Contents

Preface **xi**

1 Introduction **1**

I Fundamentals **3**

2 Probability **5**

2.1 Introduction 5
2.2 Some common probability distributions 5
2.2.1 Discrete distributions 5
2.2.2 Continuous distributions on $\mathbb{R}$ 7
2.2.3 Continuous distributions on $\mathbb{R}^+$ 9
2.2.4 Continuous distributions on $[0, 1]$ 13
2.2.5 The multivariate Gaussian (normal) distribution 13
2.2.6 Linear Gaussian systems 19
2.2.7 A general calculus for linear Gaussian systems 21
2.2.8 Some other multivariate continuous distributions 25
2.3 The exponential family 29
2.3.1 Definition 30
2.3.2 Examples 30
2.3.3 Log partition function is cumulant generating function 35
2.3.4 Canonical (natural) vs mean (moment) parameters 36
2.3.5 MLE for the exponential family 37
2.3.6 Exponential dispersion family 38
2.3.7 Maximum entropy derivation of the exponential family 38
2.4 Fisher information matrix (FIM) 39
2.4.1 Definition 39
2.4.2 Equivalence between the FIM and the Hessian of the NLL 39
2.4.3 Examples 41
2.4.4 Approximating KL divergence using FIM 42
2.4.5 Fisher information matrix for exponential family 42
2.5 Transformations of random variables 44
2.5.1 Invertible transformations (bijections) 44
2.5.2 Monte Carlo approximation 44
2.5.3 Probability integral transform 45
2.6 Markov chains 46
2.6.1 Parameterization 47
2.6.2 Application: Language modeling 49
2.6.3 Parameter estimation 49

2.6.4 Stationary distribution of a Markov chain 51
2.7 Divergence measures between probability distributions 55
2.7.1 f-divergence 55
2.7.2 Integral probability metrics 57
2.7.3 Maximum mean discrepancy (MMD) 58
2.7.4 Total variation distance 60
2.7.5 Density ratio estimation using binary classifiers 61

3 Bayesian 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 65
3.1.5 Why not just use MAP estimation? 65
3.2 Conjugate priors for simple models 70
3.2.1 The binomial model 70
3.2.2 The multinomial model 71
3.2.3 The univariate Gaussian model 73
3.3 Conjugate priors for the multivariate Gaussian 78
3.3.1 Posterior of $\boldsymbol{\mu}$ given $\boldsymbol{\Sigma}$ 78
3.3.2 Posterior of $\boldsymbol{\Sigma}$ given $\boldsymbol{\mu}$ 78
3.3.3 Posterior of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ 80
3.4 Conjugate priors for the exponential family 84
3.5 Beyond conjugate priors 86
3.5.1 Robust (heavy-tailed) priors 87
3.5.2 Priors for variance parameters 87
3.6 Noninformative priors 88
3.6.1 Maximum entropy priors 89
3.6.2 Jeffreys priors 90
3.6.3 Invariant priors 93
3.6.4 Reference priors 94
3.7 Hierarchical priors 94
3.7.1 A hierarchical binomial model 95
3.7.2 A hierarchical Gaussian model 97
3.7.3 Hierarchical conditional models 100
3.8 Empirical Bayes 101
3.8.1 EB for the hierarchical binomial model 101
3.8.2 EB for the hierarchical Gaussian model 102
3.8.3 EB for Markov models (n-gram smoothing) 103
3.9 Model selection and evaluation 105
3.9.1 Bayesian model selection 105
3.9.2 Estimating the marginal likelihood 106
3.9.3 Connection between cross validation and marginal likelihood 107
3.9.4 Bayesian leave-one-out (LOO) estimate 108
3.9.5 Information criteria 109
3.9.6 Posterior predictive checks 112
3.9.7 Bayesian p-values 113

4 Probabilistic graphical models 117

4.1 Introduction 117
4.2 Directed graphical models (Bayes nets) 117
4.2.1 Representing the joint distribution 117
4.2.2 Examples 118
4.2.3 Gaussian Bayes nets 122
4.2.4 Conditional independence properties 123
4.2.5 Generation (sampling) 129

