}, \quad j \geq 1$$

$$\sum_j \pi_j = 1.$$

Moreover, if there is no solution of the preceding linear equations, then the Markov chain is either transient or null recurrent and all  $p_{ij} = 0$ .

*Proof.*  $\pi_i$  is the long-run proportion of time you are in state  $i$ ; think of it as the long-run proportion of transitions that come from  $i$ . These transitions go to  $j$  with probability  $P_{ij}$ . So  $\pi_i P_{ij}$  is the long-run proportion of transitions that go from  $i$  to  $j$ . If we sum over all  $i$  then we get the long-run proportion of transitions that go into  $j$ , which is the same as the proportion of time you're in  $j$ . So

$$\pi_j = \sum_i \pi_i P_{ij}$$

□

For an irreducible Markov chain, let  $N_n(j)$  be the number of transitions into  $j$  by time  $n$ . Every time we reach state  $j$  we have a renewal. By the Strong Law for Renewal Processes (Theorem 7.3.6),

$$\frac{N_n(t)}{n} \xrightarrow{\text{a.s.}} \frac{1}{m_{jj}}$$

where  $m_{jj}$  is the expected number of transitions returning to  $j$  given that  $X_0 = j$ . By the Elementary Renewal Theorem (Theorem 7.3.9),

$$\lim_{s \rightarrow \infty} \frac{\mathbb{E}(N_n(s))}{n} = \frac{1}{m_{jj}}.$$

Note that

$$N_n(j) = \sum_{k=1}^n I_k$$

where  $I_k$  is an indicator variable for  $X_k = j$ .

$$\implies \frac{1}{n} \sum_{k=1}^n \Pr(X_k = j) \xrightarrow{\text{a.s.}} \frac{1}{m_{jj}}.$$

Let  $\pi_j = \frac{1}{m_{jj}}$ .

**Definition 7.31.** If state  $j$  is recurrent, call it **positive recurrent** if  $m_{jj} < \infty$  and **null recurrent** if  $m_{jj} = \infty$ .

**Remark.** Being positive recurrent is equivalent to  $\pi_j > 0$ .

**Proposition 7.6.8 (Proposition 4.3.2 in *Stochastic Processes* (p.185 of pdf)).** Positive (null) recurrence is a class property.

*Proof.* Suppose state  $i$  is positive recurrent; that is  $\pi_i > 0$ . We will show that if it communicates with another state  $j$ , that state must be recurrent. Let  $n$  be such that  $P_{ij}^n > 0$ ; such an  $n$  exists since  $i$  and  $j$  communicate.. Then  $\pi_i P_{ij}^n$  is the long-run proportion of the time the chain is in state  $i$  and will be in state  $j$   $n$  time periods later (or the proportion of time that the Markov chain is in state  $j$  and was in state  $i$   $n$  time periods earlier). Since this is less than or equal to the proportion of time the Markov chain is in state  $j$ , we have

$$\pi_i P_{ij}^n \leq \pi_j$$

But since  $\pi_i > 0$ , we have

$$0 < \pi_i P_{ij}^n \leq \pi_j \implies \pi_j > 0.$$

□

**Remark.** This also shows that null recurrence is a class property (because if one element of the class is not positive recurrent, it must be that they are all not positive recurrent).

Note that if we have a Markov chain with  $m < \infty$  states. we have to have

$$\begin{aligned} \sum_{j=1}^m N_n(0) = n &\iff \frac{1}{m} \sum_{j=1}^m N_n(0) = 1 \\ &\implies 1 = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^m N_n(0) = \sum_{j=1}^m \lim_{n \rightarrow \infty} \frac{N_n(0)}{n} = \sum_{j=1}^m \pi_j \end{aligned}$$

so all the  $\pi_j$  have to add up to 1.

**Definition 7.32.** The non-negative numbers  $x_j, j \in S$  are said to be **stationary probabilities** of a Markov chain if

$$x_j = \sum_i x_i P_{ij}, \quad i \in S$$

$$\sum_{j \in S} x_j = 1.$$

**Definition 7.33 (p. 185 of *Stochastic Processes* pdf).** A probability distribution  $\{P_j, j \geq 0\}$  is said to be **stationary** for the Markov chain if

$$P_j = \sum_{i=0}^{\infty} P_i P_{ij}, \quad j \geq 0.$$

**Lemma 7.6.9.** Suppose  $\{x_i, i \in S\}$  is a stationary probability vector for a Markov chain. If  $\Pr(X_0 = i) = x_i, i \in S$   $\Pr(X_n = j) = x_j, j \in S$  for all  $n$ .

*Proof.* We will prove this by induction on  $n$  for  $X_n$ . The case  $n = 0$  is true by assumption. Assume  $\Pr(X_n = i) = x_i, i \in S$ . Then

$$\Pr(X_{n+1} = j) = \sum_{i \in S} \Pr(X_{n+1} = j | X_n = i) \Pr(X_n = i) = \sum_{i \in S} P_{ij} x_i = x_j$$

where the last step follows by Definition 7.32.  $\square$

By the Elementary Renewal Theorem (Theorem 7.3.9),

$$\sum_{k=1}^n \frac{1}{n} \Pr(X_k = j) \xrightarrow{a.s.} \frac{1}{m_{jj}}.$$

Suppose the initial state is chosen randomly according to the probability vector.

**Proposition 7.6.10 (Stationary probability vector is  $\pi_j$ ).** If  $x_j$  is a stationary probability vector then  $x_j = \pi_j, j \in S$ .

*Proof.* Let  $x_j$  be a stationary probability vector. Suppose  $\Pr(X_0 = i) = x_i, i \in S$ . Then  $\Pr(X_k = j) = x_j$  for all  $k$  by Lemma 7.6.9. By the Elementary Renewal Theorem (Theorem 7.3.9),

$$\frac{1}{n} \Pr(X_k = j) \xrightarrow{a.s.} \pi_j.$$

But also

$$\lim_{n \rightarrow \infty} \frac{1}{n} \Pr(X_k = j) = x_j$$

so  $\pi_j = x_j$ .  $\square$

**Definition 7.34 (Ergodic Markov chain state; p.174 of pdf of *Stochastic Processes*).** A positive recurrent, aperiodic state is called **ergodic**.

**Definition 7.35 (Ergodic Markov chain; p.188 of pdf of *Stochastic Processes*).** We say a Markov chain is **ergodic** if all states are positive recurrent; that is,

$$\pi_j = \lim_{n \rightarrow \infty} P_{ij}^n > 0.$$

In this case,  $\{\pi_j, j = 0, 1, 2, \dots\}$  is a stationary distribution and there exists no other stationary distribution.

**Definition 7.36.** A Markov chain that can only return to a state in a multiple of  $d > 1$  steps is said to be **periodic** and does not have limiting probabilities. That is,  $P_{ij}^n = 0$  except when  $n$  is a multiple of  $d$  and  $d$  is the largest such integer.

An irreducible chain that is not periodic is said to be **aperiodic**.

**Lemma 7.6.11.** Periodicity is a class property. That is, if  $i$  communicates with  $j$  then  $i$  and  $j$  have the same period.

*Proof.* Requires non-trivial results from algebra to prove.

□

Also,  $P_{jj}^{nd} \rightarrow d/m_{jj}$ . For an aperiodic irreducible Markov chain,  $P_{ij}^n \rightarrow 1/m_{jj} = \pi_j$ . (Related to Blackwell's Theorem.)

**Example 7.37 (Example 4.18 in *Introduction to Probability Models*).** Stopping time:

$$N = \min\{n : X_1 + \dots + X_n = 1\} \implies 1 = X_1 + \dots + X_N$$

Apply Wald's Equation (Theorem 7.3.8):

$$1 = \mathbb{E}[X_1 + \dots + X_N] = \mathbb{E}(X)\mathbb{E}(N) = 0$$

So Wald's equation leads to a contradiction, meaning its assumptions weren't satisfied, which must be because  $\mathbb{E}(N) = \infty$ . See also Example 6.3 in the Probability notes.

**Example 7.38 (Example 4.24 in *Introduction to Probability Models*).** Not an alternating renewal process because in general it depends on which “up” state you enter into. Typically use renewal reward processes to find average length of “up” and “down” periods.

**Theorem 7.6.12 (Theorem 4.3.3 in *Stochastic Processes*, p. 186 of pdf).** Suppose we have an irreducible aperiodic Markov chain. If it is either transient or null recurrent, then there are no solutions of the equations

$$x_j = \sum_{i \in S} x_i P_{ij}, \quad j \in S,$$

$$\sum_j x_j = 1.$$

If it is positive recurrent, then  $\{\pi_j, j \in S\}$  uniquely satisfy the preceding equations. (There is no irreducible finite state Markov chain.)

*Proof.* Suppose a Markov chain is positive recurrent. Fix a state 0 and consider the Markov chain to start a new cycle each time it transitions into state 0. Think of this as a renewal reward process: suppose you earn 1 each time a transition into state  $j$  occurs (renewal each time you reach state 0 again). Then the average reward per unit time is  $\pi_j$ . But this also equals

$$\frac{\mathbb{E}(N_j)}{\mathbb{E}(N)}$$

where  $N_j$  is the number of transitions into state  $j$  during a cycle and  $N$  is the number of transitions in a cycle, so  $N = \sum_j N_j$ . Let  $N_{ij}$  be the number of transitions from state  $i$  to  $j$  in a cycle, so  $N_j = \sum_{i \in S} N_{ij}$ . Then we have

$$\mathbb{E}(N_j) = \sum_{i \in S} \mathbb{E}(N_{ij})$$

Let  $I_k$  is an indicator variable for transitioning from  $i$  to  $j$  with  $\mathbb{E}(I_k) = P_{ij}$ . Then  $N_{ij} = \sum_{k=1}^{N_i} I_k$ . Note that  $N_i$  is a stopping time for  $I_1, I_2, \dots$  because it depends on what happened before but not on the future. Therefore by Wald's Equation (Theorem 7.3.8),

$$\mathbb{E}(N_{ij}) = \mathbb{E} \sum_{k=1}^{N_i} I_k = \mathbb{E}(I_k) \mathbb{E}(N_i) = \mathbb{E}(N_i) P_{ij}.$$

Using that we have

$$\begin{aligned} \mathbb{E}(N_j) &= \sum_{i \in S} \mathbb{E}(N_{ij}) = \sum_{i \in S} \mathbb{E}(N_i) P_{ij} \\ \implies \frac{\mathbb{E}(N_j)}{\mathbb{E}(N)} &= \sum_{i \in S} \frac{\mathbb{E}(N_i)}{\mathbb{E}(N)} P_{ij} \iff \pi_j = \sum_i \pi_i P_{ij}. \end{aligned}$$

Also note that

$$\sum_j \pi_j = \sum_j \frac{\mathbb{E}(N_j)}{\mathbb{E}(N)} = \frac{\sum_j \mathbb{E}(N_j)}{\sum_j \mathbb{E}(N)} = 1.$$

□

**Example 7.39** (Similar to section 4.1.1 in *Introduction to Probability Models*). Suppose we have a 2 state Markov chain for the weather. 0 means dry, 1 means rain. dry to dry with probability 0.7, rain to rain with probability 0.6. So

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

Then  $\pi_j$  are the unique solutions to

$$\pi_j = \sum_i \pi_i P_{ij}, \quad \sum_j \pi_j = 1.$$

So we have

$$\pi_0 = \pi_0 P_{00} + \pi_1 P_{10}, \quad \pi_1 = \pi_0 P_{01} + \pi_1 P_{11}, \quad \pi_0 + \pi_1 = 1.$$

$$\implies \begin{bmatrix} (1 - P_{00}) & -P_{10} \\ -P_{01} & (1 - P_{11}) \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Solving yields  $\pi_0 = 4/7$ ,  $\pi_1 = 3/7$ .

**Definition 7.37.** A Markov chain is said to be **doubly stochastic** if the sums of the columns of the transition probability matrix also add to 1 (the rows must always add to 1). That is,

$$\sum_i P_{ij} = 1, \quad \forall j \in S$$

**Proposition 7.6.13.** Suppose you have an irreducible Markov chain with  $m < \infty$  states that is doubly stochastic. Then

$$\pi_j = \frac{1}{m}, \quad j = 1, \dots, m$$

That is, all the long-run proportions are the same.

*Proof.* Recall that  $\pi_j$  is the unique solution to  $\pi_j = \sum_i \pi_i P_{ij}$  and that  $\sum_j \pi_j = 1$ . For  $\pi_j = 1/m$ , clearly the second condition is satisfied, so we need to verify the first state. That is we need to verify that

$$\frac{1}{m} = \frac{1}{m} \sum_i P_{ij} \iff \sum_i P_{ij} = 1$$

which is true if the Markov chain is doubly stochastic.

□

**Remark.** In practice if there is a Markov chain with more than three states and you are asked to find the long-run proportions, it is likely a doubly-stochastic Markov chain.

**Example 7.40 (Example 4.25 in *Introduction to Probability Models*).** Hotel, number of new guests every day is Poisson with mean  $\lambda$ . Each guest that stays the night will independently check out the next day with probability  $p = 1 - \alpha$ . (So the number of days each guest stays are independent geometric random variables.)

$$X_n = i \implies X_{n+1} = \text{Bin}(i, \alpha) * \text{Poisson}(\lambda)$$

where  $*$  is being used for a sum to stress the independence of the random variables.

$$P_{ij} = \Pr(\text{Bin}(i, \alpha) + \text{Poisson}(\lambda) = j)$$

⋮

Part (c): Suppose the number of people initially there  $X_0$  is distributed  $\text{Poisson}(\beta)$ . Then the number of that cohort who remains the next day is  $\text{Poisson}(\alpha\beta)$ . So  $X_1 \sim \text{Poisson}(\alpha\beta + \lambda)$ . So to find the limiting distribution (where it's the same every day from now on), we need

$$\beta = \lambda + \alpha\beta \iff \beta = \frac{\lambda}{1 - \alpha} \implies \pi_j = \exp\left(-\frac{\lambda}{1 - \alpha}\right) \left(\frac{\lambda}{1 - \alpha}\right)^j / j!$$

**Proposition 7.6.14 (Proposition 4.4.2 in *Stochastic Processes*; relevant to Gambler's Ruin).** If  $j$  is a recurrent state in a Markov chain, then the set of probabilities  $\{f_{ij}, i \in T\}$  satisfies

$$f_{ij} = \sum_{k \in T} P_{ik} f_{kj} + \sum_{k \in R} P_{ik}, \quad i \in T$$

where  $R$  denotes the set of states communicating with  $j$ .

**Example 7.41 (Gambler's Ruin; Section 4.5.1 of *Introduction to Probability Models*, Example 4.4(A) in *Stochastic Processes* (p. 197 of pdf)).** Also discussed drug testing example immediately afterward in section 4.5.1 of *Introduction to Probability Models*, and what happens as  $N \rightarrow \infty$  (p. 199 of pdf of *Stochastic Processes*).

#### 7.6.4 Branching Processes (Section 4.7 of *Introduction to Probability Models*, Section 4.5 of *Stochastic Processes*)

**Proposition 7.6.15 (Theorem 4.5.1 in *Stochastic Processes* (p. 203 of pdf)).** (i) If  $\mu \leq 1$ , then  $\pi_0 = 1$ . (ii) If  $\mu > 1$ ,  $\pi_0 < 1$ . (iii)  $\pi_0$  is the smallest positive number satisfying

$$\pi_0 = \sum_{j=0}^{\infty} \pi_0^j P_j.$$

*Proof.* (i) Using

$$X_n = \sum_{i=1}^{X_{n-1}} Z_i,$$

we have

$$\mathbb{E}(X_n | X_{n-1}) = X_{n-1} \cdot \sum_{j=0}^{\infty} j P_j = X_{n-1} \mu$$

$$\implies \mathbb{E}(X_n) = \mu \mathbb{E}(X_{n-1}) = \mu^2 \mathbb{E}(X_{n-2}) = \dots = \mu^n \mathbb{E}(X_0) = \mu^n.$$

Suppose  $\mu < 1$ . By Markov's Inequality (Lemma 8.2.1),

$$\Pr(X_n \geq 1) \leq \mathbb{E}(X_n) = \mu^n \implies \lim_{n \rightarrow \infty} \Pr(X_n \geq 1) = 0.$$

When  $\mu = 1$ , can prove using a convexity argument and constraints  $P_0 > 0, P_0 + P_1 < 1$  (see *Stochastic Processes*).

(ii) In *Stochastic Processes*.

(iii) Suppose  $x > 0$  satisfies

$$x = \sum_{j=0}^{\infty} x^j P_j.$$

We will show by induction that  $x \geq \Pr(X_n = 0 \mid X_0 = 1)$  for all  $n$ . To start we need to show that  $x \geq \Pr(X_0 = 0 \mid X_0 = 1)$ . But

$$\Pr(X_0 = 0 \mid X_0 = 1) = P_0 \leq \sum_{j=0}^{\infty} x^j P_j = x$$

so follows. Now we assume  $x \geq \Pr(X_n = 0 \mid X_0 = 1)$ . We would like to show that this implies  $x \geq \Pr(X_{n+1} = 0 \mid X_0 = 1)$ . Note that

$$\Pr(X_{n+1} = 0 \mid X_0 = 1) = \sum_{j=0}^{\infty} \Pr(X_{n+1} = 0 \mid X_1 = j, X_0 = 1) = \sum_{j=0}^{\infty} \Pr(X_{n+1} = 0 \mid X_1 = j) P_j$$

But  $\Pr(X_{n+1} = 0 \mid X_1 = j) = (\Pr(X_{n+1} = 0 \mid X_1 = 1))^j$  for the following reason: if there are  $j$  people to begin with, imagine they each had their own independent branching process. The probability that you reach 0 eventually is equal to the joint probability that all  $j$  processes reach 0 eventually. So we have

$$\sum_{j=0}^{\infty} \Pr(X_{n+1} = 0 \mid X_1 = j) P_j = \sum_{j=0}^{\infty} P_j (\Pr(X_{n+1} = 0 \mid X_1 = 1))^j \leq \sum_{j=0}^{\infty} P_j x^j = x$$

where the second to last step follows by the inductive hypothesis.

□

### 7.6.5 A Markov Chain Model of Algorithmic Efficiency (Section 4.6.1 of *Stochastic Processes*)

**Definition 7.38.** A set of events  $A_2, A_3, \dots$  is said to be an **increasing sequence** if  $A_n \subseteq A_{n+1}$  and a **decreasing sequence** if  $A_n \supseteq A_{n+1}$ . If  $A_n, n \geq 1$  is increasing, define

$$\lim_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} A_n.$$

If  $A_n, n \geq 1$  is decreasing, define

$$\lim_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} A_n.$$

The **limit infimum** is the set of all points contained in all but a finite number of  $A_1, A_2, \dots$ . The **limit supremum** is the set of all points contained in an infinite number of  $A_1, A_2, \dots$ . Note that for an increasing sequence the limit infimum is a subset of the limit supremum.

**Lemma 7.6.16.** For any events  $A_1, A_2, \dots$  there are events  $B_1, B_2, \dots$  that are mutually exclusive ( $B_i \cap B_j = \emptyset, i \neq j$ ) such that

$$B_i \cap B_j = \emptyset, \quad \bigcup_{i=1}^n B_i = (\bigcup_{i=1}^n A_i \text{ for } n = 1, 2, \dots), \quad \bigcup_{i=1}^{\infty} B_i = (\bigcup_{i=1}^{\infty} A_i).$$

*Proof.* Let  $B_1 = A_1$ ,  $B_2 = A_2 \setminus A_1$ ,  $B_3 = A_3 \setminus \{A_1 \cup A_2\}$ ,  $\dots$ ,  $B_n = A_n \setminus \{A_1 \cup \dots \cup A_{n-1}\} = A_n \setminus \left\{ \bigcup_{i=1}^{n-1} A_i \right\}$ .  $\square$

**Theorem 7.6.17.** Probability is a continuous set function. That is, if  $A_n$  is an increasing or decreasing sequence then

$$\lim_{n \rightarrow \infty} \Pr(A_n) = \Pr\left(\bigcup_{n=1}^{\infty} A_n\right) = \Pr\left(\lim_{n \rightarrow \infty} A_n\right).$$

*Proof.* Let  $A_1, A_2, \dots$  be events. Per Lemma 7.6.16, define  $B_1, B_2, \dots$  such that  $B_i \cap B_j = \emptyset$  and  $\bigcup_{i=1}^n B_i = (\bigcup_{i=1}^n A_i \text{ for } n = 1, 2, \dots)$ ,  $\bigcup_{i=1}^{\infty} B_i = (\bigcup_{i=1}^{\infty} A_i)$ . Suppose  $A_n$  is increasing, so that  $A_n \subseteq A_{n+1}$ . Then

$$\Pr\left(\lim_{n \rightarrow \infty} A_n\right) = \Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \Pr\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \Pr(B_i)$$

where the last equality follows since the  $B_i$  are disjoint. Continuing we have

$$= \lim_{n \rightarrow \infty} \sum_{i=1}^n \Pr(B_i) = \lim_{n \rightarrow \infty} \Pr\left(\bigcup_{i=1}^n B_i\right) = \lim_{n \rightarrow \infty} \Pr\left(\bigcup_{i=1}^n A_i\right) = \lim_{n \rightarrow \infty} \Pr(A_n)$$

where the last step follows since  $A_n$  is an increasing sequence.

$\square$

**Remark.** Can prove Boole's Inequality by a similar logic:

$$\Pr\left(\bigcup_{i=1}^n A_i\right) = \Pr\left(\bigcup_{i=1}^n B_i\right) = \sum_{i=1}^n \Pr(B_i) \leq \sum_{i=1}^n \Pr(A_i)$$

where the last step follows since  $B_i \subseteq A_i$ .

**Remark.** Consider the Gambler's Ruin problem (Example 7.41). Let  $A_N$  be the event of reaching 0 before  $N$  given  $X_0 = i$ . Note that  $A_N \subseteq A_{N+1}$ . So we have

$$\lim_{N \rightarrow \infty} \Pr(A_N) = \Pr\left(\lim_{N \rightarrow \infty} A_n\right) = \Pr\left(\bigcup_{N=1}^{\infty} A_n\right)$$

where the last quantity is clearly the probability that the gambler eventually goes to 0, matching our interpretation in Example 7.41.

Similarly, in the proof of Proposition 7.6.15 for branching processes, we claimed

$$\Pr(\text{population dies out}) = \lim_{n \rightarrow \infty} \Pr(X_n = 0).$$

Note that  $\{X_n = 0\} \subset \{X_{n+1} = 0\}$  so

$$\lim_{n \rightarrow \infty} \{X_n = 0\} = \bigcup_{n=1}^{\infty} \{X_n = 0\}$$

again matching what we assumed.

### 7.6.6 Time Reversible Markov Chains (Section 4.8 of *Introduction to Probability Models*, Section 4.7 of *Stochastic Processes*)

**Proposition 7.6.18 (Unlabeled at beginning of each section).** Given a stationary Markov chain  $X_n, X_{n+1}, X_{n+2}, \dots$ , the reverse process  $X_n, X_{n-1}, X_{n-2}, \dots$  is a Markov chain with transition probabilities

$$Q_{ij} = P_{ij}^* = \frac{\pi_j P_{ji}}{\pi_i}.$$

*Proof.*

$$\begin{aligned} Q_{ij} &= \Pr(X_m = j \mid X_{m+1} = i) = \frac{\Pr(X_m = j \cap X_{m+1} = i)}{\Pr(X_{m+1} = i)} = \frac{\Pr(X_m = j) \Pr(X_{m+1} = i \mid X_m = j)}{\Pr(X_{m+1} = i)} \\ &= \frac{\pi_j P_{ji}}{\pi_i}. \end{aligned} \tag{7.11}$$

□

**Definition 7.39 (Time reversible Markov chain).** Suppose for a Markov chain, the reverse transition probabilities  $Q_{ij}$  equal the corresponding forward transition probabilities  $P_{ij}$  for all  $i, j$ . Then the Markov chain is said to be **time reversible**. The condition  $Q_{ij} = P_{ij}$  can also be written (per (7.11)) as

$$\pi_i P_{ij} = \pi_j P_{ji}, \quad \forall i, j.$$

**Theorem 7.6.19 (Theorem 4.7.2 in *Stochastic Processes* (p. 220 of pdf)).** A stationary Markov chain is time reversible if and only if starting in state  $i$ , any path back to  $i$  has the same probability as the reverse path, for all  $i$ . That is, if

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,i} = P_{i,i_k} P_{i_k,i_{k-1}} \cdots P_{i_1,i}$$

for all states  $i, i_1, \dots, i_k$ .

**Theorem 7.6.20 (Theorem 4.2 in *Introduction to Probability Models*).** An ergodic (all states positive recurrent; see Definition 7.35) Markov chain for which  $P_{ij} = 0$  whenever  $P_{ji} = 0$  is time reversible if and only if starting in state  $i$ , any path back to  $i$  has the same probability as the reverse path, for all  $i$ . That is, if

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,i} = P_{i,i_k} P_{i_k,i_{k-1}} \cdots P_{i_1,i}$$

for all states  $i, i_1, \dots, i_k$ .

**Proposition 7.6.21 (Proposition 4.9 in *Introduction to Probability Models*, Theorem 4.7.3 in *Stochastic Processes* (p. 222 of pdf)).** Consider an irreducible Markov chain  $\{X_n\}$  with transition probabilities  $P_{ij}$ . If  $\{X_j, j \in S\}$  is stationary, and there exist positive numbers  $x_i, i \geq 0$  and a transition probability matrix  $Q$  such that

$$x_i P_{ij} = x_j Q_{ji}, \quad i \neq j$$

$$\sum x_i = 1$$

then the Markov chain is reversible,  $\pi_j = x_j$  are the stationary probabilities for both the original and the reversed chain, and the  $Q_{ij}$  are the transition probabilities of the reversed chain.

*Proof.* Assume the above. we have

$$\sum_i x_i P_{ij} = \sum_i x_j P_{ji} = x_j$$

so these  $x_j$  satisfy the stationarity equations and the solution is unique, so  $x_j = \pi_j$ .  $\square$

**Example 7.42 (Similar to Example 4.35 in *Introduction to Probability Models*, Gambler's Ruin-type problem).**

**Example 7.43 (Similar to Example 4.36 in *Introduction to Probability Models*, Proposition 4.7.1 in *Stochastic Processes* (p. 217 of pdf; weighted graph problem)).**

**Example 7.44 (Example 4.37 in *Introduction to Probability Models*, Example 4.7(b) in *Stochastic Processes* (p. 221 of pdf; list, items requested, then move up one position in the list; average position)).** Note the similarity to *Introduction to Probability Models* problem 4.26 (from Homework 10).

### 7.6.7 Semi-Markov Processes (Section 4.8 of *Stochastic Processes* (p. 224 of pdf), Section 7.6 of *Introduction to Probability Models*)

**Definition 7.40 (Semi-Markov Process).** Suppose the stochastic process  $\{X(t), t \geq 0\}$  takes on values  $\{0, 1, \dots, M\}$ . Then  $\{X(t), t \geq 0\}$  is called a **semi-Markov process**.

Why semi-Markov? What happens next doesn't just depend on your current state, it also depends on how long you've been there. But it's semi-Markov because just after a transition you know everything.

## 7.7 ISE 620

**Exercise 14. (“Best Prize” problem.)** There are  $n$  prizes that are presented one at a time in a random order. Each prize has a defined value, and there is a defined ordering of the value of the prizes. Each time you see a prize, you only know the value of that prize relative to the prizes already seen—the value of later prizes remains unknown. Each time a prize is presented, we either accept it or reject it. You only get one prize, so once you accept it, you’re done. Once you reject a prize, you can’t get it back. Your goal is to maximize the probability of accepting the best prize. What do you do (what is the optimal policy)?

**Solution.** The only strategy that makes sense is to accept a prize if it is a *candidate*—that is, the best you’ve seen so far. You should never accept a prize that isn’t a candidate unless you get to the last prize. Let  $P_n$  be the probability of getting the best prize. Then we expect  $P_n$  to approach 0 as  $n$  approaches infinity. Note that if it were good to accept the  $k$ th prize if it were a candidate, then it would definitely be better to accept the  $k + 1$ th candidate if it were a candidate. So a good policy ( $k$ -policy) is to let  $k$  prizes go by, then accept the first candidate to come afterward.

Let  $X$  be the position of the best prize. We will condition on the best prize being in position  $i$ . Then

$$P_k(\text{best}) = \sum_{i=1}^n P_k(\text{best} \mid X = i) \Pr(X = i) = \frac{1}{n} \sum_{i=1}^n P_k(\text{best} \mid X = i)$$

Note that because in order for you to win, none of the prizes between position  $k$  and  $i$  can be candidates (or else you would accept them and not get the best prize). So the best of the first  $i - 1$  prizes must be among the first  $k$  prizes in order for you to get the best prize.

$$P_k(\text{best} \mid X = i) = \begin{cases} 0 & i \leq k \\ \Pr(\{\text{best of } 1, \dots, i-1 \text{ is among the first } k\}) & k > i \end{cases}$$

and

$$\Pr(\{\text{best of } 1, \dots, i-1 \text{ is among the first } k\}) = \frac{k}{i-1}$$

so

$$P_k(\text{best}) = \frac{1}{n} \sum_{i=k+1}^n \frac{k}{i-1} = \frac{k}{n} \sum_{j=k}^{n-1} \frac{1}{j} \approx \frac{k}{n} \int_k^{n-1} \frac{dx}{x} = \frac{k}{n} \log(x) \Big|_k^{n-1} = \frac{k}{n} \log\left(\frac{n-1}{k}\right) \approx \boxed{\frac{k}{n} \log\left(\frac{n}{k}\right)}.$$

Now we want to choose  $k$  that maximizes this quantity. Generalize to letting  $x$  equal any real number in the expression

$$f(x) = \frac{x}{n} \log\left(\frac{n}{x}\right) \implies f'(x) = \frac{x}{n} \frac{x}{n} \left(\frac{-n}{x^2}\right) + \log\left(\frac{n}{x}\right) \frac{1}{n}$$

Setting equal to 0 we have

$$\frac{1}{n} = \frac{1}{n} \log\left(\frac{n}{x}\right) \implies \frac{n}{x} = e \implies \boxed{x = \frac{n}{e}}$$

so the optimal strategy is to let about  $1/e$  of the prizes go by, then choose the first candidate to come by afterward.

**Remark.** The probability of getting the best prize is then

$$f\left(\frac{n}{e}\right) = \frac{1}{e} \log(e) = \frac{1}{e}$$

regardless of  $n!$

**Exercise 15. (Ballot problem.)** Suppose we have candidates  $A$  and  $B$  with votes counted in random order.  $A$  has  $n$  votes,  $B$  has  $m$  with  $n > m$ . What is the probability that  $A$  is always ahead in the count at every stage of the vote?

**Solution.** Let  $\Pr(\{A \text{ is always ahead}\}) = P(n, m) = P_{n,m}$ . Then

$$\begin{aligned} P_{n,m} &= \Pr(\{A \text{ is always ahead}\} \mid \{A \text{ receives the first vote}\}) \cdot \frac{n}{n+m} \\ &\quad + \Pr(\{A \text{ is always ahead}\} \mid \{B \text{ receives the first vote}\}) \cdot \frac{m}{n+m} \\ &= \Pr(\{A \text{ is always ahead}\} \mid \{A \text{ receives the first vote}\}) \cdot \frac{n}{n+m} + 0 \\ &= Q_{n-1,m} \cdot \frac{n}{n+m} \end{aligned}$$

where  $Q$  represents the probability that  $A$  is never behind (since going forward ties would be ok because  $A$  starts out ahead). We have

$$Q_{nm} = \Pr(\{A \text{ is never behind}\}) = \frac{n}{n+m} \Pr(\{A \text{ is never behind}\} \mid \{A \text{ gets first vote}\})$$

Note that  $\Pr(\{A \text{ is never behind}\} \mid \{A \text{ gets first vote}\}) = Q_{n-1,m-1}$ . But now things are more complicated because  $A$  could afford to be behind by 1 going forward. So this strategy is not working. Try instead to condition on who gets the last vote.

$$\begin{aligned} P_{n,m} &= \Pr(\{A \text{ is always ahead}\} \mid \{A \text{ receives the last vote}\}) \cdot \frac{n}{n+m} \\ &\quad + \Pr(\{A \text{ is always ahead}\} \mid \{B \text{ receives the last vote}\}) \cdot \frac{m}{n+m} \end{aligned}$$

$$= P_{n-1,m} \cdot \frac{n}{n+m} + P_{n,m-1} \cdot \frac{m}{n+m}$$

We can work this out recursively using the boundary conditions

$$P_{n,n} = 2, \quad P_{n,0} = 1, \quad n > 0$$

We will try to guess the answer.

$$P_{2,1} = \Pr(\{A \text{ gets the first two votes}\}) = \frac{1}{3}$$

$$P_{n,1} = \Pr(\{\text{first two votes are for } A\}) = \frac{n}{n+1} \cdot \frac{n-1}{n} = \frac{n-1}{n+1}$$

$$P_{3,2} = \frac{3}{5} \cdot \frac{2}{4} \cdot \Pr(\{A \text{ is not the last of the remaining votes}\}) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

$$P_{4,2} = \frac{4}{6} \cdot \frac{3}{5} \cdot \Pr(\{B \text{ does not get next two votes}\}) = \frac{4}{6} \cdot \frac{3}{5} \cdot \left(1 - \frac{2}{4} \cdot \frac{1}{3}\right) = \frac{4}{6} \cdot \frac{3}{5} \cdot \frac{5}{6} = \frac{1}{3}$$

$$P_{4,3} = \frac{4}{7} \cdot \frac{3}{6} \cdot \left( \Pr(\{\text{next is } A\}) \cdot \Pr(\{\text{not 3 } B \text{ votes in a row}\}) + \Pr(\{\text{next is } B\}) \cdot \Pr(\{\text{not 3 } A \text{ votes in a row}\}) \right)$$

$$= \frac{4}{7} \cdot \frac{3}{6} \cdot \left( \frac{2}{5} \cdot \frac{3}{4} + \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} \right)$$

⋮

It seems that

$$P_{n,m} = \frac{n-m}{n+m}$$

We can argue this solution is unique or we can argue it is true by induction.

## 7.8 DSO Statistics Group Screening Exam Problems

**Exercise 16 (2016 DSO Statistics Group In-Class Screening Exam, Question 1).** Let  $\{V_n\}$  be i.i.d. non-negative random variables. Fixing  $r > 0$  and  $q \in (0, 1]$ , consider the sequence  $W_0 = 1$  and  $W_n = (qr + (1-q)V_n)W_{n-1}, n \geq 1$ . A motivating example of  $W_n$  is recording the relative growth of a portfolio where a constant fraction  $q$  of one's wealth is re-invested each year in a risk-less asset that grows by  $r$  per year, with the remainder re-invested in a risky asset whose annual growth factors are random  $V_n$ .

- (a) Show that  $n^{-1} \log W_n \xrightarrow{a.s.} w(q)$ , where  $w(q) = \mathbb{E} \log(qr + (1 - q)V_1)$ .
- (b) Show that  $q \rightarrow w(q)$  is concave on  $(0, 1]$ .
- (c) Using Jensen's Inequality, show that  $w(q) \leq w(q)$  in case  $\mathbb{E}V_1 \leq r$ . Further, show that if  $\mathbb{E}V_1^{-1} \leq r^{-1}$  then the almost sure convergence also applies for  $q = 0$  and  $w(q) \leq w(0)$ .

**Solution.**

- (a) Note that

$$W_1 = (qr + (1 - q)V_1)W_0 = qr + (1 - q)V_1$$

$$W_2 = (qr + (1 - q)V_2)W_1 = (qr + (1 - q)V_2)(qr + (1 - q)V_1)$$

$$W_3 = (qr + (1 - q)V_3)W_2 = (qr + (1 - q)V_3)(qr + (1 - q)V_2)(qr + (1 - q)V_1) = \prod_{i=1}^3 (qr + (1 - q)V_i)$$

⋮

$$W_n = \prod_{i=1}^n (qr + (1 - q)V_i)$$

$$\implies n^{-1} \log(W_n) = \frac{1}{n} \sum_{i=1}^n (qr + (1 - q)V_i)$$

$\{qr + (1 - q)V_i\}$  is a sequence of i.i.d. nonnegative random variables, so by the Strong Law of Large Numbers, the conclusion follows if  $\mathbb{E}[qr + (1 - q)V_1] = \mathbb{E}[qr + (1 - q)V_1] = qr + (1 - q)\mathbb{E}(V_1) < \infty$ .

- (b) Since  $w(q)$  is twice differentiable, a sufficient condition for concavity is  $w''(q) \leq 0$  for all  $q \in (0, 1]$ .

$$w(q) = \mathbb{E} \log[qr + (1 - q)V_1] = \mathbb{E} \log[q(r - V_1) + V_1]$$

$$w'(q) = \frac{\partial}{\partial q} \mathbb{E} \log[q(r - V_1) + V_1] = \mathbb{E} \left( \frac{\partial}{\partial q} \log[q(r - V_1) + V_1] \right) = \mathbb{E} \left( \frac{r - V_1}{q(r - V_1) + V_1} \right) \quad (7.12)$$

$$\begin{aligned} w''(q) &= \mathbb{E} \left( \frac{\partial}{\partial q} \frac{r - V_1}{q(r - V_1) + V_1} \right) = \mathbb{E} \left( (r - V_1) \cdot \frac{\partial}{\partial q} [q(r - V_1) + V_1]^{-1} \right) \\ &= \mathbb{E} \left( (r - V_1) \cdot (-1) [q(r - V_1) + V_1]^{-2} \cdot (r - V_1) \right) = -\mathbb{E} \left( \left[ \frac{(r - V_1)}{q(r - V_1) + V_1} \right]^2 \right) \end{aligned} \quad (7.13)$$

This is  $-1$  times the expectation of a nonnegative random variable, so by Markov's Inequality we have

$$\mathbb{E} \left( \left[ \frac{(r - V_1)}{q(r - V_1) + V_1} \right]^2 \right) \geq 0 \iff w''(q) \leq 0 \quad \forall q \in (0, 1],$$

proving concavity.

(c) By Jensen's Inequality and the concavity of  $q \rightarrow \log[qr + (1 - q)V_1]$ , we have

$$w(q) = \mathbb{E} \log[qr + (1 - q)V_1] \leq \log \mathbb{E}[qr + (1 - q)V_1] = \log(qr + (1 - q)\mathbb{E}[V_1])$$

$$\leq \log(qr + (1 - q)r) = \log(r) = \mathbb{E} \log(r) = w(1).$$

where the third step used  $\mathbb{E}(V_1) \leq r$  and the second-to-last step used the fact that  $r$  is non-random.  
For  $q = 0$ , we have

$$n^{-1} \log(W_n) = \frac{1}{n} \sum_{i=1}^n V_i.$$

$\{V_i\}$  is a sequence of i.i.d. nonnegative random variables, so by the Strong Law of Large Numbers, almost sure convergence applies if  $\mathbb{E}[|V_1|] = \mathbb{E}[V_1] < \infty$ . **But this is the same condition as in the case  $q \in (0, 1]$ .**

To show  $w(q) \leq w(0)$ , we will show that  $w(q)$  is nonincreasing on  $[0, 1]$ . Recall from (7.12)

$$w'(q) = \mathbb{E} \left( \frac{r - V_1}{q(r - V_1) + V_1} \right)$$

We will show that (7.12) is upper-bounded by 0 on  $[0, 1]$ . Plugging in  $q = 0$  yields

$$w'(0) = \mathbb{E} \left( \frac{r - V_1}{V_1} \right) = \mathbb{E}(rV_1^{-1} - 1) = r\mathbb{E}(V_1^{-1}) - 1 \leq 1 - 1 = 0.$$

where we used the assumption  $\mathbb{E}(V_1^{-1}) \leq r$  on the second-to-last step. Recall that the second derivative (7.13) is continuous and nonpositive on  $(0, 1]$ . Therefore the first derivative (7.12) never exceeds  $q(0) = 0$  on  $[0, 1]$ , so  $w(q) \leq w(0)$ .

## 7.9 Simple Random Walk

**Definition 7.41.** Let  $\{X_i\}$  be i.i.d. We have

$$X_i = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } 1 - p \end{cases}$$

Let  $S_k = \sum_{i=1}^k X_i$ . Then  $\{S_k\}$  is a **simple random walk**. (It is simple because the outcomes are either 1 or -1.)

## 7.10 Martingales

**Definition.** Let  $\{y_t\}_{t=0}^\infty$  be a sequence of random variables, and let  $\Omega_t$  denote the information set available at date  $t$ , which at least contains  $\{y_t, y_{t-1}, y_{t-2}, \dots\}$ . If  $\mathbb{E}(y_t \mid \Omega_{t-1}) = y_{t-1}$  holds then  $\{y_t\}$  is a martingale process with respect to  $\Omega_t$ .

**Definition.** Let  $\{y_t\}_{t=1}^{\infty}$  be a sequence of random variables, and let  $\Omega_t$  denote the information set available at date  $t$ , which at least contains  $\{y_t, y_{t-1}, y_{t-2}, \dots\}$ . If  $\mathbb{E}(y_t | \Omega_{t-1}) = 0$ , then  $\{y_t\}$  is a martingale difference process with respect to  $\Omega_t$ .

## 7.11 Brownian Motion

**Appendix B.13, Brownian motion.** A standard Brownian motion  $b(\cdot)$  is a continuous-time stochastic process associating each date  $a \in [0, 1]$  with the scalar  $b(a)$  such that

- (i)  $b(0) = 0$
- (ii) For any dates  $0 \leq a_1 \leq a_2 \leq \dots \leq a_k \leq 1$  the changes  $[b(a_2) - b(a_1)], [b(a_3) - b(a_2)], \dots, [b(a_k) - b(a_{k-1})]$  are independent multivariate Gaussian with  $b(a) - b(s) \sim \mathcal{N}(0, a - s)$ .
- (iii) For any given realization,  $b(a)$  is continuous in  $a$  with probability 1.

Other continuous time processes can be generated from the standard Brownian motion. For example, a Brownian motion with variance  $\sigma^2$  can be obtained as

$$w(a) = \sigma b(a)$$

where  $b(a)$  is a standard Brownian motion.

The continuous time process

$$\mathbf{w}(a) = \boldsymbol{\Sigma}^{1/2} \mathbf{b}(a)$$

is a Brownian motion with covariance matrix  $\boldsymbol{\Sigma}$ .

**Definition 26 (Wiener process).** Let  $\Delta w(t)$  be the change in  $w(t)$  during the time interval  $dt$ . Then  $w(t)$  is said to follow a Wiener process if

$$\Delta w(t) = \epsilon_t \sqrt{dt}, \quad \epsilon_t \sim IID(0, 1)$$

and  $w(t)$  denotes the value of the  $w(\cdot)$  at date  $t$ . Clearly,

$$\mathbb{E}[\Delta w(t)] = 0, \text{ and } \text{Var}[\Delta w(t)] = dt$$

**Theorem 7.11.1. Donsker's Theorem, Theorem 43, p.335, Section 15.6.3.** Let  $a \in [0, 1)$ ,  $t \in [0, T]$ , and suppose  $(J-1)/T \leq a < J/T$ ,  $J = 1, 2, \dots, T$ . Define

$$R_T(a) = \frac{1}{\sqrt{T}} s_{[T^a]}$$

where

$$s_{[Ta]} = \epsilon_1 + \epsilon_2 + \dots + \epsilon_{[Ta]}$$

$[Ta]$  denotes the largest integer part of  $Ta$  and  $s_{[Ta]} = 0$  if  $[Ta] = 0$ . Then  $R_T(a)$  weakly converges to  $w(a)$ , i.e.,

$$R_T(a) \rightarrow w(a)$$

where  $w(a)$  is a Wiener process. Note that when  $a = 1$ ,  $R_T(1) = 1/\sqrt{T} \cdot S_{[T]} = 1/\sqrt{T} \cdot (\epsilon_1 + \epsilon_2 + \dots + \epsilon_T)$ . Since  $\epsilon_t$ 's are IID, by the central limit theorem,  $R_T(1) \rightarrow \mathcal{N}(0, 1)$ .

Similar (Theorem 2.1 in [Phillips and Durlauf \[1986\]](#)): Let  $\{u_t\}$  be a sequence satisfying  $\mathbb{E}(u_t) = 0$ ,  $\gamma(0) = \mathbb{E}(T^{-1}S_t^2) \rightarrow \sigma^2 < \infty$  as  $T \rightarrow \infty$ ,  $\{u_t\}$  is square summable,  $\sup_t \{\mathbb{E}(|u_t|^\beta)\} < \infty$  for some  $2 \leq \beta < \infty$  and all  $t$ ,  $\gamma(h) = \mathbb{E}(T^{-1}(y_t - y_{t-h})^2) \rightarrow K_h < \infty$  as  $\min\{h, T\} \rightarrow \infty$ . Then  $X_T(t) \Rightarrow W(t)$  as  $T \rightarrow \infty$ , where  $W(t)$  is a Wiener process.

**Theorem 7.11.2. Continuous Mapping Theorem (Theorem 44 of Pesaran in 15.6.3).** Let  $a \in [0, 1)$ ,  $i \in [0, n]$ , and suppose  $(J-1)/n \leq a < J/n$ ,  $J = 1, 2, \dots, n$ . Define  $R_n(a) = n^{-1/2}S_{[n \cdot a]}$ . If  $f(\cdot)$  is continuous over  $[0, 1)$ , then

$$f[R_n(a)] \xrightarrow{d} f[w(a)]$$

## 8 Asymptotics and Convergence

These notes are based on my notes from chapter 8 of *Time Series and Panel Data Econometrics* (1st edition) by M. Hashem Pesaran [Pesaran, 2015] and coursework for Economics 613: Economic and Financial Time Series I at USC, as well as Math 505A and Math 541A at USC and chapter 7 from *Probability and Random Processes* (Grimmett and Stirzaker) 3rd edition [Grimmett and Stirzaker, 2001].

### 8.1 Preliminaries (5.9 and 7.1, Grimmett and Stirzaker)

**Definition 8.1. Definition 7.1.4, Grimmett and Stirzaker.** If for all  $x \in [0, 1]$  the sequence  $\{f_n(x)\}$  of real numbers satisfies  $f_n(x) \rightarrow f(x)$  as  $n \rightarrow \infty$  then we say  $f_n \rightarrow f$  **pointwise**.

**Remark.** In practice pointwise convergence is often not useful for functions because a sequence of functions may be continuous while its limit is not. For instance, consider  $\{f_n : f_n = x^n \forall x \in [0, 1]\}$ . Then  $f_n$  is continuous for all  $n$  but

$$\lim_{n \rightarrow \infty} f_n = \begin{cases} 0 & x \leq 1 \\ 1 & x = 1 \end{cases}$$

Instead, the following definition is often more useful.

**Definition 8.2. (from class notes.)** We say that  $f_n$  **uniformly converges to  $f$  on  $[a, b]$**  if for every  $\epsilon > 0$  there exists  $N$  such that for every  $n > N$ ,

$$\forall x \in [a, b] \quad |f_n(x) - f(x)| < \epsilon$$

**Definition 8.3. (Definition 7.1.5, Grimmett and Stirzaker.)** Let  $V$  be a collection of functions mapping  $[0, 1]$  into  $\mathbb{R}$  and assume  $V$  is endowed with a function  $\|\cdot\| : V \rightarrow \mathbb{R}$  satisfying

- (a)  $\|f\| \geq 0$  for all  $f \in V$
- (b)  $\|f\| = 0$  if and only if  $f$  is the zero function (or equivalent to it)
- (c)  $\|af\| = |a| \cdot \|f\|$  for all  $a \in \mathbb{R}$ ,  $f \in V$
- (d)  $\|f + g\| \leq \|f\| + \|g\|$  (Triangle Inequality)

The function  $\|\cdot\|$  is called a **norm**. If  $\{f_n\}$  is a sequence of members of  $V$  then we say that  $f_n \rightarrow f$  **with respect to the norm  $\|\cdot\|$**  if  $\|f_n - f\| \rightarrow 0$  as  $n \rightarrow \infty$ .

**Definition 8.4. (Definition 7.16, Grimmett and Stirzaker.)** Let  $\epsilon > 0$  be prescribed, and define the distance between two functions  $g, h : [0, 1] \rightarrow \mathbb{R}$  by

$$d_\epsilon(g, h) = \int_E dx$$

where  $E = \{u \in [0, 1] : |g(u) - h(u)| > \epsilon\}$ . We say that  $f_n \rightarrow f$  **in measure** if

$$d_\epsilon(f_n, f) \rightarrow 0 \text{ as } n \rightarrow \infty \text{ for all } \epsilon > 0$$

**Theorem 8.1.1. Inversion Theorem (Theorem 5.9.2, Grimmett and Stirzaker).** Let  $X$  have distribution function  $F$  and characteristic function  $\phi$ . Define  $\bar{F} : \mathbb{R} \rightarrow [0, 1]$  by

$$\bar{F}(x) = \frac{1}{2} [F(x) + \lim_{y \rightarrow x^-} F(y)]$$

Then

$$\bar{F}(b) - \bar{F}(a) = \lim_{N \rightarrow \infty} \int_{-N}^N \frac{\exp(-iat) - \exp(-ibt)}{2\pi i t} \cdot \phi(t) dt$$

*Proof.* See Kingman and Taylor [1966]. □

**Corollary 8.1.1.1. Corollary 5.9.3.** Random variables  $X$  and  $Y$  have the same characteristic function if and only if they have the same distribution function.

*Proof.* Available in Grimmett and Stirzaker section 5.9, pp. 189 - 190. □

**Definition 8.5. (Definition 5.9.4, Grimmett and Stirzaker.)** We say that the sequence  $F_1, F_2, \dots$  of distribution functions **converges** to the distribution function  $F$  (written  $F_n \rightarrow F$ ) if  $F(x) = \lim_{n \rightarrow \infty} F_n(x)$  at each point  $x$  where  $F$  is continuous.

**Theorem 8.1.2. Continuity theorem (Thereom 5.9.5; in notes from Friday 10/26, Lecture 28).** Suppose that  $F_1, F_2, \dots$  is a sequence of distribution functions with corresponding characteristic functions  $\phi_1, \phi_2, \dots$

- (a) If  $F_n(x) \rightarrow F(x)$  for some distribution function  $F$  with characteristic function  $\phi$  (at  $x$  where  $F$  is continuous), then  $\phi_n(t) \rightarrow \phi(t)$  for all  $t$ .
- (b) Conversely, if  $\phi(t) = \lim_{n \rightarrow \infty} \phi_n(t)$  exists and  $\phi(t)$  is continuous at  $t = 0$ , then  $\phi$  is the characteristic function of some distribution function  $F$ , and  $F_n \rightarrow F$ .

*Proof.* See Kingman and Taylor [1966]. □

## 8.2 Inequalities (8.6 of Pesaran)

### Inequalities

- Probabilities

—

**Lemma 8.2.1. Markov's Inequality (Grimmett and Stirzaker p. 311, 319) :** Let  $X : \Omega \rightarrow [-\infty, \infty]$  be a random variable. Then for all  $a > 0$ ,

$$\Pr(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}$$

*Proof.* Note  $t \cdot \mathbf{1}_{\{|X| \geq t\}} \leq |X|$ , where  $\mathbf{1}$  is the indicator function. Dividing both sides by  $t$  and taking expectations, we have

$$\mathbb{E}(\mathbf{1}_{\{|X| \geq t\}}) \leq \frac{\mathbb{E}|X|}{t} \iff \Pr(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t}, \quad \forall t > 0.$$

□

**Corollary 8.2.1.1.** If  $n$  is a positive integer, then

$$\Pr(|X| \geq t) \leq \frac{\mathbb{E}(|X|^n)}{t^n} \quad \forall t > 0$$

*Proof.* By Markov's Inequality (Theorem 8.2.1),

$$\Pr(|X| \geq t) = \Pr(|X|^n \geq t^n) \leq \frac{\mathbb{E}(|X|^n)}{t^n}$$

□

—

**Theorem 8.2.2. Chebyshev's Inequality:** (probability p. 319) Let  $X : \Omega \rightarrow [-\infty, \infty]$  be an (integrable) random variable with  $\mathbb{E}(X^2) < \infty$ . Then for any real number  $k > 0$

$$\Pr(|X - \mathbb{E}(X)| \geq k\sqrt{\text{Var}(X)}) \leq \frac{1}{k^2}$$

This can also be written as

$$\Pr(|X - \mathbb{E}(X)| \geq k) \leq \frac{\text{Var}(X)}{k^2}$$

(Can be used to demonstrate consistency of estimators: if we can show that as  $T \rightarrow \infty$   $\text{Var}(X) = \sigma^2 \rightarrow 0$ , then this implies  $\Pr(|X - \mu| \geq k\sigma) \rightarrow 0$  as  $T \rightarrow \infty$ , showing consistency.)

**Theorem 8.2.3. Chernoff** For  $x \geq 0$ ,  $a > 0$ ,  $\forall t > 0$ ,

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}}$$

### • Moments

—

**Theorem 8.2.4 ( Cauchy-Schwarz (and Bunyakovsky)).** If  $X$  and  $Y$  are random variables with finite variance then

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Note that this can be a corollary of Theorem 8.2.7 with  $p = q = 2$ . We can also prove this theorem on its own in a different one. We first prove a useful result.

**Lemma 8.2.5.** If  $\text{Var}(X) = 0$  then  $X$  is almost surely constant; that is,  $\Pr(X = a) = 1$  for some  $a \in \mathbb{R}$ .

*Proof.* Note that because  $\text{Var}(X) = 0 < \infty$ , we know that  $\mathbb{E}(X)$  and  $\mathbb{E}(X^2)$  exist. We have

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = 0$$

Let  $Y = (X - \mathbb{E}(X))^2$ . Note that  $Y = (X - \mathbb{E}(X))^2 \geq 0$  and that  $\mathbb{E}(Y) = \text{Var}(X) = 0$ . Therefore  $\Pr(Y = 0) = 1$ , so  $\Pr(Y \neq 0) = 0$ . To see why, in the case that  $X$  is discrete,

$$\mathbb{E}(Y) = \sum_{k=0}^{\infty} k \cdot \Pr(Y = k) = \text{Var}(X) = 0$$

which is true if and only if  $\Pr(Y = k) = 0$  for all  $k > 0$ . Since we already showed that  $\Pr(Y < 0) = 0$ , it follows that  $\Pr(Y = 0) = 1$ . In the continuous case,

$$\mathbb{E}(Y) = \int_0^{\infty} y \cdot f_Y(y) dy = \text{Var}(X) = 0$$

which implies that  $f_Y(x) = 0$  for all  $x > 0$ . Again, since  $\Pr(Y < 0) = 0$ , we have  $\Pr(Y \neq 0) = 0$ . But  $Y = 0 \iff X = \mathbb{E}(X)$  so we have  $\Pr(X = \mathbb{E}(X)) = 1$ .  $\square$

**Remark.** Note that Lemma 8.2.5 along with Proposition 6.1.29 imply that  $X$  has variance 0 if and only if it is (almost surely) constant.

We are now ready to prove the Cauchy-Schwarz Inequality.

*Proof.* if  $\mathbb{E}(X^2) = 0$  or  $\mathbb{E}(Y^2) = 0$ , the Cauchy-Schwarz Inequality follows immediately. To see why, suppose without loss of generality that  $\mathbb{E}(X^2) = 0$ . Then the right side is 0. Also,  $0 \leq \text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = -\mathbb{E}(X)^2$ . Since  $\mathbb{E}(X)^2 \geq 0$ , we must have  $\mathbb{E}(X)^2 = 0$  and therefore  $\text{Var}(X) = 0$ . Therefore by Lemma 8.2.5,  $X$  is almost surely constant, which means that  $\text{Cov}(X, Y) = 0$ .

In the case that  $\mathbb{E}(X^2) > 0$  and  $\mathbb{E}(Y^2) > 0$ , for  $a, b \in \mathbb{R}$ , let  $Z = aX - bY$ . Then

$$0 \leq \mathbb{E}(Z^2) = a^2 \mathbb{E}(X^2) - 2ab \mathbb{E}(XY) + b^2 \mathbb{E}(Y^2) \quad (8.1)$$

The right side of (8.1) is quadratic in  $a$ . Because it is greater than or equal to zero, it has at most one real root, which means its discriminant must be non-positive. That is, if  $b \neq 0$ ,

$$(-2b \mathbb{E}(XY))^2 - 4b^2 \mathbb{E}(X^2) \mathbb{E}(Y^2) \leq 0 \iff \mathbb{E}(XY)^2 - \mathbb{E}(X^2) \mathbb{E}(Y^2) \leq 0$$

which yields the result. Note that equality holds if and only if  $\Pr(aX = bY) = 1$  because the discriminant is zero if and only if the quadratic has a real root, which occurs if and only if

$$\mathbb{E}[(aX - bY)^2] = 0$$

which is true if and only if  $\Pr(aX = bY) = 1$  by Lemma 8.2.5 and Proposition 6.1.29.  $\square$

– Krylov

–

**Definition 8.6.** Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . We say that  $\phi$  is **convex** if for any  $x, y \in \mathbb{R}$  and for any  $t \in [0, 1]$ , we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y)$$

**Theorem 8.2.6 (Jensen's Inequality, from Math 541A. Also Grimmett and Stirzaker p.181, 349).** Let  $X : \Omega \rightarrow [-\infty, \infty]$  be a random variable. Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be convex. If  $\mathbb{E}|X| < \infty$  and  $\mathbb{E}|\phi(X)| < \infty$ , then

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

(See also Theorem 9.1.4.)

For the definition of convexity, see Definition 9.1.

*Proof.* Note that from Theorem 9.1.2, for any  $y \in \mathbb{R}$  there exists a constant  $a$  and a function  $L$  such that

$$a(x - y) + \phi(y) \leq \phi(x) \quad \forall x \in \mathbb{R}$$

Letting  $y = \mathbb{E}(X)$  we have

$$a(X - \mathbb{E}X) + \phi(\mathbb{E}X) \leq \phi(X)$$

Since expectations preserve inequalities,

$$\mathbb{E}[a(X - \mathbb{E}X) + \phi(\mathbb{E}X)] \leq \mathbb{E}\phi(X)$$

But

$$\mathbb{E}[a(X - \mathbb{E}X) + \phi(\mathbb{E}X)] = a(\mathbb{E}X - \mathbb{E}X) + \mathbb{E}(\phi(\mathbb{E}X)) = \phi(\mathbb{E}X)$$

which yields

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

□

For some corollaries, see section 9.1.

---

**Theorem 8.2.7. [Hölder (Grimmett and Stirzaker p. p. 143, 319; Theorem 1.99 in Math 541A lecture notes) Generalization of Cauchy-Schwarz]** Let  $X, Y : \Omega \rightarrow \mathbb{R}$  be random variables. For  $p, q \geq 1$  satisfying  $1/p + 1/q = 1$  we have

$$\mathbb{E}(|XY|) \leq (\mathbb{E}(|X^p|))^{1/p}(\mathbb{E}(|X^q|))^{1/q} = \|X\|_p\|Y\|_q.$$

The equality case happens only if  $X$  is a constant multiple of  $Y$  with probability 1. Note that the case  $p = q = 2$  recovers the Cauchy-Schwarz Inequality (Theorem 8.2.4).

*Proof.* Assume without loss of generality that  $\|X\|_p = \|Y\|_q = 1$ . Also, the case  $p = 1, q = \infty$  follows from the triangle inequality, so we assume  $1 < p < \infty$ . From concavity of the log function, we have

$$\log((x^p)^{1/p}(y^q)^{1/q}) = (1/p)\log(x^p) + (1/q)\log(y^q)$$

$$\leq \log\left(\frac{1}{p}x^p + \frac{1}{q}y^q\right)$$

$$\implies (x^p)^{1/p}(y^q)^{1/q} \leq \frac{1}{p}x^p + \frac{1}{q}y^q$$

Fixing an  $\omega \in \Omega$ , we have

$$|X(\omega)Y(\omega)| = (|X(\omega)|^p)^{1/p}(|Y(\omega)|^q)^{1/q} \leq \frac{1}{p}|X(\omega)|^p + \frac{1}{q}|Y(\omega)|^q$$

Integrating we have...

□

**Theorem 8.2.8 (Hölder (vector form)).** For any  $u, v \in \mathbb{R}^n$ ,

$$|u^T v| \leq \|u\|_p \|v\|_q$$

for any  $p, q \in [0, \infty]$  satisfying  $1/p + 1/q = 1$ .

—

**Theorem 8.2.9. Minkowski** (Grimmett and Stirzaker p. p. 143) For  $p \geq 1$ ,

$$[\mathbb{E}(|X + Y|^p)]^{1/p} \leq (\mathbb{E}|X^p|)^{1/p} + (\mathbb{E}|Y^p|)^{1/p}$$

- Useful for showing lower order moments are finite (e.g. finite variance implies finite mean).

**Lemma 8.2.10. Lyapunov's Inequality (Grimmett and Stirzaker p. 143).** For  $0 < r \leq s < \infty$ ,

$$\mathbb{E}(|X|^r)^{1/r} \leq \mathbb{E}(|X|^s)^{1/s}$$

—

**Theorem 8.2.11. Triangle Inequality:** Let  $X, Y : \Omega \rightarrow \mathbb{R}$  be random variables. Let  $1 \leq p \leq \infty$ . Then

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p, 1 \leq p \leq \infty$$

*Proof.* The case  $p = \infty$  follows from the scalar triangle inequality, so assume  $1 \leq p < \infty$ . By scaling, we may assume  $\|X\|_p = 1 - t$ ,  $\|Y\|_p = t$ , for some  $t \in (0, 1)$  (zeroes and infinities being trivial). Define  $V := X/(1 - t)$ ,  $W := Y/t$ . Then by convexity of  $x \mapsto |x|^p$  on  $\mathbb{R}$ ,

$$|(1 - t)V(\omega) + t(W(\omega))|^p \leq (1 - t)|V(\omega)|^p + t|W(\omega)|^p$$

Take expectation of both sides:

$$\mathbb{E}|X + Y|^p \leq (1 - t)^{1-p}\mathbb{E}(|X|^p) + t^{1-p}\mathbb{E}(|Y|^p)$$

Since  $\|X\|_p = t$ ,  $\|Y\|_p = 1 - t$ , we have that the right side is  $(1 - t)^{1-p}t^p + t^{1-p}(1 - t)^p = 1$ . (Note:  $\|Y\|_p = t$ ,  $\mathbb{E}|Y|^p = t^p$ ,  $\|X\|_p = 1 - t$  Therefore

$$(\mathbb{E}|X + Y|^p)^{1/p} = \|X + Y\|_p \leq 1$$

□

**Remark.** See also Theorem 5.1.6 and Corollary 5.1.6.1.

**Theorem 8.2.12 (Chernoff Bound).** Let  $X$  be a random variable and let  $r > 0$ . Define  $M_X(t) := \mathbb{E}e^{tX}$  for any  $t \in \mathbb{R}$ . Then for any  $t > 0$ ,

$$\mathbb{P}(X > r) \leq e^{-tr} M_X(t).$$

*Proof.* Using Markov's Inequality (Theorem 8.2.1) on  $e^{tX}$ , we have

$$\Pr(X \geq r) = \Pr(e^{tX} \geq e^{tr}) \leq \frac{\mathbb{E}e^{tX}}{e^{tr}} = e^{-tr} M_X(t), \quad \forall t > 0.$$

□

**Remark.** Consequently, if  $X_1, \dots, X_n$  are independent random variables with the same CDF, and if  $r, t > 0$ ,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i > r\right) \leq e^{-trn} (M_{X_1}(t))^n.$$

For example, if  $X_1, \dots, X_n$  are independent Bernoulli random variables with parameter  $0 < p < 1$ , and if  $r, t > 0$ ,

$$\mathbb{P}\left(\frac{X_1 + \dots + X_n}{n} - p > r\right) \leq e^{-trn} (e^{-tp}[pe^t + (1-p)])^n.$$

And if we choose  $t$  appropriately, then the quantity  $\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - p) > r\right)$  becomes exponentially small as either  $n$  or  $r$  become large. That is,  $\frac{1}{n} \sum_{i=1}^n X_i$  becomes very close to its mean. Importantly, the Chernoff bound is much stronger than either Markov's or Chebyshev's inequality, since they only respectively imply that

$$\mathbb{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - p\right| > r\right) \leq \frac{2p(1-p)}{r}, \quad \mathbb{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - p\right| > r\right) \leq \frac{p(1-p)}{nr^2}.$$

**Monotone convergence theorem.**

**Dominated Convergence Theorem** (Theorem 6.4.2).

### 8.3 Modes of Convergence (7.2 of Grimmett and Stirzaker, 8.2 and 8.4 of Pesaran)

Let  $\{X_n\} = \{X_1, X_2, \dots\}$  and  $X$  be random variables defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

**Definition 8.7. Convergence in probability.**  $\{X_n\}$  is said to **converge in probability** to  $X$  if

- Grimmett and Stirzaker definition:

$$\lim_{n \rightarrow \infty} \Pr(|X_n - X| > \epsilon) = 0, \text{ for every } \epsilon > 0$$

- Pesaran definition:

$$\lim_{n \rightarrow \infty} \Pr(|X_n - X| < \epsilon) = 1, \text{ for every } \epsilon > 0$$

- More formal (from Math 541A):

$$\forall \epsilon > 0, \lim_{n \rightarrow \infty} \Pr(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0$$

**Remark.** This mode of convergence is also often denoted by  $X_n \xrightarrow{p} X$  and when  $X$  is a fixed constant it is referred to as the **probability limit of  $X_n$** , written as  $\text{Plim}(X_n) = x$ , as  $n \rightarrow \infty$ .

The above concept is readily extended to multivariate cases where  $\{\mathbf{X}_n, n = 1, 2, \dots\}$  denote  $m$ -dimensional vectors of random variables. Then the condition is

$$\lim_{n \rightarrow \infty} \Pr(\|\mathbf{X}_n - \mathbf{X}\| < \epsilon) = 1, \text{ for every } \epsilon > 0$$

where  $\|\cdot\|$  denotes an appropriate norm (say  $\ell_2$ ). Convergence in probability is often referred to as "weak convergence" (in contrast to convergence with probability 1, below).

**Definition 8.8. Convergence with probability 1 or almost surely.** The sequence of random variables  $\{X_n\}$  is said to **converge with probability 1** (or **almost surely**) to  $X$  if

- (505A class notes definition)

$$\Pr(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}) = 1$$

(Note: pointwise convergence can hardly ever be shown here and is not useful.)

- Grimmett and Stirzaker textbook definition:

$$\Pr(\{\omega \in \Omega : X_n(\omega) \rightarrow X(\omega) \text{ as } n \rightarrow \infty\}) = 1$$

- Pesaran textbook definition:

$$\Pr\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1$$

**Remark.** This is often written as  $X_n \xrightarrow{w.p.1} X$  or  $X_n \xrightarrow{a.s.} X$ . An equivalent condition for convergence with probability 1 is given by

$$\lim_{n \rightarrow \infty} \Pr(|X_m - X| < \epsilon, \text{ for all } m \geq n) = 1, \text{ for every } \epsilon > 0$$

which shows that convergence in probability is a special case of convergence with probability 1 (obtained by setting  $m = n$ ). Convergence with probability 1 is stronger than convergence in probability and is often referred to as "strong convergence."

**Definition 8.9. Convergence in  $r$ -th mean or convergence in  $\ell_p$ .**  $X_n \rightarrow X$  in  $r$ th mean (or in  $\ell_p$ ) where  $r \geq 1$  (or  $0 < p \leq \infty$ ) if  $\mathbb{E}|X_n|^r < \infty$  for all  $n$  and

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n - X|^r) = 0$$

or if  $\|X\|_p < \infty$  and

$$\lim_{n \rightarrow \infty} \|X_n - X\|_p = 0$$

**Remark.** Recall that  $\|X\|_p := (\mathbb{E}(X)^p)^{1/p}$  if  $0 < p < \infty$  and  $\|X\|_\infty := \inf\{c > 0 : \Pr(|X| \leq c) = 1\}$ . Note that if  $p < 1$ ,  $\|\cdot\|_p$  is no longer a norm because it does not satisfy the Triangle Inequality (Corollary ?? and Theorem 8.2.11), but this property still holds. Convergence in  $r$ th mean is often written  $X_n \xrightarrow{r} X$ .

**Definition 8.10. Convergence in Distribution.** Let  $X_1, X_2, \dots$  have distribution functions  $F_1(\cdot), F_2(\cdot), \dots$  respectively. Then  $X_n$  is said to **converge in distribution to  $X$**  if

$$\lim_{n \rightarrow \infty} \Pr(X_n \leq u) = \Pr(X \leq u)$$

for all  $u$  at which  $F_X(x) = \Pr(X \leq x)$  is continuous. This can also be written

$$\lim_{n \rightarrow \infty} F_n(u) = F(u)$$

for all  $u$  at which  $F$  is continuous.

**Remark.** Convergence in distribution is usually denoted by  $X_n \xrightarrow{d} X$ ,  $X_n \xrightarrow{L} X$ , or  $F_n \implies F$ . By the Continuity Theorem (Theorem 8.1.2, section 8.1), this is equivalent to

$$\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \phi_X(t), \quad t \in \mathbb{R}.$$

Note that the random variables are allowed to have different domains.

**Definition 8.11. (Convergence in distribution for vector-valued random variables.)** We say that random variables  $Y^{(1)}, \dots, : \Omega \rightarrow \mathbb{R}^d$  **converge in distribution** to  $Y : \Omega \rightarrow \mathbb{R}^d$  if for all  $v \in \mathbb{R}^d$ ,  $\langle v, Y^{(1)} \rangle, \langle v, Y^{(2)} \rangle, \dots$  converges in distribution to  $\langle v, Y \rangle$ .

**Theorem 8.3.1 ((Theorem 7.2.3, Grimmett and Stirzaker.)).** The following implications hold:

- (a)  $(X_n \xrightarrow{a.s.} X) \implies (X_n \xrightarrow{p} X)$
- (b)  $(X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{p} X)$  for any  $r \geq 1$
- (c)  $(X_n \xrightarrow{p} X) \implies (X_n \xrightarrow{d} X)$

Also, if  $r > s \geq 1$ , then  $(X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{s} X)$ . No other implications hold in general.

*Proof.* (a) By Markov's Inequality (Lemma 8.2.1),

$$\Pr(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}|X_n - X|}{\epsilon} \quad \text{for all } \epsilon > 0$$

Therefore if  $X_n \xrightarrow{1} X$ ; that is,  $\lim_{n \rightarrow \infty} \mathbb{E}(|X_n - X|) = 0$ , then  $\lim_{n \rightarrow \infty} \Pr(|X_n - X| > \epsilon) = 0$  for every  $\epsilon > 0$ , so  $X_n \xrightarrow{p} X$ .

(b) Let  $\epsilon > 0$  and let  $0 < r < \infty$ . From Markov's Inequality,

$$\Pr(|X_n - X| > \epsilon) = \Pr(|X_n - X|^r > \epsilon^r) \leq \frac{\mathbb{E}|X_n - X|^r}{\epsilon^r}$$

The right side converges to zero by assumption. Therefore  $X_1, X_2, \dots$  converges to  $X$  in probability. The case  $r = \infty$  follows from any case  $r < \infty$ , since for example

$$(|\mathbb{E}(X_n) - X|)^2 \leq \mathbb{E}(|X_n - X|^2) \leq \|X_n - X\|_\infty$$

which shows that  $L_\infty$  convergence implies  $L_2$  convergence. So since  $L_2$  convergence implies convergence in probability,  $L_\infty$  convergence does too. We can also see this by simply examining the definition of the  $L_\infty$  norm:

$$\|X_n - X\|_\infty := \inf\{c > 0 : \Pr(|X_n - X| \leq c) = 1\}.$$

Clearly if this number goes to 0, then  $X_n \xrightarrow{p} X$ .

(c)

□

**Remark.** Here are counterexamples showing that the converses are not always true:

(a) To see the converse fails, define an independent sequence  $\{X_n\}$  by

$$X_n = \begin{cases} n^3 & \text{with probability } n^{-2} \\ 0 & \text{with probability } 1 - n^{-2} \end{cases}$$

Then  $\Pr(|X| > \epsilon) = n^{-2}$  for all large  $n$ , and so  $X_n \xrightarrow{p} 0$ . However,  $\mathbb{E}|X_n| = n \rightarrow \infty$ .

(b) To see why the converse is false, fix  $0 < r < \infty$ , let  $\Omega := [0, 1]$  with  $\mathbb{P}$  uniform on  $\Omega$  and consider  $X_n := n^{1/r} \mathbf{1}_{\{[0, 1/n]\}}$ . Then  $X_1, X_2, \dots$  converges in probability to 0 since if  $1 > \epsilon > 0$ , then  $\mathbb{P}(|X_n - 0| > \epsilon) \leq \mathbb{P}([0, 1/n]) = 1/n \rightarrow 0$  as  $n \rightarrow \infty$ . However,  $X_1, X_2, \dots$  does not converge in  $L_r$  to 0 since  $\mathbb{E}|X_n - 0|^r = n/n = 1$  for all  $n \geq 1$ .

(c)

**Theorem 8.3.2. Some exceptions (Theorem 7.2.4).**

- (a) If  $X_n \xrightarrow{d} c$  where  $c$  is constant, then  $X_n \xrightarrow{p} c$ .
- (b) If  $X_n \xrightarrow{p} X$  and  $\Pr(|X_n| \leq k) = 0$  for all  $n$  and some  $k$ , then  $X_n \xrightarrow{r} X$  for all  $r \geq 1$ .
- (c) If  $P_n(\epsilon) = \Pr(|X_n - X| > \epsilon)$  satisfies  $\sum_n P_n(\epsilon) < \infty$  for all  $\epsilon > 0$ , then  $X_n \xrightarrow{a.s.} X$ .

*Proof.* (Part (c).) Let  $A_n(\epsilon) = \{|X_n - X| > \epsilon\}$  (so that  $P_n(\epsilon) = \Pr[A_n(\epsilon)]$ ), and let  $B_m(\epsilon) = \bigcup_{n \geq m} A_n(\epsilon)$ . Then

$$\Pr(B_m(\epsilon)) \leq \sum_{n=m}^{\infty} \Pr(A_n(\epsilon))$$

so  $\lim_{m \rightarrow \infty} \Pr(B_m(\epsilon)) = 0$  whenever  $\sum_n \Pr(A_n(\epsilon)) < \infty$ . See also Lemma 8.4.1 part (b). □

**Lemma 8.3.3. (Lemma 7.2.6 from Grimmett and Stirzaker)**

- (a) If  $r > s \geq 1$  and  $X_n \xrightarrow{r} X$ , then  $X_n \xrightarrow{s} X$ .
- (b) If  $X_n \xrightarrow{1} X$  then  $X_n \xrightarrow{p} X$ .

The converse assertions fail in general.

*Proof.* (a) Using Lyapunov's Inequality (Lemma 8.2.10), if  $r > s \geq 1$

$$[\mathbb{E}(|X_n - X|^s)]^{1/s} \leq [\mathbb{E}(|X_n - X|^r)]^{1/r}$$

Therefore if  $X_n \xrightarrow{r} X$  (meaning  $\lim_{n \rightarrow \infty} \mathbb{E}(|X_n - X|^r) = 0$ ), then  $\lim_{n \rightarrow \infty} \mathbb{E}(|X_n - X|^s) = 0$ , so  $X_n \xrightarrow{s} X$ . We show the converse fails by counterexample:

$$X_n = \begin{cases} n & \text{with probability } n^{(-1/2)(r+s)} \\ 0 & \text{with probability } 1 - n^{(-1/2)(r+s)} \end{cases}$$

Then  $\mathbb{E}|X_n^s| = n^{(1/2)(s-r)} \rightarrow 0$  and  $\mathbb{E}|X_n^r| = n^{(1/2)(r-s)} \rightarrow \infty$ .

- (b) See proof of Theorem 8.3.1(a).

□

## 8.4 More on convergence (7.2 of Grimmett and Stirzaker)

**Other theorems to include:** Fatou's Lemma, Fubini's Theorem, Kolmogorov's Maximal Inequality, Kolmogorov Three-Series Test, Lindeberg Feller Central Limit Theorem, **this and more at beginning of Mike's 505A qual solutions.**

**Definition 8.12. Cauchy Convergence.** We say that the sequence  $\{X_n : n \geq 1\}$  of random variables on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is **almost surely Cauchy convergent** if

$$\Pr(\{\omega \in \Omega : X_m(\omega) - X_n(\omega) \rightarrow 0 \text{ as } m, n \rightarrow \infty\}) = 1$$

That is, the set of points  $\omega$  of the sample space for which the real sequence  $\{X_n(\omega) : n \geq 1\}$  is Cauchy convergent is an event having probability 1.

**Lemma 8.4.1. (Lemma 7.2.10, Grimmett and Stirzaker.)** Let  $A_n(\epsilon) = \{|X_n - X| > \epsilon\}$  and  $B_m(\epsilon) = \cup_{n \geq m} A_n(\epsilon)$ . Then:

- (a)  $X_n \xrightarrow{a.s.} X$  if and only if  $\Pr(B_m(\epsilon)) \rightarrow 0$  as  $m \rightarrow \infty$  for all  $\epsilon > 0$ .
- (b)  $X_n \xrightarrow{a.s.} X$  if  $\sum_n \Pr(A_n(\epsilon)) < \infty$  for all  $\epsilon > 0$ .
- (c) If  $X_n \xrightarrow{a.s.} X$  then  $X_n \xrightarrow{p} X$ , but the converse fails in general.

*Proof.* (a)

(b) As for Theorem 8.3.2 part (c).

(c) To see the converse fails, define an independent sequence  $\{X_n\}$  by

$$X_n = \begin{cases} 1 & \text{with probability } n^{-1} \\ 0 & \text{with probability } 1 - n^{-1} \end{cases}$$

Clearly  $X_n \xrightarrow{p} 0$ . However, if  $0 < \epsilon < 1$ ,

$$\begin{aligned} \Pr(B_m(\epsilon)) &= 1 - \lim_{r \rightarrow \infty} \Pr(X_n = 0 \text{ for all } n \text{ such that } m \leq n \leq r) \text{ (by Lemma 1.3.5)} \\ &= 1 - \left(1 - \frac{1}{m}\right) \left(1 - \frac{1}{m+1}\right) \cdots \text{ (by independence)} \\ &= 1 - \lim_{M \rightarrow \infty} \left( \frac{m-1}{m} \cdot \frac{m}{m+1} \cdot \frac{m+1}{m+2} \cdots \frac{M}{M+1} \right) \\ &= 1 - \lim_{M \rightarrow \infty} \frac{m-1}{M+1} = 1 \end{aligned}$$

and so  $\{X_n\}$  does not converge almost surely.

□

**Lemma 8.4.2. (Lemma 7.2.12, Grimmett and Stirzaker.)** There exist sequences which

- (a) converge almost surely but not in mean,
- (b) converge in mean but not almost surely.

*Proof.* (a) As for Lemma 8.3.3 part (b).

□

**Theorem 8.4.3. (Theorem 7.2.13, Grimmett and Stirzaker.)** If  $X_n \xrightarrow{p} X$ , there exists a non-random increasing sequence of integers  $n_1, n_2, \dots$  such that  $X_{n_i} \xrightarrow{a.s.} X$  as  $i \rightarrow \infty$ .

**Theorem 8.4.4. Skorokhod's representation theorem (Theorem 7.2.14, Grimmett and Stirzaker).** If  $\{X_n\}$  and  $X$  with distribution functions  $\{F_n\}$  and  $F$  are such that  $X_n \xrightarrow{d} X$  (or equivalently,  $F_n \rightarrow F$ ) as  $n \rightarrow \infty$ , then there exists a probability space  $(\Omega', \mathcal{F}', \mathbb{P}')$  and random variables  $\{Y_n\}$  and  $Y$  mapping  $\Omega'$  into  $\mathbb{R}$  such that

- (a)  $\{Y_n\}$  and  $Y$  have distribution functions  $\{F_n\}$  and  $F$
- (b)  $Y_n \xrightarrow{a.s.} Y$  as  $n \rightarrow \infty$

Therefore, although  $X_n$  may fail to converge to  $X$  in any mode other than in distribution, there exists a sequence  $\{Y_n\}$  such that  $Y_n$  is distributed identically to  $X_n$  for every  $n$ , which converges almost surely to a copy of  $X$ .

**Theorem 8.4.5.** (Theorem 7.2.19, Grimmett and Stirzaker; same as Portmanteau Theorem?)

The following three statements are equivalent:

- (a)  $X_n \xrightarrow{d} X$
- (b)  $\mathbb{E}[g(X_n)] \rightarrow \mathbb{E}[g(X)]$  for all bounded continuous functions  $g$ .
- (c)  $\mathbb{E}[g(X_n)] \rightarrow \mathbb{E}[g(X)]$  for all functions  $g$  of the form  $g(x) = f(x)\mathbf{1}_{[a,b]}(x)$  where  $f$  is continuous on  $[a, b]$  and  $a$  and  $b$  are points of continuity of the distribution function of the random variable  $X$ .

**Theorem 8.4.6.** (Grimmett and Stirzaker Theorem 7.3.9.)

- (a) If  $X_n \xrightarrow{\text{a.s.}} X$  and  $Y_n \xrightarrow{\text{a.s.}} Y$  then  $X_n + Y_n \xrightarrow{\text{a.s.}} X + Y$ .
- (b) If  $X_n \xrightarrow{r} X$  and  $Y_n \xrightarrow{r} Y$  then  $X_n + Y_n \xrightarrow{r} X + Y$ .
- (c) If  $X_n \xrightarrow{p} X$  and  $Y_n \xrightarrow{p} Y$  then  $X_n + Y_n \xrightarrow{p} X + Y$ .
- (d) It is not in general true that  $X_n + Y_n \xrightarrow{d} X + Y$  whenever  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{d} Y$ .

**Proposition 8.4.7.** Almost sure convergence does not imply convergence in  $L_2$ , and convergence in  $L_2$  does not imply almost sure convergence. That is, find random variables that converge in  $L_2$  but not almost surely. Then, find random variables that converge almost surely but not in  $L_2$ .

*Proof.* (a) **Almost sure convergence does not imply  $\ell_2$  convergence:** Let  $Z_n := n\mathbf{1}_{[0,1/n]}$ . Then  $Z_1, Z_2, \dots$  converges almost surely to 0 since  $\lim_{n \rightarrow \infty} Z_n = 0$  for all  $\omega \in (0, 1]$ , but it does not converge in  $L_2$  since  $\mathbb{E}(Z_n - 0)^2 = \mathbb{E}Z_n^2 = n^2\mathbb{E}\mathbf{1}_{[0,1/n]} = n \rightarrow \infty$  as  $n \rightarrow \infty$ .

(b)  **$\ell_2$  convergence does not imply almost sure convergence:** Define an independent sequence  $\{X_n\}$  by

$$X_n = \begin{cases} 1 & \text{with probability } n^{-1} \\ 0 & \text{with probability } 1 - n^{-1} \end{cases}$$

and let  $B_m(\epsilon) = \cup_{n \geq m} \{|X_n - X| > \epsilon\}$ .  $X_n \xrightarrow{2} 0$  because

$$\lim_{n \rightarrow \infty} \|X_n - 0\|_2 = 0$$

However, if  $0 < \epsilon < 1$ ,

$$\begin{aligned} \Pr(B_m(\epsilon)) &= 1 - \lim_{r \rightarrow \infty} \Pr(X_n = 0 \text{ for all } n \text{ such that } m \leq n \leq r) \\ &= 1 - \left(1 - \frac{1}{m}\right) \left(1 - \frac{1}{m+1}\right) \cdots \text{ (by independence)} \\ &= 1 - \lim_{M \rightarrow \infty} \left(\frac{m-1}{m} \cdot \frac{m}{m+1} \cdot \frac{m+1}{m+2} \cdots \frac{M}{M+1}\right) \\ &= 1 - \lim_{M \rightarrow \infty} \frac{m-1}{M+1} = 1 \end{aligned}$$

and so  $\{X_n\}$  does not converge almost surely.

□

**Theorem 8.4.8. Borel-Cantelli lemmas (Grimmett and Stirzaker Theorem 7.3.10.)** Let  $\{A_n\}$  be an infinite sequence of events from some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Let  $A = \bigcap_n \bigcup_{m=n}^{\infty} A_m = \limsup_{n \rightarrow \infty} A_n = \{A_n \text{ i.o.}\}$  be the event that infinitely many of the  $A_n$  occur. Then:

- (a)  $\Pr(A) = 0$  if  $\sum_n \Pr(A_n) < \infty$
- (b)  $\Pr(A) = 1$  if  $\sum_n \Pr(A_n) = \infty$  and  $A_1, A_2, \dots$  are independent events.

*Proof.* (a) We have that  $A \subseteq \bigcup_{m=n}^{\infty} A_m$  for all  $n$ , so

$$\Pr(A) \leq \sum_{m=n}^{\infty} \Pr(A_m) \rightarrow 0 \text{ as } n \rightarrow \infty$$

whenever  $\sum_n \Pr(A_n) < \infty$ .

- (b) One can confirm that

$$A^c = \bigcup_n \bigcap_{m=n}^{\infty} A_m^c$$

But

$$\begin{aligned} \Pr\left(\bigcap_{m=n}^{\infty} A_m^c\right) &= \lim_{r \rightarrow \infty} \Pr\left(\bigcap_{m=n}^r A_m^c\right) = \prod_{m=n}^{\infty} [1 - \Pr(A_m)] \text{ (by independence)} \leq \prod_{m=n}^{\infty} \exp(-\Pr(A_m)) \\ &= \exp\left(-\sum_{m=n}^{\infty} \Pr(A_m)\right) = 0 \end{aligned}$$

whenever  $\sum_n \Pr(A_n) = \infty$ , where the fourth step follows since  $1 - x \leq e^{-x}$  if  $x \geq 0$ . Thus

$$\Pr(A^c) = \lim_{n \rightarrow \infty} \Pr\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 0$$

so  $\Pr(A) = 1$ .

□

**Theorem 8.4.9. Kolmogorov's Two-Series Theorem.** Let  $X_1, X_2, \dots$  be independent random variables with  $\mathbb{E}(X_n) = \mu_n$  and  $\text{Var}(X_n) = \sigma_n^2$  such that  $\sum_{n=1}^{\infty} \mu_n < \infty$  and  $\sum_{n=1}^{\infty} \sigma_n^2 < \infty$ . Then  $\sum_{n=1}^{\infty} X_n$  converges in  $\mathbb{R}$  almost surely.

*Proof.* Available on wikipedia, [https://en.wikipedia.org/wiki/Kolmogorov%27s\\_two-series\\_theorem](https://en.wikipedia.org/wiki/Kolmogorov%27s_two-series_theorem).

□

#### 8.4.1 Slutsky's Convergence Theorems (8.4.1 of Pesaran, 7.3 of Grimmett and Stirzaker)

**Theorem 8.4.10. Theorem 6 of Pesaran, Section 8.4.1, p. 173.** Let  $\{x_t, y_t\}, t = 1, 2, \dots$  be a sequence of pairs of random variables with  $y_t \xrightarrow{d} y$  and  $|y_t - x_t| \xrightarrow{p} 0$ . Then  $x_t \xrightarrow{d} y$ .

**Theorem 8.4.11. Theorem 7 in Pesaran, on p.318 (section 7.3) of Grimmett and Stirzaker.** (Section 8.4.1, p. 174) If  $x_t \xrightarrow{d} x$  and  $y_t \xrightarrow{p} c$  where  $c$  is a finite constant, then

- (i)  $x_t + y_t \xrightarrow{d} x + c$
- (ii)  $y_t x_t \xrightarrow{d} cx$
- (iii)  $x_t/y_t \xrightarrow{d} x/c$ , if  $c \neq 0$ .

**Theorem 8.4.12. on p.318 (section 7.3) of Grimmett and Stirzaker.** Suppose that  $X_n \xrightarrow{d} 0$  and  $Y_n \xrightarrow{p} Y$ , and let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  be such that  $g(x, y)$  is a continuous function of  $y$  for all  $x$ , and  $g(x, y)$  is continuous at  $x = 0$  for all  $y$ . Then  $g(X_n, Y_n) \xrightarrow{p} g(0, Y)$ .

**Theorem 8.4.13 (Continuous Mapping Theorem (Theorem 9 of Pesaran, Section 8.4.1, p. 176: convergence properties of transformed sequences.)).** Suppose  $\{\mathbf{x}_j\}$ ,  $\{\mathbf{y}_j\}$ ,  $\mathbf{x}$ , and  $\mathbf{y}$  are  $k \times 1$  vectors of random variables on a probability space, and let  $\mathbf{g}(\cdot)$  be a continuous vector-valued function. (Alternatively, suppose  $g$  has the set of discontinuity points  $D_g$  such that  $\Pr(X \in D_g) = 0$ .) Then

- (i)  $\mathbf{x}_j \xrightarrow{a.s.} \mathbf{x} \implies \mathbf{g}(\mathbf{x}_j) \xrightarrow{a.s.} \mathbf{g}(\mathbf{x})$
- (ii)  $\mathbf{x}_j \xrightarrow{p} \mathbf{x} \implies \mathbf{g}(\mathbf{x}_j) \xrightarrow{p} \mathbf{g}(\mathbf{x})$
- (iii)  $\mathbf{x}_j \xrightarrow{d} \mathbf{x} \implies \mathbf{g}(\mathbf{x}_j) \xrightarrow{d} \mathbf{g}(\mathbf{x})$
- (iv)  $\mathbf{x}_j - \mathbf{y}_j \xrightarrow{p} \mathbf{0}$  and  $\mathbf{y}_j \xrightarrow{d} \mathbf{y} \implies \mathbf{g}(\mathbf{x}_j) - \mathbf{g}(\mathbf{y}_j) \xrightarrow{d} \mathbf{0}$

where  $\mathbf{x} = (c_1, \dots, c_k) \in \mathbb{R}^k$ .

*Proof (part (b), continuous case, one-dimensional codomain).* Let  $\mathbf{x}_j = (M_{j,1}, \dots, M_{j,k})$ . We have that

$$\begin{aligned} \forall \epsilon_j > 0, \lim_{n \rightarrow \infty} \Pr(\{\omega \in \Omega : |M_{j,n}(\omega) - c_j| > \epsilon_j\}) &= 0, & \forall j \in \{1, \dots, k\}. \\ \iff \forall \epsilon_j > 0, \lim_{n \rightarrow \infty} \Pr(\{\omega \in \Omega : |M_{j,n}(\omega) - c_j| < \epsilon_j\}) &= 1, & \forall j \in \{1, \dots, k\}. \end{aligned} \tag{8.2}$$

Because  $g$  is continuous, we have that for every  $\epsilon^* > 0$  there exists a  $\delta^* > 0$  such that

$$0 < \|(M_{1,n}(\omega), \dots, M_{j,n}(\omega))\|_2 < \delta^* \implies |g(M_{1,n}, \dots, M_{j,n}) - g(c_1, \dots, c_j)| < \epsilon^*. \tag{8.3}$$

Note that since in  $\mathbb{R}$  the  $L_2$  and  $L_1$  norms are equivalent,

$$\begin{aligned}
|M_{j,n}(\omega) - c_j| < \epsilon_j &\iff \|M_{j,n}(\omega) - c_j\|_2 < \epsilon_j \implies \sum_{j=1}^k \|M_{j,n}(\omega) - c_j\|_2 < \sum_{j=1}^k \epsilon_j \\
&\implies \|(M_{1,n}(\omega), \dots, M_{j,n}(\omega))\|_2 < \sum_{j=1}^k \epsilon_j
\end{aligned}$$

where the last step follows by the Triangle Inequality. Therefore letting  $\delta^* = \sum_{j=1}^k \epsilon_j$ , we have

$$\begin{aligned}
\Pr(\{\omega \in \Omega : |M_{j,n}(\omega) - c_j| < \epsilon_j\}) &\leq \Pr(0 < \|(M_{1,n}(\omega), \dots, M_{j,n}(\omega))\|_2 < \delta^*) \\
&\leq \Pr(\{\omega \in \Omega : |g(M_{1,n}(\omega), \dots, M_{j,n}(\omega)) - g(c_1, \dots, c_j)| < \epsilon^*\})
\end{aligned}$$

where the last step follows from (8.3). So

$$\Pr(\{\omega \in \Omega : |M_{j,n}(\omega) - c_j| < \epsilon_j\}) \leq \Pr(\{\omega \in \Omega : |g(M_{1,n}(\omega), \dots, M_{j,n}(\omega)) - g(c_1, \dots, c_j)| < \epsilon^*\}). \quad (8.4)$$

Taking limits of (8.4) and substituting in (8.2), we have

$$\begin{aligned}
\forall \epsilon^* > 0, \lim_{n \rightarrow \infty} \Pr(\{\omega \in \Omega : |g(M_{1,n}(\omega), \dots, M_{j,n}(\omega)) - g(c_1, \dots, c_j)| < \epsilon^*\}) &\geq 1 \\
\iff \forall \epsilon^* > 0, \lim_{n \rightarrow \infty} \Pr(\{\omega \in \Omega : |g(M_{1,n}(\omega), \dots, M_{j,n}(\omega)) - g(c_1, \dots, c_j)| > \epsilon^*\}) &= 0 \\
\iff g(M_{1,n}, \dots, M_{j,n}) &\xrightarrow{p} g(c_1, \dots, c_j).
\end{aligned}$$

For remaining parts, see Serfling [1980] or Rao [1973]. □

See also:

**Theorem 8.4.14. (Theorem 7.2.18, Grimmett and Stirzaker.)** If  $X_n \xrightarrow{d} X$  and  $g : \mathbb{R} \rightarrow \mathbb{R}$  is continuous, then  $g(X_n) \xrightarrow{d} g(X)$ .

## 8.5 Stochastic orders $\mathcal{O}_p(\cdot)$ and $o_p(\cdot)$ (Pesaran 8.5)

**Definition 8.13 (Pesaran 8.5 Definition 6.).** Let  $\{a_t\}$  be a sequence of positive numbers and  $\{x_t\}$  be a sequence of random variables. Then

- (i)  $x_t = O_p(a_t)$ , or  $x_t/a_t$  is bounded in probability, if for every  $\epsilon > 0$  there exist real numbers  $M_\epsilon$  and  $N_\epsilon$  such that

$$\Pr\left(\frac{|x_t|}{a_t} > M_\epsilon\right) < \epsilon, \quad \text{for } t > N_\epsilon$$

- (ii)  $x_t = o_p(a_t)$  if

$$\frac{x_t}{a_t} \xrightarrow{p} 0$$

**Definition 8.14** (Ross ISE 620 Definition). We say that  $f(x)$  is  $o(h)$  if  $\lim_{h \rightarrow 0} f(h)/h = 0$ .

## 8.6 Laws of Large Numbers and Central Limit Theorems (Pesaran 8.6; Grimmett and Stirzaker 7.4, 7.5)

**Theorem 8.6.1. Weak Law of Large Numbers (Khintchine) (Pesaran 8.6 Theorem 10, Grimmett and Stirzaker Theorem 7.4.7, 541A notes Theorem 2.10).** Suppose that  $\{X_k\}$  is a sequence of (i) independent (ii) identically distributed random variables with (iii) constant means, i.e.,  $\mathbb{E}(X_k) = \mu < \infty$ . Then

$$\bar{X}_k = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{p} \mu$$

**Theorem 8.6.2. Weak Law of Large Numbers (Chebyshev) (Pesaran Section 8.6, p. 178, Theorem 11.)** Let  $\{X_k\}$  be a sequence of random variables. If (i)  $\mathbb{E}(X_k) = \mu_k$ , (ii)  $\text{Var}(X_k) = \sigma_k^2$ , and (iii)  $\text{Cov}(X_k, X_j) = 0$ ,  $k \neq j$ , and (iv)

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \sigma_k^2 < \infty$$

then we have  $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$ , where  $\bar{\mu}_n = n^{-1} \sum_{k=1}^n \mu_k$ .

