Contents

Preface xi

1 Introduction 1

I Fundamentals 3

2 Probability 5

2.1 Introduction 5

2.2 Some common univariate distributions 5

2.2.1 Some common discrete distributions 5

2.2.2 Some common continuous distributions 8

2.2.3 Pareto distribution 14

2.3 The multivariate Gaussian (normal) distribution 16

2.3.1 Definition 16

2.3.2 Moment form and canonical form 17

2.3.3 Marginals and conditionals of a MVN 17

2.3.4 Bayes' rule for Gaussians 18

2.3.5 Example: sensor fusion with known measurement noise 19

2.3.6 Handling missing data 20

2.3.7 A calculus for linear Gaussian models 20

2.4 Some other multivariate continuous distributions 23

2.4.1 Multivariate Student distribution 23

2.4.2 Circular normal (von Mises Fisher) distribution 23

2.4.3 Matrix-variate Gaussian (MVG) distribution 24

2.4.4 Wishart distribution 24

2.4.5 Dirichlet distribution 27

2.5 The exponential family 28

2.5.1 Definition 29

2.5.2 Examples 30

2.5.3 Log partition function is cumulant generating function 34

2.5.4 Canonical (natural) vs mean (moment) parameters 36

2.5.5 MLE for the exponential family 37
2.5.6 Exponential dispersion family 38
2.5.7 Maximum entropy derivation of the exponential family 38
2.6 Fisher information matrix (FIM) 39
2.6.1 Definition 39
2.6.2 Equivalence between the FIM and the Hessian of the NLL 39
2.6.3 Examples 41
2.6.4 Approximating KL divergence using FIM 42
2.6.5 Fisher information matrix for exponential family 42
2.7 Transformations of random variables 44
2.7.1 Invertible transformations (bijections) 44
2.7.2 Monte Carlo approximation 44
2.7.3 Probability integral transform 44
2.8 Markov chains 46
2.8.1 Parameterization 46
2.8.2 Application: Language modeling 48
2.8.3 Parameter estimation 49
2.8.4 Stationary distribution of a Markov chain 51
2.9 Divergence measures between probability distributions 54
2.9.1 f-divergence 55
2.9.2 Integral probability metrics 56
2.9.3 Maximum mean discrepancy (MMD) 57
2.9.4 Total variation distance 60
2.9.5 Comparing distributions using binary classifiers 60

3 Statistics 63

3.1 Introduction 63
3.1.1 Frequentist statistics 63
3.1.2 Bayesian statistics 63
3.1.3 Arguments for the Bayesian approach 64
3.1.4 Arguments against the Bayesian approach 64
3.1.5 Why not just use MAP estimation? 65
3.2 Closed-form analysis using conjugate priors 69
3.2.1 The binomial model 69
3.2.2 The multinomial model 77
3.2.3 The univariate Gaussian model 79
3.2.4 The multivariate Gaussian model 84
3.2.5 Conjugate-exponential models 90
3.3 Beyond conjugate priors 92
3.3.1 Robust (heavy-tailed) priors 92
3.3.2 Priors for variance parameters 93
3.4 Noninformative priors 94
3.4.1 Maximum entropy priors 94
3.4.2 Jeffreys priors 95
3.4.3 Invariant priors 98

3.4.4 Reference priors 99
3.5 Hierarchical priors 100
3.5.1 A hierarchical binomial model 100
3.5.2 A hierarchical Gaussian model 102
3.6 Empirical Bayes 105
3.6.1 A hierarchical binomial model 106
3.6.2 A hierarchical Gaussian model 107
3.6.3 Hierarchical Bayes for n-gram smoothing 108
3.7 Model selection and evaluation 110
3.7.1 Bayesian model selection 110
3.7.2 Estimating the marginal likelihood 111
3.7.3 Connection between cross validation and marginal likelihood 111
3.7.4 Pareto-Smoothed Importance Sampling LOO estimate 113
3.7.5 Information criteria 114
3.7.6 Posterior predictive checks 116
3.7.7 Bayesian p-values 117
3.8 Bayesian decision theory 119
3.8.1 Basics 119
3.8.2 Example: COVID-19 120
3.8.3 One-shot decision problems 121
3.8.4 Multi-stage decision problems 122

4 Probabilistic graphical models 123

4.1 Introduction 123
4.2 Directed graphical models (Bayes nets) 123
4.2.1 Representing the joint distribution 123
4.2.2 Examples 124
4.2.3 Conditional independence properties 129
4.2.4 Generation (sampling) 134
4.2.5 Inference 134
4.2.6 Learning 136
4.2.7 Plate notation 141
4.3 Undirected graphical models (Markov random fields) 144
4.3.1 Representing the joint distribution 144
4.3.2 Examples 146
4.3.3 Conditional independence properties 153
4.3.4 Generation (sampling) 155
4.3.5 Inference 155
4.3.6 Learning 156
4.4 Comparing directed and undirected PGMs 160
4.4.1 CI properties 160
4.4.2 Converting between a directed and undirected model 162
4.4.3 Combining directed and undirected graphs 163
4.4.4 Comparing directed and undirected Gaussian PGMs 165
4.4.5 Factor graphs 167

4.5 Extensions of Bayes nets 170
- 4.5.1 Probabilistic circuits 170
- 4.5.2 Relational probability models 171
- 4.5.3 Open-universe probability models 173
- 4.5.4 Programs as probability models 175

4.6 Structural causal models 175
- 4.6.1 Example: causal impact of education on wealth 176
- 4.6.2 Structural equation models 177
- 4.6.3 Do operator and augmented DAGs 177
- 4.6.4 Estimating average treatment effect using path analysis 178
- 4.6.5 Counterfactuals 179