4.2.6 Inference 129
4.2.7 Learning 130
4.2.8 Plate notation 136
4.3 Undirected graphical models (Markov random fields) 139
4.3.1 Representing the joint distribution 139
4.3.2 Fully visible MRFs (Ising, Potts, Hopfield, etc) 141
4.3.3 MRFs with latent variables (Boltzmann machines, etc) 145
4.3.4 Maximum entropy models 149
4.3.5 Gaussian MRFs 151
4.3.6 Conditional independence properties 153
4.3.7 Generation (sampling) 155
4.3.8 Inference 155
4.3.9 Learning 156
4.4 Conditional random fields (CRFs) 160
4.4.1 1d CRFs 161
4.4.2 2d CRFs 164
4.4.3 Parameter estimation 167
4.4.4 Other approaches to structured prediction 168
4.5 Comparing directed and undirected PGMs 168
4.5.1 CI properties 168
4.5.2 Converting between a directed and undirected model 169
4.5.3 Conditional directed vs undirected PGMs and the label bias problem 171
4.5.4 Combining directed and undirected graphs 172
4.5.5 Comparing directed and undirected Gaussian PGMs 174
4.6 PGM extensions 175
4.6.1 Factor graphs 176
4.6.2 Probabilistic circuits 179
4.6.3 Directed relational PGMs 179
4.6.4 Undirected relational PGMs 182
4.6.5 Open-universe probability models 184
4.6.6 Programs as probability models 185
4.7 Structural causal models 185
4.7.1 Example: causal impact of education on wealth 186
4.7.2 Structural equation models 187
4.7.3 Do operator and augmented DAGs 187
4.7.4 Counterfactuals 189

5 Information theory 193

5.1 KL divergence 193
5.1.1 Desiderata 193
5.1.2 The KL divergence uniquely satisfies the desiderata 195
5.1.3 Thinking about KL 198
5.1.4 Properties of KL 201
5.1.5 KL divergence and MLE 204
5.1.6 KL divergence and Bayesian Inference 204
5.1.7 KL divergence and Exponential Families 206
5.1.8 Bregman divergence 206
5.2 Entropy 208
5.2.1 Definition 208
5.2.2 Differential entropy for continuous random variables 208
5.2.3 Typical sets 210
5.2.4 Cross entropy and perplexity 211
5.3 Mutual information 212
5.3.1 Definition 212
5.3.2 Interpretation 212
5.3.3 Data processing inequality 213

5.3.4 Sufficient Statistics 214
5.3.5 Multivariate mutual information 214
5.3.6 Variational bounds on mutual information 217
5.3.7 Relevance networks 219
5.4 Data compression (source coding) 220
5.4.1 Lossless compression 221
5.4.2 Lossy compression and the rate-distortion tradeoff 221
5.4.3 Bits back coding 223
5.5 Error-correcting codes (channel coding) 224
5.6 The information bottleneck 226
5.6.1 Vanilla IB 226
5.6.2 Variational IB 227
5.6.3 Conditional entropy bottleneck 228

6 Optimization 231

6.1 Introduction 231
6.2 Automatic differentiation 231
6.2.1 Differentiation in functional form 231
6.2.2 Differentiating chains, circuits, and programs 235
6.3 Stochastic gradient descent 241
6.4 Natural gradient descent 241
6.4.1 Defining the natural gradient 242
6.4.2 Interpretations of NGD 243
6.4.3 Benefits of NGD 244
6.4.4 Approximating the natural gradient 244
6.4.5 Natural gradients for the exponential family 246
6.5 Gradients of stochastic functions 248
6.5.1 Minibatch approximation to finite-sum objectives 249
6.5.2 Optimizing parameters of a distribution 249
6.5.3 Score function estimator (likelihood ratio trick) 250
6.5.4 Reparameterization trick 251
6.5.5 The delta method 253
6.5.6 Gumbel softmax trick 253
6.5.7 Stochastic computation graphs 254
6.5.8 Straight-through estimator 254
6.6 Bound optimization (MM) algorithms 255
6.6.1 The general algorithm 255
6.6.2 Example: logistic regression 256
6.6.3 The EM algorithm 258
6.6.4 Example: EM for an MVN with missing data 260
6.6.5 Example: robust linear regression using Student-t likelihood 262
6.6.6 Extensions to EM 263
6.7 The Bayesian learning rule 265
6.7.1 Deriving inference algorithms from BLR 266
6.7.2 Deriving optimization algorithms from BLR 268
6.7.3 Variational optimization 272
6.8 Bayesian optimization 272
6.8.1 Sequential model-based optimization 273
6.8.2 Surrogate functions 274
6.8.3 Acquisition functions 275
6.8.4 Other issues 278
6.9 Derivative free optimization 279
6.9.1 Local search 279
6.9.2 Simulated annealing 282
6.9.3 Evolutionary algorithms 282
6.9.4 Estimation of distribution (EDA) algorithms 285