**Theorem 8.6.3. Strong Law of Large Numbers (Grimmett and Stirzakker Theorem 7.4.3).** Let  $\{X_k\}$  be a sequence of (i) independent (ii) identically distributed random variables with (iii)  $\mathbb{E}(X_k) = \mu$  and (iv)  $\mathbb{E}(X_k^2) < \infty$ . Then

$$\frac{1}{n} \sum_{k=1}^n X_k \rightarrow \mu \text{ almost surely and in mean square.}$$

**Theorem 8.6.4 (Strong Law of Large Numbers (Grimmett and Stirzakker Theorem 7.5.1, 541A notes Theorem 2.11)).** Let  $\{X_k\}$  be a sequence of (i) independent (ii) identically distributed random variables. Then if and only if (iii)  $\mathbb{E}|X_k| < \infty$ ,

$$\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{a.s.} \mu$$

**Theorem 8.6.5. Strong Law of Large Numbers 1 (Kolmogorov) (Pesaran 8.8 Theorem 12).**

Let  $\{X_k\}$  be a sequence of (i) independent random variables with (ii)  $\mathbb{E}(X_k) = \mu_k < \infty$  and (ii)  $\text{Var}(X_k) = \sigma_k^2$  such that (iii)

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} < \infty$$

Then  $\bar{X}_n - \bar{\mu}_n \xrightarrow{w.p.1} 0$ . If the independence assumption (i) is replaced by a lack of correlation (i.e.  $\text{Cov}(X_k, X_j) = 0, k \neq j$ ), the convergence of  $\bar{X}_n - \bar{\mu}_n$  with probability one requires the stronger condition

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2 (\log k)^2}{k^2} < \infty$$

**Theorem 8.6.6. Strong Law of Large Numbers 2 (Pesaran 8.8 Theorem 13)** Suppose that  $X_1, X_2, \dots$  are (i) independent random variables, and that (ii)  $\mathbb{E}(X_k) = 0$ , (iii)  $\mathbb{E}(X_k^4) \leq M \forall k$  where  $M$  is an arbitrary positive constant. Then

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{a.s.} 0$$

**Theorem 8.6.7. Central Limit Theorem (Grimmett and Stirzaker theorem 5.10.4.)** Let  $X_1, X_2, \dots$  be a sequence of independent identically distributed random variables with finite mean  $\mu$  and finite non-zero variance  $\sigma^2$ , and let  $S_n = \sum_{i=1}^n X_i$ . Then

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

**Theorem 8.6.8. (Berry-Esseen Central Limit Theorem.)** There exists  $c > 0$  such that the following holds. Let  $X_1, X_2, \dots$  be i.i.d. real-valued random variables with mean zero, variance 1, and  $\mathbb{E}(|X_1|^3) < \infty$ . Let  $Z$  be a standard Gaussian random variable. Then for any  $n \geq 1$ ,

$$\sup_{t \in \mathbb{R}} \left| \Pr \left( \frac{X_1 + \dots + X_n}{\sqrt{n}} \leq t \right) - \Pr(Z \leq t) \right| \leq c \cdot \frac{\mathbb{E}(|X_1|^3)}{\sqrt{n}}$$

**Remark.** You can look up what the  $c$  is; Heilman doesn't think it's any bigger than around 10.

**Theorem 8.6.9. (Central Limit Theorem in  $\mathbb{R}^d$ , Heilman notes Theorem 2.33.)** Let  $X^{(1)}, X^{(2)}, \dots$  be a sequence of independent identically distributed  $\mathbb{R}^d$ -valued random variables. (Notation: we write  $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)})$ .) Assume  $\mathbb{E}(X^{(n)}) = \boldsymbol{\mu}$  for all  $n \geq 1$  and for any  $1 \leq i < j \leq d$ , all of the covariances

$$a_{ij} = \mathbb{E}[(X_i^{(1)} - \mathbb{E}(X_i^{(1)}))(X_j^{(1)} - \mathbb{E}(X_j^{(1)})]$$

are finite. Let  $\mathbf{S}_n = \sum_{i=1}^n X^{(i)}$ . Then as  $n \rightarrow \infty$ ,

$$\frac{\mathbf{S}_n - n\boldsymbol{\mu}}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, [a_{ij}])$$

**Theorem 8.6.10. (Grimmett and Stirzaker theorem 5.10.5.)** Let  $X_1, X_2, \dots$  be independent random variables satisfying  $\mathbb{E}(X_j) = 0$ ,  $\text{Var}(X_j) = \sigma_j^2$ ,  $\mathbb{E}|X_j^3| < \infty$  such that

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma(n)^3} \sum_{j=1}^n \mathbb{E}|X_j^3| = 0$$

where  $\sigma(n)^2 = \text{Var}(\sum_{j=1}^n X_j) = \sum_{j=1}^n \sigma_j^2$ . Then

$$\frac{1}{\sigma(n)} \sum_{j=1}^n X_j \xrightarrow{d} \mathcal{N}(0, 1)$$

*Proof.* See Loeve [1977, p. 287] and Grimmett and Stirzaker Problem 5.12.40.  $\square$

**Lemma 8.6.11. Lindeberg's Condition:** Let  $\{X_k\}$  be a sequence of independent (not necessarily identically distributed) random variables with expectations  $\mu_k$  and finite variances  $\sigma_k^2$ . Let  $s_n^2 = \sum_{k=1}^n \sigma_k^2$ . If such a sequence of independent random variables  $X_k$  satisfies the condition

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{k=1}^n \mathbb{E}[(X_k - \mu_k)^2) \cdot \mathbf{1}_{\{|X_k - \mu_k| > \epsilon s_n\}}] = 0$$

for all  $\epsilon > 0$  then the central limit theorem holds; that is, the random variables

$$Z_n = \frac{1}{s_n} \sum_{k=1}^n (X_k - \mu_k)$$

converge in distribution to  $\mathcal{N}(0, 1)$  as  $n \rightarrow \infty$ .

## 8.7 The case of dependent and heterogeneously distributed observations (Pesaran 8.8)

**Theorem 8.7.1. Central limit theorem for martingale difference sequences (Pesaran 8.8 Theorem 28).** Let  $\{x_t\}$  be a martingale difference sequence with respect to the information set  $\Omega_t$ . Let  $\bar{\sigma}_T^2 = \text{Var}(\sqrt{T}\bar{x}_T) = T^{-1} \sum_{t=1}^T \sigma_t^2$ . If  $\mathbb{E}(|x_t|^r) < K < \infty$  for any  $r > 2$  and for all  $t$ , and

$$\frac{1}{T} \sum_{t=1}^T x_t^2 - \bar{\sigma}_T^2 \xrightarrow{p} 0$$

then  $\sqrt{T}\bar{x}_T / \bar{\sigma}_T \xrightarrow{d} \mathcal{N}(0, 1)$ .

## 8.8 Worked Examples from Math 505A Midterm 2

- (1) (a) **Fall 2010 Problem 1.** Let  $X_k$ ,  $k \geq 1$ , be i.i.d. random variables with mean 1 and variance 1.  
Show that the limit

$$\lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n X_k}{\sum_{k=1}^n X_k^2}$$

exists in an appropriate sense, and identify the limit.

- (b) **Not included on midterm or final.** Let  $(X_j)_{j \geq 1}$  be i.i.d. uniform on  $(-1, 1)$ . Let

$$Y_n = \frac{\sum_{j=1}^n X_j}{\sum_{j=1}^n X_j^2 + \sum_{j=1}^n X_j^3}$$

Prove that  $\lim_{n \rightarrow \infty} \sqrt{n} Y_n$  exists in an appropriate sense, and identify the limit.

**Solution.**

(a)

$$\lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n X_k}{\sum_{k=1}^n X_k^2} = \lim_{n \rightarrow \infty} \frac{n^{-1} \sum_{k=1}^n X_k}{n^{-1} \sum_{k=1}^n X_k^2}$$

Since  $X_1, X_2, \dots$  are i.i.d.,  $E(X_1^2) = \text{Var}(X_1) + (\mathbb{E}(X_1))^2 = 2 < \infty$ , we have

$$n^{-1} \sum_{k=1}^n X_k \xrightarrow{a.s.} \mathbb{E}(X_1) = 1 \text{ as } n \rightarrow \infty$$

by Theorem 8.6.3 (Strong Law of Large Numbers). Also,  $X_1^2, X_2^2, \dots$  are clearly identically distributed, and are independent by Theorem 4.2.3 (“If  $X$  and  $Y$  are independent, then so are  $g(X)$  and  $g(Y)$ .”). It is clear also that  $\mathbb{E}(|X_1^2|) = \mathbb{E}(X_1^2) = \text{Var}(X_1) + \mathbb{E}(X_1)^2 = 1 + 1 = 2 < \infty$ . Therefore by Theorem 8.6.4 (Strong Law of Large Numbers),

$$n^{-1} \sum_{k=1}^n X_k^2 \xrightarrow{a.s.} \mathbb{E}(X_1^2) = 2 \text{ as } n \rightarrow \infty$$

(From here I had two different ways of finishing the problem.)

- Because we have almost sure convergence in the numerator and denominator, by the Continuous Mapping Theorem (Theorem 8.4.13),

$$\lim_{n \rightarrow \infty} \frac{n^{-1} \sum_{k=1}^n X_k}{n^{-1} \sum_{k=1}^n X_k^2} = \frac{\lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n X_k}{\lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n X_k^2} \xrightarrow{a.s.} \boxed{\frac{1}{2}}$$

- Then, using one of Slutsky’s convergence theorems (Theorem 8.4.11: “If  $x_t \xrightarrow{d} x$  and  $y_t \xrightarrow{p} c$  where  $c$  is a finite constant, then  $x_t/y_t \xrightarrow{d} x/c$ , if  $c \neq 0$ .”), we have

$$\frac{n^{-1} \sum_{k=1}^n X_k}{n^{-1} \sum_{k=1}^n X_k^2} \xrightarrow{d} \frac{\mathbb{E}(X_1)}{\mathbb{E}(X_1^2)} = \frac{\mathbb{E}(X_1)}{\text{Var}(X_1) + \mathbb{E}(X_1)^2} = \frac{1}{1+1} = \frac{1}{2}$$

But then, by Theorem 8.3.2 (Theorem 7.2.4(a) in Grimmett and Stirzaker: “If  $X_n \xrightarrow{d} c$  where  $c$  is constant, then  $X_n \xrightarrow{p} c$ ”), we have  $\frac{n^{-1} \sum_{k=1}^n X_k}{n^{-1} \sum_{k=1}^n X_k^2} \xrightarrow{p} 1/2$ .

(b) (Not included on midterm or final.)

$$Y_n = \frac{\sum_{j=1}^n X_j}{\sum_{j=1}^n X_j^2 + \sum_{j=1}^n X_j^3} = \frac{n^{-1} \sum_{j=1}^n X_j}{n^{-1} \sum_{j=1}^n X_j^2 + n^{-1} \sum_{j=1}^n X_j^3}$$

Note that  $\mathbb{E}(X_1) = 0, \mathbb{E}(X_1^2) = \text{Var}(X_1) + \mathbb{E}(X_1)^2 = (1 - -1)^2/12 + 0^2 = 1/3, \mathbb{E}(X_1^3) = (1/2) \int_{-1}^1 x^3 dx = 0$ . (We derived the formulae for the first three moments of a uniform distribution on Homework 4 problem 2(2).)

$$\implies \sqrt{n}Y_n = \frac{\sqrt{1/3}(\sum_{j=1}^n X_j - n\mathbb{E}(X_1)) / \sqrt{n \cdot 1/3}}{n^{-1} \sum_{j=1}^n X_j^2 + n^{-1} \sum_{j=1}^n X_j^3}$$

By the Central Limit Theorem (Theorem 8.6.7),

$$\frac{\sum_{j=1}^n X_j - n\mathbb{E}(X_1)}{\sqrt{n \cdot 1/3}} \xrightarrow{d} \mathcal{N}(0, 1)$$

By the Law of Large Numbers (Theorem 8.6.4), since  $\mathbb{E}(|X_1^2|) = \mathbb{E}(X_1^2) = 1/3 < \infty$ ,

$$\frac{1}{n} \sum_{j=1}^n X_j^2 \xrightarrow{a.s.} \mathbb{E}(X_1^2) = 1/3$$

By the Law of Large Numbers (Theorem 8.6.4), since  $\mathbb{E}(|X_1^3|) = (1/2) \int_{-1}^1 |x^3| dx = \int_0^1 x^3 dx = 1/4 < \infty$ ,

$$\frac{1}{n} \sum_{j=1}^n X_j^3 \xrightarrow{a.s.} \mathbb{E}(X_1^3) = 0$$

In the denominator, since we have almost sure convergence, the regular rules of calculus/real analysis apply. That is, using the above results,

$$n^{-1} \sum_{j=1}^n X_j^2 + n^{-1} \sum_{j=1}^n X_j^3 \xrightarrow{a.s.} 1/3$$

Therefore

$$\sqrt{n}Y_n = \frac{\sqrt{1/3}(\sum_{j=1}^n X_j - n\mathbb{E}(X_1)) / \sqrt{n \cdot 1/3}}{n^{-1} \sum_{j=1}^n X_j^2 + n^{-1} \sum_{j=1}^n X_j^3} \xrightarrow{d} \frac{\sqrt{1/3}}{1/3} \mathcal{N}(0, 1) = \boxed{\mathcal{N}(0, 3)}$$

- (2) **Fall 2010 Problem 2.** Fix  $p \in (0, 1)$  and consider independent Poisson random variables  $X_k, k \geq 1$  with

$$\mathbb{E}X_k = \frac{p^k}{k}$$

Verify that the sum  $\sum_{k=1}^{\infty} kX_k$  converges with probability one and determine the distribution of the random variable  $Y = \sum_{k=1}^{\infty} kX_k$ .

**Solution. Melike's solution (use for midterm):** We have  $\mathbb{E}[kX_k] = p^k$  and  $\sum_{k=1}^{\infty} p^k = p/(1-p) < \infty$ , and  $\text{Var}(kX_k) = kp^k$  and

$$\sum_{k=1}^{\infty} kp^k = p \sum_{k=1}^{\infty} kp^{k-1} = p \frac{d}{dp} \sum_{k=1}^{\infty} p^k = p \frac{d}{dp} \frac{p}{1-p} = p \cdot \frac{(1-p) - p(-1)}{(1-p)^2} = \frac{p}{(1-p)^2} < \infty$$

Since the sequence  $\{Y_k\}_{k \geq 1}$  is independent, by Kolmogorov's Two Series Theorem (Theorem 8.4.9: "Let  $X_1, X_2, \dots$  be independent random variables with  $\mathbb{E}(X_n) = \mu_n$  and  $\text{Var}(X_n) = \sigma_n^2$  such that  $\sum_{n=1}^{\infty} \mu_n < \infty$  and  $\sum_{n=1}^{\infty} \sigma_n^2 < \infty$ . Then  $\sum_{n=1}^{\infty} X_n$  converges in  $\mathbb{R}$  almost surely."), we conclude that  $\sum_{k=1}^{\infty} kX_k$  converges almost surely.

To find the distribution of  $Y$ , let  $X$  be a Poisson random variable and consider its probability generating function:

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} s^k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda} e^{\lambda s} = e^{\lambda(s-1)}$$

So  $\mathbb{E}(s^{X_k}) = \exp\left(\frac{p^k}{k}(s-1)\right)$  and  $\mathbb{E}(s^{kX_k}) = \mathbb{E}[(s^k)^{X_k}] = \exp\left(\frac{p^k}{k}(s^k-1)\right)$ . Then define  $Y_n = \sum_{k=1}^n kX_k$  and consider

$$\begin{aligned} G_{Y_n}(s) &= \mathbb{E}(s^{Y_n}) = \mathbb{E}\left(\prod_{k=1}^n s^{kX_k}\right) = \prod_{k=1}^n \mathbb{E}(s^{kX_k}) = \prod_{k=1}^n \exp\left(\frac{p^k}{k}(s^k-1)\right) = \exp\left(\sum_{k=1}^n \frac{p^k}{k}(s^k-1)\right) \\ &= \exp\left(\sum_{k=1}^n \frac{(ps)^k}{k} - \sum_{k=1}^n \frac{p^k}{k}\right) \end{aligned}$$

Now, by taking limits as  $n \rightarrow \infty$  (since we are allowed to take limit inside of expectation here), we get

$$\begin{aligned} G_Y(s) &= \mathbb{E}(s^Y) = \exp\left(\sum_{k=1}^{\infty} \frac{(ps)^k}{k} - \sum_{k=1}^{\infty} \frac{p^k}{k}\right) = \exp\left(\int \sum_{k=1}^{\infty} (ps)^{k-1} dp - \int \sum_{k=1}^{\infty} p^{k-1} dp\right) \\ &= \exp\left(\int \frac{1}{1-ps} dp - \int \frac{1}{1-p} dp\right) = \exp(-\log(1-ps) + \log(1-p)), \quad -1 \leq ps < 1 \text{ and } -1 \leq p < 1 \\ &= \frac{1-p}{1-ps}, \quad -1 \leq ps < 1 \end{aligned}$$

Since we know  $\Pr(X = k) = \frac{G_X^{(k)}(0)}{k!}$ , we have

$$G_Y(s) = \frac{1-p}{1-sp}, \quad G'(s) = \frac{p(1-p)}{(1-sp)^2}, \quad G''(s) = \frac{2p^2(1-p)}{(1-sp)^3}, \quad G^{(3)}(s) = \frac{3 \cdot 2p^3(1-p)}{(1-sp)^3}, \dots$$

$$G^{(k)}(s) = \frac{k!p^k(1-p)}{(1-sp)^k} \text{ for } k = 0, 1, 2, \dots$$

So we have

$$\Pr(Y = k) = (1-p)p^k, \quad k = 0, 1, 2, \dots$$

$$= \Pr(G_1(1-p) = k+1) = \Pr(G_1(1-p) - 1 = k)$$

which means  $Y \sim G_1(1-p) - 1$ .

## (3) Spring 2017 Problem 3.

- (a) Consider the sequence  $\{X_k, k \geq 1\}$  of random variables such that  $X_1$  is uniform on  $(0, 1)$  and, given  $X_k$ , the distribution of  $X_{k+1}$  is uniform on  $(0, CX_k)$ , where  $\sqrt{3} < C < 2$ .
- (i) For  $n \geq 1$ , compute the conditional expectation  $\mathbb{E}(X_{n+1}^r | X_n)$ .
  - (ii) For  $n \geq 1$ , compute  $\mathbb{E}(X_n^r)$ .
  - (iii) Show that  $\lim_{x \rightarrow \infty} X_n = 0$  in  $\ell_1$  and with probability one, but not in  $\ell_2$ .
  - (iv) Investigate the same questions for all other values of  $C > 0$ .
- (b) Let  $a > 0$ , let  $X_n, n \geq 1$  be i.i.d. random variables that are uniform on  $(0, a)$ , and let  $Y_n = \prod_{k=1}^n X_k$ . Determine, with a proof, all values of  $a$  for which  $\lim_{n \rightarrow \infty} Y_n = 0$  with probability one.

**Solution.**

- (a) (i) We have that  $X_{n+1} | X_n \sim U(0, CX_n)$ . Therefore

$$\begin{aligned} \mathbb{E}(X_{n+1}^r | X_n) &= \frac{1}{CX_n} \int_0^{CX_n} x^r dx = \frac{1}{CX_n} \cdot \frac{x^{r+1}}{r+1} \Big|_0^{CX_n} = \frac{C^r X_n^r}{r+1} \\ \implies \mathbb{E}(X_{n+1}^r) &= \mathbb{E}[\mathbb{E}(X_{n+1}^r | X_n)] = \frac{C^r}{r+1} \cdot \mathbb{E}(X_n^r) \\ \implies \boxed{\mathbb{E}(X_{n+1}^r | X_n) = \frac{C^r}{r+1} X_n^r} \end{aligned}$$

- (ii) Note that  $E(X_1^r) = \int_0^1 x^r dr = 1/(r+1)$ . Therefore

$$\mathbb{E}(X_{n+1}^r) = \frac{C^r}{r+1} \cdot \mathbb{E}(X_n^r) = \left(\frac{C^r}{r+1}\right)^n \cdot \mathbb{E}(X_1^r) = \boxed{\left(\frac{C^r}{r+1}\right)^n \cdot \frac{1}{r+1}}$$

- (iii) We would like to show that  $X_n \xrightarrow{w.p.1} 0$  and that  $X_n \xrightarrow{1} 0$ , but that the same result does not follow for the  $\ell_2$  norm.

- **Convergence with probability one:** We seek to show that  $\Pr(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = 0\}) = 1$ . By Markov's Inequality (Lemma 8.2.1), we have

$$\begin{aligned} \Pr(|X_n| \geq a) &\leq \frac{\mathbb{E}(X_n)}{a} \quad \forall a > 0 \\ \iff \Pr(|X_n| \geq a) &\leq \left(\frac{C^1}{1+1}\right)^{n-1} \cdot \frac{1}{1+1} \cdot \frac{1}{a} = \left(\frac{C}{2}\right)^{n-1} \cdot \frac{1}{2a} \quad \forall a > 0 \end{aligned}$$

Since  $\sqrt{3} < C < 2$ ,  $\sqrt{3}/2 < C/2 < 1$ . Since  $X_n \in [0, CX_{n-1}]$ ,  $X_n \geq 0$ , so  $|X_n| = X_n$ . Therefore we have

$$\Pr(\lim_{n \rightarrow \infty} |X_n| \geq a) = \Pr(\lim_{n \rightarrow \infty} X_n \geq a) \leq \lim_{n \rightarrow \infty} \left(\frac{C}{2}\right)^{n-1} \cdot \frac{1}{2a} = 0 \quad \forall a > 0$$

Since  $|X_n| \geq 0$ , this implies that  $\Pr(\lim_{n \rightarrow \infty} X_n = 0) = \Pr(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = 0\}) = 1$ , so by the Borel-Cantelli Lemma (Theorem 8.4.8),  $X_n$  converges to 0 with probability 1.

- **Convergence in  $\ell_1$  norm:** We seek to show that  $\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|) = 0$ . Since  $X_n \in [0, CX_{n-1}]$ ,  $X_n \geq 0$ , so  $|X_n| = X_n$ . Therefore

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n) = \lim_{n \rightarrow \infty} \left(\frac{C}{2}\right)^{n-1} \cdot \frac{1}{2}$$

Since  $\sqrt{3} < C < 2$ ,  $\sqrt{3}/2 < C/2 < 1$ , so  $C/2 < 1$ . Therefore we have

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|) = \lim_{n \rightarrow \infty} \left(\frac{C}{2}\right)^{n-1} \cdot \frac{1}{2} = 0$$

so  $X_n$  converges to 0 in 1st mean.

- **Convergence in  $\ell_2$  norm:** We seek to show that  $\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|^2) \neq 0$ . We have

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|^2) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n^2) = \lim_{n \rightarrow \infty} \left(\frac{C^2}{3}\right)^{n-1} \cdot \frac{1}{3}$$

Since  $\sqrt{3} < C < 2$ ,  $3/3 < C^2/3 < 4/3$ , so  $C^2/3 > 1$ . Therefore we have

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|^2) = \lim_{n \rightarrow \infty} \left(\frac{C^2}{3}\right)^{n-1} \cdot \frac{1}{3} = \infty \neq 0$$

so  $X_n$  does not converge to 0 in 2nd mean.

- (iv) From the above, it is clear that for convergence with probability one or in 1st mean we require  $0 < C/2 < 1$  and for convergence in second mean we require  $0 < C^2/3 < 1$ . For  $0 < C < \sqrt{3}$ , we see that  $X_n$  would converge to zero in 2nd mean since this would imply that  $0 < C^2/3 < 1$ . It would also still converge to 0 in 1st mean (and with probability 1) since we would have  $(0 < C/2 < \sqrt{3}/2 < 1)$ .

For  $C = \sqrt{3}$ ,  $X_n$  would still converge to 0 with probability one and in 1st mean for the same reasons. However, it would not converge in 2nd mean because we would have

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|^2) = \lim_{n \rightarrow \infty} \left(\frac{\sqrt{3}^2}{3}\right)^{n-1} \cdot \frac{1}{3} = \frac{1}{3} \neq 0$$

For  $C \geq 2$ , it would diverge in all three cases, since in this case  $C/2 \geq 2/2 = 1$  and  $C^2/3 \geq 4/3 > 1$ .

- (b) **Probably won't be on midterm.** Note that

$$\lim_{n \rightarrow \infty} Y_n = \lim_{n \rightarrow \infty} \prod_{k=1}^n X_k = 0 \iff \log(Y_n) = \log\left(\prod_{k=1}^n X_k\right) = \sum_{k=1}^n \log(X_k) \rightarrow -\infty$$

Note that

$$\begin{aligned} \mathbb{E}[\log(Y_n)] &= \mathbb{E}\left(\sum_{k=1}^n \log(X_k)\right) = \sum_{k=1}^n \mathbb{E}[\log(X_k)] = \sum_{k=1}^n \mathbb{E}[\log(X_1)] = \sum_{k=1}^n \int_0^a (\log(x)/a) dx \\ &= \sum_{k=1}^n \frac{1}{a} [x \log x - x]_0^a = \sum_{k=1}^n \frac{a \log a - a}{a} = \sum_{k=1}^n (\log(a) - 1) = n(\log(a) - 1) \end{aligned}$$

As  $n \rightarrow \infty$  we have

$$\mathbb{E}[\log(Y_n)] = \begin{cases} -\infty & a < e \\ 0 & a = e \\ \infty & a > e \end{cases}$$

Since  $\mathbb{E}[\log(Y_n)] \rightarrow \infty$  for  $a < e$ , we have  $\lim_{n \rightarrow \infty} Y_n = 0$  for  $a < 3$ . Therefore

$$\lim_{n \rightarrow \infty} Y_n = \lim_{n \rightarrow \infty} \prod_{k=1}^n X_k = 0 \iff a < e.$$

## 8.9 Estimators and Central Limit Theorems (DSO 607)

Lyapunov (?) condition: can prove central limit theorem if we check 3rd moment. Lindeberg's Condition (Lemma 8.6.11)

## 9 Convex Optimization

These are my notes from taking EE 588 at USC taught by Mahdi Soltanolkotabi and the textbook *Convex Optimization* (Boyd and Vandenberghe) 7th printing [Boyd et al., 2004], as well as Math 541A at USC taught by Steven Heilman.

**Need to cover:**

- Update rules for optimization problems (e.g. gradient descent, be able to write down gradient, etc.)
- Know which algorithms are useful in which settings
- Homework-like problems from first part of class (no proofs though) (Boyd homework is good practice)
- Understand how to derive algorithms
- Understand how to calculate gradients, proximal functions, etc.
- Understand examples, how to run algorithms
- Only conceptual thing: duality question (write down dual)
- Formulate problems as convex optimization problems

**Do not need to cover:**

- ADMM
- Proofs from 2nd half of class (rates of convergence, etc.)
- Coding

### 9.1 Convex Functions

**Definition 9.1 (Math 541A definition).** Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . We say that  $\phi$  is **convex** if, for any  $x, y \in \mathbb{R}$  and for any  $t \in [0, 1]$ , we have

$$\phi(tx + (1 - t)y) \leq t\phi(x) + (1 - t)\phi(y).$$

**Definition 9.2 (Strict convexity, Math 541A notes definition 6.6).** Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . We say that  $\phi$  is **strictly convex** if, for any  $x, y \in \mathbb{R}, x \neq y$  and for any  $t \in (0, 1)$ , we have

$$\phi(tx + (1 - t)y) < t\phi(x) + (1 - t)\phi(y).$$

**Definition 9.3 (Convex function in  $\mathbb{R}^n$ , Math 541A Definition).** Let  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ . We say that  $\phi$  is **convex** if, for any  $x, y \in \mathbb{R}^n$  and for any  $t \in [0, 1]$ , we have

$$\phi(tx + (1 - t)y) \leq t\phi(x) + (1 - t)\phi(y). \tag{9.1}$$

**Lemma 9.1.1** (Result from Math 541A Homework 2). The slope of a convex function is nondecreasing. More formally, let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be a convex function. For any  $x \in \mathbb{R}$ , let

$$M_R := \left\{ \frac{\phi(c) - \phi(x)}{c - x} : c > x \right\}, \quad M_L := \left\{ \frac{\phi(x) - \phi(b)}{x - b} : b < x \right\}$$

be the slopes of the secant lines through  $\phi$  using points to the right and left of  $x$ , respectively. Then for any  $m \in M_R$ ,  $p \in M_L$  we have  $m \geq p$ .

*Proof.* Fix  $x \in \mathbb{R}$ . Let  $m \in M_R, p \in M_L$ . By definition, there exist  $b < x < c$  such that

$$m = \frac{\phi(c) - \phi(x)}{c - x}, \quad p = \frac{\phi(x) - \phi(b)}{x - b}.$$

Let  $t \in (0, 1)$  such that

$$tb + (1 - t)c = x. \tag{9.2}$$

Then we have

$$\begin{aligned} m \geq p &\iff \frac{\phi(c) - \phi(x)}{c - x} \geq \frac{\phi(x) - \phi(b)}{x - b} \iff (x - b)(\phi(c) - \phi(x)) \geq (c - x)(\phi(x) - \phi(b)) \\ &\iff (x - b)\phi(c) + b\phi(x) \geq c\phi(x) - (c - x)\phi(b) \end{aligned}$$

From (9.2), we have  $x - b = tb + (1 - t)c - b = (t - 1)b + (1 - t)c = (1 - t)(c - b)$  and  $t(b - c) = x - c \iff t(c - b) = c - x$ . Therefore

$$\begin{aligned} (x - b)\phi(c) + b\phi(x) \geq c\phi(x) - (c - x)\phi(b) &\iff (1 - t)(c - b)\phi(c) + b\phi(x) \geq c\phi(x) - t(c - b)\phi(b) \\ &\iff (1 - t)(c - b)\phi(c) \geq (c - b)\phi(x) - t(c - b)\phi(b) \iff (1 - t)\phi(c) + t\phi(b) \geq \phi(x) \end{aligned}$$

But  $t\phi(b) + (1 - t)\phi(c) \geq \phi(x)$  since  $\phi$  is convex. Therefore  $m \geq p$ .

□

**Theorem 9.1.2 (Result from 541A Homework 2; equivalent conditions for convexity).** Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . Then  $\phi$  is convex if and only if: for any  $y \in \mathbb{R}$ , there exists a constant  $a$  and there exists a function  $L : \mathbb{R} \rightarrow \mathbb{R}$  defined by  $L(x) = a(x - y) + \phi(y)$ ,  $x \in \mathbb{R}$ , such that  $L(y) = \phi(y)$  and such that  $L(x) \leq \phi(x)$  for all  $x \in \mathbb{R}$ . (In the case that  $\phi$  is differentiable, the latter condition says that  $\phi$  lies above all of its tangent lines.)

*Proof.*  $\implies$ : As in Lemma 9.1.1, let

$$M_R := \left\{ \frac{\phi(c) - \phi(y)}{c - y} : c > y \right\}, \quad M_L := \left\{ \frac{\phi(y) - \phi(b)}{y - b} : b < y \right\}$$

be the slopes of the secant lines through  $\phi$  using points to the right and left of  $y$ , respectively. Then by Lemma 9.1.1, for any  $m \in M_R$ ,  $p \in M_L$  we have  $m \geq p$ , so we can choose some  $a_0 \in \mathbb{R}$  such that  $p \leq a_0 \leq m$  for all  $p \in M_L, m \in M_R$ . Then let  $L : \mathbb{R} \rightarrow \mathbb{R}$  be defined by  $L(x) = a_0(x - y) + \phi(y)$ ,  $x \in \mathbb{R}$ . Note that  $L(y) = \phi(y)$ .

We argue that  $L(x) \leq \phi(x)$  for all  $x \in \mathbb{R}$  by contradiction. Suppose there is some  $z \in \mathbb{R}$  with  $L(z) > \phi(z)$ . Note that  $z \neq y$  because we have already shown that  $L(y) = \phi(y)$ . Then we have

$$L(z) > \phi(z) \iff a_0(z - y) + \phi(y) > \phi(z) \quad (9.3)$$

If  $z > y$ , then we can solve (9.3) for  $a_0$  as follows:

$$a_0 > \frac{\phi(z) - \phi(y)}{z - y}$$

But  $z \in M_R$ , so we have  $\frac{\phi(z) - \phi(y)}{z - y} > a_0$ . Contradiction. If  $z < y$ , we solve (9.3) for  $a_0$  as follows:

$$a_0 < \frac{\phi(z) - \phi(y)}{z - y} = \frac{\phi(y) - \phi(z)}{y - z}$$

But  $z \in M_L$ , so we have  $\frac{\phi(y) - \phi(z)}{y - z} < a_0$ . Contradiction. Therefore for every  $z \in \mathbb{R}$  we have  $L(z) \leq \phi(z)$  as desired.

$\Leftarrow$  : Now suppose that for any  $y \in \mathbb{R}$  there exists a constant  $a$  and a function  $L : \mathbb{R} \rightarrow \mathbb{R}$  defined by  $L(x) = a(x - y) + \phi(y)$ ,  $x \in \mathbb{R}$ , such that  $L(y) = \phi(y)$  and such that  $L(x) \leq \phi(x)$  for all  $x \in \mathbb{R}$ .

Fix  $b, c \in \mathbb{R}$  and let  $t \in (0, 1)$ . Set  $y := tb + (1 - t)c$ . Then by assumption we can write

$$a(b - y) + \phi(y) \leq \phi(b), \quad a(c - y) + \phi(y) \leq \phi(c)$$

Multiply by  $t > 0$  and  $(1 - t) > 0$  respectively to yield

$$ta(b - y) + t\phi(y) \leq t\phi(b), \quad (1 - t)a(c - y) + (1 - t)\phi(y) \leq (1 - t)\phi(c) \quad (9.4)$$

Note that

$$ta(b - y) + (1 - t)a(c - y) = a(tb - ty + (c - y - ct + yt)) = a(tb + c - y - ct)$$

$$= a(tb + c(1-t)) - tb - (1-t)c = 0$$

So adding the inequalities in (9.4) yields

$$\begin{aligned} t\phi(y) + (1-t)\phi(y) &\leq t\phi(b) + (1-t)\phi(c) \iff \phi(y) \leq t\phi(b) + (1-t)\phi(c) \\ &\iff \phi(tb + (1-t)c) \leq t\phi(b) + (1-t)\phi(c). \end{aligned}$$

□

*My original proof from submitted homework.* If  $\phi$  is differentiable at  $y$ , let  $a = \phi'(y)$ . If not, let  $a$  be any subgradient of  $\phi$  at  $y$ . Then  $L(x)$  is (a) tangent line to  $\phi$  at  $y$ , which should be lesser than or equal to  $\phi$  for all  $x \in \mathbb{R}$  if  $\phi$  is convex. If and only if this is true at every  $y \in \mathbb{R}$  (the tangent line is a global underestimator at  $y$  for every  $y \in \mathbb{R}$ ), then  $\phi$  must be convex. We proceed to show this formally:

$\implies$  : We will show that if  $\phi$  is convex; that is, if for any  $x, y \in \mathbb{R}$  and for any  $t \in [0, 1]$ , we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y) \quad (9.5)$$

then the inequality

$$\phi'(y)(x - y) + \phi(y) \leq \phi(x) \quad \forall x \in \mathbb{R} \quad (9.6)$$

holds. Starting from (9.5) note that

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y) \implies \phi(tx + (1-t)y) - \phi(x) \leq (1-t)(\phi(y) - \phi(x))$$

Suppose  $y > x$ . Then  $tx + (1-t)y - x = (1-t)(y - x) > 0$ , so we can divide by it on both sides:

$$\implies \frac{\phi(tx + (1-t)y) - \phi(x)}{tx + (1-t)y - x} \leq \frac{(1-t)(\phi(y) - \phi(x))}{(1-t)(y - x)} \implies \frac{\phi(tx + (1-t)y) - \phi(x)}{tx + (1-t)y - x} \leq \frac{\phi(y) - \phi(x)}{y - x}$$

Taking the limit as  $t \rightarrow 1$  yields

$$\phi'(x) \leq \frac{\phi(y) - \phi(x)}{y - x}$$

if  $\phi$  is differentiable, which is (equivalent to) what we hoped to prove. The case where  $x > y$  is analogous.

$\iff$  : We will show that if (9.6) holds then  $\phi$  is convex; that is, (9.5) holds for any  $x, y \in \mathbb{R}$  and for any  $t \in [0, 1]$ . Starting from (9.6) note that

$$\phi'(y)(x - y) + \phi(y) \leq \phi(x) \iff \phi'(y) \leq \frac{\phi(x) - \phi(y)}{x - y} \quad \forall x, y \in \mathbb{R}$$

□

**Theorem 9.1.3 (Global minimum of convex functions; Math 541A Homework problem).** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a convex function. Let  $x \in \mathbb{R}^n$  be a local minimum of  $f$ . Then

- (a)  $x$  is a global minimum of  $f$ .
- (b) If  $f$  is strictly convex, then there is at most one global minimum of  $f$ .
- (c) If  $f$  is a  $C^1$  function (all derivatives of  $f$  exist and are continuous), and  $x \in \mathbb{R}^n$  satisfies  $\nabla f(x) = 0$ , then  $x$  is a global minimum of  $f$ .

*Proof.* (a) Since  $x$  is a local minimum, we have that there exists  $\epsilon > 0$  such that  $f(x) \leq f(y) \forall y \in B(x, \epsilon)$  where  $B(x, \epsilon) \subseteq \mathbb{R}^n$  is an  $n$ -dimensional  $L_2$  ball of radius  $\epsilon$  centered at  $x$ . Suppose there exists some  $z \in \mathbb{R}^n$  such that  $f(z) < f(x)$ . Then by convexity of  $f$ , for  $t \in [0, 1]$ ,

$$f(tx + (1-t)z) \leq tf(x) + (1-t)f(z) < tf(x) + (1-t)f(x) = f(x)$$

which when  $t = 1$  leads to the contradiction  $f(x) < f(x)$ . (Also, for  $t = 1 - \delta$  with  $\delta$  sufficiently small, we get  $f(x') \leq tf(x) + (1-t)f(z) < f(x)$  where  $x' = tx + (1-t)z$  such that  $x' \in B(x, \epsilon)$ , contradicting the fact that  $f(x)$  is a local minimum.) Therefore there is no  $z \in \mathbb{R}^n$  such that  $f(z) < f(x)$ , so  $x$  is a global minimum.

- (b) If  $f$  is strictly convex, for any  $z \in \{\mathbb{R}^n \setminus x\}$  we have

$$f(tx + (1-t)z) < tf(x) + (1-t)f(z), \quad \forall x, z \in \mathbb{R}^n, x \neq z \quad (9.7)$$

We have already shown that there exists no  $z \in \{\mathbb{R}^n \setminus x\}$  such that  $f(z) < f(x)$ . Suppose there is more than one global minimum of  $f$ ; that is, there exists  $z \in \{\mathbb{R}^n \setminus x\}$  such that  $f(z) = f(x)$ . That is, for all  $y \in \{\mathbb{R}^n \setminus \{x, z\}\}$ ,

$$f(x) = f(z) \leq f(y). \quad (9.8)$$

But then by strict convexity,

$$f\left(\frac{x+z}{2}\right) < \frac{1}{2}f(x) + \frac{1}{2}f(z) = \frac{1}{2}f(x) + \frac{1}{2}f(x) = f(x)$$

which contradicts (9.8) if  $y = (x+z)/2$ . Therefore the global minimum of  $f$  is unique.

- (c) Recall from Exercise 4 in Homework 2 that  $f$  is convex if and only if for any  $x \in \mathbb{R}^n$  there exists a constant  $a \in \mathbb{R}^n$  and a function  $L : \mathbb{R}^n \rightarrow \mathbb{R}$  defined by  $L(y) = a^T(y - x) + f(x)$ ,  $y \in \mathbb{R}^n$  such that  $L(x) = f(x)$  and  $L(y) \leq f(y)$  for all  $y \in \mathbb{R}^n$ . Further, if  $f$  is a  $C^1$  function then this function exists for  $a = \nabla f(x)$ . That is,

$$f(y) \geq f(x) + \nabla f^T(x)(y - x), \quad \forall y \in \mathbb{R}^n.$$

Since  $\nabla f(x) = 0$ , if we plug in  $y = x$  we get

$$f(y) \geq f(x), \quad \forall y \in \mathbb{R}^n.$$

□

**Theorem 9.1.4 (Jensen's Inequality, from Math 541A).** Let  $X : \Omega \rightarrow [-\infty, \infty]$  be a random variable. Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be convex. Assume that  $\mathbb{E}|X| < \infty$  and  $\mathbb{E}|\phi(X)| < \infty$ . Then

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

*Proof.* Note that from Theorem 9.1.2, for any  $y \in \mathbb{R}$  there exists a constant  $a$  and a function  $L$  such that

$$a(x - y) + \phi(y) \leq \phi(x) \quad \forall x \in \mathbb{R}$$

Letting  $y = \mathbb{E}(X)$  we have

$$a(X - \mathbb{E}X) + \phi(\mathbb{E}X) \leq \phi(X)$$

Since expectations preserve inequalities,

$$\mathbb{E}[a(X - \mathbb{E}X) + \phi(\mathbb{E}X)] \leq \mathbb{E}\phi(X)$$

But

$$\mathbb{E}[a(X - \mathbb{E}X) + \phi(\mathbb{E}X)] = a(\mathbb{E}X - \mathbb{E}X) + \mathbb{E}(\phi(\mathbb{E}X)) = \phi(\mathbb{E}X)$$

which yields

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

□

**Corollary 9.1.4.1 (Jensen's Inequality: EE 588 Formulation).**  $f$  is convex if and only if

$$f\left(\frac{a+b}{2}\right) \leq \frac{f(a) + f(b)}{2}$$

for all  $a, b \in \text{dom}(f)$ .

*Proof.* Follows from Theorem 9.1.4 if  $X$  is a discrete random variable that equals  $a$  or  $b$  each with probability  $1/2$  and  $\phi(X) = f(X)$ . Note that  $\phi(X)$  is convex.

□

**Corollary 9.1.4.2 (Triangle Inequality).** Let  $X : \Omega \rightarrow [-\infty, \infty]$  be a random variable with  $\mathbb{E}|X| < \infty$ . Then

$$|\mathbb{E}X| \leq \mathbb{E}|X|.$$

*Proof.* Note that  $\phi(x) = |x|$  is convex by the definition of convexity: for any  $x, y \in \mathbb{R}$  and for any  $t \in (0, 1)$ , we have

$$\phi(tx + (1-t)y) = |tx + (1-t)y| \leq \dots = t|x| + (1-t)|y| = t\phi(x) + (1-t)\phi(y).$$

Then the result follows immediately from Jensen's Inequality (Theorem 9.1.4) using  $\phi(X) = |X|$ :

$$|\mathbb{E}X| \leq \mathbb{E}|X|$$

□

**Theorem 9.1.5 (Conditional Jensen Inequality).** Let  $X, Y : \Omega \rightarrow \mathbb{R}$  be random variables that are either both discrete or both continuous. Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be convex. Then

$$\phi(\mathbb{E}(X|Y)) \leq \mathbb{E}(\phi(X)|Y).$$

If  $\phi$  is strictly convex, then equality holds only if  $X$  is constant on any set where  $Y$  is constant. That is, (by an Exercise from the previous homework) equality holds only if  $X$  is a function of  $Y$ .

*Proof.* Recall that from Exercise 4 in Homework 2 that since  $\phi$  is convex, for any  $y \in \mathbb{R}$  there exists a constant  $a$  and a function  $L$  such that

$$a(x - y) + \phi(y) \leq \phi(x) \quad \forall x \in \mathbb{R}$$

Letting  $x = X$  and  $y = \mathbb{E}(X | Y)$  we have

$$a(X - \mathbb{E}(X | Y)) + \phi(\mathbb{E}(X | Y)) \leq \phi(X)$$

Since by Lemma 6.1.7 conditional expectations preserve inequalities,

$$\mathbb{E}[a(X - \mathbb{E}[X | Y]) + \phi(\mathbb{E}[X | Y]) | Y] \leq \mathbb{E}(\phi(X) | Y)$$

But

$$\mathbb{E}[a(X - \mathbb{E}[X | Y]) + \phi(\mathbb{E}[X | Y]) | Y] = a(\mathbb{E}[X | Y] - \mathbb{E}[\mathbb{E}(X | Y) | Y]) + \mathbb{E}[\phi(\mathbb{E}[X | Y]) | Y].$$

By Corollary 6.1.6.1 (letting  $h(Y) = \phi(\mathbb{E}[X | Y])$ ),  $\mathbb{E}[\phi(\mathbb{E}[X | Y]) | Y] = \phi(\mathbb{E}[X | Y])$ . By Corollary 6.1.6.2,  $\mathbb{E}[\mathbb{E}(X | Y) | Y] = \mathbb{E}(X | Y)$ . Therefore we have

$$= a(\mathbb{E}[X | Y] - \mathbb{E}[X | Y]) + \phi(\mathbb{E}[X | Y]) = \phi(\mathbb{E}(X | Y))$$

which yields

$$\phi(\mathbb{E}(X \mid Y)) \leq \mathbb{E}(\phi(X) \mid Y).$$

□

**Proposition 9.1.6 (Convexity of affine functions).** Let  $c \in \mathbb{R}^n$  and  $d \in \mathbb{R}$  be fixed. Let  $x \in \mathbb{R}^n$ . Then the function  $f(x) = c^T x + d$  is convex.

*Proof.* We will show that  $f$  satisfies (9.1) for any  $x, y \in \mathbb{R}^n$  and any  $t \in [0, 1]$ :

$$\begin{aligned} f(tx + (1-t)y) &= c^T(tx + (1-t)y) + d = c^T(tx + (1-t)y) + d = tc^T x + td + (1-t)c^T y + (1-t)d \\ &= t[c^T x + d] + (1-t)[c^T y + d] = tf(x) + (1-t)f(y). \end{aligned}$$

In particular, (9.1) is satisfied with equality.

□

**Proposition 9.1.7 (Convexity of quadratic forms).** Let  $A \in \mathbb{R}^{m \times n}$  and  $k > 0$  be fixed. Let  $x \in \mathbb{R}^n$ . Then the function  $f(x) = kx^T A^T A x$  is convex.

*Proof.* We will show that  $f$  satisfies the definition of convexity. Plugging  $\phi(x) = kx^T A^T A x$  into the left side of (9.1), we have

$$\begin{aligned} k(tx + (1-t)y)^T A^T A(tx + (1-t)y) &= k(tx^T + (1-t)y^T)A^T A(tx + (1-t)y) \\ &= k[tx^T A^T A tx + tx^T A^T A(1-t)y + (1-t)y^T A^T A tx + (1-t)y^T A^T A(1-t)y] \\ &= k[t^2 x^T A^T A x + t(1-t)x^T A^T A y + t(1-t)y^T A^T A x + (1-t)^2 y^T A^T A y] \end{aligned}$$

Note that  $x^T A^T A y \in \mathbb{R} = [x^T A^T A y]^T = y^T A^T A x$ , so we have

$$= k[t^2 x^T A^T A x + 2t(1-t)x^T A^T A y + (1-t)^2 y^T A^T A y] \quad (9.9)$$

Plugging  $\phi(x) = kx^T A^T A x$  into the right side of (9.1) yields

$$ktx^T A^T A x + k(1-t)y^T A^T A y \quad (9.10)$$

We can verify the inequality in (9.1) by subtracting (9.10) from (9.9) to see if a negative number results:

$$\begin{aligned}
& kt^2 x^T A^T A x + 2kt(1-t)x^T A^T A y + k(1-t)^2 y^T A^T A y - ktx^T A^T A x - k(1-t)y^T A^T A y \\
&= kt(t-1)x^T A^T A x + 2kt(1-t)x^T A^T A y + k(1-t)[1-t-1]y^T A^T A y \\
&= -kt(1-t)[x^T A^T A x - 2x^T A^T A y + y^T A^T A y] = -kt(1-t)[x^T A^T - y^T A^T][A x - A y] \\
&= -kt(1-t)[A(x-y)]^T A(x-y) \leq 0
\end{aligned}$$

for all  $x, y \in \mathbb{R}^n$  and any  $t \in [0, 1]$  since  $-kt(1-t) \leq 0$  (with equality only when  $t = 0$  or  $t = 1$ ) and  $[A(x-y)]^T A(x-y) \geq 0$  (with equality only when  $x = y$ ). This verifies the inequality in (9.1), which proves that  $kx^T A^T A x$  is convex.

□

**Proposition 9.1.8 (Sum of convex functions is convex).** Let  $f_1, \dots, f_n : \mathbb{R}^n \rightarrow \mathbb{R}$  be (strictly) convex functions. Then the function  $g(x) := \sum_{i=1}^n f_i(x)$  is (strictly) convex.

*Proof.* Since  $f_i$  is convex for all  $i \in \{1, \dots, n\}$ ,  $f_i$  satisfies

$$f_i(tx + (1-t)y) \leq t f_i(x) + (1-t) f_i(y), \quad \forall i \in \{1, \dots, n\}.$$

We make use of these inequalities to show that  $g$  satisfies (9.1) for any  $x, y \in \mathbb{R}^n$  and any  $t \in [0, 1]$ :

$$\begin{aligned}
g(tx + (1-t)y) &= \sum_{i=1}^n f_i(tx + (1-t)y) \leq \sum_{i=1}^n [t f_i(x) + (1-t) f_i(y)] \\
&= t \sum_{i=1}^n f_i(x) + (1-t) \sum_{i=1}^n f_i(y) = t g(x) + (1-t) g(y)
\end{aligned}$$

which proves the result. (Note that if the initial inequality is strict then strict convexity follows.)

□

**Proposition 9.1.9 (Exercise 6.43 in Math 541A Lecture Notes).** Let  $f_1, \dots, f_n : \mathbb{R} \rightarrow \mathbb{R}$  be  $n$  strictly convex functions. Define  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  by

$$g(x_1, \dots, x_n) := \sum_{i=1}^n f_i(x_i), \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Then  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex.

*Proof.* Since  $f_i$  is strictly convex for all  $i \in \{1, \dots, n\}$ , we have that for any  $x_i, y_i \in \mathbb{R}$ , for all  $t \in (0, 1)$

$$f_i(tx_i + (1-t)y_i) < tf_i(x_i) + (1-t)f_i(y_i).$$

Therefore for any  $x, y \in \mathbb{R}^n$  (where  $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n)$ ), for all  $t \in (0, 1)$

$$\begin{aligned} g(tx + (1-t)y) &= \sum_{i=1}^n f_i(tx_i + (1-t)y_i) < \sum_{i=1}^n [tf_i(x_i) + (1-t)f_i(y_i)] = t \sum_{i=1}^n f_i(x_i) + (1-t) \sum_{i=1}^n f_i(y_i) \\ &= tg(x) + (1-t)g(y). \end{aligned}$$

□

**Proposition 9.1.10 (Exercise 6.44 from Math 541A lecture notes).** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Suppose that for any fixed  $i \in \{1, \dots, n\}$  and for any  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ , the function

$$x_i \mapsto f(x_1, \dots, x_n)$$

is strictly convex. Then  $f$  has at most one global minimum.

*Proof.* An equivalent statement to our assumption is that for any  $i$ ,  $f$  is strictly convex in  $x_i$  keeping  $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$  fixed. That is, if we let  $h_i : \mathbb{R} \rightarrow \mathbb{R}$  be defined for all  $i \in \{1, \dots, n\}$  by

$$h_i(x_i \mid (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)) := f((x_1, \dots, x_n)) \quad \forall i \in \{1, \dots, n\},$$

then  $h_i$  is strictly convex for all  $i \in \{1, \dots, n\}$ . That is, for any  $x_i, y_i \in \mathbb{R}$ , for all  $t \in (0, 1)$

$$\begin{aligned} h_i(tx_i + (1-t)y_i \mid (tx_1 + (1-t)y_1, \dots, tx_{i-1} + (1-t)y_{i-1}, tx_{i+1} + (1-t)y_{i+1}, \dots, tx_n + (1-t)y_n)) \\ < th_i(x_i \mid (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)) + (1-t)h_i(y_i \mid (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)). \end{aligned} \tag{9.11}$$

By Theorem 9.1.3(b), there is at most one global minimum of  $f$  if it is strictly convex, so all we need to show is that  $f$  is strictly convex. That is, we must show that for all  $(x_1, \dots, x_n), (y_1, \dots, y_n) \in \mathbb{R}^n$ , for all  $t \in (0, 1)$ ,

$$f(t(x_1, \dots, x_n) + (1-t)(y_1, \dots, y_n)) < tf((x_1, \dots, x_n)) + (1-t)f((y_1, \dots, y_n)). \tag{9.12}$$

We will argue by contradiction. Suppose that for some  $(x_1^*, \dots, x_n^*) \in \mathbb{R}^n$  and  $(y_1^*, \dots, y_n^*) \in \mathbb{R}^n$ , (9.12) does not hold. That is,

$$f(t^*(x_1^*, \dots, x_n^*) + (1-t^*)(y_1^*, \dots, y_n^*)) \geq t^*f((x_1^*, \dots, x_n^*)) + (1-t^*)f((y_1^*, \dots, y_n^*)).$$

for some  $t^* \in (0, 1)$ . This is equivalent to

$$\begin{aligned} h_1(t^*x_1^* + (1-t^*)y_1^* | (t^*x_2^* + (1-t^*)y_2^*, \dots, t^*x_n^* + (1-t^*)y_n^*)) \\ \geq t^*h_1(x_1^* | (x_2^*, \dots, x_n^*)) + (1-t^*)h_1(y_1^* | (y_2^*, \dots, y_n^*)). \end{aligned}$$

But this contradicts (9.11). Therefore (9.12) holds for all  $(x_1, \dots, x_n), (y_1, \dots, y_n) \in \mathbb{R}^n$ , for all  $t \in (0, 1)$ , so  $f$  is strictly convex, which means (by Theorem 9.1.3(b)) that  $f$  has at most one global minimum.

□

**Proposition 9.1.11.** Let  $A$  be a real  $m \times n$  matrix. Let  $x \in \mathbb{R}^n$  and let  $b \in \mathbb{R}^m$ . Then the function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by  $f(x) = \frac{1}{2}\|Ax - b\|^2$  is convex.

*Proof.* We have

$$\begin{aligned} f(x) &= \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}(Ax - b)^T(Ax - b) = \frac{1}{2}(x^T A^T - b^T)(Ax - b) \\ &= \frac{1}{2}(x^T A^T Ax - b^T Ax - x^T A^T b + b^T b) = \frac{1}{2}x^T A^T Ax - b^T Ax + \frac{1}{2}b^T b \end{aligned}$$

where the last step follows because  $b^T Ax \in \mathbb{R} = (b^T Ax)^T = x^T A^T b$  since a real number equals its transpose. The affine function  $-b^T Ax + \frac{1}{2}b^T b$  is convex by Proposition 9.1.6, and the quadratic form  $\frac{1}{2}x^T A^T Ax$  is convex by Proposition 9.1.7. Since the sum of convex functions is convex by Proposition 9.1.8, the result follows.

□

**Remark.** Moreover,

$$\nabla f(x) = A^T(Ax - b), \quad D^2f(x) = A^T A.$$

(Here  $D^2f$  denotes the matrix of second derivatives of  $f$ .)

So, if  $\nabla f(x) = 0$ , i.e. if  $A^T Ax = A^T b$ , then  $x$  is the global minimum of  $f$ . And if  $A$  has full rank, then  $A^T A$  is invertible, so that  $x = (A^T A)^{-1}A^T b$  is the global minimum of  $f$ .

**Proposition 9.1.12 (Convexity of norms).** Every norm on  $\mathbb{R}^n$  is convex.

*Proof.* Suppose we have a generic norm  $\|\cdot\|_*$  in  $\mathbb{R}^n$ . Because  $\|\cdot\|_*$  is a norm, it satisfies the triangle inequality; that is, for all  $x, y \in \mathbb{R}^n$ ,  $\|x + y\|_* \leq \|x\|_* + \|y\|_*$ . Further, for any  $t \in [0, 1]$ , we have

$$\|tx + (1-t)y\|_* \leq \|tx\|_* + \|(1-t)y\|_* = t\|x\|_* + (1-t)\|y\|_*$$

where the last step also follows from a property of all norms.

□

**Proposition 9.1.13 (Mentioned in in-class 541A review; might have been on HW?).** If  $\phi$  is strictly convex and  $\mathbb{E}(\phi(X)) = \phi(\mathbb{E}(X))$  then  $X$  is almost surely constant.

**Proposition 9.1.14 (2018 DSO Statistics Group In-Class Screening Exam, Question 5).** The function  $f(\theta) = (\|\theta\|_1)^2$  is convex.

*Proof.* Note that

$$\|tx + (1-t)y\|_1 \leq \|tx\|_1 + \|(1-t)y\|_1 = t\|x\|_1 + (1-t)\|y\|_1 \quad (9.13)$$

where the first step follows by the Triangle Inequality (which all norms satisfy, including the  $\ell_1$  norm) and the second step follows by the homogeneity property of norms. Therefore  $\|\theta\|_1$  is convex. Next, by (9.13) and the monotonicity of  $g(\theta) = \theta^2$  when  $\theta \geq 0$ ,

$$\begin{aligned} f(tx + (1-t)y) &= (\|tx + (1-t)y\|_1)^2 \leq (t\|x\|_1 + (1-t)\|y\|_1)^2 \\ &= t^2\|x\|_1^2 + (1-t)^2\|y\|_1^2 + 2t(1-t)\|x\|_1\|y\|_1 \end{aligned}$$

and

$$tf(x) + (1-t)f(y) = t\|x\|_1^2 + (1-t)\|y\|_1^2$$

Taking the difference of these yields

$$\begin{aligned} tf(x) + (1-t)f(y) - f(tx + (1-t)y) &\geq t\|x\|_1^2 + (1-t)\|y\|_1^2 - (t^2\|x\|_1^2 + (1-t)^2\|y\|_1^2 + 2t(1-t)\|x\|_1\|y\|_1) \\ &= (t - t^2)\|x\|_1^2 + [(1-t) - (1-t)^2]\|y\|_1^2 - 2t(1-t)\|x\|_1\|y\|_1 \\ &= (t - t^2)(\|x\|_1^2 + \|y\|_1^2 - 2\|x\|_1\|y\|_1) = t(1-t)(\|x\|_1 - \|y\|_1)^2 \geq 0 \\ \iff tf(x) + (1-t)f(y) &\geq f(tx + (1-t)y) \end{aligned}$$

which proves convexity. □

## 9.2 Schur Complement Trick

### 9.2.1 Definition

For a matrix  $X \in \mathbf{S}^n$  partitioned as

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

the Schur complement is (if  $\det(A) \neq 0$ )

$$S = C - B^T A^{-1} B$$

The Schur complement has two useful properties in convex analysis.

**Theorem 9.2.1.** (a)  $X \succ 0$  if and only if  $A \succ 0$  and  $S \succ 0$ .

(b) If  $A \succ 0$ , then  $X \succeq 0$  if and only if  $S \succeq 0$ .

### 9.2.2 The Trick

Suppose we are trying to express a problem as a semidefinite program (SDP); that is, in the form

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{subject to} \quad & x_1 F_1 + \dots + x_n F_n + G \preceq 0 \\ & Ax = b \end{aligned}$$

where  $G, F_1, \dots, F_n \in \mathbf{S}^k$  and  $A \in \mathbb{R}^{p \times n}$ . If we have a constraint of the form  $c^T F(x)^{-1} c \leq t$  where  $F(x)$  is symmetric and positive definite and  $t \in \mathbb{R}$ , by Theorem 9.2.1(b) we can write

$$c^T F(x)^{-1} c \leq t \iff \begin{bmatrix} F(x) & c \\ c^T & t \end{bmatrix} \succeq 0$$

in order to get our constraint in the form required for an SDP.

### 9.2.3 Example 1: Last Year's Final, Question 2(b)

Suppose we have the constraints

$$\begin{aligned} Ax + b &\geq 0 \\ \frac{(c^T x)^2}{d^T x} &\leq t \end{aligned}$$

which we would like to express in an SDP. By Theorem 9.2.1(b) we can write

$$\frac{(c^T x)^2}{d^T x} \leq t \iff d^T x - (c^T x)^T t^{-1} c^T x \geq 0 \iff \begin{bmatrix} t & c^T x \\ c^T x & d^T x \end{bmatrix} \succeq 0$$

Since

$$Ax + b \geq 0 \iff \mathbf{diag}(Ax + b) \succeq 0$$

we can finally write our constraints as

$$\begin{bmatrix} \mathbf{diag}(Ax + b) & 0 & 0 \\ 0 & t & c^T x \\ 0 & c^T x & d^T x \end{bmatrix} \succeq 0$$

#### 9.2.4 Example 2: Last Year's Final, Question 4(b)

Suppose we have the constraints

$$\begin{aligned} Ax + b &\geq 0 \\ \frac{(c^T x)^2}{d^T x} &\leq t \end{aligned}$$

which we would like to express in an SDP. By Theorem 9.2.1(b) we can write

$$\frac{(c^T x)^2}{d^T x} \leq t \iff d^T x - (c^T x)^T t^{-1} c^T x \geq 0 \iff \begin{bmatrix} t & c^T x \\ c^T x & d^T x \end{bmatrix} \succeq 0$$

Since

$$Ax + b \geq 0 \iff \mathbf{diag}(Ax + b) \succeq 0$$

we can finally write our constraints as

$$\begin{bmatrix} \mathbf{diag}(Ax + b) & 0 & 0 \\ 0 & t & c^T x \\ 0 & c^T x & d^T x \end{bmatrix} \succeq 0$$

### 9.3 Duality

**Theorem 9.3.1.** Slater's condition/constraint qualification: Strong duality holds for a convex problem

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \leq 0, i = 1, \dots, m \\ & && Ax = b \end{aligned}$$

if it is strictly feasible, i.e., there exists at least one  $x$  in the domain of  $f_0$  such that  $f_i(x) < 0$ ,  $i = 1, 2, \dots, m$ ,  $Ax = b$ .

## 9.4 MLE estimates

For linear estimates with iid noise

$$y_i = a_i^T x + v_i, i = 1, \dots, m$$

where  $a$  is observed and  $x \in \mathbb{R}^n$  are the parameters to be estimated, the likelihood function is

$$p_x(y) = \prod_{i=1}^m \Pr(v_i = y_i - a_i^T x \mid x)$$

Therefore the log likelihood function is:

$$\ell_x(y) = \sum_{i=1}^m \log[\Pr(v_i = y_i - a_i^T x \mid x)]$$

## 9.5 Practice Final (2017 Final)

- (1) (a) Strictly convex. Multiply by  $x/x$  (allowed in this case since  $x > 0$ ) to get  $\frac{x^2}{x+1}$  which is a quadratic over linear, which is convex in  $\mathbb{R}^{++}$  according to CVX rules.
- (b) Not convex, it is convex for  $x \geq -1$ , but there is a boundary problem at  $x = -1$ . Note that Jensen's inequality (Theorem 9.1.4.1)

$$\frac{f(a) + f(b)}{2} \geq f\left(\frac{a+b}{2}\right)$$

is violated because

$$\frac{f(-1.3) + f(-0.9)}{2} = \frac{2.3 + 0}{2} = 1.15 \leq 2.2 = f(-1.1) = f\left(\frac{-1.3 + -0.9}{2}\right)$$

(c)

(d)

$$f(x) = \sup \log \left( \frac{p(t)}{q(t)} \right) = \sup \{ \log p(t) - \log q(t) \} = \sup \{ \log \left( \sum_{i=1}^n \exp(x_i \sin(it)) \right) - \sum_{i=1}^n x_i \sin(it) \}$$

(e) The proximal mapping is

$$\begin{aligned} \text{prox}_{\mathcal{R}}(z) &= \arg \min_y \frac{1}{2} \|z - y\|_2^2 + \mathcal{R}(y) = \arg \min_y \frac{1}{2} \sum_{i=1}^n (z_i - y_i)^2 + \sum_{i=1}^n w_i |y_i| \\ &= \arg \min_y \frac{1}{2} \sum_{i=1}^n [(z_i - y_i)^2 + w_i |y_i|] \end{aligned}$$

Taking the gradient of the inside quantity with respect to  $y$ , we have

$$\nabla(y) = \begin{pmatrix} \frac{1}{2} \cdot 2(z_1 - y_1) + \text{sign}(y_1)w_1 \\ \frac{1}{2} \cdot 2(z_2 - y_2) + \text{sign}(y_2)w_2 \\ \vdots \\ \frac{1}{2} \cdot 2(z_n - y_n) + \text{sign}(y_n)w_n \end{pmatrix} = \begin{pmatrix} z_1 - y_1 + \text{sign}(y_1)w_1 \\ z_2 - y_2 + \text{sign}(y_2)w_2 \\ \vdots \\ z_n - y_n + \text{sign}(y_n)w_n \end{pmatrix}$$

Setting equal to 0, we have

$$y = \begin{pmatrix} z_1 \pm w_1 \\ z_2 \pm w_2 \\ \vdots \\ z_n \pm w_n \end{pmatrix}$$

(2) (a) The constraint is convex (affine). The denominator is affine. Since  $c^T x = x^T c$ , the numerator

$$(c^T x)^2 = (c^T x)(c^T x) = x^T c c^T x = x^T (cc^T)x$$

is convex since  $cc^T$  is positive semidefinite.

(b) We start by using the epigraph trick to transform the problem:

$$\begin{aligned} & \text{minimize} && t \\ & \text{subject to} && \frac{(c^T x)^2}{d^T x} \leq t \\ & && Ax + b \geq 0 \end{aligned}$$

We are trying to express this problem as a semidefinite program (SDP); that is, in the form

$$\begin{aligned} & \text{minimize} && c^T x \\ & \text{subject to} && x_1 F_1 + \dots + x_n F_n + G \preceq 0 \\ & && Ax = b \end{aligned}$$

where  $G, F_1, \dots, F_n \in \mathbf{S}^k$  and  $A \in \mathbb{R}^{p \times n}$ . The first constraint

$$\frac{(c^T x)^2}{d^T x} \leq t$$

can be expressed in the form

$$(c^T x)^2 \leq t d^T x \iff (c^T x c^T - t d^T)x \leq 0$$

We have a constraint

$$Ax + b \geq 0$$

which can be expressed in the form

$$Ax \geq -b$$

$$c^T F(x)^{-1} c \leq t$$

where  $F(x)$  is symmetric and positive definite and  $t \in \mathbb{R}$ , by Theorem 9.2.1(b) we can write

$$c^T F(x)^{-1} c \leq t \iff \begin{bmatrix} F(x) & c \\ c^T & t \end{bmatrix} \succeq 0$$

in order to get our constraint in the form required for an SDP.

- (3) (a) Yes,  $g$  is convex over  $\mathcal{X}$  since it is quadratic over linear.
- (b) The only points satisfying the constraint have  $x_1 = 0$ . Therefore the primal optimal value (the only feasible value) is  $e^0 = \boxed{1}$ .
- (c) Lagrangian:

$$L(x, \lambda) = e^{-x_1} + \lambda(x_1^2/x_2)$$

The Lagrangian obtains its minimum value of 0 when  $x_2 = x_1^3$  and  $x_1 \rightarrow \infty$ . Thus, its dual function ( $g(\lambda) = \min_x L(x, \lambda)$ ) is

$$g(\lambda) = 0$$

The dual problem is then

maximize	0
subject to	$\lambda \geq 0$

- (d) The optimal value of the dual problem is 0. Strong duality does not hold since the optimum of the dual problem is less than the optimum of the primal problem. We can also tell this because Slater's Condition (Theorem 9.3.1) is violated; that is, there is no  $(x_1, x_2)$  that is strictly feasible since  $x_1$  must equal 0, which is on the boundary of the feasible region.
- (e) Now for the primal problem, instead of  $x_1 = 0$ , we have

$$\frac{x_1^2}{x_2} \leq u \iff x_1^2 \leq ux_2 \implies -\sqrt{ux_2} \leq x_1 \leq \sqrt{ux_2}$$

Since  $e^{-x_1}$  is minimized as  $x_1 \rightarrow \infty$ , our optimal solution is  $x_2 \rightarrow \infty, x_1 = \sqrt{ux_2} \rightarrow \infty$  yielding a primal optimal value of  $\boxed{0}$ . For the dual problem, we have

$$L(x, \lambda) = e^{-x_1} + \lambda \left( \frac{x_1^2}{x_2} - u \right)$$

Dual function ( $g(\lambda) = \min_x L(x, \lambda)$ ):

$$\frac{x_1^2}{x_2} - u = 0 \implies x_2 = \frac{x_1^2}{u}$$

and let  $x_1 \rightarrow -\infty$  to yield

$$g(\lambda) = 0$$

The dual problem is then

maximize	0
----------	---

with optimal value 0, so there is no longer a duality gap. We can also tell this because Slater's Condition (Theorem 9.3.1) is satisfied; that is, there exists an  $(x_1, x_2)$  which is strictly feasible (say  $(x_1, x_2) = (\sqrt{u}, 10)$ ).

(4) (a) Yes, the set is convex. If  $(u_i, v_i) = \mathbf{u}_i$ , each

$$\sqrt{(x - u_i)^2 + (y - v_i)^2} = \|\mathbf{x} - \mathbf{u}_i\|_2$$

is convex in  $\mathbf{x}$ . Therefore the function

$$\sum_{i=1}^k \|\mathbf{x} - \mathbf{u}_i\|_2$$

is convex. For any fixed  $d$ , this set is a sublevel set of this function, which is convex since the function is convex.

(b) This is a feasibility problem:

$$\begin{aligned} & \text{find} && \mathbf{x} \\ & \text{subject to} && \sum_{i=1}^k \|\mathbf{x} - \mathbf{u}_i\| \leq d \\ & && \sum_{i=1}^j \|\mathbf{x} - \mathbf{v}_i\| \leq e \end{aligned}$$

or

$$\begin{aligned} & \text{minimize} && 0 \\ & \text{subject to} && \sum_{i=1}^k \|\mathbf{x} - \mathbf{u}_i\| \leq d \\ & && \sum_{i=1}^j \|\mathbf{x} - \mathbf{v}_i\| \leq e \end{aligned}$$

for two sets of points in  $\mathbb{R}^2$   $\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{v}_1, \dots, \mathbf{v}_j$ . We would like to express these constraints as matrix inequalities in order to have an SDP. To do this, first rewrite the problem as

$$\begin{aligned} & \text{minimize} && 0 \\ & \text{subject to} && \|\mathbf{x} - \mathbf{u}_i\| \leq t_i, i = 1, \dots, k \\ & && \|\mathbf{x} - \mathbf{v}_i\| \leq s_i, s = 1, \dots, j \\ & && \mathbf{1}^T \mathbf{t} \leq d \\ & && \mathbf{1}^T \mathbf{s} \leq e \end{aligned}$$

Then note that we can use the Schur trick:

$$(\mathbf{x} - \mathbf{u}_i)^T I (\mathbf{x} - \mathbf{u}_i) \leq t_i \iff \begin{bmatrix} I & \mathbf{x} - \mathbf{u}_i \\ (\mathbf{x} - \mathbf{u}_i)^T & t_i \end{bmatrix} \succeq 0$$

and write the optimization problem as an SDP:

minimize	0
subject to	$\begin{bmatrix} I & \mathbf{x} - \mathbf{u}_i \\ (\mathbf{x} - \mathbf{u}_i)^T & t_i \end{bmatrix} \succeq 0, i = 1, \dots, k$ $\begin{bmatrix} I & \mathbf{x} - \mathbf{v}_i \\ (\mathbf{x} - \mathbf{v}_i)^T & s_i \end{bmatrix} \succeq 0, s = 1, \dots, j$ $\mathbf{1}^T t \leq d$ $\mathbf{1}^T s \leq e$

(5) (a) To minimize the MSE:

$$\mathcal{L}(z) = \sum_r (y_r - |a_r^T x|^2)^2$$

For MLE estimate:

$$p_x(y) = \prod_{r=1}^m \Pr(w_r = y_r - (a_r^T x)^2 \mid x) = \frac{1}{(y_r - (a_r^T x)^2)!} \cdot \exp(-(a_r^T x)^2) \cdot (a_r^T x)^{2[y_r - (a_r^T x)^2]}$$

Therefore the log likelihood function is:

$$\begin{aligned} \ell_x(y) &= \sum_{i=1}^m \log[\Pr(y_i - a_i^T x \mid x)] = \sum_{i=1}^m \log \left[ \frac{1}{(y_r - (a_r^T x)^2)!} \cdot \exp(-(a_r^T x)^2) \cdot (a_r^T x)^{2[y_r - (a_r^T x)^2]} \right] \\ &= \sum_{i=1}^m \log \left[ \frac{1}{(y_r - (a_r^T x)^2)!} \right] - (a_r^T x)^2 + 2[y_r - (a_r^T x)^2] \cdot \log[(a_r^T x)] \end{aligned}$$

- (b) b
- (c) c
- (d) d
- (e) e

## 10 Mathematical Statistics

These are my notes from taking Math 541A at USC taught by Steven Heilman as well as *Statistical Inference* (2nd edition) by Casella and Berger [Casella and Berger, 2001], Statistics 100B at UCLA taught by Nicolas Christou, ISE 620 at USC taught by Sheldon Ross, Math 505A at USC taught by Sergey Lototsky, and a few other sources I cite within the text.

### 10.1 Order Statistics

**Definition 10.1 (Order statistics (from Math 541A, more precise)).** Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable. Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from  $X$ . Define  $X_{(1)} := \min_{1 \leq i \leq n} X_i$ , and for any  $2 \leq i \leq n$ , inductively define

$$X_i := \min \left\{ \{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(i-1)}\} \right\},$$

so that

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)} = \max_{1 \leq i \leq n} X_i.$$

The random variables  $X_{(1)}, \dots, X_{(n)}$  are called the **order statistics** of  $X_1, \dots, X_n$ .

**Definition 10.2 (Order statistics (from ISE 620, more informal)).** Let  $X_1, \dots, X_n \sim \text{iid } F$  with  $F' = f$ . Define  $X_{(1)}$  as the smallest among  $X_1, \dots, X_n$ ,  $X_{(2)}$  as the 2nd smallest, and so on, up to  $X_{(n)}$ , the largest of the group. We call  $X_{(1)}, \dots, X_{(n)}$  the **order statistics** of  $X_1, \dots, X_n$ .

**Proposition 10.1.1 (Order statistics distribution function; from Math 541A).** Suppose  $X$  is a discrete random variable and we can order the values that  $X$  takes as  $x_1 < x_2 < \dots$ . For any  $i \geq 1$ , define  $p_i := \Pr(X \leq x_i)$ . Then for any  $1 \leq i, j \leq n$ ,

$$\Pr(X_{(j)} \leq x_i) = \sum_{k=j}^n \binom{n}{k} p_i^k (1-p_i)^{n-k}.$$

*Proof.* Note that  $\{X_{(j)} \leq x_i\}$  is equivalent to the event that  $j$  or more of the  $X_i$  are less than or equal to  $x_i$  regardless of order; that is,  $x_i$  is the  $k$ th smallest observed value. Let  $A_k$  be the event that exactly  $k$  of the  $X_i$  are less than or equal to  $x_i$  regardless of order. Then

$$\{X_{(j)} \leq x_i\} = \bigcup_{k=j}^n A_k.$$

Then since (by definition of  $p_i$ )

$$\Pr(A_k) = \binom{n}{k} p_i^k (1-p_i)^{n-k}$$

and using the fact that the  $\{A_k\}$  are disjoint, we have

$$\Pr(\{X_{(j)} \leq x_i\}) = \Pr(\bigcup_{k=j}^n A_k) = \sum_{k=1}^n \Pr(A_k) = \sum_{k=1}^n \binom{n}{k} p_i^k (1-p_i)^{n-k}.$$

□

**Corollary 10.1.1.1.** if  $X$  is a continuous random variable with density  $f_X$  and cumulative distribution function  $F_X$ , then for any  $1 \leq j \leq n$ ,  $F_{X(j)}$  has density

$$f_{X(j)}(x) := \frac{n!}{(j-1)!(n-j)!} f_X(x) (F_X(x))^{j-1} (1 - F_X(x))^{n-j}, \quad \forall x \in \mathbb{R}.$$

*Proof.* This follows by differentiating the identity from Proposition 10.1.1 for the cumulative distribution function.

□

**Proposition 10.1.2 (Order statistics joint density function; result from ISE 620).** The joint density of the order statistics is

$$f_{X_{(1)}, \dots, X_{(n)}}(x_1, \dots, x_n) = n! \prod_{i=1}^n f(x_i).$$

*Proof.* Start with  $n = 2$ . We seek  $f_{X_{(1)}, X_{(2)}}(x_1, x_2)$ . Note that  $X_{(1)} = x_1, X_{(2)} = x_2$  if  $X_1 = x_1, X_2 = x_2$  or if  $X_1 = x_2, X_2 = x_1$ . These are mutually exclusive events, so their density is equal to the sums of the two densities. That is,

$$f_{X_{(1)}, X_{(2)}}(x_1, x_2) = f_{X_1, X_2}(x_1, x_2) + f_{X_1, X_2}(x_2, x_1) = 2f(x_1)f(x_2)$$

where the last step follows from the i.i.d. distributions. Generalizing, we have

$$f_{X_{(1)}, \dots, X_{(n)}}(x_1, \dots, x_n) = n! \prod_{i=1}^n f(x_i).$$

□

**Proposition 10.1.3 (Distribution of order statistics of uniform random variable; from 541A).** Let  $X$  be a random variable uniformly distributed in  $[0, 1]$ . Then for any  $1 \leq j \leq n$ ,  $X_{(j)}$  is a beta distributed random variable with parameters  $j$  and  $n+1-j$ .

*Proof.* Note that for a uniform distribution on  $[0, 1]$ ,  $f_X(x) = 1, x \in [0, 1]$  and  $F_X(x) = x, x \in [0, 1]$ . Therefore by Corollary 10.1.1.1 we have

$$\begin{aligned} f_{X(j)}(x) &= \frac{n!}{(j-1)!(n-j)!} (x)^{j-1} (1-x)^{n-j}, \quad x \in [0, 1] \\ &= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n+1-j)} x^{j-1} (1-x)^{n-j} = \frac{\Gamma(j+n+1-j)}{\Gamma(j)\Gamma(n+1-j)} x^{j-1} (1-x)^{n+1-j-1} \end{aligned}$$

which is the pdf for a beta distribution with parameters  $j$  and  $n+1-j$ .

□

**Corollary 10.1.3.1.** Let  $X$  be a random variable uniformly distributed in  $[0, 1]$ . Then  $\mathbb{E}X_{(j)} = \frac{j}{n+1}$

*Proof.* Follows from Proposition 10.1.3 since the mean of such a beta distribution is  $\frac{j}{n+1}$ .

□

**Proposition 10.1.4 (Result from 541A).** Let  $a, b \in \mathbb{R}$  with  $a < b$ . Let  $U$  be the number of indices  $1 \leq j \leq n$  such that  $X_j \leq a$ . Let  $V$  be the number of indices  $1 \leq j \leq n$  such that  $a < X_j \leq b$ . Then the vector  $(U, V, n - U - V)$  is a multinomial random variable, so that for any nonnegative integers  $u, v$  with  $u + v \leq n$ , we have

$$\begin{aligned}\mathbb{P}(U = u, V = v, n - U - V = n - u - v) \\ = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b) - F_X(a))^v (1 - F_X(v))^{n-u-v}.\end{aligned}$$

Consequently, for any  $1 \leq i, j \leq n$ ,

$$\mathbb{P}(X_{(i)} \leq a, X_{(j)} \leq b) = \mathbb{P}(U \geq i, U + V \geq j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbb{P}(U = k, V = m) + \mathbb{P}(U \geq j).$$

So, it is possible to write an explicit formula for the joint distribution of  $X_{(i)}$  and  $X_{(j)}$ .

*Proof.* We can define a multinomial distribution as follows (from Sheldon Ross *Stochastic Processes*, see Definition 6.27): “Suppose that  $n$  independent trials, each of which results in either outcome  $1, 2, \dots, r$  with respective probabilities  $p_1, p_2, \dots, p_r$  (with  $\sum_i p_i = 1$ ), are performed. Let  $N_i$  denote the number of trials resulting in outcome  $i$ . Then the joint distribution of  $N_1, \dots, N_r$  is called the **multinomial distribution**.” In this case  $r = 3$ . If we define outcome 1 to be  $X_j \leq a$ , outcome 2 to be  $a < X_j \leq b$ , and outcome 3 to be  $X_j > b$ , then the counts  $(U, V, n - U - V)$  meet this definition exactly, with  $p_1 = \Pr(X_j \leq a) = F_X(a)$ ,  $p_2 = \Pr(a < X_j \leq b) = F_X(b) - F_X(a)$ ,  $p_3 = \Pr(X_j > b) = 1 - F_X(b)$ . Since the pmf of a multinomial distribution with  $r = 3$  is

$$\Pr((N_1, N_2, N_3) = (n_1, n_2, n_3)) = \binom{n}{n_1, n_2, n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3} = \frac{n!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

we have in this case

$$\Pr(U = u, V = v, n - U - V = n - u - v) = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b) - F_X(a))^v (1 - F_X(v))^{n-u-v}$$

as desired.

□

**Definition 10.3 (Median; Math 505A defintion).** The real number  $m$  is called a **median** of a random variable  $X$  if

$$\Pr(X \leq m) \geq 1/2, \quad \Pr(X \geq m) \geq 1/2.$$

**Proposition 10.1.5 (Math 505A homework problem).** (a) Every random variable has at least one median.

(b) The set of all medians is a closed interval of the real line.

*Proof.* (a) Suppose the cdf of  $X$ ,  $F_X : \mathbb{R} \rightarrow [0, 1]$ , is continuous. Because  $F_X(x) = \Pr(X \leq x)$  is a cdf, it is also monotonically increasing. By the Intermediate Value Theorem, there exists at least one  $m \in \mathbb{R}$  such that  $F_X(m) = 1/2$ . Because  $\Pr(X \leq m) = 1/2 \geq 1/2$  and  $\Pr(X \geq m) = 1 - \Pr(X < m) = 1 - 1/2 \geq 1/2$ ,  $m$  is a median.

Suppose  $F_X$  is not continuous. If it contains  $1/2$  in its range, then  $m$  such that  $F_X(m) = 1/2$  is a median. If there is no  $m \in \mathbb{R}$  such that  $F_X(m) = 1/2$ , then  $m = \inf(\{x \mid F_X(x) \geq 1/2\})$  is a median. To see why, note first that  $\Pr(X \leq m) = F_X(m) \geq 1/2$ . Second,

$$\Pr(X \geq m) = 1 - \Pr(X < m) = 1 - \lim_{x \rightarrow m^-} F_X(x) \geq 1 - 1/2 = 1/2$$

because  $F_X$  is right continuous. Therefore  $m$  is a median of  $X$ .

(b) We show that all medians of  $X$  must be in one interval by contradiction. Suppose  $a$  and  $b$  are medians of  $X$  but  $c$  is not, where  $a < c < b$ . By the definition of median,  $F_X(a) \geq 1/2$  and  $F_X(b) \geq 1/2$ . Because  $c$  is not a median,  $F_X(c) < 1/2$ . This implies that  $F_X$  is decreasing on the interval from  $a$  to  $c$ , which contradicts the fact that the distribution function  $F_X$  monotonically increases.

Finally we prove that all medians of  $X$  are in a closed interval. Let  $\mathcal{A}$  be the set of all medians of  $X$ ; that is,  $\mathcal{A} = \{x \mid \Pr(X \leq x) \geq 1/2, \Pr(X \geq x) \geq 1/2\} = \{x \mid F_X(x) \geq 1/2, \lim_{y \rightarrow x^-} F_X(y) \leq 1/2\}$ . We will show that  $\mathcal{A}$  contains its infimum and its supremum. The argument above shows that  $a = \inf(\{x \mid F_X(x) \geq 1/2\})$  satisfies  $\lim_{y \rightarrow a^-} F_X(y) \leq 1/2$ ; that is,  $a \in \mathcal{A}$ . Since there is no lower value  $k$  which satisfies  $F_X(k) \geq 1/2$ ,  $a = \inf(\mathcal{A})$ , so  $\mathcal{A}$  contains its infimum.

Let  $b = \sup\{x \mid \lim_{y \rightarrow x^-} F_X(y) \leq 1/2\}$ . Because  $b$  is the supremum of a set containing  $a$ ,  $b \geq a$ . Therefore because  $F_X$  is nondecreasing,  $F_X(b) \geq F_X(a) \geq 1/2$ , which shows that  $b \in \mathcal{A}$ . Since  $b$  is the supremum of the set of all values satisfying  $\lim_{y \rightarrow x^-} F_X(y) \leq 1/2$ ,  $b$  is the supremum of  $\mathcal{A}$ . Therefore  $\mathcal{A}$  contains its infimum and supremum, and the set of all medians of  $X$  is closed.

□

**Remark.** One example of a random variable which has a median of length  $L$ :  $X$  is a discrete random variable with the following mass function:

$$\Pr(X = 0) = 0.5$$

$$\Pr(X = L) = 0.5$$

Then  $m \in [0, L]$  are medians.

## 10.2 Random Samples

**Definition 10.4 (Random Sample).** Let  $n > 0$  be an integer. A **random sample** of size  $n$  is a sequence  $X_1, \dots, X_n$  of independent identically distributed random variables.

**Definition 10.5 (Statistic).** Let  $n, k$  be positive integers. Let  $X_1, \dots, X_n$  be a random sample of size  $n$ . Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$  be a (measureable) function. A **statistic** is a random variable of the form  $Y := f(X_1, \dots, X_n)$ . The distribution of  $Y$  is called a **sampling distribution**.

**Definition 10.6 (Sample mean).** The **sample mean** of a random sample  $X_1, \dots, X_n$  of size  $n$ , denoted  $\bar{X}$ , is the following statistic:

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i.$$

**Proposition 10.2.1.** Suppose we have a random sample of size  $n$  from an i.i.d. distribution  $X_1, X_2, \dots, X_n$  with  $\mathbb{E}(X_1) = \mu$  in  $\mathbb{R}$ ,  $\text{Var}(X_1) = \sigma^2 < \infty$ . Then

- (a)  $\mathbb{E}(\bar{X}) = \mathbb{E}(X_1)$ .
- (b)  $\text{Var}(\bar{X}) = \sigma^2/n$ .

*Proof.* (a)

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}X_i = \frac{1}{n} \cdot n\mu = \mu$$

(b) Using the independence of the  $X_i$ ,

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \boxed{\frac{\sigma^2}{n}}$$

□

**Proposition 10.2.2 (Stats 100B homework problem).** Suppose that  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  are two samples, with  $X \sim \mathcal{N}(\mu_1, \sigma_1)$  and  $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ . The difference between the sample means,  $\bar{X} - \bar{Y}$ , is then a linear combination of  $m + n$  normal random variables. Then

- a.  $\mathbb{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$ .
- b.  $\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$ .
- c. The distribution of  $\bar{X} - \bar{Y}$  is normal with mean and variance equal to the previous results.

*Proof.* a.

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i, \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$$

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m X_i - \frac{1}{n} \sum_{j=1}^n Y_j\right) = \frac{1}{m} \mathbb{E}\left(\sum_{i=1}^m X_i\right) - \frac{1}{n} \mathbb{E}\left(\sum_{j=1}^n Y_j\right)$$

$$= \frac{1}{m} \sum_{i=1}^m \mathbb{E}(X_i) - \frac{1}{n} \sum_{j=1}^n \mathbb{E}(Y_j) = \frac{1}{m} \sum_{i=1}^m \mu_1 - \frac{1}{n} \sum_{j=1}^n \mu_2 = \frac{1}{m} m \cdot \mu_1 - \frac{1}{n} n \cdot \mu_2$$

$$\implies \mathbb{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

b. Since  $X$  and  $Y$  are independent,

$$\begin{aligned} \text{Var}(\bar{X} - \bar{Y}) &= \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) \\ &= \mathbb{E}[(\bar{X} - \mathbb{E}[\bar{X}])^2] + \mathbb{E}[(\bar{Y} - \mathbb{E}[\bar{Y}])^2] \\ &= \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m X_i - \mu_1\right)^2 + \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^n Y_j - \mu_2\right)^2 \\ &= \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m (X_i - m \frac{1}{m} \mu_1)\right)^2 + \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^n (Y_j - n \frac{1}{n} \mu_2)\right)^2 \\ &= \frac{1}{m^2} \mathbb{E}\left(\sum_{i=1}^m (X_i - \mu_1)\right)^2 + \frac{1}{n^2} \mathbb{E}\left(\sum_{j=1}^n (Y_j - \mu_2)\right)^2 \end{aligned}$$

Since  $X_i$  and  $X_j$  are independent for  $i \neq j$  (and likewise for  $Y$ ),  $\text{Cov}(X_i, X_j) = 0$  for  $i \neq j$ , so

$$\mathbb{E}[(X_i - \mu_1)(X_j - \mu_1)] = 0$$

for  $i \neq j$  (and likewise for  $Y$ ). Therefore the above equation can be written as

$$\begin{aligned} &\frac{1}{m^2} \mathbb{E}\left(\sum_{i=1}^m (X_i - \mu_1)^2\right) + \frac{1}{n^2} \mathbb{E}\left(\sum_{j=1}^n (Y_j - \mu_2)^2\right) \\ &\frac{1}{m^2} \sum_{i=1}^m \mathbb{E}(X_i - \mu_1)^2 + \frac{1}{n^2} \sum_{j=1}^n \mathbb{E}(Y_j - \mu_2)^2 \\ &= \frac{1}{m^2} \left(\sum_{i=1}^m \sigma_1^2\right) + \frac{1}{n^2} \left(\sum_{j=1}^n \sigma_2^2\right) = \frac{1}{m^2} m \cdot \sigma_1^2 + \frac{1}{n^2} n \cdot \sigma_2^2 \\ &\implies \text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \end{aligned}$$

c.

$$M_{X_i}(t) = \exp\left(\mu_1 t + \frac{t^2 \sigma_1^2}{2}\right), \quad M_{Y_i}(t) = \exp\left(\mu_2 t + \frac{t^2 \sigma_2^2}{2}\right)$$

Since individual observations from  $X$  and  $Y$  are independent,

$$M_{\bar{X}}(t) = \prod_{i=1}^m M_{X_i}\left(\frac{1}{m}t\right), \quad M_{\bar{Y}}(t) = \prod_{j=1}^n M_{Y_j}\left(\frac{1}{n}t\right)$$

and

$$\begin{aligned}
M_{\bar{X}-\bar{Y}}(t) &= M_{\bar{X}}(t)M_{-\bar{Y}}(t) = M_{\bar{X}}(t)M_{\bar{Y}}(-t) = \prod_{i=1}^m M_{X_i}\left(\frac{1}{m}t\right) \prod_{j=1}^n M_{Y_j}\left(\frac{-1}{n}t\right) \\
&= \left[M_{X_i}\left(\frac{t}{m}\right)\right]^m \left[M_{Y_j}\left(\frac{-t}{n}\right)\right]^n = \left[\exp\left(\frac{\mu_1 t}{m} + \frac{t^2 \sigma_1^2}{2m^2}\right)\right]^m \left[\exp\left(\frac{-\mu_2 t}{n} + \frac{(-t)^2 \sigma_2^2}{2n^2}\right)\right]^n \\
&= \exp\left(\frac{m\mu_1 t}{m} + \frac{mt^2 \sigma_1^2}{2m^2}\right) \exp\left(\frac{-n\mu_2 t}{n} + \frac{nt^2 \sigma_2^2}{2n^2}\right) \\
\implies M_{\bar{X}-\bar{Y}}(t) &= \exp\left[(\mu_1 - \mu_2)t + \frac{1}{2}t^2\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\right]
\end{aligned}$$

This is the moment generating function of a normal distribution with mean  $\mu_1 - \mu_2$  and variance  $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$ , consistent with the results from parts (a) and (b). □

**Definition 10.7 (Sample variance).** Let  $n > 1$ . The **sample variance** of a random sample  $X_1, \dots, X_n$  of size  $n$ , denoted  $S^2$ , is the following statistic:

$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The **sample standard deviation** of a random sample of size  $n$  is  $\sqrt{S^2}$ .

**Proposition 10.2.3 (Unbiasedness of sample variance).** Suppose we have a random sample of size  $n$  from an i.i.d. distribution  $X_1, X_2, \dots, X_n$  with  $\mathbb{E}(X_1) = \mu$  in  $\mathbb{R}$ ,  $\text{Var}(X_1) = \sigma^2 < \infty$ . Then  $\mathbb{E}(S^2) = \sigma^2$ . Further,  $S^2$  is a consistent estimator of  $\sigma^2$ .

*Proof.* We have

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{1}{n-1} \mathbb{E}\left(\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + \bar{X}^2\right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^n \mathbb{E}(X_i^2) - 2\mathbb{E}\left(\bar{X} \sum_{i=1}^n X_i\right) + n\mathbb{E}\bar{X}^2\right) = \frac{1}{n-1} \left(n\mathbb{E}(X_i^2) - 2n\mathbb{E}\bar{X}^2 + n\mathbb{E}\bar{X}^2\right) \\
&= \frac{n}{n-1} (\mathbb{E}(X_i^2) - \mathbb{E}\bar{X}^2) = \frac{n}{n-1} (\text{Var}(X_i) + \mathbb{E}(X_i)^2 - [\text{Var}(\bar{X}) + \mathbb{E}(\bar{X})^2])
\end{aligned}$$

Using the results from Proposition 10.2.1, we have

$$\mathbb{E}(S^2) = \frac{n}{n-1} (\sigma^2 + \mu^2 - [\sigma^2/n + \mu^2]) = \frac{n}{n-1} \cdot \frac{(n-1)\sigma^2}{n} = \boxed{\sigma^2}$$

□

Alternative proof from Stats 100B homework.

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\right) = \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n (x_i - \mu)^2\right)$$

Assuming independence of samples, this can be written as

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}((x_i - \mu)^2) = \frac{1}{n} n \sigma^2 = \boxed{\sigma^2}$$

Since  $S^2$  is unbiased, it is a consistent estimator if we can show  $\text{Var}(S^2) \rightarrow 0$  as  $n \rightarrow \infty$ . We have

$$\text{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$$

$$\frac{(n-1)^2}{\sigma^4} \text{Var}(S^2) = 2(n-1)$$

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

$$\implies \lim_{n \rightarrow \infty} \text{Var}(S^2) = \lim_{n \rightarrow \infty} \frac{2\sigma^4}{n-1} = \boxed{0}$$

Therefore  $S^2$  is a consistent estimator of  $\sigma^2$ .

□

**Lemma 10.2.4.** Let  $X := (X_1, \dots, X_n)$  be i.i.d. mean zero, variance 1 Gaussian random variables. Let  $v_1, \dots, v_n \in \mathbb{R}^n$ . Then  $\langle X, v_1 \rangle, \dots, \langle X, v_n \rangle$  are independent if and only if  $v_1, \dots, v_m$  are pairwise orthogonal; that is,  $\langle v_i, v_j \rangle = 0 \forall 1 \leq i < j \leq m$ .