5 Information theory 183

5.1 KL divergence 183
- 5.1.1 Desiderata 184
- 5.1.2 The KL divergence uniquely satisfies the desiderata 185
- 5.1.3 Thinking about KL 188
- 5.1.4 Properties of KL 190
- 5.1.5 KL divergence and MLE 192
- 5.1.6 KL divergence and Bayesian Inference 193
- 5.1.7 KL divergence and Exponential Families 194

5.2 Entropy 195
- 5.2.1 Definition 195
- 5.2.2 Differential entropy for continuous random variables 196
- 5.2.3 Typical sets 197
- 5.2.4 Cross entropy and perplexity 199

5.3 Mutual information 200
- 5.3.1 Definition 200
- 5.3.2 Interpretation 200
- 5.3.3 Data processing inequality 201
- 5.3.4 Sufficient Statistics 202
- 5.3.5 Multivariate mutual information 202
- 5.3.6 Variational bounds on mutual information 205

5.4 Data compression (source coding) 208
- 5.4.1 Lossless compression 208
- 5.4.2 Lossy compression and the rate-distortion tradeoff 208
- 5.4.3 Bits back coding 211

5.5 Error-correcting codes (channel coding) 211

5.6 The information bottleneck 213
- 5.6.1 Vanilla IB 213
- 5.6.2 Variational IB 214
- 5.6.3 Conditional entropy bottleneck 215

6 Optimization 219

6.1 Introduction 219

6.2 Stochastic gradient descent 219
6.3 Natural gradient descent 220
6.3.1 Defining the natural gradient 221
6.3.2 Interpretations of NGD 221
6.3.3 Benefits of NGD 222
6.3.4 Approximating the natural gradient 223
6.3.5 Natural gradients for the exponential family 224
6.4 Mirror descent 227
6.4.1 Bregman divergence 227
6.4.2 Proximal point method 228
6.4.3 PPM using Bregman divergence 229
6.5 Gradients of stochastic functions 229
6.5.1 Minibatch approximation to finite-sum objectives 229
6.5.2 Optimizing parameters of a distribution 230
6.5.3 Score function estimator (likelihood ratio trick) 230
6.5.4 Reparameterization trick 232
6.5.5 The delta method 233
6.5.6 Gumbel softmax trick 234
6.5.7 Stochastic computation graphs 234
6.5.8 Straight-through estimator 234
6.6 Bound optimization (MM) algorithms 236
6.6.1 The general algorithm 236
6.6.2 Example: logistic regression 237
6.6.3 The EM algorithm 238
6.6.4 Example: EM for an MVN with missing data 240
6.6.5 Example: robust linear regression using Student-t likelihood 242
6.6.6 Extensions to EM 244
6.7 The Bayesian learning rule 246
6.7.1 Deriving inference algorithms from BLR 247
6.7.2 Deriving optimization algorithms from BLR 249
6.7.3 Variational optimization 252
6.8 Bayesian optimization 253
6.8.1 Sequential model-based optimization 253
6.8.2 Surrogate functions 255
6.8.3 Acquisition functions 256
6.8.4 Other issues 259
6.9 Optimal Transport 260
6.9.1 Warm-up: Matching optimally two families of points 260
6.9.2 From Optimal Matchings to Kantorovich and Monge formulations 261
6.9.3 Solving optimal transport 263
6.10 Submodular optimization 268
6.10.1 Intuition, Examples, and Background 269
6.10.2 Submodular Basic Definitions 271
6.10.3 Example Submodular Functions 272
6.10.4 Submodular Optimization 275

6.10.5 Applications of Submodularity in Machine Learning and AI 279
6.10.6 Sketching, CoreSets, Distillation, and Data Subset & Feature Selection 279
6.10.7 Combinatorial Information Functions 283
6.10.8 Clustering, Data Partitioning, and Parallel Machine Learning 284
6.10.9 Active and Semi-Supervised Learning 284
6.10.10 Probabilistic Modeling 285
6.10.11 Structured Norms and Loss Functions 287
6.10.12 Conclusions 287
6.11 Derivative free optimization 288

II Inference 289

7 Inference algorithms: an overview 291
7.1 Introduction 291
7.2 Common inference patterns 291
7.2.1 Global latents 292
7.2.2 Local latents 292
7.2.3 Global and local latents 293
7.3 Exact inference algorithms 293
7.4 Approximate inference algorithms 294
7.4.1 MAP estimation 294
7.4.2 Grid approximation 294
7.4.3 Laplace (quadratic) approximation 295
7.4.4 Variational inference 296
7.4.5 Markov Chain Monte Carlo (MCMC) 298
7.4.6 Sequential Monte Carlo 299
7.5 Evaluating approximate inference algorithms 299

8 State-space inference 301
8.1 Introduction 301
8.1.1 State space models 301
8.1.2 Example: casino HMM 303
8.1.3 Example: linear-Gaussian SSM for tracking in 2d 304
8.1.4 Inferential goals 304
8.2 Bayesian filtering and smoothing 307
8.2.1 The filtering equations 308
8.2.2 The smoothing equations 308
8.3 Inference for discrete SSMs 309
8.3.1 Forwards filtering 309
8.3.2 Backwards smoothing 311
8.3.3 The forwards-backwards algorithm 311
8.3.4 Two-slice smoothed marginals 313
8.3.5 Time and space complexity 314
8.3.6 The Viterbi algorithm 315