6.9.5 Cross-entropy method 287
6.9.6 Evolutionary strategies 287
6.10 Optimal Transport 288
6.10.1 Warm-up: Matching optimally two families of points 288
6.10.2 From Optimal Matchings to Kantorovich and Monge formulations 289
6.10.3 Solving optimal transport 292
6.11 Submodular optimization 296
6.11.1 Intuition, Examples, and Background 297
6.11.2 Submodular Basic Definitions 299
6.11.3 Example Submodular Functions 300
6.11.4 Submodular Optimization 303
6.11.5 Applications of Submodularity in Machine Learning and AI 307
6.11.6 Sketching, CoreSets, Distillation, and Data Subset & Feature Selection 307
6.11.7 Combinatorial Information Functions 311
6.11.8 Clustering, Data Partitioning, and Parallel Machine Learning 312
6.11.9 Active and Semi-Supervised Learning 313
6.11.10 Probabilistic Modeling 314
6.11.11 Structured Norms and Loss Functions 315
6.11.12 Conclusions 316

II Inference 317

7 Inference algorithms: an overview 319
7.1 Introduction 319
7.2 Common inference patterns 319
7.2.1 Global latents 320
7.2.2 Local latents 320
7.2.3 Global and local latents 321
7.3 Exact inference algorithms 321
7.4 Approximate inference algorithms 322
7.4.1 MAP estimation 322
7.4.2 Grid approximation 322
7.4.3 Laplace (quadratic) approximation 323
7.4.4 Variational inference 324
7.4.5 Markov Chain Monte Carlo (MCMC) 326
7.4.6 Sequential Monte Carlo 327
7.4.7 Challenging posteriors 328
7.5 Evaluating approximate inference algorithms 328

8 Inference for state-space models 331
8.1 Introduction 331
8.2 Inference based on the HMM filter 331
8.2.1 Example: casino HMM 332
8.2.2 Forwards filtering 333
8.2.3 Backwards smoothing 335
8.2.4 The forwards-backwards algorithm 337
8.2.5 Numerically stable implementation 338
8.2.6 Time and space complexity 339
8.2.7 The Viterbi algorithm 340
8.2.8 Forwards filtering, backwards sampling 343
8.3 Inference based on the Kalman filter 343
8.3.1 Examples 344
8.3.2 The Kalman filter 346
8.3.3 The Kalman (RTS) smoother 350
8.3.4 Information form filtering and smoothing 352
8.4 Inference for non-linear and/or non-Gaussian SSMs 355

8.4.1 Inference based on discretization 355
8.4.2 Inference based on Gaussian approximations 356
8.5 Inference based on local linearization 357
8.5.1 Taylor series expansion 357
8.5.2 The extended Kalman filter (EKF) 358
8.5.3 Iterated EKF 359
8.5.4 The extended Kalman smoother (EKS) 359
8.5.5 Examples 359
8.6 Other variants of the Kalman filter 360
8.6.1 Ensemble Kalman filter 360
8.6.2 Robust Kalman filters 362
8.6.3 Dual EKF 362
8.7 Inference based on the unscented transform 362
8.7.1 The unscented transform 362
8.7.2 The unscented Kalman filter (UKF) 365
8.7.3 The unscented Kalman smoother (UKS) 365
8.7.4 Examples 365
8.8 Inference based on moment matching 366
8.8.1 Gaussian moment matching 366
8.8.2 The general Gaussian filter 367
8.8.3 Implementation using numerical integration 367
8.9 Inference based on statistical linearization 367
8.9.1 Statistical linear regression 368
8.9.2 Prior linearization filter 369
8.9.3 Iterated posterior linearization filter 369
8.9.4 Iterated posterior linearization smoother 370
8.9.5 Beyond additive Gaussian noise 371
8.10 Assumed density filtering 373
8.10.1 Connection with Gaussian filtering 374
8.10.2 ADF for SLDS (Gaussian sum filter) 375
8.10.3 ADF for online logistic regression 376
8.10.4 ADF for online DNNs 380
8.11 Other inference methods for SSMs 380
8.11.1 Expectation propagation 380
8.11.2 Variational inference 380
8.11.3 MCMC 381
8.11.4 Particle filtering 381