*Proof.* By Theorem 6.3.16, we have that for any  $v \in \mathbb{R}^n$ ,  $\langle X, v \rangle$  is a mean zero Gaussian with variance  $\langle v, v \rangle$ . For notational convenience, let  $\langle X, v_k \rangle = A_k$ . Because all the  $A_k$  are Gaussian random variables by Theorem 6.3.16, the  $A_k$  are uncorrelated if and only if they are independent. That is, we would like to show that their covariances

$$\mathbb{E}[(A_k - \mathbb{E}A_k)(A_\ell - \mathbb{E}A_\ell)]$$

equal zero for all  $\{(k, \ell) : k, \ell \in \{1, 2, \dots, m\}, k \neq \ell\}$  if and only if the vectors  $v_1, \dots, v_m$  are pairwise orthogonal; that is,  $\langle v_k, v_\ell \rangle = 0$  for all  $\{(k, \ell) : k, \ell \in \{1, 2, \dots, m\}, k \neq \ell\}$ . Note that since  $A_k = \sum_{i=1}^n X_i v_{ki}$ ,  $\mathbb{E}(A_k) = \sum_{i=1}^n v_{ki} \mathbb{E}(X_i)$ . So for any  $\{(k, \ell) : k, \ell \in \{1, 2, \dots, m\}, k \neq \ell\}$  we have

$$\mathbb{E}[(A_k - \mathbb{E}A_k)(A_\ell - \mathbb{E}A_\ell)] = \mathbb{E}\left[\left(\sum_{i=1}^n X_i v_{ki} - \sum_{i=1}^n v_{ki} \mathbb{E}(X_i)\right)\left(\sum_{i=1}^n X_i v_{\ell i} - \sum_{i=1}^n v_{\ell i} \mathbb{E}(X_i)\right)\right]$$

$$\begin{aligned}
&= \mathbb{E} \left[ \left( \sum_{i=1}^n X_i v_{ki} \right) \left( \sum_{i=1}^n X_i v_{\ell i} \right) - \left( \sum_{i=1}^n X_i v_{ki} \right) \left( \sum_{i=1}^n v_{\ell i} \mathbb{E}(X_i) \right) \right. \\
&\quad \left. - \left( \sum_{i=1}^n v_{ki} \mathbb{E}(X_i) \right) \left( \sum_{i=1}^n X_i v_{\ell i} \right) + \left( \sum_{i=1}^n v_{ki} \mathbb{E}(X_i) \right) \left( \sum_{i=1}^n v_{\ell i} \mathbb{E}(X_i) \right) \right] \\
&= \mathbb{E} \left( \sum_{i=1}^n X_i^2 v_{ki} v_{\ell i} + \sum_{\{a,b \in \{1, \dots, n\}, a \neq b\}} X_a X_b v_{ka} v_{\ell b} \right) - 2 \mathbb{E} \left( \sum_{i=1}^n X_i \mathbb{E}(X_i) v_{ki} v_{\ell i} + \sum_{\{a,b \in \{1, \dots, n\}, a \neq b\}} X_a \mathbb{E}(X_b) v_{ka} v_{\ell b} \right) \\
&\quad + \mathbb{E} \left( \sum_{i=1}^n \mathbb{E}(X_i)^2 v_{ki} v_{\ell i} + \sum_{\{a,b \in \{1, \dots, n\}, a \neq b\}} \mathbb{E}(X_a) \mathbb{E}(X_b) v_{ka} v_{\ell b} \right)
\end{aligned}$$

Recall that  $\mathbb{E}(X_i) = 0$  for all  $i$ . Also, due to independence of the  $X_i$ , all of the terms that involve  $\mathbb{E}(X_a X_b)$ ,  $a \neq b$  disappear. This leaves only

$$= \mathbb{E} \left( \sum_{i=1}^n X_i^2 v_{ki} v_{\ell i} \right) = \sum_{i=1}^n \mathbb{E}(X_i^2) v_{ki} v_{\ell i} = \mathbb{E}(X_1^2) \sum_{i=1}^n v_{ki} v_{\ell i} \quad (10.1)$$

where the last step follows from the i.i.d. distributions of  $X_i$ . Recall

$$\langle v_k, v_\ell \rangle = 0 \iff \sum_{i=1}^n v_{ki} v_{\ell i} = 0.$$

Since  $\mathbb{E}(X_i^2) \neq 0$ , (10.1) equals 0 for all  $\{(k, \ell) : k, \ell \in \{1, 2, \dots, m\}, k \neq \ell\}$  if and only if  $\langle v_k, v_\ell \rangle = 0$  for all  $\{(k, \ell) : k, \ell \in \{1, 2, \dots, m\}, k \neq \ell\}$ . Therefore the random variables  $\langle X, v_1 \rangle, \dots, \langle X, v_m \rangle$  are independent if and only if the vectors  $v_1, \dots, v_m$  are pairwise orthogonal.

□

**Proposition 10.2.5 (Proposition 4.7 in 541A notes).** Let  $n \geq 2$  be an integer. Let  $X_1, \dots, X_n$  be a random sample from the Gaussian distribution with mean  $\mu \in \mathbb{R}$  and variance  $\sigma^2 > 0$ . Let  $\bar{X}$  be the sample mean and let  $S$  be the sample standard deviation. Then

- (i)  $\bar{X}$  and  $S$  are independent random variables.
- (ii)  $\bar{X}$  is a Gaussian random variable with mean  $\mu$  and variance  $\sigma^2/n$ .
- (iii)  $(n-1)S^2/\sigma^2$  is a  $\chi^2$ -distributed random variable with  $n-1$  degrees of freedom.

*Proof.* (i) Replace  $X_1, \dots, X_n$  with  $X_1 - \mu, \dots, X_n - \mu$  so that  $\mu = 0$ . Also divide by  $\sigma$  so that  $\sigma = 1$ .

Note that  $\bar{X}$  is independent of all random variables  $X_2 - \bar{X}, \dots, X_n - \bar{X}$  by Lemma 10.2.4 because for example

$$X_2 - \bar{X} = \langle X_2, e_2 - \frac{1}{n}(1, 1, \dots, 1) \rangle$$

where the second vector in the inner product is orthogonal to  $(1, 1, \dots, 1)$  (in fact,  $(1, \dots, 1)$  is orthogonal to anything in the span of these vectors). Likewise for all the remaining vectors you could use to construct  $X_i$ . (Note that the other random variables [e.g.  $X_2 - \bar{X}$  and  $X_3 - \bar{X}$ ] are not independent.)

So the proof will be complete if we can write  $S$  as a function of  $X_2 - \bar{X}, \dots, X_n - \bar{X}$ . Observe

$$\begin{aligned} (n-1)S^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + \sum_{i=2}^n (X_i - \bar{X})^2 = \left( n\bar{X} - \left[ \sum_{i=2}^n X_i \right] - \bar{X} \right)^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \\ &= \left( \sum_{i=2}^n (X_i - \bar{X}) \right)^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \end{aligned}$$

- (ii) Follows from Proposition 1.45, Example 1.108, and Exercise 1.58 in 541A notes (condense later?)
- (iii) Like above, replace  $X_1, \dots, X_n$  with  $X_1 - \mu, \dots, X_n - \mu$  so that  $\mu = 0$ . Also divide by  $\sigma$  so that  $\sigma = 1$ . We will prove by induction. Let  $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$  and let  $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ . In the case  $n = 2$  we have

$$\begin{aligned} S_2^2 &= \left( X_1 - \frac{1}{2}(X_1 + X_2) \right)^2 + \left( X_2 - \frac{1}{2}(X_1 + X_2) \right)^2 = \frac{1}{4}(X_2 - X_1)^2 + \frac{1}{4}(X_2 - X_1)^2 = \frac{1}{2}(X_2 - X_1)^2 \\ &= \left( \frac{1}{\sqrt{2}}(X_2 - X_1) \right)^2 \end{aligned}$$

Note that  $1/\sqrt{2}(X_2 - X_1)$  is a mean zero Gaussian random variable with variance 1 (see example 1.108 in 541A notes for details). So  $S_2^2$  is  $\chi_1^2$  by Definition 1.33 in 541A notes.

We now induct on  $n$ . From Lemma 4.8 in 541A notes (will prove later),

$$nS_{n+1}^2 = (n-1)S_n^2 + \frac{n}{n+1}(X_{n+1} - \bar{X}_n)^2, \quad \forall n \geq 2$$

From the first item,  $S_n$  is independent of  $\bar{X}_n$ . Also,  $X_{n+1}$  is independent of  $S_n$  by Proposition 1.61 in Math 541A notes, since  $S_n$  is a function of  $X_1, \dots, X_n$ , which are independent of  $X_{n+1}$ . So  $S_n$  is independent of  $(X_{n+1} - \bar{X}_n)^2$ . By the inductive hypothesis,  $(n-1)S_n^2$  is a  $\chi_{n-1}^2$  random variable. From Example 1.108 in Math 541A notes,  $X_{n+1} - \bar{X}_n$  is a Gaussian random variable with mean zero and variance  $1 + 1/n = (n+1)/n$  so that  $\sqrt{n/(n+1)}(X_{n+1} - \bar{X}_n)$  is a mean zero Gaussian with variance 1, implying  $n/(n+1)(X_{n+1} - \bar{X}_n)^2$  is  $\chi^2$ . Definition 1.33 in 541A notes then implies that  $nS_{n+1}$  is a  $\chi_n^2$  random variable, completing the inductive step.

□

#### Lemma 10.2.6 (Lemma 4.8 in 541A notes.).

Let  $X_1, X_2, \dots$  be random variables. For any  $n \geq 2$ , let  $\bar{X}_n := (1/n) \sum_{i=1}^n X_i$  and let  $S_n^2 := 1/(n-1) \sum_{i=1}^n (X_i - \bar{X}_n)^2$ . Then

$$nS_{n+1}^2 - (n-1)S_n^2 = \frac{n}{n+1}(X_{n+1} - \bar{X}_n)^2.$$

*Proof.*

$$nS_{n+1}^2 - (n-1)S_n^2 = \sum_{i=1}^{n+1} (X_i - \bar{X}_{n+1})^2 - \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Note:

$$(a-b)^2 - (a-c)^2 = a^2 - 2ab + b^2 - a^2 - c^2 + 2ac = b^2 - c^2 + 2a(c-b)$$

$$= (b-c)[(b+c) - 2a] = (b-c)(b+c-2a)$$

for all real  $a, b, c$ . Using  $a = X_n, b = \bar{X}_{n+1}, c = \bar{X}_n$  we have

$$\begin{aligned} &= (X_{n+1} - \bar{X}_{n+1})^2 + \sum_{i=1}^n (\bar{X}_{n+1} - \bar{X}_n)(\bar{X}_{n+1} + \bar{X}_n - 2X_i) \\ &= (X_{n+1} - \bar{X}_{n+1})^2 + (\bar{X}_{n+1} - \bar{X}_n) \sum_{i=1}^n (\bar{X}_{n+1} + \bar{X}_n - 2X_i) \\ &= (X_{n+1} - \bar{X}_{n+1})^2 + (\bar{X}_{n+1} - \bar{X}_n) \cdot n(\bar{X}_{n+1} + \bar{X}_n - 2\bar{X}_n) \\ &= (X_{n+1}(1 - 1/(n+1)) - \frac{n}{n+1}\bar{X}_n)^2 + n(\bar{X}_{n+1} - \bar{X}_n)^2 \\ &= \frac{n^2}{(n+1)^2} (X_{n+1} - \bar{X}_n)^2 + n \left( \frac{X_{n+1}}{n+1} + \left( \frac{1}{n+1} - \frac{1}{n} \right) \sum_{i=1}^n X_i \right)^2 \end{aligned}$$

Algebra:  $1/(n+1) - 1/n = \frac{n-(n+1)}{n(n+1)} = -\frac{1}{n(n+1)}$ . So we have

$$\begin{aligned} &= \frac{n^2}{(n+1)^2} (X_{n+1} - \bar{X}_n)^2 + \frac{n}{(n+1)^2} (X_{n+1} - \frac{1}{n} \sum_{i=1}^n X_i)^2 \\ &= \frac{n^2}{(n+1)^2} (X_{n+1} - \bar{X}_n)^2 + \frac{n}{(n+1)^2} (X_{n+1} - \bar{X}_n)^2 \\ &= \frac{n^2+n}{(n+1)^2} (X_{n+1} - \bar{X}_n)^2 = \frac{n}{n+1} (X_{n+1} - \bar{X}_n)^2 \end{aligned}$$

□

**Proposition 10.2.7 (Proposition 4.9 in 541A notes).** Let  $X$  be a standard Gaussian random variable. Let  $Y$  be a  $\chi_p^2$  random variable. Assume that  $X$  and  $Y$  are independent. Then  $X/\sqrt{Y/p}$  has the following density, known as **Student's  $t$ -distribution** with  $p = \text{degrees of freedom}$ : ( $p = n + 1?$ )

$$f_{X/(Y/\sqrt{p})}(t) := \frac{\Gamma((p+1)/2)}{\sqrt{p}\sqrt{\pi}\Gamma(p/2)} \left(1 + \frac{t^2}{p}\right)^{-(p+1)/2}, \quad \forall t \in \mathbb{R}$$

(should have  $p + 1$  in a bunch of the expressions above? that's what was written on board, not in notes.)

*Proof.* Let  $Z := \sqrt{Y/p}$ . We find the density of  $Z$  as follows. Let  $t > 0$ . Then

$$\begin{aligned} f_Z(y) &= \frac{d}{dy} \Big|_{y=0} \Pr(Z \leq y) = \frac{d}{dy} \Big|_{y=0} \Pr(Y \leq y^2 p) \\ &= \frac{d}{dy} \Big|_{y=0} \int_0^{y^2 p} \frac{x^{(p/2)-1} e^{-x/2}}{2^{p/2} \Gamma(p/2)} dx = 2yp \cdot p^{(p/2)-1} y^{p-2} e^{-y^2 p/2} \cdot \frac{1}{2^{p/2} \Gamma(p/2)} \\ &= p^{p/2} y^{p-1} e^{-y^2 p/2} \cdot \frac{1}{2^{p/2-1} \Gamma(p/2)} \end{aligned}$$

: skipped this stuff in class proof

$$\Pr(X/Z \leq t) = \Pr(X \leq tZ)$$

$$= (\text{by definition of joint density}) \int \int_{\{(x,y) \in \mathbb{R}^2 : x \leq ty\}} f_X(x)f_Z(y) dx dy$$

We use the change of variables formula:

$$\int \int_{\phi(U)} f(x,y) dx dy = \int \int_U f(\phi(a,b)) |\text{Jac } \phi(a,b)| da db$$

$$\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$$

$$\phi(a,b) = (ab, a)$$

$$\phi^{-1}(x,y) = (y, x/y)$$

We chose  $x/y$  as the second variable so that an upper limit of the variable will end up being  $t$  after the transformation. We need the Jacobian of  $\phi$ :

$$|\operatorname{Jac} \phi(a, b)| = \left| \det \begin{pmatrix} b & a \\ 1 & 0 \end{pmatrix} \right| = |a|$$

By the change in variables formula,

$$\begin{aligned} \int \int_{\phi(U)} f(x, y) dx dy &= \int \int_U f(\phi(a, b)) |\operatorname{Jac} \phi(a, b)| da db \\ &= \int \int_{\{(a, b) \in \mathbb{R}^2 : a \geq 0, b \leq t\}} f_X(ab) f_Z(a) |a| da db \\ \implies \Pr(X/Z \leq t) &= \int_{-\infty}^t \int_0^\infty |a| f_X(ab) f_Z(a) da db \end{aligned}$$

By the Fundamental Theorem of Calculus,

$$f_{X/Z}(t) = \frac{d}{dt} \Pr(X/Z \leq t) = \int_0^\infty |a| f_X(at) f_Z(a) da = \int_0^\infty a f_X(at) f_Z(a) da$$

By the definitions of  $X$  and  $Z$ ,

$$\begin{aligned} &= \frac{1}{2^{-/2-1}\Gamma(p/2)} \int_0^\infty a \cdot \frac{1}{\sqrt{2\pi}} e^{-(a^2 t^2)/2} \cdot p^{p/2} a^{p-1} e^{-a^2 p/2} da \\ &= \frac{p^{p/2}}{2^{-/2-1}\Gamma(p/2)\sqrt{2\pi}} \int_0^\infty e^{-[a^2(t^2+p)]/2} \cdot a^p da \end{aligned}$$

Change of variables: let  $x = a^2$ ,  $dx = 2ada$ ,  $da = \frac{1}{2a}dx = 1/(2\sqrt{x})dx$ . Then this integral is

$$= c \int_0^\infty e^{-[x(t^2+p)]/2} \cdot x^{p/2-1/2} da, \quad \text{where } c = \frac{p^{p/2}}{2^{p/2}\sqrt{2\pi}\Gamma(p/2)}$$

So the integrand is a Gamma density function with parameters  $\alpha, \beta$ :  $\alpha - 1 = p/2 - 1/2 \iff \alpha = p/2 + 1/2$ ,  $\beta = 2/(t^2 + p)$ . So if we multiply and divide  $\beta^\alpha \Gamma(\alpha)$

So

$$\begin{aligned} f_{X/Z}(t) &= \frac{p^{p/2}}{2^{p/2}\sqrt{2\pi}\Gamma(p/2)} \cdot \beta^\alpha \Gamma(\alpha) \cdot 1 = \frac{p^{p/2}\Gamma((p_1)/2)}{2^{p/2}\sqrt{2\pi}\Gamma(p/2)} \cdot \left( \frac{2}{t^2 + p} \right)^{(p-1)/2} \\ &= \frac{p^{p/2}\Gamma((p+1)/2)}{\sqrt{\pi}\Gamma(p/2)} \cdot (t^2 + p)^{-(p+1)/2} = \frac{\Gamma((p+1)/2)}{\sqrt{\pi}\Gamma(p/2)} \cdot (1 + t^2/p)^{-(p+1)/2} \end{aligned}$$

□

**Remark** (Remark 4.10 in 541A notes). If  $X_1, \dots, X_n$  is a random sample from a Gaussian distribution with mean  $\mu \in \mathbb{R}$ , standard deviation  $\sigma < 0$ , then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

also has Student's  $t$  distribution. ( $\bar{X} := n^{-1} \sum_{i=1}^n X_i, S = \sqrt{(n-1)^{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ .)

**Proposition 10.2.8 (Stats 100B homework 3 problem).** Let  $X_1, X_2$  be a random sample from a normal distribution with a mean  $\mu$  and standard deviation  $\sigma$ . Then  $(n-1)s^2/\sigma^2$  has a  $\chi_1^2$  distribution.

*Proof.*

$$\begin{aligned} s^2 &= \frac{1}{2-1} \sum_{i=1}^2 (X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = (X_1 - \frac{X_1 + X_2}{2})^2 + (X_2 - \frac{X_1 + X_2}{2})^2 \\ &= X_1^2 - 2X_1(\frac{X_1 + X_2}{2}) + (\frac{X_1 + X_2}{2})^2 + X_2^2 - 2X_2(\frac{X_1 + X_2}{2}) + (\frac{X_1 + X_2}{2})^2 \\ &= X_1^2 + X_2^2 - X_1(X_1 + X_2) - X_2(X_1 + X_2) + 2(\frac{X_1 + X_2}{2})^2 \\ &= X_1^2 + X_2^2 - (X_1 + X_2)(X_1 + X_2) + \frac{(X_1 + X_2)^2}{2} \\ &= \frac{1}{2}(2X_1^2 + 2X_2^2) - \frac{1}{2}(X_1^2 + 2X_1X_2 + X_2^2) \\ &= \frac{1}{2}(X_1^2 - 2X_1X_2 + X_2^2) \end{aligned}$$

$$s^2 = \frac{1}{2}(X_1 - X_2)^2$$

$$\implies \frac{(n-1)s^2}{\sigma^2} = (2-1)\frac{1}{2\sigma^2}(X_1 - X_2)^2 = \left(\frac{X_1 - X_2}{\sigma\sqrt{2}}\right)^2$$

Since  $X_1$  and  $X_2$  are normal,

$$X_1 - X_2 \sim \mathcal{N}(\mu - \mu, \sqrt{\sigma^2 + \sigma^2}) = \mathcal{N}(0, \sigma\sqrt{2}) \implies \frac{X_1 - X_2}{\sigma\sqrt{2}} \sim \mathcal{N}(0, 1)$$

$$\implies \left(\frac{X_1 - X_2}{\sigma\sqrt{2}}\right)^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi_1^2$$

□

**Proposition 10.2.9 (Stats 100B homework problem).** Suppose two independent random samples of  $n_1$  and  $n_2$  observations are selected from two normal populations. Further, assume that the populations possess a common variance  $\sigma^2$  which is unknown. Let the sample variances be  $S_1^2$  and  $S_2^2$  and assume they are unbiased. Then the pooled estimator for  $\sigma^2$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is unbiased and has variance  $\frac{2\sigma^4}{n_1 + n_2 - 2}$ .

*Proof.* First we show  $S^2$  is unbiased.

$$\begin{aligned} \mathbb{E}(S^2) &= \mathbb{E}\left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\right) = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbb{E}(S_1^2) + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbb{E}(S_2^2) \\ &= \frac{n_1 - 1}{n_1 + n_2 - 2}\sigma^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}\sigma^2 = \frac{(n_1 + n_2 - 2)\sigma^2}{n_1 + n_2 - 2} = \boxed{\sigma^2} \end{aligned}$$

Now we derive its variance.

$$\text{Var}(S^2) = \text{Var}\left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\right)$$

Since  $S_1$  and  $S_2$  are independent, this can be written as

$$\frac{1}{(n_1 + n_2 - 2)^2} \left( \text{Var}[(n_1 - 1)S_1^2] + \text{Var}[(n_2 - 1)S_2^2] \right)$$

Since the populations are normal, we know

$$\begin{aligned} \frac{(n_i - 1)S_i^2}{\sigma^2} &\sim \chi_{n_i - 1}^2 \implies \text{Var}\left(\frac{(n_i - 1)S_i^2}{\sigma^2}\right) = 2(n_i - 1) \\ \text{Var}(S^2) &= \frac{\sigma^4}{(n_1 + n_2 - 2)^2} \left( \text{Var}\left[\frac{(n_1 - 1)S_1^2}{\sigma^2}\right] + \text{Var}\left[\frac{(n_2 - 1)S_2^2}{\sigma^2}\right] \right) \\ &= \frac{\sigma^4}{(n_1 + n_2 - 2)^2} (2(n_1 - 1) + 2(n_2 - 1)) = \sigma^4 \frac{2(n_1 + n_2 - 2)}{(n_1 + n_2 - 2)^2} \\ &= \frac{2\sigma^4}{n_1 + n_2 - 2} \end{aligned}$$

□

**Proposition 10.2.10 (Stats 100B Homework problem).** Suppose that  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  are two samples, with  $X \sim \mathcal{N}(\mu_1, \sigma_1)$  and  $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ . The difference between the sample means,  $\bar{X} - \bar{Y}$ , is then a linear combination of  $m + n$  normal random variables.

- a.  $\mathbb{E}(\bar{X} - \bar{Y})$ .
- b.  $\text{Var}(\bar{X} - \bar{Y})$
- c. The distribution of  $\bar{X} - \bar{Y}$  is normal.

*Proof.* a.

$$\begin{aligned}\bar{X} &= \frac{1}{m} \sum_{i=1}^m X_i, \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j \\ \mathbb{E}(\bar{X} - \bar{Y}) &= \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m X_i - \frac{1}{n} \sum_{j=1}^n Y_j\right) = \frac{1}{m} \mathbb{E}\left(\sum_{i=1}^m X_i\right) - \frac{1}{n} \mathbb{E}\left(\sum_{j=1}^n Y_j\right) \\ &= \frac{1}{m} \sum_{i=1}^m \mathbb{E}(X_i) - \frac{1}{n} \sum_{j=1}^n \mathbb{E}(Y_j) = \frac{1}{m} \sum_{i=1}^m \mu_1 - \frac{1}{n} \sum_{j=1}^n \mu_2 = \frac{1}{m} m \cdot \mu_1 - \frac{1}{n} n \cdot \mu_2 \\ \boxed{\mathbb{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2}\end{aligned}$$

b. Since  $X$  and  $Y$  are independent,

$$\begin{aligned}\text{Var}(\bar{X} - \bar{Y}) &= \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) \\ &= \mathbb{E}[(\bar{X} - \mathbb{E}[\bar{X}])^2] + \mathbb{E}[(\bar{Y} - \mathbb{E}[\bar{Y}])^2] \\ &= \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m X_i - \mu_1\right)^2 + \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^n Y_j - \mu_2\right)^2 \\ &= \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m (X_i - m \frac{1}{m} \mu_1)\right)^2 + \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^n (Y_j - n \frac{1}{n} \mu_2)\right)^2 \\ &= \frac{1}{m^2} \mathbb{E}\left(\sum_{i=1}^m (X_i - \mu_1)\right)^2 + \frac{1}{n^2} \mathbb{E}\left(\sum_{j=1}^n (Y_j - \mu_2)\right)^2\end{aligned}$$

Since  $X_i$  and  $X_j$  are independent for  $i \neq j$  (and likewise for  $Y$ ),  $\text{Cov}(X_i, X_j) = 0$  for  $i \neq j$ , so

$$\mathbb{E}[(X_i - \mu_1)(X_j - \mu_1)] = 0$$

for  $i \neq j$  (and likewise for  $Y$ ). Therefore the above equation can be written as

$$\frac{1}{m^2} \mathbb{E}\left(\sum_{i=1}^m (X_i - \mu_1)^2\right) + \frac{1}{n^2} \mathbb{E}\left(\sum_{j=1}^n (Y_j - \mu_2)^2\right)$$

$$\begin{aligned}
& \frac{1}{m^2} \sum_{i=1}^m \mathbb{E}(X_i - \mu_1)^2 + \frac{1}{n^2} \sum_{j=1}^n \mathbb{E}(Y_j - \mu_2)^2 \\
&= \frac{1}{m^2} \left( \sum_{i=1}^m \sigma_1^2 \right) + \frac{1}{n^2} \left( \sum_{j=1}^n \sigma_2^2 \right) = \frac{1}{m^2} m \cdot \sigma_1^2 + \frac{1}{n^2} n \cdot \sigma_2^2 \\
&\boxed{\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}
\end{aligned}$$

c.

$$M_{X_i}(t) = \exp\left(\mu_1 t + \frac{t^2 \sigma_1^2}{2}\right), \quad M_{Y_i}(t) = \exp\left(\mu_2 t + \frac{t^2 \sigma_2^2}{2}\right)$$

Since individual observations from  $X$  and  $Y$  are independent,

$$M_{\bar{X}}(t) = \prod_{i=1}^m M_{X_i}\left(\frac{1}{m}t\right), \quad M_{\bar{Y}}(t) = \prod_{j=1}^n M_{Y_j}\left(\frac{1}{n}t\right)$$

and

$$\begin{aligned}
M_{\bar{X}-\bar{Y}}(t) &= M_{\bar{X}}(t)M_{-\bar{Y}}(t) = M_{\bar{X}}(t)M_{\bar{Y}}(-t) = \prod_{i=1}^m M_{X_i}\left(\frac{1}{m}t\right) \prod_{j=1}^n M_{Y_j}\left(\frac{-1}{n}t\right) \\
&= \left[ M_{X_i}\left(\frac{t}{m}\right) \right]^m \left[ M_{Y_j}\left(\frac{-t}{n}\right) \right]^n = \left[ \exp\left(\frac{\mu_1 t}{m} + \frac{t^2 \sigma_1^2}{2m^2}\right) \right]^m \left[ \exp\left(\frac{-\mu_2 t}{n} + \frac{(-t)^2 \sigma_2^2}{2n^2}\right) \right]^n \\
&= \exp\left(\frac{m\mu_1 t}{m} + \frac{mt^2 \sigma_1^2}{2m^2}\right) \exp\left(\frac{-n\mu_2 t}{n} + \frac{nt^2 \sigma_2^2}{2n^2}\right) \\
&\implies \boxed{M_{\bar{X}-\bar{Y}}(t) = \exp\left[(\mu_1 - \mu_2)t + \frac{1}{2}t^2\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\right]}
\end{aligned}$$

This is the moment generating function of a normal distribution with mean  $\mu_1 - \mu_2$  and variance  $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$ , consistent with the results from parts (a) and (b).

□

### 10.2.1 The Delta Method

**Theorem 10.2.11 (Delta Method, Theorem 4.14 in 541A notes, 5.5.24 in Casella and Berger [2001],).** Let  $\theta \in \mathbb{R}$ . Let  $Y_1, Y_2, \dots$  be random variables such that  $\sqrt{n}(Y_n - \theta)$  converges in distribution to a mean zero Gaussian random variable with variance  $\sigma^2 > 0$ . Let  $f : \mathbb{R} \rightarrow \mathbb{R}$ . Assume that  $f'$  exists and is continuous, and  $f'(\theta) \neq 0$ . Then

$$\sqrt{n}(f(Y_n) - f(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(f'(\theta))^2).$$

*Proof from new class notes.* Since  $f'(\theta)$  exists,  $\lim_{y \rightarrow \theta} \frac{f(y) - f(\theta)}{y - \theta}$  exists. That is, there exists  $h : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\lim_{z \rightarrow 0} \frac{h(z)}{z} = 0$  and for all  $y \in \mathbb{R}$ ,

$$f'(\theta) = \frac{f(y) - f(\theta)}{(y - \theta)} + h(y - \theta)$$

$$\iff f(y) = f(\theta) + f'(\theta)(y - \theta) + h(y - \theta).$$

In particular,

$$\sqrt{n}[f(Y_n) - f(\theta)] = \underbrace{f'(\theta)}_{\text{(constant)}} \underbrace{\sqrt{n}(Y_n - \theta)}_{\implies \mathcal{N}(0, \sigma^2)} + \underbrace{\sqrt{n}h(Y_n - \theta)}_{?}. \quad (10.2)$$

where we note that  $\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$  by assumption. Since it is multiplied by  $f'(\theta) \in \mathbb{R}$ , the product of these two terms converges to  $\mathcal{N}(0, \sigma^2[f'(\theta)]^2)$  by Slutsky's Theorem (Theorem 8.4.11(b)). We seek to show what happens to the third term of (10.2) as  $n \rightarrow \infty$  (the result follows if the term converges in probability to 0). Note that for any  $n \geq 1$  and for any  $t > 0$ ,

$$\begin{aligned} \Pr(\sqrt{n}|h(Y_n - \theta)| > t) &= \Pr\left(\sqrt{n}|h(Y_n - \theta)| > t \cap |Y_n - \theta| > \frac{t}{\sqrt{n}}\right) + \Pr\left(\sqrt{n}|h(Y_n - \theta)| > t \cap |Y_n - \theta| \leq \frac{t}{\sqrt{n}}\right) \\ &\iff \Pr(\sqrt{n}|h(Y_n - \theta)| > t) \leq \Pr(|Y_n - \theta| > t/\sqrt{n}) + \Pr(\sqrt{n}|h(Y_n - \theta)| > t \cap |Y_n - \theta| \leq t/\sqrt{n}). \end{aligned} \quad (10.3)$$

Since we already have by assumption  $\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ , it follows that  $|Y_n - \theta| \xrightarrow{p} 0$ . (For completeness, a detailed argument is included in the below lemma.) It then follows that the second term converges in probability to 0 if  $\lim_{n \rightarrow \infty} \Pr(|Y_n - \theta| > t/\sqrt{n}) = 0$  because  $\lim_{z \rightarrow 0} h(z)/z = 0$ . Therefore for any  $t > 0$ ,

$$\lim_{n \rightarrow \infty} \Pr(\sqrt{n}|h(Y_n - \theta)| > t) = 0 \iff \sqrt{n}|h(Y_n - \theta)| \xrightarrow{p} 0$$

which yields the result by (10.2). □

**Lemma 10.2.12.** Under the same assumptions and notation as in Theorem 10.2.11,

$$\lim_{n \rightarrow \infty} \Pr(|Y_n - \theta| > t/\sqrt{n}) = 0$$

*Proof.* We will examine the behavior of the right side of (10.3) as  $n \rightarrow \infty$  by looking at the first term and showing that  $Y_n - \theta$  converges in probability to 0. If  $t > 0$ , then  $\Pr(|Y_n - \theta| > t) = \Pr(\sqrt{n}|Y_n - \theta| > t\sqrt{n})$ , and if  $c > 0$  is a constant, then for sufficiently large  $n$ , the last quantity is at most  $\Pr(\sqrt{n}|Y_n - \theta| > c)$ . So we have

$$\Pr(|Y_n - \theta| > t) = \Pr(\sqrt{n}|Y_n - \theta| > t\sqrt{n}) \leq \Pr(\sqrt{n}|Y_n - \theta| > c)$$

But as  $n \rightarrow \infty$ ,  $c$  can be any constant (arbitrarily large). So

$$\lim_{n \rightarrow \infty} \Pr(\sqrt{n}|Y_n - \theta| > t) \leq \int_c^{\infty} e^{-y^2/2} \frac{1}{\sqrt{2\pi}} dy.$$

Therefore

$$\lim_{n \rightarrow \infty} \Pr(|Y_n - \theta| > t/\sqrt{n}) = 0.$$

□

**Theorem 10.2.13 (Convergence Theorem with Bounded Moment, Theorem 4.16 in 541A notes.).** Let  $X_1, X_2, \dots$  be random variables that converge in distribution to a random variable  $X$ . Assume  $\exists \epsilon > 0, c < \infty$  such that  $\mathbb{E}(|X_n|^{1+\epsilon}) \leq c, \forall n \geq 1$ . Then

$$\mathbb{E}(X) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n).$$

*Proof.* In Heilman's Graduate Probability Notes, Theorem 1.59 and Exercise 3.8(iii).

□

If  $f'(\theta) = 0$  in the Delta Method, we can instead use a second order Taylor expansion as follows.

**Theorem 10.2.14 (Second Order Delta Method, Theorem 4.17 in Math 541A Notes.).** Let  $\theta \in \mathbb{R}$ . Let  $Y_1, Y_2, \dots$  be random variables such that  $\sqrt{n}(Y_n - \theta)$  converges in distribution to a mean zero Gaussian random variable with variance  $\sigma^2 > 0$ . Let  $f : \mathbb{R} \rightarrow \mathbb{R}$ . Assume that  $f''$  exists and is continuous,  $f'(\theta) = 0$  and  $f''(\theta) \neq 0$ . Then

$$n(f(Y_n) - f(\theta))$$

converges in distribution to a  $\chi_1^2$  random variable multiplied by  $\sigma^2 \frac{1}{2}|f''(\theta)|$  as  $n \rightarrow \infty$ .

*Proof.* Using a second order Taylor expansion of  $f$ , there exists a random  $Z_n$  between  $\theta$  and  $Y_n$  such that

$$f(Y_n) = f(\theta) + f'(\theta)(Y_n - \theta) + \frac{1}{2}f''(Z_n)(Y_n - \theta)^2 = f(\theta) + \frac{1}{2}f''(Z_n)(Y_n - \theta)^2 \quad (10.4)$$

where the second equality follows because  $f'(\theta) = 0$ . As in the proof of Theorem 10.2.11,  $Z_n \xrightarrow{P} \theta$ . Since  $f''$  is continuous,  $f''(Z_n)$  converges in probability to  $f''(\theta)$  by Proposition 2.36 in the Math 541A notes (Theorem 8.4.13, continuous functions conserve convergence in probability). Therefore using (10.4),

$$n(f(Y_n) - f(\theta)) = \frac{1}{2}f''(Z_n) \cdot n(Y_n - \theta)^2$$

Note that  $\sqrt{n}(Y_n - \theta)$  converges in distribution to a mean zero Gaussian random variable by assumption, so  $n(Y_n - \theta)^2$  converges in distribution to a  $\chi_1^2$  random variable by Proposition 2.36 in the Math 541A notes (Theorem 8.4.13). So since  $f''(Z_n)$  converges in probability to a constant, by Proposition 2.36 in the Math 541A notes (Slutsky's Theorem, Theorem 8.4.11), the right side converges in probability to  $\frac{1}{2}f''(\theta)\sigma$  multiplied by a  $\chi_1^2$  random variable.

□

### 10.2.2 Simulation of Random Variables

**Proposition 10.2.15.** If  $X : \Omega \rightarrow \mathbb{R}$  is an arbitrary random variable with cumulative distribution function  $F : \mathbb{R} \rightarrow [0, 1]$ , then the function  $F^{-1}$  (if it exists) is a random variable on  $[0, 1]$  with the uniform probability law on  $(0, 1)$  that is equal in distribution to  $X$ .

*Proof.* Starting with the cdf of  $F^{-1}(u)$ ,

$$\Pr(s \in [0, 1] : F^{-1}(s) \leq t) = \Pr(s \in [0, 1] : F(t) > s) = F(t) = \Pr(\omega \in \Omega : X(\omega) \leq t)$$

where the third equality uses the definition of a uniform probability law on  $(0, 1)$ .

□

**Remark.** If  $F^{-1}$  does not exist, it can still work if you construct a generalized inverse of  $F$  as follows:

**Proposition 10.2.16 (Exercise 4.20 in Math 541A notes).** Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable on a sample space  $\Omega$  equipped with a probability law  $\mathbb{P}$ . For any  $t \in \mathbb{R}$  let  $F(t) := \mathbb{P}(X \leq t)$ . For any  $s \in (0, 1)$  define

$$Y(s) := \sup\{t \in \mathbb{R} : F(t) < s\}.$$

So  $Y$  is a random variable on  $(0, 1)$  with the uniform probability law on  $(0, 1)$ . Then  $X$  and  $Y$  are equal in distribution. That is,  $\mathbb{P}(Y \leq t) = F(t)$  for all  $t \in \mathbb{R}$ .

*Proof.* Note that since  $F$  is a cumulative distribution function,  $F$  is nondecreasing and  $F$  is right-continuous. So we have

$$\sup\{t \in \mathbb{R} : F(t) < s\} = \begin{cases} F^{-1}(s) & \text{if } F \text{ is strictly increasing (i.e. invertible) near } s \\ \inf\{x : F(x) = F(s)\} & \text{if } F \text{ is constant near } s \end{cases}$$

That is, the only time this quantity is different from  $F^{-1}(s)$  is when  $F^{-1}(\cdot)$  is undefined because  $F$  is constant on some interval around  $s$ . But if that is the case,  $F(\sup\{t \in \mathbb{R} : F(t) < s\}) = F(\inf\{x : F(x) = F(s)\}) = s$  anyway. With that in mind we proceed:

$$\mathbb{P}(Y \leq t) = \mathbb{P}(s \in (0, 1) : Y(s) \leq t) = \mathbb{P}(s \in (0, 1) : \sup\{t' \in \mathbb{R} : F(t') < s\} \leq t)$$

$$= \mathbb{P}(s \in (0, 1) : F(\sup\{t' \in \mathbb{R} : F(t') < s\}) \leq F(t)) = \mathbb{P}(s \in (0, 1) : s \leq F(t))$$

$$= \mathbb{P}(s \in (0, 1) : F(t) > s) = F(t) = \Pr(\omega \in \Omega : X(\omega) \leq t).$$

□

**Example 10.1 (Example 4.22 in Math 541A notes).** Let  $X$  be an exponential random variable with parameter 1.

$$\Pr(X \leq t) = \int_0^t e^{-x} dx = [-e^{-x}]_0^t = 1 - e^{-t} = F(t)$$

We seek  $F^{-1}(t)$ :

$$1 - e^{-y} = t \iff e^{-y} = 1 - t \iff -y = \log(1 - t) \iff y = -\log(1 - t) \implies F^{-1}(t) = -\log(1 - t)$$

So to simulate an exponential random variable with parameter 1, sample  $-\log(1 - U)$  where  $U \sim U(0, 1)$ .

**Remark.** What if the cdf is hard to compute? For example, in a Gaussian distribution:

$$F(t) = \int_{-\infty}^t (2\pi)^{-1/2} \exp(-x^2/2) dx.$$

$F^{-1}$  cannot be described using elementary formulas, so  $F^{-1}(u)$  is not the best way to simulate a Gaussian random variable. When using the Central Limit Theorem approach (see 541A notes for details), Edgeworth expansion says: if we replace  $U_1, \dots, U_n$  with i.i.d.  $X_1, \dots, X_n$  and the first  $m$  moments of  $X_1$  agree with the first  $m$  moments of Gaussian random variables, then the error in the CLT approximation to a Gaussian is  $n^{-(m-1)/2}$ . (See [https://en.wikipedia.org/wiki/Edgeworth\\_series](https://en.wikipedia.org/wiki/Edgeworth_series).) But this is still inefficient, because one Gaussian sample requires  $n$  uniform samples.

**Proposition 10.2.17 (Box-Muller Algorithm).** Let  $U_1, U_2$  be independent random variable distributed in  $(0, 1)$ . Define

$$R := \sqrt{-2 \log(U_1)}$$

this density is something like  $e^{-x^2/2}$

$$\Psi := 2\pi U_2$$

$$X := R \cos(\Psi), \quad Y := R \sin(\Psi)$$

Then  $X, Y$  are independent standard Gaussian random variables.

*Proof.* Homework problem. □

## 10.3 Data Reduction

Suppose we have some data and an exponential family. We would like to find the parameter  $\theta$  among the exponential family that fits the data well. Suppose we have a large data set, maybe so large that you can't store all the data in RAM at once. What is the "least memory" or "most efficient" method for finding  $\theta$ ? The answer: try to find a statistic that captures all the relevant information about  $\theta$ . For example, to find the mean of a Gaussian sample, use the sample mean. You don't have to store all the raw data, you can just store the sample mean. The following is a generalization of this concept:

### 10.3.1 Sufficient Statistics

**Definition 10.8 (Sufficient Statistic; definition 5.1 in Math 541A notes).** Suppose  $X_1, \dots, X_n$  is a sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$  is a family of distributions (such as an exponential family). Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}^k$  so that  $Y := g(X_1, \dots, X_n)$  is a statistic. We say that  $Y$  is a **sufficient statistic** for  $\theta$  if for every  $y \in \mathbb{R}^k$  and for every  $\theta \in \Theta$ , the conditional distribution of  $(X_1, \dots, X_n)$  given  $Y = y$  (with respect to probabilities given by  $f_\theta$ ) does not depend on  $\theta$ . That is,  $Y$  provides sufficient information to determine  $\theta$  from  $X_1, \dots, X_n$ .

**Remark.** Based on a comment Heilman made on class, this definition assumes independence of the random variables? Basically everything in this class does?

**Definition 10.9 (Sufficiency (Math 541B in-class definition)).**  $T(X_1, \dots, X_n)$  is **sufficient** for  $\theta$  (more precisely, for the statistical model  $\{P_\theta, \theta \in \Theta\}$ ) if

$$\mathbb{P}_\theta((X_1, \dots, X_n) \in A \mid T(X_1, \dots, X_n) = t) = \mu_t(A).$$

In particular, this probability does not depend on  $\theta$ .

**Goldstein lecture:** Suppose we have a model  $\{f_\theta : \theta \in \Theta\}$  which we interpret as a set of densities or mass functions. We have  $\Theta \subset \mathbb{R}^p$ , and we know the model up to  $p$  parameters. Example; we have  $X_1, X_2, \dots, X_n \sim$  i.i.d.  $f_\theta$  where  $\theta \in (\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$ , where  $f_\theta \sim \mathcal{N}(\mu, \sigma^2)$ , then

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

**Example 10.2 (Example 5.2 in 541A notes).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a Bernoulli distribution with parameter  $0 < \theta < 1$ . Then  $Y := X_1 + \dots + X_n$  is sufficient for  $\theta$ .

**Proposition 10.3.1 (Example 5.2 in 541A notes).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a Bernoulli distribution with parameter  $0 < \theta < 1$ . Let  $Y := X_1 + \dots + X_n$ . Then

$$\mathbb{P}_\theta(X = x \mid Y = y) = \begin{cases} 0 & y \neq \sum_i x_i \\ \binom{n}{y} & \sum_i x_i = y \end{cases}$$

**Remark.** If a statistic is sufficient for  $\theta$ , then we can use that sufficient statistic to re-create the data (or re-create an equivalent data set with the same statistical properties as far we are concerned with estimating the parameter of interest).

*Proof.* Let  $x_1, \dots, x_n \in [0, 1]$ . Let  $0 \leq y \leq n$  be an integer. Then  $Y$  is binomial with parameters  $n$  and  $\theta$ . We may assume  $y = x_1 + \dots + x_n$ , otherwise there is nothing to show. Using the definition of conditional probability,

$$\begin{aligned} \Pr((X_1, \dots, X_n) = (x_1, \dots, x_n) \mid Y = y) &= \frac{1}{\Pr(Y = y)} \cdot \Pr((X_1, \dots, X_n) = (x_1, \dots, x_n) \cap Y = y) \\ &= \frac{1}{\Pr(Y = y)} \cdot \Pr((X_1, \dots, X_n) = (x_1, \dots, x_n)) \end{aligned}$$

Using independence and the definition of a binomial distribution, we have

$$\begin{aligned} &= \frac{1}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \cdot \prod_{i=1}^n \Pr(X_i = x_i) = \frac{1}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \cdot \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \frac{1}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \cdot \theta^y(1-\theta)^{n-y} = \frac{1}{\binom{n}{y}}. \end{aligned}$$

Since this expression does not depend on  $\theta$ ,  $Y$  is sufficient for  $\theta$ .

□

**Example 10.3 (Example 5.3 in 541A notes).** Let  $X_1, \dots, X_n$  be a sample of size  $n$  from a Gaussian distribution with known variance  $\sigma^2 > 0$  and unknown mean  $\mu \in \mathbb{R}$ . Then  $Y := (X_1, \dots, X_n)/n$  is a sufficient statistic for  $\mu$ .

*Proof.* Note that  $Y$  is a Gaussian random variable with mean  $\mu$  and variance  $\sigma^2/n$ . Let  $x_1, \dots, x_n \in \mathbb{R}$  and let  $y = (x_1 + \dots + x_n)/n$ . Then

$$f_{X_1, \dots, X_n|Y}(x_1, \dots, x_n \mid y) = \frac{1}{f_Y(y)} \cdot f_{X_1, \dots, X_n, Y}(x_1, \dots, x_n, y) = \frac{1}{f_Y(y)} \cdot f_{X_1, \dots, X_n}(x_1, \dots, x_n, y)$$

Since

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2 - \mu^2 + 2\mu x}{2\sigma^2}\right)$$

we have

$$= \frac{1}{f_Y(y)} \cdot \prod_{i=1}^n f_{X_i}(x_i) = \frac{1}{f_Y(y)} \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}(x_1^2 + \dots + x_n^2) - \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i\right)$$

$$\begin{aligned}
&= \frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}(x_1^2 + \dots + x_n^2) - \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i\right)}{n^{1/2}(\sigma^2 2\pi)^{-1/2} \exp\left(-\frac{n}{2\sigma^2}y^2 - \frac{n}{2\sigma^2}\mu^2 + \frac{n\mu}{\sigma^2}y\right)} \\
&= \frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}(x_1^2 + \dots + x_n^2)\right)}{n^{1/2}(\sigma^2 2\pi)^{-1/2} \exp\left(-\frac{n}{2\sigma^2}y^2\right)}
\end{aligned}$$

Because  $\mu$  does not appear in this expression,  $Y$  is sufficient for  $\mu$ .

□

**Theorem 10.3.2 (Theorem 6.2.2 in Casella and Berger [2001], not in 541A lecture notes).** If  $p(x | \theta)$  is the joint pdf or pmf of a random sample  $\mathbf{X} = X_1, \dots, X_n$  and  $q(t | \theta)$  is the pdf or pmf of the statistic  $T(\mathbf{X})$ , then  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if, for every  $\mathbf{x} = x_1, \dots, x_n$  in the sample space, the ratio  $p(\mathbf{x} | \theta)/q(T(\mathbf{x} | \theta))$  is constant as a function of  $\theta$ .

**Theorem 10.3.3 (Neyman-Fisher Factorization Theorem, Theorem 5.4 in 541A notes).** Suppose  $X = (X_1, \dots, X_n)$  is a random sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$  is a family of probability density functions or probability mass functions. Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}^k$ , so  $Y := t(X_1, \dots, X_n)$  is a statistic. Then  $Y$  is sufficient for  $\theta$  if and only if there exists a nonnegative  $\{g_\theta : \theta \in \Theta\}$ ,  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $g_\theta : \mathbb{R}^k \rightarrow \mathbb{R}$  such that

$$f_\theta(x) = g_\theta(t(x))h(x), \quad \forall x \in \mathbb{R}^n, \quad \forall \theta \in \Theta. \quad (10.5)$$

*Proof.* We will prove only the discrete case to avoid measure theory. For a general case, see Keener Section 6.4.

Suppose  $Y$  is sufficient. Let  $x \in \mathbb{R}^n$ . Note that by definition and using  $Y = t(X)$ ,

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x \cap t(X) = t(x)) = \mathbb{P}_\theta(Y = t(x))\mathbb{P}_\theta(X = x | Y = t(x))$$

By sufficiency,  $\mathbb{P}_\theta(X = x | Y = t(x))$  does not depend on  $\theta$ . Therefore we can satisfy (10.5) with  $g_\theta(t(x)) = \mathbb{P}_\theta(Y = t(x))$ ,  $h(x) = \mathbb{P}_\theta(X = x | Y = t(x))$ , so the factorization holds.

Now suppose there exists a nonnegative  $\{g_\theta : \theta \in \Theta\}$ ,  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $g_\theta : \mathbb{R}^k \rightarrow \mathbb{R}$  such that

$$f_\theta(x) = g_\theta(t(x))h(x), \quad \forall x \in \mathbb{R}^n, \quad \forall \theta \in \Theta.$$

Define  $r_\theta(z) := \mathbb{P}_\theta(t(X) = z) \quad \forall z \in \mathbb{R}^k$  (the probability mass function for  $t(X)$ ). Also define  $t^{-1}t(x) := \{y \in \mathbb{R}^n ; t(y) = t(x)\} \quad \forall x \in \mathbb{R}^n$ . To show sufficiency, we need to show that  $\mathbb{P}_\theta(X = x | Y = t(x))$  does not depend on  $\theta$ . Note that

$$\mathbb{P}_\theta(X = x \mid Y = t(x)) = \frac{f_\theta(x)}{f_Y(t(x))} = \frac{f_\theta(x)}{r_\theta(t(x))}$$

Using our assumption and the Total Probability Theorem, we have

$$= \frac{g_\theta(t(x)h(x))}{\mathbb{P}_\theta(t(X) = t(x))} = \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} \mathbb{P}_\theta(X = z)} = \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} f_\theta(z)} = \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} g_\theta(t(z))h(z)}$$

By definition of  $t^{-1}t(z)$ , we can write this as

$$= \frac{g_\theta(t(x))h(x)}{\sum_{z \in t^{-1}t(x)} g_\theta(t(x))h(z)} = \frac{g_\theta(t(x))h(x)}{g_\theta(t(x)) \sum_{z \in t^{-1}t(x)} h(z)} = \frac{h(x)}{\sum_{z \in t^{-1}t(x)} h(z)}$$

where the second-to-last step follows since  $t(x)$  is constant for all  $z \in t^{-1}t(x)$ . Since this expression does not contain  $\theta$ ,  $Y$  is sufficient for  $\theta$ .

□

**Remark.** Intuition: data only cares about  $\theta$  through  $t(x)$ .

To use the Factorization Theorem (Theorem 10.5) to find a sufficient statistic, we factor the joint pdf of the sample into two parts, with one part not depending on  $\theta$ . The other part, the one that depends on  $\theta$ , usually depends on the sample only through the function  $t(x)$ , and this function is a sufficient statistic for  $\theta$ .

**Exercise 17.** Suppose  $X_1, X_2, \dots, X_n \sim$  i.i.d.  $\mathcal{N}(0, 1)$ . So density is

$$\frac{1}{\sqrt{2\pi}} e^{-1/2(x-\theta)^2}$$

Show that

$$e^{-1/2(x^2 - 2x\theta + \theta^2)} =$$

$$f_\theta(x) = \left(\frac{1}{2\pi}\right)^{n/2} e^{2/12 \sum_i X_i^2} e^{\theta \sum_i X_i - n\theta^2/2}$$

so if  $t(x) = \sum_{i=1}^n X_i$ ,  $h(x) = \left(\frac{1}{2\pi}\right)^{n/2} e^{2/12 \sum_i X_i^2}$ ,  $g_\theta(t(x)) = e^{\theta \sum_i X_i - n\theta^2/2}$ , then by the Factorization Theorem (Theorem 10.5) this  $(\bar{x})$  is a sufficient statistic.

**Remark.** In this case, if we deleted the original data we could recreate the original data by sampling from a  $\mathcal{N}(0, 1)$  distribution, then add the difference between the mean we get and the original sample mean to get an equivalent data set to the original one.

**Remark.** Suppose we define  $t(x) := x$ ,  $\forall x \in \mathbb{R}^n$ . Then  $Y = t(X_1, \dots, X_n) = (X_1, \dots, X_n)$  is (trivially) sufficient for  $\theta$ . In general there will be infinitely many sufficient statistics for  $\theta$ . For instance, in Example 10.3.1,  $(X_1 + \dots + X_n)^2$  is also sufficient. So is  $(X_1 + \dots + X_n)^3$ , etc. More generally, any invertible function of any sufficient statistic is itself sufficient.

We can see that  $(X_1, \dots, X_n)$  is sufficient for  $\theta$  if  $(t(x_1, \dots, x_n)) = (x_1, \dots, x_n)$ ,  $g_\theta = f_\theta$ ,  $h = 1$ . But this is not really helpful. We see we are interested in sufficient statistics that are smaller—reduce the data (in some sense) as much as possible.

### 10.3.2 Minimal Sufficient Statistics

**Proposition 10.3.4.** Suppose  $X = (X_1, \dots, X_n)$  is a random sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$  is a family of probability density functions or probability mass functions. Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}^k$ . Let  $Y := t(X_1, \dots, X_n)$ . Assume  $Y$  is sufficient of  $\theta$ . Let  $a : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , let  $Z := u(X_1, \dots, X_n)$ . suppose there exists  $r : \mathbb{R}^m \rightarrow \mathbb{R}^k$  such that  $r(u(x)) = t(x)$  for all  $x \in \mathbb{R}^n$ . That is, suppose  $Y = r(Z)$ . Then  $Z$  is sufficient for  $\theta$ .

*Proof.*

$$f_\theta(x) = g_\theta(t(x))h(x) = g_\theta(r(u(x)))h(x)$$

there exists  $g_\theta : \mathbb{R}^k \rightarrow [0, \infty)$ .  $Y$  is sufficient.

Define

$$\tilde{g}_\theta(y) := g_\theta(r(y)) \quad \forall y \in \mathbb{R}^m$$

So

$$f_\theta(x) = \tilde{g}_\theta(u(x))h(x) \quad \forall x \in \mathbb{R}^n.$$

□

**Definition 10.10 (Minimal sufficient statistic).** Suppose  $X = (X_1, \dots, X_n)$  is a random sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$  is a family of probability density functions or probability mass functions. Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}^k$ . Let  $Y := t(X_1, \dots, X_n)$ . Assume  $Y$  is sufficient of  $\theta$ . Then  $Y$  is a **minimal sufficient statistic** for  $\theta$  if for every statistic  $Z : \Omega \rightarrow \mathbb{R}^m$  that is sufficient for  $\theta$  there exists a function  $\mathbb{R}^m \rightarrow \mathbb{R}^k$  such that  $Y = r(Z)$ .

**Remark.** Minimal sufficient statistics are not in general unique (because if you take any one-to-one function you get another one), but they are unique up to invertible transformations. (This is true because if  $Y$  and  $Z$  are both minimal sufficient,  $Y = r(Z)$  and  $Z = s(Y)$ , so  $Y = r(s(Y))$ ,  $Z = s(r(Z))$ ). They exist under mild assumptions (for a family of densities or probability mass functions).

**Proposition 10.3.5 (Proposition Larry Goldstein gave in class; Proposition 5.12 in notes).**

Suppose  $X_1, \dots, X_n$  is a random sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$ , is a

family of probability density functions or probability mass functions ( $\Theta \in \mathbb{R}^n$ ). (In the case of probability mass functions, we also assume that the set  $\cup_{\theta \in \Theta} \{x \in \mathbb{R}^n : f_\theta(x) > 0\}$  is countable.) Then there exists a statistic  $Y$  that is minimal sufficient for  $\theta$ .

*Proof where  $\theta$  is countable.* By relabeling, let  $\Theta = \{1, 2, \dots\}$ . We say for  $x, y$  sequences, we define the equivalence relation  $x \sim y$  if  $\exists \alpha \in \mathbb{R}$  such that  $x = \alpha y$ . Finite

$$t : \mathbb{R}^n \rightarrow \mathbb{R}^m / \sim, \quad \Theta = \{1, \dots, m\}$$

$$t(x) = (f_1(x), f_2(x), \dots, f_m(x))$$

these likelihood are multiples of each other where  $\alpha$  is a constant. The likelihood ratio is a constant not depending on  $\theta$ . If they have the same  $t(x)$  then we have that.

□

**Theorem 10.3.6 (Theorem 5.8 in 541A notes).** Let  $\{f_\theta : \theta \in \Theta\}$  be a family of probability density functions or probability mass functions. Let  $X_1, \dots, X_n$  be a random sample from a member of the family. Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}^m$  and define  $Y := t(X_1, \dots, X_n)$ . Assume that  $Y$  is sufficient for  $\theta$ .  $Y$  is minimal sufficient if and only if the following condition holds for every  $x, y \in \mathbb{R}^n$ :

There exists  $c(x, y) \in \mathbb{R}$  that does not depend on  $\theta$  such that  $f_\theta(x) = c(x, y)f_\theta(y) \quad \forall \theta \in \Theta$

if and only if

$$t(x) = t(y).$$

*Proof.* We are only considering probability mass functions to make things easier. We first prove sufficiency. We will show that the condition holding implies that  $Y$  is minimal sufficient.

Recall the likelihood ratio:

$$\frac{f_\theta(x)}{f_\theta(y)}$$

Note that the condition is equivalent to the likelihood ratio not depending on  $\theta$  if and only if  $t(x) = t(y)$ . Consider the range  $R = \{t(x) : x \in \mathbb{R}^n\}$  and then for  $t \in R$  let  $S_t = \{y : S(y) = t\}$ . If  $t$  is in  $R$ , then there must be some  $z$  so that  $t(z)$  is that  $t$ . This ensures that  $S_t$  is nonempty (there is at least one  $z$  so that  $t(z) = t$ ). Let  $t(x) \in R$ , then  $S_{t(x)}$  is nonempty (in particular it contains  $x$ ). Pick any  $y$  you like in  $t(x)$ :  $y \in S$ .  $S$  depends on  $t(x)$  so we can index it by  $t(x)$ :  $y_{t(x)} \in S_{t(x)}$ . let  $y_t \in S_t$ . Note that

$$t(y_{t(x)}) = t(x)$$

But now by the assumption, we have

There exists  $c(x, y_{t(x)}) \in \mathbb{R}$  that does not depend on  $\theta$  such that  $f_\theta(x) = c(x, y_{t(x)})f_\theta(y_{t(x)}) \quad \forall \theta \in \Theta$

Then note that if  $h(x) = c(x, y_{t(x)})$ ,  $g_\theta(t) = f_\theta(y_t) \iff g_\theta(t(x)) = f_\theta(y_{t(x)})$ , we meet the conditions for the Factorization Theorem (Theorem 10.5). So using the Factorization Theorem,  $Y$  is sufficient.

⋮

**Part we did in class on Friday 02/15:** evidently (according to Goldstein) this shows that the statistic is minimal but not necessarily sufficient. Let  $Z = u(X_1, \dots, X_n)$  be any other sufficient statistic. We need to eventually show that  $Y$  is a function of  $Z$ . By the Factorization Theorem (Theorem 10.5), there exists  $h : \mathbb{R}^m \rightarrow \mathbb{R}, g_\theta : \mathbb{R}^n \rightarrow \mathbb{R}$  such that for all  $\theta \in \Theta$ ,

$$f_\theta(x) = g'_\theta(u(x))h'(x), \quad \forall x \in \mathbb{R}^n.$$

Let  $y \in \mathbb{R}^n$ . If  $h'(y) = 0$ , then  $f_\theta(y) = 0$  for all  $\theta \in \Theta$ . So,  $\mathbb{P}_\theta(y \in \mathbb{R}^n : h'(y) = 0) = 0$  for all  $\theta \in \Theta$ . So we can ignore this possibility since it's a probability 0 event and assume  $h'(y) > 0, \forall y \in \mathbb{R}^n$ .

Now let  $x, y \in \mathbb{R}^n$  such that  $u(x) = u(y)$ . **By an exercise we're going to do later**, if  $t(x) = t(y)$  then  $t$  is a function of  $u$ , so we will be done if we can show that  $t(x) = t(y)$ . Note that since  $u(x) = u(y)$ , for any  $\theta \in \Theta$

$$f_\theta(x) = g'_\theta(u(x))h'(x) = \frac{g'_\theta(u(y))h'(x)}{f_\theta(y)} \frac{h'(x)}{h'(y)} = f_\theta(y) \frac{h'(x)}{h'(y)}, \quad \text{for all } \theta \in \Theta$$

So define  $c(x, y) = h'(x)/h'(y)$ , we have

$$f_\theta(x) = f_\theta(y)c(x, y), \quad \forall \theta \in \Theta$$

Therefore  $t(x) = t(y)$ , so we're done showing that if the condition holds then  $Y$  is minimal sufficient.

Then next thing to show is that if  $Y$  is minimal sufficient then the condition holds.

⋮

For any  $z \in \{t(x) : x \in \mathbb{R}^n\}$ , let  $x_z$  be any element of  $t^{-1}(z)$

□

**Proposition 10.3.7 (Exercise 5.10 in Math 541A notes).** Let  $\{f_\theta : \theta \in \Theta\}$  be a  $k$ -parameter exponential family  $\{f_\theta : \theta \in \Theta, a(w(\theta)) < \infty\}$  of probability density functions or probability mass functions, where

$$f_\theta(x) := h(x) \exp \left( \sum_{i=1}^k w_i(\theta) t_i(x) - a(w(\theta)) \right), \quad \forall x \in \mathbb{R}.$$

For any  $\theta \in \Theta$ , let  $w(\theta) = (w_1(\theta), \dots, w_k(\theta))$ . Assume that the following subset of  $\mathbb{R}^k$  is  $k$ -dimensional:

$$\{(w_1(\theta), \dots, w_k(\theta)) \in \mathbb{R}^k : \theta \in \Theta\}.$$

That is, if  $x \in \mathbb{R}^k$  satisfies  $\langle x, y \rangle = 0$  for all  $y$  in this set, then  $x = 0$ .

Let  $X = (X_1, \dots, X_n)$  be a random sample of size  $n$  from  $f_\theta$ . Define  $t : \mathbb{R}^n \rightarrow \mathbb{R}^n$  by

$$t(X) := \sum_{j=1}^n (t_1(X_j), \dots, t_n(X_j)).$$

Then  $t(X)$  is minimal sufficient for  $\theta$ .

*Proof.* First note that  $t(X)$  is sufficient by the Factorization Theorem (Theorem 10.5) because we have for any  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$

$$\begin{aligned} f_\theta(x) &= \prod_{j=1}^n \left[ h(x_j) \exp \left( \sum_{i=1}^k w_i(\theta) t_i(x_j) - a(w(\theta)) \right) \right] = \exp \left( \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) - n \cdot a(w(\theta)) \right) \prod_{j=1}^n h(x_j) \\ &= g_\theta(t(x)) h(x), \quad \forall x \in \mathbb{R}^n \end{aligned}$$

where

$$h(x) = \prod_{j=1}^n h(x_j), \quad g_\theta(t(x)) = \exp \left( \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) \right).$$

Now we will show minimal sufficiency using Theorem 5.8 from the lecture notes (Theorem 10.3.6). We seek to show that for every  $x, y \in \mathbb{R}^n$ , the likelihood ratio  $\frac{f_\theta(x)}{f_\theta(y)}$  is constant (the constant may depend on  $x$  and  $y$ ) if and only if  $t(x) = t(y)$ . Let  $x, y \in \mathbb{R}^n$ . Suppose there is some constant  $c(x, y) > 0$  that may depend on  $x$  and  $y$  but not  $\theta$  such that

$$\begin{aligned} &\exp \left( \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) - n \cdot a(w(\theta)) \right) \prod_{j=1}^n h(x_j) \\ &= c(x, y) \exp \left( \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(y_j) - n \cdot a(w(\theta)) \right) \prod_{j=1}^n h(y_j) \end{aligned}$$

$$\iff \exp \left( \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) \right) = c_1(x, y) \exp \left( \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(y_j) \right)$$

(where cancellation of the  $h$  functions is permissible since they depend only on  $x$  and  $y$ , so we can let  $c_1(x, y) = c(x, y) \cdot \prod_{j=1}^n h(y_j) / \prod_{j=1}^n h(x_j) > 0$ )

$$\iff \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(x_j) = c_2(x, y) + \sum_{j=1}^n \sum_{i=1}^k w_i(\theta) t_i(y_j)$$

where  $c_2(x, y) = \log(c_1(x, y))$ . Then if  $\theta_0, \theta_1$  are any two points in  $\Theta$ ,

$$\sum_{j=1}^n \sum_{i=1}^k w_i(\theta_0) t_i(x_j) - \sum_{j=1}^n \sum_{i=1}^k w_i(\theta_1) t_i(x_j) = \sum_{j=1}^n \sum_{i=1}^k w_i(\theta_0) t_i(y_j) - \sum_{j=1}^n \sum_{i=1}^k w_i(\theta_1) t_i(y_j)$$

(where the  $c_2(x, y)$  terms cancel since  $(x, y)$  is held fixed.)

$$\begin{aligned} \iff \sum_{j=1}^n \sum_{i=1}^k [w_i(\theta_0) - w_i(\theta_1)] t_i(x_j) &= \sum_{j=1}^n \sum_{i=1}^k [w_i(\theta_0) - w_i(\theta_1)] t_i(y_j) \\ \iff \sum_{j=1}^n \sum_{i=1}^k [w_i(\theta_0) - w_i(\theta_1)] [t_i(x_j) - t_i(y_j)] &= 0. \end{aligned}$$

This equation holds for all  $\theta \in \Theta$  if and only if  $t_i(x_j) - t_i(y_j) = 0, i = 1, \dots, k, j = 1, \dots, n$ ; that is, it holds if and only if  $t(x) = t(y)$ . Therefore we have shown that  $t(\cdot)$  is minimal sufficient by Theorem 10.3.6.

□

**Remark.** Note that the assumption of the exercise is always satisfied for an exponential family in canonical form. From this proposition we can conclude that if we sample from a Gaussian with unknown mean  $\mu$  and variance  $\sigma^2 > 0$ , then  $\bar{X}$  is minimal sufficient for  $\theta$  and  $(\bar{X}, S)$  is minimal sufficient for  $(\mu, \sigma^2)$ .

**Proposition 10.3.8 (Exercise 5.13 in Math 541A notes).** Let  $\mathbb{P}_1, \mathbb{P}_2$  be two probability laws on the sample space  $\Omega = \mathbb{R}$ . Suppose these laws have densities  $f_1, f_2 : \mathbb{R} \rightarrow [0, \infty)$  so that

$$\mathbb{P}_i(A) = \int_A f_i(x) dx, \quad \forall i = 1, 2, \quad \forall A \subseteq \mathbb{R}.$$

Then

(a)

$$\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \frac{1}{2} \int_{\mathbb{R}} |f_1(x) - f_2(x)| dx.$$

(b) If  $\mathbb{P}_1, \mathbb{P}_2$  are probability laws on  $\Omega = \mathbb{Z}$

$$\sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \frac{1}{2} \sum_{z \in \mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)|.$$

*Proof.* (a) Note that  $\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$  returns the difference in areas under  $f_1$  and  $f_2$  in the region  $A \subset \mathbb{R}$  where that difference is positive. Suppose without loss of generality that

$$\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A \subseteq \mathbb{R}} \{\mathbb{P}_1(A) - \mathbb{P}_2(A)\}. \quad (10.6)$$

(There is no loss of generality because in the case that  $\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A \subseteq \mathbb{R}} \{\mathbb{P}_2(A) - \mathbb{P}_1(A)\}$ , we can simply switch the names of  $\mathbb{P}_1$  and  $\mathbb{P}_2$  to get the desired result.) Then the region  $A$  which maximizes the quantity on the right side of (10.6) is the suggested region,  $A := \{x \in \mathbb{R} : f_1(x) > f_2(x)\}$ . That is,

$$\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \int_A (f_1(x) - f_2(x))dx.$$

Note that

$$\begin{aligned} \int_{\mathbb{R}} |f_1(x) - f_2(x)|dx &= \int_A |f_1(x) - f_2(x)|dx + \int_{\mathbb{R} \setminus A} |f_1(x) - f_2(x)|dx \\ \iff \int_{\mathbb{R}} |f_1(x) - f_2(x)|dx &= \int_A (f_1(x) - f_2(x))dx + \int_{\mathbb{R} \setminus A} (f_2(x) - f_1(x))dx. \end{aligned} \quad (10.7)$$

Since we already have  $\int_A (f_1(x) - f_2(x))dx = \sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$ , if it is also true that  $\int_{\mathbb{R} \setminus A} (f_2(x) - f_1(x))dx = \sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$  we are done by (10.7), so this is what we will show next. Let  $\int_A f_2(x)dx = a_2$  and let  $\int_A f_1(x)dx = a_1$ , so that

$$\sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \int_A (f_1(x) - f_2(x))dx = a_1 - a_2.$$

Note that

$$1 = \int_{\mathbb{R}} f_2(x)dx = \int_A f_2(x)dx + \int_{\mathbb{R} \setminus A} f_2(x)dx = a_2 + \int_{\mathbb{R} \setminus A} f_2(x)dx \iff \int_{\mathbb{R} \setminus A} f_2(x)dx = 1 - a_2,$$

and similarly  $\int_{\mathbb{R} \setminus A} f_1(x)dx = 1 - a_1$ . Therefore

$$\int_{\mathbb{R} \setminus A} (f_2(x) - f_1(x))dx = \int_{\mathbb{R} \setminus A} f_2(x)dx - \int_{\mathbb{R} \setminus A} f_1(x)dx = 1 - a_2 - (1 - a_1) = a_1 - a_2 = \sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|.$$

So by (10.7), we have

$$\int_{\mathbb{R}} |f_1(x) - f_2(x)|dx = 2 \sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| \iff \sup_{A \subseteq \mathbb{R}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \frac{1}{2} \int_{\mathbb{R}} |f_1(x) - f_2(x)|dx.$$

(b) Analogous to the proof of (a). Note that  $\sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$  returns the difference in probabilities for those numbers in the set  $A \subset \mathbb{Z}$  where that difference is positive. Suppose without loss of generality that

$$\sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A \subseteq \mathbb{Z}} \{\mathbb{P}_1(A) - \mathbb{P}_2(A)\}. \quad (10.8)$$

(There is no loss of generality because in the case that  $\sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sup_{A \subseteq \mathbb{Z}} \{\mathbb{P}_2(A) - \mathbb{P}_1(A)\}$ , we can simply switch the names of  $\mathbb{P}_1$  and  $\mathbb{P}_2$  to get the desired result.) Then the set  $A$  which maximizes the quantity on the right side of (10.8) can be defined as (similarly to part (a))  $A := \{z \in \mathbb{Z} : \mathbb{P}_1(z) > \mathbb{P}_2(z)\}$ . That is,

$$\sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sum_{z \in A} \mathbb{P}_1(z) - \mathbb{P}_2(z).$$

Note that

$$\begin{aligned} \sum_{z \in \mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| &= \sum_{z \in A} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| + \sum_{z \in \{\mathbb{Z} \setminus A\}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| \\ \iff \sum_{z \in \mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| &= \sum_{z \in A} \mathbb{P}_1(z) - \mathbb{P}_2(z) + \sum_{z \in \{\mathbb{Z} \setminus A\}} (\mathbb{P}_2(z) - \mathbb{P}_1(z)). \end{aligned} \quad (10.9)$$

Since we already have  $\sum_{z \in A} \mathbb{P}_1(z) - \mathbb{P}_2(z) = \sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$ , if it is also true that  $\sum_{z \in \{\mathbb{Z} \setminus A\}} (\mathbb{P}_2(z) - \mathbb{P}_1(z)) = \sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$  we are done by (10.9), so this is what we will show next. Let  $\sum_{z \in A} \mathbb{P}_2(z) = a_2$  and let  $\sum_{z \in A} \mathbb{P}_1(z) = a_1$ , so that

$$\sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \sum_{z \in A} \mathbb{P}_1(z) - \mathbb{P}_2(z) = a_1 - a_2.$$

Note that

$$1 = \sum_{z \in \mathbb{Z}} \mathbb{P}_2(z) = \sum_{z \in A} \mathbb{P}_2(z) + \sum_{z \in \{\mathbb{Z} \setminus A\}} \mathbb{P}_2(z) = a_2 + \sum_{z \in \{\mathbb{Z} \setminus A\}} \mathbb{P}_2(z) \iff \sum_{z \in \{\mathbb{Z} \setminus A\}} \mathbb{P}_2(z) = 1 - a_2,$$

and similarly  $\sum_{z \in \{\mathbb{Z} \setminus A\}} \mathbb{P}_1(z) = 1 - a_1$ . Therefore

$$\sum_{z \in \{\mathbb{Z} \setminus A\}} (\mathbb{P}_2(z) - \mathbb{P}_1(z)) = \sum_{z \in \{\mathbb{Z} \setminus A\}} \mathbb{P}_2(z) - \sum_{z \in \{\mathbb{Z} \setminus A\}} \mathbb{P}_1(z) = 1 - a_2 - (1 - a_1) = a_1 - a_2 = \sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|.$$

So by (10.9), we have

$$\sum_{z \in \mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)| = 2 \sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| \iff \sup_{A \subseteq \mathbb{Z}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| = \frac{1}{2} \sum_{z \in \mathbb{Z}} |\mathbb{P}_1(z) - \mathbb{P}_2(z)|.$$

□

### 10.3.3 Ancillary Statistics

**Definition 10.11 (Ancillary Statistic).** Suppose  $X_1, \dots, X_n$  is a random sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$  is a family of distributions. A statistic  $Y = t(X_1, \dots, X_n)$ ,  $t : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is **ancillary** for  $\theta$  if the distribution of  $Y$  does not depend on  $\theta$ .

**Example 10.4. [Example from 541B lecture]**

$X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ . Then  $V(x) := X_1 - \bar{X}$  is ancillary. Proof: let  $X_d = \mu + X'_d, X'_d \sim \mathcal{N}(0, 1)$ . Then

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j = \mu + \bar{X}'.$$

so

$$V(X_1, \dots, X_n) = \mu + X'_1 - \mu - \bar{X}' = X'_1 - \bar{X}'$$

Since  $X'$  does not depend on  $\mu$ ,  $V(X_1, \dots, X_n)$  is ancillary for  $\mu$ .

**Example 10.5 (Example 5.15 from 541A notes).** Let  $X_1, \dots, X_n$  be random sample of size  $n$  from the location family for the Cauchy distribution:

$$f_\theta(x) := \prod_{i=1}^n \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2}, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \forall \theta \in \mathbb{R}.$$

Then the order statistics  $X_{(1)} \leq \dots \leq X_{(n)}$  are minimal sufficient for  $\theta$ .

*Proof.* Sufficiency follows by the Factorization Theorem (Theorem 10.5, Theorem 5.4 in Math 541A notes) (**usually the easiest way to prove sufficiency**) since if  $t(x) := (x_{(1)}, \dots, x_{(n)})$ , then  $f_\theta(t(x)) = f_\theta(x)$  (because  $t(x)$  is just a permutation so the product won't change).

To get minimal sufficiency, apply Theorem 10.3.6 (Theorem 5.8 in 541A notes). Recall we have minimal sufficiency if the following condition holds for every  $x, y \in \mathbb{R}^n$ :

There exists  $c(x, y) \in \mathbb{R}$  that does not depend on  $\theta$  such that  $f_\theta(x) = c(x, y)f_\theta(y) \quad \forall \theta \in \Theta$

if and only if

$$t(x) = t(y).$$

Let's try to show it.

$$\frac{f_\theta(x)}{f_\theta(y)} = \prod_{i=1}^n \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2} \Bigg/ \prod_{i=1}^n \frac{1}{\pi} \frac{1}{1 + (y_i - \theta)^2} = \prod_{i=1}^n [1 + (y_i - \theta)^2] \Bigg/ \prod_{i=1}^n [1 + (x_i - \theta)^2] \quad (10.10)$$

Keep  $x, y$  fixed,  $\theta \in \mathbb{R}$  variable. Then the likelihood ratio (10.10) does not depend on  $\theta$  if and only if the roots in  $\theta$  on top are equal to the roots on bottom. Roots on top:  $\theta = y_i \pm \sqrt{-i}, 1 \leq i \leq n$ . Roots on bottom:  $\theta = x_i \pm \sqrt{-i}, 1 \leq i \leq n$ . So we can see this is true if and only if the vector  $(x_1, \dots, x_n)$  is a permutation of  $(y_1, \dots, y_n)$ , which is exactly the case if  $t(x) = t(y)$ .

□

However, this statistic has ancillary information. Specifically,  $X_{(n)} - X_{(1)}$  is ancillary (its distribution does not depend on  $\theta$ ).

*Proof.* Let  $Z_1, \dots, Z_n$  be independent centered Cauchy random variables; that is, they all have density  $\frac{1}{\pi} \frac{1}{1+a^2}$ ,  $a \in \mathbb{R}$ . Then  $X_i = Z_i + \theta$ ,  $\forall 1 \leq i \leq n, \forall \theta \in \mathbb{R}$ . Also,  $X_{(i)} = Z_{(i)} + \theta$ . So  $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$  does not depend on  $\theta \in \mathbb{R}$ . So,  $X_{(n)} - X_{(1)}$  is ancillary for  $\theta$ . That is, there exists a constant  $c$  that does not depend on  $\theta$  such that  $\mathbb{E}_\theta[X_{(n)} - X_{(1)} - c] = 0$  for all  $\theta \in \mathbb{R}$ .

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x_1, \dots, x_n) = x_n - x_1 - c \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n$ . Then  $\mathbb{E}_\theta f(Y) = 0, \forall \theta \in \Theta, Y = (X_{(1)}, \dots, X_{(n)})$ . Note that  $f \neq 0$  (in fact  $f(Y) \neq 0$  with probability 1).

□

#### 10.3.4 Complete Statistics

**Definition 10.12 (Complete statistic; definition 5.16 in 541A notes).** Suppose  $X_1, \dots, X_n$  is a random sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$  is a family of distributions. A statistic  $Y = t(X_1, \dots, X_n), t : \mathbb{R}^n \rightarrow \mathbb{R}^k$  is **complete** for  $\theta$  if the following holds:

For any  $f : \mathbb{R}^k \rightarrow \mathbb{R}$  such that  $\mathbb{E}_\theta f(Y) = 0 \quad \forall \theta \in \Theta$ , it holds that  $f(Y) = 0$ .

Intuition:  $Y$  has no “excess information.”

**Definition 10.13 (541B definition).**  $T$  is **complete** for  $\{f_\theta, \theta \in \Theta\}$  if

$$\mathbb{E}_\theta \phi(T) = 0 \quad \forall \theta \implies \phi(T) = 0 \text{ almost surely.}$$

**Remark.** If a statistic has ancillary information, then it is not complete. Therefore if a statistic is complete, it is not ancillary. Minimal sufficient statistics pretty much always exist, but a complete sufficient statistic might not exist ([see example from homework 4](#)).

**Remark.** A complete statistic may not be sufficient (example: a constant).

**Example 10.6 (Exercise 5.19 in Math 541A notes).** The following is an example of a statistic  $Y$  that is complete and nonconstant, but not sufficient. Suppose  $X_1, \dots, X_n$  is a random sample of known size  $n$  from a Bernoulli distribution with unknown probability parameter  $\theta \in (0, 1)$ . Let

$$Y = t(X_1, \dots, X_n) = \sum_{i=1}^{n-1} X_i.$$

Then  $Y$  is complete for  $\mu$  because for any  $f : \mathbb{R}^m \rightarrow \mathbb{R}$ , suppose

$$0 = \mathbb{E}_\theta f(Y) = \sum_{j=0}^{n-1} f(j) \Pr(Y = j \mid \theta) = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \theta^j (1-\theta)^{n-1-j}, \quad \forall \theta \in (0, 1).$$

Divide by  $(1 - \theta)^{n-1}$  and let  $\alpha = \theta/(1 - \theta)$  for notational ease. (Note that since  $\theta \in (0, 1)$ ,  $\alpha > 0$ .)

$$0 = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \theta^j \frac{(1-\theta)^{n-1-j}}{(1-\theta)^{n-1}} = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \left(\frac{\theta}{1-\theta}\right)^j = \sum_{j=0}^{n-1} f(j) \binom{n-1}{j} \alpha^j, \quad \forall \alpha > 0.$$

The sum on the right side is a polynomial in  $\alpha > 0$ . That means the sum on the right can only equal 0 if every coefficient on the polynomial equals zero.  $\binom{n}{j}$  is of course nonzero for all  $j \in 0, \dots, n-1$ . Therefore for all  $\alpha > 0$  we have that  $\mathbb{E}_\theta f(Y) = 0$  only if  $f(Y) = 0$ , so  $Y$  is complete.

However,  $Y$  is not sufficient for  $\mu$ . Using the definition of conditional probability,

$$\begin{aligned} \Pr((X_1, \dots, X_n) = (x_1, \dots, x_n) \mid Y = y) &= \frac{1}{\Pr(Y = y)} \cdot \Pr((X_1, \dots, X_n) = (x_1, \dots, x_n) \cap Y = y) \\ &= \frac{1}{\Pr(Y = y)} \cdot \Pr((X_1, \dots, X_n) = (x_1, \dots, x_n)) \end{aligned}$$

Using independence and the definition of a binomial distribution, we have

$$\begin{aligned} &= \frac{1}{\binom{n-1}{y} \theta^y (1-\theta)^{n-1-y}} \cdot \prod_{i=1}^n \Pr(X_i = x_i) = \frac{1}{\binom{n-1}{y} \theta^y (1-\theta)^{n-1-y}} \cdot \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \frac{1}{\binom{n-1}{y} \theta^y (1-\theta)^{n-1-y}} \cdot \theta^y (1-\theta)^{n-y} = \frac{1-\theta}{\binom{n-1}{y}}. \end{aligned}$$

Because  $\Pr((X_1, \dots, X_n) = (x_1, \dots, x_n) \mid Y = y) = (1-\theta)/\binom{n-1}{y}$  depends on  $\theta$ ,  $Y$  is not sufficient for  $\theta$ .

**Exercise 18 (Exercise 5.20 in Math 541A notes).** This exercise shows that a complete sufficient statistic might not exist.

Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from the uniform distribution on the three points  $\{\theta, \theta + 1, \theta + 2\}$ , where  $\theta \in \mathbb{Z}$ .

- (a) Show that the vector  $Y := (X_{(1)}, X_{(n)})$  is minimal sufficient for  $\theta$ .
- (b) Show that  $Y$  is not complete by considering  $X_{(n)} - X_{(1)}$ .
- (c) Using minimal sufficiency, conclude that any sufficient statistic for  $\theta$  is not complete.

*Proof.* (a) First we need to show that  $Y$  is sufficient for  $\theta$ . Informally, it makes sense that this would be the case because there are three possibilities:

- (1) If  $\theta$  and  $\theta + 2$  appear in the data set, we can identify  $\theta$  with certainty as the smallest of them. Simply observing  $x_{(1)}$  and  $x_{(n)}$  would show us the smallest and largest observations, which would be 2 units apart. Then we would know that the these observations are  $\theta$  and  $\theta + 2$ , and we could identify  $\theta$  with certainty. (Note that it doesn't matter in this case whether we observe  $\theta + 1$ .)

- (2) If only the values  $\{\theta, \theta + 1\}$  or  $\{\theta + 1, \theta + 2\}$  appear in the data set, we have to guess which one of these pairs we observed. If we guess that we have observed  $\theta + 1$  and  $\theta + 2$  and subtract one from the smallest observation to estimate  $\theta$ , or we can guess that we have observed  $\theta$  and  $\theta + 1$  and use the smallest value as our estimate for  $\theta$ . (Or we can hedge and take the mean of these values.) In any case, simply observing  $x_{(1)}$  and  $x_{(n)}$  would show us the smallest and largest observations, which would be 1 apart, leaving us in the same position as if we had all the data.
- (3) If only one value appears, we have to guess if it is  $\theta, \theta + 1$ , or  $\theta + 2$  in a way similar to if we only observe two values. But if we observe that the smallest and largest values are equal, we are observing the same information and we are in the same position for estimating  $\theta$  as if we had all of the data.

We can formally show that  $Y$  is sufficient using the Factorization Theorem (Theorem 10.5). Note that because on each trial we observe  $\theta, \theta + 1$ , or  $\theta + 2$  with equal probability, the mass function for the unordered observations  $f_{u,\theta} : \mathbb{Z}^n \rightarrow \mathbb{R}$  is the same as that of a multinomial distribution with three outcomes with equal probabilities. That is, if  $n_0 = \sum_{i=1}^n \mathbf{1}_{\{x_i=\theta\}}$  (where  $\mathbf{1}_{\{x_i=\theta\}}$  is an indicator variable for the  $i$ th observation having value  $\theta$ ),  $n_1 = \sum_{i=1}^n \mathbf{1}_{\{x_i=\theta+1\}}$ , and  $n_2 = \sum_{i=1}^n \mathbf{1}_{\{x_i=\theta+2\}}$ , we have

$$f_{u,\theta}(x) = \binom{n}{n_0, n_1, n_2} \left(\frac{1}{3}\right)^n = \frac{n!}{n_0! n_1! n_2!} \left(\frac{1}{3}\right)^n.$$

But taking into account the order in which we observe the samples, the probability of observing any one sample of size  $n$  ( $x = (x_1, \dots, x_n)$ ) is simply

$$f_\theta(x) = \begin{cases} \left(\frac{1}{3}\right)^n & x_1 \in \{\theta, \theta + 1, \theta + 2\}, \dots, x_n \in \{\theta, \theta + 1, \theta + 2\} \\ 0 & \text{otherwise.} \end{cases}$$

We have  $t : \mathbb{Z}^n \rightarrow \mathbb{Z}^2$  is given by

$$t(X_1, \dots, X_n) = (X_{(1)}, X_{(n)}).$$

Choose

$$h(x) = (1/3)^n, \quad g_\theta(t(x)) = \begin{cases} 1 & t(x) \in \{\theta, \theta + 1, \theta + 2\} \times \{\theta, \theta + 1, \theta + 2\} \\ 0 & \text{otherwise.} \end{cases} \quad (10.11)$$

Then we have  $f_\theta(x) = g_\theta(t(x))h(x)$ , as desired. Now we will use Theorem 10.3.6 (Theorem 5.8 from the lecture notes) to show that  $Y$  is not only sufficient, but is also minimal sufficient. Let  $x, z \in \mathbb{Z}^n$ , and let  $y_x = (x_{(1)}, x_{(n)}), y_z = (z_{(1)}, z_{(n)})$ . We seek to show that for every  $x, z \in \mathbb{Z}^n$ , the likelihood ratio  $\frac{f_\theta(x)}{f_\theta(z)}$  is constant (the constant may depend on  $x$  and  $z$ ) if and only if  $y_x = y_z$ . (We only need to consider  $x, z \in \mathbb{Z}^n$  rather than all of  $\mathbb{R}^n$  because since  $\theta \in \mathbb{Z}, X_i \in \mathbb{Z} \forall i \in \{1, \dots, n\}, \forall \theta \in \Theta$ .) Using the expressions in (10.11), we can write the equation in (10.12) as

$$f_\theta(x) = c(x, y)f_\theta(z) \iff g_\theta(t(x)) \left(\frac{1}{3}\right)^n = c(x, z)g_\theta(t(z)) \left(\frac{1}{3}\right)^n \iff g_\theta(t(x)) = c(x, z)g_\theta(t(z)) \quad (10.12)$$

I will argue that the equality in (10.12) only holds for some  $c(x, z) \in \mathbb{R}$  if  $t(x) = t(z)$ . Suppose we have observed data  $x$  and  $z$  from a distribution with a specific  $\theta_0$ . There are three cases to consider:

- (1)  $t(x) = \{\theta_0, \theta_0 + 2\}$  (**full information**): Then the only  $\theta$  for which  $g_\theta(t(x)) \neq 0$  is  $\theta = \theta_0$ . (This corresponds to situation (1) above.)
- (2)  $t(x) = \{\theta_0, \theta_0 + 1\}$  **or**  $\{\theta_0, \theta_0 + 1\}$ : Then  $g_\theta(t(x)) \neq 0$  for two values of  $\theta$ . In the first case, those two values will be  $\theta_0 - 1$  and  $\theta_0$ . In the second case, those two values will be  $\theta_0$  and  $\theta_0 + 1$ . (This corresponds to situation (2) above.)
- (3)  $t(x) = \{\theta_0, \theta_0\}, \{\theta_0 + 1, \theta_0 + 1\}$ , **or**  $\{\theta_0 + 2, \theta_0 + 2\}$ : Then  $g_\theta(t(x)) \neq 0$  for three values of  $\theta$ . In the first case, those three values will be  $\theta_0 - 2, \theta_0 - 1$ , and  $\theta_0$ . In the second case, those three values will be  $\theta_0 - 1, \theta_0$ , and  $\theta_0 + 1$ . In the last case, those three values will be  $\theta_0, \theta_0 + 1$ , and  $\theta_0 + 2$ . (This corresponds to situation (3) above.)

I have enumerated all possible values of  $t(x)$  or  $t(z)$  for a given true  $\theta = \theta_0$ , and note that there is no overlap among any of the possibilities for what values of  $\theta$  will yield identical values of  $g_\theta(t(x))$  and  $g_\theta(t(z))$  for all  $\theta$ . That is, that is true (and (10.12) holds for all  $\theta \in \Theta = \mathbb{Z}$ ) if and only if  $t(x) = t(z)$ , which is what we were trying to show. So  $Y$  is minimal sufficient.

- (b) Recall the definition of a complete statistic:

**Definition 10.14 (Complete statistic; definition 5.16 in 541A notes).** Suppose  $X_1, \dots, X_n$  is a random sample of size  $n$  from a distribution  $f$  where  $f \in \{f_\theta : \theta \in \Theta\}$  is a family of distributions. A statistic  $Y = t(X_1, \dots, X_n), t : \mathbb{R}^n \rightarrow \mathbb{R}^k$  is **complete** for  $\theta$  if the following holds:

For any  $f : \mathbb{R}^k \rightarrow \mathbb{R}$  such that  $\mathbb{E}_\theta f(Y) = 0 \forall \theta \in \Theta$ , it holds that  $f(Y) = 0$ .

We will show that  $Y$  is not complete by showing that  $Y$  contains ancillary information. Specifically, we will show that  $\mathbb{E}_\theta[f(Y)] \neq 0$  where  $f(Y) = X_{(n)} - X_{(1)} - c$  for some  $c \in \mathbb{Z}$ .

Let  $Z_1, \dots, Z_n$  be a random sample of size  $n$  from the uniform distribution on the three points  $\{0, 1, 2\}$ . Then  $X_i = Z_i + \theta, \forall 1 \leq i \leq n, \forall \theta \in \mathbb{Z}$ . Also,  $X_{(i)} = Z_{(i)} + \theta$ . So  $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$  does not depend on  $\theta \in \mathbb{Z}$ . So,  $X_{(n)} - X_{(1)}$  is ancillary for  $\theta$ . That is, there exists a constant  $c \in \mathbb{Z}$  that does not depend on  $\theta$  such that  $\mathbb{E}_\theta[X_{(n)} - X_{(1)} - c] = 0$  for all  $\theta \in \mathbb{Z}$ .

Define this  $c$  and let  $f : \mathbb{Z}^2 \rightarrow \mathbb{Z}$  be  $f(x_1, x_n) := x_n - x_1 - c \forall (x_1, x_n) \in \mathbb{R}^n$ . Then  $\mathbb{E}_\theta f(Y) = 0, \forall \theta \in \Theta, Y = (X_{(1)}, X_{(n)})$ .

- (c) Let  $S$  be any sufficient statistic for  $\theta$ . Since  $Y$  is minimal sufficient, there exists a function  $\phi$  such that  $Y = \phi(\theta)$ . Therefore  $S$  is not complete because  $\mathbb{E}_\theta(f(\phi(S))) = \mathbb{E}_\theta(f(Y)) = 0$  for all  $\theta \in \mathbb{Z}$ . So any sufficient statistic for  $\theta$  is not complete.

□

**Example 10.7 (Discrete RV example, Example 5.21 in Math 541A notes; return to Example 10.3.1).** Suppose we take a sample of size  $n$  from a Bernoulli distribution with parameter  $0 < \theta < 1$ . We already showed  $Y := X_1 + \dots + X_n$  is sufficient for  $\theta$ . Now we show  $Y$  is complete.

*Proof.* Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  with  $\mathbb{E}_\theta f(Y) = 0 \forall \theta \in \Theta$ . Since  $Y$  is binomial,

$$0 = \mathbb{E}_\theta f(Y) = \sum_{j=0}^n f(j) \binom{n}{j} \theta^j (1-\theta)^{n-j}, \quad \forall \theta \in (0, 1)$$

Let  $\alpha = \theta/(1 - \theta)$  and divide by  $(1 - \theta)^n$ ;

$$0 = \sum_{j=0}^n f(j) \binom{n}{j} \alpha^j, \quad \forall \alpha > 0.$$

The sum on the right side is a polynomial in  $\alpha > 0$ . That means the sum on the right can only equal 0 if every coefficient on the polynomial equals zero.  $\binom{n}{j}$  is of course nonzero for all  $j \in 0, \dots, n-1$ . Therefore for all  $\alpha > 0$  we have that  $\mathbb{E}_\theta f(Y) = 0$  only if we have  $f(j) = 0$  for all  $0 \leq j \leq n$ , so  $f(Y) = 0$  so  $Y$  is complete.

□

**Example 10.8 (Continuous RV example; return to Example 10.3).** For a random sample from a Gaussian distribution with known variance  $\sigma^2 > 0$  and unknown  $\mu \in \mathbb{R}$ , we showed that  $Y = (X_1 + \dots + X_n)/n$  is sufficient for  $\mu$ . Now we will show it is complete.

*Proof. Found this proof confusing* For simplicity, let  $\sigma = 1, n = 1$ . Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  with  $\mathbb{E}_\mu |f(Y)| < \infty$  for all  $\mu \in \mathbb{R}$ . Then

$$0 = \mathbb{E}_\mu(f(Y)) = \int_{-\infty}^{\infty} f(y) \exp\left(-\frac{(y-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} dy, \quad \forall \mu \in \mathbb{R}$$

Multiplying both sides by  $e^{\mu^2/2} \sqrt{2\pi}$  yields

$$0 = \int_{-\infty}^{\infty} f(y) e^{-y^2/2} e^{y\mu} dy, \quad \forall \mu \in \mathbb{R} \tag{10.13}$$

If  $f(y) \geq 0$ , we are done since (10.13) is the moment generating function of a random variable with density

$$\frac{f(y) e^{-y^2/2}}{\int_{\mathbb{R}} f(x) e^{-x^2/2} dx}$$

Then by Theorem 9.2 from the appendix of the Math 541A notes (uniqueness of moment-generating functions), this describes a unique random variable. But this is a contradiction because we can't divide by 0. (???)

In the case that  $f$  is positive and negative at different points, we write  $f = f_+ - f_-$  where  $f_+(x) := \max\{f(x), 0\}$  and  $f_-(x) := \max\{-f(x), 0\}$ . use (10.13) for any  $\mu$  and divide by case  $\mu = 0$ ,

$$\int_{-\infty}^{\infty} f_-(y) e^{-y^2/2} e^{y\mu} dy, \quad \int_{-\infty}^{\infty} f_-(y) e^{-y^2/2} dy$$

which yields

$$\int_{-\infty}^{\infty} f_+(y) e^{-y^2/2} e^{y\mu} dy \Big/ \int_{-\infty}^{\infty} f_-(y) e^{-y^2/2} dy$$

so we are done again by Theorem 9.2. So  $f_y = f_-$ ,  $f = f_+ - f_- = 0$ .

So basically we started by assuming that the expression equals zero and concluded that  $f$  must equal 0; therefore the statistic is complete.

□

**Remark.** Later we will show that complete and sufficient statistics are minimal sufficient.

**Exercise 19 (Conditional expectation exercise relevant to proof of Bahadur's Theorem).** Let  $X, Y : \Omega \rightarrow \mathbb{R}$  both be discrete or both continuous. For all  $y$  in the range of  $Y$ , define  $g(y) := \mathbb{E}(X | Y = y)$ . Define the conditional expectation of  $X$  given  $Y$ , denoted  $\mathbb{E}(X | Y)$ , as the random variable  $g(Y)$ .

Solution: Theorem 6.1.6

**Theorem 10.3.9 (Bahadur's Theorem; Theorem 5.25 in Math 541A notes).** If  $Y$  is a complete sufficient statistic for a family  $\{f_\theta : \theta \in \Theta\}$  of probability densities or probability mass functions, then  $Y$  is a minimal sufficient statistic for  $\theta$ .

**Remark.** By Remark 5.11 in Math 541A notes, a complete sufficient statistic is unique up to an invertible map. Also by Example 5.15 in Math 541A notes, the converse of Bahadur's Theorem is false.