8.3.7 Forwards filtering, backwards sampling 318
8.3.8 Application to discretized state spaces 318
8.4 Inference for linear-Gaussian SSMs 319
8.4.1 The Kalman filter 319
8.4.2 Kalman filtering for linear regression (recursive least squares) 324
8.4.3 Predictive coding as Kalman filtering 326
8.4.4 The Kalman (RTS) smoother 328
8.5 Inference based on local linearization 329
8.5.1 Taylor series expansion 329
8.5.2 The extended Kalman filter (EKF) 332
8.5.3 The extended Kalman smoother 335
8.5.4 Exponential-family EKF 335
8.6 Inference based on the unscented transform 337
8.6.1 The unscented transform 338
8.6.2 The unscented Kalman filter (UKF) 339
8.6.3 The unscented Kalman smoother 341
8.7 Other variants of the Kalman filter 342
8.7.1 Ensemble Kalman filter 342
8.7.2 Robust Kalman filters 343
8.7.3 Gaussian filtering 343
8.8 Assumed density filtering 346
8.8.1 The ADF algorithm 346
8.8.2 Connection with Gaussian filtering 347
8.8.3 The Gaussian sum filter for switching SSMs 347
8.8.4 ADF for training logistic regression 350

9 Message passing inference 355
9.1 Introduction 355
9.2 Belief propagation on trees 356
9.2.1 BP for polytrees 356
9.2.2 BP for undirected graphs with pairwise potentials 359
9.2.3 BP for factor graphs 360
9.2.4 Max product belief propagation 361
9.2.5 Gaussian and non-Gaussian belief propagation 363
9.3 Loopy belief propagation 363
9.3.1 Convergence 364
9.3.2 Accuracy 366
9.3.3 Connection with variational inference 367
9.3.4 Generalized belief propagation 367
9.3.5 Application: error correcting codes 367
9.3.6 Application: Affinity propagation 369
9.3.7 Emulating BP with graph neural nets 370
9.4 The variable elimination (VE) algorithm 371
9.4.1 Derivation of the algorithm 371
9.4.2 Computational complexity of VE 372

9.4.3 Computational complexity of exact inference 374
9.4.4 Drawbacks of VE 375
9.5 The junction tree algorithm (JTA) 376
9.5.1 Creating a junction tree 376
9.5.2 Running belief propagation on a junction tree 381
9.5.3 The generalized distributive law 382
9.5.4 Other applications of the JTA 383
9.6 Inference as backpropagation 383

10 Variational inference 387

10.1 Introduction 387
10.1.1 Variational free energy 387
10.1.2 Evidence lower bound (ELBO) 388
10.2 Mean field VI 389
10.2.1 Coordinate ascent variational inference (CAVI) 389
10.2.2 Example: CAVI for the Ising model 390
10.2.3 Variational Bayes 392
10.2.4 Example: VB for a univariate Gaussian 393
10.2.5 Variational Bayes EM 396
10.2.6 Example: VBEM for a GMM 397
10.2.7 Variational message passing (VMP) 403
10.2.8 Autoconj 404
10.3 Fixed-form VI 404
10.3.1 Black-box variational inference 404
10.3.2 Stochastic variational inference 406
10.3.3 Reparameterization VI 407
10.3.4 Gaussian VI 408
10.3.5 Automatic differentiation VI 412
10.3.6 Beyond Gaussian posteriors 413
10.3.7 Amortized inference 415
10.3.8 Exploiting partial conjugacy 416
10.3.9 Online variational inference 420
10.4 More accurate variational posteriors 423
10.4.1 Structured mean field 424
10.4.2 Hierarchical (auxiliary variable) posteriors 424
10.4.3 Normalizing flow posteriors 424
10.4.4 Implicit posteriors 426
10.4.5 Combining VI with MCMC inference 427
10.5 Lower bounds 427
10.5.1 Multi-sample ELBO (IWAE bound) 427
10.5.2 The thermodynamic variational objective (TVO) 428
10.6 Upper bounds 428
10.6.1 Minimizing the χ-divergence upper bound 429
10.6.2 Minimizing the evidence upper bound 430
10.7 Expectation propagation (EP) 431

10.7.1 Minimizing forwards vs reverse KL 431
10.7.2 EP as generalized ADF 433
10.7.3 Algorithm 433
10.7.4 Example 434
10.7.5 Optimization issues 434
10.7.6 Power EP and α-divergence 435
10.7.7 Stochastic EP 435
10.7.8 Applications 436

11 Monte Carlo inference 437
11.1 Introduction 437
11.2 Monte Carlo integration 437
11.2.1 Example: estimating π by Monte Carlo integration 438
11.2.2 Accuracy of Monte Carlo integration 438
11.3 Generating random samples from simple distributions 440
11.3.1 Sampling using the inverse cdf 440
11.3.2 Sampling from a Gaussian (Box-Muller method) 441
11.4 Rejection sampling 441
11.4.1 Basic idea 442
11.4.2 Example 443
11.4.3 Adaptive rejection sampling 443
11.4.4 Rejection sampling in high dimensions 444
11.5 Importance sampling 444
11.5.1 Direct importance sampling 445
11.5.2 Self-normalized importance sampling 445
11.5.3 Choosing the proposal 446
11.5.4 Annealed importance sampling (AIS) 446
11.6 Controlling Monte Carlo variance 448
11.6.1 Rao-Blackwellisation 448
11.6.2 Control variates 449
11.6.3 Antithetic sampling 450
11.6.4 Quasi Monte Carlo (QMC) 451

12 Markov Chain Monte Carlo inference 453
12.1 Introduction 453
12.2 Metropolis Hastings algorithm 453
12.2.1 Basic idea 454
12.2.2 Why MH works 455
12.2.3 Proposal distributions 456
12.2.4 Initialization 459
12.2.5 Simulated annealing 459
12.3 Gibbs sampling 461
12.3.1 Basic idea 462
12.3.2 Gibbs sampling is a special case of MH 462
12.3.3 Example: Gibbs sampling for Ising models 463