9 Inference for graphical models 383

9.1 Introduction 383
9.2 Belief propagation on trees 383
9.2.1 Directed vs undirected trees 384
9.2.2 Sum-product algorithm 385
9.2.3 Max-product algorithm 387
9.3 Loopy belief propagation 389
9.3.1 Loopy BP for pairwise undirected graphs 389
9.3.2 Loopy BP for factor graphs 390
9.3.3 Gaussian belief propagation 391
9.3.4 Convergence 392
9.3.5 Accuracy 395
9.3.6 Generalized belief propagation 395
9.3.7 Convex BP 395
9.3.8 Application: error correcting codes 396
9.3.9 Application: Affinity propagation 397
9.3.10 Emulating BP with graph neural nets 398
9.4 The variable elimination (VE) algorithm 399

9.4.1 Derivation of the algorithm 399
9.4.2 Computational complexity of VE 401
9.4.3 Picking a good elimination order 403
9.4.4 Computational complexity of exact inference 403
9.4.5 Drawbacks of VE 404
9.5 The junction tree algorithm (JTA) 405
9.6 Inference as optimization 406
9.6.1 Inference as backpropagation 406
9.6.2 Perturb and MAP 407

10 Variational inference 409
10.1 Introduction 409
10.1.1 Variational free energy 409
10.1.2 Evidence lower bound (ELBO) 410
10.2 Mean field VI 411
10.2.1 Coordinate ascent variational inference (CAVI) 411
10.2.2 Example: CAVI for the Ising model 413
10.2.3 Variational Bayes 415
10.2.4 Example: VB for a univariate Gaussian 416
10.2.5 Variational Bayes EM 419
10.2.6 Example: VBEM for a GMM 420
10.2.7 Variational message passing (VMP) 426
10.2.8 Autoconj 427
10.3 Fixed-form VI 427
10.3.1 Stochastic variational inference 427
10.3.2 Black-box variational inference 428
10.3.3 Reparameterization VI 430
10.3.4 Gaussian VI 433
10.3.5 Automatic differentiation VI 434
10.3.6 Amortized inference 437
10.4 More accurate variational posteriors 438
10.4.1 Structured mean field 438
10.4.2 Hierarchical (auxiliary variable) posteriors 438
10.4.3 Normalizing flow posteriors 439
10.4.4 Implicit posteriors 439
10.4.5 Combining VI with MCMC inference 439
10.5 Tighter bounds 440
10.5.1 Multi-sample ELBO (IWAE bound) 440
10.5.2 The thermodynamic variational objective (TVO) 441
10.5.3 Minimizing the evidence upper bound 441
10.6 Wake-sleep algorithm 441
10.6.1 Wake phase 442
10.6.2 Sleep phase 443
10.6.3 Daydream phase 443
10.6.4 Summary of algorithm 444
10.7 Expectation propagation (EP) 445
10.7.1 Algorithm 445
10.7.2 Example 446
10.7.3 EP as generalized ADF 447
10.7.4 Optimization issues 447
10.7.5 Power EP and α-divergence 448
10.7.6 Stochastic EP 448

11 Monte Carlo inference 451
11.1 Introduction 451
11.2 Monte Carlo integration 451
11.2.1 Example: estimating π by Monte Carlo integration 452

11.2.2 Accuracy of Monte Carlo integration 452
11.3 Generating random samples from simple distributions 454
11.3.1 Sampling using the inverse cdf 454
11.3.2 Sampling from a Gaussian (Box-Muller method) 455
11.4 Rejection sampling 455
11.4.1 Basic idea 456
11.4.2 Example 457
11.4.3 Adaptive rejection sampling 457
11.4.4 Rejection sampling in high dimensions 458
11.5 Importance sampling 458
11.5.1 Direct importance sampling 459
11.5.2 Self-normalized importance sampling 459
11.5.3 Choosing the proposal 460
11.5.4 Annealed importance sampling (AIS) 460
11.6 Controlling Monte Carlo variance 462
11.6.1 Common random numbers 462
11.6.2 Rao-Blackwellisation 462
11.6.3 Control variates 463
11.6.4 Antithetic sampling 464
11.6.5 Quasi Monte Carlo (QMC) 465