*Proof.* By Proposition 10.3.5 (Proposition 5.12 in Math 541A notes), there exists a minimal sufficient statistic  $Z$  for  $\theta$ . To show that  $Y$  is minimal sufficient, it suffices to find a function  $r$  such that  $Y = r(Z)$ . Define  $r(Z) = \mathbb{E}_\theta(Y | Z)$ . Since  $Z$  is minimal sufficient and  $Y$  is sufficient by assumption, there exists a function  $u$  such that  $Z = u(Y)$ . By conditioning on  $Y$  we have by Exercise 19 (Exercise 5.24 in the Math 541A notes)

$$\begin{aligned} \mathbb{E}_\theta(r(u(Y))) &= \mathbb{E}_\theta(r(Z)) = \mathbb{E}_\theta[\mathbb{E}_\theta(r(Z) | Y)] = \mathbb{E}_\theta[\mathbb{E}_\theta(\mathbb{E}_\theta(Y | Z) | Y)] = \mathbb{E}_\theta[\mathbb{E}_\theta(\mathbb{E}_\theta(Y | u(Y)) | Y)] \\ &= \mathbb{E}_\theta[\mathbb{E}_\theta(Y | u(Y))] = \mathbb{E}_\theta(Y). \end{aligned}$$

That is,  $\mathbb{E}_\theta(r(u(Y)) - Y) = 0$  for all  $\theta \in \Theta$ . Since  $Y$  is complete, we conclude that  $r(u(Y)) = Y$ , and since  $r(u(Y)) = r(Z)$ , we have  $r(Z) = Y$ , as desired.

□

Basu's theorem tells us that a complete sufficient statistic implies independence from any ancillary statistic. So complete sufficient statistics have no ancillary information, unlike minimal sufficient statistics.

**Theorem 10.3.10 (Basu's Theorem, Theorem 5.27 in Math 541A notes).** Let  $Y : \Omega \rightarrow \mathbb{R}^k$  and  $Z : \Omega \rightarrow \mathbb{R}^m$  be statistics. If  $Y$  is a complete sufficient statistic for  $\{f_\theta : \theta \in \Theta\}$  and  $Z$  is ancillary for  $\theta$ , then for all  $\theta \in \Theta$ ,  $Y$  and  $Z$  are independent with respect to  $f_\theta$ .

*Proof.* Let  $A \subseteq \mathbb{R}^k$  and  $B \subseteq \mathbb{R}^m$ . We need to show that

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{P}_\theta(Y \in A)\mathbb{P}_\theta(Z \in B), \quad \forall \theta \in \Theta.$$

Note that

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{E}_\theta \mathbf{1}_{\{Y \in A\}} \mathbf{1}_{\{Z \in B\}} = \mathbb{E}_\theta \mathbb{E}_\theta(\mathbf{1}_{\{Y \in A\}} \mathbf{1}_{\{Z \in B\}} \mid Y) = \mathbb{E}_\theta[\mathbf{1}_{\{Y \in A\}} \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)].$$

Let  $g(Y) := \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)$ . Then

$$\mathbb{E}_\theta(g(Y)) = \mathbb{E}_\theta(\mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)) = \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}}) = \mathbb{P}_\theta(Z \in B). \quad (10.14)$$

Let  $c := \mathbb{P}_\theta(Z \in B) = \mathbb{E}_\theta(g(Y)) = \mathbb{E}_\theta[\mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)]$ . Then  $c$  does not depend on  $\theta$  since  $Z$  is ancillary by assumption. Then  $\mathbb{E}_\theta(g(Y) - c) = 0, \forall \theta \in \Theta$  for all  $\theta \in \Theta$ . Note that  $g(Y) := \mathbb{E}_\theta(\mathbf{1}_{\{Z \in B\}} \mid Y)$  does not depend on  $\theta$  since  $Y$  is sufficient. Since  $Y$  is complete,  $g(Y) - c = 0 \iff g(Y) = c$ , so  $Y$  is constant. Therefore by (10.14)

$$c = \mathbb{E}_\theta(c) = \mathbb{E}_\theta(g(Y)) = \mathbb{P}_\theta(Z \in B),$$

so we have

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{P}_\theta(Y \in A)g(Y) = \mathbb{P}_\theta(Y \in A)c = \mathbb{P}_\theta(Y \in A)\mathbb{P}_\theta(Z \in B), \quad \forall \theta \in \Theta.$$

as desired. □

**Example 10.9 (Example from 541B).**  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ , i.i.d. Then  $\bar{X}_n$  is complete sufficient for  $\mu$ . Therefore by Basu's Theorem (Theorem 10.3.10) and Example 10.4,  $\bar{X}_n$  and  $X_1 - \bar{X}_n$  are independent.

**Example 10.10.** By Basu's Theorem (Theorem 10.3.10), the sample mean and sample variance of a random sample of a Gaussian random variable are independent.

## 10.4 Point Estimation

**Definition 10.15 (Point estimator).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a family of distribution  $\{f_\theta : \theta \in \Theta\}$ . If  $Y$  is a statistic that is used to estimate the parameter  $\theta$  that fits the data at hand, we then refer to  $Y$  as a **point estimator** or **estimator**.

### 10.4.1 Heuristic Principles for Finding Good Estimators

**Definition 10.16 (Likelihood, Definition 6.1 in Math 541A notes).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a family of distributions  $\{f_\theta : \theta \in \Theta\}$ . If we have data  $x \in \mathbb{R}^n$ , then the function  $L : \Theta \rightarrow [0, \infty)$  defined by  $L(\theta) := f_\theta(x)$  is called the **likelihood function**.

- **Likelihood principle:** All data relevant to estimating the parameter  $\theta$  is contained in the likelihood function.
- **Sufficiency principle:** If  $Y = t(X_1, \dots, X_n)$  is a sufficient statistic and if we have two results  $x, y \in \mathbb{R}^n$  from an experiment with the same statistics  $t(x) = t(y)$ , then our estimate of the parameter  $\theta$  should be the same for either experimental result.
- **Equivariance principle:** If the family of distributions  $\{f_\theta : \theta \in \Theta\}$  is invariant under some symmetry, then the estimator of  $\theta$  should respect the same symmetry. (For example, a location family is invariant under translation, so an estimator for the location parameter should commute with translations.)

#### 10.4.2 Evaluating Estimators

We can enumerate several desirable properties for estimators.

**Definition 10.17 (Unbiasedness, Definition 6.2 in Math 541A notes).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a family of distributions  $\{f_\theta : \theta \in \Theta\}$ . Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}^k$  and let  $Y := t(X_1, \dots, X_n)$  be an estimator for  $g(\theta)$ . Let  $g : \Theta \rightarrow \mathbb{R}^k$ . We say that  $Y$  is **unbiased** for  $g(\theta)$  if  $\mathbb{E}_\theta Y = g(\theta)$  for all  $\theta \in \Theta$ .

One common way to check the quality of an estimator is the mean squared error, or squared  $L_2$  norm, of the estimator minus  $\theta$ ,  $\mathbb{E}_\theta(Y - g(\theta))^2$ . If the estimator is unbiased, this quantity is equal to the variance of  $Y$ .

**Definition 10.18 (UMVU, sometimes called MVUE (minimum variance unbiased estimator); Definition 6.3 in Math 541A Notes).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a family of distributions  $\{f_\theta : \theta \in \Theta\}$ . Let  $g : \Theta \rightarrow \mathbb{R}$ . Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $Y := t(X_1, \dots, X_n)$  be an unbiased estimator for  $g(\theta)$ . We say that  $Y$  is **uniformly minimum variance unbiased (UMVU)** if, for any other unbiased estimator  $Z$  for  $g(\theta)$ , we have  $\text{Var}_\theta(Y) \leq \text{Var}_\theta(Z)$  for all  $\theta \in \Theta$ .

**Remark.** The “uniform” property has to do with the fact that this inequality must hold for every  $\theta \in \Theta$  (as opposed to for a particular  $\theta$ , or averaged over all  $\theta \in \Theta$ , or something like that).

**Example 10.11 (Math 541B example).** Suppose  $X \sim \text{Bin}(3, \theta)$ ,  $\theta \in [0, 1]$ . Does there exist an unbiased estimator of  $\theta^5$ ? No. Assume that  $T(X)$  is unbiased for  $\theta^5$ , so that

$$\mathbb{E}_\theta T(x) = \sum_{j=0}^3 T(j) \binom{3}{j} \theta^j (1-\theta)^{3-j} = \theta^5 \quad \forall \theta \in [0, 1].$$

But this cannot be, because the quantity on the left side of the last equality is a third degree polynomial in  $\theta$  and the quantity on the right side is a 5th degree polynomial. So the only powers of  $\theta$  that are possible to estimate unbiasedly in this case are powers less than or equal to 3.

**Example 10.12 (Math 541B example).**  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ ,  $\mu \in \mathbb{R}$ . Let  $g(\mu) = \mathbb{P}_\mu(X_1 \leq x)$ . Find a UMVUE of  $g(\mu)$ .

**Solution.**

Consider  $S(X_1, \dots, X_n) = I\{X_1 \leq x\}$  (an indicator function). Then  $\mathbb{E}_\mu S(X_1, \dots, X_n) = g(\mu)$ . Define

$$\hat{S}(X_1, \dots, X_n) := \mathbb{E} [I\{X_1 \leq x\} \mid \bar{X}_n].$$

Then by Lehmann-Scheffe (Theorem 10.4.3),  $\hat{S}$  is UMVU. Note that

$$\hat{S}(X_1, \dots, X_n) = \mathbb{E} [I\{X_1 - \bar{X}_n \leq x - \bar{X}_n\} \mid \bar{X}_n]$$

By Example 10.4,  $X_1 - \bar{X}_n$  is independent of  $\bar{X}_n$ , so we can write

$$= \mathbb{E} [I\{X_1 - \bar{X}_n \leq x - \bar{X}_n\}] = \mathbb{P} [X_1 - \bar{X}_n \leq x - \bar{X}_n] = \phi \left( \sqrt{\frac{n}{n+1}}(x - \bar{X}_n) \right)$$

where  $\phi(\cdot)$  is the cdf of a standard Gaussian random variable.

More generally, given a family of distributions  $\{f_{\tilde{\theta}} : \tilde{\theta} \in \Theta\}$ , we could be given a **loss function**  $L(\theta, y) : \Theta \times \mathbb{R}^k \rightarrow \mathbb{R}$  and be asked to minimize the **risk function**  $r(\theta, Y) := \mathbb{E}_{\tilde{\theta}}(\ell(\theta, Y))$  over all possible estimators  $Y$ . In the case of mean squared error loss, we have  $L(\theta, y) := (y - g(\theta))^2$  for all  $y, \theta \in \mathbb{R}$ .

The Rao-Blackwell Theorem says that if  $L(\theta, y)$  is convex in  $y$  then we can create an optimal estimator for  $g(\theta)$  from a sufficient statistic and any estimator for  $g(\theta)$  (we can lower the risk of an estimator  $Y$  by conditioning on a sufficient statistic  $Z$ ).

**Theorem 10.4.1 (Rao-Blackwell; Theorem 6.4 in Math 541A notes).** Let  $Z$  be a sufficient statistic for  $\{f_\theta : \theta \in \Theta\}$  and let  $Y$  be an estimator for  $g(\theta)$ . Define  $W := \mathbb{E}_\theta(Y \mid Z)$ . Let  $\theta \in \Theta$ . Then

$$\text{Var}_\theta(W) \leq \text{Var}_\theta(Y).$$

Further, let  $r(\theta, y) < \infty$  and such that  $\ell(\theta, y)$  is convex in  $y$ . Then

$$r(\theta, W) \leq r(\theta, Y).$$

*Proof.* Note that since  $Z$  is sufficient,  $W$  does not depend on  $\theta$ . By the Conditional Jensen's Inequality (Theorem 9.1.5) and using the convexity of  $\ell(\theta, y)$  in  $y$ ,

$$\ell(\theta, w) = \ell(\theta, \mathbb{E}_{\tilde{\theta}}(Y \mid Z)) \leq \mathbb{E}_{\tilde{\theta}}[\ell(\theta, Y) \mid Z].$$

Take expectations of both sides to get

$$\mathbb{E}_{\tilde{\theta}}\ell(\theta, w) = r(\theta, W) \leq \mathbb{E}_{\tilde{\theta}}\mathbb{E}_{\tilde{\theta}}[\ell(\theta, Y) \mid Z] = \mathbb{E}_{\tilde{\theta}}\ell(\theta, Y) = r(\theta, Y).$$

□

**Definition 10.19 (Definition 6.4 in 541A notes; MISSED SOME NOTES TODAY).** We say  $Y$  is **uniformly minimum risk unbiased** (UMRU) if for any other unbiased estimator  $Z$  for  $g(\theta)$ ,

$$r(\theta, Y) \leq r(\theta, z), \quad \forall \theta \in \Theta$$

**Remark.** Unfortunately, UMRU or UMVU may not exist. More fundamentally, an unbiased estimator for  $g(\theta)$  may not exist. For example, let  $X$  be a binomial random variable with known  $n$ , unknown  $0 < \theta < 1$ , and  $g(\theta) = \theta/(1 - \theta)$ . Then no unbiased estimator exists for  $g(\theta)$ . Why?

$$\mathbb{E}_\theta t(X) = \sum_{j=0}^n t(j) \binom{n}{j} \theta^j (1 - \theta)^{n-j}, \quad \forall \theta \in \Theta, \text{ by definition of } X.$$

where the summation is a polynomial of degree at most  $n$  in  $\theta$ . Then it is impossible to have  $\mathbb{E}_\theta t(x) = g(\theta)$  when  $g(\theta) = \theta/(1 - \theta)$ .

Recall the definition of strict convexity (Definition 9.2).

**Theorem 10.4.2 (Rao-Blackwell restated; Theorem 6.7 in Math 541A notes).** Let  $Z$  be a sufficient statistic for  $\{f_\theta : \theta \in \Theta\}$  and let  $Y$  be an estimator for  $g(\theta)$ . Define  $W := \mathbb{E}_\theta(Y | Z)$ . Let  $\theta \in \Theta$  with  $r(\theta, y) < \infty$  and such that  $\ell(\theta, y)$  is convex in  $y$ . Then

$$r(\theta, W) \leq r(\theta, Y).$$

Further, if  $\ell(\theta, y)$  is strictly convex in  $y \in \mathbb{R}$ , then  $r(\theta, W) < r(\theta, Y)$  unless  $W = Y$  (that is, there is a unique minimizer of the risk).

So  $Z$  makes the estimator better. Question: can we construct  $\mathbb{E}_\theta(Y | Z)$  to be UMRU?

**Remark (Remark 6.9 in Math 541A notes).**  $\mathbb{E}_\theta W = \mathbb{E}_\theta \mathbb{E}_\theta(Y | Z) = \mathbb{E}_\theta Y$ . So if  $Y$  is unbiased for  $g(\theta)$ , then so is  $W$ .

**Remark.** What happens if  $Z$  is constant in Rao-Blackwell? Then in general  $Z$  will not be sufficient, so  $W$  might depend on  $\theta$  which is not allowed. Put another way, if  $Z$  has insufficient information, then  $W$  gets messed up (???).

**Example 10.13.** Let  $X_1, \dots, X_n$  be a random sample with unknown mean  $\mu \in \mathbb{R}$ . We want to construct an estimator for  $\mu$  using Rao-Blackwell. Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}$  so that  $t(x_1, \dots, x_n) = x_1$  for all  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ . Let  $Y = t(X_1, \dots, X_n) = X_1$ . Note that  $Y$  is unbiased. First use of Rao-Blackwell: use  $Z = (X_1, \dots, X_n)$ . Then by Exercise 5.24,

$$W := \mathbb{E}_\mu(X_1 | (X_1, \dots, X_n)) = \mathbb{E}(X_1 | X_1) = X_1.$$

We can think of this as failing to improve the estimator because we used “too much” information. Second try: use  $Z = \sum_{i=1}^n X_i$ . Note that in the Gaussian case  $Z$  is sufficient for  $\mu$  and unbiased for  $n\mu$ . Since  $X_1, \dots, X_n$  are i.i.d. for all  $1 \leq k \leq \ell \leq n$  the joint distribution of  $(X_k, \sum_{i=1}^n X_i)$  is the same as the joint distribution of  $(X_\ell, \sum_{i=1}^n X_i)$ . So

$$\mathbb{E}(X_k \mid \sum_{i=1}^n X_i) = \mathbb{E}(X_\ell \mid \sum_{i=1}^n X_i).$$

So we have

$$W := \mathbb{E}_\mu \left( X_1 \mid \sum_{i=1}^n X_i \right) = \frac{1}{n} \sum_{j=1}^n \mathbb{E}_\mu \left( X_j \mid \sum_{i=1}^n X_i \right) = \frac{1}{n} \mathbb{E}_\mu \left( \sum_{j=1}^n X_j \mid \sum_{i=1}^n X_i \right) = \frac{1}{n} \sum_{i=1}^n X_i.$$

So we started with a trivial estimator  $X_1$  and ended up with the sample mean using Rao-Blackwell.

**Theorem 10.4.3 (Lehmann-Scheffe, Theorem 6.13 in Math 541A notes).** Let  $Z$  be a complete sufficient statistic for a family of distributions  $\{f_\theta : \theta \in \Theta\}$ . Let  $Y$  be an unbiased estimator for  $g(\theta)$ . Define  $W := \mathbb{E}_\theta(Y \mid Z)$ . (Since  $Z$  is sufficient,  $W$  does not depend on  $\theta$ .) Then  $W$  is UMVRU for  $g(\theta)$ . Further, if  $\ell(\theta, y)$  is strictly convex in  $y$  for all  $\theta \in \Theta$ , then  $W$  is unique. In particular,  $W$  is the unique UMVU for  $g(\theta)$ .

*Proof.*  $W$  is unbiased by Remark 6.10 in math 541A notes (“ $\mathbb{E}_\theta W = \mathbb{E}_\theta \mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta Y$ . So if  $Y$  is unbiased for  $g(\theta)$ , then so is  $W$ .”) We first show  $W$  does not depend on  $Y$ . Let  $Y'$  be an unbiased estimator for  $g(\theta)$ . We show that  $\mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta(Y' \mid Z)$  for all  $\theta \in \Theta$ . Note that

$$\mathbb{E}_\theta(\mathbb{E}_\theta(Y \mid Z) - \mathbb{E}_\theta(Y' \mid Z)) = \mathbb{E}_\theta(Y - Y') = g(\theta) - g(\theta) = 0, \forall \theta \in \Theta$$

Note that  $\mathbb{E}_\theta(Y \mid Z)$  and  $\mathbb{E}_\theta(Y' \mid Z)$  are functions of  $Z$ . Therefore since  $Z$  is complete,  $\mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta(Y' \mid Z)$  for all  $\theta \in \Theta$ .

Next, by Rao Blackwell,

$$r(\theta, Y') = r(\theta, \mathbb{E}_\theta(Y' \mid Z)) = r(\theta, \mathbb{E}_\theta(Y \mid Z)) = r(\theta, W), \forall \theta \in \Theta.$$

□

**Remark (Remark 6.14 in Math 541A notes, Theorem 7.3.23 in Casella and Berger [2001, p. 347]).** Let  $Z : \Omega \rightarrow \mathbb{R}^k$  be a complete sufficient statistic for  $\{f_\theta : \theta \in \Theta\}$  and let  $h : \mathbb{R}^k \rightarrow \mathbb{R}^m$ . Let  $g(\theta) := \mathbb{E}_\theta h(Z)$  for all  $\theta \in \Theta$ . Then  $h(Z)$  is unbiased for  $g(\theta)$ , since  $\mathbb{E}_\theta h(Z) = g(\theta) = \mathbb{E}_\theta(g(\theta))$ . Applying Theorem 10.4.3, we have

$$W := \mathbb{E}_\theta(h(Z) \mid Z) = \mathbb{E}_\theta(\mathbb{E}_\theta[h(Z) \mid h(Z)] \mid Z) = \mathbb{E}_\theta[h(Z) \mid h(Z)] = h(Z).$$

Therefore by Theorem 10.4.3,  $h(Z)$  is UMVU for  $g(\theta)$ . That is, any function of a complete sufficient statistic is UMVU for its expected value. So one way to find a UMVU is to come up with a function of a complete sufficient statistic that is unbiased for a given function  $g(\theta)$ .

**Summary of methods for finding UMVU (given a complete sufficient statistic  $Z$ , want to estimate  $g(\theta)$ )**

- (1) (**Condition method/Rao-Blackwell**): Follow Theorem 10.4.3: find an unbiased  $Y$  and let  $W := \mathbb{E}_\theta(Y | Z)$ . (problem: can be hard to find an unbiased  $Y$ .)
- (2) Solve for  $h : \mathbb{R}^k \rightarrow \mathbb{R}$  satisfying

$$\mathbb{E}_\theta h(Z) = g(\theta) \quad (10.15)$$

by the above remark, (10.15) will also give you the UMVU. Then  $h(Z)$  is UMVU for  $g(\theta)$ . By “solve”, consider that we have  $g$  and  $Z$  and somehow solve for the  $h$  satisfying (10.15). For example if  $Z$  is binomial the left side of (10.15) will be the sum of a bunch of numbers. Find the  $h$  values that satisfy (10.15), if possible.

- (3) (**Luck method**): Somehow guess the  $h$  such that (10.15) is satisfied.

**Example 10.14.** Suppose we are sampling from a Gaussian distribution with unknown mean and variance. By the Factorization Theorem and Exercise 5.23 (**on homework 4**), we know  $(\bar{X}, S^2)$  is complete sufficient of  $(\mu, \sigma^2)$  (sufficiency follows by the Factorization Theorem (Theorem 10.5), completeness follows by the exercise). For example, using method (2) above,  $\bar{X}$  is UMVU for  $\mu$  (with finite  $\sigma$ ) by method (3) above, since  $\bar{X}$  is a function of  $(\bar{X}, S^2)$ ,  $h(x, y) := x$ ,  $g(\mu, \sigma^2) := \mu$  (then (10.15) is satisfied). Similarly,  $S^2$  is UMVU for  $\sigma^2$  by (3) using  $h(x, y) := y$ ,  $g(\mu, \sigma^2) := \sigma^2$  (then (10.15) is satisfied).

Suppose we want a UMVU for  $\mu^2$ . Try guessing  $(\bar{X})^2$  as an estimator. Note that

$$\mathbb{E}[(\bar{X})^2] = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2 = \frac{1}{n^2} \left( \sum_{i=1}^n \mathbb{E}X_i^2 + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}X_i \mathbb{E}X_j \right) = \dots = \mu^2 + \sigma^2/n$$

So,

$$\mathbb{E}\left(\bar{X}^2 - S^2/n\right) = \mu^2$$

which means that  $\bar{X}^2 - S^2/n$  is UMVU since it is a function of  $(\bar{X}, S^2)$ .

**Example 10.15.** Try Method (2). Let  $X$  be a binomial random variable with known parameter  $n$  and unknown  $0 < \theta < 1$ . Suppose we want to estimate  $g(\theta) := \theta(1 - \theta)$ . Solve (10.15): find the  $h$  satisfying

$$\theta(1 - \theta) = \mathbb{E}_\theta h(X)$$

$$\iff \theta(1 - \theta) = \sum_{j=0}^n h(j) \binom{n}{j} \theta^j (1 - \theta)^{n-j}, \quad \forall \theta \in (0, 1)$$

For convenience, let  $a := \theta/(1 - \theta)$ , so that  $a(1 - \theta) = \theta \iff \theta = a/(1 + a)$ ,  $1 - \theta = 1/(1 + a)$ . Then we have

$$(1 - \theta)^{-n} = \sum_{j=0}^n h(j) \binom{n}{j} a^j = \theta(1 - \theta)^{1-n} = \frac{a}{1+a} \left(\frac{1}{1+a}\right)^{1-n} = a(1+a)^{n-2} \quad (10.16)$$

So we want to solve for  $h(j)$  so that for all  $a > 0$ ,

$$\sum_{j=0}^n h(j) \binom{n}{j} a^j = a(1+a)^{n-2} = \text{(RHS of (10.16), by binomial theorem)} a \sum_{j=0}^{n-2} \binom{n-2}{j} a^j = \sum_{j=1}^{n-1} \binom{n-2}{j} - 1 a^j$$

We need the coefficients to match up one by one. So  $h(0) = h(n) = 0$ , and then it works if  $h(j) \binom{n}{j} = \binom{n-2}{j-1}$  for all  $j \in 1, \dots, n-1$ . So

$$h(j) = \frac{\binom{n-2}{j-1}}{\binom{n}{j}} = \frac{(n-2)!}{n!} \frac{(n-j)!j!}{| } (n-j-1)!(j-1)! = \frac{(n-j)j}{n(n-1)}$$

So in fact,

$$h(j) = \frac{(n-j)j}{n(n-1)}, \quad \forall 0 \leq j \leq n$$

Therefore the UMVU for  $\theta(1-\theta)$  is

$$\frac{X(n-X)}{n(n-1)}.$$

**Example 10.16.** Try Method (1). Suppose we have  $n$  independent samples  $X_1, \dots, X_n$  from a Bernoulli distribution with unknown  $\theta \in (0, 1)$ . From Example 6.9 (Example 3.15 in Math 541A notes) and Exercise 5.23 in Math 541A notes, a complete sufficient statistic is  $Z := \sum_{i=1}^n X_i$  is complete and sufficient for  $\theta$ . Also,  $(1/n) \sum_{i=1}^n X_i$  is unbiased for  $\theta$ . So,  $(1/n) \sum_{i=1}^n X_i$  is UMVU for  $\theta$ . Suppose we want to estimate  $\theta^2$ . We need an unbiased estimator  $Y$  for  $\theta^2$ . Let  $Y = X_1 X_2$ . Then  $\mathbb{E}(Y) = \mathbb{E}(X_1)\mathbb{E}(X_2) = \theta^2$ . By Theorem 10.4.3,  $W := \mathbb{E}_\theta(Y | Z)$  is UMVU for  $\sigma^2$ . Note,  $Y = 1$  when  $X_1 = X_2 = 1$  and 0 otherwise. So

$$\begin{aligned} \mathbb{E}_\theta(Y | Z = z) &= \mathbb{E}_\theta(\mathbf{1}_{\{X_1=X_2=1\}} | Z = z) = \mathbb{P}_\theta(X_1 = X_2 = 1 | Z = z) = \mathbb{P}_\theta(X_1 = X_2 = 1 | \sum_{i=1}^n X_i = z) \\ &= \frac{1}{\mathbb{P}_\theta\left(\sum_{i=1}^n X_i = z\right)} \cdot \mathbb{P}_\theta\left(X_1 = X_2 = 1 \cap \sum_{i=1}^n X_i = z\right) \\ &= \frac{1}{\mathbb{P}_\theta\left(\sum_{i=1}^n X_i = z\right)} \cdot \mathbb{P}_\theta\left(X_1 = X_2 = 1 \cap \sum_{i=3}^n X_i = z-2\right) \\ &= \frac{\theta^2 \binom{n-2}{z-2} \theta^{z-2} (1-\theta)^{n-z}}{\binom{n}{z} \theta^z (1-\theta)^{n-z}} = \frac{\binom{n-z}{z-2}}{\binom{n}{z}} = \frac{(n-z)!(n-z)!z!}{n!(n-z)!(z-2)!} = \frac{z(z-1)}{n(n-1)} \end{aligned}$$

So we have shown that for all  $0 \leq Z \leq n$ ,

$$\mathbb{E}_\theta(Y \mid Z = z) = \frac{z(z-1)}{n(n-1)}.$$

So, by Theorem 10.4.3,

$$W := \mathbb{E}_\theta(Y \mid Z) = \frac{Z(Z-1)}{n(n-1)}$$

is UMVU for  $\theta^2$ .

**Question:** If  $W_1$  is UMVU for  $g_1(\theta)$  and  $W_2$  is UMVU for  $g_2(\theta)$ , is  $W_1 + W_2$  UMVU for  $g_1(\theta) + g_2(\theta)$ ? If there is a complete sufficient statistic, then by Lehmann-Scheffe (Theorem 10.4.3),  $W_1 = \mathbb{E}_\theta(Y_1 \mid Z)$ ,  $W_2 = \mathbb{E}(Y_2 \mid Z)$  where  $Y_1$  is unbiased for  $g_1$ ,  $Y_2$  is unbiased for  $g_2$ . Then

$$W_1 + W_2 = \mathbb{E}_\theta(Y_1 + Y_2 \mid Z).$$

Note:  $\mathbb{E}_\theta(Y_1 + Y_2) = \mathbb{E}_\theta Y_1 + \mathbb{E}_\theta Y_2 = g_1(\theta) + g_2(\theta) = g_Z(2\theta)$ . So  $W_1 + W_2$  is UMVU by Lehmann-Scheffe (Theorem 10.4.3) for  $g_1(\theta) + g_2(\theta)$ . But is it true if we don't have a complete sufficient statistic (and this argument doesn't apply)? **yes, by the theorem below; this condition will clearly hold across sums.**

**Theorem 10.4.4 (Alternate Characterization of UMVU; Theorem 6.18 in Math 541A notes).** Let  $f \in \{f_\theta : \theta \in \Theta\}$  be a family of distributions and let  $g : \Theta \rightarrow \mathbb{R}$ . Let  $W$  be an unbiased estimator for  $g(\theta)$  (note that the existence of an unbiased estimator is a nontrivial assumption). Let  $L_2(\Omega)$  be the set of statistics with finite second moment. Then  $W \in L_2(\Omega)$  is UMVU for  $g(\theta)$  if and only if for any  $\theta \in \Theta$ ,

$$\mathbb{E}_\theta(WU) = 0, \quad \forall U \in L_2(\Omega) \text{ that are unbiased estimators of 0}$$

Thinking of this as an inner product, we have to be orthogonal to all such  $U$ .

*Proof.* Assume  $W$  is UMVU for  $g(\theta)$ . Let  $U$  be an unbiased estimator of 0. Let  $s \in \mathbb{R}$ , consider  $W + sU$ . Note that  $W + sU$  is also unbiased for  $g(\theta)$ . Since  $W$  is UMVU,

$$\begin{aligned} \text{Var}_\theta(W) &\leq \text{Var}_\theta(W + sU) = \text{Var}_\theta(W) + s^2 \text{Var}_\theta(U) + 2s \text{Cov}_\theta(W, U) \\ &= \text{Var}_\theta(W) + s^2 \text{Var}(U) + 2s \mathbb{E}_\theta[(W - \mathbb{E}_\theta(W))U], \quad \forall \theta \in \Theta. \end{aligned}$$

Note that we have equality when  $s = 0$ . Also, the derivative of the right side with respect to  $s$  must be 0 when  $s = 0$  or else the inequality does not hold (the minimum value occurs at  $s = 0$  if and only if the derivative of the right side in  $s$  is 0 at  $s = 0$ ). Note that the derivative of the right side is

$$0 = 2\mathbb{E}_\theta[(W - \mathbb{E}_\theta W)U] = 2\mathbb{E}_\theta(WU).$$

The converse is also true because this reasoning can be reversed, since if  $Y$  is any unbiased estimator for  $g(\theta)$ , then  $U := W - Y$  is an unbiased estimator for 0, and  $Y = W + sU$  with  $s = 1$ . We have

$$\text{Var}_\theta(Y) = \text{Var}_\theta(W - U) = \text{Var}_\theta W + \text{Var}_\theta U + 2\text{Cov}_\theta(W, U) = \text{Var}_\theta W + \text{Var}_\theta U + 2\mathbb{E}_\theta(WU)$$

So  $\text{Var}_\theta Y \geq \text{Var}_\theta W$  for all  $\theta \in \Theta$ .

□

**Remark.** If we have a complete sufficient statistic, better to use the earlier methods in general (unless it is really complicated to work with). If we don't have a complete sufficient statistic, use this theorem.

#### 10.4.3 Efficiency of an Estimator

Another desirable property of an estimator is high efficiency—"good" with a small number of samples. One way to quantify this notion is to define a notion of "information" and try to maximize the information content of the estimator.

**Definition 10.20 (Fisher Information, Definition 6.19 in Math 541A notes).** Let  $f \in \{f_\theta : \theta \in \Theta\}$  be a family of multivariate probability densities or probability mass functions. Assume  $\Theta \subseteq \mathbb{R}$  (this is a one-parameter situation). Let  $X$  be a random variable with distribution  $f_\theta$ . Define the **Fisher information** of the family to be

$$I(\theta) = I_X(\theta) := \mathbb{E}_\theta \left( \frac{d}{d\theta} \log f_\theta(X) \right)^2, \quad \forall \theta \in \Theta$$

if this quantity exists and is finite.

(See also Section 11.21 for connections to generalized linear models.)

**Remark.** Note that if  $X$  is continuous,

$$\mathbb{E}_\theta \left( \frac{d}{d\theta} \log f_\theta(X) \right) = \int_{\mathbb{R}^n} \frac{1}{f_\theta(x)} \frac{d}{d\theta} f_\theta(x) \cdot f_\theta(x) dx = \int_{\mathbb{R}^n} \frac{d}{d\theta} f_\theta(x) dx = \frac{d}{d\theta} \int_{\mathbb{R}^n} f_\theta(x) dx = \frac{d}{d\theta} 1 = 0.$$

So we could have equivalently defined the Fisher information as

$$I_X(\theta) = \text{Var}_\theta \left( \frac{d}{d\theta} \log f_\theta(X) \right)$$

**Example 10.17 (Example 6.20 in Math 541A notes).** Let  $\theta > 0$ , let

$$f_\theta(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp \left( -\frac{(x-\theta)^2}{2\sigma^2} \right), \quad \forall \theta \in \mathbb{R} = \Theta, \quad \forall x \in \mathbb{R}.$$

Then

$$I(\theta) = \text{Var}_\theta \left( \frac{d}{d\theta} - \frac{(x-\theta)^2}{2\sigma^2} \right) = \frac{1}{\sigma^4} \text{Var}_\theta(x-\theta) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Observe that a small  $\sigma$  means large  $I(\theta)$  in this case.

**Proposition 10.4.5 (Proposition 6.21 in Math 541A notes).** Let  $X$  be a random variable with distribution from  $\{f_\theta : \theta \in \Theta\}$  (densities or mass functions). Let  $Y$  be a random variable with distribution from  $\{g_\theta : \theta \in \Theta\}$  (densities or mass functions). Assume  $\Theta \subseteq \mathbb{R}$  (one parameter in  $\theta$ ). If  $X$  and  $Y$  are independent, then

$$I_{(X,Y)}(\theta) = I_X(\theta)I_Y(\theta).$$

*Proof.* This proof will just be for the case of densities (continuous random variables; the case for probability mass functions is similar). Since  $X$  and  $Y$  are independent,  $(X, Y)$  has a distribution with density  $f_\theta(x)g_\theta(y)$ , for all  $x, y \in \mathbb{R}$ . Also,  $\frac{d}{d\theta} \log f_\theta(X)$  and  $\frac{d}{d\theta} \log g_\theta(Y)$  are independent for all  $\theta \in \Theta$ . So,

$$I_{(X,Y)}(\theta) = \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_\theta(X)g_\theta(Y)] \right) = \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_\theta(X)] + \frac{d}{d\theta} \log[g_\theta(Y)] \right)$$

By independence we can write

$$= \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_\theta(X)] \right) + \text{Var}_\theta \left( \frac{d}{d\theta} \log[g_\theta(Y)] \right) = I_X(\theta) + I_Y(\theta).$$

□

**Remark.** This is consistent with a notion of “information” since if variables are independent, the information is the sum of the information of each variable. This proof also shows the main reason why the logarithm is in the definition of the Fisher information—it brings a product to a sum.

**Proposition 10.4.6 (Exercise 6.22 in Math 541A notes).** Let  $X$  be a random variable with distribution from  $\{f_\theta : \theta \in \Theta\}$  (densities or mass functions). Let  $Y$  be a random variable with distribution from  $\{g_\theta : \theta \in \Theta\}$  (densities or mass functions). Then

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_{Y|X=x}(\theta), \quad \forall \theta \in \Theta, x \in \mathbb{R}.$$

*Proof.* Recall that  $Y | X$  has density  $f_{X,Y}(x, y)/f_X(x)$  for any fixed  $x$ . And if  $X, Y$  are discrete random variables, recall that  $Y | X$  has mass function  $\mathbb{P}(X = x, Y = y)/\mathbb{P}(Y = y)$ .

$$\begin{aligned} I_{(X,Y)}(\theta) &= \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_{\theta(X,Y)}(x, y)] \right) = \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_\theta(x)f_{\theta(Y|X=x)}(y)] \right) \\ &= \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_\theta(X)] + \frac{d}{d\theta} \log[f_{\theta(Y|X=x)}(y)] \right) \end{aligned}$$

Note that  $Y | X = x$  is independent of  $X$  (because we have conditioned out the dependence). Therefore we have

$$= \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_\theta(X)] \right) + \text{Var}_\theta \left( \frac{d}{d\theta} \log[f_{\theta(Y|X=x)}(y)] \right) = I_X(\theta) + I_{Y|X=x}(\theta).$$

□

**Theorem 10.4.7 (Cramer-Rao/Information Inequality, Theorem 6.23 in Math 541A Notes).**

Let  $X : \Omega \rightarrow \mathbb{R}^n$  be a random variable with distribution from a family of multivariable probability densities or probability mass functions  $\{f_\theta : \theta \in \Theta\}$  with  $\Theta \subseteq \mathbb{R}$  and  $\Theta$  an open interval. Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $Y := t(X)$  be a statistic. For any  $\theta \in \Theta$  let  $g(\theta) := \mathbb{E}_\theta Y$ . Assume  $g'(\theta)$  and  $\frac{\partial p_\theta(x)}{\partial \theta}$  exist for all  $x$ . Assume the set  $A = \{x : p_\theta(x) = 0\}$  does not depend on  $\theta$ , and assume the Fisher information  $I_X(\theta)$  is finite and nonzero. Lastly, assume

$$\frac{d}{d\theta} \int T(x)p_\theta(x) dx = \int T(x) \frac{d}{d\theta} p_\theta(x) dx.$$

Then

$$\text{Var}_\theta(Y) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}, \quad \forall \theta \in \Theta.$$

In particular, if  $Y$  is unbiased for  $\theta$ , then  $g(\theta) = \theta$ , so,

$$\text{Var}_\theta(Y) \geq \frac{1}{I_X(\theta)}, \quad \forall \theta \in \Theta.$$

Equality occurs for some  $\theta \in \Theta$  only when  $\frac{d}{d\theta} \log f_\theta(x)$  and  $Y - \mathbb{E}_\theta Y$  are multiples of each other.

**Remark.** For a one-parameter family of distributions, the equality case of Theorem 10.4.7 gives a new way to find a UMVU that avoids any discussion of complete sufficient statistics. This is another way to find a UMVU ( $\frac{d}{d\theta} \log f_\theta(X)$ ) that sidesteps the need for a complete sufficient statistic. That is, to find a UMVU, we look for affine functions of  $\frac{d}{d\theta} \log f_\theta(X)$ .

*Proof.*

$$|g'(\theta)| = \left| \frac{d}{d\theta} \int_{\mathbb{R}} f_\theta(x)t(x)dx \right| = \left| \int_{\mathbb{R}} \left( \frac{d}{d\theta} \log f_\theta(x) \right) t(x)f_\theta(x)dx \right| = \left| \mathbb{E}_\theta \frac{d}{d\theta} \log f_\theta(X) t(X) \right|$$

Note that  $\mathbb{E}_\theta \frac{d}{d\theta} \log f_\theta(X) = 0$ , so this can be written as

$$= \left| \text{Cov}_\theta \left( \frac{d}{d\theta} \log f_\theta(X), t(X) \right) \right|$$

Then by Remark 1.63 in math 541A notes, by the Cauchy-Schwarz inequality,

$$\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \leq \sqrt{\text{Var}_\theta(X)\text{Var}_\theta(Y)},$$

so we have

$$\left| \text{Cov}_\theta \left( \frac{d}{d\theta} \log f_\theta(X), t(X) \right) \right| \leq \sqrt{\text{Var}_\theta \left( \frac{d}{d\theta} \log f_\theta(X) \right) \text{Var}_\theta(Y)} = \sqrt{I_X(\theta)} \sqrt{\text{Var}_\theta(Y)}$$

$$\iff |g'(\theta)|^2 \leq I_X(\theta) \text{Var}_\theta(Y) \iff \text{Var}_\theta(Y) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}, \quad \forall \theta \in \Theta.$$

Recall that equality occurs in the Cauchy-Schwarz inequality if and only if  $\frac{d}{d\theta} \log f_\theta(x)$  is a constant multiple of  $Y - \mathbb{E}_\theta Y$  with probability 1. (See also Corollary 7.3.15 in Casella and Berger [2001, p. 341].)

□

**Example 10.18 (Example 6.24).** Suppose  $f_\theta(x) := \theta x^{\theta-1} \mathbf{1}_{0 < x < 1}$  for all  $x \in \mathbb{R}, \theta > 0$ . (This is a beta distribution with  $\beta = 1$ .) We have

$$\frac{d}{d\theta} \log f_\theta(x) = \frac{1}{\theta} + \log x, \quad \forall 0 < x < 1.$$

A vector  $X = (X_1, \dots, X_n)$  of  $n$  independent samples from  $f_\theta$  is distributed according to the product  $\prod_{i=1}^n f_\theta(x_i)$ , so that

$$\frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(x_i) = \frac{d}{d\theta} \sum_{i=1}^n \log f_\theta(x_i) = \sum_{i=1}^n \left( \frac{1}{\theta} + \log x_i \right) = n \left( \frac{1}{\theta} + \frac{1}{n} \log \prod_{i=1}^n x_i \right), \quad \forall 0 < x_i < 1, 1 \leq i \leq n.$$

Then by Theorem 10.4.7 (Theorem 6.23 in Math 541A notes), any function of  $\frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(X_i)$  (plus a constant) is UMVU of its expectation. So, e.g.

$$Y := -\frac{1}{n} \log \prod_{i=1}^n X_i$$

is UMVU of its expectation, and  $\mathbb{E}_\theta Y = \theta^{-1}$  since  $\mathbb{E}_\theta \frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(X_i) = \mathbb{E} n \left( \frac{1}{\theta} + \frac{1}{n} \log \prod_{i=1}^n x_i \right) = 0$ .

**Definition 10.21 (Efficiency, Definition 6.25 in Math 541A notes).** Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable with distribution from a family of multivariable probability densities or probability mass functions  $\{f_\theta : \theta \in \Theta\}$  with  $\Theta \in \mathbb{R}$ . Let  $t : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $Y := t(X)$  be a statistic. Define the **efficiency** of  $Y$  to be

$$\frac{1}{I_X(\theta) \text{Var}_\theta(Y)}, \quad \forall \theta \in \Theta$$

if this quantity exists and is finite. If  $Z$  is another statistic, we define the **relative efficiency** of  $Y$  to  $Z$  to be

$$\frac{I_X(\theta) \text{Var}_\theta(Z)}{I_X(\theta) \text{Var}_\theta(Y)} = \frac{\text{Var}_\theta(Z)}{\text{Var}_\theta(Y)}, \quad \forall \theta \in \Theta.$$

**Definition 10.22 (Efficient estimator; Math 541B definition).** We say that  $\hat{\theta}_n$  is an **efficient** estimator of  $\theta$  if and only if its variance achieves the Cramer-Rao lower bound (**and it is unbiased?**).

**Definition 10.23 (Efficient estimator; Math 541B definition).** We say that  $\hat{\theta}_n$  is an **efficient** estimator of  $\theta$  if and only if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right).$$

**Superefficiency (Math 541B).** Let  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ . Then  $\bar{X}_n$  is efficient, and it is also asymptotically efficient. Now consider **Hodge's Estimator**:

$$T'(X_1, \dots, X_n) := \begin{cases} \bar{X}_n & |\bar{X}_n| \geq n^{-1/4} \\ 0 & |\bar{X}_n| < n^{-1/4}. \end{cases}$$

Claim:  $\sqrt{n}(T' - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$  for  $\theta \neq 0$ . Moreover, for  $r_n(T' - \theta) \xrightarrow{p} 0$  for any sequence  $r_n$  when  $\mu = 0$ .

#### 10.4.4 Bayes Estimation

In Bayes estimation, the parameter  $\theta \in \Theta$  is regarded as a random variable  $\Psi$ . The distribution of  $\Psi$  reflects our prior knowledge about the probable values of  $\Psi$ . Then, given that  $\Psi = \theta$ , the conditional distribution of  $X \mid \Psi = \theta$  is assumed to be  $\{f_\theta : \theta \in \Theta\}$ , where  $f_\theta : \mathbb{R}^n \rightarrow [0, \infty)$ . Suppose  $t : \mathbb{R}^n \rightarrow \mathbb{R}^k$  and we have a statistic  $Y := t(X)$  and a loss function  $\ell : \mathbb{R}^k \times \mathbb{R}^k \rightarrow \mathbb{R}$ . Let  $g : \Theta \rightarrow \mathbb{R}^k$ .

**Definition 10.24 (Bayes estimator, Defintion 6.26 in Math 541A notes).** A **Bayes estimator**  $Y$  for  $g(\theta)$  with respect to  $\Psi$  is defined such that

$$\mathbb{E}\ell(g(\Psi), Y) \leq \mathbb{E}\ell(g(\Psi), Z)$$

for all estimators  $Z$ . Here the expectation is with respect to both  $\Psi$  and  $Y$ . Note that we have not made any assumptions about bias for  $Y$  or  $Z$ . To find a Bayes estimator, it is sufficient to minimize the conditional risk.

**Remark.**  $t(X)$  can depend on  $\Psi$ .

**Proposition 10.4.8 (Proposition 6.27 in Math 541A notes).** Suppose there exists  $t : \mathbb{R}^k \rightarrow \mathbb{R}$  such that for almost every  $x \in \mathbb{R}^n$ ,  $Y := t(X)$  minimizes

$$\mathbb{E}(\ell(g(\Psi), Z) \mid X = x)$$

over all estimators  $Z$ . Then  $t(X)$  is a Bayes estimator for  $g(\theta)$  with respect to  $\Psi$ .

*Proof.* By assumption,

$$\mathbb{E}(\ell(g(\Psi), t(X)) \mid X = x) \leq \mathbb{E}(\ell(g(\Psi), Y) \mid X = x)$$

for any estimator  $Y$  and for almost every  $x$ . Taking expected values of both sides, we get

$$\mathbb{E}\ell(g(\Psi, t(X))) \leq \mathbb{E}\ell(g(\Psi, Y)).$$

□

**Example 10.19 (Example 6.29 in Math 541A notes).** Suppose  $n = 1$ ,  $g(\theta) = \theta$ , and  $\ell(\Psi, Y) = (\Psi - Y)^2$ . The minimum value of

$$\mathbb{E}[(\Psi - Y(X))^2 | X = x] = \mathbb{E}(\Psi^2 - 2\Psi t(X) + (t(X))^2 | X = x)$$

$$= \mathbb{E}(\Psi^2 | X = x) - 2t(X)\mathbb{E}(\Psi | X = x) + (t(X))^2$$

(where we remove expressions with  $x$  from the expectation because we take  $x$  to be fixed) occurs when  $t(x) = \mathbb{E}(\Psi | X = x)$ . So in this specific case the Bayes estimator is  $Y = t(X) = \mathbb{E}(\Psi | X)$ .

Given that  $\Psi = \theta < 0$ , suppose  $X$  is uniform on the interval  $[0, \theta]$ . (Suppose  $X$  is a single sample  $n = 1$  from this distribution.) Also assume that  $\Psi$  has the gamma distribution with parameters  $\alpha = 2$  and  $\beta = 1$ , so that  $\Psi$  has density  $\theta e^{-\theta} \mathbf{1}_{\{\theta>0\}}$ . The joint distribution of  $X$  and  $\Psi$  is then

$$f_{\Psi, X}(\theta, x) := \frac{1}{\theta} \mathbf{1}_{\{0 < x < \theta\}} \theta e^{-\theta} \mathbf{1}_{\{\theta>0\}} = \mathbf{1}_{\{0 < x < \theta\}} e^{-\theta}.$$

The marginal distribution of  $X$  is then

$$f_X(x) = \mathbf{1}_{\{x>0\}} \int_{-\infty}^{\infty} f_{\Psi, X}(\theta, x) d\theta = \mathbf{1}_{\{x>0\}} \int_x^{\infty} e^{-\theta} d\theta = \mathbf{1}_{\{x>0\}} e^{-x}.$$

So the conditional distribution of  $\Psi$  given  $X$  is

$$f_{\Psi|X=x}(\theta | x) = \frac{f_{\Psi, X}(\theta, x)}{f_X(x)} = \frac{e^{-\theta} \mathbf{1}_{\{0 < x < \theta\}}}{e^{-x} \mathbf{1}_{\{x>0\}}} = e^{x-\theta} \mathbf{1}_{\{0 < x < \theta\}}.$$

So,

$$\mathbb{E}(\Psi | X = x) = \int_{-\infty}^{\infty} \theta f_{\Psi|X=x}(\theta | x) d\theta = e^x \int_x^{\infty} \theta e^{-\theta} d\theta = e^x ((x+1)e^{-x}) = x+1.$$

So the Bayes estimator is  $Y = t(X) = \mathbb{E}(\Psi | X) = X + 1$ . This estimator minimizes  $\mathbb{E}(Y - Z)^2$  over all estimators  $Z$ . ( $\ell(a, b) = (a - b)^2$ ,  $g(\theta) = \theta$ ,  $\mathbb{E}\ell(g(\Psi), Z)$ ).

In contrast, the UMVU for one sample is  $2X$  by Theorem 10.4.3, since  $2X$  is complete sufficient and unbiased for  $\theta$  and  $\mathbb{E}_\theta(2X | 2X) = 2X$ . ( $X$  uniform on  $[0, \theta]$ ,  $\theta$  unknown,  $\mathbb{E}_\theta X = \theta/2, \forall \theta < 0$ .)

(Not obvious) For  $n$  samples,  $(1 + n^{-1})X_{(n)}$  is the UMVU for  $\theta$ . Note that  $2X$  is recovered when  $n = 1$ . Remarks: this estimator seems to be sufficient because you could factorize it as in the Factorization Theorem (Theorem 10.5).

**Exercise 20 (2018 DSO Statistics Group In-Class Screening Exam, Question 3).** Suppose that given the vector  $\mu$ , the random vector  $X$  has a normal distribution in  $\mathbb{R}^n$  with mean  $\mu$  and identity covariance matrix. We want to make inference about  $\|\mu\|^2$ .

- (a) Find an unbiased estimate of  $\|\mu\|^2$ . Call this estimator  $\hat{\delta}_{\text{unbiased}}$ .
- (b) Suppose that a Bayesian has a proper prior distribution for  $\mu$  that is Gaussian with mean vector 0 and covariance  $kI$ , where  $k$  is any fixed positive real number and  $I$  is the identity matrix. He wants to minimize mean squared error (MSE). The estimator minimizing the MSE is the posterior mean of  $\|\mu\|^2$ , i.e.,  $\mathbb{E}(\|\mu\|^2 | X)$ . Find this estimator. Call this estimator  $\hat{\delta}_{\text{proper}}$ .
- (c) Suppose now the Bayesian uses the uniform prior (which is also called a “flat” or “noninformative” prior) for  $\mu$ . Report  $\mathbb{E}(\|\mu\|^2 | X)$  in this case. Call it  $\hat{\delta}_{\text{flat}}$ . Report  $\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}}$ .
- (d) Now, if the true distribution of  $\mu$  is indeed Gaussian with mean vector 0 and covariance  $kI$ , then show that with respect to the unconditional (i.e. marginal) distribution of  $X$ , the Bayes estimator  $\hat{\delta}_{\text{proper}}$  is closer in Euclidean distance to  $\hat{\delta}_{\text{unbiased}}$  than it is to  $\hat{\delta}_{\text{flat}}$  when  $n$  is large. That is, show

$$\mathbb{E} \left( \hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}} \right)^2 < \mathbb{E} \left( \hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} \right)^2$$

for large  $n$ , where the expectation is over the unconditional distribution of  $X$ , which is

$$\int_{\mathbb{R}^n} f(x | \mu) \pi(\mu) d\mu$$

with  $f(x | \mu) = \mathcal{N}_n(\mu, I)$  and  $\pi(\mu) = \mathcal{N}_n(0, kI)$ . (Hint: let  $\hat{D} = \hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}}$ . Compute the mean and variance of  $\hat{D}$  under the unconditional distribution of  $X$ .)

### Solution.

- (a) We have

$$X | \mu \sim \mathcal{N}(\mu, I_n)$$

Let  $X = (X_1, \dots, X_n)^T$  and let  $\mu = (\mu_1, \dots, \mu_n)^T$ . Notice that

$$\begin{aligned} \mathbb{E}(X^T X) &= \mathbb{E}[\mathbb{E}(X^T X | \mu)] = \mathbb{E} [\mathbb{E} (X_1^2 + X_2^2 + \dots + X_n^2 | \mu)] = \mathbb{E} \left[ \sum_{i=1}^n \mathbb{E}(X_i^2 | \mu) \right] \\ &= \mathbb{E} \left[ \sum_{i=1}^n \text{Var}(X_i | \mu) + \mathbb{E}(X_i | \mu)^2 \right] = \mathbb{E} \left[ \sum_{i=1}^n 1 + \mu_i^2 \right] = n + \mathbb{E}\|\mu\|_2^2 \\ \implies \mathbb{E}(X^T X - n) &= \mathbb{E}\|\mu\|_2^2 \end{aligned}$$

Therefore  $\hat{\delta}_{\text{unbiased}} = X^T X - n$  is unbiased for  $\mathbb{E}\|\mu\|_2^2$  (and given  $\mu$  it is unbiased for  $\|\mu\|_2^2$ ).

(b) We will begin by finding the posterior distribution of  $\mu$ . The prior distribution of  $\mu$  is

$$f(\boldsymbol{\mu}) = (2\pi)^{-n/2} |k\mathbf{I}_n|^{-1/2} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T(k\mathbf{I}_n)^{-1}\boldsymbol{\mu}\right) = \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2k}\boldsymbol{\mu}^T\boldsymbol{\mu}\right).$$

The likelihood is

$$\begin{aligned} f_{\mathbf{X}|\boldsymbol{\mu}}(\mathbf{x} | \boldsymbol{\mu}) &= (2\pi)^{-n/2} |\mathbf{I}_n|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{I}_n)^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})\right). \end{aligned}$$

So the unconditional distribution of  $\mathbf{X}$  is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \int_{\mathbb{R}^n} f_{\mathbf{X}|\boldsymbol{\mu}}(\mathbf{x} | \boldsymbol{\mu}) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ &= \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})\right) \cdot \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2k}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) d\boldsymbol{\mu} \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\left[\boldsymbol{\mu}^T\boldsymbol{\mu} + \frac{1}{k}\boldsymbol{\mu}^T\boldsymbol{\mu} - 2\mathbf{x}^T\boldsymbol{\mu} + \mathbf{x}^T\mathbf{x}\right]\right) d\boldsymbol{\mu} \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\boldsymbol{\mu}^T\boldsymbol{\mu} - \frac{2k}{k+1}\mathbf{x}^T\boldsymbol{\mu}\right] - \frac{1}{2}\mathbf{x}^T\mathbf{x}\right) d\boldsymbol{\mu} \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\boldsymbol{\mu}^T\boldsymbol{\mu} - \frac{2k}{k+1}\mathbf{x}^T\boldsymbol{\mu} + \left(\frac{k}{k+1}\right)^2\mathbf{x}^T\mathbf{x}\right] - \left(-\frac{k+1}{2k}\right)\left(\frac{k}{k+1}\right)^2\mathbf{x}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{x}\right) d\boldsymbol{\mu} \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\left(\boldsymbol{\mu} - \frac{k}{k+1}\mathbf{x}\right)^T\left(\boldsymbol{\mu} - \frac{k}{k+1}\mathbf{x}\right)\right] - \frac{1}{2}\left[-\left(\frac{k}{k+1}\right)\mathbf{x}^T\mathbf{x} + \frac{k+1}{k+1}\mathbf{x}^T\mathbf{x}\right]\right) d\boldsymbol{\mu} \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\left(\boldsymbol{\mu} - \frac{k}{k+1}\mathbf{x}\right)^T\left(\boldsymbol{\mu} - \frac{k}{k+1}\mathbf{x}\right)\right]\right) \exp\left(-\frac{1}{2}\frac{1}{k+1}\mathbf{x}^T\mathbf{x}\right) d\boldsymbol{\mu} \\ &= \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2}\frac{1}{k+1}\mathbf{x}^T\mathbf{x}\right) \left(\frac{k}{k+1}\right)^{n/2} \\ &\quad \cdot \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n}} \cdot \left(\frac{k}{k+1}\right)^{-n/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu} - \frac{k}{k+1}\mathbf{x}\right)^T\left(\frac{k}{k+1}\mathbf{I}_n\right)^{-1}\left(\boldsymbol{\mu} - \frac{k}{k+1}\mathbf{x}\right)\right) d\boldsymbol{\mu} \end{aligned}$$

The second row is the integral over  $\mathbb{R}^n$  of an  $n$ -dimensional multivariate Gaussian distribution with mean  $k/(k+1)\mathbf{x}$  and covariance  $k/(k+1)\mathbf{I}_n$ , so it equals 1. Then we are left with

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n}} \frac{1}{k^{n/2}} \exp\left(-\frac{1}{2}\frac{1}{k+1}\mathbf{x}^T\mathbf{x}\right) \left(\frac{k}{k+1}\right)^{n/2}$$

$$= \frac{1}{\sqrt{(2\pi)^n}} [(k+1)^n]^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T ((k+1)\mathbf{I}_n)^{-1} \mathbf{x}\right) \quad (10.17)$$

which is the density of an  $n$ -dimensional multivariate Gaussian random variable with mean  $\mathbf{0}$  and covariance  $(k+1)\mathbf{I}_n$ . Therefore the posterior distribution of  $\boldsymbol{\mu}$  is

$$\begin{aligned} f_{\boldsymbol{\mu}|\mathbf{X}}(\boldsymbol{\mu} | \mathbf{x}) &= \frac{f_{\mathbf{X}|\boldsymbol{\mu}}(\mathbf{x} | \boldsymbol{\mu})f(\boldsymbol{\mu})}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \left[ \frac{1}{(2\pi\sqrt{k})^n} \exp\left(-\frac{1}{2} \left[ \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{k} \boldsymbol{\mu}^T \boldsymbol{\mu} - 2\mathbf{x}^T \boldsymbol{\mu} + \mathbf{x}^T \mathbf{x} \right] \right) \right] \Bigg/ \left[ \frac{1}{\sqrt{(2\pi)^n}} [(k+1)^n]^{-1/2} \exp\left(-\frac{1}{2} \frac{1}{k+1} \mathbf{x}^T \mathbf{x}\right) \right] \\ &= \frac{1}{\sqrt{(2\pi)^n}} \left( \frac{k+1}{k} \right)^{n/2} \exp\left(-\frac{k+1}{2k} \left[ \left( \boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right)^T \left( \boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right) \right] - \frac{1}{2} \frac{1}{k+1} \mathbf{x}^T \mathbf{x} + \frac{1}{2} \frac{1}{k+1} \mathbf{x}^T \mathbf{x} \right) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \left( \frac{k+1}{k} \right)^{n/2} \exp\left(-\frac{k+1}{2k} \left( \boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right)^T \left( \boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right) \right) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \cdot \left( \frac{k}{k+1} \right)^{-n/2} \exp\left(-\frac{1}{2} \left( \boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right)^T \left( \frac{k}{k+1} \mathbf{I}_n \right)^{-1} \left( \boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right) \right) \end{aligned}$$

which is an  $n$ -dimensional multivariate Gaussian distribution with mean  $k/(k+1)\mathbf{x}$  and covariance  $k/(k+1)\mathbf{I}_n$ . That is, conditional on  $\mathbf{X}$ ,  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ , where

$$\begin{aligned} \mu_i | \mathbf{X} &\stackrel{i.i.d.}{\sim} \mathcal{N}\left(\frac{k}{k+1}X_i, \frac{k}{k+1}\right). \\ \iff \left( \mu_i - \frac{k}{k+1}X_i \right) \frac{k+1}{k} | \mathbf{X} &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1) \iff \frac{k+1}{k}\mu_i - X_i | \mathbf{X} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1) \\ \implies \mathbb{E}\left(\left[\frac{k+1}{k}\mu_i - X_i\right]^2 | \mathbf{X}\right) = 1 &\iff \mathbb{E}\left(\left[\frac{k+1}{k}\right]^2 \mu_i^2 + X_i^2 - 2\frac{k+1}{k}\mu_i X_i | \mathbf{X}\right) = 1 \quad (10.18) \end{aligned}$$

So,

$$\begin{aligned} \hat{\delta}_{\text{proper}} &= \mathbb{E}(\|\boldsymbol{\mu}\|_2^2 | \mathbf{X}) = \mathbb{E}\left(\sum_{i=1}^n \mu_i^2 | \mathbf{X}\right) = \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n \left[\frac{k+1}{k}\right]^2 \mu_i^2 | \mathbf{X}\right) \\ &= \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n \left[\frac{k+1}{k}\right]^2 \mu_i^2 + X_i^2 - 2\frac{k+1}{k}\mu_i X_i | \mathbf{X}\right) - \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n X_i^2 - 2\frac{k+1}{k}\mu_i X_i | \mathbf{X}\right) \\ &= \left[\frac{k}{k+1}\right]^2 - \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 + 2\frac{k+1}{k} \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i \mathbb{E}(\mu_i | \mathbf{X}) \end{aligned}$$

$$\begin{aligned}
&= \left[ \frac{k}{k+1} \right]^2 - \left[ \frac{k}{k+1} \right]^2 \sum_{i=1}^n X_i^2 + 2 \frac{k}{k+1} \sum_{i=1}^n X_i \cdot \frac{k}{k+1} X_i \\
&= \left[ \frac{k}{k+1} \right]^2 + 2 \left[ \frac{k}{k+1} \right]^2 \sum_{i=1}^n X_i^2 - \left[ \frac{k}{k+1} \right]^2 \sum_{i=1}^n X_i^2 = \left[ \frac{k}{k+1} \right]^2 (1 + \mathbf{X}^T \mathbf{X}).
\end{aligned}$$

- (c) We will again begin by finding the posterior distribution of  $\mu$ . The (improper) prior distribution of  $\mu$  is constant; that is, for some  $c \in \mathbb{R}$ ,

$$f(\boldsymbol{\mu}) = c, \quad \forall \boldsymbol{\mu} \in \mathbb{R}^n.$$

The likelihood is

$$\begin{aligned}
f_{\mathbf{X}|\boldsymbol{\mu}}(\mathbf{x} | \boldsymbol{\mu}) &= (2\pi)^{-n/2} |\mathbf{I}_n|^{-1/2} \cdot \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{I}_n)^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \\
&= (2\pi)^{-n/2} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}) \right).
\end{aligned}$$

So the unconditional distribution of  $\mathbf{X}$  is

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}) &= \int_{\mathbb{R}^n} f_{\mathbf{X}|\boldsymbol{\mu}}(\mathbf{x} | \boldsymbol{\mu}) f(\boldsymbol{\mu}) d\boldsymbol{\mu} = c \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}) \right) d\boldsymbol{\mu} \\
&= c \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp \left( -\frac{1}{2} (\boldsymbol{\mu} - \mathbf{x})^T (\boldsymbol{\mu} - \mathbf{x}) \right) d\boldsymbol{\mu}
\end{aligned}$$

The expression inside the integral is the density of a Gaussian random variable with mean  $\mathbf{x}$  and covariance  $\mathbf{I}_n$ , so the integral evaluates to 1. Therefore the unconditional distribution of  $\mathbf{X}$  is also flat. Therefore the posterior distribution of  $\boldsymbol{\mu}$  is the same as the likelihood:

$$f_{\boldsymbol{\mu}|\mathbf{X}}(\boldsymbol{\mu} | \mathbf{x}) = \frac{f_{\mathbf{X}|\boldsymbol{\mu}}(\mathbf{x} | \boldsymbol{\mu}) f(\boldsymbol{\mu})}{f_{\mathbf{X}}(\mathbf{x})} = (2\pi)^{-n/2} \exp \left( -\frac{1}{2} (\boldsymbol{\mu} - \mathbf{x})^T (\boldsymbol{\mu} - \mathbf{x}) \right);$$

that is, conditional on  $\mathbf{X}$ ,  $\boldsymbol{\mu}$  is normally distributed with mean  $\mathbf{x}$  and covariance  $\mathbf{I}_n$ . So conditional on  $\mathbf{X}$ ,  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ , where

$$\mu_i | \mathbf{X} \stackrel{i.i.d.}{\sim} \mathcal{N}(X_i, 1).$$

$$\iff \mu_i - X_i | \mathbf{X} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1) \implies \mathbb{E}([\mu_i - X_i]^2 | \mathbf{X}) = 1 \iff \mathbb{E}(\mu_i^2 + X_i^2 - 2\mu_i X_i | \mathbf{X}) = 1 \tag{10.19}$$

So,

$$\hat{\delta}_{\text{flat}} = \mathbb{E}(\|\boldsymbol{\mu}\|_2^2 | \mathbf{X}) = \mathbb{E} \left( \sum_{i=1}^n \mu_i^2 | \mathbf{X} \right) = \mathbb{E} \left( \sum_{i=1}^n \mu_i^2 + X_i^2 - 2\mu_i X_i | \mathbf{X} \right) - \mathbb{E} \left( \sum_{i=1}^n X_i^2 - 2\mu_i X_i | \mathbf{X} \right)$$

$$= 1 - \sum_{i=1}^n X_i^2 + 2 \sum_{i=1}^n X_i \mathbb{E}(\mu_i | \mathbf{X}) = 1 - \sum_{i=1}^n X_i^2 + 2 \sum_{i=1}^n X_i^2 = 1 + \mathbf{X}^T \mathbf{X}.$$

So,

$$\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} = 1 + \mathbf{X}^T \mathbf{X} - (\mathbf{X}^T \mathbf{X} - n) = 1 + n. \quad (10.20)$$

(d) If the true distribution of  $\boldsymbol{\mu}$  is the prior from part (b), then the marginal distribution of  $\mathbf{X}$  is (10.17):

$$= \frac{1}{\sqrt{(2\pi)^n}} [(k+1)^n]^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T ((k+1)\mathbf{I}_n)^{-1} \mathbf{x}\right);$$

that is,  $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, (k+1)\mathbf{I}_n)$ . (Note that this means  $(k+1)^{-1}X_i$  is standard Gaussian and i.i.d. for all  $i \in \{1, \dots, n\}$ .) Per the suggestion, let

$$\begin{aligned} \hat{D} &= \hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}} = \left[\frac{k}{k+1}\right]^2 (1 + \mathbf{X}^T \mathbf{X}) - (\mathbf{X}^T \mathbf{X} - n) = \frac{k^2 - (k^2 + 2k + 1)}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1}\right]^2 + n \\ &= -\frac{2k+1}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1}\right]^2 + n \end{aligned} \quad (10.21)$$

Since from (10.20) we have

$$\mathbb{E}(\hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}})^2 = n^2 + 2n + 1,$$

we seek

$$\mathbb{E}(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}})^2 = \mathbb{E}(\hat{D}^2) = \text{Var}(\hat{D}) + [\mathbb{E}(\hat{D})]^2.$$

Note that since  $(k+1)^{-1}X_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$  for all  $i \in \{1, \dots, n\}$ ,

$$\sum_{i=1}^n ((k+1)^{-1}X_i)^2 \sim \chi_n^2,$$

so

$$\mathbb{E}\left[\sum_{i=1}^n \left(\frac{1}{k+1}X_i\right)^2\right] = n \iff \frac{1}{(k+1)^2} \mathbb{E}\left[\sum_{i=1}^n X_i^2\right] = n \iff \mathbb{E}\left[\sum_{i=1}^n X_i^2\right] = n(k+1)^2,$$

and

$$\text{Var}\left[\sum_{i=1}^n \left(\frac{1}{k+1}X_i\right)^2\right] = 2n \iff \frac{1}{(k+1)^4} \text{Var}\left[\sum_{i=1}^n X_i^2\right] = 2n \iff \text{Var}\left[\sum_{i=1}^n X_i^2\right] = 2n(k+1)^4.$$

Under the unconditional distribution of  $\mathbf{X}$ ,

$$\begin{aligned}
\mathbb{E}(\hat{D}) &= \mathbb{E}\left(-\frac{2k+1}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1}\right]^2 + n\right) = -\frac{2k+1}{(k+1)^2} \mathbb{E}(\mathbf{X}^T \mathbf{X}) + \left[\frac{k}{k+1}\right]^2 + n \\
&= -\frac{2k+1}{(k+1)^2} \mathbb{E}\left(\sum_{i=1}^n X_i^2\right) + \left[\frac{k}{k+1}\right]^2 + n = -\frac{2k+1}{(k+1)^2} n(k+1)^2 + \left[\frac{k}{k+1}\right]^2 + n \\
&= -n \frac{(2k+1)(k^2+2k+1)+k^2}{(k+1)^2} + n = n \left[ \frac{k^2+2k+1}{(k+1)^2} - \frac{2k^3+4k^2+2k+k^2+2k+1+k^2}{(k+1)^2} \right] \\
&= n \left[ \frac{-2k^3-5k^2-2k}{(k+1)^2} \right] = -n \left[ \frac{2k^3+5k^2+2k}{(k+1)^2} \right]
\end{aligned}$$

Next,

$$\begin{aligned}
\text{Var}(\hat{D}) &= \text{Var}\left(-\frac{2k+1}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1}\right]^2 + n\right) = \frac{(2k+1)^2}{(k+1)^4} \text{Var}\left(\sum_{i=1}^n X_i^2\right) \\
&= \frac{(2k+1)^2}{(k+1)^4} 2n(k+1)^4 = 2n(2k+1)^2
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}})^2 &= 2n(2k+1)^2 + \left(-n \left[ \frac{2k^3+5k^2+2k}{(k+1)^2} \right]\right)^2 = n^2 \left[ \frac{2k^3+5k^2+2k}{(k+1)^2} \right]^2 + n \cdot 2(2k+1)^2 \\
&\approx 4k^2 n^2 + 8k^2 n,
\end{aligned}$$

so in general when  $k \geq 1$  and  $n$  is large,,  $\mathbb{E}(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}})^2 > \mathbb{E}(\hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}})^2$ ; that is, the flat prior Bayes estimator is further from the unbiased estimator than the proper prior Bayes estimator.

#### 10.4.5 Method of Moments

**Definition 10.25 (Consistency, Definition 6.30 in Math 541A notes).** Let  $\{f_\theta : \theta \in \Theta\}$  be a family of distributions. Let  $Y_1, Y_2, \dots$  be a sequence of estimators of  $g(\theta)$ . We say that  $Y_1, Y_2, \dots$  is **consistent** for  $g(\theta)$  if for any  $\theta \in \Theta$ ,  $Y_1, Y_2, \dots$  converges in probability to the constant value  $g(\theta)$  with respect to the probability distribution  $f_\theta$ . That is,  $Y_n \xrightarrow{P} g(\theta)$ . (Typically we will take  $Y_n$  to be a function of a random sample of size  $n$  for all  $n \geq 1$ .)

**Example 10.20 (Example 6.31 in Math 541A notes).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  with distribution  $f_\theta$ . The Weak Law of Large Numbers (Theorem 8.6.1) says that the sample mean is consistent when  $\mathbb{E}_\theta|X_1| < \infty$  for all  $\theta \in \Theta$ . More generally, if  $j \geq 1$  is a positive integer such that  $\mathbb{E}_\theta|X_1|^j < \infty$  for all  $\theta \in \Theta$ , then the  $j$ th sample moment

$$M_j = M_{j,n}(\theta) := \frac{1}{n} \sum_{i=1}^n X_i^j$$

is a consistent estimator for  $\mu_j(\theta) := \mathbb{E}X_1^j$ .

**Definition 10.26 (Method of Moments, Definition 6.32 in Math 541A notes).** Let  $g : \Theta \rightarrow \mathbb{R}^k$ . Suppose we want to estimate  $g(\theta)$  for any  $\theta \in \Theta$ . Suppose there exists  $h : \mathbb{R}^j \rightarrow \mathbb{R}^k$  such that  $g(\theta) = h(\mu_1, \dots, \mu_j)$ . Then the estimator

$$h(M_1, \dots, M_j)$$

is a **method of moments** estimator for  $g(\theta)$ , where  $M_j$  is the  $j$ th sample moment

$$M_j = M_{j,n}(\theta) := \frac{1}{n} \sum_{i=1}^n X_i^j$$

**Example 10.21 (Example 6.33 in Math 541A notes).** Recall that the standard deviation is

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

To estimate the standard deviation, we can use  $\Theta = \mathbb{R} \times (0, \infty) = \{(\mu_1, \mu_2) : \mu_1 \in \mathbb{R}, \mu_2 > 0\}$ ,  $j = 2$ , and  $h(\mu_1, \mu_2) = \sqrt{\mu_2 - \mu_1^2}$ , so that the method of moments estimator of the standard deviation is  $\sqrt{M_2 - M_1^2}$ .

**Remark.** The method of moments estimator is not necessarily unbiased.

#### 10.4.6 Maximum likelihood estimator

**Definition 10.27 (Maximum Likelihood Estimator (Math 541B Definition)).**  $P = \{P_\theta, \theta \in \Theta\}$ ,  $X_1, \dots, X_n$  i.i.d.  $P_{\theta_0}, \theta_0 \in \Theta$ . The likelihood function

$$L_n(\theta | X_1, \dots, X_n) := \prod_{j=1}^n p_\theta(x_j)$$

$$\ell_n(\theta | X_1, \dots, X_n) = \log(L_n(\theta | X_1, \dots, X_n)) = \sum_{i=1}^n \log p_\theta(X_j)$$

Then the MLE  $\hat{\theta}_n$  is defined as

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \{L_n(\theta)\} = \arg \max_{\theta \in \Theta} \{\ell_n(\theta)\}.$$

**Remark.** Under some reasonable assumptions, the maximum likelihood estimator is consistent. However, the MLE does not always exist and might not be unique. See Keener for details.

**Remark (Math 541 B remarks).** Equivalently,  $\hat{\theta}_n$  maximizes

$$\prod_{j=1}^n \frac{p_\theta(X_j)}{p_{\theta_0}(X_j)}$$

$$\iff \hat{\theta}_n = \arg \max_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \log \left[ \frac{p_\theta(X_j)}{p_{\theta_0}(X_j)} \right] \right\}$$

As  $n \rightarrow \infty$ , for every  $\theta \in \Theta$ , by the Law of Large Numbers

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \log \left[ \frac{p_\theta(X_j)}{p_{\theta_0}(X_j)} \right] &\xrightarrow{a.s.} \mathbb{E}_{\theta_0} \log \left( \frac{p_\theta(X)}{p_{\theta_0}(X)} \right) \\ \iff \hat{\theta}_n &\approx \arg \min_{\theta \in \Theta} \mathbb{E}_{\theta_0} \log \left( \frac{p_{\theta_0}(X)}{p_\theta(X)} \right) =: KL(p_{\theta_0} || p_\theta) \end{aligned}$$

where  $KL(p_{\theta_0} || p_\theta)$  is the **Kullback-Leibler Divergence**. Facts:

1.

$$KL(p_{\theta_0} || p_\theta) = 0 \iff p_{\theta_0} = p_\theta \iff \theta = \theta_0 \text{ (identifiability)}$$

*Proof.*

$$KL(p_{\theta_0} || p_\theta) = \mathbb{E}_{\theta_0} - \log \left( \frac{p_\theta(X)}{p_{\theta_0}(X)} \right) \geq \text{(by Jensen's Inequality)} - \log \left( \mathbb{E}_{\theta_0} \left[ \log \left( \frac{p_\theta(X)}{p_{\theta_0}(X)} \right) \right] \right) = 0$$

for Jensen's Inequality, equality holds if and only if

$$\mathbb{P} \left( \frac{p_\theta(X)}{p_{\theta_0}(X)} = 1 \right) = 1$$

□

2.  $KL(p || q)$  is not a distance:  $KL(p || q) \neq KL(q || p)$ , and the KL divergence does not satisfy the Triangle Inequality.
3. **total variational distance**: (in principal, existence of density is not required, but typically we assume a density exists)

$$TV(P, Q) := \sup_A |P(A) - Q(A)| = \frac{1}{2} \int |p - q| d\mu$$

**Pinsker's Inequality:**

$$TV(P, Q) \leq \sqrt{\frac{KL(P || Q)}{2}}.$$

**Hellinger's Distance:** (need to assume existence of density)

$$H^2(P, Q) := \int (\sqrt{p} - \sqrt{q})^2 d\mu$$

( $\ell_2$  distance between square roots of densities)

**Le Cann's Inequality:**

$$TV(P, Q) \geq \frac{1}{2} H^2(P, Q)$$

Using Pinsker's Inequality and Le Cann's Inequality, you can relate KL divergence to Hellinger's Distance.

For more information on KL Divergence, see Sections 6.5 and 11.11.1).

**Proposition 10.4.9 (Math 541A Proposition 6.40).** For all  $i \in 1, \dots, n$ , if  $\Theta \rightarrow f_\theta(x_i)$  is strictly log-concave, then  $\ell(\theta)$  has at most one maximum value.

**Remark.** One example of a function whose log likelihood has no maximum value is  $\exp(-e^{-\theta})$  (as in an extreme value distribution).

**Remark.** Intuition of Lemma 6.50 in Math 541A: if distribution follows  $\theta$  then it is more likely to match distribution function of  $f_\theta$  than  $f_\omega$ .

Note on Theorem 6.53: exponential family is an example of a family that satisfies condition (a).

Note from proof:

$$\sqrt{n}\ell_n(\theta) = \frac{1}{\sqrt{n}} \left( \sum_{i=1}^n \log f_\theta(x_i) - \mathbb{E}(\log f_\theta(x_i)) \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_{X_1}(\theta)}\right)$$

by the Central Limit Theorem (and Assumption 3 for the variance) since  $\mathbb{E}(\log f_\theta(x_i)) = 0$ . Also,

$$\ell''_n(\theta') = \frac{1}{n} \sum_{i=1}^n \frac{d^2}{d[\theta']^2} \log f_{\theta'}(x_i)$$

which explains the applicability of the Weak Law of Large Numbers.

**Proposition 10.4.10 (Stats 100B homework problem).** Suppose  $X_1, X_2, \dots, X_n$  is a random sample from a Bernoulli( $p$ ) distribution. Let  $X = \sum_{i=1}^n X_i$ . Then

- (a) The maximum likelihood estimator of  $p$  is  $\hat{p} = X/n$ .
- (b) The maximum likelihood estimator attains the Cramer-Rao lower bound.
- (c) The maximum likelihood estimator is a consistent estimator for  $p$ .
- (d)  $\frac{\hat{p}(1-\hat{p})}{n-1}$  is an unbiased estimator for  $\text{Var}(\hat{p}) = p(1-p)/n$ .

*Proof.* a. Bernoulli random variable:

$$P(X_i = x) = p^x (1-p)^{1-x}$$

Assuming independent samples,

$$L = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i} (1-p)^{\sum_{i=1}^n 1-X_i}$$

$$\log(L) = \sum_{i=1}^n X_i \log(p) + \left( \sum_{i=1}^n 1 - X_i \right) \log(1 - p)$$

$$\frac{d \log(L)}{dp} = \frac{1}{p} \sum_{i=1}^n X_i - \frac{1}{1-p} \sum_{i=1}^n (1 - X_i) = 0$$

$$\frac{1}{p} \sum_{i=1}^n X_i = \frac{1}{1-p} \sum_{i=1}^n (1 - X_i)$$

$$(1 - \hat{p}) \sum_{i=1}^n X_i = \hat{p} \sum_{i=1}^n (1 - X_i)$$

$$\sum_{i=1}^n X_i = \hat{p} \sum_{i=1}^n (X_i + 1 - X_i)$$

$$\sum_{i=1}^n X_i = n\hat{p}$$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

b.

$$\text{Var}(\hat{p}) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right)$$

Since  $X_i$  are independent, we can write this as

$$\frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i)$$

And since  $X_i$  is Bernoulli,  $\text{Var}(X_i) = p(1 - p)$ .

$$= \frac{1}{n^2} \sum_{i=1}^n p(1 - p) = \frac{1}{n^2} np(1 - p) = \boxed{\frac{p(1 - p)}{n}}$$

Cramer-Rao lower bound:

$$\text{Var}(\hat{\theta}) \geq 1 / \left( -n \mathbb{E} \left[ \frac{\partial^2 \log(f(X; \theta))}{\partial \theta^2} \right] \right)$$

$$\frac{\partial}{\partial p} \log(p^x(1 - p)^{1-x}) = \frac{\partial}{\partial p} (x \log(p) + (1 - x) \log(1 - p)) = \frac{x}{p} - \frac{1 - x}{1 - p}$$

$$\frac{\partial^2 \log(f(X; \theta))}{\partial \theta^2} = \frac{\partial}{\partial p} \left( \frac{x}{p} - \frac{1 - x}{1 - p} \right) = -\frac{x}{p^2} - \frac{1 - x}{(1 - p)^2}$$

$$\mathbb{E} \left[ \frac{\partial^2 \log(f(X; \theta^2))}{\partial \theta} \right] = \mathbb{E} \left( -\frac{x}{p^2} - \frac{1 - x}{(1 - p)^2} \right) = -\frac{1}{p^2} \mathbb{E}(x) - \frac{1}{(1 - p)^2} \mathbb{E}(1 - x) = -\frac{1}{p^2} p - \frac{1}{(1 - p)^2} (1 - p)$$

$$\begin{aligned}
&= -\frac{1}{p} - \frac{1}{1-p} = -\frac{1-p}{p(1-p)} - \frac{p}{p(1-p)} = \frac{-1}{p(1-p)} \\
\implies \text{Var}(\hat{p}) &\geq 1 / \left( -n \left( \frac{-1}{p(1-p)} \right) \right) = \frac{p(1-p)}{n} = \text{Var}(\hat{p})
\end{aligned}$$

c. (1) **Unbiased:**

$$E\left(\frac{X}{n}\right) = \frac{np}{p} = p$$

(2)  $\text{Var}(\hat{\theta}) \rightarrow 0$  as  $n \rightarrow \infty$  :

$$\lim_{n \rightarrow \infty} \text{Var}\left(\frac{X}{n}\right) = \lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \lim_{n \rightarrow \infty} \frac{1}{n^2} \cdot np(1-p) = \lim_{n \rightarrow \infty} \frac{p(1-p)}{n} = \boxed{0}$$

Therefore  $\frac{X}{n}$  is a consistent estimator of  $p$ .

d.

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2) &= \mathbb{E}\left[\frac{1}{n}\left(\frac{X}{n}\left(1-\frac{X}{n}\right)\right)\right] = \mathbb{E}\left[\frac{X(n-X)}{n^3}\right] = \frac{1}{n^3}\mathbb{E}[nX - (X)^2] = \frac{1}{n^2}\mathbb{E}(X) - \frac{1}{n^3}\mathbb{E}(X^2) \\
&= \frac{1}{n^2} \cdot np - \frac{1}{n^3}(\text{Var}(X) + (E(X))^2) = \frac{p}{n} - \frac{np(1-p)}{n^3} - \frac{p^2 n^2}{n^3} = \frac{pn - p + p^2 - p^2 n}{n^2} \\
&= \frac{p(n-1+p-pn)}{n^2} = \frac{p(n-1)(1-p)}{n^2}
\end{aligned}$$

This is a biased estimator since  $\text{Var}(X) = \frac{p(1-p)}{n}$  (since  $X$  is binomial).

$$c \cdot \frac{p(n-1)(1-p)}{n^2} = \frac{p(1-p)}{n} \implies \boxed{c = \frac{n}{n-1}}$$

□

**Proposition 10.4.11 (Stats 100B homework problem).** Suppose that  $X$  follows a geometric distribution and we take an i.i.d. sample of size  $n$ . Then the maximum likelihood estimator of  $p$  is

$$\hat{p} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}}.$$

*Proof.* Since sample is i.i.d.:

$$L = \prod_{i=1}^n p(1-p)^{X_i-1} = p^n(1-p)^{-n+\sum_{i=1}^n X_i}$$

$$\log(L) = n \log(p) + \left( -n + \sum_{i=1}^n X_i \right) \log(1-p)$$

$$\frac{d \log(L)}{dp} = \frac{n}{p} - \frac{1}{1-p} \left( -n + \sum_{i=1}^n X_i \right) = 0$$

$$\frac{n}{\hat{p}} = \frac{1}{1-\hat{p}} \left( -n + \sum_{i=1}^n X_i \right)$$

$$(1-\hat{p})\hat{p} = -n\hat{p} + \hat{p} \sum_{i=1}^n X_i$$

$$n = \hat{p} \sum_{i=1}^n X_i$$

$$\hat{p} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}}$$

□

**Proposition 10.4.12 (Stats 100B homework problem).** Suppose  $X_1, X_2, \dots, X_n$  is a random sample from a  $\text{Poisson}(\lambda)$  distribution. Then

(a) The maximum likelihood estimator of  $\lambda$  is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i.$$

(b) The variance of the maximum likelihood estimator is

$$\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}$$

(c) The maximum likelihood estimator is a minimum variance unbiased estimator.

(d) The maximum likelihood estimator is consistent.

*Proof.* (a)

$$f(X_i; \lambda) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

Assuming the samples are independent,

$$L = \prod_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} = \left( e^{-n\lambda} \lambda^{\sum_{i=1}^n X_i} \right) / \prod_{i=1}^n X_i!$$

$$\log(L) = -n\lambda + \left( \sum_{i=1}^n X_i \right) \log(\lambda) - \sum_{i=1}^n \log(X_i!)$$

$$\frac{d \log(L)}{d\lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^n X_i = 0$$

$$\implies \hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{x}$$

(b)

$$\text{Var}(\hat{\lambda}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right)$$

Since  $X_i$  are i.i.d. this can be written as

$$= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \lambda = \frac{\lambda}{n}$$

(c) Cramer-Rao lower bound:

$$\text{Var}(\hat{\lambda}) \geq 1 / \left( -n \mathbb{E} \left[ \frac{\partial^2 \log(f(X; \lambda))}{\partial \lambda^2} \right] \right)$$

$$\log(f(X; \lambda)) = \log \left( \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \right) = X_i \log(\lambda) - \lambda - \log(X_i!)$$

$$\frac{\partial}{\partial \lambda} \log(f(X; \lambda)) = \frac{1}{\lambda} X_i - 1$$

$$\frac{\partial^2 \log(f(X; \lambda))}{\partial \lambda^2} = -\frac{1}{\lambda^2} X_i$$

$$\mathbb{E} \left[ \frac{\partial^2 \log(f(X; \lambda^2))}{\partial \lambda} \right] = -\frac{1}{\lambda^2} \mathbb{E}(X_i) = -\frac{1}{\lambda^2} \lambda = -\frac{1}{\lambda}$$

$$\implies \text{Var}(\hat{\lambda}) \geq 1 / \left( -n \mathbb{E} \left[ \frac{\partial^2 \log(f(X; \lambda))}{\partial \lambda^2} \right] \right) = \frac{1}{n/\lambda} = \boxed{\frac{\lambda}{n} = \text{Var}(\hat{\lambda})}$$

Since  $\text{Var}(\hat{\lambda})$  equals the Cramer-Rao lower bound,  $\hat{\lambda}$  is a MVUE.

(d) We already know the MLE is unbiased. To show consistency, we show  $\text{Var}(\hat{\theta}) \rightarrow 0$  as  $n \rightarrow \infty$ .

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\lambda}) = \lim_{n \rightarrow \infty} \frac{\lambda}{n} = \boxed{0}$$

Therefore  $\hat{\lambda}$  is a consistent estimator of  $\lambda$ .

□

**Proposition 10.4.13 (Stats 100B homework problem, similar to Math 541A example 6.47).**

Suppose  $X_1, X_2, \dots, X_n$  is a random sample from a  $\text{Exponential}(\lambda)$  distribution. Then the maximum likelihood estimator of  $\lambda$  is

$$\hat{\lambda} = n / \sum_{i=1}^n X_i = \frac{1}{\bar{X}}.$$

*Proof.*

$$f(X_i; \lambda) = \lambda e^{-\lambda X_i}$$

Assuming the samples are independent,

$$L = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n \exp(-\lambda \sum_{i=1}^n X_i)$$

$$\log(L) = n \log(\lambda) + -\lambda \sum_{i=1}^n X_i$$

$$\frac{d \log(L)}{d \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n X_i = 0$$

$$\implies \hat{\lambda} = n / \sum_{i=1}^n X_i = \frac{1}{\bar{X}}$$

□

**Proposition 10.4.14 (Stats 100B homework problem; similar to Math 541A Example 6.45).**

Let  $X_1, X_2, \dots, X_n$  be an i.i.d. random sample from a normal population with mean zero and unknown variance  $\sigma^2$ . Then

- (a) The maximum likelihood estimator of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

- (b) The maximum likelihood estimator of  $\sigma^2$  is biased, but asymptotically unbiased.

- (c) The maximum likelihood estimator of  $\sigma^2$  has variance

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{2\sigma^4}{n}$$

and is consistent.

- (d) The variance of the maximum likelihood estimator of  $\sigma^2$  reaches the Cramer-Rao lower bound.

- (e) The maximum likelihood estimator of  $\mu$  is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

and it is unbiased and UMVU/MVUE.

*Proof.* a. Since sample is i.i.d.  $\mathcal{N}(0, \sigma^2)$ :

$$\begin{aligned} L &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{X_i - \mu}{\sigma}\right]^2\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\right) \\ \log(L) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - (\sigma^2)^{-1} \frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2 \\ \frac{\partial \log(L)}{\partial \sigma^2} &= -\frac{n}{2} \frac{1}{\sigma^2} + (\sigma^2)^{-2} \frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2 = 0 \\ \frac{\sum_{i=1}^n (X_i - \mu)^2}{(\hat{\sigma}^2)^2} &= \frac{n}{\hat{\sigma}^2} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \end{aligned}$$

b. something wrong with this part

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n X_i^2\right)$$

Since the sample is i.i.d., this can be written as

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i^2)$$

Since  $X_i \sim \mathcal{N}(0, \sigma^2)$ ,  $X_i^2 / \sigma^2 \sim \chi_1^2$ . So we have

$$\mathbb{E}\left(\frac{X_i^2}{\sigma^2}\right) = 1$$

$$\frac{1}{\sigma^2} \mathbb{E}(X_i^2) = 1$$

$$\mathbb{E}(X_i^2) = \sigma^2$$

Therefore

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{1}{n} \sum_{i=1}^n \sigma^2 = \frac{1}{n} n \sigma^2 = \boxed{\sigma^2}$$

However it is asymptotically biased: that is,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}(\hat{\sigma}^2)}{\sigma^2} = 1$$

c.

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i^2\right)$$

Since  $X_i$  is i.i.d. this can be written as

$$\frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i^2)$$

Again, ,  $X_i^2/\sigma^2 \sim \chi_1^2$ , so we have

$$\text{Var}\left(\frac{X_i^2}{\sigma^2}\right) = 2$$

$$\frac{1}{\sigma^4} \text{Var}(X_i^2) = 2$$

$$\text{Var}(X_i^2) = 2\sigma^4$$

Therefore

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{1}{n^2} \sum_{i=1}^n 2\sigma^4 = \frac{2n\sigma^4}{n^2} = \boxed{\frac{2\sigma^4}{n}}$$

Test for consistency (already known that estimate is unbiased):

$$\lim_{n \rightarrow \infty} \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = \lim_{n \rightarrow \infty} \frac{2\sigma^4}{n} = \boxed{0}$$

So this is a consistent estimator of  $\sigma^2$ .

d. Cramer-Rao lower bound:

$$\text{Var}(\hat{\theta}) \geq 1 / \left( -n \mathbb{E} \left[ \frac{\partial^2 \log(f(X; \theta))}{\partial \theta^2} \right] \right)$$

$$\log(f(X; \theta)) = \log \left[ \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left[ \frac{X_i}{\sigma} \right]^2 \right) \right] = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2} X_i^2 (\sigma^2)^{-1}$$

$$\frac{\partial}{\partial \sigma^2} \left( -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2} X_i^2 (\sigma^2)^{-1} \right) = -\frac{1}{2\sigma^2} + \frac{1}{2} (X_i)^2 (\sigma^2)^{-2}$$

$$\frac{\partial^2 \log(f(X; \theta))}{\partial (\sigma^2)^2} = \frac{1}{2} (\sigma^2)^{-2} - X_i^2 (\sigma^2)^{-3}$$

$$\mathbb{E} \left[ \frac{\partial^2 \log(f(X; \theta^2))}{\partial \theta} \right] = \mathbb{E} \left[ \frac{1}{2} (\sigma^2)^{-2} - X_i^2 (\sigma^2)^{-3} \right] = \frac{1}{2\theta^4} - \frac{1}{\theta^6} \mathbb{E}(X_i^2) = \frac{1}{2\theta^4} - \frac{\theta^2}{\theta^6} = -\frac{1}{2\theta^4}$$

$$\implies \text{Var}(\hat{\sigma}^2) \geq 1 / \left( -n \mathbb{E} \left[ \frac{\partial^2 \log(f(X; \theta))}{\partial \theta^2} \right] \right) = \frac{1}{n/(2\theta^4)} = \boxed{\frac{2\theta^4}{n}}$$

Therefore the variance of this estimator is equal to the Cramer-Rao lower bound.

Alternative solution (Math 541A):

$$\begin{aligned}
 I_X(\sigma) &= I_{(X_1, \dots, X_n)}(\sigma) = (\text{by Proposition 10.4.5 (Proposition 6.21 in Math 541A notes)}) nI_{X_1}(\sigma) \\
 &= n\text{Var}_\sigma \left( \frac{d}{d\sigma} \log \left( \frac{1}{\sigma\sqrt{2\pi}} \exp \left( -\frac{(X_1 - \mu)^2}{2\sigma^2} \right) \right) \right) \text{ (by definition of Fisher information)} \\
 &= n\text{Var}_\sigma \left[ -\frac{1}{\sigma} - \frac{d}{d\sigma} \left( \frac{(-X_1 - \mu)^2}{2\sigma^2} \right) \right] = n\sigma^{-6}\text{Var}_\sigma((X_1 - \mu)^2) \\
 &= n\sigma^{-6}(\mathbb{E}(X_1 - \mu)^4 - [\mathbb{E}(X_1 - \mu)^2]^2) = n\sigma^{-6}\sigma^4(3 - 1) = \frac{2n}{\sigma^2}.
 \end{aligned}$$

By Cramer-Rao,

$$\text{Var}_\sigma(z) \geq \frac{|g'(0)|^2}{I_X(0)}$$

Note that  $g(\sigma) = \mathbb{E}Y = \frac{n-1}{n}\sigma^2$  and that  $\mathbb{E} \sum_{j=1}^n (X_j - \bar{X})^2 = \sigma^2(n-1)$ . If  $Z$  is unbiased for  $\sigma^2$ ,

$$g'(\sigma) = \frac{2\sigma(n-1)}{n}|g'(\sigma)|^2 = \frac{4\sigma^2(n-1)^2}{n^2}$$

So,

$$\text{Var}_\sigma(z) \geq \frac{4\sigma^2(n-1)^2}{n^2 2n\sigma^{-2}} = \frac{2(n-1)^2\sigma^4}{n^3}.$$

Note that

$$\frac{n-1}{n}\sigma^2 \sum_{j=1}^n (X_j - \bar{X})^2 \sim \chi_{n-1}^2$$

⋮

Note that

$$\begin{aligned}
 \frac{(n-1)S_n^2}{\sigma^2} &\sim \chi_n^2 \implies \text{Var} \left( \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \right) = (n-1) \cdot 2 \\
 \implies \text{Var}(\hat{\sigma}^2) &= \text{Var}_\sigma \left( \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right) = \frac{1}{n^2} \sigma^4 \text{Var}_\sigma \left( \sigma^{-2} \sum_{i=1}^n (X_i - \bar{X})^2 \right) = \frac{2\sigma^4(n-1)}{n^2}
 \end{aligned}$$

e.

$$\begin{aligned} \log(L) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - (\sigma^2)^{-1} \frac{1}{2} \sum_{i=1}^n X_i^2 \\ \frac{\partial \log(L)}{\partial \mu} &= \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0 \iff n\mu = \sum_{i=1}^n x_i \iff \mu = \frac{1}{n} \sum_{i=1}^n x_i \\ &\implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \end{aligned}$$

Also note that

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \cdot n\mu = \mu.$$

□

**Example 10.22 (Example 6.46, Math 541A notes).** Alternative solution at <https://math.stackexchange.com/questions/49543/maximum-estimator-method-more-known-as-mle-of-a-uniform-distribution>

**Proposition 10.4.15 (Functional Equivariance of the MLE; Proposition 6.49 from Lecture Notes, Theorem 7.2.10 in Casella and Berger [2001],).** Let  $g : \Theta \rightarrow \Theta'$  be a bijection. Suppose  $Y$  is the MLE of  $\theta$ . Then  $g(Y)$  is the MLE of  $g(\theta)$ .