12.3.4 Example: Gibbs sampling for Potts models 464
12.3.5 Example: Gibbs sampling for GMMs 465
12.3.6 Sampling from the full conditionals 466
12.3.7 Blocked Gibbs sampling 467
12.3.8 Collapsed Gibbs sampling 468
12.4 Auxiliary variable MCMC 470
12.4.1 Slice sampling 471
12.4.2 Swendsen Wang 472
12.5 Hamiltonian Monte Carlo (HMC) 474
12.5.1 Hamiltonian mechanics 474
12.5.2 Integrating Hamilton's equations 475
12.5.3 The HMC algorithm 476
12.5.4 Tuning HMC 477
12.5.5 Riemann Manifold HMC 478
12.5.6 Langevin Monte Carlo (MALA) 479
12.5.7 Connection between SGD and Langevin sampling 480
12.5.8 Applying HMC to constrained parameters 482
12.5.9 Speeding up HMC 482
12.6 MCMC convergence 483
12.6.1 Mixing rates of Markov chains 484
12.6.2 Practical convergence diagnostics 484
12.6.3 Improving speed of convergence 491
12.6.4 Non-centered parameterizations and Neal's funnel 492
12.7 Stochastic gradient MCMC 493
12.7.1 Stochastic Gradient Langevin Dynamics (SGLD) 494
12.7.2 Preconditionining 494
12.7.3 Reducing the variance of the gradient estimate 495
12.7.4 SG-HMC 496
12.7.5 Underdamped Langevin Dynamics 497
12.8 Reversible jump (trans-dimensional) MCMC 498
12.8.1 Basic idea 498
12.8.2 Example 500
12.8.3 Discussion 501
12.9 Annealing methods 501
12.9.1 Parallel tempering 501

13 Sequential Monte Carlo inference 503
13.1 Introduction 503
13.1.1 Problem statement 503
13.1.2 Particle filtering for state-space models 503
13.1.3 SMC samplers for static parameter estimation 505
13.2 Basics of SMC 505
13.2.1 Importance sampling 505
13.2.2 Sequential importance sampling 506
13.2.3 Sequential importance sampling with resampling 507

13.2.4 Resampling methods 510
13.2.5 Adaptive resampling 512
13.3 Some applications of particle filtering 513
13.3.1 1d pendulum model with outliers 513
13.3.2 Visual object tracking 514
13.3.3 Robot localization 515
13.3.4 Online parameter estimation 516
13.4 Proposal distributions 517
13.4.1 Locally optimal proposal 518
13.4.2 Proposals based on the Laplace approximation 518
13.4.3 Proposals based on the extended and unscented Kalman filter 520
13.4.4 Proposals based on SMC 520
13.4.5 Neural adaptive SMC 521
13.4.6 Amortized adaptive SMC 521
13.4.7 Variational SMC 522
13.5 Rao-Blackwellised particle filtering (RBPF) 523
13.5.1 Mixture of Kalman filters 523
13.5.2 FastSLAM 525
13.6 SMC samplers 527
13.6.1 Ingredients of an SMC sampler 527
13.6.2 Likelihood tempering (geometric path) 528
13.6.3 Data tempering 531
13.6.4 Sampling rare events and extrema 532
13.6.5 SMC-ABC and likelihood-free inference 533
13.6.6 SMC^2 534
13.7 Particle MCMC methods 534
13.7.1 Particle Marginal Metropolis Hastings 534
13.7.2 Particle Independent Metropolis Hastings 535
13.7.3 Particle Gibbs 536

III Prediction 537

14 Prediction models: an overview 539
14.1 Introduction 539
14.1.1 Types of model 539
14.1.2 Model fitting using ERM, MLE and MAP 540
14.1.3 Model fitting using Bayes, VI and generalized Bayes 541
14.2 Evaluating predictive models 542
14.2.1 Proper scoring rules 542
14.2.2 Calibration 542
14.2.3 Beyond evaluating marginal probabilities 546
14.3 Conformal prediction 549
14.3.1 Conformalizing classification 550
14.3.2 Conformalizing regression 551

14.3.3 Conformalizing Bayes 552
14.3.4 What do we do if we don't have a calibration set? 552

15 Generalized linear models 553

15.1 Introduction 553
15.1.1 Examples 553
15.1.2 GLMs with non-canonical link functions 556
15.1.3 Maximum likelihood estimation 556
15.1.4 Bayesian inference 557
15.2 Linear regression 558
15.2.1 Conjugate priors 558
15.2.2 Uninformative priors 560
15.2.3 Informative priors 562
15.2.4 Spike and slab prior 564
15.2.5 Laplace prior (Bayesian lasso) 565
15.2.6 Horseshoe prior 566
15.2.7 Automatic relevancy determination 567
15.3 Logistic regression 569
15.3.1 Binary logistic regression 570
15.3.2 Multinomial logistic regression 570
15.3.3 Priors 571
15.3.4 Posteriors 572
15.3.5 Laplace approximation 572
15.3.6 MCMC inference 575
15.3.7 Variational inference 576
15.4 Probit regression 576
15.4.1 Latent variable interpretation 576
15.4.2 Maximum likelihood estimation 577
15.4.3 Bayesian inference 578
15.4.4 Ordinal probit regression 579
15.4.5 Multinomial probit models 580
15.5 Multi-level GLMs 580
15.5.1 Generalized linear mixed models (GLMMs) 580
15.5.2 Model fitting 581
15.5.3 Example: radon regression 581

16 Deep neural networks 585

16.1 Introduction 585
16.2 Building blocks of differentiable circuits 585
16.2.1 Linear layers 586
16.2.2 Non-linearities 586
16.2.3 Convolutional layers 587
16.2.4 Residual (skip) connections 588
16.2.5 Normalization layers 589
16.2.6 Dropout layers 589