12 Markov Chain Monte Carlo inference 467

12.1 Introduction 467
12.2 Metropolis Hastings algorithm 468
12.2.1 Basic idea 468
12.2.2 Why MH works 469
12.2.3 Proposal distributions 470
12.2.4 Initialization 472
12.3 Gibbs sampling 473
12.3.1 Basic idea 473
12.3.2 Gibbs sampling is a special case of MH 473
12.3.3 Example: Gibbs sampling for Ising models 474
12.3.4 Example: Gibbs sampling for Potts models 476
12.3.5 Example: Gibbs sampling for GMMs 476
12.3.6 Metropolis within Gibbs 478
12.3.7 Blocked Gibbs sampling 478
12.3.8 Collapsed Gibbs sampling 479
12.4 Auxiliary variable MCMC 481
12.4.1 Slice sampling 482
12.4.2 Swendsen Wang 483
12.5 Hamiltonian Monte Carlo (HMC) 484
12.5.1 Hamiltonian mechanics 485
12.5.2 Integrating Hamilton's equations 485
12.5.3 The HMC algorithm 487
12.5.4 Tuning HMC 488
12.5.5 Riemann Manifold HMC 489
12.5.6 Langevin Monte Carlo (MALA) 489
12.5.7 Connection between SGD and Langevin sampling 490
12.5.8 Applying HMC to constrained parameters 491
12.5.9 Speeding up HMC 492
12.6 MCMC convergence 492
12.6.1 Mixing rates of Markov chains 492
12.6.2 Practical convergence diagnostics 494
12.6.3 Effective sample size 497
12.6.4 Improving speed of convergence 499
12.6.5 Non-centered parameterizations and Neal's funnel 500

12.7 Stochastic gradient MCMC 501
12.7.1 Stochastic Gradient Langevin Dynamics (SGLD) 501
12.7.2 Preconditioning 502
12.7.3 Reducing the variance of the gradient estimate 502
12.7.4 SG-HMC 503
12.7.5 Underdamped Langevin Dynamics 503
12.8 Reversible jump (trans-dimensional) MCMC 504
12.8.1 Basic idea 504
12.8.2 Example 505
12.8.3 Discussion 507
12.9 Annealing methods 507
12.9.1 Simulated annealing 508
12.9.2 Parallel tempering 510

13 Sequential Monte Carlo inference 511

13.1 Introduction 511
13.1.1 Problem statement 511
13.1.2 Particle filtering for state-space models 511
13.1.3 SMC samplers for static parameter estimation 513
13.2 Particle filtering 513
13.2.1 Importance sampling 513
13.2.2 Sequential importance sampling 514
13.2.3 Sequential importance sampling with resampling 516
13.2.4 Resampling methods 519
13.2.5 Adaptive resampling 520
13.3 Proposal distributions 521
13.3.1 Locally optimal proposal 522
13.3.2 Proposals based on the extended and unscented Kalman filter 522
13.3.3 Proposals based on the Laplace approximation 523
13.3.4 Proposals based on SMC (nested SMC) 524
13.4 Rao-Blackwellised particle filtering (RBPF) 524
13.4.1 Mixture of Kalman filters 525
13.4.2 Example: tracking a maneuvering object 526
13.4.3 Example: FastSLAM 527
13.5 Extensions of the particle filter 531
13.6 SMC samplers 531
13.6.1 Ingredients of an SMC sampler 531
13.6.2 Likelihood tempering (geometric path) 532
13.6.3 Data tempering 535
13.6.4 Sampling rare events and extrema 536
13.6.5 SMC-ABC and likelihood-free inference 536
13.6.6 SMC^2 537
13.6.7 Variational filtering SMC 537
13.6.8 Variational smoothing SMC 538

III Prediction 541

14 Predictive models: an overview 543

14.1 Introduction 543
14.1.1 Types of model 543
14.1.2 Model fitting using ERM, MLE and MAP 544
14.1.3 Model fitting using Bayes, VI and generalized Bayes 545
14.2 Evaluating predictive models 546
14.2.1 Proper scoring rules 546
14.2.2 Calibration 546

14.2.3 Beyond evaluating marginal probabilities 550
14.3 Conformal prediction 553
14.3.1 Conformalizing classification 554
14.3.2 Conformalizing regression 555