*Proof (case when  $g$  is invertible).* Note that if  $\ell(\theta)$  is the likelihood function for  $\theta$ , then the likelihood function for  $g(\theta)$  can be expressed as

$$\ell(g(\theta)) = \prod_{i=1}^n f_\theta(x_i | g^{-1}(g(\theta))) = \ell(g^{-1}(g(\theta))) = \ell(g^{-1}(\theta'))$$

where  $\theta' = g(\theta)$ . By definition of the MLE,  $Y = t(X_1, \dots, X_n)$  achieves the maximum value of  $\theta \mapsto \ell(\theta)$ . Therefore we can equivalently say  $g(Y) = g(t(X_1, \dots, X_n))$  achieves the maximum value of  $\theta' \mapsto \ell(g^{-1}(\theta'))$ .

(For a proof when  $g$  is not invertible, see Theorem 7.2.10 in Casella and Berger [2001][p.320].)

□

**Proposition 10.4.16 (Math 541A Homework Problem).** Let  $X_1, \dots, X_n$  be a random sample of size  $n$ , so that  $X_1$  has the Laplace density  $\frac{1}{2}e^{-|x-\theta|}$  for all  $x \in \mathbb{R}$ , where  $\theta \in \mathbb{R}$  is unknown. Then the MLE of  $\theta$  is the median.

*Proof.*

$$f(x_i; \theta) = \frac{1}{2}e^{-|x_i - \theta|}$$

$$\implies L = \prod_{i=1}^n \frac{1}{2}e^{-|x_i - \theta|} = 2^{-n} \exp\left(-\sum_{i=1}^n |x_i - \theta|\right)$$

$$\implies \log(L) = -n \log(2) - \sum_{i=1}^n |x_i - \theta| \implies \frac{d \log(L)}{d\theta} = - \sum_{i=1}^n \frac{d}{d\theta} |x_i - \theta| = - \sum_{i=1}^n \text{sgn}(x_i - \theta)$$

since  $\frac{d|x|}{dx} = \text{sgn}(x)$ . Next set this equal to 0 and solve:

$$- \sum_{i=1}^n \text{sgn}(x_i - \hat{\theta}_{MLE}) = 0$$

Notice that if  $n$  is even then the median set as  $\hat{\theta}_{MLE}$  satisfies the above equation. If  $n$  is odd, the median is still the best we can do. So the MLE is the median.

□

#### 10.4.7 Bayes estimator

#### 10.4.8 EM Algorithm

**Remark (Correction to Remark 6.57).** If  $Y$  is constant, the algorithm just outputs  $\theta_0$  in one step by the Likelihood Inequality (Lemma 6.50 in lecture notes):

$$\mathbb{E}_\theta \log \left( \frac{f_\theta(X)}{f_\omega(X)} \right) \geq 0 \iff \mathbb{E}_\theta \log f_\theta(X) - \mathbb{E}_\theta \log f_\omega(X) \geq 0$$

has equality only when  $\omega = \theta$  (if  $\mathbb{P}_\theta \neq \mathbb{P}_\omega \forall \theta \neq \omega$ ). So

$$\mathbb{E}_\theta \log f_\theta(X) \geq \mathbb{E}_\theta \log f_\omega(X).$$

**Remark (note on proof of Lemma 6.58).**

$$Y = t(X), \quad f_{X,Y}(x, y) = f_X(x) \mathbf{1}_{y=t(x)}.$$

#### 10.4.9 Comparison of estimators

### 10.5 Resampling and Bias Reduction

#### 10.5.1 Jackknife Resampling

$$Z_n := Y_n + (n-1) \left( Y_n - \frac{1}{n} \sum_{i=1}^n t_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \right)$$

### 10.5.2 Bootstrapping

## 10.6 Some Concentration of Measure

### 10.6.1 Concentration for Independent Sums

Can generate similar results for other random variables—just need different bound on moment-generating function (generally is fine as long as values of random variable are bounded between two real numbers, but more work to prove). However, doesn't work when values aren't bounded (e.g. for Gaussian random variables).

- What about other unbounded random variables?
- What about dependent random variables?

General question: **how far is a random variable from its mean?** Will first address functions of independent Gaussian random variables.

**Definition 10.28 (Lipschitz functions).** A real-valued function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is called **Lipschitz continuous** or  **$L$ -Liptschitz** if there exists a positive real constant  $L$  such that for all  $x_1, x_2 \in \mathbb{R}^n$ ,

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|_2.$$

**Theorem 10.6.1 (Theorem 8.5 in 541A notes).** Note from proof: use the fact that since  $f$  is 1-Lipschitz,  $\|\frac{df(X)}{dx_i}\|_2^2 \leq 1$ .

**Theorem 10.6.2 (Theorem 8.6 in 541A notes).** Notes from proof: change bounds on integral to  $8a/\pi$ . Then since  $u \geq 8a/\pi \iff -a \geq -\pi u/8$ , we have

$$\|\Pi(X)\| > u \implies \|\Pi(X)\| - a > u - a \geq u - \pi/8u > u/2.$$

Therefore

$$\Pr(\|\Pi(X)\| > u) = \Pr(\|\Pi(X)\| > u) \geq \Pr(|\|\Pi(X)\| - a| > u/2).$$

⋮

Loose bound:  $a^2 e^{-a^2} \leq 10$  for all  $a > 0$ . ( $k > 1$ ).

⋮

$$\mathbb{E}\|\Pi(X)\|^4 \leq 2^{12}a^4 + 10^6k^2$$

then by Jensen's Inequality,

$$a^4 = (\mathbb{E}\|\Pi(X)\|)^4 \leq (\mathbb{E}\|\Pi(X)\|^2)^2$$

So  $2^{12}a^4 \leq 2^{12}(\mathbb{E}\|\Pi(X)\|^2)^2$ . Then we chose  $10^{10}$  to make things easy and say

$$2^{12}a^4 + 10^6k^2 \leq 10^{10}(\mathbb{E}\|\Pi(X)\|^2)^2.$$

Summary:

$$Z := \|\Pi(X)\|^2, \quad \mathbb{E}(Z^2) = k, \quad \mathbb{E}(Z^4) \leq 10^{10}(\mathbb{E}(Z^2))^2.$$

⋮

Union bound (Boole's Inequality)

## 10.7 Math 541B

**Definition 10.29 (Statistical model).** A **statistical model** is a family of probability distributions  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ . Assume that  $X_1, \dots, X_n$  are i.i.d. from  $P_{\theta_0}, \theta_0 \in \Theta$ .

**Definition 10.30 (Identifiability).**

$$\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}.$$

**Definition 10.31 (Estimator).** An **estimator**  $\hat{\theta}(X_1, \dots, X_n)$  is any measurable function of the data  $(X_1, \dots, X_n)$ .

**Definition 10.32 (Bias).** The **bias** of an estimator is defined via

$$B(\hat{\theta}) := \mathbb{E}_{\theta_0} \hat{\theta}(X_1, \dots, X_n) - \theta_0.$$

**Example 10.23.** Alice wants to send a message to Bob. A message is a sequence of symbols from  $\{a, b, \dots, z\} = \mathcal{A}$ . **prefix-free codebook:**  $\mathcal{A}$ : sequence of 0, 1.

**Theorem 10.7.1 (Shannon's Theorem).** Assume that  $X$  has values in  $\mathcal{A}$  and has distribution  $D$ . To minimize  $\mathbb{E}_\rho$ (number of bits), the codebook has to use about  $\log_2(1/P(\beta))$  digits to encode  $\rho \in \mathcal{A}$ . Then

$$\mathbb{E}_\rho[\text{number of bits}] = \sum_{\beta \in \mathcal{A}} D(\beta) \log_2(1/p(\rho))$$

this is the entropy of  $D$ .

Assume that the codebook is created assuming that the true distribution of symbols is  $P$  but actually the true distribution is  $Q$ . In this case, the expected number of bits is

$$\mathbb{E}_Q[\text{number of bits}] = \sum_{\beta \in \mathcal{A}} Q(\beta) \log_2(1/p(\beta)).$$

we can also write this as the entropy of  $Q$  etc.

$$= \underbrace{\sum_{\beta \in \mathcal{A}} Q(\beta) \log_2(1/Q(\beta))}_{\text{entropy of } Q, \text{ Ent}(Q)} + \underbrace{\sum_{\beta} Q(\beta) \log(Q(\beta)/p(\beta))}_{KL(Q||P)}$$

so the roots of MLE are in information theory.

## 10.8 Hypothesis Testing

Why is the likelihood ratio useful for hypothesis testing? It is a sufficient statistic. Suppose we have  $H_0 : X \sim P$ ,  $H_1 : X \sim Q$ , and  $\theta = 1 \implies \mathbb{P}_\theta = Q$ ,  $\theta = 0 \implies \mathbb{P}_\theta = P$ . Then the likelihood function is

$$L(\theta | X) = q(x)^{I\{\theta=1\}} p(x)^{1-I\{\theta=1\}} = \left(\frac{q(x)}{p(x)}\right)^{I\{\theta=1\}} \cdot p(x),$$

so the likelihood ratio is a sufficient statistic by the Factorization Theorem.

### Bayes testing:

Prior probability  $\pi$  on  $H_0$  (that distribution is on  $P$ ),  $1 - \pi$  on  $H_a$  (distribution on  $Q$ ). The if  $\alpha_\phi := \mathbb{E}_P(\phi(X))$  is the probability of a type I error and  $1\beta_\phi := 1 - \mathbb{E}_Q(\phi(X))$  is the probability of a type II error, we want to minimize the Bayes risk:

$$R_\pi(\phi) := \pi\alpha_\phi + (1 - \pi)(1 - \beta_\phi).$$

The Bayes test is

$$\phi_\pi^*(X) = \begin{cases} 1 & L(X) < (1 - \pi)/\pi \\ \gamma & L(X) = (1 - \pi)/\pi \\ 0 & L(X) > (1 - \pi)/\pi \end{cases}$$

where  $L(X) = P(X)/Q(X)$  is the likelihood ratio.

**Mimimax:** minimize worst case:

$$\min \{ \max \{ \alpha_\phi, 1 - \beta_\phi \} \}.$$

**Neymann-Pearson:** acceptable Type I error rate  $\alpha$ . So we have the constraint  $\mathbb{E}_P(\phi(X)) \leq \alpha$ . Maximize power subject to this constraint.

All give roughly similar tests.

**Theorem 10.8.1.** Assume that for some  $\pi_* \in [0, 1]$  that the Bayes test  $\phi_{\pi_*}^*$  satisfies

$$\alpha_{\phi_{\pi_*}}^* = 1 - \beta_{\phi_{\pi_*}}^*. \quad (10.22)$$

Then  $\phi_{\pi_*}^*$  is minimax. Moreover, a  $\pi_*$  with this desired property always exists.

*Proof.* Assume that  $\phi_{\pi_*}^*$  is a Bayes test satisfying (10.22), but is not minimax. Then there must exist  $\tilde{\phi}$  such that

$$\min \left\{ \max \left\{ \alpha_{\tilde{\phi}}, 1 - \beta_{\tilde{\phi}} \right\} \right\} < \min \left\{ \max \left\{ \alpha_{\phi_{\pi_*}}^*, 1 - \beta_{\phi_{\pi_*}}^* \right\} \right\}$$

and note that

$$\max \left\{ \alpha_{\tilde{\phi}}, 1 - \beta_{\tilde{\phi}} \right\} < \max \left\{ \alpha_{\phi_{\pi_*}}^*, 1 - \beta_{\phi_{\pi_*}}^* \right\} = \phi_{\pi_*}^* \text{ (by (10.22))}$$

We have

$$R_{\pi_*}(\tilde{\phi}) = \pi_* \alpha_{\tilde{\phi}} + (1 - \pi_*)(1 - \beta_{\tilde{\phi}}) < \pi_* \alpha_{\phi_{\pi_*}}^* + (1 - \pi_*)(1 - \beta_{\phi_{\pi_*}}^*) = R_{\pi_*}(\phi_{\pi_*}^*)$$

But this contradicts the fact that the Bayes test minimizes this ...

To prove existence, consider for  $\pi \in [0, 1]$  the non-randomized test

$$\phi_\pi = \begin{cases} 1 & L(X) \leq (1 - \pi)/\pi \\ 0 & L(x) > (1 - \pi)/\pi \end{cases}$$

Since the range of  $\pi$  varies the Type I and Type II error rate, if we could show this function is continuous in  $\pi$  then we would be done by the Intermediate Value Theorem (the point where both error rates are equal would exist). We can accommodate the possibility of a jump discontinuity right at what would have been the best value by using a randomized test that takes an appropriate convex combination

□

**Theorem 10.8.2 (same theorem, re-stated on 09/13).**  $X : \Omega \rightarrow S$ .  $H_0 : X \sim P$ .  $X_a : X \sim Q$ . Assume  $\mu$  is a probability measure on  $\Omega$ . And  $P$  has density  $p$  with respect to  $\mu$ ,  $Q$  has density  $q$  with respect to  $\mu$ . Let  $\phi : S \rightarrow [0, 1]$  be a measurable function (randomized statistical test; the value it takes is interpreted as a probability). Let  $\alpha_\phi := \mathbb{E}_P \phi(X)$  be the probability of a Type I error, and let  $\beta_\phi := \mathbb{E}_Q \phi(X)$ .

Assume  $\pi_* \in [0, 1]$ . Then the Bayes optimal test  $\phi_{\pi_*}^*$  is such that  $\alpha_{\phi_{\pi_*}^*} = 1 - \beta_{\phi_{\pi_*}^*}$ . Then  $\phi_{\pi_*}^*$  is minimax optimal. Moreover,  $\pi_*$  with required properties holds.

*Proof of second claim.* We now show that  $\pi_*$  and  $\phi_{\pi_*}^*$  exist. Let  $\pi \in [0, 1]$ . Consider the non-randomized

$$\phi_\pi(x) = \begin{cases} 1 & \text{if } \frac{p(x)}{q(x)} \leq \frac{1-\pi}{\pi} \\ 0 & \text{if } \frac{p(x)}{q(x)} > \frac{1-\pi}{\pi} \end{cases}$$

Let

$$L(x) := \frac{p(x)}{q(x)}.$$

Consider

$$\begin{aligned} G(\pi) &:= \underbrace{Q(L(x) > (1-\pi)/\pi)}_{\mathbb{P} \text{ type II errors of } \pi_\pi} - \underbrace{P(L(x) \leq (1-\pi)/\pi)}_{\mathbb{P} \text{ type I error of } \phi_\pi} \\ &= Q(L(x) > (1-\pi)/\pi) + P(L(x) > (1-\pi)/\pi) - 1 \end{aligned}$$

Observe:

- (a)  $\lim_{\pi \rightarrow 0^+} G(\pi) = -1$ , since the probability of being over infinity is 0.
- (b)  $G(1) = Q(L(x) > 0) + P(L(x) > 0) - 1 = Q(L(x) > 0) \geq 0$ .
- (c)  $G$  is nondecreasing, continuous from the left (because a CDF is right continuous, but the arguments are decreasing as  $\pi$  increases, so instead it is left continuous.) So,  $G$  will either intersect the  $\pi$  axis or it will jump over it. That is, there exists some  $\pi_* \in (0, 1]$  such that either (i)  $G(\pi_*) = 0$  or (ii)  $G$  jumps over the  $\pi$  axis;  $G(\pi_*) < 0$  and  $\lim_{\pi \rightarrow \pi_*^+} G(\pi) > 0$ .

**Case (i).** In this case we are done.  $\phi_{\pi_*}^*$  is minimax optimal by the first part of the theorem.

**Case (ii).** Let  $\gamma := G(\pi_*^+)/G(\pi_*^+) - G(\pi_*) \in (0, 1)$ . Let

$$\phi'_{\pi_*} = \begin{cases} 1 & L(x) < (1-\pi_*)/\pi_* \\ \gamma & L(x) = (1-\pi_*)/\pi_* \\ 0 & L(x) > (1-\pi_*)/\pi_* \end{cases}$$

$\phi'_{\pi_*}$  is (randomized) Bayes optimal for the prior  $\pi_*$ . Note: The difference of type II and type II errors of  $\phi'_{\pi_*}$  is

$$\underbrace{Q(L(x) > (1-\pi_*)/\pi_*) + (1-\gamma)Q(L(x) = (1-\pi_*)/\pi_*)}_{\text{Type II error}} - \underbrace{P(L(x) < (1-\pi_*)/\pi_*) - \gamma P(L(x) = (1-\pi_*)/\pi_*)}_{\text{Type I error}}$$

Using  $P(L(x) < (1-\pi_*)/\pi_*) = 1 - P(L(x) > (1-\pi_*)/\pi_*) - P(L(x) = (1-\pi_*)/\pi_*)$ , we can write this as

$$= Q(L(x) > (1-\pi_*)/\pi_*) + P(L(x) > (1-\pi_*)/\pi_*) - 1 + (1-\gamma) [Q(L(x) = (1-\pi_*)/\pi_*) + P(L(x) = (1-\pi_*)/\pi_*)] \quad (10.23)$$

For the  $\gamma$  we defined, (10.23) is 0. Why?

$$1 - \gamma = -G(\pi_*) / (G(\pi_*^+) - G(\pi_*))$$

Recall that a jump in a cdf occurs when the corresponding random variable takes on a fixed value with nonzero probability. Compute

$$G(\pi_*^+) - G(\pi_*) = Q(L(x) = (1-\pi)/\pi) - P(L(x) = (1-\pi)/\pi).$$

so we see everything cancels;

$$-(1-\gamma) [Q(L(x) = (1-\pi_*)/\pi_*) + P(L(x) = (1-\pi_*)/\pi_*)] = Q(L(x) > (1-\pi_*)/\pi_*) + P(L(x) > (1-\pi_*)/\pi_*) - 1$$

and (10.23) equals 0. That is, the difference is zero, so  $\phi'_{\pi_*}$  is minimax optimal by the first part of the theorem.

□

### 10.8.1 Neyman-Pearson Tests

**Definition 10.33 (Power function; definition 8.3.1 in Casella and Berger [2001]).** The **power function** of a hypothesis test with rejection region  $R$  is the function of  $\theta$  defined by

$$\beta(\theta) := \mathbb{P}_\theta(x \in R).$$

Motivation: some errors are worse than others. Goal: control Type I error, keeping it small for any sample size.

Define the **test function**  $\phi(x)$  to be an indicator of  $x$  being in the rejection region for a test:

$$\phi(x) = \begin{cases} 1 & x \in R \\ 0 & x \notin R \end{cases}.$$

Let  $\Phi$  be a set of such tests.

**Definition 10.34 (UMP; Definition 8.3.11 in Casella and Berger [2001]).**  $\phi^*$  is **uniformly most powerful** (UMP) of size  $\alpha \in [0, 1]$  if it achieves the maximum value of  $B_\phi$  subject to  $\alpha_\phi \leq \alpha$  over all randomized tests  $\phi \in \Phi$ . That is, it is uniformly most powerful if and only if it solves

$$\beta_\phi = \mathbb{E}_Q \phi(x) \rightarrow \max_{\phi \in \Phi} \text{ s.t. } \alpha_\phi = \mathbb{E}_p \phi(x) \leq \alpha.$$

or

**Lemma 10.8.3 (Neyman-Pearson; Theorem 8.3.12, p. 388 in Casella and Berger [2001]).** Consider testing  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta = \theta_1$ , where the pdf or pmf corresponding to  $\theta_i$  is  $f(x | \theta_i)$ ,  $i \in \{0, 1\}$ . Suppose we have a test  $\phi^*$  such that

$$\phi^* = \begin{cases} 1, & f(x | \theta_1) > cf(x | \theta_0) \quad (q(x) > cp(x)) \\ 0, & f(x | \theta_1) < cf(x | \theta_0) \quad (q(x) < cp(x)) \end{cases} \quad (10.24)$$

(recall that  $\phi^*(x) = 1$  when  $x \in R$  and 0 otherwise) for some  $c \geq 0$  and

$$\alpha_{\phi^*} = \mathbb{P}_{\theta_0}(x \in R) = \beta^*(\theta_0) = \alpha. \quad (10.25)$$

(where  $\beta^*$  is the power test for  $\phi^*$ ). Then

1. Any test  $\phi^*$  that satisfies (10.24) and (10.25) is a UMP level  $\alpha$  test.
2. There exist  $c \geq 0$  and  $\phi^*$  that satisfy (10.24) and (10.25).
3. If  $\tilde{\phi}$  is UMP of size  $\alpha$ , then it satisfies (10.24) almost surely. Moreover,  $\alpha_{\tilde{\phi}} = \alpha$  unless  $p_{\tilde{\phi}} = 1$ . (In other words, the only complication/exception is if you have a test that has power 1 and Type I error rate less than  $\alpha$ ; then obviously you choose the test with power 1 and minimal Type I error rate, which will be less than  $\alpha$ .)

(Claim from Casella and Berger [2001]) If there exists a test satisfying (10.24) and (10.25), then every UMP level  $\alpha$  test is a size  $\alpha$  test (satisfies (10.25)) and every UMP level  $\alpha$  test satisfies (10.24) except perhaps on a set  $A$  satisfying  $\mathbb{P}_{\theta_0}(X \in A) = \mathbb{P}_{\theta_1}(x \in A) = 0$ .

*Proof (in-class, 541B).* 1. Suppose that  $\phi$  is the test function for another level  $\alpha$  test; that is, if  $\beta$  is the power function for  $\phi$ ,

$$\sup_{\theta \in \{\theta_0\}} \beta(\theta) = \beta(\theta_0) \leq \alpha. \quad (10.26)$$

Note that

$$\int_S (\phi^* - \phi) [f(x | \theta_1) - cf(x | \theta_0)] d\mu \geq 0 \quad (10.27)$$

for the following reason: by (10.24), when  $x \in R$ ,  $\phi^* = 1$ , so  $\phi^* - \phi \in [0, 1]$ . Also, in this region,  $f(x | \theta_1) > cf(x | \theta_0)$ , so the second term is positive. On the other hand, when  $x \in R^C$ ,  $\phi^* = 0$ , so  $\phi^* - \phi \in [-1, 0]$ , and  $f(x | \theta_1) < cf(x | \theta_0)$  so the second term is negative.

On the other hand, we can rewrite (10.27) as follows:

$$0 \leq \int_S \phi^* f(x | \theta_1) d\mu - \int_S \phi f(x | \theta_1) d\mu - c \int_S \phi^* f(x | \theta_0) d\mu + c \int_S \phi f(x | \theta_0) d\mu$$

$$\iff 0 \leq \beta^*(\theta_1) - \beta(\theta_1) - c[\beta^*(\theta_0) - \beta(\theta_0)] \quad (10.28)$$

(where  $\beta^*$  is the power function for  $\phi^*$  and  $\beta$  is the power function for  $\phi$ ) (since  $\phi^*$  and  $\phi$  are indicator functions for being in the rejection region). Note that  $c[\beta^*(\theta_0) - \beta(\theta_0)] \geq 0$  since  $c \geq 0$  by assumption,  $\beta^*(\theta_0) = \alpha$  by (10.25), and  $\beta(\theta_0) \leq \alpha$  by (10.26). Therefore by (10.28),

$$0 \leq \beta^*(\theta_1) - \beta(\theta_1) - c[\beta^*(\theta_0) - \beta(\theta_0)] \leq \beta^*(\theta_1) - \beta(\theta_1);$$

that is,  $\beta^*$  is UMP over all level  $\alpha$  tests.

(Notes from class:)

$$= \mathbb{E}_Q \phi^* - \mathbb{E}_Q \phi - c(\mathbb{E}_P \phi^* - \mathbb{E}_P \phi) =$$

. Since  $\mathbb{E}_P \phi^* = \alpha$  and  $\mathbb{E}_P \phi \leq \alpha$ ,  $\mathbb{E}_P \phi^* - \mathbb{E}_P \phi \geq 0$ . But  $\mathbb{E}_Q \phi^* - \mathbb{E}_Q \phi \geq c(\mathbb{E}_P \phi^* - \mathbb{E}_P \phi)$  (**why?**), so

$$\mathbb{E}_Q \phi^* = \beta_{\phi^*} \geq \mathbb{E}_Q \phi = \beta_\phi$$

So the power of  $\phi^*$  is greater than or equal to the power of  $\phi$ .

2. We will use the randomization possible when  $q(x) = cp(x)$  to make the Type I error rate equal  $\alpha$  exactly. Consider the test

$$\phi^*(x) = \begin{cases} 1, & q(x) > cp(x) \\ \gamma, & q(x) = cp(x) \\ 0, & q(x) < cp(x) \end{cases} \quad (10.29)$$

for some  $\gamma \in [0, 1]$  and some  $c \geq 0$ . The size of  $\phi^*$  is

$$\alpha_{\phi^*} = \mathbb{P}_P(x : q(x) > cp(x)) + \gamma \mathbb{P}(x : q(x) = cp(x))$$

$$= \int_S \phi^*(x)p(x) d\mu$$

so it remains to show that for any  $\alpha \in [0, 1]$ , there exists  $c, \gamma$  such that  $\alpha_{\phi^*} = \alpha$ . Let  $H(c) := \mathbb{P}_P(x : q(x) \leq cp(x))$ . Then

- (a)  $H(c)$  is non-decreasing.
- (b)  $H(c)$  is right-continuous (like a cdf, because of the  $\leq$ ).

Then  $\alpha_{\phi^*} = (1 - H(c)) + \gamma[H(c) - \lim_{y \rightarrow c^+} H(y)]$ , and we want this to equal  $\alpha$ . (Notice that this equation is very similar to the ones we encountered when studying minimax tests.) We will consider two cases.

- (a) If there exists a  $c$  such that  $1 - H(c) = \alpha$ , simply take  $\gamma = 0$ .
- (b) If there does not exist a  $c$  such that  $1 - H(c) = \alpha$ ; that is,  $1 - H(c) < \alpha$ ,  $\lim_{y \rightarrow c^+} 1 - H(y) > \alpha$ , we can set  $\gamma$  appropriately. (Exercise: find the  $\gamma$ .)

More notes:

$$H(c) = \mathbb{P}(x : q(x) \leq cp(x)) = \mathbb{P}(x : q(x)/p(x) \leq c).$$

We can divide by  $p(x)$  because the probability that  $p(x) = 0$  is 0. Now  $H$  is kind of like a distribution function. Consider:

$c < 0 \implies H(c) = 0$ . Also, as  $c \rightarrow \infty$ ,  $H \rightarrow 1$ . So we can always solve for  $c \geq 0$ ,  $\gamma \in [0, 1]$  such that

$$1 - H(c) + \gamma[H(c) - H(c^-)] = \alpha.$$

(Not a difficult exercise, but requires careful attention to detail.)

3. Consider again (10.27), except now let  $\phi$  be any UMP level  $\alpha$  test. Since by part (1)  $\phi^*$  is also UMP level  $\alpha$ , we have  $\beta^*(\theta_1) = \beta(\theta_1)$ , so from (10.28) we have

$$0 \leq \beta^*(\theta_1) - \beta(\theta_1) - c[\beta^*(\theta_0) - \beta(\theta_0)] = c[\beta(\theta_0) - \beta^*(\theta_0)] \implies \beta(\theta_0) - \beta^*(\theta_0) \geq 0$$

Recall that  $\beta^*(\theta_0) = \alpha$  by (10.25), so  $\beta(\theta_0) \geq \alpha$ . But  $\beta(\theta_0) \leq \alpha$  since it is UMP level  $\alpha$ . Therefore we must have  $\beta(\theta_0) = \alpha$ , so  $\phi$  satisfies (10.25) (unless there is a set  $A$  such that  $\int_S f(x | \theta_1) dx = \int_S f(x | \theta_0) dx = 0$ ).

(Notes from class) Consider

$$A = \int_S (\phi^* - \tilde{\phi}) \underbrace{(q - cp)}_{\text{positive on } c_+} d\mu \quad (10.30)$$

where  $\phi^*$  is the UMP test (10.29) from part (2). Then (10.30) is greater than or equal to 0 ( $A \geq 0$ ) by part 1 of the lemma. On the other hand,

$$A = \underbrace{\beta_{\phi^*} - \beta_{\tilde{\phi}}}_{=0 \text{ if they are both UMP}} - \underbrace{c}_{\geq 0} \begin{pmatrix} \alpha_{\phi^*} - \alpha_{\tilde{\phi}} \\ =\alpha \quad \leq\alpha \end{pmatrix}.$$

They must both be UMP tests, so the above notes follow. So either  $c = 0$  or  $\alpha_{\phi^*} = \alpha_{\tilde{\phi}}$ . If  $c = 0$ ,

$$\beta_{\phi^*} = \int \phi^*(x)q(x) d\mu = 1$$

$$\tilde{\phi} = \begin{cases} 1, & q(x) > 0 \\ 0, & q(x) \leq 0 \end{cases}$$

(but of course  $q(x)$  is never negative). Then  $\beta_{\tilde{\phi}} = \int \tilde{\phi}(x)q(x) d\mu = 1$ .

Note that we don't need to know the  $\gamma$ , because

The other possibility is  $\alpha_{\phi^*} = \alpha_{\tilde{\phi}} = \alpha$ . Consider the set  $c_+ = \{x : q(x) > cp(x)\} \implies \tilde{\phi}(x) = 1$  on  $c_+$ . And  $c_- = \{x : q(x) < cp(x)\} \implies \tilde{\phi}(x) = 0$  on  $c_-$ . Finally, since  $\alpha_{\tilde{\phi}} = \alpha_{\phi^*} = \alpha$ ,  $\tilde{\phi}(x) = \gamma$  when  $q(x) = cp(x)$ , then  $\gamma$  is the same for  $\phi^*$ , so the tests must be equal.

□

Geometric illustration (Figure 2): consider the mapping  $T : \Phi \rightarrow [0, 1] \times [0, 1]$  where  $\Phi$  is the set of all possible tests where  $T(\phi) = (\alpha_\phi, \beta_\phi)$ . Observation:  $(\alpha, \beta) \in \text{im}(T) \implies (1 - \alpha, 1 - \beta) \in \text{im}(T)$  because if  $T(\phi) = (\alpha, \beta)$ , then  $T(1 - \phi) = (1 - \alpha, 1 - \beta)$ . So we have symmetry around the point  $(1/2, 1/2)$ . This picture shows why there is an exception if  $\beta = 1$ ; then you want the test with minimum  $\alpha$  given maximum power. But in practical terms this case is not really important; this is more of a technical detail.

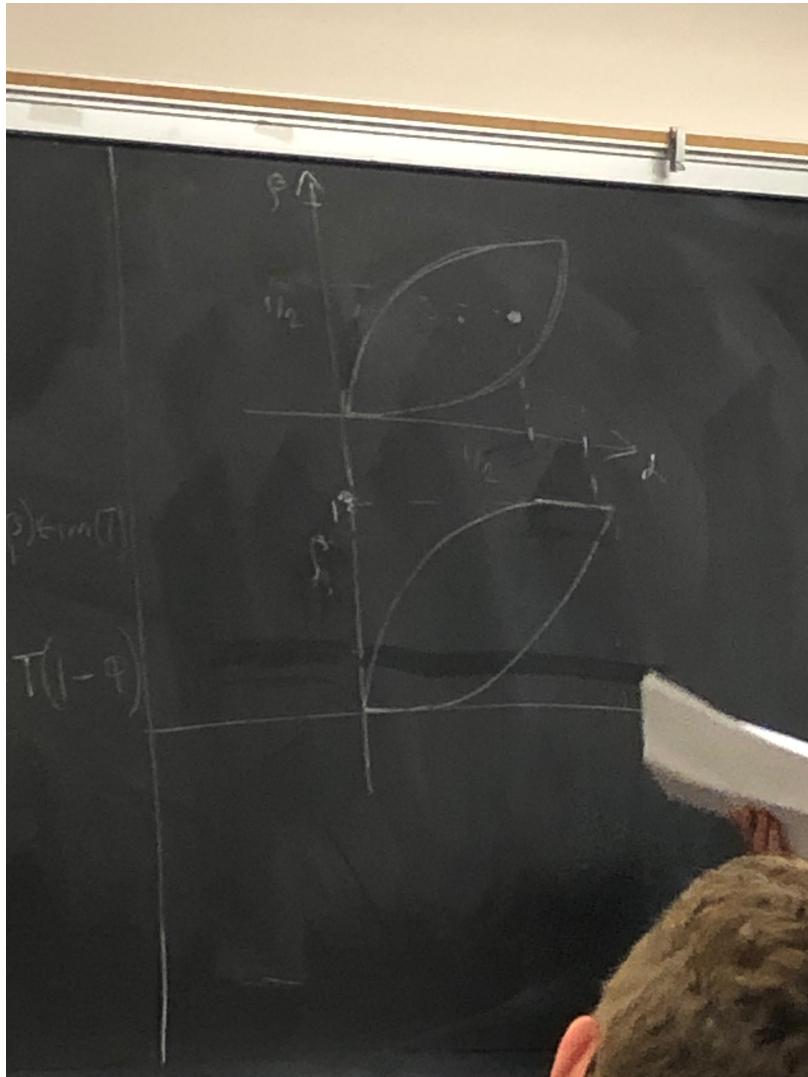


Figure 2: Sets of possible Neyman-Pearson tests; similar to Figure 3.1 in Lehmann and Romano.

**Exercise 21.** 1. Let  $\phi^*$  be the MP test of size  $\alpha$ . Then  $\beta_{\phi^*} \geq \alpha$ . Moreover,  $\beta_{\phi^*}$  is strictly greater than  $\alpha$  unless  $P = Q$  (Use N-P lemma, part 1).

2. Let  $X_1, \dots, X_n$  be i.i.d. uniform.  $H_0 : X_1, \dots, X_n \sim U(0, 1)$ .  $H_a : X_1, \dots, X_n \sim U(1/3, 2/3)$ . Derive a UMP test of size  $\alpha$  for all values of  $\alpha$ .

**Solution.**

1.

2.

$$q(x_1, \dots, x_n) = e^n \prod_{i=1}^n I\{x_i \in [1/3, 2/3]\} = e^n I\{x_1, \dots, x_n \in [1/3, 2/3]\}$$

Similarly,

$$p(x_1, \dots, x_n) = I\{x_1, \dots, x_n \in [0, 1]\}.$$

Note that the ratio  $q/p$  only takes on three values. By the Neyman-Pearson Lemma, the most powerful test  $\phi^*$  is either

$$\phi_1^*(x_1, \dots, x_n) = \begin{cases} 1, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 3^n \\ \gamma, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 0 \end{cases}$$

or

$$\phi_2^*(x_1, \dots, x_n) = \begin{cases} \gamma, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 3^n \\ 0, & q(x_1, \dots, x_n)/p(x_1, \dots, x_n) = 0 \end{cases}$$

Now

$$\begin{aligned} \alpha_{\phi_1^*} &= \mathbb{E}_P \phi_1^*(X_1, \dots, X_n) = 1 \cdot \mathbb{P}(X_1, \dots, X_n \in [1/3, 2/3]) + \gamma \mathbb{P}(\text{at least one } X_j \text{ is outside } [1/3, 2/3]) \\ &= 3^{-n} + \gamma(1 - 3^{-n}) = \alpha \end{aligned}$$

which has a solution for  $\gamma(\alpha)$  for  $\alpha \geq 3^{-n}$ . On the other hand,

$$\begin{aligned} \alpha_{\phi_2^*} &= \mathbb{E}_P \phi_2^*(X_1, \dots, X_n) = \gamma \mathbb{P}(X_1, \dots, X_n \in [1/3, 2/3]) + 0 \cdot \mathbb{P}(\text{at least one } X_j \text{ is outside } [1/3, 2/3]) \\ &= \gamma 3^{-n} = \alpha \end{aligned}$$

which has a unique solution  $\gamma(\alpha)$  for  $\alpha \leq 3^{-n}$ . So for a test with a very small  $\alpha$ , you need to consider a strictly randomized test like  $\alpha_{\phi_2^*}$ .

**Theorem 10.8.4 (Went over last time).** We have a sequence  $X_1, \dots, X_n$  distributed i.i.d. Last time we showed that the following hypothesis testing problem:

$$H_0 : X_1, \dots, X_n \sim P, \quad H_A : X_1, \dots, X_n \sim Q$$

with

$$\frac{dP}{d\mu} = p, \quad \frac{dQ}{d\mu} = q$$

using the test

$$\phi^*(x_1, \dots, x_n) = \begin{cases} 1, & \prod_{i=1}^n q(x_i) > c \prod_{i=1}^n p(x_i) \\ \gamma, & \prod_{i=1}^n q(x_i) = c \prod_{i=1}^n p(x_i) \\ 0, & \prod_{i=1}^n q(x_i) < c \prod_{i=1}^n p(x_i) \end{cases} \quad (10.31)$$

is UMP of size  $\alpha = \mathbb{E}_P(\phi^*)(X_1, \dots, X_n)$  (given the correct choice of  $\gamma$ ).

**Remark.** Intuition: if the likelihood function for  $Q$  exceeds the likelihood function of  $P$  by a reasonable amount (determined by  $c$ ), choose  $Q$  to be more likely than  $P$ .

Also,  $\phi^*$  of this kind is referred to as a *Neyman-Pearson test*.

### 10.8.2 Consistency of Neyman-Pearson Tests

We want the test to be consistent—as we collect more data and  $n \rightarrow \infty$ , errors of both types go to 0. We hope to show the consistency of Neyman-Pearson tests.

**Definition 10.35 (Consistency of a hypothesis test).** Consider the framework of the Neyman-Pearson lemma (described above). A sequence of tests  $\{\phi_n\}_{n \geq 1}$ , where  $\phi_n = \phi_n(x_1, \dots, x_n)$  is *consistent* if and only if

$$\alpha_{\phi_n} = \mathbb{E}_P \phi_n(X_1, \dots, X_n) \rightarrow 0, \quad \text{and } \beta_{\phi_n} = \mathbb{E}_Q \phi_n(X_1, \dots, X_n) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

It turns out that  $c$  grows with  $n$ , and the growth condition of the size of this  $c_n$  is related to a measure of distance between  $P$  and  $Q$  (Hellinger distance).

Recall the following definition:

**Definition 10.36 (Hellinger Distance and Affinity (see Section 13.1 of Lehmann and Romano)).** The *Hellinger distance* between probability laws  $P$  and  $Q$  is defined as

$$H^2(P, Q) := \int_S (\sqrt{p} - \sqrt{q})^2 d\mu = 2 \int_S (1 - \sqrt{pq}) d\mu = 2 - 2 \int_{\mathbb{R}} \sqrt{pq} d\mu.$$

Further,

$$A(P, Q) := \int \sqrt{pq} d\mu$$

is known as the *Hellinger affinity*.

**Remark.** This is a true distance—it satisfies the Triangle Inequality,  $H(P, Q) = 0$  if and only if  $P = Q$  almost everywhere.

**Lemma 10.8.5.** Let  $X \sim P$  and  $Y \sim Q$ . Let  $X_1, \dots, X_n$  be i.i.d. copies of  $X$  and likewise for  $Y_1, \dots, Y_n$ , so  $(X_1, \dots, X_n) \sim P^n$  and  $(Y_1, \dots, Y_n) \sim Q^n$ . Then

$$A(P^n, Q^n) = [A(P, Q)]^n.$$

*Proof.*

$$A(P^n, Q^n) = \int \sqrt{p(x_1, \dots, x_n)} \sqrt{q(x_1, \dots, x_n)} d\mu^n = \prod_{j=1}^n \int \sqrt{p(x_j)q(x_j)} d\mu = [A(P, Q)]^n.$$

□

**Theorem 10.8.6 (Consistency of Neyman-Pearson tests).** Assume that  $\rho := A(P, Q) < 1$ . Then take any sequence  $\{c_n\} \subset \mathbb{R}_+$  such that

$$\rho^{2n} \ll c_n \ll \rho^{-2n}. \quad (10.32)$$

Then any sequence  $\{\phi_n\}_{n \geq 1}$  of Neyman-Pearson tests corresponding to  $\{c_n\}_{n \geq 1}$  is consistent.

*Proof.* First, we will show that  $\alpha_{\phi_n} \rightarrow 0$  under these conditions. Recall the definition of  $\phi$  from (10.31).

$$\begin{aligned} \alpha_{\phi_n} &= \int \phi(x_1, \dots, x_n) p(x_1, \dots, x_n) d\mu^n \\ &\leq \int I \left\{ \frac{\prod_{i=1}^n q(x_i)}{\prod_{i=1}^n p(x_i)} \geq c_n \right\} p(x_1) \cdots p(x_n) d\mu^n \\ &\leq \int \sqrt{\frac{q(x_1) \cdots q(x_n)}{c_n p(x_1) \cdots p(x_n)}} p(x_1) \cdots p(x_n) d\mu^n \\ &= \frac{1}{\sqrt{c_n}} \int \sqrt{q(x_1) \cdots q(x_n)} \sqrt{p(x_1) \cdots p(x_n)} d\mu^n = \rho^n c_n^{-1/2} \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

(The second inequality follows because when the indicator equals 0 the square root term is greater than or equal to 0 but when the indicator equals 1 the square root term is greater than or equal to 1. The last step follows by assumption of the lemma.)

Now consider the Type II error  $1 - \beta_{\phi_n}$ .

$$\begin{aligned} 1 - \beta_{\phi_n} &= \int [1 - \phi(x_1, \dots, x_n)] q(x_1) \cdots q(x_n) d\mu^n \\ &\leq \int I \{q(x_1) \cdots q(x_n) \leq c p(x_1) \cdots p(x_n)\} q(x_1) \cdots q(x_n) d\mu^n \\ &\leq \int \sqrt{\frac{p(x_1) \cdots p(x_n)}{q(x_1) \cdots q(x_n)}} \cdot c_n \cdot q(x_1) \cdots q(x_n) d\mu^n \\ &= \sqrt{c_n} A(P^n, Q^n) = c - n^{1/2} \rho^n \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

□

**Remark.** Intuition:  $A(P, Q) < 1 \iff H(P, Q) > 0 \iff P \neq Q$ . Also,  $a_n \ll b_n \iff a_n = o(b_n)$ , and  $a_n \gg b_n \iff b_n \ll a_n$ . Almost all reasonable sequences satisfy (10.32)—this range is very large.

Basically, this theorem says the Type I and Type II errors converge to 0 geometrically fast (as  $\rho^n$ ) if  $P$  and  $Q$  are not too close together.

### 10.8.3 Composite Hypothesis Testing

We will start with one-dimensional family (one parameter), then expand to multiple parameters using tricks based on conditioning and using the principles of sufficiency and completeness.

**Definition 10.37 (Composite Hypothesis Test).** Assume we have a statistical model  $\{P_\theta, \theta \in \Theta\}$ . Let  $\Theta_0, \Theta_1 \subset \Theta$  be such that  $\Theta_0 \cap \Theta_1 = \emptyset$  and  $\Theta_0 \cup \Theta_1 = \Theta$ . Assume  $X \sim P_\theta$  for some  $\theta \in \Theta$  (where  $X$  may be multivariate). We would like to test the hypotheses

$$H_0 : \theta \in \Theta_0, \quad H_a : \theta \in \Theta_1.$$

Such a test is a *composite hypothesis test*.

The Neyman-Pearson Lemma does not directly apply to composite hypothesis tests, but it turns out we can use the Neyman-Pearson lemma to prove some results about some composite hypothesis tests under some circumstances.

**Definition 10.38 (Power function; definition 8.3.1 in Casella and Berger).** The *power function* of a test  $\phi$  is defined as

$$\beta_\phi(\theta) := \mathbb{E}_\theta \phi(X).$$

**Remark.** For  $\theta \in \Theta_0$ ,  $\beta_\phi(\theta)$  is the probability of a Type 1 error; for  $\theta \in \Theta_1$ ,  $\beta_\phi(\theta)$  is the power of  $\phi$ .

**Definition 10.39 (Test size; definition 8.3.6 in Casella and Berger).** The test  $\phi$  is of *size*  $\alpha$  if and only if

$$\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha.$$

**Definition 10.40 (Uniformly most powerful test; definition 8.3.11 in Casella and Berger).**  $\phi^*$  is the *uniformly most powerful test* of size  $\alpha$  if  $\phi^*$  is of size  $\alpha$  and  $\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$  for any other test  $\phi$  of size  $\alpha$  for all  $\theta \in \Theta_1$ .

We will be interested in families with *monotone likelihood ratios*. (Note: for the rest of the class today,  $\Theta = [a, b] \subseteq \mathbb{R}$  with  $a, b \in [-\infty, \infty]$ ).

**Definition 10.41 (Monotone likelihood ratio; definition 8.3.16 in Casella and Berger).** Let  $T(X)$  be a statistic. Consider  $\Theta \subseteq \mathbb{R}$  as a segment on the real line. We say that the family of pdfs  $\{p_\theta, \theta \in \Theta\}$  has a *monotone likelihood ratio (MLR)* with respect to  $T(x)$  if and only if for any  $\theta' > \theta''$ ,

$$\frac{p_{\theta'}(x)}{p_{\theta''}(x)} = \psi_{\theta', \theta''}(T(x))$$

where  $\psi_{\theta', \theta''}(\cdot)$  is non-decreasing. (That it is non-decreasing is without loss of generality; in the case that  $\psi$  is non-increasing, simply replace  $T(x)$  with  $-T(x)$  to achieve a non-decreasing  $\psi$ ).

**Example 10.24.** Let  $X \sim \text{Poisson}(\lambda)$ , so

$$p_\lambda(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x \in \mathbb{Z}_+.$$

Then  $\{\text{Poisson}(\lambda), \lambda \in \mathbb{R}_+\}$  has a monotone likelihood ratio with respect to  $T(x) = x$  by the following argument. Take  $\lambda' > \lambda''$ . Then

$$\frac{e^{-\lambda'} (\lambda')^x x!}{x! e^{-\lambda''} (\lambda'')^x} = e^{-(\lambda' - \lambda'')} \left( \frac{\lambda'}{\lambda''} \right)^x$$

and

$$\psi_{\lambda', \lambda''}(x) = e^{-(\lambda' - \lambda'')} \left( \frac{\lambda'}{\lambda''} \right)^x$$

is non-decreasing.

**Theorem 10.8.7 (Karlin-Rubin; similar to Theorem 8.3.17 in Casella and Berger, Theorem 3.4.1 in Lehmann and Romano).** Assume  $\{p_\theta, \theta \in [a, b]\}$  has a monotone likelihood ratio with respect to some  $T(x)$ . Suppose we want to test

$$H_0 : \theta \leq \theta_0, \quad H_a : \theta > \theta_0.$$

Then there exists  $c \geq 0$  and  $\gamma \in [0, 1]$  such that

$$\phi^*(x) = \begin{cases} 1, & T(x) > c \\ \gamma, & T(x) = c \\ 0, & T(x) < c \end{cases}$$

is the uniformly most powerful test of size  $\alpha$ . (In other words, the UMP test will always be of this form.)

*Proof.* We will show that this is the uniformly most powerful test over the larger family of tests that satisfy  $\mathbb{E}_{\theta_0} \phi(x) = \alpha$  ( $\beta_\phi(\theta_0) \leq \alpha$ ), which contains tests of size  $\alpha$ . So if we can find the UMP test for this larger family (and it is a test of size  $\alpha$ ), then that is also the UMP test for test of size  $\alpha$  (satisfying  $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha$ ).

First note that we can always find  $c$  and  $\gamma$  such that

$$\mathbb{E}_{\theta_0}(\phi^*(x)) = \alpha.$$

Define  $F(c) := \mathbb{P}_{\theta_0}(\{x : T(x) \leq c\})$ . Note that

$$\mathbb{E}_{\theta_0} \psi^*(x) = F(c) + \gamma[F(c) - F(c^+)]$$

where  $F(c) - F(c^+)$  is the size of the jump. (The proof of this claim is an exercise.)

Next, take  $\theta' > \theta_0$  and consider the simple hypothesis test

$$H'_0 : \theta = \theta_0, \quad H'_a : \theta = \theta'.$$

By the Neyman-Pearson Lemma, the most powerful test for this is

$$\phi'(x) = \begin{cases} 1, & p_{\theta'}(x) > c' p_{\theta_0}(x) \\ \gamma', & p_{\theta'}(x) = c' p_{\theta_0}(x) \\ 0, & p_{\theta'}(x) < c' p_{\theta_0}(x) \end{cases} \iff \phi'(x) = \begin{cases} 1, & p_{\theta'}(x)/p_{\theta_0}(x) > c' \\ \gamma', & p_{\theta'}(x)/p_{\theta_0}(x) = c' \\ 0, & p_{\theta'}(x)/p_{\theta_0}(x) < c' \end{cases}$$

But since  $p_{\theta'}(x)/p_{\theta_0}(x) = \psi(T(x))$ , by the assumption of the monotone likelihood ratio property,

$$\frac{p_{\theta'}(x)}{p_{\theta_0}(x)} > c' \iff T(x) > \psi^{-1}(c').$$

Set  $\psi^{-1}(c') = c$  and  $\psi^{-1}(\gamma') = \gamma$  (using the values of  $\gamma$  and  $c$  from earlier in the proof). Then we have

$$\phi'(x) = \begin{cases} 1, & T(x) > c \\ \gamma', & T(x) = c \\ 0, & T(x) < c \end{cases}$$

as desired. (Moreover,  $c$  and  $\gamma$  are uniquely determined by  $\mathbb{E}_{\theta_0}(\phi^*) = \alpha$ . Note that  $\theta'$  does not appear anywhere in this test. Since  $\theta'$  was arbitrary, the proof is almost complete.

The final argument is to show that this test has size  $\alpha$ . Namely, we must show

$$\sup_{\theta \leq \theta_0} \beta_{\phi^*}(\theta) = \sup_{\theta \leq \theta_0} \mathbb{E}_\theta \phi^*(x) \leq \alpha.$$

It is sufficient to show that  $\beta_{\phi^*}(\theta)$  is monotone. Take  $\theta_1 < \theta_2$ . Then by the Neyman-Pearson Lemma, we know that  $\phi^*$  is the UMP test for testing

$$H''_0 : \theta = \theta_1 \quad H''_a : \theta = \theta_2$$

of size  $\beta_{\phi^*}(\theta)$ . Since it is the most powerful test, its power  $\mathbb{E}_{\theta_2} \phi^*(x)$  is at least as powerful than the test which is identically equal to  $\beta_{\phi^*}(\theta)$ ; that is,  $\mathbb{E}_{\theta_2}(\phi^*(x)) \geq \mathbb{E}_{\theta_1}(\phi^*(x))$ .

□

**Example 10.25.**  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ . Test  $H_0 : \mu \leq 0$  against  $H_1 : \mu > 0$ . Find the UMP test.

**Solution.**

Take  $\mu_1 > \mu_2$ . Then

$$\begin{aligned} \frac{p_{\mu_1}(x_1, \dots, x_n)}{p_{\mu_2}(x_1, \dots, x_n)} &= \exp \left\{ -\frac{1}{2} \left[ \sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_2)^2 \right] \right\} \\ &= \exp \left\{ -\frac{1}{2} \left[ n(\mu_1^2 - \mu_2^2) + (\mu_1 - \mu_2) \sum_{i=1}^n x_i \right] \right\} \end{aligned}$$

This is an increasing function with respect to  $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ . Therefore by Theorem 10.8.7, we have that the UMP test is

$$\phi(x_1, \dots, x_n) = \begin{cases} 1, & \sum_{i=1}^n x_i \geq c_\alpha \\ 0, & \sum_{i=1}^n x_i < c_\alpha \end{cases} = \begin{cases} 1, & n^{-1/2} \sum_{i=1}^n x_i \geq c'_\alpha \\ 0, & n^{-1/2} \sum_{i=1}^n x_i < c'_\alpha \end{cases} \quad (10.33)$$

Note that

$$\mathbb{E}_{\mu=0} \phi(X_1, \dots, X_n) = \alpha \iff \mathbb{P} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \geq c'_\alpha \right) = \alpha$$

and  $n^{-1/2} \sum_{i=1}^n X_i \sim \mathcal{N}(0, 1)$ . Therefore  $c'_\alpha = z_{1-\alpha}$  (the  $1 - \alpha$  quantile of a standard Gaussian distribution). Substituting this into (10.33) defines the UMP test.

**Remark.** The Karlin-Rubin Theorem (Theorem 10.8.7) still applies if the inequalities in the hypothesis test are reversed.

**Example 10.26.**  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ . Test  $H_0 : \mu \in [0, 1]$  against  $H_1 : \mu < 0 \cup \mu > 1$ . Does there exist a UMP test?

### Solution.

No. Consider  $H'_0 : \mu \in [0, 1]$ ,  $H'_a : \mu > 1$  and  $H''_0 : \mu \in [0, 1]$ ,  $H''_a : \mu < 1$ . If the UMP test for our test exists, it must be identical to the UMP tests for both of these tests (since it would have to be UMP over every  $\theta$ ). By Theorem 10.8.7, the UMP test in the first case is

$$\phi'(x_1, \dots, x_n) = \begin{cases} 1, & \sum_{i=1}^n x_i \geq c \\ 0, & \sum_{i=1}^n x_i < c. \end{cases}$$

The UMP test in the second case is

$$\phi'(x_1, \dots, x_n) = \begin{cases} 1, & \sum_{i=1}^n x_i \leq c \\ 0, & \sum_{i=1}^n x_i > c. \end{cases}$$

Since these tests are not identical, there is no UMP test for our test.

**Theorem 10.8.8 (Generalized Neyman-Pearson Lemma, Theorem 3.6.1 in Lehmann and Romano).** Assume that  $f_0, f_1, \dots, f_n : \mathbb{R}^d \rightarrow \mathbb{R}$  such that

$$\int |f_j| d\mu < \infty, \quad j \in \{0, \dots, N\}.$$

1. We seek to maximize the  $\int \phi f_0 d\mu$  over all tests  $\phi$  subject to  $\int \phi f_j d\mu = \alpha_j, j \in \{1, \dots, N\}$ .
2. We seek to maximize the  $\int \phi f_0 d\mu$  over all tests  $\phi$  subject to  $\int \phi f_j d\mu \leq \alpha_j, j \in \{1, \dots, N\}$ .

Suppose there exists  $k_1, \dots, k_N$  such that

$$\phi^* = \begin{cases} 1, & f_0(x) > k_1 f_1(x) + \dots + k_N f_N(x) \\ 0, & f_0(x) \leq k_1 f_1(x) + \dots + k_N f_N(x) \end{cases}$$

satisfies  $\int \phi^* f_j d\mu = \alpha_j, j \in \{1, \dots, N\}$ . Then  $\phi^*$  solves the first problem. Moreover, if  $k_1, \dots, k_N \geq 0$ , then  $\phi^*$  also solves the second problem.

**Example 10.27 (Like Theorem 3.7.1 in Lehmann and Romano).**  $\{P_\theta, \theta \in \Theta\}$ .  $H_0 : \theta \in [\theta_1, \theta_2]$ .  $H_a : \theta \notin [\theta_1, \theta_2]$ . We have

$$p_\theta(x) = \frac{1}{c(\theta)} e^{\theta T(x)}$$

(one-parameter exponential family). Find the UMP unbiased test.

### Solution.

For the test to be unbiased, the Type I error rate has to be less than or equal to  $\alpha$  (for all  $\theta \in [\theta_1, \theta_2]$ ) and the power has to be greater than or equal to  $\alpha$  (for all  $\theta$  in the rejection region). We will use the Generalized Neyman-Pearson Lemma to show the solution is of the form

$$\phi^*(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & T(x) \in (c_1, c_2) \end{cases}$$

such that  $\mathbb{E}_{\theta_1} \phi^*(X) = \mathbb{E}_{\theta_2} \phi^*(X) = \alpha$ . By the Generalized Neyman-Pearson Lemma, we look for the solution in the form

$$\tilde{\phi}(x) = \begin{cases} 1, & p_{\theta'}(x) > k_1 p_{\theta_1}(x) + k_2 p_{\theta_2}(x) \\ 0, & p_{\theta'}(x) \leq k_1 p_{\theta_1}(x) + k_2 p_{\theta_2}(x). \end{cases}$$

This is equivalent to

$$\tilde{\phi}(x) = \begin{cases} 1, & 1 > \tilde{k}_1 e^{(\theta_1 - \theta')T(x)} + \tilde{k}_2 e^{(\theta_2 - \theta')T(x)} \\ 0, & 1 < \tilde{k}_1 e^{(\theta_1 - \theta')T(x)} + \tilde{k}_2 e^{(\theta_2 - \theta')T(x)}. \end{cases} \quad (10.34)$$

Let's analyze the inequality  $g(t) < 1$  where

$$g(t) = \tilde{k}_1 e^{a_1 t} + \tilde{k}_2 e^{a_2 t}$$

with  $a_1 = \theta_1 - \theta' > 0, a_2 = \theta_2 - \theta' > 0, a_1 < a_2$ . Consider the second case.

- (a)  $\tilde{k}_1 > 0, \tilde{k}_2 > 0 \implies g(t)$  is increasing.  $g(t) < 1 \iff t < c$  for some  $c \in \mathbb{R}$ . From the Karlin-Rubin Theorem, in this case  $\beta_{\tilde{\phi}}(\theta)$  is monotone. But this is impossible, as  $\beta_{\tilde{\phi}}(\theta_1) = \beta_{\tilde{\phi}}(\theta_2)$ .
- (b)  $\tilde{k}_1 < 0, \tilde{k}_2 < 0 \implies g(t) < 1$ . Examining (10.34) shows that in this case  $\tilde{\phi}(x) = 1$  for all  $x$  (so  $\tilde{\phi}$  always rejects), so we can't have controlled Type I error.
- (c)  $\tilde{k}_1 < 0, \tilde{k}_2 > 0$ . In this case,  $g'(t) = \tilde{k}_1 a_1 e^{a_1 t} + \tilde{k}_2 a_2 e^{a_2 t}$ . Then  $g'(t) = 0$  has a unique solution (see Figure 3), and  $g(t) < 1 \iff t \in (c_1, c_2)$  (rejection only happens in this bounded interval). But this can't be a UMP unbiased test because the power function of this test is less than a constant test. We should have high power as our test statistic goes off to infinity or negative infinity, but examining (10.34), we see that  $\tilde{\phi}(x) = 0$  as  $T(x) \rightarrow \infty$  or  $-\infty$ .

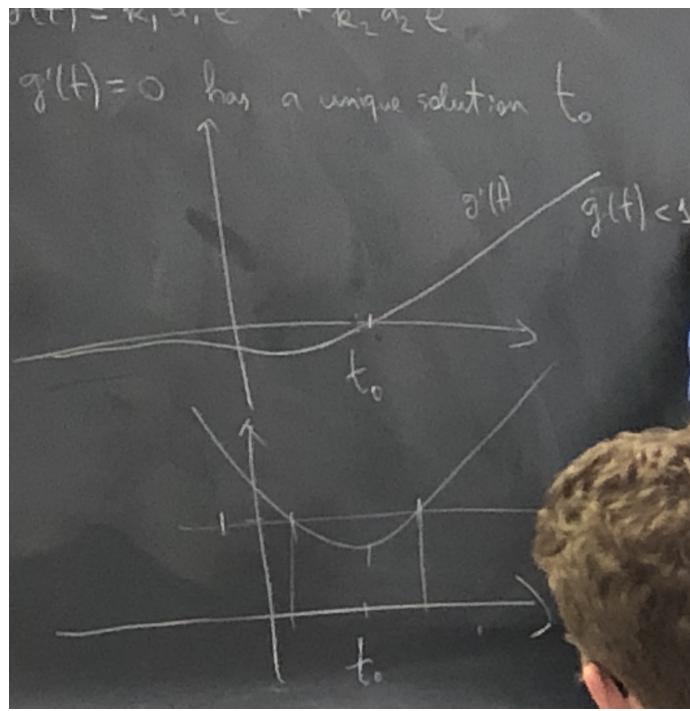


Figure 3: Figure for Example 10.27, case (c). This can't be the UMP unbiased test because our power is approaches 0 the further away we get from this narrow interval (we accept the null hypothesis when this function is more than 1).

- (d)  $\tilde{k}_1 > 0, \tilde{k}_2 < 0$ . Then  $g'(t)$  has a unique zero (see Figure 4), and  $g(t) < 1 \iff t \in (c_1, c_2)$  which gives us the form of the test that we were looking for (we always accept the null hypothesis when this function is more than 1 and reject it when it is less than 1).

It remains to show that  $\beta_{\phi^*}(\theta) \leq \alpha$  for all  $\theta \in [\theta_1, \theta_2]$  (the power function appears as in Figure 5). We can show this in one of two ways.

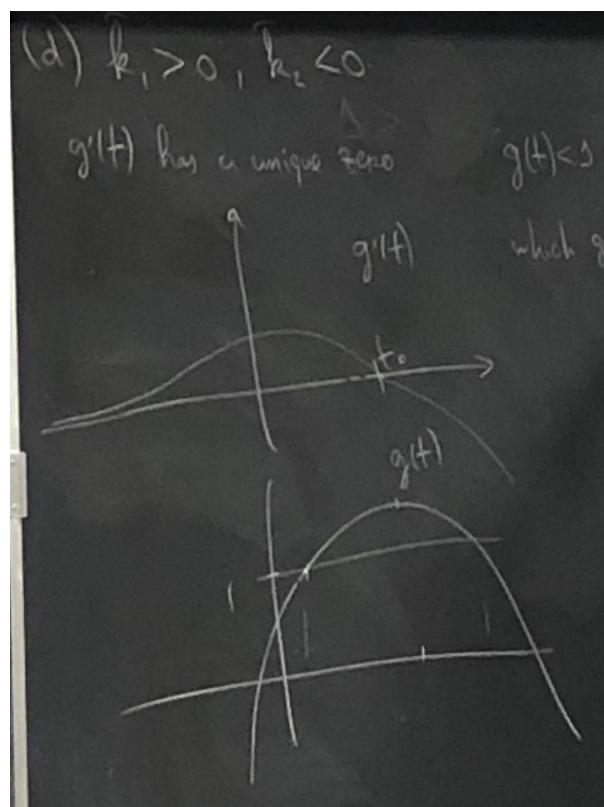


Figure 4: Second figure for Example 10.27, case (d). This is the form of the UMP unbiased test.

- (a)  $\beta'_{\phi^*}(\theta) = 0$  has a unique solution (then by a similar argument to that used in parts (c) and (d) of the previous part, the power function will have the form we want).
- (b) Lemma: the test  $\phi^*$  minimizes the power at any  $\theta' \in (\theta_1, \theta_2)$  among all SOB tests. (SOB means "similar on the boundary:" the power function is essentially constant on the boundary) (proof: exercise, similar mechanics to this).

We can prove this in a way similar to Theorem 3.6.1 in Lehmann and Romano, although the proof is a little different.

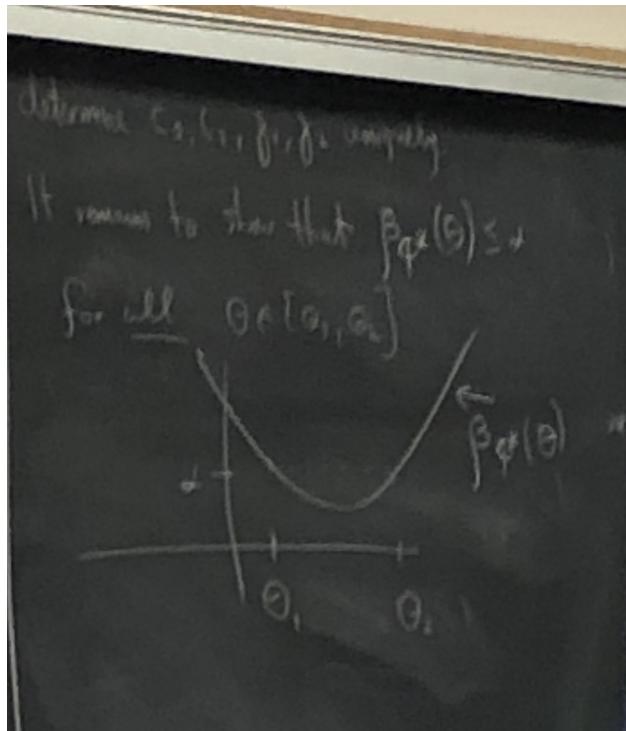


Figure 5: Third figure for Example 10.27, the desired form of a power function for our selected test.

**Theorem 10.8.9 (Theorem 3.7.1 in Lehmann and Romano).** The most powerful unbiased test for a one-parameter exponential family is

$$\phi^*(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & T(x) \in (c_1, c_2) \end{cases}$$

such that

- (a)  $\mathbb{E}_{\theta_0}(\phi^*(X)) = \alpha$ .
- (b)  $\frac{d}{d\theta}\beta_{\phi^*}(\theta)|_{\theta=\theta_0} = 0$  ( $\mathbb{E}_{\theta_0}T(X)\phi^*(X) = \alpha\mathbb{E}_{\theta_0}T(X)$ ).

*Proof of constraint (b).* For any unbiased test  $\phi$  of size  $\alpha$ ,

$$p_\theta(x) = \frac{1}{c(\theta)} e^{\theta T(x)} = \tilde{c}(\theta) e^{\theta T(x)}$$

(letting  $\tilde{c}(\theta) = \frac{1}{c(\theta)}$ ). Then

$$\beta_\phi(\theta) = \int \phi(x) p_\theta(x) d\mu = \int \phi(x) \tilde{c}(\theta) e^{\theta T(x)} d\mu$$

so the derivative is

$$\begin{aligned} \beta'_\phi(\theta) &= \int \phi(x) [\tilde{c}'(\theta) + T(x)\tilde{c}(\theta)] \tilde{c}(\theta) e^{\theta T(x)} d\mu \\ &= \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \int \phi(x) \tilde{c}(\theta) e^{\theta T(x)} d\mu + \int \phi(x) T(x) \tilde{c}(\theta) e^{\theta T(x)} d\mu = \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \mathbb{E}_\theta \phi(X) + \mathbb{E}_\theta \phi(X) T(X) \\ &= \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \alpha + \alpha \mathbb{E}_{\theta_0} T(X) = 0 \implies \frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} = -E_{\theta_0} T(X) \end{aligned}$$

Hence,

$$\frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} \underbrace{\mathbb{E}_{\theta_0} \phi(X)}_\alpha + \mathbb{E}_{\theta_0} \phi(X) T(X) = 0$$

implies that  $-\alpha \mathbb{E}_{\theta_0} T(X) + \mathbb{E}_{\theta_0} \phi(X) T(X) = 0$  so it follows that  $\mathbb{E}_{\theta_0} T(X) \phi(X) = \alpha \mathbb{E}_{\theta_0} T(X)$ .

⋮

For any unbiased test  $\phi$ ,  $\beta'_\phi(\theta) = 0$ . (We know the test is unbiased:  $\phi_\alpha = \alpha$ .) Plug  $\phi(x) = \alpha$  in to this equation to get

$$\frac{\tilde{c}'(\theta)}{\tilde{c}(\theta)} = -\mathbb{E}_{\theta_0} T(X).$$

□

**Example 10.28.** Let  $X_1, \dots, X_n$  be i.i.d.  $\mathcal{N}(0, \sigma^2)$ ,  $\sigma^2 > 0$ . Find the UMP unbiased test for  $H_0 : \sigma^2 = \sigma_0^2$ ,  $H_a : \sigma^2 \neq \sigma_0^2$ . Assuming that  $n$  is large, find the approximate values  $c_1(\alpha)$ ,  $c_2(\alpha)$ .

Before solving this we will prove a lemma that we have been using implicitly.

**Lemma 10.8.10.** Suppose that  $T(X)$  is sufficient for  $\{p_\theta, \theta \in \Theta\}$ , and let  $\phi$  be a test. Denote

$$\psi(X) := \mathbb{E}_\theta[\phi(X) \mid T(X)].$$

Then

- (a)  $\psi(X)$  is a test.
- (b)  $\mathbb{E}_\theta \psi(X) = \mathbb{E}_\theta \phi(X)$ .

*Proof.* (a) Since  $T(X)$  is sufficient, the expectation over  $\theta$  does not matter because we are conditioning on  $T(X)$  anyway. That is, the distribution of  $\phi(X)$  conditional on  $T(X)$  does not depend on  $\theta$ . Therefore the expectation over  $\theta$  does not either. (A test takes values between 0 and 1 and doesn't depend on  $\theta$ , only depends on the data.)

(b)

$$\mathbb{E}_\theta \psi(X) = \mathbb{E}_\theta (\mathbb{E}_\theta [\phi(X) \mid T(X)]) = \mathbb{E}_\theta \phi(X).$$

□

**Definition 10.42 (Boundably complete).**  $T(X)$  is **boundably complete** if and only if for any function  $G$  such that  $|G| \leq 1$  (bounded),

$$\{\mathbb{E}_\theta G(T) = 0 \quad \forall \theta \in \Theta\} \implies G(T) = 0 \text{ a.e.}$$

So if a statistic is complete, it is boundably complete (if it is complete, the equation holds for all functions  $G$ , not just bounded ones). Example:

$$X \in \{-1, 0, 1, 2, \dots\}, \quad \mathbb{P}_\theta(X = k) = \begin{cases} \theta, & k = -1 \\ (1 - \theta)^2 \theta^k, & k \in \{0, 1, \dots\} \end{cases}$$

Then  $T(X) = X$  is boundably complete but not complete.

We want to be able to work with distributions that are more than one dimensional. We can do this by reducing the family to a one-parameter family. We do this in the following way, with tests with Neyman structure. Assume we have  $X \sim P_\theta, \theta \in \Theta$ . The problem is  $H_0 : \theta \in \Theta_0, H_a : \theta \in \Theta_a$ , where  $\Theta_B := \overline{\Theta}_0 \cap \overline{\Theta}_a$  (the intersection of the closures of these sets).

Example:  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ ,  $H_0 : \mu \leq 0, H_a : \mu > 0$ . Then if we plot  $\mathbb{R}^2$  with  $\mu$  as the horizontal axis and  $\sigma^2$  as the vertical axis, the acceptance region is Quadrant II (including the top part of the vertical axis) and the rejection region is Quadrant I (not including the top part of the vertical axis). The boundary is the vertical axis ( $\Theta_B = \{\mathcal{N}(0, \sigma^2), \sigma^2 > 0\}$ ).

Approach: take a complete sufficient statistic for the boundary family. We will condition on it and then find tests of size  $\alpha$ .

**Definition 10.43.** Let  $T(X)$  be sufficient for the boundary family  $\{p_\theta, \theta \in \Theta_B\}$ . Test  $\phi$  has **Neyman structure with respect to  $T$**  if and only if for some  $\alpha \in [0, 1]$

$$\mathbb{E}_\theta[\phi(X) \mid T(X)] = \alpha, \quad \forall \theta \in \Theta_B.$$

Note that  $\phi$  has Neyman structure implies it is SOB of size  $\alpha$  (SOB means "similar on the boundary:" the power function is essentially constant on the boundary).

**Theorem 10.8.11.** All SOB tests of size  $\alpha$  have Neyman structure if and only if  $T$  is boundably complete (and sufficient).

*Proof.* First we will prove this direction: suppose that  $T$  is sufficient and boundably complete. Then

$$\mathbb{E}_\theta \phi(X) = \alpha \quad \forall \theta \in \Theta_B.$$

But then

$$\begin{aligned} \alpha = \mathbb{E}_\theta \phi(X) &= \mathbb{E}_\theta \mathbb{E}[\phi(X) \mid T(X)] = \mathbb{E}_\theta G(T) \implies \mathbb{E}_\theta(G(T) - \alpha) = 0, \quad \theta \in \Theta_0 \\ &\implies G(T) - \alpha = 0 \text{ a.s.} \implies G(T) = \alpha \text{ a.s.} \end{aligned}$$

so the result follows. (Note that we only need boundable completeness since we must have  $0 \leq |G| \leq 1 \forall T$ .) Next, assume that all SOB tests of size  $\alpha$  have Neyman structure. We will prove by contradiction that then  $T$  must be boundably complete. Suppose  $T$  is not boundably complete. Then there must exist some function  $\psi$  such that  $|\psi| \leq 1$  and  $\mathbb{E}_\theta \psi(T) = 0 \forall \theta \in \Theta_B$  but  $\psi(T)$  is not identically 0. Define  $\phi := \alpha + c\psi(T)$ , where we choose  $c \neq 0$  such that  $0 \leq |\phi| \leq 1$  so that  $\phi$  is a test. Note that  $\mathbb{E}_\theta \phi = \alpha$  for all  $\theta \in \Theta_B$ , so  $\phi$  is SOB of size  $\alpha$  (it is SOB because this result holds for all  $\theta \in \Theta_B$ , where  $\Theta_B$  is the boundary family). But

$$\mathbb{E}(\phi(x) \mid T(x)) = \alpha + c\psi(T(x)) \neq \alpha$$

since  $\psi$  is not identically 0 by assumption. □

**Theorem 10.8.12.** Assume that  $X \sim P_\theta$  where

$$p_\theta(x) = \frac{1}{c(\theta)} h(x) \cdot \exp \left\{ \sum_{j=1}^k \theta_j T_j(x) \right\}$$

(canonical exponential family). Then  $T(X) = (T_1(X), \dots, T_k(X))$  is a sufficient statistic for the family  $\{p_\theta, \theta \in \Theta\}$  provided that  $\Theta$  has non-empty interior. (See also Proposition 10.3.7.)

Conditioning method: assume that  $\beta_\phi(\theta)$  is continuous for any test  $\phi$ .

(1)  $\phi$  is unbiased, so  $\phi$  is SOB.

- (2) UMP SOB test of size  $\alpha$  is UMP unbiased test of size  $\alpha$ .  
(3) Neyman structure implies that it suffices to find the test that maximizes the conditional power  $\mathbb{E}_\theta(\phi | T)$  for all  $\theta \in \Theta_a$ .

**Example 10.29.**  $X \sim \text{Poisson}(\mu)$ ,  $Y \sim \text{Poisson}(\nu)$ .  $H_0 : \mu \leq \nu$ ,  $H_a : \nu > \mu$ . Find the UMP unbiased test.

**Solution.**

$$p_{\mu,\nu}(x,y) = \frac{e^{-(\mu+\nu)} \mu^x \nu^y}{x!y!} = e^{-(\mu+\nu)} e^{x \log \mu + y \log \nu} \cdot \frac{1}{x!y!} = \frac{1}{x!y!} e^{-(\mu+\nu)} e^{x \log(\mu/\nu) + (x+y) \log \nu}$$

Let  $\xi := \log(\mu/\nu)$ ,  $\eta = \log \nu$ . Then  $H_0 : \xi \leq 0$ ,  $H_a : \xi > 0$ . We can use an SOB thing where  $\Theta_B = \{(\xi, \eta) : \eta = 0\}$  is the boundary family. Then  $T(X, Y) = X + Y$  is complete sufficient for the boundary family  $\{p_{\xi, \eta} : \xi = 0\}$  by Theorem 10.8.12 (Proposition 10.3.7). Let's condition everything on  $T = X + Y$ . We need to find  $\mathbb{P}(X = x | T = t)$ . Note that

$$\mathbb{P}_{\mu,\nu}(X = x | T = t) = \frac{\mathbb{P}_{\mu,\nu}(X = x, Y = t - x)}{\mathbb{P}_{\mu,\nu}(T = t)} = \frac{\mu^x}{x!} e^{-\mu} \cdot \frac{\nu^{t-x}}{(t-x)!} e^{-\nu} \cdot \frac{t! e^{\mu+\nu}}{(\mu+\nu)^t}$$

where we used  $T \sim \text{Poisson}(\mu + \nu)$  (exercise; use characteristic functions)

$$= \binom{t}{x} \left( \frac{\mu}{\mu+\nu} \right)^x \left( \frac{\nu}{\mu+\nu} \right)^{t-x}$$

so  $\{X | T = t\} \sim \text{Bin}(t, \mu/(\mu + \nu))$ .

Note that  $\mu \leq \nu \iff p \leq 1/2$ , so for our new problem, we have  $X \sim \text{Binom}(t, p)$ , and we will test  $H'_0 : p \leq 1/2$ ,  $H'_a : p > 1/2$ . For this problem we can find the UMP test (not just the UMP unbiased test) because the binomial family has a monotone likelihood ratio with respect to  $T(X) = X$  when  $p \in (0, 1)$  and  $t$  fixed (very straightforward to check).

Then

$$\phi(x) = \begin{cases} 1, & x > c(t) \\ \gamma, & x = c(t) \\ 0, & x < c(t) \end{cases}$$

is UMP. To find  $c(t, \alpha)$ , use

$$\mathbb{E}_{p=1/2} \phi(X) = \alpha \iff \sum_{x>c(t)} \binom{t}{x} (1/2)^t + \gamma \binom{t}{c(t)} (1/2)^{c(t)} = \alpha. \quad (10.35)$$

(hard to compute in practice; need a computer). So the test for the original problem is

$$\phi(X, Y) = \begin{cases} 1, & X > c(t) \\ \gamma, & x = c(t) \\ 0, x < c(t) \end{cases}$$

where  $t = X + Y$  and  $c(t)$  solves (10.35).

**Proposition 10.8.13 (Stats 100B Homework problem).** Let  $Y_1, Y_2, \dots, Y_n$  be the outcomes of  $n$  independent Bernoulli trials. Then by the Neyman-Pearson lemma, the best critical region for testing

$$H_0 : p = p_0 \quad H_a : p > p_0$$

is

$$\frac{y}{n} = \frac{1}{n} \sum Y_i > \frac{\log(K) + n \log\left(\frac{1-p_a}{1-p_0}\right)}{n \log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)}.$$

*Proof.*

$$\Pr(\sum Y_i = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

Using the Neyman-Pearson lemma (let  $p_a$  be some particular value of  $p > p_0$ ):

$$\frac{L(p_0)}{L(p_a)} = \frac{\binom{n}{y} p_0^y (1-p_0)^{n-y}}{\binom{n}{y} p_a^y (1-p_a)^{n-y}} < K$$

$$\left(\frac{p_0}{p_a}\right)^y \left(\frac{1-p_0}{1-p_a}\right)^n \left(\frac{1-p_0}{1-p_a}\right)^{-y} < K$$

$$\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)^y < K \left(\frac{1-p_a}{1-p_0}\right)^n$$

$$y \log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right) < \log(K) + n \log\left(\frac{1-p_a}{1-p_0}\right)$$

Aside:

$$\frac{p_0(1-p_a)}{p_a(1-p_0)} = \frac{p_0 - p_0 p_a}{p_a - p_0 p_a} < 1$$

since by assumption  $p_a > p_0$ . Therefore  $\log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right) < 0$ . So we have

$$\frac{y}{n} = \frac{1}{n} \sum Y_i > \frac{\log(K) + n \log\left(\frac{1-p_a}{1-p_0}\right)}{n \log\left(\frac{p_0(1-p_a)}{p_a(1-p_0)}\right)}$$

as the form for our critical region.

□

## 11 Linear Regression

These notes are based on my notes from *Time Series and Panel Data Econometrics* (1st edition) by M. Hashem Pesaran [Pesaran, 2015] and coursework for Economics 613: Economic and Financial Time Series I at USC taught by M. Hashem Pesaran, DSO 607 at USC taught by Jinchi Lv, Statistics 100B at UCLA taught by Nicolas Christou, and the Coursera MOOC “Econometrics: Methods and Applications” from Erasmus University Rotterdam. I also borrowed from some other sources which I mention when I use them.

### 11.1 Chapter 1: Linear Regression

#### 11.1.1 Preliminaries

Suppose the true model is  $y_i = \alpha + \beta x_i + \epsilon_i$ . Classical assumptions:

- (i)  $\mathbb{E}(\epsilon_i) = 0$
- (ii)  $\text{Var}(\epsilon_i | x_i) = \sigma^2$  (constant)
- (iii)  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$  if  $i \neq j$
- (iv)  $\epsilon_i$  is uncorrelated to  $x_i$ , or  $\mathbb{E}(\epsilon_i | x_j) = 0$  for all  $i, j$ .

#### 11.1.2 Estimation

$$\hat{\beta} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

or

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

or

$$\hat{\beta} = r \frac{S_{YY}}{S_{XX}}$$

where  $r$  is the correlation coefficient.

Let

$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

so that

$$\hat{\beta} = \sum_{i=1}^n w_i(y_i - \bar{y}) = \sum_{i=1}^n w_i y_i - \bar{y} \frac{\sum_{i=1}^n x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i y_i$$

since  $\sum_{i=1}^n x_i - \bar{x} = 0$ . Then a simple expression for  $\text{Var}(\hat{\beta})$  is

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^n w_i^2 \text{Var}(y_i | x_i) = \sum_{i=1}^n w_i^2 \text{Var}(\epsilon | x_i) = \sigma^2 \sum_{i=1}^n w_i^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{XX}}$$

We can estimate these quantities as follows:

$$\hat{\sigma}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Note that

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{t=1}^T (y_t - \hat{\alpha} - \hat{\beta} x_t)^2 = \frac{1}{n-2} \sum_{t=1}^T [(y_t - (\bar{y} - \hat{\beta} \bar{x}) - \hat{\beta} x_t)^2] = \frac{1}{n-2} \sum_{t=1}^T (y_t - \bar{y} - \hat{\beta}(x_t - \bar{x}))^2 \\ &= \frac{1}{n-2} \sum_{t=1}^T (y_t - \bar{y})^2 - 2\hat{\beta}(x_t - \bar{x})(y_t - \bar{y}) + \hat{\beta}^2(x_t - \bar{x})^2 \end{aligned}$$

In the case where there is no intercept, we have

$$\hat{\sigma}^2 = \frac{1}{T-1} \sum_{t=1}^T (y_t - \hat{\beta} x_t)^2 = \frac{1}{T-1} \sum_{t=1}^T \left( y_t^2 - 2r \frac{S_{YY}}{S_{XX}} x_t y_t + r^2 \frac{S_{YY}^2}{S_{XX}^2} x_t^2 \right)$$

Also,

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{\hat{\sigma}^2}{S_{XX}} = \frac{1}{n-2} \cdot \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Correlation coefficient:

$$r^2 = \frac{(\sum_{t=1}^T x_t y_t)^2}{\sum_{t=1}^T x_t^2 \sum_{t=1}^T y_t^2}$$

$$r = \frac{1}{T-1} \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

**Remark.** The formulas for the coefficients in univariate OLS can also be derived by considering  $(x, y)$  as a bivariate normal distribution and calculating the conditional expectation of  $y$  given  $x$ . (See Proposition (6.3.29).)

**Proposition 11.1.1 (Stats 100B homework problem).** Consider the regression model  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  with  $x_i$  fixed and  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ ,  $\epsilon_i$  i.i.d. Let  $e_i = y_i - \hat{y}_i$  be the residuals.

(a)

$$\sum_{i=1}^n e_i = 0$$

(b)  $\text{Cov}(\bar{Y}, \hat{\beta}_1) = 0$  where  $\bar{Y}$  is the sample mean of the  $y$  values.

(c)

$$\text{Cov}(e_i, e_j) = \sigma^2 \left( -\frac{1}{n} - \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2} \right)$$

(d) We can construct a confidence interval for  $\sigma^2$  as

$$\Pr \left( \frac{\sum_{i=1}^n e_i^2}{\chi^2_{1-\frac{\alpha}{2}; n-2}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n e_i^2}{\chi^2_{\frac{\alpha}{2}; n-2}} \right) = 1 - \alpha$$

*Proof.* (a)

$$\begin{aligned} \sum_{i=1}^n e_i &= \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - [\bar{y} + \hat{\beta}_1(x_i - \bar{x})]) \\ &= \sum_{i=1}^n \left( y_i - \bar{y} - \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} (x_i - \bar{x}) \right) = \sum_{i=1}^n y_i - n\bar{y} - \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) \\ &= \sum_{i=1}^n y_i - n \frac{1}{n} \sum_{i=1}^n y_i - \left( \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \right) \left[ \sum_{i=1}^n \left( x_i - \frac{1}{n} \sum_{i=1}^n x_i \right) \right] \\ &= \sum_{i=1}^n (y_i - \bar{y}) - \left( \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \right) \left[ \sum_{i=1}^n x_i - \frac{1}{n} \cdot n \sum_{i=1}^n x_i \right] = 0 - 0 = \boxed{0} \end{aligned}$$

Or:

$$\begin{aligned} \sum_{i=1}^n e_i &= \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^n (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) = 0 \end{aligned}$$

(b)

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = \text{Cov} \left( \frac{1}{n} \sum_{i=1}^n Y_i, \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \text{Cov} \left( \sum_{i=1}^n Y_i, \sum_{i=1}^n (x_i - \bar{x})Y_i \right)$$

$x_i$  is fixed,  $\text{Cov}(Y_i, Y_j) = 0$  for  $i \neq j$  by assumption of the model,  $\text{Var}(Y_i) = \sigma^2$  by assumption of the model.

$$= \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n [(x_i - \bar{x}) \text{Var}(Y_i)] = \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) = \boxed{0}$$

(c)

$$\begin{aligned} \text{Cov}(e_i, e_j) &= \text{Cov}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}), y_j - \bar{y} - \hat{\beta}_1(x_j - \bar{x})) \\ &= \text{Cov}(y_i, y_j) - \text{Cov}(y_i, \bar{y}) - \text{Cov}(y_i, \hat{\beta}_1(x_j - \bar{x})) - \text{Cov}(\bar{y}, y_j) + \text{Cov}(\bar{y}, \bar{y}) + \text{Cov}(\bar{y}, \hat{\beta}_1(x_j - \bar{x})) - \text{Cov}(\hat{\beta}_1(x_i - \bar{x}), y_j) \\ &\quad + \text{Cov}(\hat{\beta}_1(x_i - \bar{x}), \bar{y}) + \text{Cov}(\hat{\beta}_1(x_i - \bar{x}), \hat{\beta}_1(x_j - \bar{x})) \end{aligned}$$

By assumption of the model,  $\text{Cov}(y_i, y_j) = 0$ .

$$\begin{aligned} &= 0 - \text{Cov}(y_i, \bar{y}) - (x_j - \bar{x})\text{Cov}(y_i, \hat{\beta}_1) - \text{Cov}(\bar{y}, y_j) + \text{Var}(\bar{y}) + (x_j - \bar{x})\text{Cov}(\bar{y}, \hat{\beta}_1) - (x_i - \bar{x})\text{Cov}(\hat{\beta}_1, y_j) \\ &\quad + (x_i - \bar{x})\text{Cov}(\hat{\beta}_1, \bar{y}) + (x_i - \bar{x})(x_j - \bar{x})\text{Cov}(\hat{\beta}_1, \hat{\beta}_1) \end{aligned}$$

In part 7(b) we showed  $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$ .  $\text{Var}(\bar{y}) = \sigma^2/n$ .  $\text{Cov}(\hat{\beta}_1, \hat{\beta}_1) = \text{Var}(\hat{\beta}_1) = \sigma^2 / \sum(x_k - \bar{x})^2$ . So this simplifies to

$$\begin{aligned} &= -\text{Cov}(y_i, \bar{y}) - (x_j - \bar{x})\text{Cov}(y_i, \hat{\beta}_1) - \text{Cov}(y_j, \bar{y}) + \frac{\sigma^2}{n} + 0 - (x_i - \bar{x})\text{Cov}(y_j, \hat{\beta}_1) + 0 + (x_i - \bar{x})(x_j - \bar{x}) \frac{\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2} \\ &= -\text{Cov}(y_i, \bar{y}) - (x_j - \bar{x})\text{Cov}(y_i, \hat{\beta}_1) - \text{Cov}(y_j, \bar{y}) + \frac{\sigma^2}{n} - (x_i - \bar{x})\text{Cov}(y_j, \hat{\beta}_1) + (x_i - \bar{x})(x_j - \bar{x}) \frac{\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2} \end{aligned} \tag{11.1}$$

Find  $\text{Cov}(y_i, \bar{y})$ ,  $\text{Cov}(y_j, \bar{y})$ ,  $\text{Cov}(y_i, \hat{\beta}_1)$ , and  $\text{Cov}(y_j, \hat{\beta}_1)$ :

Using that  $x_i$  is fixed,  $\text{Cov}(Y_i, Y_j) = 0$  for  $i \neq j$  by assumption of the model,  $\text{Var}(Y_i) = \sigma^2$  by assumption of the model:

$$\text{Cov}(y_i, \bar{y}) = \text{Cov}\left(y_i, \frac{1}{n} \sum_{k=1}^n y_k\right) = \frac{1}{n} \text{Cov}(y_i, y_i) = \frac{\sigma^2}{n}$$

Similarly,

$$\text{Cov}(y_j, \bar{y}) = \frac{\sigma^2}{n}$$

$$\begin{aligned} \text{Cov}(y_i, \hat{\beta}_1) &= \text{Cov}\left(y_i, \frac{\sum_{k=1}^n (x_k - \bar{x}) y_k}{\sum_{k=1}^n (x_k - \bar{x})^2}\right) = \frac{1}{\sum_{k=1}^n (x_k - \bar{x})^2} \text{Cov}\left(y_i, \sum_{k=1}^n (x_k - \bar{x}) y_k\right) \\ &= \frac{1}{\sum_{k=1}^n (x_k - \bar{x})^2} \text{Cov}(y_i, (x_i - \bar{x}) y_i) = \frac{x_i - \bar{x}}{\sum_{k=1}^n (x_k - \bar{x})^2} \text{Var}(y_i) = \frac{x_i - \bar{x}}{\sum_{k=1}^n (x_k - \bar{x})^2} \sigma^2 \end{aligned}$$

Similarly,

$$\text{Cov}(y_j, \hat{\beta}_1) = \frac{x_j - \bar{x}}{\sum_{k=1}^n (x_k - \bar{x})^2} \sigma^2$$

Plugging these in to equation (11.1) yields

$$\begin{aligned}
 \text{Cov}(e_i, e_j) &= -\frac{\sigma^2}{n} - (x_j - \bar{x}) \frac{(x_i - \bar{x})\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2} - \frac{\sigma^2}{n} + \frac{\sigma^2}{n} - (x_i - \bar{x}) \frac{(x_j - \bar{x})\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2} \\
 &\quad + (x_i - \bar{x})(x_j - \bar{x}) \frac{\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2} \\
 &= \frac{-\sigma^2}{n} - \sigma^2 \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2} \\
 \text{Cov}(e_i, e_j) &= \sigma^2 \left( -\frac{1}{n} - \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2} \right)
 \end{aligned}$$

(d) From class notes 08/29:

$$\begin{aligned}
 \frac{(n-2)S_e^2}{\sigma^2} &\sim \chi_{n-2}^2 \\
 \implies \Pr \left( \chi_{\frac{\alpha}{2}; n-2}^2 \leq \frac{(n-2)S_e^2}{\sigma^2} \leq \chi_{1-\frac{\alpha}{2}; n-2}^2 \right) &= 1 - \alpha \\
 \implies \boxed{\Pr \left( \frac{(n-2)S_e^2}{\chi_{1-\frac{\alpha}{2}; n-2}^2} \leq \sigma^2 \leq \frac{(n-2)S_e^2}{\chi_{\frac{\alpha}{2}; n-2}^2} \right) = 1 - \alpha}
 \end{aligned}$$

Since

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

this interval can be expressed as

$$\Pr \left( \frac{\sum_{i=1}^n e_i^2}{\chi_{1-\frac{\alpha}{2}; n-2}^2} \leq \sigma^2 \leq \frac{\sum_{i=1}^n e_i^2}{\chi_{\frac{\alpha}{2}; n-2}^2} \right) = 1 - \alpha$$

□

**Proposition 11.1.2 (Stats 100B homework problem).** Suppose  $Y_i = \beta_1 x_i + \epsilon_i$  (no intercept). Suppose  $x_i$  is fixed and  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ .

(a) The maximum likelihood estimator of  $\beta_1$  is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

which is unbiased. Its variance is  $\frac{\sigma^2}{\sum_{i=1}^n x_i^2}$  and it is normally distributed.

(b) The maximum likelihood estimator of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_1 x_i)^2.$$

*Proof.* (a) First we find the likelihood function to find the MLE. Assuming the  $n$  observations are independent,

$$\begin{aligned} L &= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (y_i - \beta_1 x_i)^2 \right) \\ &= (2\sigma^2 \pi)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 x_i)^2 \right) \end{aligned}$$

Next,

$$\begin{aligned} \log(L) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 x_i)^2 \\ \frac{d \log(L)}{d\beta_1} &= \frac{d}{d\beta_1} \left( -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 x_i)^2 \right) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n x_i (y_i - \beta_1 x_i) = 0 \\ \sum_{i=1}^n x_i y_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0 \\ \implies \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \end{aligned}$$

Next we show that this estimator is unbiased.

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) = \frac{1}{\sum_{i=1}^n x_i^2} \mathbb{E}\left(\sum_{i=1}^n x_i (\beta_1 x_i + \epsilon_i)\right) = \frac{1}{\sum_{i=1}^n x_i^2} \left[ \mathbb{E}\left(\sum_{i=1}^n x_i^2 \beta_1\right) + \mathbb{E}\left(\sum_{i=1}^n x_i \epsilon_i\right) \right]$$

Since  $x_i$  and  $\beta_1$  are non-random and  $\epsilon_i$  are independent, this can be written as

$$\frac{1}{\sum_{i=1}^n x_i^2} \left[ \sum_{i=1}^n x_i^2 \beta_1 + \sum_{i=1}^n x_i \mathbb{E}(\epsilon_i) \right] = \frac{1}{\sum_{i=1}^n x_i^2} \beta_1 \sum_{i=1}^n x_i^2 = \beta_1$$

Next we find the variance.

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) = \frac{1}{(\sum_{i=1}^n x_i^2)^2} \text{Var}\left(\sum_{i=1}^n x_i (\beta_1 x_i + \epsilon_i)\right) \\ &= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \left[ \text{Var}\left(\sum_{i=1}^n x_i^2 \beta_1\right) + \text{Var}\left(\sum_{i=1}^n x_i \epsilon_i\right) \right] \end{aligned}$$

Since  $x_i$  and  $\beta_1$  are non-random and  $\epsilon_i$  are independent, this can be written as

$$\frac{1}{(\sum_{i=1}^n x_i^2)^2} \left[ 0 + \sum_{i=1}^n x_i^2 \text{Var}(\epsilon_i) \right] = \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sigma^2 \sum_{i=1}^n x_i^2 = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

$\beta_1$  is a linear combination of  $y_i$  which is normally distributed, therefore  $\beta_1$  is normally distributed.

$$\implies \beta_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^n x_i^2}}\right)$$

(b)

$$\begin{aligned} \frac{d \log(L)}{d\sigma^2} &= \frac{d}{d\sigma^2} \left( -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 x_i)^2 \right) \\ &= -\frac{n}{2} \frac{1}{2\pi\sigma^2} 2\pi - \frac{1}{2} \left( -\frac{1}{(\sigma^2)^2} \right) \sum_{i=1}^n (y_i - \beta_1 x_i)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (y_i - \beta_1 x_i)^2 = 0 \\ &\quad \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (y_i - \beta_1 x_i)^2 = \frac{n}{2\hat{\sigma}^2} \\ &\quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_1 x_i)^2 \end{aligned}$$

□

**Remark.** More details on this problem available in Math 541A Homework 7.

## 11.2 Chapter 2: Multiple Regression

General OLS:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) = \text{Var}(\beta) + \text{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) = 0 + \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{u}\mathbf{u}' \mid \mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2 \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'I_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2 \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Or:

$$\begin{aligned} \text{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Var}[\mathbf{y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[\sigma^2 I_n]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k}$$

**Theorem 11.2.1 (Gauss-Markov Theorem, as stated in Pesaran [2015]).** Suppose we have data generated by

$$y = X\beta + \epsilon$$

and we make the following assumptions.