16.2.7 Attention layers 590
16.2.8 Recurrent layers 592
16.2.9 Multiplicative layers 593
16.2.10 Implicit layers 594
16.3 Canonical examples of neural networks 594
16.3.1 Multi-layer perceptrons (MLP) 594
16.3.2 Convolutional neural networks (CNN) 595
16.3.3 Recurrent neural networks (RNN) 595
16.3.4 Transformers 597
16.3.5 Graph neural networks (GNNs) 601
16.4 Automatic differentiation 604
16.4.1 Differentiation in functional form 605
16.4.2 Differentiating chains, circuits, and programs 609

17 Bayesian neural networks 617

17.1 Introduction 617
17.2 Priors for BNNs 617
17.2.1 Gaussian priors 618
17.2.2 Sparsity-promoting priors 619
17.2.3 Learning the prior 619
17.2.4 Priors in function space 620
17.2.5 Architectural priors 620
17.3 Likelihoods for BNNs 621
17.4 Posteriors for BNNs 622
17.4.1 Laplace approximation 622
17.4.2 Variational inference 623
17.4.3 Expectation propagation 624
17.4.4 Last layer methods 624
17.4.5 Dropout 624
17.4.6 MCMC methods 625
17.4.7 Methods based on the SGD trajectory 625
17.4.8 Deep ensembles 626
17.4.9 Approximating the posterior predictive distibution 629
17.5 Generalization in Bayesian deep learning 630
17.5.1 Sharp vs flat minima 631
17.5.2 Effective dimensionality of a model 632
17.5.3 The hypothesis space of DNNs 633
17.5.4 Double descent 634
17.5.5 A Bayesian Resolution to Double Descent 637
17.5.6 PAC-Bayes 638
17.5.7 Out-of-Distribution Generalization for BNNs 639
17.6 Online inference 641
17.6.1 Extended Kalman Filtering for DNNs 642
17.6.2 Assumed Density Filtering for DNNs 644
17.6.3 Sequential Laplace for DNNs 645

17.6.4 Variational methods 646
17.7 Hierarchical Bayesian neural networks 646
17.7.1 Solving multiple related classification problems 646

18 Gaussian processes 651

18.1 Introduction 651
18.2 Mercer kernels 653
18.2.1 Some popular Mercer kernels 654
18.2.2 Mercer’s theorem 659
18.2.3 Kernels from Spectral Densities 660
18.3 GPs with Gaussian likelihoods 662
18.3.1 Predictions using noise-free observations 662
18.3.2 Predictions using noisy observations 663
18.3.3 Weight space vs function space 664
18.3.4 Semi-parametric GPs 665
18.3.5 Marginal likelihood 666
18.3.6 Computational and numerical issues 666
18.3.7 Kernel ridge regression 667
18.4 GPs with non-Gaussian likelihoods 670
18.4.1 Binary classification 670
18.4.2 Multi-class classification 672
18.4.3 GPs for Poisson regression (Cox process) 672
18.5 Scaling GP inference to large datasets 673
18.5.1 Subset of data 674
18.5.2 Nyström approximation 674
18.5.3 Inducing point methods 675
18.5.4 Sparse variational methods 678
18.5.5 Exploiting parallelization and structure via kernel matrix multiplies 682
18.6 Learning the kernel 684
18.6.1 Empirical Bayes for the kernel parameters 685
18.6.2 Bayesian inference for the kernel parameters 687
18.6.3 Multiple kernel learning for additive kernels 688
18.6.4 Automatic search for compositional kernels 690
18.6.5 Spectral mixture kernel learning 692
18.6.6 Deep kernel learning 694
18.6.7 Functional kernel learning 696
18.7 GPs and DNNs 696
18.7.1 Kernels derived from random DNNs (NN-GP) 696
18.7.2 Kernels derived from trained DNNs (neural tangent kernel) 700
18.7.3 Deep GPs 702

19 Structured prediction 709

19.1 Introduction 709
19.2 Conditional random fields (CRFs) 709
19.2.1 1d CRFs 709

19.2.2 2d CRFs 713
19.2.3 Parameter estimation 715
19.3 Time series forecasting 716
19.3.1 Structural time series models 716
19.3.2 Prophet 722
19.3.3 Gaussian processes for timeseries forecasting 723
19.3.4 Neural forecasting methods 725
19.3.5 Causal impact of a time series intervention 726

20 Beyond the iid assumption 731
20.1 Introduction 731
20.2 Distribution shift 731
20.2.1 Motivating examples 731
20.2.2 A causal view of distribution shift 733
20.2.3 Covariate shift 734
20.2.4 Domain shift 734
20.2.5 Label / prior shift 735
20.2.6 Concept shift 735
20.2.7 Manifestation shift 735
20.2.8 Selection bias 735
20.3 Training-time techniques for distribution shift 736
20.3.1 Importance weighting for covariate shift 736
20.3.2 Domain adaptation 738
20.3.3 Domain randomization 738
20.3.4 Data augmentation 739
20.3.5 Unsupervised label shift estimation 739
20.3.6 Distributionally robust optimization 739
20.4 Test-time techniques for distribution shift 740
20.4.1 Detecting shifts using two-sample testing 740
20.4.2 Detecting single out-of-distribution (OOD) inputs 740
20.4.3 Selective prediction 743
20.4.4 Open world recognition 745
20.4.5 Online adaptation 745
20.5 Learning from multiple distributions 745
20.5.1 Transfer learning 746
20.5.2 Few-shot learning 747
20.5.3 Multi-task learning 748
20.5.4 Domain generalization 748
20.5.5 Invariant risk minimization 750
20.6 Meta-learning 751
20.6.1 Meta-learning as probabilistic inference for prediction 751
20.6.2 Gradient-based meta-learning 753
20.6.3 Metric-based few-shot learning 753
20.6.4 VERSA 753
20.6.5 Neural processes 754