15 Generalized linear models 557
15.1 Introduction 557
15.1.1 Examples 557
15.1.2 GLMs with non-canonical link functions 560
15.1.3 Maximum likelihood estimation 560
15.1.4 Bayesian inference 561
15.2 Linear regression 562
15.2.1 Conjugate priors 562
15.2.2 Uninformative priors 564
15.2.3 Informative priors 566
15.2.4 Spike and slab prior 568
15.2.5 Laplace prior (Bayesian lasso) 569
15.2.6 Horseshoe prior 570
15.2.7 Automatic relevancy determination 571
15.2.8 Multivariate linear regression 573
15.3 Logistic regression 575
15.3.1 Binary logistic regression 575
15.3.2 Multinomial logistic regression 576
15.3.3 Priors 577
15.3.4 Posteriors 578
15.3.5 Laplace approximation 578
15.3.6 MCMC inference 580
15.3.7 Variational inference 582
15.3.8 Assumed density filtering 582
15.4 Probit regression 582
15.4.1 Latent variable interpretation 582
15.4.2 Maximum likelihood estimation 583
15.4.3 Bayesian inference 585
15.4.4 Ordinal probit regression 585
15.4.5 Multinomial probit models 586
15.5 Multi-level (hierarchical) GLMs 586
15.5.1 Generalized linear mixed models (GLMMs) 587
15.5.2 Model fitting 587
15.5.3 Example: radon regression 588

16 Deep neural networks 591
16.1 Introduction 591
16.2 Building blocks of differentiable circuits 591
16.2.1 Linear layers 592
16.2.2 Non-linearities 592
16.2.3 Convolutional layers 593
16.2.4 Residual (skip) connections 594
16.2.5 Normalization layers 595
16.2.6 Dropout layers 595
16.2.7 Attention layers 596
16.2.8 Recurrent layers 598
16.2.9 Multiplicative layers 599
16.2.10 Implicit layers 600
16.3 Canonical examples of neural networks 600
16.3.1 Multi-layer perceptrons (MLP) 600
16.3.2 Convolutional neural networks (CNN) 601
16.3.3 Autoencoders 602

16.3.4 Recurrent neural networks (RNN) 604
16.3.5 Transformers 604
16.3.6 Graph neural networks (GNNs) 605

17 Bayesian neural networks 607

17.1 Introduction 607
17.2 Priors for BNNs 607
17.2.1 Gaussian priors 608
17.2.2 Sparsity-promoting priors 610
17.2.3 Learning the prior 610
17.2.4 Priors in function space 610
17.2.5 Architectural priors 610
17.3 Posteriors for BNNs 611
17.3.1 MC dropout 611
17.3.2 Laplace approximation 612
17.3.3 Variational inference 613
17.3.4 Expectation propagation 614
17.3.5 Last layer methods and SNGP 614
17.3.6 SNGP 614
17.3.7 MCMC methods 615
17.3.8 Methods based on the SGD trajectory 616
17.3.9 Deep ensembles 617
17.3.10 Approximating the posterior predictive distibution 621
17.3.11 Tempered and cold posteriors 622
17.4 Generalization in Bayesian deep learning 622
17.4.1 Sharp vs flat minima 623
17.4.2 Mode connectivity and the loss landscape 624
17.4.3 Effective dimensionality of a model 624
17.4.4 The hypothesis space of DNNs 625
17.4.5 PAC-Bayes 626
17.4.6 Out-of-Distribution generalization for BNNs 627
17.4.7 Model Selection for BNNs 629
17.5 Online inference 630
17.5.1 Sequential Laplace for DNNs 630
17.5.2 Extended Kalman Filtering for DNNs 630
17.5.3 Assumed Density Filtering for DNNs 633
17.5.4 Online variational inference for DNNs 634
17.6 Hierarchical Bayesian neural networks 635
17.6.1 Example: multi-moons classification 635

18 Gaussian processes 639

18.1 Introduction 639
18.1.1 GPs: What and why? 639
18.2 Mercer kernels 641
18.2.1 Some popular Mercer kernels 642
18.2.2 Mercer's theorem 649
18.2.3 Kernels from Spectral Densities 650
18.3 GPs with Gaussian likelihoods 651
18.3.1 Predictions using noise-free observations 651
18.3.2 Predictions using noisy observations 652
18.3.3 Weight space vs function space 653
18.3.4 Semi-parametric GPs 654
18.3.5 Marginal likelihood 655
18.3.6 Computational and numerical issues 655
18.3.7 Kernel ridge regression 656
18.4 GPs with non-Gaussian likelihoods 659
18.4.1 Binary classification 660