1.  $\mathbb{E}(\epsilon) = 0$ .
2. Homoskedasticity:  $\text{Var}(\epsilon_i | X) = \sigma^2 > 0 \quad \forall i$
3. Uncorrelated errors:  $\text{Cov}(\epsilon_i, \epsilon_j | X) = 0, \quad \forall i \neq j$ .
4. Orthogonality:  $\mathbb{E}(\epsilon_i | X) = 0, \quad \forall i$ .

Then if  $X\beta$  is estimable, then the best linear unbiased estimator (BLUE) of  $\beta$  is  $\hat{\beta}_{OLS}$ , the least squares estimate. That is, if  $\tilde{\beta}$  is an alternative linear unbiased estimator, where  $\tilde{\beta} = \hat{\beta}_{OLS} + C^T y$  with  $C \in \mathbb{R}^{n \times p}$  and  $\mathbb{E}(\tilde{\beta}) = \beta$  for all  $\beta$ , then  $\text{Var}(\tilde{\beta} | X) \geq \text{Var}(\hat{\beta}_{OLS} | X)$ .

*Proof (adapted from Pesaran [2015]).* Note that  $\text{Var}(\hat{\beta}_{OLS} | X) \leq \text{Var}(\tilde{\beta} | X) \iff \text{Var}(\tilde{\beta} | X) - \text{Var}(\hat{\beta}_{OLS} | X)$  is a positive semidefinite matrix. We have

$$\tilde{\beta} = [(X^T X)^{-1} X^T + C^T] y = [(X^T X)^{-1} X^T + C^T](X\beta + \epsilon) = \beta + (X^T X)^{-1} X^T \epsilon + C^T X \beta + C^T \epsilon$$

Since  $\mathbb{E}(\tilde{\beta}) = \beta$  for all  $\beta$ , we have  $C^T X \beta = 0$  for all  $\beta$ , so  $C^T X = 0$ . (Note that since  $n \geq p$ ,  $C^T \in \mathbb{R}^{p \times n}$  has at least an  $n - p$ -dimensional nullspace.)

$$\implies \tilde{\beta} = \beta + (X^T X)^{-1} X^T \epsilon + C^T \epsilon \iff \tilde{\beta} - \beta = [(X^T X)^{-1} X^T + C^T] \epsilon \quad (11.2)$$

Then

$$\text{Var}(\tilde{\beta}) = \mathbb{E}[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T] = \mathbb{E}[\tilde{\beta}\tilde{\beta}^T - \tilde{\beta}\beta^T - \beta\tilde{\beta}^T + \beta\beta^T] \quad (11.3)$$

Using (11.2) we have

$$\begin{aligned} \mathbb{E}[\tilde{\beta}\tilde{\beta}^T] &= \mathbb{E}[(\beta + (X^T X)^{-1} X^T \epsilon + C^T \epsilon)(\beta + (X^T X)^{-1} X^T \epsilon + C^T \epsilon)^T] \\ &= \mathbb{E}[\beta\beta^T + \beta\epsilon^T X(X^T X)^{-1} + \beta\epsilon^T X + (X^T X)^{-1} X^T \epsilon\beta^T + (X^T X)^{-1} X^T \epsilon\epsilon^T X(X^T X)^{-1} \\ &\quad + (X^T X)^{-1} X^T \epsilon\epsilon^T C + C^T \epsilon\beta^T + C^T \epsilon\epsilon^T X(X^T X)^{-1} + C^T \epsilon\epsilon^T C] \end{aligned}$$

$$\begin{aligned}
&= \beta\beta^T + (X^T X)^{-1} X^T \mathbb{E} [\epsilon\epsilon^T] X (X^T X)^{-1} + (X^T X)^{-1} X^T \mathbb{E} [\epsilon\epsilon^T] C + C^T \mathbb{E} [\epsilon\epsilon^T] X (X^T X)^{-1} + C^T \mathbb{E} [\epsilon\epsilon^T] C \\
&= \beta\beta^T + \sigma^2(X^T X)^{-1} + \sigma^2(X^T X)^{-1} X^T C + \sigma^2 C^T X (X^T X)^{-1} + \sigma^2 C^T C \\
&= \beta\beta^T + \sigma^2(X^T X)^{-1} + \sigma^2 C^T C
\end{aligned} \tag{11.4}$$

Also,

$$\begin{aligned}
\tilde{\beta}\beta^T &= (\beta + (X^T X)^{-1} X^T \epsilon + C^T \epsilon) \beta^T = \beta\beta^T + (X^T X)^{-1} X^T \epsilon \beta^T + C^T \epsilon \beta^T \\
\implies \mathbb{E}(\tilde{\beta}\beta^T) &= \beta\beta^T, \quad \mathbb{E}(\beta\tilde{\beta}^T) = \mathbb{E}(\tilde{\beta}\beta^T) = \beta\beta^T.
\end{aligned} \tag{11.5}$$

Substituting (11.4) and (11.5) into (11.3), we have

$$\begin{aligned}
\text{Var}(\tilde{\beta}) &= \beta\beta^T + \sigma^2(X^T X)^{-1} + \sigma^2 C^T C - 2\beta\beta^T + \beta\beta^T \\
&= \sigma^2 [(X^T X)^{-1} + C^T C]
\end{aligned}$$

Therefore (using  $\text{Var}(\hat{\beta}_{OLS}) = \sigma^2(X^T X)^{-1}$ )

$$\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta}_{OLS}) = \sigma^2 [(X^T X)^{-1} + C^T C - (X^T X)^{-1}] = \sigma^2 C^T C$$

which is a positive semidefinite matrix since the inner product of a matrix with itself is positive semidefinite (see Proposition 2.3.1).

□

**Theorem 11.2.2 (Gauss-Markov Theorem, as stated in Faraway [2002]).** Suppose

$$y = X\beta + \epsilon$$

with  $X \in \mathbb{R}^{n \times p}$  a fixed full rank matrix and  $n \geq p$ ,  $\beta \in \mathbb{R}^p$ , and  $\epsilon \in \mathbb{R}^n$  with  $\mathbb{E}(\epsilon) = 0$  and  $\text{Var}(\epsilon) = \sigma^2 I_n$ . Suppose  $X$  is fixed ( $\mathbb{E}(Y) = X\beta$ ). Let  $\psi : \mathbb{R}^p \rightarrow \mathbb{R}$  be an estimable function:  $\psi := c^T \beta$  for some  $c \in \mathbb{R}^p$ . Then in the class of all unbiased linear estimates of  $\psi$ ,  $\hat{\psi} = c^T \hat{\beta}$  has the minimum variance and is unique.

*Proof (adapted from Faraway [2002]).* Suppose for some  $a \in \mathbb{R}^n$ ,  $a^T y$  is another unbiased estimator of  $c^T \beta$  so that

$$\mathbb{E}(a^T y) = a^T X \beta = c^T \beta, \quad \forall \beta \in \mathbb{R}^p.$$

This implies

$$a^T X = c^T \iff X^T a = c, \quad (11.6)$$

Since  $X$  has full column rank,  $X^T X$  is full rank and its column space is all of  $\mathbb{R}^p$ , so there exists a unique  $\lambda \in \mathbb{R}^p$  such that

$$c = X^T X \lambda = X^T a \quad (11.7)$$

(Note that  $a = X\lambda + k$  where  $k \in \mathbb{R}^n$  is any vector in the  $n - p$ -dimensional nullspace of  $X^T$ .) Then

$$\begin{aligned} \text{Var}(a^T y) &= \text{Var}(a^T y - c^T \hat{\beta} + c^T \hat{\beta}) = \text{Var}(a^T y - \lambda^T X^T \hat{y} + c^T \hat{\beta}) \\ &= \text{Var}(a^T y - \lambda^T X^T \hat{y}) + \text{Var}(c^T \hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}) \geq \text{Var}(c^T \hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}), \end{aligned} \quad (11.8)$$

so we are done if the covariance in (11.8) is nonnegative.

$$\begin{aligned} \text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}) &= \mathbb{E} \left[ (a^T y - \lambda^T X^T \hat{y} - \mathbb{E}[a^T y - \lambda^T X^T \hat{y}]) (c^T \hat{\beta} - \mathbb{E}[c^T \hat{\beta}]) \right] \\ &= \mathbb{E} \left[ (a^T (X\beta + \epsilon) - \lambda^T X^T X \hat{\beta} - a^T X\beta + \lambda^T X^T X \beta) (c^T \hat{\beta} - c^T \beta) \right] \\ &= \mathbb{E} \left[ (a^T \epsilon - \lambda^T X^T X (\hat{\beta} - \beta)) c^T (\hat{\beta} - \beta) \right] = \mathbb{E} [(a^T \epsilon - \lambda^T X^T X (X^T X)^{-1} X^T \epsilon) c^T (X^T X)^{-1} X^T \epsilon] \\ &= \mathbb{E} [a^T \epsilon c^T (X^T X)^{-1} X^T \epsilon] - \mathbb{E} [\lambda^T X^T \epsilon c^T (X^T X)^{-1} X^T \epsilon] \\ &= (a^T - \lambda^T X^T) \mathbb{E} [\epsilon c^T (X^T X)^{-1} X^T \epsilon] = (a^T - \lambda^T X^T) \mathbb{E} [\epsilon (\lambda^T X^T \epsilon)] \end{aligned}$$

Note that

$$\begin{aligned} \lambda^T X^T \epsilon &= \sum_{j=1}^n (\lambda^T X^T)_j \epsilon_j \in \mathbb{R}^n \\ \implies (a^T - \lambda^T X^T) \mathbb{E} [\epsilon (\lambda^T X^T \epsilon)] &= (a^T - \lambda^T X^T) \mathbb{E} \begin{bmatrix} \epsilon_1 [\lambda^T X^T \epsilon] \\ \vdots \\ \epsilon_n [\lambda^T X^T \epsilon] \end{bmatrix} \end{aligned}$$

$$= (a^T - \lambda^T X^T) \mathbb{E} \begin{bmatrix} \epsilon_1 \sum_{j=1}^n (\lambda^T X^T)_j \epsilon_j \\ \vdots \\ \epsilon_n \sum_{j=1}^n (\lambda^T X^T)_j \epsilon_j \end{bmatrix}$$

Using independence of  $\epsilon_i$  and  $\epsilon_j$  for  $i \neq j$ , we can write this as

$$\begin{aligned} &= (a^T - \lambda^T X^T) \mathbb{E} \begin{bmatrix} \epsilon_1 (\lambda^T X^T)_1 \epsilon_1 \\ \vdots \\ \epsilon_n (\lambda^T X^T)_n \epsilon_n \end{bmatrix} = (a^T - \lambda^T X^T) \sigma^2 \begin{bmatrix} (\lambda^T X^T)_1 \\ \vdots \\ (\lambda^T X^T)_n \end{bmatrix} = \sigma^2 (a^T - \lambda^T X^T) X \lambda \\ &= \sigma^2 (a^T X - \lambda^T X^T X) \lambda = 0 \end{aligned}$$

since by (11.7)  $c^T = a^T X = \lambda^T X^T X$ .

□

**Theorem 11.2.3 (Gauss-Markov Theorem, as stated in Faraway [2002], more general case).**  
Suppose

$$y = X\beta + \epsilon$$

with  $X \in \mathbb{R}^{n \times p}$  a fixed full rank matrix and  $n \geq p$ ,  $\beta \in \mathbb{R}^p$ , and  $\epsilon \in \mathbb{R}^n$  with  $\mathbb{E}(\epsilon) = 0$  and  $\text{Var}(\epsilon) = \sigma^2 I_n$ . Suppose  $X$  is fixed ( $\mathbb{E}(Y) = X\beta$ ). Let  $\psi : \mathbb{R}^p \rightarrow \mathbb{R}^k$  be an estimable function:  $\psi := c^T \beta$  for some full rank  $c \in \mathbb{R}^{p \times k}$ . Then in the class of all unbiased linear estimates of  $\psi$ ,  $\hat{\psi} = c^T \hat{\beta}$  has the minimum variance and is unique. That is, for some  $a \in \mathbb{R}^{n \times k}$  such that  $a^T y$  is another unbiased estimator of  $c^T \beta$ ,

$$\text{Var}(a^T y) \succeq \text{Var}(c^T \hat{\beta}).$$

*Proof (adapted from Faraway [2002]).* First, note that

$$\mathbb{E}(a^T y) = a^T X \beta = c^T \beta, \quad \forall \beta \in \mathbb{R}^p.$$

This implies

$$a^T X = c^T \iff X^T a = c, \tag{11.9}$$

Since  $X$  has full column rank,  $X^T X$  is full rank and its column space is all of  $\mathbb{R}^p$ , so there exists a unique  $\lambda \in \mathbb{R}^{p \times k}$  such that

$$c = X^T X \lambda = X^T a \tag{11.10}$$

(Note that  $a = X\lambda + b$  for some  $b \in \mathbb{R}^{n \times k}$  whose columns lie in the  $(n - p)$ -dimensional nullspace of  $X^T$ .) Then

$$\begin{aligned} \text{Var}(a^T y) &= \text{Var}(a^T y - c^T \hat{\beta} + c^T \hat{\beta}) = \text{Var}(a^T y - \lambda^T X^T \hat{y} + c^T \hat{\beta}) \\ &= \text{Var}(a^T y - \lambda^T X^T \hat{y}) + \text{Var}(c^T \hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}) \succeq \text{Var}(c^T \hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}), \end{aligned} \quad (11.11)$$

so we are done if the covariance matrix in (11.11) is positive semidefinite.

$$\begin{aligned} \text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}) &= \mathbb{E} \left[ (a^T y - \lambda^T X^T \hat{y} - \mathbb{E}[a^T y - \lambda^T X^T \hat{y}]) (c^T \hat{\beta} - \mathbb{E}[c^T \hat{\beta}])^T \right] \\ &= \mathbb{E} \left[ (a^T (X\beta + \epsilon) - \lambda^T X^T X \hat{\beta} - a^T X\beta + \lambda^T X^T X \beta) (c^T [\hat{\beta} - \beta])^T \right] \\ &= \mathbb{E} \left[ (a^T \epsilon - \lambda^T X^T X (\hat{\beta} - \beta)) (\hat{\beta} - \beta)^T c \right] = \mathbb{E} [(a^T \epsilon - \lambda^T X^T X (X^T X)^{-1} X^T \epsilon) \epsilon^T X (X^T X)^{-1} c] \\ &= \mathbb{E} [a^T \epsilon \epsilon^T X (X^T X)^{-1} c] - \mathbb{E} [\lambda^T X^T \epsilon \epsilon^T X (X^T X)^{-1} c] \\ &= \sigma^2 a^T I_n X (X^T X)^{-1} c - \sigma^2 \lambda^T X^T I_n X (X^T X)^{-1} c = \sigma^2 (a^T X - \lambda^T X^T X) \lambda = 0 \end{aligned}$$

since by (11.10)  $c^T = a^T X = \lambda^T X^T X$ .

□

### 11.3 Chapter 3: Hypothesis testing in regression

In this section, I borrow from C. Flinn's notes "Asymptotic Results for the Linear Regression Model," available online at <http://www.econ.nyu.edu/user/flinnc/notes1.pdf>.

**Lemma 11.3.1.**

$$\frac{1}{n} \cdot X' \epsilon \xrightarrow{p} 0$$

*Proof.* Note that  $\mathbb{E} \frac{1}{n} \cdot X' \epsilon = 0$  for any  $n$ . Then we have

$$\text{Var}\left(\frac{1}{n} \cdot X' \epsilon\right) = \mathbb{E}\left(\frac{1}{n} \cdot X' \epsilon\right)^2 = n^{-2} \mathbb{E}(X' \epsilon \epsilon' X) = n^{-2} \mathbb{E}(\epsilon \epsilon') X' X = \frac{\sigma^2}{n} \frac{X' X}{n}$$

implying that  $\lim_{n \rightarrow \infty} \text{Var}\left(\frac{1}{n} \cdot X' \epsilon\right) = 0$ . Therefore the result follows from Chebyshev's Inequality (Theorem 8.2.2). □

**Lemma 11.3.2.** If  $\epsilon$  is i.i.d. with  $E(\epsilon_i) = 0$  and  $\mathbb{E}(\epsilon_i^2) = \sigma^2$  for all  $i$ , the elements of the matrix  $X$  are uniformly bounded so that  $|X_{ij}| < U$  for all  $i$  and  $j$  and for  $U$  finite, and  $\lim_{n \rightarrow \infty} X'X/n = Q$  is finite and nonsingular, then

$$\frac{1}{\sqrt{n}} X' \epsilon \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q)$$

*Proof.* If we have one regressor, then  $n^{-1/2} \sum_{i=1}^n X_i \epsilon_i$  is a scalar. Let  $G_i$  be the cdf of  $X_i \epsilon_i$ . Let

$$S_n^2 = \sum_{i=1}^n \text{Var}(X_i \epsilon_i) = \sigma^2 \sum_{i=1}^n X_i^2$$

In this scalar case,  $Q = \lim_{n \rightarrow \infty} n^{-1} \sum_i X_i^2$ . By the Lindeberg-Feller Theorem, a necessary and sufficient condition for  $Z_n \rightarrow \mathcal{N}(0 \sigma^2 Q)$  is

$$\lim_{n \rightarrow \infty} \frac{1}{S_n^2} \sum_{i=1}^n \int_{|\omega| > \nu S_n} \omega^2 dG_i(\omega) = 0$$

for all  $\nu > 0$ . Now  $G_i(\omega) = F(\omega/|X_i|)$ . Then rewrite the above equation as

$$\lim_{n \rightarrow \infty} \frac{n}{S_n^2} \sum_{i=1}^n \frac{X_i^2}{n} \int_{|\omega/X_i| > \nu S_n/|X_i|} \left( \frac{\omega}{X_i} \right)^2 dF(\omega/|X_i|) = 0$$

Since  $\lim_{n \rightarrow \infty} S_n^2 = \lim_{n \rightarrow \infty} n \sigma^2 \sum_{i=1}^n X_i^2 / n = n \sigma^2 Q$ , we have  $\lim_{n \rightarrow \infty} n/S_n^2 = (\sigma^2 Q)^{-1}$ , which is a finite and nonzero scalar. Then we need to show

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i^2 \delta_{i,n} = 0$$

where

$$\delta_{i,n} = \int_{|\omega/X_i| > \nu S_n/|X_i|} \left( \frac{\omega}{X_i} \right)^2 dF(\omega/|X_i|)$$

But  $\lim_{n \rightarrow \infty} \delta_{i,n} = 0$  for all  $i$  and any fixed  $\nu$  since  $|X_i|$  is bounded while  $\lim_{n \rightarrow \infty} X_n = \infty$ , so the measure of the set  $\{|\omega/X_i| > \nu S_n/|X_i|\}$  goes to 0 asymptotically. Since  $\lim_{n \rightarrow \infty} n^{-1} \sum_i X_i^2$  is finite and  $\lim_{n \rightarrow \infty} \delta_{i,n} = 0$  for all  $i$ ,  $\lim_{n \rightarrow \infty} n^{-1} \sum_i X_i^2 \delta_{i,n} = 0$ , so  $\frac{1}{n} \cdot X' \epsilon \xrightarrow{P} 0$ .

□

**Theorem 11.3.3.** Under the conditions of Lemma 11.3.2 ( $\epsilon$  is i.i.d. with  $E(\epsilon_i) = 0$  and  $\mathbb{E}(\epsilon_i^2) = \sigma^2$  for all  $i$ , the elements of the matrix  $X$  are uniformly bounded so that  $|X_{ij}| < U$  for all  $i$  and  $j$  and for  $U$  finite, and  $\lim_{n \rightarrow \infty} X'X/n = Q$  is finite and nonsingular),

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1})$$

*Proof.*

$$\sqrt{n}(\hat{\beta} - \beta) = \left( \frac{X'X}{n} \right)^{-1} \frac{1}{\sqrt{n}} X' \epsilon$$

Since  $\lim_{n \rightarrow \infty} (X'X/n)^{-1} = Q^{-1}$  and by Lemma 11.3.2

$$\frac{1}{\sqrt{n}} X' \epsilon \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q)$$

then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1} Q Q^{-1}) = \mathcal{N}(0, \sigma^2 Q^{-1})$$

□

*t*-test statistic:

$$t = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

*F*-test statistic:

$$F = \left( \frac{T - k - 1}{r} \right) \left( \frac{SSR_R - SSR_U}{SSR_U} \right)$$

Since

$$R^2 = \frac{\sum_t (y_t - \bar{y})^2 - \sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \bar{y})^2} = \frac{\sum_t (y_t - \bar{y})^2 - SSR_U}{\sum_t (y_t - \bar{y})^2}$$

we have

$$SSR_U = \sum_t (y_t - \bar{y})^2 - R^2 \sum_t (y_t - \bar{y})^2 = (1 - R^2) \sum_t (y_t - \bar{y})^2$$

yielding

$$F = \left( \frac{T - k - 1}{r} \right) \left( \frac{\sum_t (y_t - \bar{y})^2 - (1 - R^2) \sum_t (y_t - \bar{y})^2}{(1 - R^2) \sum_t (y_t - \bar{y})^2} \right) = \left( \frac{T - k - 1}{r} \right) \left( \frac{R^2}{1 - R^2} \right)$$

**Confidence interval for sums of coefficients.** (Two coefficient case.) Suppose we want to test  $H_0 : \beta_1 + \beta_2 = k$ . Let  $\delta = \beta_1 + \beta_2 - k$ ,  $\hat{\delta} = \hat{\beta}_1 + \hat{\beta}_2 - k$ . Note that under the null hypothesis  $\delta = 0$ . We can construct a *t*-statistic

$$t_{\hat{\delta}} = \frac{\hat{\delta} - 0}{\sqrt{\text{Var}(\hat{\delta})}} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - k}{\sqrt{\text{Var}(\hat{\delta})}}$$

where

$$\hat{\text{Var}}(\hat{\delta}) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) + 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$

This means that a 95% confidence interval for  $\delta$  can be constructed in the following way:

$$\hat{\delta} \pm t^* \sqrt{\hat{\text{Var}}(\hat{\delta})}$$

where  $t^*$  is the 95% critical value for the  $t$ -distribution.

### 11.3.1 ANOVA

$$g_1 = lm(R \sim x_1 + x_2 + x_3 + x_4), \quad g_2 = lm(R \sim x_2 + x_3),$$

$$y = \beta_0 + \sum_{i=1}^4 \beta_i x_i$$

$$H_0 : \beta_1 = \beta_4 = 0, \quad H_A : \text{at least one of } \beta_1, \beta_4 \text{ is not zero.}$$

## 11.4 Chapter 4: Heteroskedasticity

Under heteroskedasticity, the OLS estimator  $\hat{\beta} = (X'X)^{-1}X'y$  is unbiased, but the true covariance matrix of  $\hat{\beta}$  no longer matches the OLS formula. For instance, suppose we have

$$y_t = \sum_{i=1}^K \beta_i x_{ti} + u_t$$

where  $\text{Var}(u_t) = \sigma^2 z_t^2$ .

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \beta + (X'X)^{-1}X'u$$

$$\implies \mathbb{E}(\hat{\beta}) = \mathbb{E}[\beta] + (X'X)^{-1}X'\mathbb{E}[u] = \beta$$

since  $\mathbb{E}(u)$  is still 0. However,

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \mathbb{E}[(\hat{\beta} - \mathbb{E}(\hat{\beta}))(\hat{\beta} - \mathbb{E}(\hat{\beta}))'] = \mathbb{E}[(\beta + (X'X)^{-1}X'u - \beta)(\beta + (X'X)^{-1}X'u - \beta)'] \\ &= \mathbb{E}[(X'X)^{-1}X'u((X'X)^{-1}X'u)'] = \mathbb{E}[(X'X)^{-1}X'u u' X ((X'X)^{-1})'] \end{aligned}$$

$$\begin{aligned}
&= (X'X)^{-1} X' \mathbb{E}[uu' | X] X (X'X)^{-1} \\
&= (X'X)^{-1} X' \begin{bmatrix} \sigma^2 z_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 z_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 z_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 z_T^2 \end{bmatrix} X (X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1} X' \begin{bmatrix} z_1^2 & 0 & 0 & \dots & 0 \\ 0 & z_2^2 & 0 & \dots & 0 \\ 0 & 0 & z_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & z_T^2 \end{bmatrix} X (X'X)^{-1}
\end{aligned}$$

which is different from the OLS estimator of the covariance matrix  $\sigma^2(X'X)^{-1}$ . Therefore the estimate of the variances of  $\hat{\beta}$  will be biased if the OLS formulas are used, and the usual  $t$  and  $F$  tests for  $\hat{\beta}$  will be invalid.

## 11.5 Chapter 5: Autocorrelated disturbances

**Generalized least squares model:**

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where

$$\mathbb{E}(\mathbf{u} | \mathbf{X}) = 0 \quad \forall t$$

$$\mathbb{E}(\mathbf{u}\mathbf{u}' | \mathbf{X}) = \boldsymbol{\Sigma}$$

where  $\boldsymbol{\Sigma}$  is a positive definite matrix.

Suppose

$$y = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

where  $\epsilon \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ . Then

$$\boldsymbol{\Sigma}^{-1/2}y = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \tilde{\epsilon}$$

where  $\tilde{\epsilon} = \boldsymbol{\Sigma}^{-1/2}\epsilon \sim \mathcal{N}(0, 1)$ . Now we can do ordinary least squares since the errors of this transformed model are i.i.d.

$$\hat{\beta}_{GLS} = \left( \left[ \Sigma^{-1/2} X \right]^T \Sigma^{-1/2} X \right)^{-1} \left[ \Sigma^{-1/2} X \right]^T \Sigma^{-1/2} y = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y$$

$$\text{Var}(\hat{\beta}_{GLS}) = (X' \Sigma^{-1} X)^{-1}$$

Example application: spatial data.

## 11.6 Quantile Regression

Estimate the conditional *median* rather than the conditional mean (as in least squares). Least absolute deviation:

$$\arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n |y_i - x'_i \beta|$$

Problems: no closed form solution; have to solve by linear programming. Good: robust; less changed by fluctuations of outliers. Asymmetric loss:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (1 - \tau) \underbrace{(y_i - x'_i \beta)_+}_{\text{under-estimation}} + \tau \underbrace{(x'_i \beta - y_i)_+}_{\text{over-estimation}}$$

In fixed dimensions:

$$\hat{\beta}(\tau) \xrightarrow{d} \mathcal{N} \left( \beta, (X^T X)^{-1} \frac{\tau(1 - \tau)}{f_\epsilon^2(0)} \right)$$

Suppose  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ ; then  $f_\epsilon(0) = 1/\sqrt{2\pi\sigma^2}$ , so we have

$$\hat{\beta}(\tau) \xrightarrow{d} \mathcal{N} \left( \beta, 2\pi\sigma^2 (X^T X)^{-1} \tau(1 - \tau) \right)$$

This is then maximized for  $\tau = 1/2$ , and the max value is  $1/4 \cdot 2\pi = \pi/2$ . For more information on other loss functions, see also Section 11.14.

### 11.6.1 Detecting outliers in multiple dimensions

Hard because of curse of dimensionality. One thing: project onto lower-dimensional space and then compute distance there. multi-dimensional scaling or dimension reduction methods.

1. random projections
2. nonlinear methods: Iso-map, local linear embedding

Half-space depth (Tukey): make polygons of observations (outer one is the convex hull). Then create convex hulls of inside, keep going. See Figure 6.

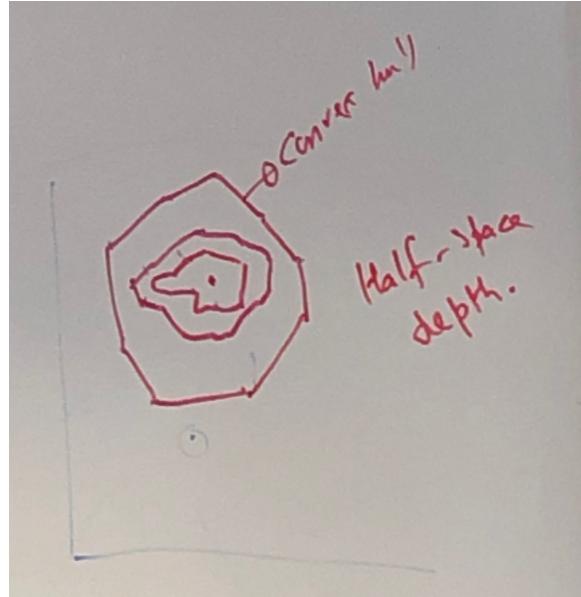


Figure 6: Illustration of half-space depth method for detecting outliers.

## 11.7 Transformed Linear Models

### 11.7.1 Transformations of response

Consider the transformation

$$t(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y) = \lim_{\lambda \rightarrow 0} \frac{y^\lambda - 1}{\lambda}, & \lambda = 0 \end{cases}$$

One example: **log linear model**.

$$\log y_i = \beta_0 + \beta_1 x_i + \epsilon.$$

Multiplicative model in the true response:

$$y_i = \exp(\beta_0) \cdot \exp(x_1 \beta_1) \cdot \exp(\epsilon_i)$$

If you increase  $y_i$  by one unit, then  $y_i$  is multiplied by  $\exp(\beta_1)$ . So in this case rather than talking about linear effects, you talk about percentage changes in the response:  $(e^{\beta_1} - 1) \cdot 100\%$ .

Square root transformations work well in Poisson models—variance stabilizes.

Consider

$$t_\lambda = x' \beta + \epsilon.$$

Then we have

$$SSE = (t_\lambda - x'_i \hat{\beta}_{OLS})^T (t_\lambda - x'_i \hat{\beta}_{OLS})$$

Try

$$\frac{n}{2} \log \left( \frac{SSE}{n} \right) + \left( \sum_{i=1}^n \log(y_i) \right).$$

for  $\lambda = -2, -2, -1/2, 0, 1/2, 1, 2$ . We have  $2L(\hat{\lambda}) - 2L(\lambda_{true}) \sim \chi_1^2$ . If 1 is inside the 95% confidence interval, don't transform; if 1 isn't, do. In a lot of cases, changing the response can be a pain for interpretability.

### 11.7.2 Transforming predictor values

Add polynomial terms:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d + \epsilon.$$

Want to maintain hierarchy in selecting variables (backward elimination). Advantages: nonlinear curvature. Global: all data have influence on every predicted point.

Broken stick regression/segmented regression: useful if data has two different groups.

$$y = B_\ell(x) + B_r(x) = \beta_0 + \beta_1(c - x)_+ + \beta_2(x - c)_+$$

where  $c$  is the breaking point and is fixed (if you use data to choose  $c$ , becomes a nonlinear problem). Advantages: more localized

Splines: like combination of broken stick regression and polynomial regression. fit a polynomial in each segment.

## 11.8 Segmented regression, local regression, splines

for splines: knot selection is important (bias/variance tradeoff)

### 11.8.1 Local Regression

Local regression: weight chosen by kernel:

$$w_i = \frac{K((x_i - x_0)/\sqrt{v})}{\sum_{i=1}^{M(s)} K((x_i - x_0)/\sqrt{v})}$$

$$\text{dnorm}(x_i, \text{mean} = x_0; \text{sd} = \sqrt{v})$$

$$K((x_i - x_0)/\sqrt{v}) = \phi((x_i - x_0)/\sqrt{v}) \cdot v^{-1/2} = \frac{1}{\sqrt{2\pi v}} \exp - \left( \frac{(x_i - x_0)^2}{2v} \right)$$

so if the distance between two points increases than the weight decreases. This is called kernel density with a Gaussian kernel.

$$x = x_0; \hat{\beta}(x_0); \hat{\sigma}(x_0)$$

Two tuning parameters:  $v$  (bandwidth) and  $s$ . Disadvantages: hard to interpret, starts to do poorly in high dimensions.

### 11.8.2 Curse of Dimensionality (brief discussion)

Suppose  $x_i : i \in [n] \sim F$  i.i.d. with fixed dimension, and  $F$  has bounded support  $[a, b]^d$ . Let  $x \in \text{supp}(F)$ . Then

$$\min_x d_2(x, x_i) \sim \frac{1}{n^{1/d}}$$

so goes to 0 as  $n \rightarrow \infty$  for fixed  $d$ . But bad in  $d$ ; for example, if  $d = 5$ , need  $n = 10^5$  for  $d = 0.1$ . As this minimum distance increases, regularity becomes more difficult.

## 11.9 Dimension Reduction methods

### 11.9.1 Principal components regression

(See Section 2.4.2 for more details on principal components.) Given data  $(y, X)$ , don't look at  $y$  for now. Look for maximum variability direction of  $X$ :

$$P_1 = \arg \max_{\|a\|_2=1} \text{Var}(a^T x) = a^T (X^T X) a.$$

then (letting  $\mathcal{P}_1$  be the projection matrix for projecting onto  $P_1$ )

$$P_2 = \arg \max_{\|a\|_2=1, \mathcal{P}_1 a=0} \text{Var}(a^T x) = a^T (X^T X) a.$$

and so on. We will regress  $y$  against the first  $M$  principal components of  $X$  for some  $X \leq p$ , where the principal components are the columns of  $U$ :

$$\hat{y}_{(M)}^{\text{PCR}} = \bar{y} + \sum_{m=1}^M \hat{\theta}_m z_m,$$

where  $\hat{\theta}_m = (z_m^T y) / (z_m^T z_m)$  since all the  $z_m$  are orthogonal. Also, since the  $z_m$  are linear combinations of the original  $x_j$ , the solution can be expressed in terms of coefficients in the original feature space:

$$\hat{\beta}^{\text{PCR}}(M) = \sum_{m=1}^M \hat{\theta}_m v_m,$$

Works well when high variability directions in the explanatory variables are also interesting attributes to study.

Sometimes if data are in a manifold that isn't linearly parsed, good to use an ISOMAP or LLE (local linear embedding).

(read: starting on p.63 of ESL)

### 11.9.2 Partial least squares

Does look at  $y$ , unlike PCR.  $X_{n \times p} \rightarrow W_{n \times 3}$ . Start by standardizing all  $x_j$  to have mean 0 and unit variance. Then compute  $\hat{\phi}_{1j} = \langle x_j, y \rangle$  for each  $j$ . Then the derived input  $z_1 = \sum_j \hat{\phi}_{1j} x_j$  is the first partial least squares direction. Then  $y$  is regressed on  $z_1$  giving coefficient  $\hat{\theta}_1$ , and then  $x_1, \dots, x_j$  are orthogonalized with respect to  $z_1$ . Continue until  $M \leq p$  directions have been obtained. (see p. 80 of ESL)

⋮

notes from GSBA 604: First reduced dimension:

$$w_{n \times 1}^{(1)} = X_{n \times 1}^{(1)} + X_{n \times 1}^{(2)} + \dots + X_{n \times 1}^{(p)}$$

proportional correlation  $(y, x^{(1)})$  to

### 11.9.3 Dimension reduction by random matrix

Let  $R \in \mathbb{R}^{p \times d}$ ,  $d < p$ , with  $R_{ij} \sim \mathcal{N}(0, 1)$ . Let  $\tilde{X} = XR$ . By the Johnson-Lindenstrauss lemma, for any  $\epsilon > 0$  there exists an  $R$  such that

$$(1 - \epsilon) \|\tilde{X}_i - \tilde{X}_j\|_2^2 \leq \|X_i - X_j\|_2^2 \leq (1 + \epsilon) \|\tilde{X}_i - \tilde{X}_j\|_2^2.$$

(isometric transformation—distances are maintained)

## 11.10 Goodness of fit, residuals, residual diagnostics, leverage

Goodness of fit:  $F$  test. Assume  $\text{Var}(\epsilon) = \sigma^2 I$  and recall  $\hat{\epsilon} = (I - H)\epsilon$ . So  $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H)$  and  $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - H_{ii})$ . So  $h_i := H_{ii}$  is called the **leverage** for the  $i$ th case. Some properties: if  $X$  is **ill-conditioned** (condition number—ratio of largest to smallest eigenvalue of Gram matrix  $X^T X$ —is high), then  $H_{ii}$  will vary a lot.

Properties:  $\sum_i h_i = p + 1$  (because  $H_{ii}$  is idempotent, its trace is equal to  $p + 1$ —all its eigenvalues are 0 or 1, and it is nonsingular, so all of its eigenvalues are 1). And, all  $h_i \geq 1/n$ .

Larger  $h_i$  result in smaller  $\text{Var}(\hat{\epsilon}_i)$ , which forces the fit to be close to  $y_i$ . Average leverage:  $(p + 1)/n$ . Rule of thumb: leverages of more than  $2(p + 1)/n$  should be looked at closely (they have a large influence on the slope, so if they are incorrect then it's a big problem for the fit).

Standardized or Studentized results:

$$r_i = \frac{\hat{\epsilon}_i}{\text{se}(\hat{\epsilon}_i)} = \frac{\hat{\epsilon}_i}{\hat{\sigma} \sqrt{1 - h_i}}$$

Let  $\hat{\beta}_i$  and  $\hat{\sigma}_i^2$  be the estimates from the regression with the  $i$ th case excluded. Let  $x_i^T$  denote the  $i$ th row of  $X$ , write  $X_{(i)}$  for  $X$  without the  $i$ th row.

⋮

Outlier test: can test for outliers using the fact that each  $t_i \sim t(n - p - 2)$  if no outliers are present. Need to do a multiplicity correction for multiple testing (say Bonferroni correction). have to do residual diagnostics.

### 11.10.1 Residual diagnostics

Consider leverage points (outliers, influential points: points that change the slope a lot).

1. **Partial residual plot:**  $k$ th variable.  $\hat{\epsilon}^{(k)} = y - \sum_{j \neq k} \beta_j x_j$ : residuals without the  $k$ th explanatory variable. Plot this against  $x_k$  and regress. If the resulting slope  $\hat{\beta}_k$  is large, this is an influential point.
2. **Added variable plot:** Regress  $x_k$  by other explanatory variables  $\hat{\delta}^{(k)}$ . Plot  $\hat{\epsilon}^{(k)}$  against  $\hat{\delta}^{(k)}$ .

## 11.11 DSO 607

Generalized linear models:

$$f_n(z, \beta) = \prod_{i=1}^n \exp [\theta_i z_i - b(\theta_i) h(z_i)], \quad z = (z_1, \dots, z_n)^T$$

Natural parameter  $\theta_i$ :  $\theta_i = x_i^T \beta$ ,  $x_i = \{x_{ij} : j \in \mathcal{M}\}$

$h(z_i)$ : normalization constant

linear regression:  $b(\theta) = \frac{1}{2}\theta^2$

other:  $b(\theta) = \log(1 + e^\theta)$

If  $Y = (Y_1, \dots, Y_n)^T \sim F_n(\cdot, \beta)$ , then  $\mathbb{E}(Y) = (b'(\theta_1), \dots, b'(\theta_n))^T = \mu(\theta)$  and

$\text{Cov}(Y) = \text{diag}\{b''(\theta_1), \dots, b''(\theta_n)\} = \Sigma(\theta)$  where  $\theta = X\beta$  and  $X = (x_1, \dots, x_n)^T$  is the  $n \times d$  design matrix.

Quasi-log-likelihood (“quasi” because error may be misspecified):

$$\ell_n(y, \beta) = y^T X\beta - \mathbf{1}^T b(X\beta) + \mathbf{1}^T h(y)$$

Like MLE, maximizing  $\ell_n(y, \beta)$  with respect to  $\beta$  gives the quasi-MLE  $\hat{\beta}_n$ . Solution exists and is unique due to strict convexity of  $b$ , solves the score equation

$$\frac{\partial \ell_n(y, \beta)}{\partial \beta} = x^T [y - \mu(X\beta)] = \mathbf{0}$$

(Intuition of score equation: the columns of  $X$  are all orthogonal to the errors (uncorrelated if  $X$  is random)).

### 11.11.1 Akaike Information Criterion (AIC)

AIC: proposed by [Akaike \[1973\]](#) to choose a model by minimizing the Kullback-Leibler (KL) divergence of the fitted model from the true model (or equivalently, maximize the expected log-likelihood). Recall the KL Divergence

$$I(\theta; \theta_0) := 2\mathbb{E}_{\theta_0} [\log(f(X | \theta_0))] - 2\mathbb{E}_{\theta_0} [\log(f(X | \theta))].$$

We will try to maximizing the KL Divergence by estimating  $\theta_0$  as best as we can by maximizing the **probabilistic negentropy**

$$\mathbb{E}_Z I(\theta; \hat{\theta}_0(Z)) := 2\mathbb{E}_{\theta_0} [\log(f(X | \theta_0))] - 2\mathbb{E}_{\theta_0, Z} [\log(f(X | \hat{\theta}_0(Z)))] .$$

Because the true model  $\theta_0$  is unknown we cannot carry out this maximization directly. Note that as the number of independent observations increases, the **mean log-likelihood ratio**

$$\hat{I}(\theta; \theta_0) := \frac{2}{n} \sum_{i=1}^n \log \frac{f(x_i | \theta_0)}{f(x_i | \theta)} \xrightarrow{P} I(\theta; \theta_0).$$

Because of this, Akaike reasons that maximizing the mean log-likelihood ratio over  $\theta_0$  (i.e. computing the maximum likelihood estimate) tend to maximize the entropy. So the maximum likelihood estimate  $\hat{\theta}_0(Z)$  is substituted for the unknown  $\theta_0$ .

Way we wrote KL Divergence in DSO 607: density  $f$  from density  $g$ :

$$I(g_n; f_n(\cdot, \beta)) = \int [\log g(z)]g(z)dz - \int [\log f(z)]g(z)dz$$

[Akaike \[1973\]](#) found that up to an additive constant, the KL divergence of the fitted model from the true model can be asymptotically expanded as

$$-\ell_n(\hat{\theta}) + \lambda \dim(\hat{\theta}) = -\ell_n(\hat{\theta}) + \lambda \sum_{j=1}^p \mathbf{1}_{\{\hat{\theta}_j \neq 0\}}$$

where  $\ell_n(\theta)$  is the log-likelihood function and  $\lambda = 1$ . This leads to the Akaike information criterion (AIC) for comparing models:

$$AIC(\hat{\theta}_k(Z)) := n\hat{I}(\hat{\theta}_k(Z; \hat{\theta}_0(Z))) + 2\|\hat{\theta}_k(Z)\|_0 = 2 \sum_{i=1}^n \log \frac{f(x_i | \hat{\theta}_0(Z))}{f(x_i | \hat{\theta}_k(Z))} + 2\|\hat{\theta}_k(Z)\|_0$$

Way we wrote this is DSO 607:

$$AIC(\hat{\theta}) := -2\ell_n(\hat{\theta}) + 2\|\hat{\theta}\|_0$$

Intuition:  $\log g(x)$  is the log likelihood. Penalty term can be interpreted as penalty, or as a bias correction since you are doing training and feature selection simultaneously on the same data.

$$I(g_n; f_n(\cdot, \beta)) = \sum_{i=1}^n \left[ \int \right]$$

To minimize the KL divergence

$$\frac{\partial I(g_n; f_n(\cdot, \beta))}{\partial \beta} = -X^T[\mathbb{E}(Y) - \mu(X\beta)] = 0$$

the inverse of the Fisher information matrix is the covariance of the MLE (?).

⋮

(For more information on KL Divergence, see Sections 6.5 and 10.4.6). For AIC, we minimize the KL divergence. For BIC, we maximize the Bayes factor (posterior probability for the model).

### 11.11.2 Bayesian Information Criterion (BIC)

A typical Bayesian model selection procedure is to first give nonzero prior probability  $\alpha_M$  on each model  $M$  and then prescribe a prior distribution  $\mu_M$  for the parameter vector in the corresponding model. The Bayesian principle of model selection is to choose the most probable model *a posteriori*; that is, to choose a model that maximizes the log-marginal likelihood (or the Bayes factor)

$$\log \int \alpha_M \exp[\ell_n(\theta)] d\mu_m(\theta).$$

Schwarz [1978] took a Bayesian approach with prior distributions that have nonzero prior probabilities on some lower dimensional subspaces of  $\mathbb{R}^p$  and showed that the negative log-marginal likelihood can be asymptotically expanded as

$$-\ell_n(\hat{\theta}) + \lambda \|\hat{\theta}\|_0$$

where  $\lambda = (\log n)/2$ . This asymptotic expansion leads to the Bayesian information criterion (BIC) for comparing models:

$$BIC(\hat{\theta}) := -2 \log \left( f(x | \hat{\theta}; \hat{\theta}_{MLE}) \right) + (\log n) \|\hat{\theta}\|_0.$$

where  $f$  is the density function parameterized by  $\hat{\theta}_{MLE}$ , the maximum likelihood estimate for the density given the data  $x$ .

Way we wrote this in DSO 607:

$$BIC(\hat{\theta}) := -2\ell_n(\hat{\theta}) + (\log n) \|\hat{\theta}\|_0.$$

⋮

$$B_n^{1/2} A_n (\hat{\beta}_n - \beta_{n,0}) = W_n \xrightarrow{D} \mathcal{N}(0, I_d)$$

$$\hat{\beta}_n - \beta_{n,0} = A_n^{-1} B_n^{1/2} W_n \implies \text{Cov}(\hat{\beta}_n) = \text{Cov}(\hat{\beta} - n - \beta_{n,0})$$

$$= \text{Cov}(A_n^{-1} B^{1/2} W_n) = A_n^{-1} B_n^{1/2} \text{Cov}(W_n) B_n^{1/2} A_n^{-1} = A_n^{-1} B_n^{1/2} I_d B_n^{1/2} A_n^{-1} = \boxed{A_n^{-1} B_n A_n^{-1}}$$

Note that if the model is correct,  $A_n = B_n$  so this reduces to conventional asymptotic MLE theory ( $\text{Cov}(\hat{\beta}_n) = A_n^{-1}$ ).

⋮

$A_n$  from working model,  $B_n$  from true model (unknown).

GBIC in misspecified models:  $H_n = A_n^{-1} B_n$  (covariance contrast matrix). Note that when model is specified,  $H_n = I_d$  so the log of its determinant is 0 so it vanishes. If not, then it is a misspecification penalty.

⋮

Note:  $\log(y, \hat{\beta}_n) > \log(y, \beta_{n,0})$  because  $\hat{\beta}_n$  is by definition the MLE on the observed data. But  $\mathbb{E}(\log(\tilde{y}, \beta_{n,0})) > \mathbb{E}(\log(\tilde{y}, \hat{\beta}_n))$  because  $\beta_{n,0}$  is the true parameter. We have a systematic upward bias when we use the empirical estimate. (p.18 of week 2-2 slides)

**Proposition 11.11.1 (Result from “Econometrics: Methods and Applications” homework).**

Consider the usual linear model, where  $y = X\beta + \epsilon$ . Suppose we compare two regressions, which differ in how many variables are included in the matrix  $X$ . In the full (unrestricted) model  $p_1$  regressors are included. In the restricted model only a subset of  $p_0 < p_1$  regressors are included. Then for large  $n$ , selection based on AIC corresponds to an  $F$ -test with a critical value of approximately 2.

*Proof.* Let  $e_R$  be the vector of residuals for the restricted model with  $p_0$  parameters and  $e_U$  the vector of residuals for the full unrestricted model with  $p_1$  parameters. Then we have the sample standard deviations

$$s_0^2 = \frac{1}{n - p_0} e_R' e_R, s_1^2 = \frac{1}{n - p_1} e_U' e_U \quad (11.12)$$

Recall the AIC:

$$\log(s^2) + \frac{2k}{n}$$

where  $k$  is the number of regressors included in the model.

For the small model, we have

$$AIC_0 = \log(s_0^2) + \frac{2p_0}{n}.$$

For the big model, we have

$$AIC_1 = \log(s_1^2) + \frac{2p_1}{n}.$$

Therefore the smallest model is preferred according to the AIC if

$$AIC_0 < AIC_1$$

$$\begin{aligned} \iff \log(s_0^2) + \frac{2p_0}{n} < \log(s_1^2) + \frac{2p_1}{n} \iff \log(s_0^2) - \log(s_1^2) < \frac{2p_1}{n} - \frac{2p_0}{n} \iff \log\left(\frac{s_0^2}{s_1^2}\right) < \frac{2}{n}(p_1 - p_0) \\ \iff \frac{s_0^2}{s_1^2} < e^{\frac{2}{n}(p_1 - p_0)} \end{aligned} \quad (11.13)$$

If  $n$  is very large,  $\frac{2}{n}(p_1 - p_0)$  is small. Therefore, using the first order Taylor approximation  $e^x \approx 1 + x$  we can approximate that

$$e^{\frac{2}{n}(p_1 - p_0)} \approx 1 + \frac{2}{n}(p_1 - p_0)$$

(if  $n$  is very large.) Substituting this expression into the right side of (11.13) yields

$$\begin{aligned} \frac{s_0^2}{s_1^2} < 1 + \frac{2}{n}(p_1 - p_0) \iff \frac{s_0^2}{s_1^2} - 1 < \frac{2}{n}(p_1 - p_0) \iff \frac{s_0^2}{s_1^2} - \frac{s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0) \\ \iff \frac{s_0^2 - s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0) \end{aligned}$$

for  $n$  very large. Plugging in the expressions from (11.12), we have

$$\frac{\frac{1}{n-p_0}e_R'e_R - \frac{1}{n-p_1}e_U'e_U}{\frac{1}{n-p_1}e_U'e_U} < \frac{2}{n}(p_1 - p_0).$$

For large values of  $n$ ,  $n - p_0 \approx n - p_1 \approx n$ . This yields

$$\begin{aligned} \frac{\frac{1}{n}e_R'e_R - \frac{1}{n}e_U'e_U}{\frac{1}{n}e_U'e_U} &< \frac{2}{n}(p_1 - p_0) \\ = \frac{e_R'e_R - e_U'e_U}{e_U'e_U} &< \frac{2}{n}(p_1 - p_0) \end{aligned} \quad (11.14)$$

Now recall the  $F$  statistic:

$$F = \frac{(e_R'e_R - e_U'e_U)/g}{e_U'e_U/(n-k)} \quad (11.15)$$

where  $k$  is the number of explanatory factors in the unrestricted model, and  $g$  is the number of explanatory factors removed from the unrestricted model to create the restricted model. Under this test, we believe there is significant evidence to suggest that  $\beta \neq 0$  (so the unrestricted model is preferred) if  $F > F_{critical}$ . Therefore a larger model is preferred if  $F > F_{critical}$ , and we stay with (prefer) a smaller model if  $F < F_{critical}$ .

Let  $F_{critical} = 2$ . Then a smaller model is preferred if  $F < 2$ :

$$\frac{(e_R'e_R - e_U'e_U)/g}{e_U'e_U/(n-k)} < 2$$

In this case, with  $p_1$  factors in the unrestricted model and  $p_0$  in the restricted model, we get

$$\frac{(e_R'e_R - e_U'e_U)/(p_1 - p_0)}{e_U'e_U/(n - p_1)} < 2$$

$$\frac{(e_R'e_R - e_U'e_U)}{e_U'e_U} < \frac{2(p_1 - p_0)}{n - p_1}$$

If  $n$  is very large,  $n - p_1 \approx n$ . Substituting this in yields

$$\frac{(e_R'e_R - e_U'e_U)}{e_U'e_U} < \frac{2(p_1 - p_0)}{n} \quad (11.16)$$

which equals (11.14). Our condition for preferring a restricted model when doing an F-test with  $F_{critical} = 2$  (and when  $n$  is very large) is approximately the same as our condition for preferring a restricted model when using the AIC (when  $n$  is very large).

□

## 11.12 Ridge Regression

If  $p$  is large, it tends to be better to shrink the last squares estimator. (Even though this introduces bias, it will likely reduce variance, and the tradeoff will often help for some amount of shrinkage.) This is related to the Stein estimator.

Suppose  $\beta \in \mathbb{R}^p$  is an unknown vector, and for all  $1 \leq i \leq n$ , there are known vectors  $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$ . Our observed data are  $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ . Let  $\mathbf{X}$  be the  $n \times p$  matrix so that the  $i^{th}$  row of  $\mathbf{X}$  is the row vector  $x^{(i)}$ . Assume that  $p \leq n$  and the matrix  $\mathbf{X}$  has full rank. Let  $\lambda > 0$  and consider the quantity

$$\sum_{i=1}^n (y_i - x^{(i)\top} \beta)^2 + \lambda \|\beta\|_2^2 = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad (11.17)$$

The term  $\|\beta\|_2^2$  penalizes  $\beta$  from having large entries. By Lagrange Multipliers, a critical point  $\beta$  of the constrained minimization problem

$$\text{minimize } \sum_{i=1}^n (\mathbf{y}_i - \langle x^{(i)}, \beta \rangle)^2 \quad \text{subject to } \|\beta\|_2^2 \leq 1$$

is equivalent to the existence of a  $\lambda \in \mathbb{R}$  such that  $\beta$  is a critical point of (11.17). We call the  $\hat{\beta}$  that minimizes (11.17) the **ridge regression** estimator for  $\beta$ .

**Proposition 11.12.1 (Math 541A Homework Problem).** The value of  $\hat{\beta} \in \mathbb{R}^p$  that minimizes (11.17) is  $\hat{\beta}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$ .

*Proof.*

$$\sum_{i=1}^n \left( y_i - x^{(i)T} \beta \right)^2 + \lambda \|\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta$$

$$= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\beta - \beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X}\beta + \lambda \beta^T \beta = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\beta + \beta^T \mathbf{X}^T \mathbf{X}\beta + \lambda \beta^T \beta$$

where  $\mathbf{y}^T \mathbf{X}\beta = \beta^T \mathbf{X}^T \mathbf{y}$  because a scalar equals its transpose. Differentiating with respect to  $\beta$  yields

$$-2\mathbf{y}^T \mathbf{X} + 2\beta^T \mathbf{X}^T \mathbf{X} + 2\lambda \beta^T = 0 \iff \beta^T (2\mathbf{X}^T \mathbf{X} + 2\lambda \mathbf{I}_p) = 2\mathbf{y}^T \mathbf{X}$$

$$\iff (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p) \beta = \mathbf{X}^T \mathbf{y} \iff \hat{\beta}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{y}$$

where  $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p$  is invertible by the following argument.  $\mathbf{X}^T \mathbf{X}$  must be positive semidefinite. In fact, it is positive definite because  $\mathbf{X} \in \mathbb{R}^{n \times p}$  has full rank; that is,  $\text{rank}(\mathbf{X}) = p$ , so  $\mathbf{X}^T \mathbf{X} \in \mathbb{R}^{p \times p}$  has rank  $p$  (full rank) and is invertible. So  $\mathbf{X}^T \mathbf{X}$  is positive definite (all positive eigenvalues). Then since  $\text{Tr}(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p) > \text{Tr}(\mathbf{X}^T \mathbf{X})$ , the eigenvalues of  $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p$  are also all positive, which means the determinant of  $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p$  is nonzero, which means it is invertible.

□

**Proposition 11.12.2 (DSO 607 Homework Problem).** Suppose  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  where  $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ .

- (a) The asymptotic behavior of the ridge estimator is as follows: as  $\lambda \rightarrow \infty$ ,  $\hat{\beta}_{\text{ridge}} \rightarrow \mathbf{0}$ , and as  $\lambda \rightarrow 0$ ,  $\hat{\beta}_{\text{ridge}} \rightarrow X^\dagger(\mathbf{X}\beta + \epsilon)$ .
- (b) For any fixed  $\lambda > 0$ , the probability that each component of the ridge estimator  $\hat{\beta}_{\text{ridge}}$  equals 0 is 0.

*Proof.* (a) Since  $X$  is fixed, as  $\lambda \rightarrow \infty$  we have

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y \rightarrow (\lambda I_p)^{-1} X^T y = \frac{1}{\lambda} I_p X^T (X\beta + \epsilon) = \frac{1}{\lambda} (X^T X\beta + X^T \epsilon) \rightarrow \mathbf{0}$$

where  $\mathbf{0}$  is a  $p$ -dimensional vector of zeroes. As  $\lambda \rightarrow 0^+$  we have

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y \rightarrow X^\dagger y = X^\dagger(X\beta + \epsilon)$$

where we substitute the pseudoinverse instead of the inverse because since  $X^T X$  is rank deficient,  $(X^T X)^{-1}$  does not exist (and because the Moore-Penrose pseudoinverse minimizes the  $\ell_2$  norm, exactly what the ridge solution will do).

- (b) We have

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y = (X^T X + \lambda I_p)^{-1} X^T (X\beta + \epsilon) = (X^T X + \lambda I_p)^{-1} X^T X\beta + (X^T X + \lambda I_p)^{-1} X^T \epsilon$$

Let  $e_i$  be a selection vector, with the  $i$ th entry equal to 1 and all other entries equal to 0. Let the  $i$ th entry of  $\hat{\beta}_{\text{ridge}}$  be  $\hat{\beta}_{\text{ridge}}^{(i)} = e_i^T \hat{\beta}_{\text{ridge}}$ . We have

$$\Pr(\hat{\beta}_{\text{ridge}}^{(i)} = 0) = \Pr(e_i^T [(X^T X + \lambda I_p)^{-1} X^T X \beta + (X^T X + \lambda I_p)^{-1} X^T \epsilon] = 0)$$

$$= \Pr(e_i^T (X^T X + \lambda I_p)^{-1} X^T \epsilon = -e_i^T (X^T X + \lambda I_p)^{-1} X^T X \beta)$$

Since every entry of  $\epsilon$  is distributed continuously, the probability of it equaling a particular value is 0. Therefore the probability that each component of the ridge estimator equals 0 is 0. (For an intuitive argument as to why this is, see Figure 7.)

□

**GSBA 604:** using SVD  $X = UDV^T$ , we have

$$\begin{aligned} \hat{y}_{\text{ridge}}(\lambda) &= X \hat{\beta}_{\text{ridge}}(\lambda) = X(X^T X + \lambda I)^{-1} X^T y = UDV^T (VD^2 V^T + \lambda I)^{-1} VDU^T y \\ &= UDV^T (V(D^2 + \lambda I)V^T)^{-1} VDU^T y = UDV^T [(D^2 + \lambda I)V^T]^{-1} V^T VDU^T y = UDV^T V(D^2 + \lambda I)^{-1} D U^T y \\ &= UD(D^2 + \lambda I)^{-1} DU^T y = \sum_{j=1}^n \frac{d_j^2}{d_j^2 + \lambda} u_j u_j^T y \end{aligned} \tag{11.18}$$

also,

$$\hat{\beta}_{LS} = \sum_{j=1}^n v_j (u_j^T y)$$

so by comparison, what ridge regression is doing is changing the weights of the columns (by weights that are between 0 and 1 for any  $\lambda > 0$ ). Higher  $d_j$ 's means higher weights. So the directions that have higher variability are shrunk less.

So in the limit of (11.18) as  $\lambda \rightarrow 0^+$ , we have

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y \rightarrow X^\dagger y = X^\dagger (X \beta + \epsilon) \tag{11.19}$$

where we substitute the pseudoinverse instead of the inverse because since  $X^T X$  is rank deficient,  $(X^T X)^{-1}$  does not exist (and because the Moore-Penrose pseudoinverse minimizes the  $\ell_2$  norm, exactly what the ridge solution will do).

$$\hat{y}_{\text{ridge}}(\lambda) = X \hat{\beta}_{\text{ridge}}(\lambda) \rightarrow \sum_{j=1}^n u_j u_j^T y = UU^T y,$$

which is the least squares solution in the case that  $p \leq n$  and this exists, or (11.19) in the general case.

### 11.13 Lasso

From KKT theory, the correlation between all selected features and residual will be  $\lambda$  (see the remark in Section 11.13.4 for an explanation why).

Consider the linear regression model  $y = X\beta + \epsilon$ . If we assume the errors  $\epsilon$  have a multivariate Gaussian distribution, that is,

$$f_\epsilon(t) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left( -\frac{t^T t}{2\sigma^2} \right), \quad t = (t_1, \dots, t_n)^T$$

then the log likelihood is

$$\log(f(t)) = n \log[(2\pi\sigma^2)^{-1/2}] - t^T t / (2\sigma^2)$$

Suppose we want the MLE estimator. When we maximize the log likelihood, we can disregard the first term which does not include  $t$  (it is constant). So we seek

$$\arg \max_{\beta \in \mathbb{R}^p} \{-t^T t / (2\sigma^2)\} = \arg \max_{\beta \in \mathbb{R}^p} \{-\|y - X\beta\|_2^2 / (2\sigma^2)\}$$

which is the same as

$$\arg \min_{\beta \in \mathbb{R}^p} \{\|y - X\beta\|_2^2 / (2\sigma^2)\}$$

We commonly scale this with an  $n$  in the denominator to match the empirical risk; note that this does not affect the arguments which minimize the quantity. When the design matrix  $X$  multiplied by  $n^{-1/2}$  is orthonormal ( $(X^T X = nI_p)$ , the penalized least squares reduces to the minimization of

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\hat{\beta}\|_2^2 + \frac{1}{2} \|\hat{\beta} - \beta\|_2^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\}$$

where  $\hat{\beta} = (X^T X)^{-1} X^T y = nX^T y$  is the OLS estimator. Disregarding the first term which does not contain  $\beta$ , we have a **separable** loss function (we can solve for one parameter at a time):

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\hat{\beta} - \beta\|_2^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\}.$$

So we can consider the univariate penalized least squares function

$$\hat{\theta}(z) = \arg \min_{\theta \in \mathbb{R}} \left\{ \frac{1}{2} (z - \theta)^2 + p_\lambda(|\theta|) \right\}.$$

Antoniadis and Fan [2001] showed that the PLS estimator  $\hat{\theta}$  possesses the following properties:

- *sparsity* if  $\min_{t \geq 0} \{t + p'_\lambda(t)\} > 0$ ;
- *approximate unbiasedness* if  $p'_\lambda(t) = 0$  for large  $t$ ;
- *continuity* if and only if  $\arg \min_{t \geq 0} \{t + p'_\lambda(t)\} = 0$ . Intuition: if you perturb data a little, the solution should remain similar.

In general, the singularity of the penalty function at the origin (i.e.,  $p'_\lambda(0+) < 0$ ) is needed for generating sparsity in variable selection and the concavity is needed to reduce the bias.

To recap: constrained version:

$$\begin{aligned}\hat{\beta}_{\text{lasso}} = \arg \min_{\beta \in \mathbb{R}^p} \quad & \frac{1}{2n} \|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq t\end{aligned}$$

Unconstrained version:

$$\hat{\beta}_{\text{lasso}} = \arg \min \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

Use  $1/n$  to rescale RSS due to  $\|1\| - 2 = \sqrt{n}$ .

**Proposition 11.13.1 (Math 541A Homework Problem).** Suppose  $\beta \in \mathbb{R}^p$  is an unknown vector, and for all  $1 \leq i \leq n$ , there are known vectors  $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$ . Our observed data are  $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ . Let  $X$  be the  $n \times p$  matrix so that the  $i^{\text{th}}$  row of  $X$  is the row vector  $x^{(i)}$ . Assume that  $p \leq n$  and the matrix  $X$  has full rank. Let  $\lambda > 0$  and consider the quantity

$$\sum_{i=1}^n \left( y_i - x^{(i)T} \beta \right)^2 + \lambda \sum_{i=1}^p |\beta_i| \tag{11.20}$$

Then there exists a  $\hat{\beta} \in \mathbb{R}^p$  that minimizes this quantity (this  $\hat{\beta}$  is known as the LASSO, or least absolute shrinkage and selection operator).

*Proof.* We can write (11.20) as

$$\begin{aligned}\sum_{i=1}^n \left( y_i - x^{(i)T} \beta \right)^2 + \lambda \sum_{i=1}^p |\beta_i| &= \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \\ &= \|y - X\beta\|_2^2 + \lambda \|\beta\|_1\end{aligned}\tag{11.21}$$

By Proposition 9.1.11,  $\|y - X\beta\|_2^2$  is convex, and by Proposition 9.1.12,  $\lambda \|\beta\|_1$  is convex. Therefore by Proposition 9.1.8, (11.21) is convex. Differentiating and setting equal to 0 yields

$$-2\mathbf{y}^T \mathbf{X} + 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} + \lambda [\text{sgn}(\boldsymbol{\beta}_i)] = 0 \quad (11.22)$$

where  $[\text{sgn}(\boldsymbol{\beta}_i)]$  is vector resulting from the sgn function being applied elementwise to  $\boldsymbol{\beta}$ . Since (11.22) is linear in  $\boldsymbol{\beta}$ , it has one solution. Since (11.21) is convex, any solution to (11.22) minimizes (11.20).

□

**Remark.** The  $L_1$  penalization term in (11.20) is better at penalizing large entries of  $\boldsymbol{\beta}$  (a similar observation applies in the compressed sensing literature). Unfortunately, there is no closed form solution to (11.20) in general. The constrained minimization problem

$$\text{minimize } \sum_{i=1}^n (\mathbf{y}_i - \langle \mathbf{x}^{(i)}, \boldsymbol{\beta} \rangle)^2 \quad \text{subject to } \sum_{i=1}^n |\boldsymbol{\beta}_i| \leq 1$$

is morally equivalent to (11.20), but technically Lagrange Multipliers does not apply since the constraint is not differentiable everywhere.

### 11.13.1 Soft Thresholding

Classical ideas of nonparametric models: kernels (locally constant/linear), splines (smooth basis functions). But wavelets are non-smooth. Why is this beneficial? Some real life functions are non-smooth. (example; image data with noise. There will be non-smooth edges to objects.) Also, the wavelet basis functions are orthonormal (which is closely related to the assumption we made above about the orthonormal design matrix). So when working with wavelets, we have a separable optimization problem. Soft thresholding is something like the lasso idea for wavelets (but before the lasso was developed).

Suppose we wish to recover an unknown function  $f$  on  $[0, 1]$  from noisy data

$$d_i = f(t_i) + \sigma z_i, \quad i = 0, \dots, n-1$$

where  $t_i = i/n$  and  $z_i \sim \mathcal{N}(0, 1)$ . The term de-noising is to optimize the mean squared error  $n^{-1} E \|\hat{f} - f\|_2^2$ . [Donoho and Johnstone \[1994\]](#) proposed a soft-thresholding estimator

$$\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

where  $\gamma$  is some small number. (So estimator gets shrunk by  $\gamma$ , and if  $\gamma$  is bigger than the original estimator, we set it equal to 0.) They applied this estimator to the coefficients of a wavelet transform of a function measured with noise, then back-transformed to obtain a smooth estimate of the function.

**Example 11.1.** Suppose we have an image in data in the form of  $X \in \mathbb{R}^n$ . We have a wavelet basis  $W \in \mathbb{R}^{n \times n}$  where  $W$  is orthonormal. We transform the image into the frequency domain by

$$Wx \rightarrow \tilde{x}$$

where  $\tilde{x}$  is the frequency domain representation. Then we apply soft-thresholding to  $\tilde{x}$  to yield  $\tilde{x}^*$ , which we hope is de-noised. Finally, we bring the image back into the original domain according to

$$\hat{x} = W^{-1}\tilde{x}^* = W^T\tilde{x}^*.$$

The asymptotic risk of this estimator is

$$[2(\log p) + 1](\sigma^2 + R_{DP})$$

Note that the  $2 \log p$  term is related to the result (described informally) below:

**Proposition 11.13.2.** if we have  $n$  i.i.d.  $\mathcal{N}(0, 1)$  random variables, the maximum of them is near  $\sqrt{2 \log n}$  if  $n$  is large. (The order is this large with high probability)

**Remark.** In the language of wavelets, sometimes  $\ell_0$  penalization is called “hard-thresholding.”

### 11.13.2 Lasso theory

Drawbacks of previous techniques that lasso helps with: subset selection is interpretable but computationally intensive and not stable because it is a discrete process (small changes in the data can result in very different models being selected). Ridge regression is a continuous process and more stable, but it does not set any coefficients equal to 0 and hence does not give an easily interpretable model.

In the orthonormal design case  $X^T X = nI_p$ , the lasso solution can be shown to be the same as soft thresholding:

$$\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

where  $\gamma \geq 0$  is determined by the condition  $\sum_{j=1}^p |\beta_j| = t$ .

Geometry: the criterion  $\sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2$  equals the quadratic function (plus a constant)

$$(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0).$$

*Proof.*

$$\begin{aligned} \sum_{i=1}^n \left( y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 &= \sum_{i=1}^n \left( y_i - X_i \hat{\beta} \right)^2 = (\mathbf{y} - \mathbf{X} \hat{\beta})^T (\mathbf{y} - \mathbf{X} \hat{\beta}) = [\mathbf{X}(\beta^0 - \hat{\beta})]^T [\mathbf{X}(\beta^0 - \hat{\beta})] \\ &= (\beta^0 - \hat{\beta})^T \mathbf{X}^T \mathbf{X} (\beta^0 - \hat{\beta}) \end{aligned}$$

□

The contours (level sets) are therefore elliptical and centered at the OLS estimates. If the constraint region does not have corners, as in ridge regression, zero solutions result with probability zero (see Proposition 11.12.2 and Figure 7).

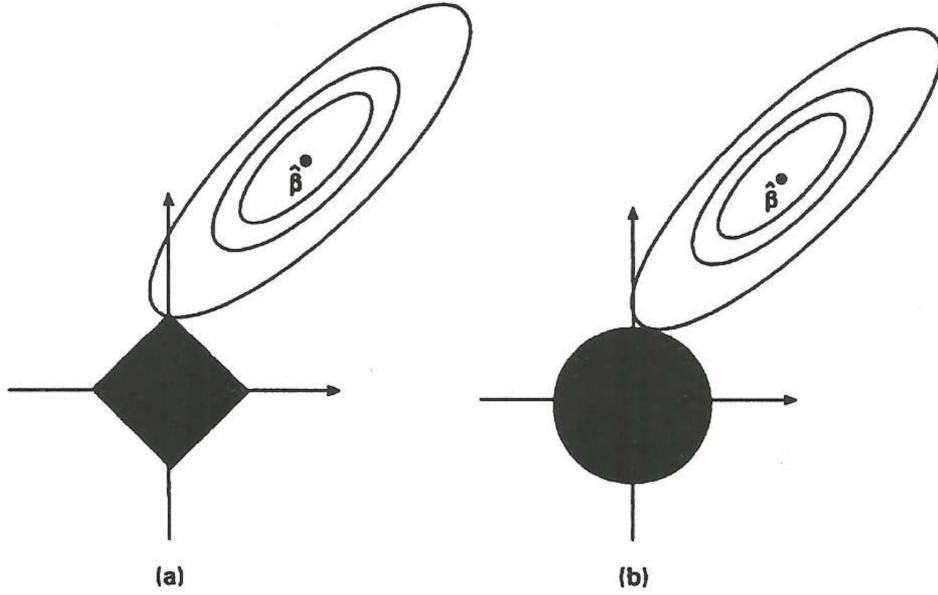


Figure 7: Level sets of least squares loss function with feasible sets for (a) lasso and (b) ridge regression in the case of  $\beta \in \mathbb{R}^2$ .

**Proposition 11.13.3 (2018 DSO Statistics Group In-Class Screening Exam, Question 5).** Consider the optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad (11.23)$$

where  $y \in \mathbb{R}^n$ ,  $X \in \mathbb{R}^{n \times p}$ , and  $\lambda > 0$ .

(a) The following problem is a dual of (11.23):

$$\underset{u \in \mathbb{R}^n}{\text{maximize}} \quad \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 \quad \text{subject to} \quad \|X^T u\|_\infty \leq \lambda.$$

Also,  $\hat{u} = y - X\hat{\beta}$ , where  $\hat{\beta}$  is a solution of (11.23) and  $\hat{u}$  is a solution of the dual.

(b)  $\hat{\beta}$  is not necessarily unique, but  $\hat{u}$ ,  $\|y - X\hat{\beta}\|_2^2$ , and  $\|\hat{\beta}\|_1$  are.

(c) Suppose  $y = X\beta^* + \epsilon$ , and suppose the tuning parameter  $\lambda$  is chosen to satisfy  $\lambda \geq \|X^T \epsilon\|_\infty$ . Then

(i)

$$\frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{2} \|\epsilon\|_2^2 + \lambda \|\beta^*\|_1.$$

(ii)

$$\frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \geq \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|X\beta^*\|_2^2.$$

(iii)

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \geq \frac{1}{2}\|\epsilon\|_2^2 - \lambda\|\beta^*\|_1.$$

**Remark.** We can express the original optimization problem (11.23) as

$$\begin{aligned} & \underset{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \\ & \text{subject to} \quad z = X\beta. \end{aligned} \tag{11.24}$$

We will also refer to another expression of the lasso optimization problem,

$$\begin{aligned} & \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 \\ & \text{subject to} \quad \|\beta\|_1 \leq t \end{aligned} \tag{11.25}$$

for some  $t > 0$ .

Before proving the main results, we will show a few simpler results. Whenever  $\lambda > 0$ , the lasso objective function (11.23) is the Lagrangian of (11.25). We will prove a useful lemma about the relationship between these functions.

**Lemma 11.13.4.** For a given  $\lambda > 0$ , let  $\hat{\beta}$  minimize (11.23). Then there is exactly one  $t = \|\hat{\beta}\|_1$  such that any  $\hat{\beta}$  minimizing (11.23) also minimizes (11.25).

*Proof.* This must be true by contradiction. First of all, since the objective function of (11.25) is continuous and the feasible region  $\|\beta\|_1 \leq t$  is compact, a minimum of (11.25) is guaranteed to exist. Now suppose  $\hat{\beta}$  minimizes (11.23) for a fixed  $\lambda$ , with  $\|\hat{\beta}\|_1 = t$ , but there is a different solution  $\hat{\beta}^*$  that is feasible for (11.25) and achieves a lower value. That is,

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2.$$

and  $\|\hat{\beta}^*\|_1 \leq \|\hat{\beta}\|_1 = t$ . Since  $\lambda > 0$ ,  $\|\hat{\beta}\|_1 < \|\hat{\beta}_{\text{global}}\|_1$ , where  $\hat{\beta}_{\text{global}}$  is a global minimum for  $\frac{1}{2}\|y - X\hat{\beta}\|_2^2$ . Since (11.25) is convex and all global minima lie outside the feasible region,  $\hat{\beta}^*$  lies on the boundary; that is,  $\|\hat{\beta}^*\|_1 = \|\hat{\beta}\|_1 = t$ . But then

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 \iff \frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 + \lambda\|\hat{\beta}^*\|_1 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$$

which contradicts the fact that  $\hat{\beta}$  minimizes (11.23).

□

Another useful result follows in a simple way from Lemma 11.13.4.

**Proposition 11.13.5.** Let  $\mathcal{B}$  be the set of all  $\hat{\beta}$  that minimize (11.23) for some fixed  $\lambda > 0$ . Then for any two  $\hat{\beta}_1, \hat{\beta}_2 \in \mathcal{B}$ ,  $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$ . That is,  $\|\hat{\beta}\|_1$  is unique.

*Proof.* Suppose  $\hat{\beta}_1$  and  $\hat{\beta}_2$  both minimize (11.23), and (without loss of generality)  $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$ . By Lemma 11.13.4, these values both minimize (11.25) with  $t = \|\hat{\beta}_2\|_1$  (we cannot choose  $t = \|\hat{\beta}_1\|_1$  because  $\hat{\beta}_1$  is not feasible for that problem). Because the global minimum of (11.25) lies outside the feasible region and (11.25) is convex, all solutions to (11.25) lie on the boundary of the feasible region. But  $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$ , so  $\hat{\beta}_1$  is not on the boundary of the feasible region, contradiction. Therefore  $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$  for all solutions  $\hat{\beta}_1, \hat{\beta}_2$  to (11.23); that is,  $\|\hat{\beta}\|_1$  is unique. (See Osborne et al. [2000] for more details.)

□

Now we are ready to prove Proposition 11.13.3.

*Proof of Proposition 11.13.3.* (a) The Lagrangian of (11.24) is

$$\mathcal{L}(\beta, z, u) = \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 + u^T(z - X\beta),$$

so the Lagrange dual function is

$$\begin{aligned} \inf_{\beta, z} \{\mathcal{L}(x, u)\} &= \inf_{\beta, z} \left\{ \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 + u^T(z - X\beta) \right\} \\ &= \inf_{\beta, z} \left\{ \frac{1}{2} (y - z)^T(y - z) + u^T z + \lambda \|\beta\|_1 - u^T X \beta \right\} \end{aligned}$$

This minimization is separable:

$$= \inf_z \left\{ \frac{1}{2} (y^T y - 2y^T z + z^T z) + u^T z \right\} + \inf_{\beta} \{ \lambda \|\beta\|_1 - u^T X \beta \} \quad (11.26)$$

We will handle each part of (11.26) separately. First, the left side:

$$\inf_z \left\{ \frac{1}{2} (y^T y - 2y^T z + z^T z) + u^T z \right\} = \inf_z \left\{ \frac{1}{2} z^T z + (u - y)^T z + \frac{1}{2} y^T y \right\}$$

Since this is a convex quadratic form, differentiate with respect to  $z$  and set equal to zero:

$$z + (u - y) = 0 \implies z = y - u \quad (11.27)$$

$$\begin{aligned} \implies \inf_z \left\{ \frac{1}{2} z^T z + (u - y)^T z + \frac{1}{2} y^T y \right\} &= \frac{1}{2} (y - u)^T (y - u) + (u - y)^T (y - u) + \frac{1}{2} y^T y \\ &= \frac{1}{2} (y^T y - 2u^T y + u^T u) + 2u^T y - y^T y - u^T u + \frac{1}{2} y^T y = -\frac{1}{2} u^T u + u^T y = \frac{1}{2} y^T y - \frac{1}{2} y^T y + u^T y - \frac{1}{2} u^T u \\ &= \frac{1}{2} y^T y - \frac{1}{2} (y^T y - 2u^T y + u^T u) = \frac{1}{2} y^T y - \frac{1}{2} (y - u)^T (y - u) = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 \end{aligned}$$

Next we will minimize the right side of (11.26):

$$\inf_{\beta} \{ \lambda \|\beta\|_1 - u^T X \beta \} = \inf_{\beta} \left\{ \lambda \sum_{i=1}^p |\beta_i| - \sum_{i=1}^p [u^T X]_i \beta_i \right\} = \inf_{\beta} \left\{ \sum_{i=1}^p (\lambda |\beta_i| - [u^T X]_i \beta_i) \right\}$$

$$= \inf_{\beta} \left\{ \sum_{i=1}^p \left( \text{sgn}(\beta_i) \lambda - [u^T X]_i \right) \beta_i \right\} = \sum_{i=1}^p \inf_{\beta_i} \left\{ \left( \text{sgn}(\beta_i) \lambda - [u^T X]_i \right) \beta_i \right\}.$$

Notice that when  $\beta_i$  is negative, if  $\left( \text{sgn}(\beta_i) \lambda - [u^T X]_i \right) = -\left( \lambda + [u^T X]_i \right)$  is positive there is no lower bound on the quantity we are minimizing; otherwise, when  $\beta_i$  is negative the infimum is 0. When  $\beta_i$  is positive, if  $\left( \text{sgn}(\beta_i) \lambda - [u^T X]_i \right) = \left( \lambda - [u^T X]_i \right)$  is negative there is no lower bound on the quantity we are minimizing; otherwise, when  $\beta_i$  is negative the infimum is 0. That is, the only dual feasible points satisfy for all  $i$

$$-\left( \lambda + [u^T X]_i \right) \leq 0, \quad \lambda - [u^T X]_i \geq 0 \iff [u^T X]_i \geq -\lambda, \quad [u^T X]_i \leq \lambda$$

which is equivalent to the condition

$$\|u^T X\|_\infty \leq \lambda.$$

Therefore the Lagrange dual function is

$$\inf_{\beta, z} \{\mathcal{L}(x, u)\} = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 \quad (11.28)$$

subject to the constraint  $\|u^T X\|_\infty \leq \lambda$ . This quantity represents a lower bound on the minimum value of the original optimization problem for all  $u \in \mathbb{R}^p$ . The dual problem is to find the best lower bound by maximizing over  $u$ ; that is, the dual problem is

$$\begin{aligned} & \underset{u \in \mathbb{R}^p}{\text{maximize}} \quad \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 \\ & \text{subject to} \quad \|u^T X\|_\infty \leq \lambda. \end{aligned} \quad (11.29)$$

Lastly, suppose  $\hat{\beta}$  and  $\hat{u}$  satisfy

$$\begin{aligned} \hat{\beta} &= \arg \min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \\ \hat{u} &= \arg \max_{u \in \mathbb{R}^p} \quad \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 = \arg \min_{u \in \mathbb{R}^p} \quad -\frac{1}{2} \|y\|_2^2 + \frac{1}{2} \|y - u\|_2^2 \\ &\text{subject to} \quad \|u^T X\|_\infty \leq \lambda \quad \text{subject to} \quad \|u^T X\|_\infty \leq \lambda \end{aligned}$$

Then since (11.27) is a requirement for dual feasibility of  $u$  and strong duality applies, we have  $\hat{u} = y - X\hat{\beta}$ .

**Remark.** Note that we could also find the arguments maximizing (11.29) by

$$\begin{aligned} \arg \min_{u \in \mathbb{R}^p} \quad & -\frac{1}{2} \|y\|_2^2 + \frac{1}{2} \|y - u\|_2^2 = \arg \min_{u \in \mathbb{R}^p} \quad \frac{1}{2} \|y - u\|_2^2 \\ & \text{subject to} \quad \|u^T X\|_\infty \leq \lambda. \quad \text{subject to} \quad \|u^T X\|_\infty \leq \lambda. \end{aligned}$$

where the first step follows from the fact that arguments that maximize a function are the same as the arguments that minimize the negative of a function, and the second step follows from the fact that the  $-\frac{1}{2} \|y\|_2^2$  term does not include  $u$ . Therefore we see that the residual vector  $u$  from a lasso fit can be thought of as the projection of  $y$  onto the convex polyhedron  $C \subset \mathbb{R}^n$  defined by  $C := \{u : \|X^T u\|_\infty \leq \lambda\}$ .

Another way of saying this is that the lasso estimate  $\hat{y} = X\hat{\beta}_{\text{lasso}}$  itself is the residual from projecting  $y$  onto  $C$ ; that is,

$$X\hat{\beta}_{\text{lasso}} = (I - P_C)y,$$

where  $P_C$  is the operator projecting  $y$  onto  $C$ .

- (b) (i) **Not necessarily unique.** Per Tibshirani [2013], if  $\text{rank}(X) < p$ , the lasso solution is not necessarily unique. Intuitively, this is because the columns of  $X$  are linearly dependent, so there may exist more than one linear combination of the columns that minimizes (11.23). **Jacob's suggestion: counterexample.**  $X$  is two columns that are equal; then convex combinations of two solutions are equal as long as same sign (can't be opposite sign because then  $\ell_1$  could be smaller by setting one equal to 0).
- (ii) **Necessarily unique.** The dual problem (11.29) is strictly concave, so the value  $\hat{u}$  that maximizes it is unique.
- (iii) **Necessarily unique** (except in the trivial case  $\lambda = 0$ ). Per part 5(b)(iv),  $\|\hat{\beta}\|_1$  is unique. (11.23) is convex, so the minimum  $\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$  is unique. Therefore  $\|y - X\hat{\beta}\|_2^2$  must be unique. **Jacob's solution:** Since  $\hat{u}$  is unique and by (11.27)  $\hat{u} = y - X\hat{\beta}$ , we must have that  $\|\hat{u}\| = \|y - X\hat{\beta}\|$  is unique.
- (iv) **Necessarily unique** (except in the trivial case  $\lambda = 0$ ). This is immediate from Proposition 11.13.5.
- (c) (i) Since  $\beta^*$  is clearly feasible for (11.23) and  $\hat{\beta}$  achieves the minimum, we have

$$\frac{1}{2}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1 \geq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \iff \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1$$

- (ii) We know that the expression in the dual problem (11.29) is a lower bound for the solution of the primal problem (11.23) for any  $u$  feasible for (11.29) (that is, any  $u$  satisfying  $\|u^T X\|_\infty \leq \lambda$ ). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2.$$

Since by assumption  $\lambda \geq \|X^T \epsilon\|_\infty$ ,  $\epsilon$  is feasible for (11.29). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - \epsilon\|_2^2 = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \quad (11.30)$$

as desired.

- (iii) We can rewrite the right side of (11.30) as

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|X\beta^*\|_2^2 + \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^* - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^*. \quad (11.31)$$

By assumption, we have

$$\lambda \geq \|X^T \epsilon\|_\infty \iff \lambda \mathbf{1} - X^T \epsilon \succeq 0 \implies \lambda \mathbf{1} \beta^* - X^T \epsilon \beta^* \succeq 0$$

$$\iff -\lambda \|\beta^*\|_1 \leq \epsilon^T X \beta^* \leq \lambda \|\beta^*\|_1.$$

By Hölder's Inequality, we have for any two vectors  $u, v \in \mathbb{R}^n$ ,  $|u^T v| \leq \|u\|_\infty \|v\|_1$ . Therefore

$$|\epsilon^T X \beta^*| = |(X^T \epsilon)^T \beta^*| \leq \|X^T \epsilon\|_\infty \|\beta^*\|_1 \leq \lambda \|\beta^*\|_1$$

where the last step used the assumption  $\|X^T \epsilon\|_\infty \leq \lambda$ . So we have

$$\frac{1}{2} \|\epsilon\|_2^2 + \lambda \|\beta^*\|_1 \leq \frac{1}{2} \|\epsilon\|_2^2 + \epsilon^T X \beta^*.$$

Substituting in to (11.30), using the identity in (11.31), and using the result from part 5(c)(iii) yields

$$\frac{1}{2} \|\epsilon\|_2^2 + \lambda \|\beta^*\|_1 \leq \frac{1}{2} \|\epsilon\|_2^2 + \epsilon^T X \beta^* = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|X \beta^*\|_2^2 \leq \frac{1}{2} \|y - X \hat{\beta}\|_2^2 + \lambda \|\beta\|_1$$

as desired.

(iv) We see from parts (i) and (iii) that

$$\begin{aligned} \frac{1}{2} \|y - X \hat{\beta}\|_2^2 + \lambda \|\beta\|_1 - \lambda \|\beta^*\|_1 &\leq \frac{1}{2} \|\epsilon\|_2^2 \leq \frac{1}{2} \|y - X \hat{\beta}\|_2^2 + \lambda \|\beta\|_1 + \lambda \|\beta^*\|_1 \\ \iff \frac{1}{n} \|y - X \hat{\beta}\|_2^2 + \frac{2}{n} \lambda \|\beta\|_1 - \frac{2}{n} \lambda \|\beta^*\|_1 &\leq \frac{1}{n} \|\epsilon\|_2^2 \leq \frac{1}{n} \|y - X \hat{\beta}\|_2^2 + \frac{2}{n} \lambda \|\beta\|_1 + \frac{2}{n} \lambda \|\beta^*\|_1 \end{aligned}$$

that is, we can lower bound and upper bound  $\frac{1}{n} \|\epsilon\|_2^2$  by taking the quantity  $\frac{1}{n} \|y - X \hat{\beta}\|_2^2 + \frac{2}{n} \lambda \|\beta\|_1$  and adding or subtracting  $\frac{2}{n} \lambda \|\beta^*\|_1$ . Therefore it seems that the quantity in the middle of this interval,  $\frac{1}{n} \|y - X \hat{\beta}\|_2^2 + \frac{2}{n} \lambda \|\beta\|_1$ , is a reasonable estimator for  $\sigma^2 = \mathbb{E}[n^{-1} \|\epsilon\|_2^2]$ .

□

### 11.13.3 Non-Negative Garotte

This idea inspired the lasso. Proposed by Breiman [1995]. It minimizes

$$\sum_{i=1}^n \left( y_i - \alpha - \sum_{j=1}^p c_j \hat{\beta}_j^o x_{ij} \right)^2 \text{ subject to } c_j \geq 0, \sum_{j=1}^p c_j \leq t$$

It starts with OLS estimates and shrinks them by non-negative factors whose sum is constrained. It depends on both the sign and magnitude of OLS estimates. In contrast, lasso avoids the explicit use of OLS estimates.

### 11.13.4 LARS—Preliminaries and Intuition

Intuition: the algorithm takes steps from a model where all coefficients are 0 to the biggest model (the unpenalized OLS model). Covariates are considered from the highest correlation with  $y$  to the least. (The variable most highly correlated with  $y$  is the one at the “least angle” from  $y$ .) Recall the original definition of the lasso estimator:

$$\hat{\beta}_{lasso} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t \quad (11.32)$$

The more common version now:

$$\hat{\beta}_{lasso} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \quad (11.33)$$

One form can be changed to the other by applying Lagrangians<sup>2</sup>. Have to be careful because this is a convex program (quadratic with “linear” constraint—use a slack variable).

Taking the gradient of the loss function in (11.33) yields

$$\begin{aligned} \nabla \left( \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right) &= \nabla \left( \frac{1}{2n} \|y - X\beta\|_2^2 \right) + \lambda \nabla (\|\beta\|_1) \\ &= -\frac{1}{n} X^T (y - X\beta) + \lambda \nabla (\|\beta\|_1) \end{aligned} \quad (11.34)$$

We set this equal to zero. If the first term equals 0, the residual has to equal 0. For the second part to equal zero, we have to account for the fact that the gradient doesn’t exist at 0. In the one-dimensional case  $g(t) = |t|$ , we have

$$g'(t) = \begin{cases} -1 & t < 0 \\ 1 & t > 0 \end{cases}$$

but it doesn’t exist at 0. Instead of using the gradient, we will use  $\partial$ , the subdifferential, which is the set of all subgradients. We have a solution if 0 is in the subdifferential. We can rewrite (11.34) using the subdifferential instead of the gradient:

$$\partial \left( \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right) = \nabla \left( \frac{1}{2n} \|y - X\beta\|_2^2 \right) + \lambda \partial (\|\beta\|_1) = -\frac{1}{n} X^T (y - X\beta) + \lambda \partial (\|\beta\|_1)$$

Then rather than setting the gradient equal to 0, our condition is

$$0 \in -\frac{1}{n} X^T (y - X\beta) + \lambda \partial (\|\beta\|_1)$$

Note that

$$\partial g(t) = \begin{cases} -1 & t < 0 \\ [-1, 1] & t = 0 \\ 1 & t > 0 \end{cases} = \begin{cases} \text{sgn}(t) & t \neq 0 \\ [-1, 1] & t = 0 \end{cases}$$

so we have

---

<sup>2</sup>However, the correspondence between  $t$  and  $\lambda$  is **not** one-to-one. Because with  $t = \infty$ ,  $\lambda = 0$ . But a slightly smaller  $t$  would result in the same solution.

$$0 \in -\frac{1}{n}X^T(y - X\beta) + \lambda \cdot \begin{bmatrix} \begin{cases} \text{sgn}(\beta_j) & t \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases} \end{bmatrix} \quad (11.35)$$

where

$$\begin{bmatrix} \begin{cases} \text{sgn}(\beta_j) & t \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases} \end{bmatrix} \in \mathbb{R}^p$$

is a vector with each entry as specified.

**Remark.** (1) Examining the  $j$ th component of the separable equation (11.35), if  $\beta_j \neq 0$ , we have

$$0 = -\frac{1}{n}X_j^T(y - X\beta) + \lambda \cdot \text{sgn}(\beta_j) \iff \frac{1}{n}X_j^T(y - X\beta) = \lambda \cdot \text{sgn}(\beta_j)$$

Note that the left side contains the correlation between  $X_j$  and  $e = y - X\beta$ , the residual vector. **So if lasso chooses  $k$  variables, all  $k$  of them will have the same correlation with the residual ( $\lambda$ ).**

(2) If  $\beta_j = 0$ , we have

$$0 \in -\frac{1}{n}X^T(y - X\beta) + \lambda \cdot [-1, 1] \iff \left| \frac{1}{n}X^T(y - X\beta) \right| \leq \lambda$$

**So for unselected features, the (absolute) correlation should be bounded by  $\lambda$ .**

These two conditions relate to the KKT conditions (first order conditions).

So if we start with  $\lambda$  very large and gradually decrease it, we will let in as the first feature the one that is most highly correlated with  $y$ —that is, the feature with the *least angle* between it and  $y$ .

### 11.13.5 LARS

In Figure 8, note that we choose feature  $X_1$  first because it has the highest correlation with  $y$ . As the coefficient on  $X_1$  increases, the correlation between  $X_1$  and the residual with  $y$  decreases, while the correlation between  $X_2$  and the residual remains constant (**increases?**). When the correlation between  $X_1$  and the residual becomes equal to the correlation between  $X_2$  and the residual,  $X_2$  enters the lasso path.

**Remark.** Just like in lasso, in LARS the correlation between all included features and the residual are equal (see the remark in Section 11.13.4). However, LARS is a stepwise procedure—once we add a feature, it stays in the model. In the lasso, features can be dropped later in the path after they are selected—whenever  $\beta_j$  becomes 0, it is dropped from the current active set. A feature’s sign cannot change in lasso—it is not possible. If we modify the LARS algorithm to have this property (“lasso modification”), then the result is the lasso estimator.

The LARS algorithm for lasso has order  $\mathcal{O}(np \cdot \min\{n, p\})$ . In particular, if  $p > n$  it has order  $\mathcal{O}(n^2p)$ .

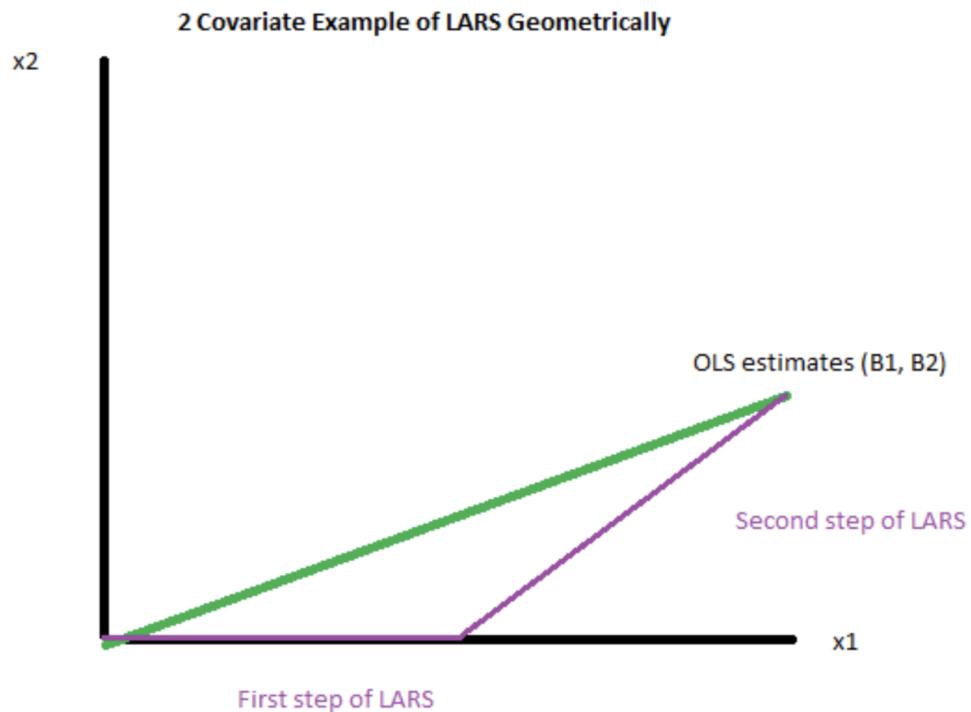


Figure 8: LARS figure in 2d case.