20.7 Continual learning 754
20.7.1 Domain drift 754
20.7.2 Concept drift 754
20.7.3 Task incremental learning 756
20.7.4 Catastrophic forgetting 757
20.7.5 Online learning 759
20.8 Adversarial examples 760
20.8.1 Whitebox (gradient-based) attacks 762
20.8.2 Blackbox (gradient-free) attacks 762
20.8.3 Real world adversarial attacks 764
20.8.4 Defenses based on robust optimization 764
20.8.5 Why models have adversarial examples 765

IV Generation 767

21 Generative models: an overview 769
21.1 Introduction 769
21.2 Types of generative model 769
21.3 Goals of generative modeling 771
21.3.1 Generating data 771
21.3.2 Density estimation 772
21.3.3 Imputation 773
21.3.4 Structure discovery 773
21.3.5 Latent space interpolation 774
21.3.6 Representation learning 775
21.4 Evaluating generative models 775
21.4.1 Likelihood 776
21.4.2 Distances and divergences in feature space 778
21.4.3 Precision and recall metrics 779
21.4.4 Statistical tests 780
21.4.5 Challenges with using pretrained classifiers 780
21.4.6 Using model samples to train classifiers 780
21.4.7 Assessing overfitting 781
21.4.8 Human evaluation 781

22 Variational autoencoders 783
22.1 Introduction 783
22.2 VAE basics 783
22.2.1 Modeling assumptions 784
22.2.2 Evidence lower bound 785
22.2.3 Optimization 786
22.2.4 The reparameterization trick 786
22.2.5 Computing the reparameterized ELBO 788
22.2.6 Comparison of VAEs and autoencoders 790

22.2.7 VAEs optimize in an augmented space 791
22.3 VAE generalizations 793
22.3.1 σ-VAE 793
22.3.2 β-VAE 794
22.3.3 InfoVAE 796
22.3.4 Multi-modal VAEs 798
22.3.5 VAEs with missing data 801
22.3.6 Semi-supervised VAEs 803
22.3.7 VAEs with sequential encoders/decoders 804
22.4 Avoiding posterior collapse 807
22.4.1 KL annealing 808
22.4.2 Lower bounding the rate 808
22.4.3 Free bits 808
22.4.4 Adding skip connections 808
22.4.5 Improved variational inference 808
22.4.6 Alternative objectives 809
22.4.7 Enforcing identifiability 809
22.5 VAEs with hierarchical structure 810
22.5.1 Bottom-up vs top-down inference 811
22.5.2 Example: Very deep VAE 812
22.5.3 Connection with autoregressive models 814
22.5.4 Variational pruning 815
22.5.5 Other optimization difficulties 815
22.6 Vector quantization VAE 816
22.6.1 Autoencoder with binary code 816
22.6.2 VQ-VAE model 817
22.6.3 Learning the prior 819
22.6.4 Hierarchical extension (VQ-VAE-2) 819
22.6.5 Discrete VAE 820
22.6.6 VQ-GAN 821
22.7 Wake-sleep algorithm 821
22.7.1 Wake phase 822
22.7.2 Sleep phase 823
22.7.3 Daydream phase 824
22.7.4 Summary of algorithm 824

23 Auto-regressive models 827
23.1 Introduction 827
23.2 Neural autoregressive density estimators (NADE) 828
23.3 Causal CNNs 828
23.3.1 1d causal CNN (Convolutional Markov models) 829
23.3.2 2d causal CNN (PixelCNN) 829
23.4 Transformer decoders 830
23.4.1 Text generation (GPT) 831
23.4.2 Music generation 831

23.4.3 Text-to-image generation (DALL-E) 832

24 Normalizing Flows 835

24.1 Introduction 835
24.1.1 Preliminaries 835
24.1.2 Example 837
24.1.3 How to train a flow model 838
24.2 Constructing Flows 839
24.2.1 Affine flows 839
24.2.2 Elementwise flows 840
24.2.3 Coupling flows 842
24.2.4 Autoregressive flows 844
24.2.5 Residual flows 848
24.2.6 Continuous-time flows 851
24.3 Applications 852
24.3.1 Density estimation 852
24.3.2 Generative Modeling 853
24.3.3 Inference 853

25 Energy-based models 855

25.1 Introduction 855
25.1.1 Example: Products of experts (PoE) 856
25.1.2 Computational difficulties 856
25.2 Maximum Likelihood Training 857
25.2.1 Gradient-based MCMC methods 858
25.2.2 Contrastive divergence 858
25.3 Score Matching (SM) 861
25.3.1 Basic score matching 862
25.3.2 Denoising Score Matching (DSM) 862
25.3.3 Sliced Score Matching (SSM) 864
25.3.4 Connection to Contrastive Divergence 865
25.3.5 Score-Based Generative Models 866
25.4 Noise Contrastive Estimation 868
25.4.1 Connection to Score Matching 870
25.5 Other Methods 871
25.5.1 Minimizing Differences/Derivatives of KL Divergences 871
25.5.2 Minimizing the Stein Discrepancy 871
25.5.3 Adversarial Training 872

26 Denoising diffusion models 875

26.1 Model definition 875
26.2 Examples 877
26.3 Model training 878
26.4 Connections with other generative models 880
26.4.1 Connection with score matching 880