18.4.2 Multi-class classification 660
18.4.3 GPs for Poisson regression (Cox process) 661
18.4.4 Other likelihoods 662
18.5 Scaling GP inference to large datasets 662
18.5.1 Subset of data 663
18.5.2 Nyström approximation 664
18.5.3 Inducing point methods 665
18.5.4 Sparse variational methods 668
18.5.5 Exploiting parallelization and structure via kernel matrix multiplies 671
18.5.6 Converting a GP to a SSM 674
18.6 Learning the kernel 674
18.6.1 Empirical Bayes for the kernel parameters 674
18.6.2 Bayesian inference for the kernel parameters 677
18.6.3 Multiple kernel learning for additive kernels 678
18.6.4 Automatic search for compositional kernels 679
18.6.5 Spectral mixture kernel learning 682
18.6.6 Deep kernel learning 683
18.7 GPs and DNNs 685
18.7.1 Kernels derived from infinitely wide DNNs (NN-GP) 685
18.7.2 Neural tangent kernel (NTK) 687
18.7.3 Deep GPs 688
18.8 Gaussian processes for timeseries forecasting 689
18.8.1 Example: Mauna Loa 689

19 Beyond the iid assumption 691

19.1 Introduction 691
19.2 Distribution shift 691
19.2.1 Motivating examples 691
19.2.2 A causal view of distribution shift 693
19.2.3 The four main types of distribution shift 694
19.2.4 Selection bias 696
19.3 Detecting distribution shifts 696
19.3.1 Detecting shifts using two-sample testing 697
19.3.2 Detecting single out-of-distribution (OOD) inputs 697
19.3.3 Selective prediction 700
19.3.4 Open world recognition 701
19.4 Robustness to distribution shifts 701
19.4.1 Data augmentation 702
19.4.2 Distributionally robust optimization 702
19.5 Adapting to distribution shifts 702
19.5.1 Supervised adaptation using transfer learning 702
19.5.2 Weighted ERM for covariate shift 704
19.5.3 Unsupervised domain adaptation for covariate shift 705
19.5.4 Unsupervised techniques for label shift 706
19.5.5 Test-time adaptation 706
19.6 Learning from multiple distributions 707
19.6.1 Multi-task learning 707
19.6.2 Domain generalization 708
19.6.3 Invariant risk minimization 710
19.6.4 Meta-learning 711
19.7 Continual learning 714
19.7.1 Domain drift 714
19.7.2 Concept drift 715
19.7.3 Task incremental learning 716
19.7.4 Catastrophic forgetting 717
19.7.5 Online learning 719

19.8 Adversarial examples 720
19.8.1 Whitebox (gradient-based) attacks 722
19.8.2 Blackbox (gradient-free) attacks 723
19.8.3 Real world adversarial attacks 724
19.8.4 Defenses based on robust optimization 724
19.8.5 Why models have adversarial examples 725

IV Generation 727

20 Generative models: an overview 729

20.1 Introduction 729
20.2 Types of generative model 729
20.3 Goals of generative modeling 731
20.3.1 Generating data 731
20.3.2 Density estimation 733
20.3.3 Imputation 734
20.3.4 Structure discovery 734
20.3.5 Latent space interpolation 735
20.3.6 Latent space arithmetic 736
20.3.7 Generative design 737
20.3.8 Model-based reinforcement learning 737
20.3.9 Representation learning 737
20.3.10 Data compression 737
20.4 Evaluating generative models 738
20.4.1 Likelihood-based evaluation 738
20.4.2 Distances and divergences in feature space 740
20.4.3 Precision and recall metrics 741
20.4.4 Statistical tests 742
20.4.5 Challenges with using pretrained classifiers 742
20.4.6 Using model samples to train classifiers 742
20.4.7 Assessing overfitting 742
20.4.8 Human evaluation 743

21 Variational autoencoders 745

21.1 Introduction 745
21.2 VAE basics 745
21.2.1 Modeling assumptions 746
21.2.2 Evidence lower bound (ELBO) 747
21.2.3 Evaluating the ELBO 748
21.2.4 Optimizing the ELBO 748
21.2.5 Using the reparameterization trick to compute ELBO gradients 749
21.2.6 Comparison of VAEs and autoencoders 752
21.2.7 VAEs optimize in an augmented space 753
21.3 VAE generalizations 755
21.3.1 β-VAE 756
21.3.2 InfoVAE 757
21.3.3 Multi-modal VAEs 759
21.3.4 VAEs with missing data 761
21.3.5 Semi-supervised VAEs 763
21.3.6 VAEs with sequential encoders/decoders 764
21.4 Avoiding posterior collapse 767
21.4.1 KL annealing 768
21.4.2 Lower bounding the rate 768
21.4.3 Free bits 769
21.4.4 Adding skip connections 769