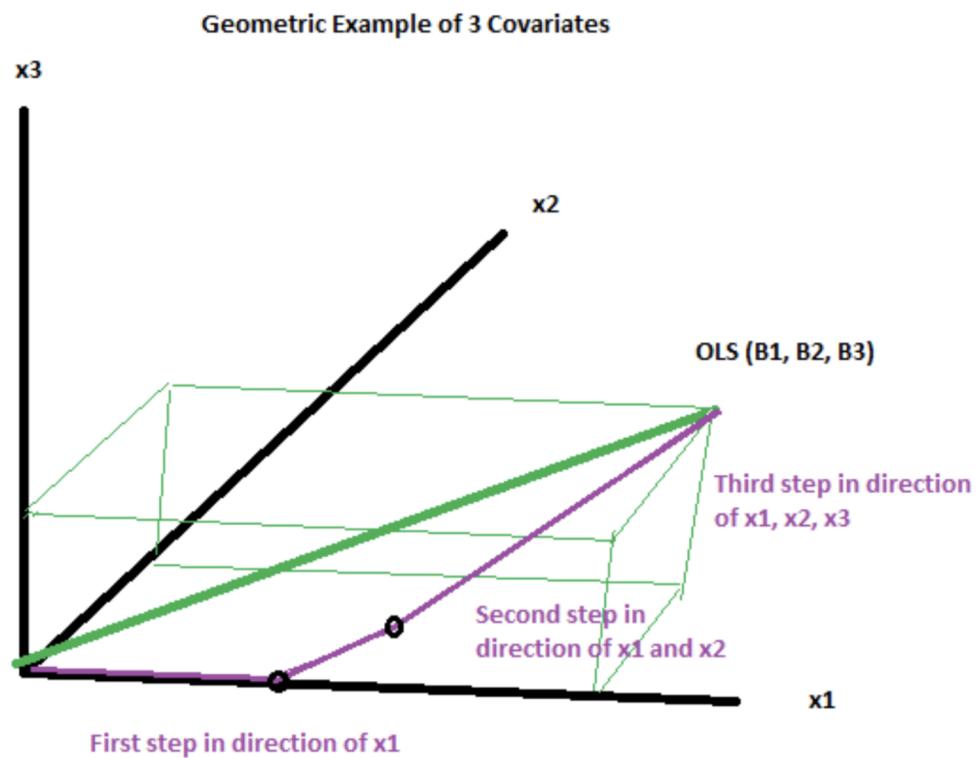


Figure 9: LARS figure in 3d case.

## 11.14 Loss Functions

Asymmetric loss: may have asymmetry between overestimation and underestimation. For example: in a supply chain, estimate inventory level  $y_1, \dots, y_n$ . Let  $y_i = x'_i \beta + \epsilon_i$ . Underestimating is really bad because then you don't have enough product for customers and they might not come back; overestimating not so bad. Then our loss function:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (1 - \tau) \underbrace{(y_i - x'_i \beta)_+}_{\text{under-estimation}} + \tau \underbrace{(x'_i \beta - y_i)_+}_{\text{over-estimation}}$$

you could choose  $\tau = 0.1$  for a 9 to 1 ratio of loss (i.e., underestimation is 9 times as costly as overestimation).

**Theorem 11.14.1 (Loss: quadratic).** Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable with  $\mathbb{E}X^2 < \infty$ . Then  $\mathbb{E}(X - t)^2$  is minimized for  $t \in \mathbb{R}$  uniquely when  $t = \mathbb{E}X$ .

*Proof.* We seek

$$\arg \min_t \mathbb{E}(X - t)^2 = \arg \min_t [\mathbb{E}(X^2) - 2t\mathbb{E}(X) + t^2] = \arg \min_t [t^2 - 2t\mathbb{E}(X)]$$

where the last step follows because  $\mathbb{E}(X^2)$  is independent of  $t$ . This expression is quadratic in  $t$ . Differentiating with respect to  $t$  and setting equal to 0, we have

$$2t - 2\mathbb{E}(X) = 0 \implies \boxed{\arg \min_t \mathbb{E}(X - t)^2 = \mathbb{E}(X)}$$

□

Huber loss: combines benefits of squared loss (unbiasedness) with MAD loss (robust).

$$L_\delta^H(y, \hat{y}) = (y - \hat{y})^2 \cdot I\{|y - \hat{y}| \leq \delta\} + |y - \hat{y}| \cdot I\{|y - \hat{y}| > \delta\}.$$

Only downside: not that easy to estimate (loss function is not differentiable). Instance of **M-estimation** (more generalized than regression):

$$\hat{\beta}_M = \arg \min_{\beta, \sigma} \sum_{i=1}^n \rho_M \left( \frac{y_i - x'_i \beta}{\sigma} \right)$$

where  $\rho_M$  is a loss function. Suppose  $y_i = \sum_{j=1}^p x_{ij} \beta_j + \epsilon_i$ . Then differentiating the loss function with respect to  $\beta$  yields

$$\frac{\partial}{\partial \beta_k} \left[ \sum_{i=1}^n \rho_M \left( \frac{y_i - x'_i \beta}{\sigma} \right) \right] = 0 \quad \text{for } k \in \{1, \dots, p\}.$$

$$= \sum_{i=1}^n \rho'(y_i - x_i' \beta) \cdot \frac{(-x_{ij})}{\sigma}$$

Let  $u_i := (y_i - x_i' \beta) / \hat{\sigma}$ .

$$\implies \sum_{i=1}^n \underbrace{\frac{\rho'(u_i)}{u_i}}_{\text{weight}} \cdot x_{ij}(y_i - x_i' \beta) = 0$$

Compare to least squares: normal equations  $X^T(y - X^T \beta_{OLS}) = 0$ , or

$$\sum_{i=1}^n x_{ij}(y_i - x_i' \beta) = 0$$

Now we have this extra weighting term. Algorithm for computing:

1. Use  $\hat{\beta}_{OLS}$  as the initial solution; then compute

$$u_i^{(0)} = \frac{y_i' - x_i' \hat{\beta}_{OLS}}{\hat{\sigma}_{OLS}}$$

2. With  $w_i^{(0)} = \rho'(u_i)/u_i$ , compute

$$\hat{\beta}_{LS}^{(1)} = (X^T W X)^{-1} X^T W y.$$

3. Update weights with  $\hat{\beta}_{LS}^{(1)}$ .
4. Repeat until convergence. (Convergence is a little tricky but should work.)

For absolute deviation loss, see Section 11.6 on quantile regression.

#### 11.14.1 Feature Selection properties

Model selection consistency:  $\Pr(\text{supp}(\hat{\beta}) = \text{supp}(\beta_0)) \rightarrow 1$ .

Oracle property: model selection consistency, asymptotic efficiency as efficient as if true model were known (“efficiency” having to do with the variance given  $n$ ).

**Definition 11.1 (Oracle property).** Let  $\beta^0$  denote the true parameter vector for data generated from a linear model. Let  $S_0$  be the true support; that is,  $S_0 = \{j : \beta_j^0 \neq 0, j = 1, \dots, p\}$ . Denote  $\hat{\beta}(\delta)$  the coefficient estimator for fitting procedure  $\delta$ . We call  $\delta$  an **oracle procedure** if  $\hat{\beta}(\delta)$  asymptotically has the following properties:

- Identifies the right subset model (consistency):  $\{j : \hat{\beta}_j \neq 0\} = S_0$ .

- Has the optimal estimation rate:  $\sqrt{n}(\hat{\beta}(\delta)_{S_0} - \beta_{S_0}^0) \xrightarrow{d} \mathcal{N}(0, \Sigma_0)$  where  $\Sigma_0$  is the covariance matrix knowing the true subset model.

The lasso problem is convex but not necessarily strictly convex if  $p > n$ . That is, there is some flat region, so the minimizer may not be unique. Consider the KKT conditions from convex optimization:

$$g(\beta) = \arg \min \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} = \arg \min \{f_1(\beta) + f_2(\beta)\}$$

Then  $\hat{\beta}$  is a lasso solution if and only if 0 is in the subdifferential of  $g(\hat{\beta})$ . Note that

$$\partial g(\hat{\beta}) = \nabla f_1 + \partial f_2 = \frac{1}{n} X^T(X\beta - y) + \lambda \begin{bmatrix} \vdots \\ \partial|\beta_j| \\ \vdots \end{bmatrix} = \frac{1}{n} X^T(X\beta - y) + \lambda \begin{bmatrix} \vdots \\ \begin{cases} \text{sgn}(\beta_j) & \beta_j \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases} \\ \vdots \end{bmatrix}$$

Now assume  $\text{supp}(\hat{\beta}) = \text{supp}(\beta_0)$  (that is, assume lasso recovers the correct support). Suppose the first  $s$  features are nonzero and consider one of them (so we know that we should have  $\hat{\beta}_j \neq 0$ ):

$$0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) = \left[ \frac{1}{n} X^T(X\beta - y) \right]_j + \lambda \text{sgn}(\hat{\beta}_j)$$

Therefore

$$\frac{1}{n} X_A^T(X\hat{\beta} - y) + \lambda \text{sgn}(\hat{\beta}_j) = 0 \tag{11.36}$$

where  $X_A$  is a submatrix of  $X$  containing the columns corresponding to the features in the true support, is our first condition. Next, consider what happens for  $j > s$  (features not in the true support). We have

$$\begin{aligned} 0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) &= \left[ \frac{1}{n} X^T(X\beta - y) \right]_j + \lambda[-1, 1] \\ &\implies \left\| \frac{1}{n} X_{A^C}^T(X\hat{\beta} - y) \right\|_\infty \leq \lambda \end{aligned} \tag{11.37}$$

where  $X_{A^C}$  is a submatrix of  $X$  containing the columns corresponding to the features not in the true support, is our boundary condition. Recall the true model

$$y = X\beta_0 + \epsilon$$

and consider the case  $X = [X_1 \ X_2]$  where  $X_1$  are the features in the true model and  $X_2$  are noise features; that is,  $\beta_0 = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix}$ . Then we are assuming

$$\hat{\beta}_{\text{lasso}} = \begin{bmatrix} \hat{\beta}_1 \\ 0 \end{bmatrix}.$$

We have from (11.36)

$$\begin{aligned} 0 &= \frac{1}{n} X_1^T (X \hat{\beta} - y) + \lambda \text{sgn}(\hat{\beta}_1) = \frac{1}{n} X_1^T (X_1 \hat{\beta}_1 - X_1 \beta_1 - \epsilon) + \lambda \text{sgn}(\hat{\beta}_1) \\ &\iff \frac{1}{n} X_1^T X_1 (\hat{\beta}_1 - \beta_1) = \frac{1}{n} X_1^T \epsilon - \lambda \text{sgn}(\hat{\beta}_1) \end{aligned}$$

Let's assume that  $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta_0)$  (sign consistency).

$$\iff \frac{1}{n} X_1^T X_1 (\hat{\beta}_1 - \beta_1) = \frac{1}{n} X_1^T \epsilon - \lambda \text{sgn}(\beta_1)$$

which is linear in  $\hat{\beta}$ . Solving, we have

$$\iff \hat{\beta}_1 - \beta_1 = (X_1^T X_1)^{-1} (X_1^T \epsilon - n \lambda \text{sgn}(\beta_1)) \iff \hat{\beta}_1 = \beta_1 + (n^{-1} X_1^T X_1)^{-1} (n^{-1} X_1^T \epsilon - \lambda \text{sgn}(\beta_1)) \quad (11.38)$$

Looking at the second (boundary) condition (11.37), we have

$$\left\| \frac{1}{n} X_2^T (X \hat{\beta} - y) \right\|_{\infty} \leq \lambda. \quad (11.39)$$

Consider that

$$X \hat{\beta} - y = X_1 \hat{\beta}_1 - X_1 \beta_1 - \epsilon = X_1 (\hat{\beta}_1 - \beta_1) - \epsilon$$

Substituting in the result from (11.38) yields

$$X \hat{\beta} - y = X_1 [(n^{-1} X_1^T X_1)^{-1} (n^{-1} X_1^T \epsilon - \lambda \text{sgn}(\beta_1))] - \epsilon$$

which when we plug into (11.39) yields

$$\begin{aligned} &\left\| \frac{1}{n} X_2^T [X_1 (n^{-1} X_1^T X_1)^{-1} (n^{-1} X_1^T \epsilon - \lambda \text{sgn}(\beta_1)) - \epsilon] \right\|_{\infty} \leq \lambda. \\ &\iff \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} (n^{-1} X_1^T \epsilon - \lambda \text{sgn}(\beta_1)) - \frac{1}{n} X_2^T \epsilon \right\|_{\infty} \leq \lambda. \end{aligned}$$

Using the Triangle Inequality, we have

$$\begin{aligned}
& \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} (n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)) - \frac{1}{n} X_2^T \epsilon \right\|_{\infty} \\
& \leq \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} (n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)) \right\|_{\infty} + \left\| \frac{1}{n} X_2^T \epsilon \right\|_{\infty} \\
& \leq \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \right\|_{\infty} \cdot \|n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\|_{\infty} + \left\| \frac{1}{n} X_2^T \epsilon \right\|_{\infty} \tag{11.40}
\end{aligned}$$

Assume that the  $j$ th column of  $X$  has  $L_2$  norm  $n^{1/2}$  (as it would if all entries equaled 1). We have

$$\|n^{-1} X_1^T \epsilon\|_{\infty} \leq \lambda/2, \quad \|n^{-1} X_2^T \epsilon\|_{\infty} \leq \lambda/2$$

$$\|n^{-1} X^T \epsilon\|_{\infty} \leq \lambda/2 \text{ with large probability}$$

Recall that  $\lambda = \sigma \sqrt{\frac{c \log p}{n}}$  for some  $c > 2$ . Then we have (continuing from (11.40)), and using  $\|n^{-1} X_2^T \epsilon\| \leq \lambda/2$ ,

$$\begin{aligned}
& \leq \|n^{-1} X_1^T \epsilon\|_{\infty} + \|\lambda \operatorname{sgn}(\beta_1)\|_{\infty} \\
& \|n^{-1} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1}\|_{\infty} \cdot \underbrace{\|\cdot\|_{\infty}}_{3/2\lambda} + \underbrace{\|\cdot\|_{\infty}}_{\lambda/2} \leq \lambda \\
& \left\| \underbrace{n^{-1} X_2^T X_1}_{\text{corr. between noise and true sample covariance matrix}} \left( \underbrace{n^{-1} X_1^T X_1}_{\text{sample covariance matrix}} \right)^{-1} \right\|_{\infty} \leq 1/3 \tag{11.41}
\end{aligned}$$

It turns out we're fine as long as it's less than or equal to 1. This is known as the **irrepresentable condition**. Note that the sample covariance matrix is the same as the sample correlation since the columns are standardized. So this is the correlation between the true variables. Note that this matrix has dimension  $(p-s) \times s$  where  $s$  is the dimension of the true support. Note that

$$n^{-1} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} = [X_2^T X_1 (X_1^T X_1)^{-1}]^T = (X_1^T X_1)^T X_1^T X_2$$

which is ordinary least squares for regressing  $X_2$  on  $X_1$ . In the end, the irrepresentable condition says the correlation between the noise and true variables can't be too high.

### 11.15 Dantzig Selector

Dantzig selector:

$$\begin{aligned}\hat{\beta}_{\text{Dantzig}} &= \arg \min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \\ \text{subject to } & \|n^{-1}X^T(y - X\beta)\|_\infty \leq \lambda\end{aligned}$$

Can be recast as a linear program:

$$\begin{aligned}\hat{\beta}_{\text{Dantzig}} &= \arg \min_{u \in \mathbb{R}^p} \sum_{i=1}^p u_i \\ \text{subject to } & -u \leq \beta \leq u \\ & -\lambda_p \sigma \mathbf{1} \leq n^{-1}X^T(y - X\beta) \leq \lambda_p \sigma \mathbf{1}\end{aligned}\tag{11.42}$$

where  $|u|$  denotes the absolute value of  $u$  componentwise. (This is a benefit because linear programming is easy to use and very popular in industry and other applications.) Note that  $n^{-1}X^t(y - X\beta)$  corresponds to the correlations between the residuals and the design matrix. Recall that in OLS this correlation is 0—the design matrix is orthogonal to the residuals. In the Dantzig selector we relax this, bounding the  $L_\infty$  norm by  $\lambda$ . Recall that the gradient of the log-likelihood is the **score function**, in this case  $n^{-1}X^T(y - X\beta)$ . For example, the score equation in linear regression is  $n^{-1}X^T y = n^{-1}X^T X\beta$ . Note:

$$\nabla \left( \frac{1}{2n} \|y - X\beta\|_2^2 \right) = \frac{1}{n} X^T(X\beta - y)$$

Note for Theorem 1: in original paper, assumed columns had  $L_2$  norm 1, resulting in  $\lambda_p = \sqrt{2 \log p}$ . We are instead assuming each column has  $L_2$  norm  $\sqrt{n}$ , which results in  $\lambda = \sigma \cdot \sqrt{\frac{c \log p}{n}}$ . Intuition of  $\log p$  term:

By a theorem in [James et al. \[2009\]](#), the lasso and Dantzig selector estimates equal each other under certain conditions:

**Theorem 11.15.1.** Let  $I_l$  be the support of the lasso estimate  $\hat{\beta}_{\text{lasso}}$ . Let  $\mathbf{X}_L$  be the  $n \times |I_L|$  matrix constructed by taking  $\mathbf{X}_{I_L}$  and multiplying its columns by the signs of the corresponding coefficients in  $\hat{\beta}_{\text{lasso}}$ . Suppose that  $\lambda_{\text{lasso}} = \lambda_{\text{Dantzig}}$ . Then  $\hat{\beta}_{\text{lasso}} = \hat{\beta}_{\text{Dantzig}}$  if  $\mathbf{X}_L$  has full rank and

$$\mathbf{u} = (\mathbf{X}_L^T \mathbf{X}_L)^{-1} \mathbf{1} \succeq 0 \text{ and } \|\mathbf{X}^T \mathbf{X}_L \mathbf{u}\|_\infty \leq 1$$

where  $\mathbf{1}$  is an  $|I_L|$ -vector of ones and the vector inequality is understood componentwise.

**Corollary 11.15.1.1.** If  $\mathbf{X}$  is orthonormal ( $\mathbf{X}^T \mathbf{X} = \mathbf{I}_p$ ), then the entire lasso and Dantzig selector coefficient paths are identical.

*Proof.* For each index set  $\mathbf{I}$ ,  $\mathbf{X}^T \mathbf{X} = \mathbf{I}_{|\mathbf{I}|}$ , so clearly both of the conditions of Theorem 11.15.1 are satisfied. □

The entire paths can be identical under another condition presented in the same paper.

**Theorem 11.15.2.** Suppose that all pairwise correlations between the columns of  $\mathbf{X}$  are equal to the same value  $\rho$  where  $0 \leq \rho < 1$ . Then the entire Lasso and Dantzig selector coefficient paths are identical. In addition, when  $p = 2$ , the same holds for every  $\rho \in (-1, 1)$ .

## 11.16 Coordinate Descent

Start with  $\beta_1$  varying and all other  $\beta$ s fixed. Optimize  $\beta_1$ . Then cycle through each  $\beta_j$ , run until convergence.

## 11.17 Total Variational Distance

## 11.18 Non-parametric regression

One example: LOESS

Another: GAM

### 11.18.1 Generalized additive models

Suppose  $y_i = f(x_i) + \epsilon_i$ . Suppose  $0 \leq x_1 \leq \dots \leq x_n \leq 1$ . Then express

$$f(x) = \sum_{i=1}^{\infty} \theta_i \phi_i(x)$$

where  $\int_{\mathbb{R}} f^2(x) dx < \infty$  and the  $\phi_i$  form an orthonormal basis. Assume sparsity:

$$f(x) \approx \sum_{i=1}^p \theta_i \phi_i(x).$$

So

$$y_i = \sum_{i=1}^p \theta_i \phi_i(x) + \epsilon_i, \quad i \in \{1, \dots, n\}$$

and if  $f$  is smooth then the coefficient basis is sparse, so  $\theta$  is sparse, so

$$\|\theta\|_p = \left[ \sum_{i=1}^n |\theta_i|^p \right]^{1/p}$$

is small.

Can choose a wavelet basis.

### 11.19 Mixture regression

Suppose

$$y_i \sim \sum_{k=1}^K \pi_k \cdot \mathcal{N}(X'_i \beta_k, \sigma_k^2 I).$$

Mixture modeling is very useful. Typically estimated by expectation maximization (see Section 10.4.8). One example: spatial analysis. See Figure 10. In the black region, slopes and standard deviations will be much higher; outside, will be much lower.

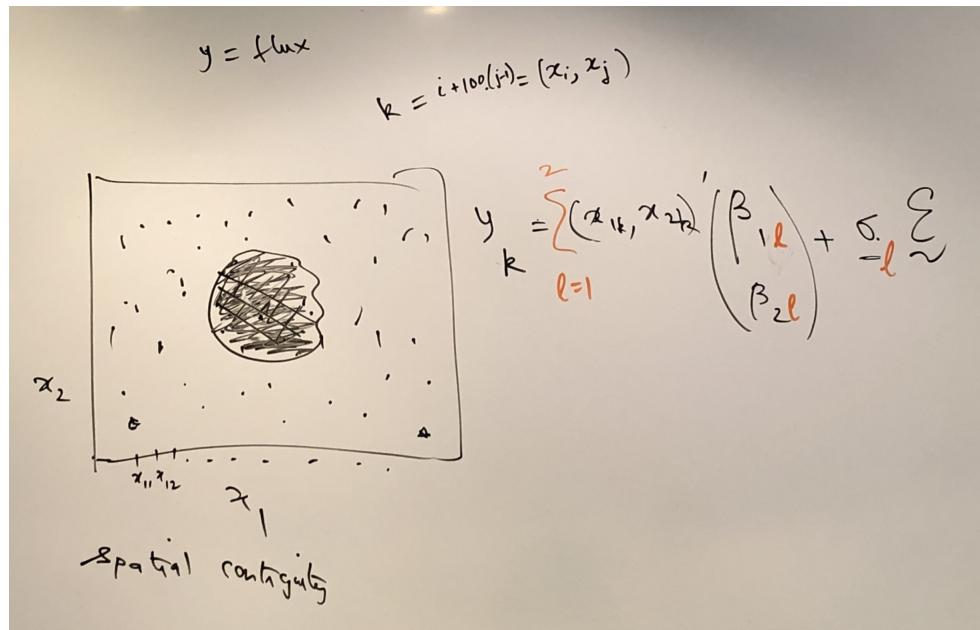


Figure 10: Mixture model example.

### 11.20 Missing observations

### 11.21 Generalized linear models

Exponential family: binomial, Poisson, Beta, negative binomial, etc. See Section 6.4 for more information on exponential families. The density is

$$g_\eta(y) = c(y) \exp\{\eta y - \psi(\eta)\}$$

where  $c(y)$  is the **carrier density**,  $\eta$  is the **natural parameter**,  $y$  is a **sufficient statistic**, and  $\psi(\cdot)$  is

the **normalizing function** (or **cumulant generating function**), chosen so that it is a proper density: that is, choose  $\psi(\eta)$  such that

$$\int_{\mathbb{R}} g_\eta(y) dy = 1,$$

so

$$e^{-\psi(y)} = \left( \int_{\mathbb{R}} c(y) e^{\eta y} dy \right)^{-1}$$

The density is simplest to express in the natural parameter.

Skewness: third centered moment.

$$\frac{\mathbb{E}(Y - \mathbb{E}(Y))^3}{\text{Var}(Y)^{3/2}}$$

kurtosis: 4th centered moment

$$\frac{\mathbb{E}(Y - \mathbb{E}(Y))^4}{\text{Var}(Y)}$$

get excess kurtosis by subtracting 3 (which is Gaussian kurtosis).

In general: cumulant, either centered or raw.

$$K_\gamma = \mathbb{E}(Y - \mathbb{E}(Y))^\gamma, \quad K_\gamma = \mathbb{E}(Y^\gamma)$$

Cumulant generating function:

$$\psi(\eta) - \psi(\eta_0) = \sum_{\gamma=1}^{\infty} K_\gamma \frac{(\eta - \eta_0)^\gamma}{\gamma!}$$

if you know all the cumulants, you know the distribution exactly.

1. **Gaussian variables:** Suppose  $y_1, y_2, \dots, y_n$  are i.i.d.  $\mathcal{N}(\mu, 1)$ .
2. **Gaussian response:** Suppose  $(y_i, x_i)$  are i.i.d. pairs in  $\mathbb{R}^{p+1}$  with  $y_i \sim \mathcal{N}(x_i' \beta, 1)$ .
3. **Exponential family variables:** Suppose  $Y_1, \dots, Y_n$  are i.i.d. in an exponential family. Then the density of each is  $g_\eta(y) = e^{\eta y_0 \psi(\eta)} c(y)$  and the joint density is

$$\prod_{i=1}^n g_\eta(y_i) = \exp \left( \eta \sum_{i=1}^n y_i - n\psi(\eta) \right) \cdot \prod_{i=1}^n c(y_i). \quad (11.43)$$

Then the density of  $\bar{y}$  is simply

$$\exp\{n\eta\bar{y} - n\psi(\eta)\}$$

now the natural parameter is  $n\eta$  and the normalizing function is  $n\psi(\eta)$ . So

$$\bar{Y} \sim \left[ n \frac{d\psi(\eta)}{d\eta}, n \frac{d\psi^2(\eta)}{d\eta^2} \right].$$

The log likelihood is the log of (11.43):

$$\log L(\eta) = \ell(\eta) = \sum_{i=1}^n \eta y_i - n\psi(\eta) + \sum_{i=1}^n \log c(y_i).$$

The maximum likelihood estimate solves

$$\hat{\eta}_{MLE} := \arg \max_{\eta} \ell(\eta)$$

we have

$$\frac{d}{d\eta} \ell(\eta) = \sum_{i=1}^n y_i - n \frac{d}{d\eta} \psi(\eta), \quad \frac{d^2}{d\eta^2} \ell(\eta) = -n \frac{d^2}{d\eta^2} \psi(\eta) = -n \text{Var}(\eta) < 0$$

so since this function is concave down it has a unique maximum at

$$\sum_{i=1}^n y_i - n \frac{d}{d\eta} \psi(\hat{\eta}_{MLE}) \implies \psi(\hat{\eta}_{MLE}) = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} \implies \hat{\mu}_{MLE} = \bar{y}$$

That is, the MLE of the mean parameter in an exponential family is the sample mean. Due to the equivariance property of the maximum likelihood estimator, the MLE for any function of the mean parameter is simply the function of the MLE for the mean parameter (see Proposition 10.4.15).

We call the derivative  $\frac{d}{d\eta} \ell(\eta)$  the **score function**. If  $\xi = h(\eta)$ , then by the chain rule

$$\frac{d}{d\xi} \ell(\eta) = \frac{d}{d\eta} \ell(\eta) \Big/ \frac{d\xi}{d\eta}$$

For the mean parameter in particular, we have

$$\frac{d}{d\mu} \ell(\eta) = \frac{d}{d\eta} \ell(\eta) \Big/ \frac{d\mu}{d\eta} = n(\bar{y} - \mu)/\text{Var}_\eta(Y) \quad (11.44)$$

(since  $\frac{d}{d\eta} \ell(\eta) = n\bar{y} - n \frac{d}{d\eta} \psi(\eta)$  and  $\text{Var}_\eta(Y) = \frac{d\mu}{d\eta}$ ). Also, since the expectation of the score function  $\mathbb{E}(\frac{d}{d\eta} \ell(\eta)) = 0$ , we have

$$\text{Var}\left(\frac{d}{d\eta} \ell(\eta)\right) = \mathbb{E}\left[\left(\frac{d}{d\eta} \ell(\eta)\right)^2\right] - \left[\mathbb{E}\left(\frac{d}{d\eta} \ell(\eta)\right)\right]^2 = \mathbb{E}\left[\left(\frac{d}{d\eta} \ell(\eta)\right)^2\right]$$

We call this the **Fisher information**:

$$i_\eta^{(n)}(\xi) := \mathbb{E}\left[\left(\frac{d}{d\eta} \ell(\eta)\right)^2\right]. \quad (11.45)$$

(See also Definition 10.20 and Section 10.4.3 for more on this.) In the case of the mean, we can write this as

$$i_\eta^{(n)}(\eta) = \mathbb{E} \left[ \left( n\bar{y} - n \frac{d}{d\eta} \psi(\eta) \right)^2 \right] = \text{Var} \left( n\bar{y} - n \frac{d}{d\eta} \psi(\eta) \right) = \text{Var}(n\bar{y}) = n^2 \text{Var}(\bar{y}) = n^2 \text{Var}_\eta(y)/n = n \text{Var}_\eta(y)$$

Then substituting (11.44) into (11.45) we have

$$i_\eta^{(n)}(\mu) = \mathbb{E} \left[ \frac{\left( n\bar{y} - n \frac{d}{d\eta} \psi(\eta) \right)^2}{\text{Var}_\eta(y)^2} \right] = \frac{\text{Var} \left[ n\bar{y} - n \frac{d}{d\eta} \psi(\eta) \right]}{\text{Var}_\eta(y)^2} = \frac{n \text{Var}_\eta(y)}{\text{Var}_\eta(y)^2} = \frac{n}{\text{Var}_\eta(y)}. \quad (11.46)$$

Note that as  $n$  increases, the Fisher information increases linearly. The Fisher information is inversely proportional to the underlying variance. Again, by Proposition 10.4.15, the plug-in estimator for the Fisher Information using the maximum likelihood estimator for the variance is the MLE for the Fisher information:

$$i_\eta^{(n)}(\mu) \Big|_{\eta=\hat{\eta}_{MLE}} = \widehat{\text{Var}}_\eta(y) = \mathbb{E} \left[ \left( \frac{d}{d\mu} \ell(\eta) \right)^2 \right] = -\mathbb{E} \left[ \frac{d^2}{d\mu^2} \ell(\eta) \right] \Big|_{\eta=\hat{\eta}_{MLE}}$$

In the case of the mean parameter, we can write this as

$$= -\frac{d^2}{d\mu^2} \ell(\eta) \Big|_{\eta=\hat{\eta}_{MLE}}$$

**Cramer-Rao lower bound** (see Theorem 10.4.7):

$$\text{Var}_\eta(\hat{\xi}) \geq \frac{\left( \frac{d}{d\eta} [\mathbb{E}_\eta(\hat{\eta})] \right)^2}{i_\eta^{(n)}(\xi)}$$

Any estimator that reaches (or approaches asymptotically) this lower bound is **(asymptotically) efficient**. Consider an unbiased estimator  $\hat{\xi}$  ( $\mathbb{E}(\hat{\xi}) = \xi$ ): then the bound reduces to

$$\text{Var}_\eta(\hat{\xi}) \geq \frac{1}{i_\eta^{(n)}(\xi)}$$

Further, suppose we have an unbiased estimator of the mean parameter  $\hat{\mu}$ . Then using (11.46) we have

$$\text{Var}_\eta(\hat{\mu}) \geq \frac{1}{i_\eta^{(n)}(\mu)} = \frac{\text{Var}_\eta(y)}{n}$$

Since we saw earlier that the MLE has exactly this variance ( $\text{Var}(\bar{y}) = \text{Var}_\eta(y)/n$ ), it is the efficient unbiased estimator for  $\mu$ . (However, it turns out that the plug-in MLE estimator for a function of this mean parameter is not in general the efficient unbiased estimator.)

#### 4. Exponential family model (GLM):

## 12 Time Series

These notes are based on my notes from *Time Series and Panel Data Econometrics* (1st edition) by M. Hashem Pesaran [Pesaran, 2015] as well as coursework for Economics 613: Economic and Financial Time Series I at USC.

### 12.1 Chapter 6: ARDL Models

In an ARDL model, if the error are serially correlated, then the coefficient estimates are biased (even as  $T \rightarrow \infty$ ).

### 12.2 Chapters 12 and 13: Intro to Stochastic Processes and Spectral Analysis

**Stationarity conditions:**  $\{X_t\}$  is **strictly stationary** if the joint distribution functions of  $\{X_{t_1}, X_{t_2}, \dots, X_{t_k}\}$  and  $\{X_{t_1+h}, X_{t_2+h}, \dots, X_{t_k+h}\}$  are identical for all values of  $t_1, t_2, \dots, t_k$  and  $h$  and all positive integers  $k$ .

**Definition 12.1.**  $X_t$  is **weakly (or covariance) stationary** if it has a constant mean and variance and its covariance function  $\gamma(t_1, t_2)$  depends only on the absolute difference  $|t_1 - t_2|$ , namely  $\gamma(t_1, t_2) = \gamma(|t_1 - t_2|)$ .

**Definition 12.2.**  $X_t$  is said to be **trend stationary** if  $y_t = X_t - d_t$  is covariance stationary, where  $d_t$  is the perfectly predictable component of  $X_t$ .

The process  $\{\epsilon_t\}$  is said to be a **white noise process** if it has mean zero, a constant variance, and  $\epsilon_t$  and  $\epsilon_s$  are uncorrelated for all  $s \neq t$ .

**Autocovariance generating function:** The autocovariance generating function for the general linear stationary process  $y_t = \sum_{i=0}^{\infty} a_i \epsilon_{t-i}$  is given by:

$$G(z) = \sigma^2 a(z) a(z^{-1})$$

where  $a(z) = \sum_{i=0}^{\infty} a_i z^i$ .

**Wold's Decomposition** (Theorem 42, p. 275, Section 12.5) Any trend-stationary process  $\{y_t\}$  can be represented in the form of  $y_t = d_t + \sum_{i=0}^{\infty} \alpha_i \epsilon_{t-i}$  where  $\alpha_0 = 1$  and  $\sum_{i=0}^{\infty} \alpha_i^2 < K < \infty$ . The term  $d_t$  is a deterministic component, while  $\{\epsilon_t\}$  is a serially uncorrelated process:  $\epsilon_t = y_t - \mathbb{E}(y_t | y_{t-1}, y_{t-2}, \dots)$ .

**Stationarity conditions for an ARMA( $p, q$ ) process:** Consider the ARMA( $p, q$ ) process

$$y_t = \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=0}^q \theta_i \epsilon_{t-i}, \quad \theta_0 = 1$$

The MA part is stationary for any finite  $q$ . The AR part is stationary if the roots of the characteristic equation

$$\lambda^t = \sum_{i=1}^p \phi_i \lambda^{t-i}$$

lie strictly inside the unit circle. Alternatively, in terms of  $z = \lambda^{-1}$ , the process is stationary if the roots of

$$1 - \sum_{i=1}^p \phi_i z^i = 0$$

lie outside the unit circle. The ARMA process is **invertible** (so that  $y_t$  can be solved uniquely in terms of its past values) if all the roots of

$$1 - \sum_{i=1}^p \theta_i z^i = 0$$

fall outside the unit circle.

**Spectral Density Function:** Definition (Equation 13.3):

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{ih\omega}, \omega \in (-\pi, \pi)$$

Equation (13.5):

$$f(\omega) = \frac{1}{2\pi} \left[ \gamma(0) + 2 \sum_{h=1}^{\infty} \gamma(h) \cos(h\omega) \right], \quad \omega \in [0, \pi]$$

Can also be found using the autocovariance generating function. We have (Equation 13.6, section 13.3.1)

$$f(\omega) = \frac{1}{2\pi} G(e^{i\omega}) = \frac{\sigma^2}{2\pi} a(e^{i\omega}) a(e^{-i\omega})$$

**Properties of spectral density function:**

- (1)  $f(\omega)$  always exists and is bounded if  $\gamma(h)$  is absolutely summable.
- (2)  $f(\omega)$  is symmetric.
- (3) The spectrum of a stationary process is finite at zero frequency; that is,  $f(0) < \infty$ .

Linear (time-domain) processes don't have to be stationary, but to write something as a frequency-domain process, it must be stationary.

### 12.2.1 Worked Examples

#### Midterm Problem 2 part (1) (chapter 12 exercise 6)

Midterm Problem 2 part (2) (exercise 7 in chapter 12; similar to exercise 1 in chapter 14.

Suppose  $\{y_t\}$  has the following general linear process

$$y_t = \mu + \alpha(L)\epsilon_t, \quad \epsilon_t \sim i.i.d. (0, \sigma^2)$$

where  $\alpha(L) = \alpha_0 + \alpha_1 L + \alpha_2 L^2 + \dots$ ;  $\alpha_0 = 1$ . Let

$$\bar{y}_T = \frac{1}{T} \sum_{t=1}^T y_t$$

$$\gamma(h) = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)]$$

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=h+1}^T (y_t - \bar{y}_T)(y_{t-h} - \bar{y}_T)$$

Derive the conditions under which

- (a)  $\bar{y}_T$  is a consistent estimator of  $\mu$  as  $T \rightarrow \infty$
- (b) For fixed  $h$ ,  $\hat{\gamma}(h)$  is a consistent estimator of  $\gamma(h)$  as  $T \rightarrow \infty$ .

#### Solution.

- (a) This is an MA( $\infty$ ) process. By Chebyshev's Inequality (Theorem 8.2.2),  $\bar{y}_T$  is a consistent estimator of  $\mu$  as  $T \rightarrow \infty$  if  $\lim_{T \rightarrow \infty} \mathbb{E}(\bar{y}_T) = \mathbb{E}(y_T) = \mu$  and  $\lim_{T \rightarrow \infty} \text{Var}(\bar{y}_T) = 0$ . In this case in particular (MA( $\infty$ ) process), we can write

$$\bar{y}_T = \frac{1}{T} \sum_{t=1}^T (\mu + \alpha(L)\epsilon_t) = \frac{1}{T} \cdot T\mu + \frac{1}{T} \sum_{t=1}^T \alpha(L)\epsilon_t = \mu + \frac{1}{T} \sum_{t=1}^T \alpha(L)\epsilon_t$$

Then we have

$$\mathbb{E}(\bar{y}_T) = \mu + \frac{1}{T} \mathbb{E} \left( \sum_{t=1}^T \alpha(L)\epsilon_t \right) = \mu + \frac{1}{T} \sum_{t=1}^T \mathbb{E}(\alpha(L)\epsilon_t) = \mu$$

$$\text{Var}(\bar{y}_T) = 0 + \frac{1}{T^2} \text{Var} \left( \sum_{t=1}^T \alpha(L)\epsilon_t \right) = \frac{1}{T^2} \sum_{t=1}^T \text{Var}[\alpha(L)\epsilon_t] = \frac{1}{T^2} \sum_{t=1}^T \mathbb{E}[\alpha(L)\epsilon_t]^2 = \frac{1}{T} \alpha(1)^2 \mathbb{E}[\epsilon_t]^2$$

$$= \frac{\sigma^2}{T} \alpha(1)^2$$

Therefore a sufficient condition for consistency is

$$\lim_{T \rightarrow \infty} \frac{\sigma^2}{T} \alpha(1)^2 = 0 \iff \alpha(1)^2 < \infty \iff \boxed{\sum_{i=0}^{\infty} \alpha_i = 0}$$

(b) See section 12.4.2.

## 12.3 Some time series and their properties

### 12.3.1 White noise process:

$$x_t = \epsilon_t, \epsilon_t \sim IID(0, \sigma^2)$$

- Autocovariances:

$$\gamma(0) = \sigma^2$$

$$\gamma(h) = 0, \quad \forall h \neq 0$$

- Spectral density function:

$$f_x(\omega) = \frac{1}{2\pi} \cdot \sigma^2 = \frac{\sigma^2}{2\pi} \text{ (flat spectrum)}$$

### 12.3.2 MA(1) process:

$$x_t = \epsilon_t + \theta \epsilon_{t-1} \text{ with } \epsilon_t \sim iid(0, \sigma^2), |\rho| < 1.$$

- Autocovariances: By Equation (12.2), the autocovariance function is

$$\begin{aligned} \text{Cov}(u_t, u_{t-h}) &= \gamma(h) = \sigma^2 \sum_{i=0}^{1-|h|} a_i a_{i+|h|} \text{ if } 0 \leq |h| \leq 1 \\ &\implies \mathbb{E}(x_t^2) = \gamma(0) = (1 + \theta^2)\sigma^2 \end{aligned}$$

$$\mathbb{E}(x_t x_{t-1}) = \gamma(1) = \theta \sigma^2$$

$$\gamma(h) = 0 \quad \forall |h| > 1$$

So the covariance matrix is

$$\begin{pmatrix} \sigma^2(1+\theta^2) & \sigma^2\theta & 0 & 0 & \cdots & 0 \\ \sigma^2\theta & \sigma^2(1+\theta^2) & \sigma^2\theta & 0 & \cdots & 0 \\ 0 & \sigma^2\theta & \sigma^2(1+\theta^2) & \sigma^2\theta & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma^2\theta & \sigma^2(1+\theta^2) & \sigma^2\theta \\ 0 & 0 & \cdots & 0 & \sigma^2\theta & \sigma^2(1+\theta^2) \end{pmatrix}$$

$$= \sigma^2(1+\theta^2)I_T + \sigma^2\theta A$$

where  $A$  is defined as in section 14.3.2 (p. 304).

- Spectral density function:

$$f(\omega) = \frac{\sigma^2}{2\pi} [1 + 2\theta \cos(\omega) + \rho^2], \quad \omega \in [0, \pi]$$

### 12.3.3 MA( $\infty$ ) process:

This process is covariance stationary.

- Autocovariances:

### 12.3.4 AR(1) process:

$$x_t = \phi x_{t-1} + \epsilon_t, \quad |\phi| < 1, \quad \epsilon_t \sim IID(0, \sigma^2).$$

- Yule-Walker Equations:

$$\mathbb{E}[x_t x_{t-h}] = \mathbb{E}[\phi x_{t-1} x_{t-h}] + \mathbb{E}[\epsilon x_{t-h}]$$

$$\gamma_h = \phi \gamma_{h-1} + \mathbb{E}[\epsilon x_{t-h}]$$

$$\implies \gamma_0 = \phi \gamma_1 + \sigma^2, \quad \gamma_h = \phi \gamma_{h-1} \quad \forall h \geq 1$$

- Autocovariances:

$$\gamma(0) = \frac{\sigma^2}{1 - \phi^2}$$

$$\gamma_h = \frac{\sigma^2 \phi^h}{1 - \rho^2} = \phi^h \gamma(0) \quad \forall h \geq 1$$

$$\implies \text{Cov}(x) =$$

$$\begin{pmatrix} \sigma^2/(1-\phi^2) & \sigma^2\phi/(1-\phi^2) & \sigma^2\phi^2/(1-\phi^2) & \sigma^2\phi^3/(1-\phi^2) & \dots & \sigma^2\phi^{T-1}/(1-\phi) \\ \sigma^2\phi/(1-\phi) & \sigma^2/(1-\phi^2) & \sigma^2\phi/(1-\phi^2) & \sigma^2\phi^2/(1-\phi^2) & \dots & \sigma^2\phi^{T-2}/(1-\phi^2) \\ \sigma^2\phi^2/(1-\phi^2) & \sigma^2\phi/(1-\phi^2) & \sigma^2/(1-\phi^2) & \sigma^2\phi/(1-\phi^2) & \dots & \sigma^2\phi^{T-3}/(1-\phi^2) \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ \sigma^2\phi^{T-2}/(1-\phi^2) & \sigma^2\phi^{T-3}/(1-\phi^2) & \dots & \sigma^2\phi/(1-\phi^2) & \sigma^2/(1-\phi^2) & \sigma^2\phi/(1-\phi^2) \\ \sigma^2\phi^{T-1}/(1-\phi^2) & \sigma^2\phi^{T-2}/(1-\phi^2) & \dots & \sigma^2\phi^2/(1-\phi^2) & \sigma^2\phi/(1-\phi^2) & \sigma^2/(1-\phi^2) \end{pmatrix}$$

- If stationary, can be written as an infinite MA process with absolutely summable coefficients

$$x_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i} = \left( \frac{1}{1-\phi L} \right) \epsilon_t$$

- Autocovariance generating function:

$$G(z) = \left( \frac{\sigma^2}{1-\phi^2} \right) \left( 1 + \sum_{h=1}^{\infty} \phi^h (z^h + z^{-h}) \right)$$

- Spectral density function:

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \frac{\sigma^2 \phi^{|h|}}{(1-\phi^2)} (e^{i\omega})^h = \frac{1}{2\pi} \frac{\sigma^2}{(1-\phi e^{i\omega})(1-\phi e^{-i\omega})} = \frac{1}{2\pi} \frac{\sigma^2}{1-2\phi \cos(\omega) + \phi^2}$$

### 12.3.5 AR(2) process:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon, |\phi_1| < 1, |\phi_2| < 1, \epsilon_t \sim IID(0, \sigma^2).$$

Can be written as

$$x_t = \frac{1}{1-\phi L} \epsilon_t = \epsilon_t + \phi \epsilon_{t-1} + \phi^2 \epsilon_{t-2} + \dots$$

- Yule-Walker equations:

$$\mathbb{E}[x_t x_{t-h}] = \mathbb{E}[\phi_1 x_{t-1} x_{t-h}] + \mathbb{E}[\phi_2 x_{t-2} x_{t-h}] + \mathbb{E}[\epsilon x_{t-h}]$$

$$\gamma_h = \phi_1 \gamma_{h-1} + \phi_2 \gamma_{h-2} + \mathbb{E}[\epsilon x_{t-h}]$$

$$\implies \boxed{\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2, \quad \gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1, \quad \gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0}$$

- Autocovariances:

### 12.3.6 AR( $p$ ) process:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \epsilon, |\phi_i| < 1, \epsilon_t \sim IID(0, \sigma^2).$$

- Stationary if the eigenvalues of  $\Phi$  lie inside the unit circle, which is equivalent to all the roots of

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p = 0$$

being strictly larger than unity. Under this condition the AR process has the infinite-order MA representation'

$$x_t = \sum_{i=0}^{\infty} \alpha_i \epsilon_{t-i}$$

where  $\alpha_i = \phi_1 \alpha_{i-1} + \dots + \phi_p \alpha_{i-p}$ .

- Autocovariance generating function:

$$G(z) = \frac{\sigma^2}{\phi(z)\phi(z^{-1})}$$

### 12.3.7 ARMA(1, 1) process:

$$x_t = \phi x_{t-1} + \epsilon_t + \theta \epsilon_{t-1}, \text{ with } |\phi| < 1 \text{ (implying stationarity)}, \mathbb{E}(\epsilon_t^2) = \sigma^2, \mathbb{E}(\epsilon_t \epsilon_s) = 0 \text{ for } t \neq s.$$

- Yule-Walker Equations:

$$\gamma(0) = \phi \gamma(1) + \sigma^2(1 + \theta^2)$$

$$\gamma(1) = \phi \gamma(0) + \sigma^2 \phi^2$$

$$\gamma(h) = \phi \gamma(h-1) \quad \forall h \geq 2$$

- Autocovariances:

$$\gamma(0) = \sigma^2 \left( 1_{\frac{(\phi+\theta)^2}{1-\phi^2}} \right)$$

$$\gamma(1) = \sigma^2 \left( \phi + \theta + \frac{(\phi + \theta)^2 \phi}{1 - \phi^2} \right)$$

$$\gamma(2) = \phi^{h-1} \gamma(1) \quad \forall h \geq 2$$

- Autocorrelation function:

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \frac{(\phi+\theta)(1+\phi\theta)}{1+2\phi\theta+\theta^2} & h = 1 \\ \phi^{h-1}\rho(1) & h \geq 2 \end{cases}$$

- Autocovariance generating function: the autocovariance function of an ARMA( $p, q$ ) process  $\phi(L)y_t = \theta(L)\epsilon_t$  is given by

$$f(\omega) = \sigma^2 \frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})}$$

Plugging in for the ARMA(1,1) case yields (**double-check**)

$$f(\omega) = \sigma^2 \frac{(1+\theta)^2}{(1-\rho)^2}$$

- Spectral Density Function: the spectral density function of an ARMA( $p, q$ ) process  $\phi(L)y_t = \theta(L)\epsilon_t$  is given by

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{\theta(e^{i\omega})\theta(e^{-i\omega})}{\phi(e^{i\omega})\phi(e^{-i\omega})}, \quad \omega \in [0, 2\pi]$$

Plugging in for the ARMA(1,1) case yields

$$f_x(\omega) = \frac{\sigma^2}{2\pi} \frac{(e^{i\omega} - \theta e^{i\omega})(e^{-i\omega} - \theta e^{-i\omega})}{(e^{i\omega} - \phi e^{i\omega})(e^{-i\omega} - \phi e^{-i\omega})} = \frac{\sigma^2}{2\pi} \frac{1 - 2\theta + \theta^2}{1 - 2\phi + \phi^2}$$

- If  $\phi = \theta$ , the ARMA(1,1) process becomes a white noise process. We can see this two ways. The ARMA(1, 1) process can be represented in the following way:

$$(1 - \phi L)y_t = (1 - \theta L)\epsilon_t$$

Therefore  $\phi(L) = \theta(L)$  yields  $y_t = \epsilon_t$ .

We can also see that when  $\phi = \theta$ , an ARMA(1,1) process is equivalent to a white noise process as follows. Plugging in  $\phi = \theta$  to the spectral density function, we have

$$f_x(\omega) = \frac{\sigma^2}{2\pi} \frac{1 - 2\theta + \theta^2}{1 - 2\theta + \theta^2} = \frac{\sigma^2}{2\pi}$$

showing that if  $\theta = \phi$ , the spectral density function is constant and independent of  $\theta$  and  $\phi$ . We can see that it in fact is a white noise process. Since a white noise process has the following covariances:

$$\gamma(0) = \sigma^2$$

$$\gamma(h) = 0, \quad \forall h \neq 0$$

for a white noise process we have

$$f_x(\omega) = \frac{1}{2\pi} \cdot \sigma^2 = \frac{\sigma^2}{2\pi}$$

## 12.4 Chapter 14: Estimation of Stationary Time Series Processes

### 12.4.1 Sufficient conditions for ergodicity of mean. (Book section 14.2.1)

By Chebyshev's Inequality (see section 8.2),  $\bar{y}_T$  is a consistent estimator of  $\mu$  as  $T \rightarrow \infty$  if  $\lim_{T \rightarrow \infty} \mathbb{E}(\bar{y}_T) = \mathbb{E}(y_T) = \mu$  and  $\lim_{T \rightarrow \infty} \text{Var}(\bar{y}_T) = 0$ . We have

$$\begin{aligned}\mathbb{E}(\bar{y}_T) &= \frac{1}{T} \mathbb{E}\left(\sum_{t=1}^T y_t\right) = \frac{1}{T} \sum_{t=1}^T \mathbb{E}(y_t) = \mu \\ \text{Var}(\bar{y}_T) &= \frac{1}{T^2} \text{Var}\left(\sum_{t=1}^T y_t\right) = \frac{1}{T^2} \left( \sum_{t=1}^T \text{Var}(y_t) + 2 \sum_{0 \leq i < j \leq T} \text{Cov}(y_i, y_j) \right) \\ &= \frac{1}{T^2} \left( \sum_{t=1}^T \gamma(0) + 2 \sum_{0 \leq i < j \leq T} \gamma(j-i) \right) = \frac{1}{T^2} \left( T\gamma(0) + 2 \sum_{h=1}^{T-1} (T-h)\gamma(h) \right) \\ &= \frac{1}{T} \left[ \gamma(0) + 2 \sum_{h=1}^{T-1} \left( 1 - \frac{h}{T} \right) \gamma(h) \right] = \frac{1}{T^2} \mathbf{1}' \text{Var}(\mathbf{y}) \mathbf{1}\end{aligned}$$

where  $\mathbf{1}$  is a vector of ones and

$$\text{Var}(\mathbf{y}) = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(T-2) & \gamma(T-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(T-3) & \gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(T-2) & \gamma(T-3) & \cdots & \gamma(0) & \gamma(1) \\ \gamma(T-1) & \gamma(T-2) & \cdots & \gamma(1) & \gamma(0) \end{pmatrix}$$

Notice that

$$\left| \gamma(0) + 2 \sum_{h=1}^{T-1} \left( 1 - \frac{h}{T} \right) \gamma(h) \right| < \left| 2 \sum_{h=0}^{T-1} \gamma(h) \right| \leq 2 \sum_{h=0}^{T-1} |\gamma(h)|$$

Therefore

$$\sum_{h=0}^{\infty} |\gamma(h)| < \infty$$

is a sufficient condition for

$$\lim_{T \rightarrow \infty} \text{Var}(\bar{y}_T) = \lim_{T \rightarrow \infty} \frac{1}{T} \left[ \gamma(0) + 2 \sum_{h=1}^{T-1} \left( 1 - \frac{h}{T} \right) \gamma(h) \right] = 0$$

### 12.4.2 Estimation of autocovariances (Book section 14.2.2).

A moment estimator of  $\gamma(h) = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)]$  is

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=h+1}^T (y_t - \bar{y}_T)(y_{t-h} - \bar{y}_T)$$

By Chebyshev's Inequality (Theorem 8.2.2),  $\hat{\gamma}(h)$  is a consistent estimator of  $\gamma(h)$  as  $T \rightarrow \infty$  if  $\lim_{T \rightarrow \infty} \mathbb{E}(\hat{\gamma}(h)) = \gamma(h)$  and  $\lim_{T \rightarrow \infty} \text{Var}(\hat{\gamma}(h)) = 0$ .

$$\begin{aligned} \hat{\gamma}(h) &= \frac{1}{T} \sum_{t=h+1}^T (y_t - \bar{y}_T)(y_{t-h} - \bar{y}_T) = \frac{1}{T} \sum_{t=h+1}^T (y_t - \mu + \mu - \bar{y}_T)(y_{t-h} - \mu + \mu - \bar{y}_T) \\ &= \frac{1}{T} \sum_{t=h+1}^T (y_t - \mu)(y_{t-h} - \mu) + (y_t - \mu)(\mu - \bar{y}_T) + (\mu - \bar{y}_T)(y_{t-h} - \mu) + (\mu - \bar{y}_T)^2 \\ &= \frac{1}{T} \sum_{t=h+1}^T (y_t - \mu)(y_{t-h} - \mu) + (\mu - \bar{y}_T) \frac{1}{T} \sum_{t=h+1}^T (y_t - \mu) + (\mu - \bar{y}_T) \frac{1}{T} \sum_{t=h+1}^T (y_{t-h} - \mu) + \frac{1}{T} (T-h)(\mu - \bar{y}_T)^2 \\ &\quad \vdots \end{aligned}$$

Because where does this line come from? on page 300 of book/331 of pdf.

$$\bar{y}_T = \mu + \mathcal{O}_p(T^{-1/2})$$

and for any fixed  $h$

$$T^{-1/2} \sum_{t=h+1}^T (y_t - \mu) = \mathcal{O}_p(1)$$

it follows that

$$(\mu - \bar{y}_T) \frac{1}{T} \sum_{t=h+1}^T (y_t - \mu) = \frac{\mu}{T} \sum_{t=h+1}^T (y_t - \mu) - \frac{\bar{y}_T}{\sqrt{T}} \cdot \frac{1}{\sqrt{T}} \sum_{t=h+1}^T (y_t - \mu) = \mathcal{O}_p(T^{-1})$$

$$(\mu - \bar{y}_T) \frac{1}{T} \sum_{t=h+1}^T (y_{t-h} - \mu) = \mathcal{O}_p(T^{-1})$$

$$\frac{1}{T} (T-h)(\mu - \bar{y}_T)^2 = (\mu - \bar{y}_T)^2 - \frac{h}{T} (\mu - \bar{y}_T)^2 = \mathcal{O}_p(T^{-1})$$

$$\implies \hat{\gamma}(h) = \frac{1}{T} \sum_{t=h+1}^T (y_t - \mu)(y_{t-h} - \mu) + \mathcal{O}_p(T^{-1})$$

For  $\hat{\gamma}(h)$  to be consistent, we need

$$\frac{1}{T} \sum_{t=h_1}^T (y_t - \mu)(y_{t-h} - \mu) \xrightarrow{p} \gamma(h)$$

First we show that  $(y_t - \mu)(y_{t-h} - \mu)$  is a martingale difference process:

$$\mathbb{E}[(y_t - \mu)(y_{t-h} - \mu) | F_{t-h}] = (y_{t-h} - \mu)\mathbb{E}[y_t - \mu | F_{t-h}] = 0$$

We need to show that

$$\mathbb{E}[(y_t - \mu)^2(y_{t-h} - \mu)^2] = \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha_i \epsilon_{t-i}\right)^2 \left(\sum_{j=0}^{\infty} \alpha_j \epsilon_{t-h-j}\right)^2\right] < \infty$$

By the Cauchy-Schwarz Inequality (Theorem 8.2.4), we have

$$\begin{aligned} \mathbb{E}\left|\left[\left(\sum_{i=0}^{\infty} \alpha_i \epsilon_{t-i}\right)^2 \left(\sum_{j=0}^{\infty} \alpha_j \epsilon_{t-h-j}\right)^2\right]\right|^2 &\leq \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha_i \epsilon_{t-i}\right)^2\right]^2 \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \alpha_j \epsilon_{t-h-j}\right)^2\right]^2 \\ &< \infty \iff \mathbb{E}\left[\sum_{i=0}^{\infty} \alpha_i \epsilon_{t-i}\right]^4 < \infty, \quad \mathbb{E}\left[\sum_{j=0}^{\infty} \alpha_j \epsilon_{t-h-j}\right]^4 < \infty \end{aligned}$$

These conditions hold if  $\mathbb{E}(\epsilon_t^4) < \infty$  and  $\sum_{i=0}^{\infty} |\alpha_i| < \infty$ . Then  $\mathbb{E}[(y_t - \mu)^2(y_{t-h} - \mu)^2] < \infty$  holds and

$$\hat{\gamma} \xrightarrow{p} \gamma(h)$$

### 12.4.3 Worked examples

**Midterm Problem 3 parts (3) and (4) (similar to 14.7 and 14.8 material).** Consider the following ARMA(1, 1) model

$$y_t = \phi y_{t-1} + u_t + \theta u_{t-1}, \text{ for } t = -\infty, \dots, -1, 0, 1, \dots$$

where  $|\theta| < 1$ ,  $|\phi| < 1$ , and  $u_t$  is i.i.d. with mean zero and variance  $\sigma_u^2$ ,  $\mathbb{E}(u_t^4) < \infty$ .

- (1) Suppose that we have the data  $\{y_t : t = 0, 1, \dots, T\}$ . Consider the following estimator of  $\phi$ :

$$\hat{\phi}_T = \frac{\sum_{t=2}^T y_t y_{t-2}}{\sum_{t=2}^T y_{t-1} y_{t-2}}$$

Show that  $\hat{\phi}$  is a consistent estimator of  $\phi$  and derive the asymptotic distribution of  $\sqrt{T}(\hat{\phi}_T - \phi)$ . Comment on the case where  $\theta = \phi$ .

- (2) Suppose that  $\sigma_u^2 = 1$  is known. Show that  $\theta$  can be consistently estimated by

$$\hat{\theta}_T = \frac{1}{T} \sum_{t=1}^T y_t y_{t-1} - \frac{\hat{\phi}_T}{T} \sum_{t=1}^T y_{t-1}^2$$

**Solution.**

- (1) From the results in Question 2 part 2(b) (in section 12.2.1), since  $\mathbb{E}(y_t) = \mathbb{E}(y_{t-1}) = \mathbb{E}(y_{t-2}) = 0$ , we know that

$$\hat{\phi}_T = \frac{\sum_{t=2}^T y_t y_{t-2}}{\sum_{t=2}^T y_{t-1} y_{t-2}} = \frac{T^{-1} \sum_{t=2}^T y_t y_{t-2}}{T^{-1} \sum_{t=2}^T y_{t-1} y_{t-2}} \xrightarrow{p} \frac{\gamma(2)}{\gamma(1)}$$

By the result from Question 3 part (2), we have  $\gamma(h) = \phi\gamma(h-1)$  for  $h \geq 2$ . Therefore  $\gamma(2)/\gamma(1) = \phi$ , so  $\hat{\phi}_T$  is a consistent estimator for  $\phi$ . To obtain the asymptotic distribution, note that

$$\begin{aligned} \sqrt{T}(\hat{\phi}_T - \phi) &= \sqrt{T} \left( \frac{T^{-1} \sum_{t=2}^T y_t y_{t-2}}{T^{-1} \sum_{t=2}^T y_{t-1} y_{t-2}} - \phi \right) \\ &= \frac{T^{-1/2} \sum_{t=2}^T (\phi y_{t-1} + u_t + \theta u_{t-1}) y_{t-2}}{T^{-1} \sum_{t=2}^T y_{t-1} y_{t-2}} - \frac{\phi T^{-1/2} \sum_{t=2}^T y_{t-1} y_{t-2}}{T^{-1} \sum_{t=2}^T y_{t-1} y_{t-2}} \\ &= \frac{T^{-1/2} \sum_{t=2}^T (u_t + \theta u_{t-1}) y_{t-2}}{T^{-1} \sum_{t=2}^T y_{t-1} y_{t-2}} \end{aligned}$$

In Question 2 part 2(b) (in section 12.2.1), we showed that

$$\frac{1}{T} \sum_{t=h_1}^T (y_t - \mu)(y_{t-h} - \mu) \xrightarrow{p} \gamma(h)$$

Therefore in the denominator, since  $\mathbb{E}(y_{t-1}) = \mathbb{E}(y_{t-h}) = 0$ , we have

$$T^{-1} \sum_{t=2}^T y_{t-1} y_{t-2} \xrightarrow{p} \gamma(1)$$

In the numerator,

$$\begin{aligned} T^{-1/2} \sum_{t=2}^T (u_t + \theta u_{t-1}) y_{t-2} &= \frac{1}{\sqrt{T}} \sum_{t=2}^T [u_t y_{t-2} + \theta u_{t-1} y_{t-2}] \\ &= \frac{1}{\sqrt{T}} \sum_{t=2}^T u_t y_{t-2} + \frac{1}{\sqrt{T}} \sum_{t=2}^T \theta u_{t-1} y_{t-2} = \frac{1}{\sqrt{T}} \left( \sum_{t=2}^{T-1} u_t y_{t-2} + u_T y_{T-2} \right) + \frac{1}{\sqrt{T}} \sum_{t'=1}^{T-1} \theta u_{t'} y_{t'-1} \end{aligned}$$

$$= \frac{1}{\sqrt{T}} \left( \sum_{t=2}^{T-1} u_t y_{t-2} + u_T y_{T-2} \right) + \frac{1}{\sqrt{T}} \left( \theta u_1 y_0 + \sum_{t=2}^{T-1} \theta u_t y_{t-1} \right) = \frac{1}{\sqrt{T}} \left( \sum_{t=2}^{T-1} u_t (y_{t-2} + \theta y_{t-1}) + \theta u_1 y_0 + u_T y_{T-2} \right)$$

Since  $\mathbb{E}(u_t(y_{t-2} + \theta y_{t-1}) | F_{t-1}) = 0$ . Further,  $T^{-1/2}(\theta u_1 y_0 + u_{T-1} y_{T-2}) = o_p(1)$ . Then by the Central Limit Theorem in martingale difference processes (Theorem (8.7.1)):

**Theorem 28 (Central limit theorem for martingale difference sequences).** Let  $\{x_t\}$  be a martingale difference sequence with respect to the information set  $\Omega_t$ . Let  $\bar{\sigma}_T^2 = \text{Var}(\sqrt{T}\bar{x}_T) = T^{-1} \sum_{t=1}^T \sigma_t^2$ . If  $\mathbb{E}(|x_t|^r) < K < \infty$ ,  $r > 2$  and for all  $t$ , and

$$\frac{1}{T} \sum_{t=1}^T x_t^2 - \bar{\sigma}_T^2 \xrightarrow{p} 0$$

then

$$\sqrt{T} \cdot \frac{\bar{x}_T}{\bar{\sigma}_T} \xrightarrow{d} \mathcal{N}(0, 1)$$

we have

$$\sqrt{T} \cdot \frac{\bar{x}_T}{T^{-1/2} \sqrt{\sum_{t=1}^T \sigma_t^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

⋮

$$\frac{1}{\sigma^2} \frac{\gamma(1)^2}{(1+\theta)^2 \gamma(0) + 2\theta \gamma(1)} \sqrt{T} (\hat{\phi}_T - \phi) \xrightarrow{d} \mathcal{N}(0, 1)$$

$$\iff \sqrt{T} (\hat{\phi}_T - \phi) \xrightarrow{d} \mathcal{N} \left( 0, \sigma^2 \frac{(1+\theta)^2 \gamma(0) + 2\theta \gamma(1)}{\gamma(1)^2} \right)$$

(2) From the results of Question 2 part 2(b) (in section 12.2.1), where we showed that

$$\frac{1}{T} \sum_{t=h_1}^T (y_t - \mu)(y_{t-h} - \mu) \xrightarrow{p} \gamma(h)$$

(and since  $\mathbb{E}(y_{t-1}) = \mathbb{E}(y_{t-h}) = 0$ ,

$$T^{-1} \sum_{t=2}^T y_t y_{t-1} \xrightarrow{p} \gamma(1), \quad T^{-1} \sum_{t=2}^T y_{t-1}^2 \xrightarrow{p} \gamma(0)$$

and by the Weak Law of Large Numbers (Theorem 8.6.1) we have

$$\hat{\theta}_T = \frac{1}{T} \sum_{t=1}^T y_t y_{t-1} - \frac{\hat{\phi}_T}{T} \sum_{t=1}^T y_{t-1}^2 \xrightarrow{p} \gamma(1) - \phi \gamma(0) = \phi \gamma(0) + \theta \sigma^2 - \phi \gamma(0) = \theta$$

**Chapter 14 Question 3.** The time series  $\{y_t\}$  and  $\{x_t\}$  are independently generated according to the following schemes:

$$y_t = \lambda y_{t-1} + \epsilon_{1t}, \quad |\lambda| < 1$$

$$x_t = \rho x_{t-1} + \epsilon_{2t}, \quad |\rho| < 1$$

for  $t = 1, 2, \dots, T$ , where  $\epsilon_{1t}$  and  $\epsilon_{2t}$  are non-autocorrelated and distributed independently of each other with zero means and variances equal to  $\sigma_1^2$  and  $\sigma_2^2$  respectively. An investigator estimates the simple regression

$$y_t = \beta x_t + u_t \quad t = 1, 2, \dots, T$$

by the OLS method. Show that

$$(a) \hat{\beta} \xrightarrow{P} 0 \text{ as } T \rightarrow \infty$$

(b)

$$t_{\hat{\beta}}^2 = \frac{\hat{\beta}^2}{\widehat{\text{Var}}(\hat{\beta})} = \frac{(T-1)r^2}{1-r^2}$$

(c)

$$Tr^2 \xrightarrow{P} \frac{1+\lambda\rho}{1-\lambda\rho} \text{ as } T \rightarrow \infty$$

where  $\hat{\beta}$  is the OLS estimator of  $\beta$ ,  $\widehat{\text{Var}}(\hat{\beta})$  is the estimated variance of  $\hat{\beta}$ , and  $r$  is the sample correlation coefficient between  $x$  and  $y$ , i.e.

$$r^2 = \left( \sum_{t=1}^T x_t y_t \right)^2 \Bigg/ \left( \sum_{t=1}^T x_t^2 \sum_{t=1}^T y_t^2 \right)$$

What are the implications of these results for problems of spurious correlation in economic time series analysis?

**Solution.**

(a)

$$\hat{\beta} = \frac{\sum_{t=1}^T x_t y_t}{\sum_{t=1}^T x_t^2} = \frac{T^{-1} \sum_{t=1}^T x_t y_t}{T^{-1} \sum_{t=1}^T x_t^2}$$

By the Weak Law of Large Numbers (Theorem 8.6.1), we have

$$T^{-1} \sum_{t=1}^T x_t y_t = T^{-1} \sum_{t=1}^T x_t y_t \xrightarrow{P} \mathbb{E}(x_t y_t) = \text{Cov}(x_t, y_t) = 0$$

because of the i.i.d. distributions of  $\epsilon_1$  and  $\epsilon_{2t}$ . By the Law of Large Numbers (Theorem 8.6.4),

$$T^{-1} \sum_{t=1}^T x_t^2 \xrightarrow{a.s.} \mathbb{E}(x_t^2) = \gamma_x(0) > 0$$

Therefore

$$\hat{\beta} \xrightarrow{p} \frac{0}{\gamma_x(0)} = 0$$

(b)

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{\hat{\sigma}^2}{S_{XX}}$$

where

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{T-1} \sum_{t=1}^T (y_t - \hat{\beta}x_t)^2 = \frac{1}{T-1} \sum_{t=1}^T (y_t^2 - 2y_t \hat{\beta}x_t + \hat{\beta}^2 x_t^2) \\ &= \frac{1}{T-1} \sum_{t=1}^T y_t^2 - \frac{1}{T-1} \sum_{t=1}^T \left( 2y_t x_t \frac{\sum_{t=1}^T x_t y_t}{\sum_{t=1}^T x_t^2} \right) + \frac{1}{T-1} \sum_{t=1}^T \left( \left[ \frac{\sum_{t=1}^T x_t y_t}{\sum_{t=1}^T x_t^2} \right]^2 x_t^2 \right) \\ &= \frac{1}{T-1} \sum_{t=1}^T y_t^2 - 2 \cdot \frac{1}{T-1} \frac{(\sum_{t=1}^T x_t y_t)^2}{\sum_{t=1}^T x_t^2} + \frac{1}{T-1} \frac{(\sum_{t=1}^T x_t y_t)^2}{\sum_{t=1}^T x_t^2} \\ &= \frac{1}{T-1} \left( \sum_{t=1}^T y_t^2 - \frac{(\sum_{t=1}^T x_t y_t)^2}{\sum_{t=1}^T x_t^2} \right) \\ \implies \hat{t}^2 &= \frac{(\sum_{t=1}^T x_t y_t)^2}{(\sum_{t=1}^T x_t^2)^2} \cdot \sum_{t=1}^T x_t^2 \cdot (T-1) \Bigg/ \left( \sum_{t=1}^T y_t^2 - \frac{(\sum_{t=1}^T x_t y_t)^2}{\sum_{t=1}^T x_t^2} \right) \\ &= \frac{(\sum_{t=1}^T x_t y_t)^2}{\sum_{t=1}^T x_t^2} \cdot (T-1) \Bigg/ \left( \sum_{t=1}^T y_t^2 - \frac{(\sum_{t=1}^T x_t y_t)^2}{\sum_{t=1}^T x_t^2} \right) \\ &= \left( \sum_{t=1}^T x_t y_t \right)^2 \cdot (T-1) \Bigg/ \left( \sum_{t=1}^T y_t^2 \sum_{t=1}^T x_t^2 - \left( \sum_{t=1}^T x_t y_t \right)^2 \right) \end{aligned}$$

Note that

$$\begin{aligned} r^2 &= \left( \sum_{t=1}^T x_t y_t \right)^2 \Bigg/ \left( \sum_{t=1}^T x_t^2 \sum_{t=1}^T y_t^2 \right), \quad 1 - r^2 = \left[ \sum_{t=1}^T x_t^2 \sum_{t=1}^T y_t^2 - \left( \sum_{t=1}^T x_t y_t \right)^2 \right] \Bigg/ \left( \sum_{t=1}^T x_t^2 \sum_{t=1}^T y_t^2 \right) \\ \implies \frac{r^2}{1-r^2} &= \left( \sum_{t=1}^T x_t y_t \right)^2 \Bigg/ \left[ \sum_{t=1}^T x_t^2 \sum_{t=1}^T y_t^2 - \left( \sum_{t=1}^T x_t y_t \right)^2 \right] \end{aligned}$$

Therefore

$$\hat{t}^2 = \left( \sum_{t=1}^T x_t y_t \right)^2 \cdot (T-1) \Bigg/ \left( \sum_{t=1}^T y_t^2 \sum_{t=1}^T x_t^2 - \left( \sum_{t=1}^T x_t y_t \right)^2 \right) = \boxed{\frac{(T-1)r^2}{1-r^2}}$$

(c)

$$\begin{aligned} Tr^2 &= T \cdot \frac{(T^{-1} \sum_{t=1}^T x_t y_t)^2}{T^{-1} \sum_{t=1}^T x_t^2 \cdot T^{-1} \sum_{t=1}^T y_t^2} = \frac{T^{-1} (2 \sum_{1 \leq i < j \leq T} x_i x_j y_i y_j + \sum_{t=1}^T x_t^2 y_t^2)}{T^{-1} \sum_{t=1}^T x_t^2 \cdot T^{-1} \sum_{t=1}^T y_t^2} \\ &= \frac{2T^{-1} \sum_{1 \leq i < j \leq T} x_i x_j y_i y_j + T^{-1} \sum_{t=1}^T x_t^2 y_t^2}{T^{-1} \sum_{t=1}^T x_t^2 \cdot T^{-1} \sum_{t=1}^T y_t^2} \end{aligned}$$

Note that

$$2T^{-1} \sum_{1 \leq i < j \leq T} (x_i x_j)(y_i y_j) = \sum_{h=1}^{T-1} 2T^{-1}(T-h)(x_t x_{t-h})(y_t y_{t-h})$$

By the Weak Law of Large Numbers (Theorem 8.6.1) and the independence of  $x_t$  and  $y_t$ ,

$$\sum_{h=1}^{T-1} 2T^{-1}(T-h)(x_t x_{t-h})(y_t y_{t-h}) \approx 2T^{-1} \sum_{h=1}^{T-1} (x_t x_{t-h})(y_t y_{t-h}) \xrightarrow{P} 2 \sum_{h=1}^{\infty} \gamma_x(h) \gamma_y(h)$$

Recall that for an AR(1) process with coefficient  $\phi$ ,  $\gamma(h) = \phi^h \gamma(0)$ . (See section 12.3.4).

$$= 2 \sum_{h=1}^{\infty} \rho^h \gamma_x(0) \lambda^h \gamma_y(0) = 2\gamma_x(0) \gamma_y(0) \sum_{h=1}^{\infty} (\rho \lambda)^h$$

By the Weak Law of Large Numbers (Theorem 8.6.1) and the independence of  $x_t$  and  $y_t$ ,

$$T^{-1} \sum_{t=1}^T x_t^2 y_t^2 \xrightarrow{a.s.} \mathbb{E}(x_t^2 y_t^2) = \mathbb{E}(x_t^2) \mathbb{E}(y_t^2) = \text{Var}(x_t) \text{Var}(y_t) = \sigma_1^2 \sigma_2^2 = \gamma_x(0) \gamma_y(0)$$

Therefore in the numerator, we have

$$\begin{aligned} 2T^{-1} \sum_{1 \leq i < j \leq T} x_i x_j y_i y_j + T^{-1} \sum_{t=1}^T x_t^2 y_t^2 &\xrightarrow{P} 2\gamma_x(0) \gamma_y(0) \sum_{h=1}^{\infty} (\rho \lambda)^h + \gamma_x(0) \gamma_y(0) \\ &= \gamma_x(0) \gamma_y(0) \left( 2 \frac{\rho \lambda}{1 - \rho \lambda} + 1 \right) = \gamma_x(0) \gamma_y(0) \left( \frac{1 + \rho \lambda}{1 - \rho \lambda} \right) \end{aligned}$$

Next we examine the denominator. By the Law of Large Numbers (Theorem 8.6.4),

$$T^{-1} \sum_{t=1}^T x_t^2 \xrightarrow{a.s.} \mathbb{E}(x_t^2) = \text{Var}(x_t) = \gamma_x(0) = \sigma_2^2$$

By the Law of Large Numbers (Theorem 8.6.4),

$$T^{-1} \sum_{t=1}^T y_t^2 \xrightarrow{a.s.} \mathbb{E}(y_t^2) = \gamma_y(0) = \sigma_1^2$$

Therefore in the denominator, we have

$$T^{-1} \sum_{t=1}^T x_t^2 \cdot T^{-1} \sum_{t=1}^T y_t^2 \xrightarrow{a.s.} \gamma_x(0) \gamma_y(0) = \sigma_1^2 \sigma_2^2$$

This yields

$$Tr^2 \xrightarrow{P} \frac{1+\rho\lambda}{1-\rho\lambda}$$

**Lidan's explanation:** Because as  $T \rightarrow \infty$

$$t_{\hat{\beta}} \rightarrow \sqrt{Tr^2} = \sqrt{\frac{1+\lambda\rho}{1-\lambda\rho}}$$

so if  $\lambda\rho \approx 1$ , at very high probability we will reject the null  $\hat{\beta} = 0$  when in fact  $\hat{\beta} \xrightarrow{P} 0$ .

**My original explanation:** Because  $\hat{\beta}$  converges in probability to 0 and  $t_{\hat{\beta}}^2$  is proportional to  $r^2$  (which we would expect to be close to 0), this suggests that a regression of uncorrelated variables should result in an insignificant  $\hat{\beta}$ , but if  $r^2$  is high due to a spurious correlation, the  $\hat{\beta}$  could be found to be statistically significant even if there is no meaningful relationship between  $x$  and  $y$ .

## 12.5 Chapter 15: Unit Root Processes

Need to review concepts in this chapter.

### 12.5.1 Worked Problems

**Problem 3.** Suppose that a time series of interest can be decomposed into a deterministic trend, a random walk component, and stationary errors:

$$y_t = \alpha + \delta t + \gamma_t + v_t \quad (12.1)$$

$$\gamma_t = \gamma_{t-1} + u_t$$

with

$$v_t \sim iid \mathcal{N}(0, \sigma_v^2), \quad u_t \sim iid \mathcal{N}(0, \sigma_u^2), \quad u_t \perp v_t$$

Let  $\lambda = \sigma_u^2 / \sigma_v^2$ .

- (a) Show that under  $\lambda = 0$ ,  $y_t$  reduces to a trend stationary process.
- (b) Alternatively, suppose that  $y_t$  follows an ARIMA(0,1,1) process of the form

$$y_t = \delta + y_{t-1} + w_t \quad (12.2)$$

$$w_t = \epsilon_t + \theta \epsilon_{t-1}$$

where  $\epsilon_t$  are iid  $\mathcal{N}(0, \sigma_\epsilon^2)$ . In this case show that under  $\theta = -1$ ,  $y_t$  is a trend stationary process.

- (c) Derive a relation between  $\lambda$  and the MA(1) parameter  $\theta$ , and hence or otherwise show that a test of  $\theta = -1$  in (12.2) is equivalent to a test of  $\lambda = 0$  in (12.1).

- (d) Show that (12.2) as a characterization of (12.1) implies  $\theta < 0$ .

**Solution.**

$$(a) \lambda = 0 \implies \sigma_u^2 = 0 \implies u_t = 0 \text{ (constant)}$$

$$\implies \gamma_t = \gamma_{t-1} + 0 \iff \gamma_t = \gamma_0$$

$$\implies y_t = \alpha + \delta t + \gamma_0 + v_t = (\alpha + \gamma_0) + \delta t + v_t$$

which is trend stationary because  $d_t = (\alpha + \gamma_0) + \delta t$  is perfectly predictable, and  $y_t - d_t = v_t$  is covariance stationary.

Also note that  $\text{Var}(y_t - \delta t) = \sigma_v^2$ ,  $\text{Cov}([y_t - \delta t][y_{t-h} - \delta(t-h)]) = 0$ , which implies trend stationarity of  $y_t$ . (Recall Definition 12.2: “ $X_t$  is said to be **trend stationary** if  $y_t = X_t - d_t$  is covariance stationary, where  $d_t$  is the perfectly predictable component of  $X_t$ .”)

$$(b) \theta = -1 \implies w_t = \epsilon_t - \epsilon_{t-1} = (1 - L)\epsilon_t$$

$$\implies (1 - L)y_t = \delta + (1 - L)\epsilon_t \iff y_t = (1 - L)^{-1}\delta + \epsilon_t$$

which is trend stationary because  $d_t = (1 - L)^{-1}\delta$  is perfectly predictable, and  $y_t - d_t = \epsilon_t$  is covariance stationary.

- (c) **Lidan's solution:** From (12.1), let

$$z_t = y_t - y_{t-1} = \alpha + \delta t + \gamma_{t-1} + u_t + v_t - (\alpha + \delta(t-1) + \gamma_{t-1} + v_{t-1})$$

$$= \delta + u_t + v_t - v_{t-1}$$

From (12.2), let

$$a_t = y_t - y_{t-1} = \delta + y_{t-1} + w_t - y_{t-1} = \delta + \epsilon_t + \theta \epsilon_{t-1}$$

Calculate the autocovariances for each and set them equal (using the serial independence of  $u_t$ ,  $v_t$ , and  $\epsilon_t$  as well as the independence of  $u_t$ ,  $v_t$ , and  $\epsilon_t$  for all  $t$ ):

•

$$\gamma(0) : \text{Var}(z_t) = \text{Var}(a_t) \iff \text{Var}(u_t + v_t - v_{t-1}) = \text{Var}(\epsilon_t + \theta\epsilon_{t-1})$$

$$\iff \sigma_u^2 + 2\sigma_v^2 = \sigma_\epsilon^2(1 + \theta^2) \quad (12.3)$$

•

$$\gamma(1) : \text{Cov}(z_t, z_{t-1}) = \text{Cov}(a_t, a_{t-1}) \iff \text{Cov}(-v_{t-1}, v_{t-1}) = \text{Cov}(\theta\epsilon_{t-1}, \epsilon_{t-1})$$

$$\iff -\sigma_v^2 = \theta\sigma_\epsilon^2 \quad (12.4)$$

•

$$\gamma(h) \ (h \geq 2) : 0 = 0$$

Plugging (12.4) into (12.3) and using  $\lambda = \sigma_u^2/\sigma_v^2$  we have

$$\sigma_u^2 + 2\sigma_v^2 = -\frac{\sigma_v^2}{\theta}(1 + \theta^2) \iff \sigma_u^2 = \sigma_v^2(-1/\theta - \theta - 2) \iff \lambda = -1/\theta - \theta - 2$$

$$\iff \theta^2 + (2 + \lambda)\theta + 1 = 0 \iff \theta = \frac{-(2 + \lambda) \pm \sqrt{(2 + \lambda)^2 - 4}}{2}$$

$$\iff \boxed{\theta = \frac{-(2 + \lambda) \pm \sqrt{\lambda^2 + 4\lambda}}{2}}$$

Clearly if  $\lambda = 0$  then  $\theta = -1$ . The reverse is also true:

$$\theta = -1 \implies \lambda = -1/(-1) - (-1) - 2 = 1 + 1 - 2 = 0$$

Therefore a test of  $\theta = -1$  in (12.2) is equivalent to a test of  $\lambda = 0$  in (12.1).

**original solution:**

$$y_t = \alpha + \delta t + \gamma_{t-1} + u_t + v_t$$

$$y_{t-1} = \alpha + \delta(t-1) + \gamma_{t-1} + v_{t-1}$$

$$\implies y_t = y_{t-1} + \delta + u_t - v_{t-1} + v_t$$

Comparing this to the second expression,  $y_t = \delta + y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$ , they match if  $u_t - v_{t-1} + v_t = \epsilon_t + \theta\epsilon_{t-1}$ . Since the distributions of  $\epsilon_t$  and  $v_t$  are both i.i.d. normal, this is the case if  $\lambda = 0$  (so that  $\sigma_u^2 = 0$  and  $u_t = 0 \forall t$ ),  $\theta = -1$ , and  $\sigma_\epsilon^2 = \sigma_v^2$ .

(d) **Lidan's Solution:** By (12.4) (which follows from (12.2)),

$$-\sigma_v^2 = \theta\sigma_\epsilon^2 \implies \theta < 0$$

**My solution:** Again consider (2)

$$y_t = \delta + y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$$

which is clearly a unit root process with a drift, compared to a re-written version of (1)

$$y_t = \delta + y_{t-1} + u_t + v_t - v_{t-1}$$

which has a strong resemblance to a unit root process with a drift. These match up if  $\epsilon_t = u_t + v_t$  and  $\theta < 0$ .

#### Problem 4

**Homework 4 Problem 3.** Let  $\{u_t\}$  be an i.i.d. sequence with mean zero and variance  $\sigma^2$ , and let

$$y_t = u_1 + u_2 + \dots + u_t$$

with  $y_0 = 0$ .

(a) Show that

$$T^{-3/2} \sum_{t=1}^T y_{t-1} = T^{-1/2} \sum_{t=1}^{T-1} u_t - T^{-3/2} \sum_{t=1}^{T-1} tu_t$$

(b) Show that

$$\begin{bmatrix} T^{-1/2} \sum_{t=1}^T u_t \\ T^{-3/2} \sum_{t=1}^{T-1} y_{t-1} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}\right)$$

(c) Use the functional central limit theorem (Donsker's Theorem, Theorem 7.11.1) to show that

$$T^{-3/2} \sum_{t=1}^T y_{t-1} \xrightarrow{d} \sigma \int_0^1 W(r) dr$$

where  $W(\cdot)$  is a standard Brownian motion process.

(d) Use parts (a) - (c) to show that

$$T^{-3/2} \sum_{t=1}^T tu_t \xrightarrow{d} \sigma \cdot W(a) - \sigma \int_0^1 W(r) dr$$

**Solution.**

$$y_t = u_1 + u_2 + \dots + u_t = \sum_{i=1}^t u_i = \sum_{i=0}^t u_i$$

where the last equality follows because  $y_0 = 0 \iff u_0 = 0$ .

$$u_t \sim iid (0, \sigma^2), \quad y_0 = 0$$

(a)

$$\begin{aligned} y_T &= \sum_{t=1}^T u_t \\ T^{-3/2} \sum_{t=1}^T y_{t-1} &= T^{-3/2} \sum_{t=1}^T \left( \sum_{i=0}^{t-1} u_i \right) = T^{-3/2} \sum_{t=1}^T (T - (t-1)) u_{t-1} = T^{-3/2} \sum_{t'=0}^{T-1} (T - t') u_{t'} \\ &= T^{-3/2} \cdot T \sum_{t'=0}^{T-1} u_{t'} - T^{-3/2} \sum_{t'=0}^{T-1} t' u_{t'} = T^{-1/2} \sum_{t'=1}^{T-1} u_{t'} - T^{-3/2} \sum_{t'=1}^{T-1} t' u_{t'} \\ \implies \boxed{T^{-3/2} \sum_{t=1}^T y_{t-1}} &= T^{-1/2} \sum_{t'=1}^{T-1} u_{t'} - T^{-3/2} \sum_{t'=1}^{T-1} t' u_{t'} \end{aligned} \tag{12.5}$$

(b) By the Central Limit Theorem (Theorem 8.6.7), since  $u_t \sim (0, \sigma^2)$ , we have

$$\frac{1}{\sqrt{T}\sigma^2} \sum_{t=1}^T u_t \xrightarrow{d} \mathcal{N}(0, 1)$$

which implies

$$\boxed{\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t \xrightarrow{d} \mathcal{N}(0, \sigma^2)} \tag{12.6}$$

The distribution of  $T^{-3/2} \sum_{t=1}^T y_{t-1}$  is trickier. Note that from (12.5),

$$T^{-3/2} \sum_{t=1}^T y_{t-1} = T^{-1/2} \left( \sum_{t=0}^{T-1} u_t - \sum_{t=0}^{T-1} \frac{tu_t}{T} \right)$$

Note that  $x_t = (tu_t)/T$  is a martingale difference process because

$$\mathbb{E} \left( \frac{tu_t}{T} \mid u_{t-1}, u_{t-2}, \dots \right) = 0$$

and  $\text{Var}(x_t) = \frac{1}{T^2} \text{Var}(tu_t) = \frac{t^2}{T^2} \sigma^2 < \infty$ . By the Central Limit Theorem in martingale difference processes (Theorem (8.7.1)):

**Theorem 28 (Central limit theorem for martingale difference sequences).** Let  $\{x_t\}$  be a martingale difference sequence with respect to the information set  $\Omega_t$ . Let  $\bar{\sigma}_T^2 = \text{Var}(\sqrt{T}\bar{x}_T) = T^{-1} \sum_{t=1}^T \sigma_t^2$ . If  $\mathbb{E}(|x_t|^r) < K < \infty$  for any  $r > 2$  and for all  $t$ , and

$$\frac{1}{T} \sum_{t=1}^T x_t^2 - \bar{\sigma}_T^2 \xrightarrow{p} 0$$

then

$$\sqrt{T} \cdot \frac{\bar{x}_T}{\bar{\sigma}_T} \xrightarrow{d} \mathcal{N}(0, 1)$$

since

$$T^{-1} \sum_{t=1}^T \sigma_t^2 = T^{-1} \sum_{t=1}^T \frac{t^2}{T^2} \sigma^2 = \frac{\sigma^2}{T^3} \cdot \frac{T(T+1)(2T+1)}{6} = \frac{\sigma^2}{6} \cdot \frac{2T^2 + 3T + 1}{T^2},$$

$$\mathbb{E}(|x_t|^r) = \mathbb{E}\left(\left|\frac{tu_t}{T}\right|^r\right) = \frac{t^r}{T^r} \mathbb{E}(|u_t|^r)$$

which is finite for  $r = 4$  if  $u_t$  has finite fourth moment, and since by the Weak Law of Large Numbers (Theorem 8.6.1)

$$\begin{aligned} T^{-1} \sum_{t=1}^T x_t^2 - T^{-1} \sum_{t=1}^T \frac{t^2}{T^2} \sigma^2 &= T^{-1} \sum_{t=1}^T \frac{t^2}{T^2} (u_t^2 - \sigma^2) \\ \left| T^{-1} \sum_{t=1}^T \frac{t^2}{T^2} (u_t^2 - \sigma^2) \right| &\leq \left| T^{-1} \sum_{t=1}^T (u_t^2 - \sigma^2) \right| \leq T^{-1} \sum_{t=1}^T |u_t^2 - \sigma^2| \xrightarrow{p} 0 \\ \implies T^{-1} \sum_{t=1}^T x_t^2 - T^{-1} \sum_{t=1}^T \frac{t^2}{T^2} \sigma^2 &\xrightarrow{p} 0 \end{aligned}$$

we have (if  $u_t$  has finite fourth moment)

$$\begin{aligned} &\sqrt{T} \cdot \frac{1}{T} \left( \sum_{t=1}^T \frac{tu_t}{T} \right) \Bigg/ \sqrt{\frac{\sigma^2}{6} \cdot \frac{2T^2 + 3T + 1}{T^2}} \xrightarrow{d} \mathcal{N}(0, 1) \\ \iff T^{-3/2} \left( \sum_{t=1}^T tu_t \right) \cdot \sqrt{\frac{6}{\sigma^2} \cdot \frac{1}{2T^2 + 3T + 1}} &\xrightarrow{d} \mathcal{N}(0, 1) \implies \sqrt{T} \left( \sum_{t=1}^T tu_t \right) \cdot \sqrt{\frac{6}{\sigma^2} \cdot \frac{1}{2T^2}} \xrightarrow{d} \mathcal{N}(0, 1) \\ \iff T^{-3/2} \left( \sum_{t=1}^T \frac{tu_t}{T} \right) \cdot \sqrt{\frac{3}{\sigma^2}} &\xrightarrow{d} \mathcal{N}(0, 1) \implies T^{-1/2} \frac{\sqrt{3}}{\sigma} \sum_{t=0}^{T-1} \frac{tu_t}{T} \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{3}\right) \\ \iff T^{-1/2} \sum_{t=0}^{T-1} \frac{tu_t}{T} &\xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{3}\right) \end{aligned} \tag{12.7}$$

To get the covariance between these distributions, we have (using the serial independence of  $u_t$ )

$$\begin{aligned} \text{Cov}\left(T^{-1/2} \sum_{t=0}^{T-1} u_t, T^{-3/2} \sum_{t=0}^{T-1} t u_t\right) &= \mathbb{E}\left[T^{-1/2} \sum_{t=0}^{T-1} u_t \cdot T^{-3/2} \sum_{t=0}^{T-1} t u_t\right] = T^{-2} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1} u_t \cdot \sum_{t=0}^{T-1} t u_t\right] \\ &= T^{-2} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1} t u_t^2\right] = T^{-2} \sum_{t=0}^{T-1} t \mathbb{E}(u_t^2) = \frac{\sigma^2}{T^2} \cdot \frac{T(T-1)}{2} \rightarrow \boxed{\frac{\sigma^2}{2}} \end{aligned} \quad (12.8)$$

Putting (12.6), (12.7), and (12.8) together, we have  $T^{-3/2} \sum_{t=1}^T y_{t-1} \xrightarrow{d} \mathcal{N}(0, \sigma^2 + \sigma^2/3 - 2 \cdot \sigma^2/2) = \boxed{\mathcal{N}(0, \sigma^2/3)}.$

Lastly, to get the covariance between these distributions, we have (again using the serial independence of  $u_t$ )

$$\begin{aligned} \text{Cov}\left(T^{-1/2} \sum_{t=1}^{T-1} u_t, T^{-3/2} \sum_{t=1}^T y_{t-1}\right) &= T^{-2} \text{Cov}\left(\sum_{t=1}^{T-1} u_t, \sum_{t=1}^T (T-(t-1)) u_{t-1}\right) \\ &= T^{-2} \text{Cov}\left(\sum_{t=1}^{T-1} u_t, \sum_{t'=0}^{T-1} (T-t') u_{t'}\right) = T^{-2} \mathbb{E}\left(\sum_{t=1}^{T-1} (T-t) u_t^2\right) = \frac{(T-1)T - T(T-1)/2}{T^2} \sigma^2 \\ &= \frac{(T-1)T}{2T^2} \sigma^2 \rightarrow \boxed{\sigma^2/2} \end{aligned} \quad (12.9)$$

which yields the result.

(c) Let  $r \in [0, 1)$ ,  $t \in [0, T]$ . Define

$$R_T(r) = \frac{1}{\sigma \sqrt{T}} y_{[rT]}$$

where  $[rT]$  denotes the largest integer part of  $rT$  and  $y_{[rT]} = 0$  if  $[rT] = 0$ . That is,

$$R_T(r) = \begin{cases} 0 & 0 \leq r < 1/T \\ \frac{y_1}{\sigma \sqrt{T}} & 1/T \leq r < 2/T \\ \frac{y_2}{\sigma \sqrt{T}} & 2/T \leq r < 3/T \\ \vdots & \vdots \\ \frac{y_{T-1}}{\sigma \sqrt{T}} & (T-1)/T \leq r < 1 \end{cases}$$

We have

$$\begin{aligned} T^{-3/2} \sum_{t=1}^T y_{t-1} &= T^{-3/2} \sum_{t'=0}^{T-1} y_{t'} = T^{-3/2} \sum_{t'=0}^{T-1} (y_{t'-1} + u_{t'}) = \sigma \sum_{t'=0}^{T-1} \int_{t'/T}^{(t'+1)/T} R_T(r) dr + o_p(1) \\ &= \sigma \int_0^1 R_T(r) dr + o_p(1) \implies \sigma \int_0^1 W(r) dr \text{ as } T \rightarrow \infty \end{aligned}$$

where the last step follows from Donsker's Theorem (Theorem 7.11.1, the functional central limit theorem) and the continuous mapping theorem (Theorem 7.11.2) and  $W(r)$  is a standard Weiner process.

**Donsker's Theorem, Theorem 43, p.335, Section 15.6.3.** Let  $a \in [0, 1)$ ,  $t \in [0, T]$ , and suppose  $(J - 1)/T \leq a < J/T$ ,  $J = 1, 2, \dots, T$ . Define

$$R_T(a) = \frac{1}{\sqrt{T}} s_{[Ta]}$$

where

$$s_{[Ta]} = \epsilon_1 + \epsilon_2 + \dots + \epsilon_{[Ta]}$$

$[Ta]$  denotes the largest integer part of  $Ta$  and  $s_{[Ta]} = 0$  if  $[Ta] = 0$ . Then  $R_T(a)$  weakly converges to  $w(a)$ , i.e.,

$$R_T(a) \rightarrow w(a)$$

where  $w(a)$  is a Wiener process. Note that when  $a = 1$ ,  $R_T(1) = 1/\sqrt{T} \cdot S_{[T]} = 1/\sqrt{T} \cdot (\epsilon_1 + \epsilon_2 + \dots + \epsilon_T)$ . Since  $\epsilon_t$ 's are IID, by the central limit theorem,  $R_T(1) \rightarrow \mathcal{N}(0, 1)$ .