26.4.2 Connection with VAEs 881
26.4.3 Connection with flow models 881

27 Generative adversarial networks 883
27.1 Introduction 883
27.2 Learning by Comparison 884
27.2.1 Guiding principles 885
27.2.2 Class probability estimation 886
27.2.3 Bounds on f-divergences 889
27.2.4 Integral probability metrics 890
27.2.5 Moment matching 892
27.2.6 On density ratios and differences 893
27.3 Generative Adversarial Networks 894
27.3.1 From learning principles to loss functions 895
27.3.2 Gradient Descent 896
27.3.3 Challenges with GAN training 897
27.3.4 Improving GAN optimization 899
27.3.5 Convergence of GAN training 899
27.4 Conditional GANs 902
27.5 Inference with GANs 904
27.6 Neural architectures in GANs 904
27.6.1 The importance of discriminator architectures 905
27.6.2 Architectural inductive biases 905
27.6.3 Attention in GANs 905
27.6.4 Progressive generation 906
27.6.5 Regularization 907
27.6.6 Scaling up GAN models 908
27.7 Applications 908
27.7.1 GANs for image generation 909
27.7.2 Video generation 911
27.7.3 Audio generation 912
27.7.4 Text generation 912
27.7.5 Imitation Learning 913
27.7.6 Domain Adaptation 914
27.7.7 Design, Art and Creativity 914

V Discovery 915

28 Discovery methods: an overview 917
28.1 Introduction 917
28.2 Overview of Part V 918

29 Latent variable models 919
29.1 Introduction 919
29.2 Mixture models 919

29.2.1 Gaussian mixture models (GMMs) 920
29.2.2 Bernoulli mixture models 922
29.2.3 Gaussian scale mixtures 922
29.2.4 Using GMMs as a prior for inverse imaging problems 924
29.3 Factor analysis 927
29.3.1 Vanilla factor analysis 927
29.3.2 Probabilistic PCA 931
29.3.3 Factor analysis models for paired data 934
29.3.4 Factor analysis with exponential family likelihoods 937
29.3.5 Factor analysis with DNN likelihoods 938
29.3.6 Factor analysis with GP likelihoods (GP-LVM) 939
29.4 Mixture of factor analysers 941
29.4.1 Model definition 941
29.4.2 Model fitting 942
29.4.3 MixFA for image generation 943
29.5 LVMs with non-Gaussian priors 947
29.5.1 Non-negative matrix factorization (NMF) 947
29.5.2 Multinomial PCA 948
29.5.3 Latent Dirichlet Allocation (LDA) 950
29.6 Independent components analysis (ICA) 951
29.6.1 Noiseless ICA model 952
29.6.2 The need for non-Gaussian priors 952
29.6.3 Maximum likelihood estimation 953
29.6.4 Alternatives to MLE 954
29.6.5 Sparse coding 956
29.6.6 Nonlinear ICA 957

30 Hidden Markov models 959

30.1 Introduction 959
30.2 HMMs: parameterization 959
30.2.1 Transition model 959
30.2.2 Observation model 960
30.3 HMMs: Applications 963
30.3.1 Segmentation of time series data 963
30.3.2 Spelling correction 965
30.3.3 Protein sequence alignment 968
30.4 HMMs: parameter learning 969
30.4.1 The Baum-Welch (EM) algorithm 969
30.4.2 Parameter estimation using SGD 973
30.4.3 Parameter estimation using spectral methods 975
30.4.4 Bayesian parameter inference 976
30.5 HMMs: Generalizations 976
30.5.1 Hidden semi-Markov model (HSMM) 977
30.5.2 HSMMs for changepoint detection 979
30.5.3 Hierarchical HMMs 982

30.5.4 Factorial HMMs 984
30.5.5 Coupled HMMs 986
30.5.6 Dynamic Bayes nets (DBN) 987

31 State-space models 989
31.1 Introduction 989
31.2 Linear dynamical systems 989
31.2.1 Example: Noiseless 1d spring-mass system 990
31.2.2 Example: Noisy 2d tracking problem 991
31.2.3 Example: Online linear regression 994
31.2.4 Example: structural time series forecasting 996
31.2.5 Parameter estimation 996
31.3 Non-linear dynamical systems 998
31.3.1 Example: nonlinear 2d tracking problem 999
31.3.2 Example: Simultaneous localization and mapping (SLAM) 999
31.3.3 Example: stochastic volatility models 1001
31.3.4 Example: Multi-target tracking 1002
31.4 Other kinds of SSM 1004
31.4.1 Exponential family SSM 1004
31.4.2 Bayesian SSM 1008
31.4.3 GP-SSM **(Unfinished)** 1008
31.5 Deep state space models **(Unfinished)** 1008
31.5.1 Deep Markov models 1009
31.5.2 Recurrent SSM 1010
31.5.3 Improving multi-step predictions 1011
31.5.4 Variational RNNs 1012
31.5.5 Structured State Space Sequence model (S4) 1013

32 Graph learning 1017
32.1 Introduction 1017
32.2 Latent variable models for graphs 1017
32.2.1 Stochastic block model 1017
32.2.2 Mixed membership stochastic block model 1019
32.2.3 Infinite relational model 1021
32.3 Graphical model structure learning 1022
32.3.1 Applications 1023
32.3.2 Relevance networks 1024
32.3.3 Learning sparse PGMs 1026

33 Non-parametric Bayesian models 1027
33.1 Introduction 1027
33.2 Dirichlet process 1028
33.2.1 Definition 1028
33.2.2 Stick breaking construction of the DP 1030
33.2.3 The Chinese restaurant process (CRP) 1031