**Continuous Mapping Theorem (Theorem 44 of Pesaran in 15.6.3).** Let  $a \in [0, 1)$ ,  $i \in [0, n]$ , and suppose  $(J - 1)/n \leq a < J/n$ ,  $J = 1, 2, \dots, n$ . Define  $R_n(a) = n^{-1/2} S_{[n \cdot a]}$ . If  $f(\cdot)$  is continuous over  $[0, 1]$ , then

$$f[R_n(a)] \xrightarrow{d} f[w(a)]$$

- (d) We have  $T^{-1/2} \sum_{t=1}^T u_t = \sigma \cdot R_T(1) \implies \sigma \cdot W(1)$  by Donsker's Theorem. Using that and the result from (c), we have

$$T^{-3/2} \sum_{t=1}^{T-1} t u_t = T^{-1/2} \sum_{t=1}^{T-1} u_t - T^{-3/2} \sum_{t=1}^T y_{t-1} \xrightarrow{d} \sigma \cdot W(1) - \sigma \int_0^1 W(r) dr$$

## 12.6 Chapter 17: Introduction to Forecasting

Feel pretty good on concepts

### 12.6.1 17.7: Iterated and direct multi-step AR methods

Suppose  $y_t$  follows the AR(1) model:

$$y_t = a + \phi y_{t-1} + \epsilon_t, \quad |\phi| < 1, \epsilon_t \sim iid(0, \sigma_\epsilon^2) \quad (12.10)$$

$$\iff y_t = \frac{a}{1-\phi} + \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}$$

$$\iff y_t = a \left( \frac{1 - \phi^h}{1 - \phi} \right) + \phi^h y_{t-h} + \sum_{j=0}^{h-1} \phi^j \epsilon_{t-j} \quad (12.11)$$

We have two methods for forecasting  $y_{t+h}$   $h > 1$  steps ahead.

- (1) **Iterated method:** In this method, we first calculate the OLS estimates of  $\hat{a}_T$  and  $\hat{\phi}_T$  in Equation (12.12) using all available data  $\Omega_T$ . Then we use the form of Equation (12.11):

$$\hat{y}_{T+h|T}^* = \hat{a}_T \left( \frac{1 - \hat{\phi}_T^h}{1 - \hat{\phi}_T} \right) + \hat{\phi}_T^h y_T$$

- (2) **Direct method:** We directly calculate OLS estimates of the parameters in Equation (12.11) using all available data  $\Omega_T$ :

$$\tilde{y}_{T+h|T}^* = \tilde{a}_{h,T} + \tilde{\phi}_{h,T} y_T$$

**Proposition 12.6.1. (Pesaran Chapter 17 Proposition 45.)** Suppose data is generated by Equation (12.12). If  $u_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}$  and  $v_t = \sum_{j=0}^{h-1} \phi^j \epsilon_{t-j}$  are symmetrically distributed around zero and have finite second moments, and if  $\mathbb{E}(\hat{\phi}_T)$  and  $\mathbb{E}(\tilde{\phi}_{h,T})$  exist, then for any finite  $T$  and  $h$  we have

$$\mathbb{E}(\hat{y}_{T+h|T}^* - y_{T+h}) = \mathbb{E}(\tilde{y}_{T+h|T}^* - y_{T+h}) = 0$$

### 12.6.2 Worked Problems

**Problem 4 (Homework 5 Question 3—fine on all but part (c), which even Lidan is hazy on.**

Consider the AR(1) model

$$y_t = \phi y_{t-1} + u_t, \quad u_t \sim iid(0, \sigma^2) \quad (12.12)$$

- (a) Derive iterated and direct forecasts of  $y_{T+2}$  condition on  $y_T$ , and show that they can be estimated as

$$\text{Iterated: } \hat{y}_{T+2|T}^{(it)} = \hat{\phi}^2 y_T$$

$$\text{Direct: } \hat{y}_{T+2|T}^{(d)} = \hat{\phi}_2 y_T$$

where  $\hat{\phi}$  and  $\hat{\phi}_2$  are OLS coefficients in the regressions of  $y_t$  on  $y_{t-1}$  and  $y_{t-2}$ , respectively, using the  $M$  observations  $y_{T-M+1}, y_{T-M+2}, \dots, y_T$ .

- (b) Show that conditional on  $y_t$ ,

$$\mathbb{E}(y_{T+2} - \hat{y}_{T+2|T}^{(it)})^2 = \mathbb{E}(\phi^2 - \hat{\phi}^2)^2 y_T^2 + (1 + \phi^2)\sigma^2$$

$$\mathbb{E}(y_{T+2} - \hat{y}_{T+2|T}^{(d)})^2 = \mathbb{E}(\phi^2 - \hat{\phi}_2)^2 y_T^2 + (1 + \phi^2)\sigma^2$$

(c) Hence, or otherwise, show that

$$\lim_{M \rightarrow \infty} \mathbb{E}(d_{T+2}) = 0$$

where  $d_{T+2}$  is the loss differential of the two forecasting methods, defined by

$$d_{T+2} = (y_{T+2} - \hat{y}_{T+2|T}^{(it)})^2 + (y_{T+2} - \hat{y}_{T+2|T}^{(d)})^2$$

**Solution.**

(a) Note that we can write

$$\begin{aligned} y_t &= \phi[y_{t-2} + u_{t-1}] + u_t = \phi^2 y_{t-2} + u_t + \phi u_{t-1} \\ \implies y_{T+2} &= \phi^2 y_T + u_{T+2} + \phi u_{T+1} \\ \implies \mathbb{E}(y_{T+2} \mid y_{T-M+1}, y_{T-M+2}, \dots, y_T) &= \phi^2 y_T + \mathbb{E}(u_{T+2} \mid y_{T-M+1}, \dots, y_T) + \phi \mathbb{E}(u_{T+1} \mid y_{T-M+1}, \dots, y_T) \\ &= \phi^2 y_T \\ \iff \mathbb{E}(y_{T+2} \mid y_{T-M+1}, y_{T-M+2}, \dots, y_T) &= \phi^2 y_T \end{aligned} \tag{12.13}$$

If we calculate the OLS estimate  $\hat{\phi}$  in Equation (12.12), we can substitute that into Equation (12.13) to obtain the iterated estimate of  $y_{T+2}$ :

$$\hat{y}_{T+2|T}^{(it)} = \hat{\phi}^2 y_T$$

Since we have no intercept term, the OLS estimate would be simply

$$\hat{\phi} = (x'x)^{-1}x'y = \frac{\sum_{t=T-M+2}^T y_t y_{t-1}}{\sum_{t=T-M+1}^{T-1} y_t^2}$$

Alternatively, we could directly calculate the OLS estimate  $\hat{\phi}_2$  of  $\phi^2$  in Equation (12.13)

$$\hat{y}_{T+2|T}^{(d)} = \hat{\phi}_2 y_T$$

The OLS estimate would be simply

$$\hat{\phi}_2 = (x'x)^{-1}x'y = \frac{\sum_{t=T-M+3}^T y_t y_{t-2}}{\sum_{t=T-M+1}^{T-2} y_t^2}$$

(b) We have

$$\begin{aligned}
& \mathbb{E}(y_{T+2} - \hat{y}_{T+2|T}^{(it)})^2 = \mathbb{E}(y_{T+2}^2) + \mathbb{E}((\hat{y}_{T+2|T}^{(it)})^2) - 2\mathbb{E}(y_{T+2} \cdot \hat{y}_{T+2|T}^{(it)}) \\
&= \mathbb{E}((\phi^2 y_T + u_{T+2} + \phi u_{T+1})^2) + \mathbb{E}((\hat{\phi}^2 y_T)^2) - 2\mathbb{E}((\phi^2 y_T + u_{T+2} + \phi u_{T+1}) \cdot \hat{\phi}^2 y_T) \\
&= \mathbb{E}(\phi^4 y_T^2 + (u_{T+2} + \phi u_{T+1})^2 + 2\phi^2 y_T(u_{T+2} + \phi u_{T+1})) + y_T^2 \mathbb{E}(\hat{\phi}^2) - 2\phi^2 y_T \mathbb{E}(\hat{\phi}^2 y_T) \\
&= \mathbb{E}(\phi^4) y_T^2 + \mathbb{E}(u_{T+2}^2 + \phi^2 u_{T+1}^2) + y_T^2 \mathbb{E}(\hat{\phi}^2) - 2\phi^2 y_T^2 \mathbb{E}(\hat{\phi}^2) = \mathbb{E}(\phi^4) y_T^2 + (1 + \phi^2) \sigma^2 + y_T^2 \mathbb{E}(\hat{\phi}^2) - 2\phi^2 y_T^2 \mathbb{E}(\hat{\phi}^2) \\
&= y_T^2 [\mathbb{E}(\phi^4 + \hat{\phi}^2 - 2\phi^2 \hat{\phi}^2)] + (1 + \phi^2) \sigma^2 = [\mathbb{E}(\phi^2 - \hat{\phi}^2)^2 y_T^2 + (1 + \phi^2) \sigma^2] \\
& \mathbb{E}(y_{T+2} - \hat{y}_{T+2|T}^{(d)})^2 = \mathbb{E}(y_{T+2}^2) + \mathbb{E}((\hat{y}_{T+2|T}^{(d)})^2) - 2\mathbb{E}(y_{T+2} \cdot \hat{y}_{T+2|T}^{(d)}) \\
&= \mathbb{E}((\phi^2 y_T + u_{T+2} + \phi u_{T+1})^2) + \mathbb{E}((\hat{\phi}_2 y_T)^2) - 2\mathbb{E}((\phi^2 y_T + u_{T+2} + \phi u_{T+1}) \cdot \hat{\phi}_2 y_T) \\
&= \mathbb{E}(\phi^4 y_T^2 + (u_{T+2} + \phi u_{T+1})^2 + 2\phi^2 y_T(u_{T+2} + \phi u_{T+1})) + y_T^2 \mathbb{E}(\hat{\phi}_2) - 2\phi^2 y_T \mathbb{E}(\hat{\phi}_2 y_T) \\
&= \mathbb{E}(\phi^4) y_T^2 + \mathbb{E}(u_{T+2}^2 + \phi^2 u_{T+1}^2) + y_T^2 \mathbb{E}(\hat{\phi}_2) - 2\phi^2 y_T^2 \mathbb{E}(\hat{\phi}_2) = \mathbb{E}(\phi^4) y_T^2 + (1 + \phi^2) \sigma^2 + y_T^2 \mathbb{E}(\hat{\phi}_2) - 2\phi^2 y_T^2 \mathbb{E}(\hat{\phi}_2) \\
&= y_T^2 [\mathbb{E}(\phi^4 + \hat{\phi}_2 - 2\phi^2 \hat{\phi}_2)] + (1 + \phi^2) \sigma^2 = [\mathbb{E}(\phi^2 - \hat{\phi}_2)^2 y_T^2 + (1 + \phi^2) \sigma^2]
\end{aligned}$$

(c) Begin by expanding  $\hat{\phi}^2$  in a first order Taylor series about  $\phi$ :

$$\hat{\phi} = \phi + \mathcal{O}_p(M^{-1/2}) \implies \hat{\phi}^2 = \phi^2 + 2\phi(\hat{\phi} - \phi) + \mathcal{O}_p(M^{-1})$$

Then

$$\begin{aligned}
\mathbb{E}[(\hat{\phi}^2 - \phi^2)^2 | y_T] &= \mathbb{E}[(2\phi(\hat{\phi} - \phi) + \mathcal{O}_p(M^{-1}))^2 | y_T] = 4\phi^2 \mathbb{E}[(\hat{\phi} - \phi + \mathcal{O}_p(M^{-1}))^2 | y_T] \\
&= 4\phi^2 \mathbb{E}[(\hat{\phi} - \phi + \mathcal{O}_p(M^{-1}))^2 | y_T]
\end{aligned}$$

## 12.7 Chapter 18: Measurement and Modeling of Volatility

Maybe review a little, but feel pretty okay on concepts

GARCH(1, 1) model (Pesaran Equation 18.5):

$$h_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \phi_1 h_{t-1}^2, \quad \alpha_0 > 0 \quad (12.14)$$

- This process is unconditionally stationary if  $|\alpha_1 + \phi_1| < 1$ .
- The unconditional variance exists and is fixed if  $|\alpha_1 + \phi_1| < 1$ .
- The case where  $\alpha_1 + \phi_1 = 1$  is known as the Integrated GARCH(1,1), or IGARCH(1,1) for short. The RiskMetrics exponentially weighted formulation of  $h_t^2$  for large  $H$  is a special case of the IGARCH(1,1) model where  $\alpha_0$  is set to 0. RiskMetrics formulation avoids the variance non-existence problem by focusing on  $H$  fixed.

### 12.7.1 Higher order GARCH models (Pesaran Section 18.4.2)

The various members of the GARCH and GARCH-M class of models can be written compactly as

$$y_t = \beta' \mathbf{x}_{t-1} + \gamma h_t^2 + \epsilon_t \quad (12.15)$$

where

$$h_t^2 = \text{Var}(\epsilon_t | \Omega_{t-1}) = \mathbb{E}(\epsilon_t^2 | \Omega_{t-1}) = \alpha_0 + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^p \phi_i h_{t-i}^2 \quad (12.16)$$

and  $\Omega_{t-1}$  is the information set at time  $t - 1$  containing at least  $(\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \dots, y_{t-1}, y_{t-2}, \dots)$ . The unconditional variance of  $\epsilon_t$  is determined by

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^q \alpha_i \sigma_{t-i}^2 + \sum_{i=1}^p \phi_i \sigma_{t-i}^2$$

and yields a stationary outcome if all the roots of

$$1 - \sum_{i=1}^q \alpha_i \lambda^i + \sum_{i=1}^p \phi_i \lambda^i = 0$$

lie outside the unit circle. In that case

$$\text{Var}(\epsilon_t) = \sigma^2 = \frac{\alpha_0}{1 - \sum_{i=1}^q \alpha_i - \sum_{i=1}^p \phi_i} > 0 \quad (12.17)$$

Clearly the necessary condition for (12.16) to be covariance stationary is given by

$$\sum_{i=1}^q \alpha_i + \sum_{i=1}^p \phi_i < 1$$

### 12.7.2 Testing for GARCH effects (Pesaran Section 18.5.1)

If we consider  $y_t$  as periodic data defined by  $y_t = r_t - \bar{r}$  with  $r_t$  representing, say, asset return and  $\bar{r}$  representing the unconditional mean, then we have the GARCH(1,1) representation of volatility:

$$\text{Var}(y_t | \Omega_{t-1}) = h_t^2 = \bar{\sigma}^2(1 - \alpha - \beta) + \alpha y_{t-1}^2 + \beta h_{t-1}^2$$

Then the test for GARCH effects would test

$$H_0 : \alpha = 0$$

against

$$H_1 : \alpha \neq 0$$

GARCH(1,1) can be approximated by ARCH( $q$ ):

$$\begin{aligned} \text{Var}(y_t | \Omega_{t-1}) &= \frac{\bar{\sigma}^2(1 - \alpha - \beta)}{1 - \beta} + \alpha y_{t-1}^2 + \alpha \beta y_{t-2}^2 + \dots + \alpha \beta^{q-1} y_{t-q}^2 \\ &= \tilde{\alpha}_0 + \tilde{\alpha}_1 y_{t-1}^2 + \tilde{\alpha}_2 z_{t-2}^2 + \dots + \tilde{\alpha}_q z_{t-q}^2 \end{aligned}$$

which means that we can approximate this hypothesis test by instead using the Lagrange multiplier test proposed by Engle:

$$H_0 : \tilde{\alpha}_1 = \tilde{\alpha}_2 = \dots = \tilde{\alpha}_q$$

against

$$H_1 : \tilde{\alpha}_1 \neq 0, \tilde{\alpha}_2 \neq 0, \dots, \tilde{\alpha}_q \neq 0$$

### 12.7.3 Worked Problems

**Problem 1.** Consider the generalized autoregressive heteroskedastic model

$$y_t = h_t z_t$$

where

$$z_t \mid \Omega_{t-1} \sim IID\mathcal{N}(0, 1) \quad (12.18)$$

$$h_t^2 = \text{Var}(y_t \mid \Omega_{t-1}) = \mathbb{E}(y_t^2 \mid \Omega_{t-1}) = \bar{\sigma}^2(1 - \alpha - \beta) + \alpha y_{t-1}^2 + \beta h_{t-1}^2 \quad (12.19)$$

and  $\Omega_t$  is the information set that contains at least  $y_t$  and its lagged values.

- (a) Derive the conditions under which  $\{y_t\}$  is a stationary process.
- (b) Are the observations  $\{y_t\}$  serially independent and/or serially uncorrelated?
- (c) Develop a test of the GARCH effect and discuss the estimation of the above model by the maximum likelihood method.
- (d) Discuss the relevance of GARCH models for the analysis of financial time series data.

### Solution.

- (a) Note that (using the fact that  $y_t$  and  $h_t$  are conditionally independent given  $\Omega_{t-1}$ )

$$\mathbb{E}(y_t) = \mathbb{E}(h_t z_t) = \mathbb{E}[\mathbb{E}(h_t z_t \mid \Omega_{t-1})] = \mathbb{E}[\mathbb{E}(h_t \mid \Omega_{t-1})\mathbb{E}(z_t \mid \Omega_{t-1})] = \mathbb{E}[\mathbb{E}(h_t \mid \Omega_{t-1}) \cdot 0] = 0$$

We have

$$\begin{aligned} \text{Var}(y_t) &= \mathbb{E}[(y_t - \mathbb{E}(y_t))^2] = \mathbb{E}[y_t^2] = \mathbb{E}[\mathbb{E}(y_t^2 \mid \Omega_{t-1})] = \mathbb{E}(h_t^2) = \mathbb{E}(\bar{\sigma}^2(1 - \alpha - \beta) + \alpha y_{t-1}^2 + \beta h_{t-1}^2) \\ &= \mathbb{E}(\bar{\sigma}^2(1 - \alpha - \beta)) + \alpha \mathbb{E}(y_{t-1}^2) + \beta \mathbb{E}(h_{t-1}^2) = \bar{\sigma}^2(1 - \alpha - \beta) + (\alpha + \beta)\text{Var}(y_{t-1}) \end{aligned}$$

Therefore in order for this to be a stationary process, we require  $|\alpha + \beta| < 1$ .

### more stuff I did on original homework

By Definition 12.1  $\{y_t\}$  is a stationary process if it has constant mean and its covariance function depends only on the absolute difference  $|t_1 - t_2|$ ; that is,

$$\text{Cov}(y_{t_1}, y_{t_2}) = \gamma(t_1, t_2) = \gamma(|t_1 - t_2|) \text{ for all } t_1, t_2$$

⋮

In order for this to be true, we must first show that  $h_t$  is finite for all  $t$ . It is sufficient to find the conditions that make  $h_t$  stationary. Using Equation (12.16)

$$h_t^2 = \alpha_0 + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^p \phi_i h_{t-i}^2$$

From equation (12.19) we have

$$h_t^2 = \bar{\sigma}^2(1 - \alpha - \beta) + \alpha y_{t-1}^2 + \beta h_{t-1}^2$$

Therefore we note that  $p = q = 1$  and we have a GARCH(1,1) model. From Equation (12.14):

$$h_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \phi_1 h_{t-1}^2, \quad \alpha_0 > 0$$

is unconditionally stationary if  $|\alpha_1 + \phi_1| < 1$ . In this case, we require  $|\alpha + \beta| < 1$  for stationarity of  $h_t$ .

⋮

Assume without loss of generality that  $t_2 \geq t_1$ . (Note that this implies  $z_{t_2} \perp\!\!\!\perp z_{t_1}, h_{t_2}, h_{t_1} \mid \Omega_{t_2-1}$ .)

$$\begin{aligned} \text{Cov}(y_{t_1}, y_{t_2}) &= \mathbb{E}[(y_{t_1} - \mathbb{E}(y_{t_1}))(y_{t_2} - \mathbb{E}(y_{t_2}))] = \mathbb{E}[y_{t_1} y_{t_2}] = \mathbb{E}[h_{t_1} h_{t_2} z_{t_1} z_{t_2}] \\ &= \mathbb{E}[\mathbb{E}(h_{t_1} h_{t_2} z_{t_1} z_{t_2} \mid \Omega_{t_2-1})] = \mathbb{E}[\mathbb{E}(h_{t_1} h_{t_2} z_{t_1} \mid \Omega_{t_2-1}) \mathbb{E}(z_{t_2} \mid \Omega_{t_2-1})] \\ &= \mathbb{E}[\mathbb{E}(h_{t_1} h_{t_2} z_{t_1} \mid \Omega_{t_2-1}) \cdot 0] = [0] \end{aligned}$$

Therefore given that  $h_t$  is finite (that is, given  $|\alpha + \beta| < 1$ ), we have that  $\mathbb{E}(y_t) = 0$ ,  $\text{Cov}(y_{t_1}, y_{t_2}) = 0 \quad \forall t_1, t_2$  which implies that under these conditions  $y_t$  is stationary.

- (b)  $y_t$  is serially uncorrelated because  $\text{Cov}(y_{t_1}, y_{t_2}) = 0 \quad \forall t_1, t_2$ . However, it is clear that  $y_t$  is not serially independent since past values of  $y_t$  affect  $h_t^2 = \text{Var}(y_t)$ . Observe that

$$\Pr(y_t \leq y \mid y_{t-1}) = \Pr(h_t z_t \leq y \mid y_{t-1}) = \Pr\left(z_t \sqrt{\bar{\sigma}^2(1 - \alpha - \beta) + \alpha y_{t-1}^2 + \beta h_{t-1}^2} \leq y \mid y_{t-1}\right)$$

In other words, even though the mean of  $y_t$  remains constant, because  $y_{t-1}$  affects the variance of  $y_t$ , it affects the heaviness of the tails of  $y_t$ , changing the probability distribution of  $y_t$ . Therefore the conditional cumulative distribution function of  $y_t$  given  $y_{t-1}$  is not equal to the unconditional cdf, so  $y_t$  and  $y_{t-1}$  are not independent.

**Lidan's Explanation:** From (a) we know that  $\{y_t\}$  is serially uncorrelated, but they are not necessarily independent because from Equation (12.19) we know that  $h_t$  is not independent from  $h_{t-1}$ .

- (c) See section 12.7.2.
- (d) **Lidan:** Usually financial time series data are fat-tailed. So the series may not look stationary, and therefore the local variance would be clustered in some very low and very high values. To capture this serial correlation and heterogeneity in volatility, we need to use a GARCH model.

**Book:** In financial econometrics, ARCH and GARCH are fundamental tools for analyzing the time-variation of conditional variance. In many applications in finance, the assumption that the conditional variance of the disturbances is constant over time is not valid. GARCH models allow for time variation in volatility, relating (unobserved) volatility to squares of past innovations in price changes. However, this approach only partly overcomes the deficiency of the historical measure and continues to respond very slowly when volatility undergoes rapid changes.

## 12.8 Chapter 21: Vector Autoregressive Models

Feel ok on concepts.

### 12.8.1 Worked Problems

**Problem 1.** Consider the bivariate autoregressive model:

$$Y_t = \Phi Y_{t-1} + U_t, \quad U_t \sim IID\mathcal{N}(0, \Sigma) \quad (12.20)$$

where

$$Y_t = (y_{1t}, y_{2t})', \quad Y_{t-1} = (y_{1,t-1}, y_{2,t-1})', \quad U_t = (u_{1t}, u_{2t})'$$

and

$$\Phi = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \succ 0$$

- (a) Derive the conditional mean and variance of  $y_{1t}$  with respect to  $y_{2t}$  and lagged values of  $y_{1t}$  and  $y_{2t}$ .
- (b) Show that the univariate representation of  $y_{1t}$  is an ARMA(2,1) process.

### Solution

- (a) Note that

$$y_{1t} = \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + u_{1t} \quad (12.21)$$

$$y_{2t} = \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + u_{2t} \quad (12.22)$$

First,

$$\mathbb{E}(y_{1t} | y_{2t}, y_{2,t-1}, y_{2,t-2}, \dots, y_{1,t-1}, y_{1,t-2}, \dots) = \mathbb{E}(y_{1t} | y_{2t}, \Omega_{t-1})$$

$$\begin{aligned}
&= \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \mathbb{E}(u_{1t} | y_{2t}) = \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \mathbb{E}(u_{1t} | \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + u_{2t}) \\
&= \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \mathbb{E}(u_{1t} | u_{2t})
\end{aligned}$$

Recall Proposition 6.3.29: for a bivariate normal distribution with mean 0, the conditional distribution of  $u_{1t}$  given  $u_{2t}$  is

$$u_{1t} | u_{2t} \sim \mathcal{N}\left(\rho \frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}} u_{2t}, (1 - \rho^2)\sigma_{11}\right)$$

where  $\rho = \sigma_{12}/\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$ .

$$\begin{aligned}
&\iff u_{1t} | u_{2t} \sim \mathcal{N}\left(\frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} \cdot \frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}} u_{2t}, \left[1 - \left(\frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}\right)^2\right]\sigma_{11}\right) \\
&\iff u_{1t} | u_{2t} \sim \mathcal{N}\left(\frac{\sigma_{12}}{\sigma_{22}} u_{2t}, \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right) \\
&\implies \mathbb{E}(u_{1t} | u_{2t}) = \frac{\sigma_{12}}{\sigma_{22}} u_{2t}, \quad \text{Var}(u_{1t} | u_{2t}) = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} \\
&\implies \mathbb{E}(y_{1t} | y_{2t}, \Omega_{t-1}) = \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \frac{\sigma_{12}}{\sigma_{22}} u_{2t} \\
&= \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \frac{\sigma_{12}}{\sigma_{22}}(y_{2t} - \phi_{21}y_{1,t-1} - \phi_{22}y_{2,t-1}) \\
&= \boxed{\left(\phi_{11} - \frac{\sigma_{12}}{\sigma_{22}}\phi_{21}\right)y_{1,t-1} + \left(\phi_{12} - \frac{\sigma_{12}}{\sigma_{22}}\phi_{22}\right)y_{2,t-1} + \frac{\sigma_{12}}{\sigma_{22}}y_{2t}}
\end{aligned}$$

Second,

$$\begin{aligned}
\text{Var}(y_{1t} | y_{2t}, \Omega_{t-1}) &= \phi_{11}^2 \text{Var}(y_{1,t-1} | y_{2t}, \Omega_{t-1}) + \phi_{12}^2 \text{Var}(y_{2,t-1} | y_{2t}, \Omega_{t-1}) + \text{Var}(u_{1t} | y_{2t}, \Omega_{t-1}) \\
&= \text{Var}(u_{1t} | u_{2t}) = \boxed{\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}}
\end{aligned}$$

(b) Again, using equations (12.21) and (12.22),

$$\begin{aligned}
y_{1t} &= \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + u_{1t}, \quad y_{2t} = \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + u_{2t} \\
&\implies (1 - \phi_{11}L)y_{1t} = \phi_{12}Ly_{2t} + u_{1t} \tag{12.23}
\end{aligned}$$

$$(1 - \phi_{22}L)y_{2t} = \phi_{21}Ly_{1t} + u_{2t} \tag{12.24}$$

We can multiply  $(1 - \phi_{22}L)$  on both sides of (12.23) to yield

$$(1 - \phi_{22}L)(1 - \phi_{11}L)y_{1t} = \phi_{12}L(1 - \phi_{22}L)y_{2t} + (1 - \phi_{22}L)u_{1t}$$

Using (12.24), we have

$$\begin{aligned} & (1 - (\phi_{22} + \phi_{11})L + \phi_{11}\phi_{22}L^2)y_{1t} = \phi_{12}L[\phi_{21}Ly_{1t} + u_{2t}] + (1 - \phi_{22}L)u_{1t} \\ \iff & (1 - (\phi_{22} + \phi_{11})L + (\phi_{11}\phi_{22} - \phi_{12}\phi_{21})L^2)y_{1t} = \phi_{12}Lu_{2t} + (1 - \phi_{22}L)u_{1t} \end{aligned} \quad (12.25)$$

If we can show that the left side of (12.25) is an AR(2) process and the right side is an MA(1) process, we are done. **Lidan:** “It is obvious that the left hand side of (12.25) is an AR(2) process.” The left side is a stationary AR(2) process if the absolute values of all the roots of

$$1 - (\phi_{22} + \phi_{11})z + (\phi_{11}\phi_{22} - \phi_{12}\phi_{21})z^2 = 0$$

are greater than 1. The roots are

$$\begin{aligned} z &= \frac{\phi_{22} + \phi_{11} \pm \sqrt{(\phi_{22} + \phi_{11})^2 - 4(\phi_{11}\phi_{22} - \phi_{12}\phi_{21})}}{2(\phi_{11}\phi_{22} - \phi_{12}\phi_{21})} = \frac{\phi_{22} + \phi_{11} \pm \sqrt{\phi_{22}^2 + \phi_{11}^2 - 2\phi_{22}\phi_{11} + 4\phi_{12}\phi_{21}}}{2(\phi_{11}\phi_{22} - \phi_{12}\phi_{21})} \\ &= \frac{\phi_{22} + \phi_{11} \pm \sqrt{(\phi_{22} - \phi_{11})^2 + 4\phi_{12}\phi_{21}}}{2(\phi_{11}\phi_{22} - \phi_{12}\phi_{21})} \\ &\vdots \end{aligned}$$

To check if the right side of (12.25) is MA(1), we will write  $x_t = \phi_{12}Lu_{2t} + (1 - \phi_{22}L)u_{1t}$  as an MA(1) process; that is,

$$x_t = \phi_{12}u_{2,t-1} + u_{1t} - \phi_{22}u_{1,t-1} = \xi_t + \theta\xi_{t-1}$$

with  $\xi_t \sim iid(0, \sigma_\xi^2)$ ,  $|\theta| < 1$ . If  $x_t$  is an MA(1) process, it must satisfy

$$\gamma(0) = \mathbb{E}(x_t^2) = (1 + \theta^2)\sigma_\xi^2, \quad \gamma(1) = \mathbb{E}(x_t x_{t-1}) = \theta\sigma_\xi^2$$

We have (using the serial independence of the  $u_{1t}, u_{2t}$ )

$$\begin{aligned} \mathbb{E}(x_t^2) &= \mathbb{E}([\phi_{12}u_{2,t-1} + u_{1t} - \phi_{22}u_{1,t-1}]^2) = \mathbb{E}(\phi_{12}^2u_{2,t-1}^2 - 2\phi_{12}\phi_{22}u_{2,t-1}u_{1,t-1} + \phi_{22}^2u_{1,t-1}^2 + u_{1t}^2) \\ &= \phi_{12}^2\mathbb{E}(u_{2,t-1}^2) - 2\phi_{12}\phi_{22}\mathbb{E}(u_{2,t-1}u_{1,t-1}) + \phi_{22}^2\mathbb{E}(u_{1,t-1}^2) + \mathbb{E}(u_{1t}^2) \\ &= \phi_{12}^2\sigma_{22} - 2\phi_{12}\phi_{22}\sigma_{12} + \phi_{22}^2\sigma_{11} + \sigma_{11} = (1 + \phi_{22}^2)\sigma_{11} - 2\phi_{12}\phi_{22}\sigma_{12} + \phi_{12}^2\sigma_{22} \end{aligned}$$

$$\mathbb{E}(x_t x_{t-1}) = \mathbb{E}([\phi_{12}u_{2,t-1} + u_{1t} - \phi_{22}u_{1,t-1}][\phi_{12}u_{2,t-2} + u_{1,t-1} - \phi_{22}u_{1,t-2}])$$

$$= \mathbb{E}(\phi_{12}u_{2,t-1}u_{1,t-1} - \phi_{22}u_{1,t-1}^2) = \phi_{12}\sigma_{12} - \phi_{22}\sigma_{11}$$

Therefore we require

$$(1 + \theta^2)\sigma_\xi^2 = (1 + \phi_{22}^2)\sigma_{11} - 2\phi_{12}\phi_{22}\sigma_{12} + \phi_{12}^2\sigma_{22} \quad (12.26)$$

$$\theta\sigma_\xi^2 = \phi_{12}\sigma_{12} - \phi_{22}\sigma_{11} \quad (12.27)$$

Dividing (12.26) by (12.27) we have

$$\frac{1 + \theta^2}{\theta} = \frac{(1 + \phi_{22}^2)\sigma_{11} - 2\phi_{12}\phi_{22}\sigma_{12} + \phi_{12}^2\sigma_{22}}{\phi_{12}\sigma_{12} - \phi_{22}\sigma_{11}} \quad (12.28)$$

For simplicity, let the right side of (12.28) be  $A \in \mathbb{R}$ . Then we have

$$\theta^2 - A\theta + 1 = 0 \iff \boxed{\theta_1 = \frac{1}{2}(A + \sqrt{A^2 - 4}), \theta_2 = \frac{1}{2}(A - \sqrt{A^2 - 4})}$$

Then the corresponding  $\sigma_\xi^2$  are

$$\boxed{\sigma_{\xi,1}^2 = \sigma_{\xi,2}^2 = \frac{\phi_{12}\sigma_{12} - \phi_{22}\sigma_{11}}{\theta_1}}$$

As a double check,

$$\mathbb{E}(x_t x_{t-2}) = \mathbb{E}([\phi_{12}u_{2,t-1} + u_{1t} - \phi_{22}u_{1,t-1}][\phi_{12}u_{2,t-3} + u_{1,t-2} - \phi_{22}u_{1,t-3}]) = 0$$

as expected for an MA(1) process. Therefore the right side of (12.25) is an MA(1) process (provided that  $0 < |\theta_1| < 1$  and/or  $0 < |\theta_2| < 1$ ). Since the left side is an AR(2) process, this proves that the univariate representation (12.25) of  $y_{1t}$  is an ARMA(2,1) process.

**Problem 2.** Consider the VAR(2) model in the  $m$ -dimensional vector  $Y_t$ :

$$Y_t = \mu + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + U_t, \quad U_t \sim (0, \Sigma) \quad (12.29)$$

where  $\mu$  is an  $m$ -vector of fixed constants.

- (a) Derive the conditions under which the VAR(2) model defined in (12.29) is stationary.
- (b) Derive the error correction form of (12.29) and discuss what is meant by the process  $Y_t$  being cointegrated.
- (c) Suppose now that one or more elements of  $Y_t$  is I(1). Derive suitable restrictions on the intercepts  $\mu$  such that despite the I(1) nature of the variables in (12.29),  $Y_t$  has a fixed mean. Discuss the importance of such restrictions for the analysis of cointegration.

**Solution.**

(a) Let

$$Y_t^* = Y_t - (I - \Phi_1 - \Phi_2)^{-1}\mu. \quad (12.30)$$

Then

$$\begin{aligned} Y_t^* &= \mu + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + U_t - (I - \Phi_1 - \Phi_2)^{-1}\mu \\ &= \mu + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + U_t - [(I - \Phi_1 - \Phi_2) + \Phi_1 + \Phi_2](I - \Phi_1 - \Phi_2)^{-1}\mu \\ &= \mu - (I - \Phi_1 - \Phi_2)(I - \Phi_1 - \Phi_2)^{-1}\mu + \Phi_1(Y_{t-1} - (I - \Phi_1 - \Phi_2)^{-1}\mu) + \Phi_2(Y_{t-2} - (I - \Phi_1 - \Phi_2)^{-1}\mu) + U_t \\ &= \Phi_1 Y_{t-1}^* + \Phi_2 Y_{t-2}^* + U_t \end{aligned} \quad (12.31)$$

Equation (12.31) can be rewritten in the companion form as follows:

$$\begin{bmatrix} Y_t^* \\ Y_{t-1}^* \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{bmatrix} \begin{bmatrix} Y_{t-1}^* \\ Y_{t-2}^* \end{bmatrix} + \begin{bmatrix} U_t \\ 0 \end{bmatrix} \quad (12.32)$$

which is a VAR(1) model. If  $Y_{-M+1}, Y_{-M+2}$  are given, equation (12.32) can be solved iteratively from  $t = -M + 2$  to obtain

$$\begin{bmatrix} Y_t^* \\ Y_{t-1}^* \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{bmatrix}^{t+M-2} \begin{bmatrix} Y_{-M+2}^* \\ Y_{-M+1}^* \end{bmatrix} + \sum_{j=0}^{t+M-3} \begin{bmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{bmatrix}^j \begin{bmatrix} U_{t-j} \\ 0 \end{bmatrix} \quad (12.33)$$

Then the condition for (12.33) to be covariance stationary is for all the eigenvalues of

$$\Phi = \begin{bmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{bmatrix}$$

to lie inside the unit circle; that is, the solutions of

$$|\Phi - \lambda I_4| = 0$$

must satisfy  $|\lambda| < 1$ . Equivalently, the stability condition can be written in terms of the roots of the determinantal equation

$$|I_2 - \Phi_1 z - \Phi_2 z^2| = 0$$

in which case the process  $Y_t^*$  will be stationary if all the roots lie outside the unit circle ( $|z| > 1$ ). Then if  $Y_t^*$  is stationary, so is  $Y_t$ , so either of these conditions (plus the invertibility of  $(I - \Phi_1 - \Phi_2)$ ) ensure stationarity of  $Y_t$ .

(b) From (12.31) we have

$$Y_t^* = \Phi_1 Y_{t-1}^* + \Phi_2 Y_{t-2}^* + U_t$$

Note that (letting  $\Delta Y_t^* = Y_t^* - Y_{t-1}^*$ )

$$Y_t^* - Y_{t-1}^* + Y_{t-1}^* = \Phi_1 Y_{t-1}^* + \Phi_2(Y_{t-2}^* - Y_{t-1}^* + Y_{t-1}^*) + U_t$$

$$\Leftrightarrow \Delta Y_t^* + Y_{t-1}^* = \Phi_1 Y_{t-1}^* + \Phi_2(Y_{t-1}^* - \Delta Y_{t-1}^*) + U_t$$

$$\Leftrightarrow \Delta Y_t^* = -(I - \Phi_1 - \Phi_2)Y_{t-1}^* - \Phi_2 \Delta Y_{t-1}^* + U_t \Leftrightarrow \boxed{\Delta Y_t^* = -\Pi Y_{t-1}^* + \Gamma \Delta Y_{t-1}^* + U_t}$$

where  $\Pi = (I - \Phi_1 - \Phi_2)$  and  $\Gamma = -\sum_{i=2}^2 \Phi_i = -\Phi_2$ . Because

$$\Delta Y_t^* = Y_t^* - Y_{t-1}^* = Y_t - (I - \Phi_1 - \Phi_2)^{-1}\mu - [Y_{t-1} - (I - \Phi_1 - \Phi_2)^{-1}\mu] = Y_t - Y_{t-1} = \Delta Y_t,$$

this can also be written as

$$\begin{aligned} \Delta Y_t &= -\Pi[Y_{t-1} - (I - \Phi_1 - \Phi_2)^{-1}\mu] + \Gamma \Delta Y_{t-1} + U_t \\ &= -\Pi[Y_{t-1} - \Pi^{-1}\mu] + \Gamma \Delta Y_{t-1} + U_t = \Pi \cdot \Pi^{-1}\mu - \Pi Y_{t-1} + \Gamma \Delta Y_{t-1} + U_t \\ &\Rightarrow \boxed{\Delta Y_t = \mu - \Pi Y_{t-1} + \Gamma \Delta Y_{t-1} + U_t} \end{aligned} \quad (12.34)$$

For the definition of cointegration, see Definition 12.3 and Section 12.9.1. In this particular case, if  $Y_{t-1}^* \sim I(1)$  and the linear combinations  $\Pi Y_{t-1}^*$  of  $Y_{t-1}^*$  are covariance stationary (that is,  $\Pi Y_{t-1}^* \sim I(0)$ ), we say  $Y_t^*$  is cointegrated (and therefore so is  $Y_t$ ).

(c) Since (12.34) is  $I(0)$ , for  $Y_t$  to have fixed mean, take expectations on both sides of (12.34):

$$\mu - (I_m - \Phi_1 - \Phi_2)\mathbb{E}(Y_{t-1}) = 0$$

If this restriction is violated, then (12.34) becomes a stationary process with a drift, which implies (12.29) has a time trend.

## 12.9 Chapter 22: Cointegration Analysis

Feel pretty good except for long run effects, examples we went over in class, the restrictions, and the 5 cases. Don't need to understand SURE.

### 12.9.1 22.4 Cointegrating VAR: multiple cointegrating relations and 22.5: Identification of long-run effects

**Definition 12.3.** We say that the  $m$  variables in  $Y_t$  are *cointegrated* if they are individually integrated (or have a random walk component) but there exist linear combinations of them which are stationary. That is,  $y_{it} \sim I(1)$  for  $i = 1, 2, \dots, m$ , but there exists an  $m \times r$  matrix  $\beta$  such that  $\beta' Y_t = \xi_t \sim I(0)$ .

- In this case  $r$  denotes the number of cointegrating vectors, also known as the dimension of the cointegration space.
- The cointegrating relations summarized in the  $r \times 1$  vector  $\beta' Y_t$  are also known as long-run relations.
- $r = \text{rank}(\Pi)$  is the dimension of the cointegration space.
- Cointegration is present if  $\Pi$  is rank-deficient; that is,  $r < m$ .

When  $\text{rank}(\Pi) = r < m$ , we can write  $\Pi$  as

$$\Pi = \alpha\beta' \quad (12.35)$$

where  $\alpha$  and  $\beta$  are  $m \times r$  matrices of full column rank. Then

$$\Pi y_{t-1} = \alpha\beta' y_{t-1} \sim I(0)$$

and the VECM can be written as

$$\Delta y_t = -\alpha\beta' y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t \quad (12.36)$$

Since  $\alpha$  is full rank, we have

$$\beta' y_{t-1} \sim I(0)$$

where  $\beta' y_t$  is the  $r$ -vector of cointegrating relations, also known as the long-run relations.

However,  $\beta$  as defined above is not uniquely determined. Consider a linear transformation of  $\beta$  by a non-singular  $r \times r$  matrix  $Q$ :  $\tilde{\beta} = \beta Q$ . Then since  $\text{rank}(\Pi) = r < m$ ,  $\Pi$  can be expressed as  $\Pi = \alpha\beta'$  where  $\alpha$  is a rank  $r$   $m \times r$  matrix. But consider that

$$\Pi = \alpha\beta = (\alpha Q'^{-1})(Q'\beta') = \tilde{\alpha}\tilde{\beta}'$$

with  $\tilde{\alpha} = \alpha Q'^{-1}$ . Therefore  $\beta$  is not uniquely determined without  $r^2$  exact- or just-identifying restrictions,  $r$  restrictions on each of the  $r$  cointegrating relations.

### 12.9.2 Worked Problems

**Problem 2:** See Chapter 21 Problem 2.

## 12.10 Chapter 23: VARX Modeling

Lidan says will not be on final.

### 12.10.1 Worked Problems

## 12.11 Chapter 24: Impulse Response Analysis

Feel ok on concepts, should review and do an example problem.

### 12.11.1 Worked Problems

**Chapter 24 Problem 1.** Consider the VAR(2) model

$$x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \epsilon_t, \quad \epsilon_t \sim IID(0, \Sigma)$$

in the  $m \times 1$  vector of random variables  $x_t$  and  $\Sigma$  is the covariance matrix of the errors with typical element  $\sigma_{ij}$ .

- (a) Derive the conditions under which this process is stationary, and show that it has the following moving average representation:

$$x_t = \sum_{j=0}^{\infty} A_j \epsilon_{t-j} \tag{12.37}$$

- (b) Derive the coefficient matrices  $A_j$  in terms of  $\Phi_1$  and  $\Phi_2$ .  
(c) Using the above result, write down the orthogonalized (OIR) and generalized impulse (GIR) response functions of one standard error shock (i.e.  $\sqrt{\sigma_{ii}}$  to the error of the  $i$ th equation,  $\epsilon_{it} = s'_i \epsilon_t$ , where  $s_i$  is an  $m \times 1$  selection vector).  
(d) What are the main differences between OIR and GIR functions?

### Chapter 24 Problem 1 Solution.

- (a) For stationarity conditions, see Ch. 21 Problem 2 in Section 12.8.1. To get the MA representation

$$x_t = \sum_{j=0}^{\infty} A_j \epsilon_{t-j} \tag{12.38}$$

of (12.37) note that

$$(I - \Phi_1 L - \Phi_2 L^2)x_t = \epsilon_t$$

$$\implies x_t = (I - \Phi_1 L - \Phi_2 L^2)^{-1}\epsilon_t = \sum_{j=0}^{\infty} A_j \epsilon_{t-j}$$

for some  $\{A_j\}$ .

- (b) Now we seek to evaluate  $(I - \Phi_1 L - \Phi_2 L^2)^{-1}$ . To do this, let  $(I - \Phi_1 L - \Phi_2 L^2)^{-1} = A_0 + A_1 L + A_2 L^2 + A_3 L^3 + \dots$  and note that

$$\begin{aligned} & (I - \Phi_1 L - \Phi_2 L^2)(I - \Phi_1 L - \Phi_2 L)^{-1} = I \\ \iff & (I - \Phi_1 L - \Phi_2 L^2)(A_0 + A_1 L + A_2 L^2 + A_3 L^3 + \dots + A_j L^j + \dots) = I \\ \iff & A_0 + (A_1 - \Phi_1)L + (A_2 - \Phi_1 A_1 - \Phi_2)L^2 + (A_3 - \Phi_1 A_2 - \Phi_2 A_1)L^3 + \dots \\ & + (A_j - \Phi_1 A_{j-1} + \Phi_2 A_{j-2})L^j + \dots = I \end{aligned}$$

In order for this equation to hold, all the lag terms must equal zero and the constant matrix  $A_0$  must equal  $I$ .

$$\implies [A_0 = I]$$

$$A_1 - \Phi_1 = 0 \iff [A_1 = \Phi_1]$$

$$A_2 - \Phi_1 A_1 - \Phi_2 = 0 \iff A_2 - \Phi_1^2 - \Phi_2 = 0 \iff [A_2 = \Phi_1^2 + \Phi_2]$$

$$A_3 - \Phi_1 A_2 - \Phi_2 A_1 = 0 \iff A_3 - \Phi_1^3 - \Phi_1 \Phi_2 - \Phi_2 \Phi_1 = 0 \iff [A_3 = \Phi_1^3 + \Phi_1 \Phi_2 + \Phi_2 \Phi_1]$$

and, in general,

$$[A_j = \Phi_1 A_{j-1} + \Phi_2 A_{j-2}]$$

- (c) • **OIR:** We employ the Cholesky decomposition of  $\Sigma$ :

$$\Sigma = PP' \tag{12.39}$$

where  $P$  is a lower-triangular matrix. Then the MA representation (12.38) can be written as

$$x_t = \sum_{j=0}^{\infty} (A_j P)(P^{-1} u_{t-j}) = \sum_{j=0}^{\infty} B_j \eta_{t-j} \tag{12.40}$$

where  $B_j = A_j P$ ,  $\eta_t = P^{-1} u_t$ , so we have

$$\mathbb{E}(\eta_t \eta'_t) = P^{-1} \mathbb{E}(u_t u'_t)(P^{-1})' = P^{-1} \Sigma (P^{-1})' = P^{-1} P P' (P^{-1})' = I_m$$

so the new errors  $\eta_{1t}, \eta_{2t}, \dots, \eta_{mt}$  are contemporaneously uncorrelated. Then the orthogonalized impact of a unit shock at time  $t$  to the  $i$ th equation on  $y$  at time  $t + n$  is given by

$$B_n e_i, n = 0, 1, \dots \quad (12.41)$$

where  $e_i$  is an  $m \times 1$  selection vector. Written more compactly, the orthogonalized impulse response function of a unit (one standard error) shock to the  $i$ th variable on the  $j$ th variable is given by

$$OI_{ij,n} = e'_j A_n P e_i, \quad i, j = 1, 2, \dots, m \quad (12.42)$$

These orthogonalized impulse responses are not unique and depend on the particular ordering of the variables in the VAR. The orthogonalized responses are invariant to the ordering of the variables only if  $\Sigma$  is diagonal.

- **GIR:** If the VAR model is perturbed by a shock of size  $\delta_i = \sqrt{\sigma_{ii}}$  to its  $i$ th equation at time  $t$ , by the definition of the generalized IR function we have

$$GI_y(n, \delta_i, \Omega_{t-1}^0) = \mathbb{E}(y_t | u_{it} = \delta_i, \Omega_{t-1}^0) - \mathbb{E}(y_t | \Omega_{t-1}^0) \quad (12.43)$$

Once again using the MA( $\infty$ ) representation (12.38) we obtain

$$GI_y(n, \delta_i, \Omega_{t-1}^0) = A_n \mathbb{E}(u_t | u_{it} = \delta_i) \quad (12.44)$$

which is history invariant (i.e. does not depend on  $\Omega_{t-1}^0$ ). The computation of the conditional expectations  $\mathbb{E}(u_t | u_{it} = \delta_i)$  depends on the nature of the multivariate distribution assumed for the disturbances  $u_t$ . In the case where  $u_t \sim IID\mathcal{N}(0, \Sigma)$ , we have

$$\mathbb{E}(u_t | u_{it} = \delta_i) = \begin{bmatrix} \sigma_{1i}/\sigma_{ii} \\ \sigma_{2i}/\sigma_{ii} \\ \vdots \\ \sigma_{mi}/\sigma_{ii} \end{bmatrix} \delta_i \quad (12.45)$$

where as before  $\Sigma = [\sigma_{ij}]$ . Hence for a unit shock  $\delta_i = \sqrt{\sigma_{ii}}$  we have

$$GI_y(n, \delta_i = \sqrt{\sigma_{ii}}, \Omega_{t-1}^0) = \frac{A_n \Sigma e_i}{\sqrt{\sigma_{ii}}}, \quad i, j = 1, 2, \dots, m \quad (12.46)$$

The GIRF of a unit shock to the  $i$ th equation in the VAR( $p$ ) model

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \dots + \Phi_p y_{t-p} + u_t, \quad u_t \sim IID(0, \Sigma) \quad (12.47)$$

on the  $j$ th variable at horizon  $n$  is given by the  $j$ th element of (12.46), expressed more compactly by

$$GI_y(n, \delta_i = \sqrt{\sigma_{ii}}, \Omega_{t-1}^0) = \frac{e'_j A_n \Sigma e_i}{\sqrt{\sigma_{ii}}}, \quad i, j = 1, 2, \dots, m \quad (12.48)$$

- (d) The GIRF circumvents the problem of the dependence of the orthogonalized impulse responses to the ordering of the variables in the VAR. Unlike the OIR responses in (12.42), the GIR responses in (12.48) are invariant to the ordering of the variables in the VAR. The two responses coincide only for the first variable in the VAR or when  $\Sigma$  is diagonal.

**Chapter 24 Problem 4.**

**Chapter 24 Problem 4 Solution.**

### 12.12 Chapter 33: Theory and Practice of GVAR Modeling

Lidan says will not be on final.

## 13 Statistical Learning

These notes are based on my notes from Math 547 at USC taught by Steven Heilman. I also borrowed from some other sources which I mention when I use them.

### 13.1 Math 547

Exercise 3.16: this inequality talks about the number of misclassifications, not the probability of misclassification under any distribution.

#### 13.1.1 Perceptron Algorithm

**Remark.** Note that the run times only depends on the  $\ell_2$  norm of the solution loadings  $w$  and the  $\ell_2$  norm of the longest vector in the data set. That sounds good since it doesn't depend on the size of the data, but in the worse case  $\theta$  can grow exponentially in the dimension of the data.

Also, note that the actual run time is at least linear in the size of the data, since on each iteration the algorithm checks some multiple of  $m$  points.

#### 13.1.2 Mercer's Theorem

How is this an infinite-dimensional version of the Exercise? Let  $M$  be a  $k \times k$  real symmetric matrix. By the Spectral Theorem, there exists an orthogonal  $Q$  and a diagonal  $D$  such that  $M = Q^T D Q$ . For all  $1 \leq p \leq k$ , let  $\lambda_p$  denote the  $p$ th diagonal entry of  $D$ . Let  $\psi_i^p \in \mathbb{R}^k$  denote the  $i$ th row of  $Q$ . Then

$$m_{ij} = \sum_{p=1}^k \lambda_p \psi_i^{(p)} \psi_j^{(p)}, \quad \forall 1 \leq i, j \leq k.$$

Also,  $m(x, y)$  is called a **kernel**.

How to get  $\psi$  from  $m(x, y)$ ? Define

$$\ell_2 := \{\}$$

Note: if we could write the algorithm in terms of  $m(x, y)$ , we don't need to specify this embedding  $\phi$  at all.

### 13.2 Norms

**Proposition 13.2.1.** Let  $C \subseteq \mathbb{R}^n$  be a symmetric (if  $c \in C$  then  $-c \in C$ ) convex set containing 0. Define for all  $x \in \mathbb{R}^n$ ,

$$\|x\|_C := \inf \left\{ t > 0 : \frac{x}{t} \in C \right\}$$

Then  $\|x\|_C$  is a norm.

## 14 Abstract Algebra

These are my notes from reading *Elementary Abstract Algebra* by W. Edwin Clark, available for free download on his website: [http://shell.cas.usf.edu/~wclark/#ELEMENTARY\\_ABSTRACT\\_ALGEBRA](http://shell.cas.usf.edu/~wclark/#ELEMENTARY_ABSTRACT_ALGEBRA)

### 14.1 Chapter 1: Binary Operations

**Definition 1.1** A **binary operation**  $*$  on a set  $S$  is a function from  $S \times S$  to  $S$ . If  $(a, b) \in S \times S$  then we write  $a * b$  to indicate the image of the element  $(a, b)$  under the function  $*$ .

The following lemma explains in more detail exactly what this definition means.

**Lemma 1.1** A binary operation  $*$  on a set  $S$  is a rule for combining two elements of  $S$  to produce a third element of  $S$ . This rule must satisfy the following conditions:

- (a)  $a \in S$  and  $b \in S \implies a * b \in S$ . [ $S$  is closed under  $*$ .]
- (b) For all  $a, b, c, d$  in  $S$   
 $a = c$  and  $b = d \implies a * b = c * d$ . [Substitution is permissible.]
- (c) For all  $a, b, c, d$  in  $S$   
 $a = b \implies a * c = b * c$ .
- (d) For all  $a, b, c, d$  in  $S$   
 $c = d \implies a * c = a * d$ .

**Definition:** A **function**  $f$  from the set  $A$  to the set  $B$  is a rule which assigns to each element  $a \in A$  an element  $f(a) \in B$  in such a way that the following condition holds for all  $x, y \in A$ :

$$x = y \implies f(x) = f(y)$$

To indicate that  $f$  is a function from  $A$  to  $B$  we write  $f : A \rightarrow B$ . The set  $A$  is called the **domain** of  $f$  and the set  $B$  is called the **codomain** of  $f$ .

A function  $f : A \rightarrow B$  is said to be **one-to-one** or **injective** if the following condition holds for all  $x, y \in A$ :

$$f(x) = f(y) \implies x = y$$

A function  $f : A \rightarrow B$  is said to be **onto** or **surjective** if the following condition holds:

$$\forall b \in B \exists a \in A \mid f(a) = b$$

A function  $f : A \rightarrow B$  is said to be **bijective** if it is both one-to-one and onto. Then  $f$  is sometimes said to be a **bijection** or a **one-to-one correspondence** between  $A$  and  $B$ .

15. Let  $S$ ,  $T$ , and  $U$  be nonempty sets, and let  $f : S \rightarrow T$  and  $g : T \rightarrow U$  be functions such that the function  $g \circ f : S \rightarrow U$  is one-to-one (injective). Which of the following must be true?
- $f$  is one-to-one.
  - $f$  is onto.
  - $g$  is one-to-one.
  - $g$  is onto.
  - $g \circ f$  is onto.

**Solution 15.** (A) For a composition of functions, if the first function isn't one-to-one, there's no way the composite is. It's worth mentioning here that the opposite is true for onto: the second function had better be onto.

Let  $S$  be a set. The **power set**  $\mathcal{P}(S)$  of  $S$  is the set of all subsets of  $S$  (including  $S$  itself).

**Definition 1.2** Assume that  $*$  is a binary operation on the set  $S$ .

1. We say that  $*$  is **associative** if

$$x * (y * z) = (x * y) * z \quad \text{for all } x, y, z \in S.$$

2. We say that an element  $e$  in  $S$  is an **identity** with respect to  $*$  if

$$x * e = x \text{ and } e * x = x \quad \text{for all } x \text{ in } S.$$

3. Let  $e \in S$  be an identity with respect to  $*$ . Given  $x \in S$  we say that an element  $y \in S$  is an **inverse** of  $x$  if both

$$x * y = e \text{ and } y * x = e.$$

4. We say that  $*$  is **commutative** if

$$x * y = y * x \quad \text{for all } x, y \in S.$$

5. We say that an element  $a$  of  $S$  is **idempotent** with respect to  $*$  if

$$a * a = a.$$

6. We say that an element  $z$  of  $S$  is a **zero** with respect to  $*$  if

$$z * x = z \text{ and } x * z = z \quad \text{for all } x \in S.$$

For each integer  $n \geq 2$  define the set

$$\mathbb{Z}_n = \{0, 1, 2, \dots, n - 1\}$$

For all  $a, b \in \mathbb{Z}_n$  let

$$a + b = \text{remainder when the ordinary sum of } a \text{ and } b \text{ is divided by } n$$

and

$$a \cdot b = \text{remainder when the ordinary product of } a \text{ and } b \text{ is divided by } n.$$

These binary operations are referred to as **addition modulo  $n$**  and **multiplication modulo  $n$** . The integer  $n$  in  $\mathbb{Z}_n$  is called the **modulus**. The plural of modulus is **moduli**.

Let  $K$  denote any one of the following:  $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{Z}_n$ .

$$M_n(K)$$

is the set of all  $n \times n$  matrices containing elements of  $K$ .

$$GL(n, K)$$

is the set of all matrices in  $M_n(K)$  with non-zero determinant.  $(GL(n, k), \cdot)$  is called the **general linear group of degree  $n$  over  $K$** . It is non-abelian.

$$SL(n, K) = \{A \in GL(n, K) \mid \det(A) = 1\}$$

$SL(n, K)$  is called the **Special Linear Group of degree  $n$  over  $K$** .

## 14.2 Chapter 2: Groups

**Definition** A **group** is an ordered pair  $(G, *)$  where  $G$  is a set and  $*$  is a binary operation on  $G$  satisfying the following properties:

1. The binary operation is associative on  $G$ :  $\forall x, y, z \in G$ ,

$$x * (y * z) = (x * y) * z$$

2. The binary operation contains a (unique) identity in  $G$ :  $\exists e \in G \mid \forall x \in G$

$$e * x = x, x * e = x$$

3. Every element in  $G$  has a (unique) inverse on  $*$  in  $G$ :  $\forall x \in G \exists y \in G \mid$

$$x * y = e, y * x = e$$

A group  $(G, *)$  is said to be **abelian** if  $\forall x, y \in G$ ,  $x * y = y * x$ . A group is said to be **non-abelian** if it is not abelian.

**Theorem 2.2** Let  $(G, *)$  be a group with identity  $e$ . Then the following hold for all elements  $a, b, c, d$  in  $G$ :

1. If  $a * c = a * b$ , then  $c = b$ . [Left cancellation law for groups.]
2. If  $c * a = b * a$ , then  $c = b$ . [Right cancellation law for groups.]
3. Given  $a$  and  $b$  in  $G$  there is a unique element  $x$  in  $G$  such that  $a * x = b$ .
4. Given  $a$  and  $b$  in  $G$  there is a unique element  $x$  in  $G$  such that  $x * a = b$ .
5. If  $a * b = e$  then  $a = b^{-1}$  and  $b = a^{-1}$ . [Characterization of the inverse of an element.]
6. If  $a * b = a$  for just one  $a$ , then  $b = e$ .
7. If  $b * a = a$  for just one  $a$ , then  $b = e$ .
8. If  $a * a = a$ , then  $a = e$ . [The only idempotent in a group is the identity.]
9.  $(a^{-1})^{-1} = a$ .
10.  $(a * b)^{-1} = b^{-1} * a^{-1}$ .

### 14.3 Chapter 3: The Symmetric Groups

If  $n$  is a positive integer,

$$[n] = \{1, 2, \dots, n\}$$

A **permutation** of  $[n]$  is a one-to-one, onto function from  $[n]$  to  $[n]$ , and

$$S_n$$

is the set of all permutations of  $[n]$ .

The identity of  $S_n$  is the so-called **identity function**

$$\iota : [n] \rightarrow [n]$$

which is defined by the rule

$$\iota(x) = x, \quad \forall x \in [n]$$

**The inverse of an element  $\sigma \in S_n$ :** Suppose  $\sigma \in S_n$ . Since  $\sigma$  is by definition one-to-one and onto, the rule

$$\sigma^{-1}(y) = x \iff \sigma(x) = y$$

defines a function  $\sigma^{-1} : [n] \rightarrow [n]$ . This function  $\sigma^{-1}$  is also one-to-one and onto and satisfies

$$\sigma\sigma^{-1} = \iota \text{ and } \sigma^{-1}\sigma = \iota$$

so it is the inverse of  $\sigma$  in the group sense also.

Since the binary operation of composition on  $S_n$  is associative  $[(\gamma\beta)\alpha = \gamma(\beta\alpha)]$ ,  $S_n$  under the binary operation of composition is a group (it is associative, it has an inverse, and it has an identity).

**Definition 3.2** Let  $i_1, i_2, \dots, i_k$  be a list of  $k$  distinct elements from  $[n]$ . Define a permutation  $\sigma$  in  $S_n$  as follows:

$$\begin{array}{rcl} \sigma(i_1) & = & i_2 \\ \sigma(i_2) & = & i_3 \\ \sigma(i_3) & = & i_4 \\ & \vdots & \vdots \\ \sigma(i_{k-1}) & = & i_k \\ \sigma(i_k) & = & i_1 \end{array}$$

and if  $x \notin \{i_1, i_2, \dots, i_k\}$  then

$$\sigma(x) = x$$

Such a permutation is called a **cycle** or a  **$k$ -cycle** and is denoted by

$$(i_1 \ i_2 \ \cdots \ i_k).$$

If  $k = 1$  then the cycle  $\sigma = (i_1)$  is just the identity function, i.e.,  $\sigma = \iota$ .

Two cycles  $(i_1 \ i_2 \ \dots \ i_k)$  and  $(j_1 \ j_2 \ \dots \ j_l)$  are said to be **disjoint** if the sets  $\{i_1, i_2, \dots, i_k\}$  and  $\{j_1, j_2, \dots, j_l\}$  are disjoint.

So for example, the cycles  $(1 \ 2 \ 3)$  and  $(4 \ 5 \ 8)$  are disjoint, but the cycles  $(1 \ 2 \ 3)$  and  $(4 \ 2 \ 8)$  are not disjoint.

If  $\sigma$  and  $\tau$  are disjoint cycles, then  $\sigma\tau = \tau\sigma$ .

**Theorem 3.4** Every element  $\sigma \in S_n$ ,  $n \geq 2$ , can be written as a product

$$\sigma = \sigma_1 \sigma_2 \cdots \sigma_m \tag{3.1}$$

where  $\sigma_1, \sigma_2, \dots, \sigma_m$  are pairwise disjoint cycles, that is, for  $i \neq j$ ,  $\sigma_i$  and  $\sigma_j$  are disjoint. If all 1-cycles of  $\sigma$  are included, the factors are unique except for the order. ■

The factorization (3.1) is called the **disjoint cycle decomposition of  $\sigma$** .

An element of  $S_n$  is called a **transposition** if and only if it is a 2-cycle.

Every element of  $S_n$  can be written as a product of transpositions. The factors of such a product are not unique. However, if  $\sigma \in S_n$  can be written as a product of  $k$  transpositions and if the same  $\sigma$  can also be written as a product of  $l$  transpositions, then  $k$  and  $l$  have the same parity.

A permutation is **even** if it is a product of an even number of transpositions and **odd** if it is a product of an odd number of transpositions. We define the function  $\text{sign} : S_n \rightarrow \{1, -1\}$  by

$$\text{sign}(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd} \end{cases}$$

If  $n = 1$  then there are no transpositions. In this case, to be complete we define the identity permutation  $\iota$  to be even.

If  $\sigma$  is a  $k$ -cycle, then  $\text{sign}(\sigma) = 1$  if  $k$  is odd and  $\text{sign}(\sigma) = -1$  if  $k$  is even.

**Remark.** Let  $A = [a_{ij}]$  be an  $n \times n$  matrix. The determinant of  $A$  may be defined by the sum

$$\det(A) = \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}.$$

For example, if  $n = 2$  we have only two permutations  $\iota$  and  $(1 \ 2)$ . Since  $\text{sign}(\iota) = 1$  and  $\text{sign}((1 \ 2)) = -1$  we obtain

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}.$$

**Definition:** If  $(G, *)$  is a group, the number of elements in  $G$  is called the **order** of  $G$ . We use  $|G|$  to denote the order of  $G$ . Note that  $|G|$  may be finite or infinite.

Let

$$A_n$$

be the set of all even permutations in the group  $S_n$ .  $A_n$  is called the **alternating group of degree  $n$** .

## 14.4 Chapter 4: Subgroups

**Definition:** Let  $G$  be a group. A **subgroup** of  $G$  is a subset  $H$  of  $G$  which satisfies the following three conditions:

1.  $e \in H$
2.  $a, b \in H \implies ab \in H$
3.  $a \in H \implies a^{-1} \in H$

If  $H$  is a subgroup of  $G$ , we write  $H \leq G$ . The subgroups  $\{e\}$  and  $G$  are said to be **trivial** subgroups of  $G$ .

Every finite subgroup may be thought of as a subgroup of one of the groups  $S_n$ .

Let  $A_n$  be the set of all even permutations in the group  $S_n$ .  $A_n$  is then a subgroup of  $S_n$ .  $A_n$  is called the **alternating group of degree  $n$** .

Let  $a$  be an element of the group  $G$ . If  $\exists n \in \mathbb{N} \mid a^n = e$  we say that  $a$  has **finite order** and we define

$$\text{o}(a) = \min\{n \in \mathbb{N} \mid a^n = e\}$$

If  $a^n \neq e \forall n \in \mathbb{N}$  we say that  $a$  has **infinite order** and we define

$$\text{o}(a) = \infty$$

In either case we call  $\text{o}(a)$  the **order** of  $a$ . Note carefully the difference between the order of a group and the order of an element of a group. Note also that  $a = e \iff \text{o}(a) = 1$ . So every element of a group other than  $e$  has order  $n \geq 2$  or  $\infty$ .

Let  $a$  be an element of group  $G$ . Define

$$\langle a \rangle = \{a^i : i \in \mathbb{Z}\}$$

We call  $\langle a \rangle$  the **subgroup of  $G$  generated by  $a$** . Note that  $e = a^0$  and  $a^{-1}$  are in  $\langle a \rangle$ .

**Theorem.** For each  $a \in G$ ,  $\langle a \rangle$  is a subgroup of  $G$ .  $\langle a \rangle$  contains  $a$  and is the smallest subgroup of  $G$  containing  $a$ .

**Proof of second statement.** If  $H$  is any subgroup of  $G$  containing  $a$ ,  $\langle a \rangle \subseteq H$  since  $H$  is closed under taking products and inverses. That is, every subgroup of  $G$  containing  $a$  also contains  $\langle a \rangle$ . This implies that  $\langle a \rangle$  is the smallest subgroup of  $G$  containing  $a$ .

**Theorem.** Let  $G$  be a group and let  $a \in G$ . If  $\text{o}(a) = 1$ , then  $\langle a \rangle = \{e\}$ . If  $\text{o}(a) = n$  where  $n \geq 2$ , then

$$\langle a \rangle = \{e, a, a^2, \dots, a^{n-1}\}$$

and the elements  $e, a, a^2, \dots, a^{n-1}$  are distinct; that is,

$$\text{o}(a) = |\langle a \rangle|$$

**Proof** Assume that  $\text{o}(a) = n$ . The case  $n = 1$  is left to the reader. Suppose  $n \geq 2$ . We must prove two things.

1. If  $i \in \mathbb{Z}$  then  $a^i \in \{e, a, a^2, \dots, a^{n-1}\}$ .
2. The elements  $e, a, a^2, \dots, a^{n-1}$  are distinct.

To establish 1 we note that if  $i$  is any integer we can write it in the form  $i = nq + r$  where  $r \in \{0, 1, \dots, n - 1\}$ . Here  $q$  is the quotient and  $r$  is the remainder when  $i$  is divided by  $n$ . Now using Theorem 2.4 we have

$$a^i = a^{nq+r} = a^{nq}a^r = (a^n)^qa^r = e^qa^r = ea^r = a^r.$$

This proves 1. To prove 2, assume that  $a^i = a^j$  where  $0 \leq i < j \leq n - 1$ . It follows that

$$a^{j-i} = a^{j+(-i)} = a^j a^{-i} = a^i a^{-i} = a^0 = e.$$

But  $j - i$  is a positive integer less than  $n$ , so  $a^{j-i} = e$  contradicts the fact that  $\text{o}(a) = n$ . So the assumption that  $a^i = a^j$  where  $0 \leq i < j \leq n - 1$  is false. This implies that 2 holds. It follows that  $\langle a \rangle$  contains exactly  $n$  elements, that is,  $\text{o}(a) = |\langle a \rangle|$ .

**Theorem.** If  $G$  is a finite group, then every element of  $G$  has finite order.

**49.** What is the largest order of an element in the group of permutations of 5 objects?

- (A) 5      (B) 6      (C) 12      (D) 15      (E) 120

**Solution 49.** (B) The greatest order is given by the product of a 2-cycle and a 3-cycle acting on disjoint elements. That gives order 6.

## 14.5 Chapter 5: The Group of Units of $\mathbb{Z}_n$

Let  $n \geq 2$ . An element  $a \in \mathbb{Z}_n$  is said to be a **unit** if  $\exists b \in \mathbb{Z}_n \mid ab = 1$  (where the product is multiplication modulo  $n$ ).

The set of all units in  $\mathbb{Z}_n$  is denoted by

$$U_n$$

and is a group under multiplication modulo  $n$  called the **group of units of  $\mathbb{Z}_n$** .

**Theorem.** For  $n \geq 2$ ,  $U_n = \{a \in \mathbb{Z}_n : \gcd(a, n) = 1\}$

**Theorem.**  $p$  is a prime  $\implies \exists a \in U_p \mid U_p = \langle a \rangle$

**Theorem.** If  $n \geq 2$  then  $U_n$  contains an element  $a$  satisfying  $U_n = \langle a \rangle$  if and only if  $a$  has one of the following forms: 2, 4,  $p^k$ , or  $2p^k$  where  $p$  is an odd prime and  $k \in \mathbb{N}$ .

## 14.6 Chapter 6: Direct Products of Groups

If  $G_1, G_2, \dots, G_n$  is a list of  $n$  groups we make the Cartesian product  $G_1 \times G_2 \times \dots \times G_n$  into a group by defining the binary operation

$$(a_1, a_2, \dots, a_n) \cdot (b_1, b_2, \dots, b_n) = (a_1 \cdot b_1, a_2 \cdot b_2, \dots, a_n \cdot b_n)$$

Here for each  $i \in \{1, 2, \dots, n\}$  the product  $a_i \cdot b_i$  is the product of  $a_i$  and  $b_i$  in the group  $G_i$ . We call this group the **direct product** of the groups  $G_1, G_2, \dots, G_n$ .

The direct product contains an identity and an inverse, and is associative (since it is composed of groups which must themselves be associative), so it is a group per below:

**Theorem.** If  $G_1, G_2, \dots, G_n$  is a list of  $n$  groups, the direct product  $G = G_1 \times G_2 \times \dots \times G_n$  as defined above is a group. Moreover, if for each  $i$ ,  $e_i$  is the identity of  $G_i$ , then  $e_1, e_2, \dots, e_n$  is the identity of  $G$ , and if

$$\mathbf{a} = (a_1, a_2, \dots, a_n) \in G$$

then the inverse of  $\mathbf{a}$  is given by

$$\mathbf{a}^{-1} = (a_1^{-1}, a_2^{-1}, \dots, a_n^{-1})$$

where  $a_i^{-1}$  is the inverse of  $a_i$  in the group  $G_i$ .

## 14.7 Chapter 7: Isomorphism of Groups

Let  $G = \{g_1, g_2, \dots, g_n\}$ . Let  $\text{o}(g_i) = k_i$  for  $i = 1, 2, \dots, n$ . We say that the sequence  $(k_1, k_2, \dots, k_m)$  is the **order sequence** of the group  $G$ . To make the sequence unique we assume the elements are ordered so that  $k_1 \leq k_2 \leq \dots \leq k_n$ .

Let  $(G, *)$  and  $(H, \bullet)$  be groups. A function  $f : G \rightarrow H$  is said to be a **homomorphism** from  $G$  to  $H$  if

$$f(a * b) = f(a) \bullet f(b)$$

for all  $a, b \in G$ . If in addition  $f$  is one-to-one and onto,  $f$  is said to be an **isomorphism** from  $G$  to  $H$ .

We say that  $G$  and  $H$  are **isomorphic** if and only if there is an isomorphism from  $G$  to  $H$ . We write  $G \cong H$  to indicate that  $G$  is isomorphic to  $H$ .

**Isomorphism is an equivalence relation:** If  $G, H$ , and  $K$  are groups then

1.  $G \cong G$
2. If  $G \cong H$  then  $H \cong G$ , and
3. If  $G \cong H$  and  $H \cong K$ , then  $G \cong K$ .

**Theorem.** Let  $(G, *)$  and  $(H, \bullet)$  be groups and let  $f : G \rightarrow H$  be a homomorphism. Let  $e_G$  denote the identity of  $G$ , and let  $e_H$  denote the identity of  $H$ . Then

1.  $f(e_G) = e_H$

*Proof:* Let  $x_G \in G$  and let  $f(x_G) = x_H \in H$ . Then  

$$x_H = f(x_G) = f(e_G * x_G) = f(e_G) \bullet f(x_G) = f(e_G) \bullet x_H = e_H \bullet x_H.$$

2.  $f(a^{-1}) = f(a)^{-1}$

*Proof:*  $f(a)^{-1} \bullet f(a) = e_H = f(e_G) = f(a^{-1} * a) = f(a^{-1}) \bullet f(a)$

3.  $f(a^n) = f(a)^n \forall n \in \mathbb{Z}$

*Proof by induction.*

**Theorem.** Let  $(G, *)$  and  $(H, \bullet)$  be groups and let  $f : G \rightarrow H$  be an isomorphism. Then  $\text{o}(a) = \text{o}(f(a)) \forall a \in G$ . It follows that  $G$  and  $H$  have the same number of elements of each possible order.

**Theorem.** If  $G$  and  $H$  are isomorphic groups, and  $G$  is abelian, then so is  $H$ .

*Proof:* Let  $a_G, b_G \in G$  and let  $f(a_G) = a_H \in H, f(b_G) = b_H \in H$ .  

$$a_H \bullet b_H = f(a_G) \bullet f(b_G) = f(a_G * b_G) = f(b_G * a_G) = f(b_G) \bullet f(a_G) = b_H \bullet a_H.$$

A group  $G$  is **cyclic** if there is an element  $a \in G$  |  $\langle a \rangle = G$ . If  $\langle a \rangle = G$  then we say that  $a$  is a **generator** for  $G$ .

**Theorem.** If  $G$  and  $H$  are isomorphic groups and  $G$  is cyclic then  $H$  is cyclic.

**Theorem.** Let  $a$  be an element of group  $G$ .

1.  $\text{o}(a) = \infty \implies \langle a \rangle \cong \mathbb{Z}$ .
2.  $\text{o}(a) = n \in \mathbb{N} \implies \langle a \rangle \cong \mathbb{Z}_n$

**Cayley's Theorem.** If  $G$  is a finite group of order  $n$ , then there is a subgroup  $H$  of  $S_n$  such that  $G \cong H$ .

**66.** Let  $\mathbb{Z}_{17}$  be the ring of integers modulo 17, and let  $\mathbb{Z}_{17}^\times$  be the group of units of  $\mathbb{Z}_{17}$  under multiplication.

Which of the following are generators of  $\mathbb{Z}_{17}^\times$ ?

- I. 5
- II. 8
- III. 16

- (A) None      (B) I only      (C) II only      (D) III only      (E) I, II, and III

**Solution 66.** (B) We need to pick elements of order 16 in  $\mathbb{Z}/17^\times$ . It is easy to rule out 16  $\equiv -1$ , since  $-1$  has order 2. We see that  $5^2 = 25 \equiv 8$ , so there's no way that 8 can be a generator. We just need to verify that the order of 5 is more than 8, so we can check  $5^8$ :

$$5^4 = 8^2 = 64 \equiv -4, \quad 5^8 = (-4)^2 = 16 \neq 1.$$

That makes 5 a generator.

## 14.8 Chapter 8: Cosets and Lagrange's Theorem

Let  $G$  be a group and let  $H$  be subgroup of  $G$ . For each element  $a$  of  $G$  we define

$$aH = \{ah \mid h \in H\}$$

We call  $aH$  the **coset of  $H$  in  $G$  generated by  $a$** .

Let  $a, b \in G$ . Then

1.  $a \in aH$  (since  $H$  must contain an identity; specifically, the identity of  $G$ )
2.  $|aH| = |H|$  (since  $ah$  is unique)
3.  $aH \cap bH \neq \emptyset \implies aH = bH$

**Lagrange's Theorem.** If  $G$  is a finite group and  $H \leq G$  then  $|H|$  divides  $|G|$ .

Any group of prime order is cyclic; therefore, there is only one such group up to isomorphism.

**Exercise 3.** Use Lagrange's theorem to prove that any group of prime order is cyclic.

*Proof.* Let  $G$  be a group whose order is a prime  $p$ . Since  $p > 1$ , there is an element  $a \in G$  such that  $a \neq e$ . The group  $\langle a \rangle$  generated by  $a$  is a subgroup of  $G$ . By Lagrange's theorem, the order of  $\langle a \rangle$  divides  $|G|$ . But the only divisors of  $|G| = p$  are 1 and  $p$ . Since  $a \neq e$  we have  $|\langle a \rangle| > 1$ , so  $|\langle a \rangle| = p$ . Hence  $\langle a \rangle = G$  and  $G$  is cyclic.  $\square$

We say that there are  $k$  **isomorphism classes of groups of order  $n$**  if there are  $k$  groups  $G_1, G_2, \dots, G_k$  such that

1. if  $i \neq j$  then  $G_i$  and  $G_j$  are not isomorphic, and
2. Every group of order  $n$  is isomorphic to  $G_i$  for some  $i \in \{1, 2, \dots, k\}$ .

This is sometimes expressed by saying that "there are  $k$  groups of order  $n$  up to isomorphism" or that "there are  $k$  non-isomorphic groups of order  $n$ ."

12. For which integers  $n$  such that  $3 \leq n \leq 11$  is there only one group of order  $n$  (up to isomorphism) ?
- (A) For no such integer  $n$
  - (B) For 3, 5, 7, and 11 only
  - (C) For 3, 5, 7, 9, and 11 only
  - (D) For 4, 6, 8, and 10 only
  - (E) For all such integers  $n$

**Solution 12.** (B) Any group of prime order is necessarily cyclic, and hence there is only one up to isomorphism. This limits our choices to (B), (C), and (E). But there are two groups of order 9 (at least):  $\mathbb{Z}/3 \times \mathbb{Z}/3$  and  $\mathbb{Z}/9$ . This makes (B) our only option.

In more advanced courses in algebra, it is shown that the number of isomorphism classes of groups of order  $n$  for  $n \leq 17$  is given by the following table:

<i>Order :</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
<i>Number :</i>	1	1	1	2	1	2	1	5	2	2	1	5	1	2	1	14	1

This table means, for example, that one may find 14 groups of order 16 such that every group of order 16 is isomorphic to one and only one of these 14 groups.

There is only one isomorphism class of groups of order  $n$  if  $n$  is prime. But there are some non-primes that have this property; for example, 15.

**The Fundamental Theorem of Finite Abelian Groups.** If  $G$  is a finite abelian group of order at least 2, then

$$G \cong \mathbb{Z}_{p_1^{n_1}} \times \mathbb{Z}_{p_2^{n_2}} \times \cdots \times \mathbb{Z}_{p_s^{n_s}}$$

where for each  $i$ ,  $p_i$  is a prime and  $n_i$  is a positive integer. Moreover, the prime powers  $p_i^{n_i}$  are unique except for the order of the factors.

If the group  $G$  in the above theorem has order  $n$  then

$$n = p_1^{n_1} p_2^{n_2} \cdots p_s^{n_s}$$

So the  $p_i$  may be obtained from the prime factorization of the order of the group  $G$ . These primes are not necessarily distinct, so we cannot say what the  $n_i$  are. However, we can find all possible choices for the  $n_i$ . For example, if  $G$  is an abelian group of order  $72 = 3^2 \cdot 2^3$  then  $G$  is isomorphic to one and only one of the following groups. Note that each corresponds to a way of factoring 72 as a product of prime powers.

$\mathbb{Z}_9 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$	$72 = 9 \cdot 2 \cdot 2 \cdot 2$
$\mathbb{Z}_9 \times \mathbb{Z}_4 \times \mathbb{Z}_2$	$72 = 9 \cdot 4 \cdot 2$
$\mathbb{Z}_9 \times \mathbb{Z}_8$	$72 = 9 \cdot 8$
$\mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$	$72 = 3 \cdot 3 \cdot 2 \cdot 2 \cdot 2$
$\mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_4 \times \mathbb{Z}_2$	$72 = 3 \cdot 3 \cdot 4 \cdot 2$
$\mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_8$	$72 = 3 \cdot 3 \cdot 8$

Thus there are exactly 6 non-isomorphic abelian groups of order 72.

**Corollary.** For  $n \geq 2$ , the number of isomorphism classes of abelian groups of order  $n$  is equal to the number of ways to factor  $n$  as a product of prime powers (where the order of the factors does not count).

## 14.9 Chapter 9: Introduction to Ring Theory

**Definition 9.1** A **ring** is an ordered triple  $(R, +, \cdot)$  where  $R$  is a set and  $+$  and  $\cdot$  are binary operations on  $R$  satisfying the following properties:

**A1**  $a + (b + c) = (a + b) + c$  for all  $a, b, c$  in  $R$ .

**A2**  $a + b = b + a$  for all  $a, b$  in  $R$ .

**A3** There is an element  $0 \in R$  satisfying  $a + 0 = a$  for all  $a$  in  $R$ .

**A4** For every  $a \in R$  there is an element  $b \in R$  such that  $a + b = 0$ .

**M1**  $a \cdot (b \cdot c) = (a \cdot b) \cdot c$  for all  $a, b, c$  in  $R$ .

**D1**  $a \cdot (b + c) = a \cdot b + a \cdot c$  for all  $a, b, c$  in  $R$ .

**D2**  $(b + c) \cdot a = b \cdot a + c \cdot a$  for all  $a, b, c$  in  $R$ .

**Terminology** If  $(R, +, \cdot)$  is a ring, the binary operation  $+$  is called *addition* and the binary operation  $\cdot$  is called *multiplication*. In the future we will usually write  $ab$  instead of  $a \cdot b$ . The element  $0$  mentioned in A3 is called the **zero** of the ring. Note that we have not assumed that  $0$  behaves like a *zero*, that is, we have not assumed that  $0 \cdot a = a \cdot 0 = 0$  for all  $a \in R$ . What A3 says is that  $0$  is an identity with respect to addition. Note that *negative* (as the opposite of *positive*) has no meaning for most rings. We do not assume that multiplication is commutative and we have not assumed that there is an identity for multiplication, much less that elements have inverses with respect to multiplication.

23. Let  $(\mathbb{Z}_{10}, +, \cdot)$  be the ring of integers modulo 10, and let  $S$  be the subset of  $\mathbb{Z}_{10}$  represented by  $\{0, 2, 4, 6, 8\}$ . Which of the following statements is FALSE?

- (A)  $(S, +, \cdot)$  is closed under addition modulo 10.
- (B)  $(S, +, \cdot)$  is closed under multiplication modulo 10.
- (C)  $(S, +, \cdot)$  has an identity under addition modulo 10.
- (D)  $(S, +, \cdot)$  has no identity under multiplication modulo 10.
- (E)  $(S, +, \cdot)$  is commutative under addition modulo 10.

**Solution 23.** (D) Examining the choices, we see  $S \subset \mathbb{Z}/10$  is a subgroup of an abelian group. Therefore it still have an additive identity and the operation is commutative. It is also closed under addition and multiplication. While  $S$  does not contain the multiplicative identity of  $\mathbb{Z}/10$ , it does have a multiplicative identity.  $6 \in S$  is such an identity, as

$$6x = (5 + 1)x = 5x + x.$$

Since  $x \in S$  are all even,  $5x = 0$ , so  $6x = x$ .

50. Let  $R$  be a ring and let  $U$  and  $V$  be (two-sided) ideals of  $R$ . Which of the following must also be ideals of  $R$ ?

- I.  $U + V = \{u + v : u \in U \text{ and } v \in V\}$
  - II.  $U \cdot V = \{uv : u \in U \text{ and } v \in V\}$
  - III.  $U \cap V$
- (A) II only      (B) III only      (C) I and II only      (D) I and III only      (E) I, II, and III

**Solution 50.** (D) The sum of the ideals is still an ideal: it is clearly closed under addition (using commutativity of addition), and still under left and right multiplication due to the distributive property. The intersection of ideals is still an ideal, which is not too hard to work out. The product of ideals, however, need not be closed under addition. Consider, for example,  $R = \mathbb{Z}[X]$ ,  $U = (2, X)$ , and  $V = (3, X)$  (the ideals generated by two elements). Then we know that  $-2X \in U \cdot V$  and  $3X \in U \cdot V$ , and hence we should expect  $3X - 2X = X \in U \cdot V$ . However, there is no way to get  $X$  as the product of an element of  $U$  and an element of  $V$ .

18. Let  $V$  be the real vector space of all real  $2 \times 3$  matrices, and let  $W$  be the real vector space of all real  $4 \times 1$  column vectors. If  $T$  is a linear transformation from  $V$  onto  $W$ , what is the dimension of the subspace  $\{\mathbf{v} \in V : T(\mathbf{v}) = \mathbf{0}\}$ ?

- (A) 2      (B) 3      (C) 4      (D) 5      (E) 6

**Solution 18.** (A) We see that  $\dim V = 6$  and  $\dim W = 4$ . Since  $\dim \text{im } T = \dim W = 4$ , we must have  $\dim \ker T = 6 - 4 = 2$ .

## 15 Miscellaneous

### 15.1 Set Theory

**Proposition 15.1.1 (Math 541A Homework Problem).** Let  $A, B, \Omega$  be sets. Let  $u : \Omega \rightarrow A$  and let  $t : \Omega \rightarrow B$ . Assume that, for every  $x, y \in \Omega$ , if  $u(x) = u(y)$ , then  $t(x) = t(y)$ . Show that there exists a function  $s : A \rightarrow B$  such that

$$t = s(u).$$

*Proof.* Let  $B' \subseteq B$  be the image of  $\Omega$  under  $t$  (so that  $t$  is surjective onto  $B'$ ). Fix  $x \in \Omega$  and let  $y \in \Omega$  range over  $\{\Omega \setminus x\}$ . Let  $Y \subseteq \Omega$  be the set of values such that  $t(y) = t(x)$  for all  $y \in Y$ . Then let  $s$  map  $u(y) \in A$  to  $t(y) = t(x) \in B$  for every  $y \in Y$ .

(Note that if  $x$  is not the only value in  $\Omega$  that  $t$  maps to  $t(x)$ , then  $Y$  contains elements other than  $x$ ; otherwise,  $Y = \{x\}$ . In either case this mapping is fine. Note that  $u(y_1)$  does not necessarily equal  $u(y_2)$  for every  $y_1 \neq y_2 \in Y$ , but again this does not pose any difficulties for the mapping.)

Since this argument is true for every  $x \in \Omega$ , we can argue by contradiction that  $s$  is surjective onto  $B'$ . Let  $A' \subseteq A$  be the image of  $\Omega$  under  $u$  (so that  $u$  is surjective onto  $A'$ ). Suppose there is some  $b \in B'$  such that there is no unused  $a \in A'$  to correspond to it. That is, there are some  $y, z \in \Omega$  such that  $t(z) \neq t(y)$  but  $u(z) = u(y)$ . In that case the mapping  $s$  would map  $u(z)$  and  $u(y)$  both to the same value in  $B'$ , so one of the values  $t(z)$  or  $t(y)$  would necessarily be missed. But by assumption there is no  $z \in \Omega$  such that  $t(z) \neq t(y)$  but  $u(z) = u(y)$ . (Note the contrapositive of the assumption: “for every  $x, y \in \Omega$ , if  $t(x) \neq t(y)$ , then  $u(x) \neq u(y)$ .”)

□

**Remark.** By showing this mapping exists, we have shown that the cardinality of  $B'$  is less than or equal to the cardinality of  $A'$ .

### 15.2 Other

6. Which of the following circles has the greatest number of points of intersection with the parabola  $x^2 = y + 4$ ?
- (A)  $x^2 + y^2 = 1$
  - (B)  $x^2 + y^2 = 2$
  - (C)  $x^2 + y^2 = 9$
  - (D)  $x^2 + y^2 = 16$
  - (E)  $x^2 + y^2 = 25$

**Solution 6.** (C) We can try to do this algebraically, but non-algebraically is simpler. Graphing  $y = x^2 - 4$  shows that the graph crosses the  $x$ -axis at  $\pm 2$ . Therefore a circle of radius 1 or  $\sqrt{2}$  will not intersect the parabola at all. A circle of radius 3 will intersect four times – twice above and twice below the  $x$ -axis. A circle of radius 4 will only intersect at one point below the  $x$ -axis (and twice above), and a circle of radius 5 will only intersect at the two points above.

19. If  $z$  is a complex variable and  $\bar{z}$  denotes the complex conjugate of  $z$ , what is  $\lim_{z \rightarrow 0} \frac{(\bar{z})^2}{z^2}$ ?

- (A) 0      (B) 1      (C)  $i$       (D)  $\infty$       (E) The limit does not exist.

**Solution 19.** (E) Let us represent  $z = a + bi$ . Then our limit becomes

$$\lim_{(a,b) \rightarrow 0} \frac{(a - bi)^2}{(a + bi)^2} = \lim_{(a,b) \rightarrow 0} \frac{a^2 - b^2 - 2abi}{a^2 - b^2 + 2abi}.$$

If we let  $a = 0$  (for instance), it is easy to see that the limit is equal to 1. However, if we let  $a = b$ , then our limit becomes

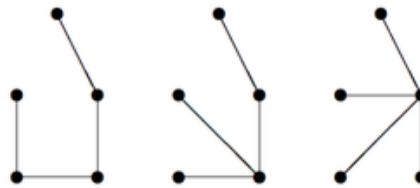
$$\lim_{a \rightarrow 0} \frac{-2a^2i}{2a^2i} = -1.$$

Therefore the limit does not exist.

29. A tree is a connected graph with no cycles. How many nonisomorphic trees with 5 vertices exist?

- (A) 1      (B) 2      (C) 3      (D) 4      (E) 5

**Solution 29.** (C) It's probably easiest to draw this out for yourself. The maximum degree of any vertex is 2, 3, or 4. If there is a vertex of degree 4, then our tree looks like a star. If the maximum degree of any vertex is 2, then we have a straight line. In the middle case, we obtain a 3-pointed star to which we attach one more vertex – the choice of branch yields isomorphic graphs. See Figure 1.



38. The maximum number of acute angles in a convex 10-gon in the Euclidean plane is

- (A) 1      (B) 2      (C) 3      (D) 4      (E) 5

**Solution 38.** (C) The total angle measure of a 10-gon is  $180 \cdot 8 = 1440^\circ$ . If the polygon is to be convex, all angles must be less than  $180^\circ$ . If we have 5 acute angles, then the remaining 5 angles would have to make up for  $> 1440 - 5 \cdot 90 = 990$  degrees. This is impossible to do and remain convex. If we have 4 acute angles, the remaining 6 angles need to make up for  $> 1440 - 4 \cdot 90 = 1080$  degrees. This is our edge case, so the answer must be 3 acute angles.

45. How many positive numbers  $x$  satisfy the equation  $\cos(97x) = x$ ?

- (A) 1      (B) 15      (C) 31      (D) 49      (E) 96

**Solution 45.** (C) Certainly our solutions are concentrated in  $[0, 1]$ . We know that every  $2\pi/97$  units in  $x$ , we get another period of  $\cos(97x)$ , and each period must meet  $y = x$  twice. Therefore there are

$$\frac{1}{2\pi/97} = \frac{97}{2\pi} \approx \frac{97}{6.3} \approx 15$$

periods in  $[0, 1]$  and about 30 meetings. There's only one answer in that range, so we'll stick with it.

## References

- H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Tsahkadsor, Armenia, USSR, 1973.
- A. Antoniadis and J. Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96:939–967, 2001. ISSN 0162-1459. doi: 10.1198/016214501753208942. URL <https://www.tandfonline.com/action/journalInformation?journalCode=uasa20>.
- S. Boyd, S. Boyd, L. Vandenberghe, and C. U. Press. *Convex Optimization*. Berichte über verteilte messysteme. Cambridge University Press, 2004. ISBN 9780521833783. URL <https://books.google.com/books?id=mYm0bLd3fc0C>.
- L. Breiman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4):373–384, 1995. URL <https://www-jstor-org.libproxy2.usc.edu/stable/pdf/1269730.pdf?refreqid=excelsior%3A76eea9bd08301e990d7d6edd86067262>.
- G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.
- D. L. Donoho and I. M. Johnstone. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, 81(3):425–455, 1994. URL <https://www-jstor-org.libproxy2.usc.edu/stable/pdf/2337118.pdf?refreqid=excelsior%3Afb36dc2b9ad4d3d57225eebb75e05df2>.
- J. J. Faraway. *Practical Regression and Anova using R*. 2002.
- G. Grimmett and D. Stirzaker. *Probability and random processes*, volume 80. Oxford university press, 2001. URL [http://scholar.google.com/scholar.bib?q=info:xzStZXK20NkJ:scholar.google.com&output=citation&hl=en&as\\_sdt=0,5&ct=citation&cd=0](http://scholar.google.com/scholar.bib?q=info:xzStZXK20NkJ:scholar.google.com&output=citation&hl=en&as_sdt=0,5&ct=citation&cd=0).
- G. M. James, P. Radchenko, and J. Lv. DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(1):127–142, 2009. URL

- <https://rss-onlinelibrary-wiley-com.libproxy1.usc.edu/doi/pdf/10.1111/j.1467-9868.2008.00668.x>.
- I. M. Johnstone. On The Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics*, 29(2):295–327, 2001. URL [https://projecteuclid.org/download/pdf\\_1/euclid-aos/1009210544](https://projecteuclid.org/download/pdf_1/euclid-aos/1009210544).
- J. Kingman and S. Taylor. *Introduction to Measure and Probability*. Cambridge University Press, 1966. URL <https://books.google.com/books?id=JbtEAAAIAAJ>.
- M. Loeve. *Probability Theory I*. Comprehensive Manuals of Surgical Specialties. Springer, 1977. ISBN 9780387902104. URL [https://books.google.com/books?id=\\_9xWB1vUEuIC](https://books.google.com/books?id=_9xWB1vUEuIC).
- M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000. ISSN 1537-2715. doi: 10.1080/10618600.2000.10474883. URL <https://www.tandfonline.com/action/journalInformation?journalCode=ucgs20>.
- M. H. Pesaran. *Time Series and Panel Data Econometrics*. Number 9780198759980 in OUP Catalogue. Oxford University Press, 2015. ISBN ARRAY(0x3bdaf68). URL <https://ideas.repec.org/b/oxp/oobooks/9780198759980.html>.
- P. C. B. Phillips and S. N. Durlauf. Multiple Time Series Regression with Integrated Processes. *The Review of Economic Studies*, 53(4):473–495, 1986. ISSN 00346527. doi: 10.2307/2297602.
- C. Rao. *Linear statistical inference and its applications*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley, 1973. ISBN 9780471708230. URL <https://books.google.com/books?id=lPhQAAAAMAAJ>.
- S. Ross. *Stochastic Processes*. Wiley series in probability and statistics. Wiley India Pvt. Limited, 2nd ed. edition, 2008. ISBN 9788126517572. URL <https://books.google.com/books?id=HVHqPgAACAAJ>.
- S. Ross. *Introduction to Probability Models*. Academic Press, Boston, eleventh edition edition, 2014. ISBN 978-0-12-407948-9. doi: <https://doi.org/10.1016/B978-0-12-407948-9.00012-8>. URL <http://www.sciencedirect.com/science/article/pii/B9780124079489000128>.
- G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978. URL <https://www.andrew.cmu.edu/user/kk3n/simplicity/schwarzbic.pdf>.
- R. Serfling. *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley, New York, NY [u.a.], [nachdr.] edition, 1980. ISBN 0471024031. URL [http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024353353&sourceid=fbw\\_bibsonomy](http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024353353&sourceid=fbw_bibsonomy).
- R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(1):1456–1490, 2013. ISSN 19357524. doi: 10.1214/13-EJS815.