33.2.4 Dirichlet process mixture models 1033
33.3 Generalizations of the Dirichlet process 1038
33.3.1 Pitman-Yor process 1039
33.3.2 Dependent random probability measures 1040
33.4 The Indian buffet process and the Beta process 1042
33.5 Small-variance asymptotics 1045
33.6 Completely random measures 1048
33.7 Lévy processes 1049
33.8 Point processes with repulsion and reinforcement 1051
33.8.1 Poisson process 1051
33.8.2 Renewal process 1052
33.8.3 Hawkes process 1053
33.8.4 Gibbs point process 1055
33.8.5 Determinantal point process 1056

34 Representation learning (Unfinished) 1059
34.1 CLIP 1059

35 Interpretability 1061
35.1 Introduction 1061
35.1.1 The Role of Interpretability 1062
35.1.2 Terminology and Framework 1063
35.2 Methods for Interpretable Machine Learning 1067
35.2.1 Inherently Interpretable Models: The Model is its Explanation 1067
35.2.2 Semi-Inherently Interpretable Models: Example-Based Methods 1069
35.2.3 Post-hoc or Joint training: The Explanation gives a Partial View of the Model 1070
35.2.4 Transparency and Visualization 1074
35.3 Properties: The Abstraction Between Context and Method 1075
35.3.1 Properties of Explanations from Interpretable Machine Learning 1075
35.3.2 Properties of Explanations from Cognitive Science 1078
35.4 Evaluation of Interpretable Machine Learning Models 1079
35.4.1 Computational Evaluation: Does the Method have Desired Properties? 1080
35.4.2 User Study-based Evaluation: Does the Method Help a User Perform a Task? 1084
35.5 Discussion: How to Think about Interpretable Machine Learning 1088

VI Decision making 1095

36 Multi-step decision problems 1097
36.1 Introduction 1097
36.2 Decision (influence) diagrams 1097
36.2.1 Example: oil wildcatter 1097
36.2.2 Information arcs 1098
36.2.3 Value of information 1099

36.2.4 Computing the optimal policy 1100
36.3 A/B testing 1100
36.3.1 A Bayesian approach 1101
36.3.2 Example 1104
36.4 Contextual bandits 1105
36.4.1 Types of bandit 1105
36.4.2 Applications 1107
36.4.3 Exploration-exploitation tradeoff 1107
36.4.4 The optimal solution 1107
36.4.5 Upper confidence bounds (UCB) 1109
36.4.6 Thompson sampling 1111
36.4.7 Regret 1112
36.5 Markov decision problems 1113
36.5.1 Basics 1114
36.5.2 Partially observed MDPs 1115
36.5.3 Episodes and returns 1116
36.5.4 Value functions 1116
36.5.5 Optimal value functions and policies 1117
36.6 Planning in an MDP 1118
36.6.1 Value iteration 1119
36.6.2 Policy iteration 1120
36.6.3 Linear programming 1121

37 Reinforcement learning 1123
37.1 Introduction 1123
37.1.1 Overview of methods 1123
37.1.2 Value based methods 1124
37.1.3 Policy search methods 1124
37.1.4 Model-based RL 1125
37.1.5 Exploration-exploitation tradeoff 1125
37.2 Value-based RL 1127
37.2.1 Monte Carlo RL 1127
37.2.2 Temporal difference (TD) learning 1128
37.2.3 TD learning with eligibility traces 1129
37.2.4 SARSA: on-policy TD control 1129
37.2.5 Q-learning: off-policy TD control 1130
37.2.6 Deep Q-network (DQN) 1133
37.3 Policy-based RL 1134
37.3.1 The policy gradient theorem 1134
37.3.2 REINFORCE 1135
37.3.3 Actor-critic methods 1135
37.3.4 Bound optimization methods 1137
37.3.5 Deterministic policy gradient methods 1139
37.3.6 Gradient-free methods 1140
37.4 Model-based RL 1140

37.4.1 Model predictive control (MPC) 1141
37.4.2 Combining model-based and model-free 1142
37.4.3 MBRL using Gaussian processes 1143
37.4.4 MBRL using DNNs 1144
37.4.5 MBRL using latent-variable models 1145
37.4.6 Robustness to model errors 1147
37.5 Off-policy learning 1147
37.5.1 Basic techniques 1148
37.5.2 The curse of horizon 1151
37.5.3 The deadly triad 1152
37.6 Control as inference 1154
37.6.1 Maximum entropy reinforcement learning 1154
37.6.2 Active inference 1156
37.6.3 Other approaches 1157
37.6.4 Imitation learning 1158

38 Causality 1161
38.1 Introduction 1161
38.1.1 Why is causality different than other forms of ML? 1161
38.2 Causal Formalism 1163
38.2.1 Structural Causal Models 1163
38.2.2 Causal DAGs 1165
38.2.3 Identification 1167
38.2.4 Counterfactuals and the Causal Hierarchy 1168
38.3 Randomized Control Trials 1170
38.4 Confounder Adjustment 1171
38.4.1 Causal Estimand, Statistical Estimand, and Identification 1171
38.4.2 ATE Estimation with Observed Confounders 1174
38.4.3 Uncertainity Quantification 1179
38.4.4 Matching 1180
38.4.5 Practical Considerations and Procedures 1181
38.4.6 Summary and Practical Advice 1184
38.5 Instrumental Variable Strategies 1185
38.5.1 Additive Unobserved Confounding 1187
38.5.2 Instrument Monotonicity and Local Average Treatment Effect 1188
38.5.3 Two Stage Least Squares 1192
38.6 Difference in Differences 1192
38.6.1 Estimation 1196
38.7 Credibility Checks 1196
38.7.1 Placebo Checks 1197
38.7.2 Sensitivity Analysis to Unobserved Confounding 1197
38.8 The Do Calculus 1205
38.8.1 The three rules 1205
38.8.2 Revisiting Backdoor Adjustment 1206
38.8.3 Frontdoor Adjustment 1207

38.9 Further Reading 1209

Bibliography **1223**