21.4.5 Improved variational inference 769
21.4.6 Alternative objectives 770
21.5 VAEs with hierarchical structure 771
21.5.1 Bottom-up vs top-down inference 771
21.5.2 Example: Very deep VAE 772
21.5.3 Connection with autoregressive models 773
21.5.4 Variational pruning 774
21.5.5 Other optimization difficulties 775
21.6 Vector quantization VAE 775
21.6.1 Autoencoder with binary code 776
21.6.2 VQ-VAE model 776
21.6.3 Learning the prior 778
21.6.4 Hierarchical extension (VQ-VAE-2) 778
21.6.5 Discrete VAE 779
21.6.6 VQ-GAN 780

22 Auto-regressive models 781
22.1 Introduction 781
22.2 Neural autoregressive density estimators (NADE) 782
22.3 Causal CNNs 782
22.3.1 1d causal CNN (Convolutional Markov models) 783
22.3.2 2d causal CNN (PixelCNN) 783
22.4 Transformer decoders 784
22.4.1 Text generation (GPT) 785
22.4.2 Music generation 785
22.4.3 Text-to-image generation (DALL-E) 786

23 Normalizing Flows 789
23.1 Introduction 789
23.1.1 Preliminaries 789
23.1.2 How to train a flow model 791
23.2 Constructing Flows 792
23.2.1 Affine flows 792
23.2.2 Elementwise flows 792
23.2.3 Coupling flows 795
23.2.4 Autoregressive flows 796
23.2.5 Residual flows 802
23.2.6 Continuous-time flows 804
23.3 Applications 806
23.3.1 Density estimation 806
23.3.2 Generative Modeling 806
23.3.3 Inference 807

24 Energy-based models 809
24.1 Introduction 809
24.1.1 Example: Products of experts (PoE) 809
24.1.2 Computational difficulties 810
24.2 Maximum Likelihood Training 810
24.2.1 Gradient-based MCMC methods 812
24.2.2 Contrastive divergence 812
24.3 Score Matching (SM) 815
24.3.1 Basic score matching 816
24.3.2 Denoising Score Matching (DSM) 816
24.3.3 Sliced Score Matching (SSM) 818
24.3.4 Connection to Contrastive Divergence 819
24.3.5 Score-Based Generative Models 820
24.4 Noise Contrastive Estimation 823

24.4.1 Connection to Score Matching 824
24.5 Other Methods 825
24.5.1 Minimizing Differences/Derivatives of KL Divergences 825
24.5.2 Minimizing the Stein Discrepancy 826
24.5.3 Adversarial Training 826

25 Diffusion models 829

25.1 Variational diffusion models 829
25.1.1 Encoder 829
25.1.2 Decoder 831
25.1.3 Model fitting 833
25.1.4 Connection to DDPM 836
25.1.5 2d Example 837
25.1.6 Image generation 837
25.2 Conditional diffusion models 838
25.2.1 Classifier guidance 838
25.2.2 Classifier-free guidance 839
25.2.3 Conditional image generation 840
25.2.4 Other forms of conditional generation 840
25.3 Speeding up the generation process 840

26 Generative adversarial networks 843

26.1 Introduction 843
26.2 Learning by Comparison 844
26.2.1 Guiding principles 845
26.2.2 Density ratio estimation using binary classifiers 846
26.2.3 Bounds on f-divergences 849
26.2.4 Integral probability metrics 850
26.2.5 Moment matching 852
26.2.6 On density ratios and differences 853
26.3 Generative Adversarial Networks 854
26.3.1 From learning principles to loss functions 854
26.3.2 Gradient Descent 856
26.3.3 Challenges with GAN training 857
26.3.4 Improving GAN optimization 858
26.3.5 Convergence of GAN training 859
26.4 Conditional GANs 862
26.5 Inference with GANs 863
26.6 Neural architectures in GANs 864
26.6.1 The importance of discriminator architectures 865
26.6.2 Architectural inductive biases 865
26.6.3 Attention in GANs 865
26.6.4 Progressive generation 866
26.6.5 Regularization 868
26.6.6 Scaling up GAN models 868
26.7 Applications 869
26.7.1 GANs for image generation 869
26.7.2 Video generation 871
26.7.3 Audio generation 872
26.7.4 Text generation 873
26.7.5 Imitation Learning 874
26.7.6 Domain Adaptation 874
26.7.7 Design, Art and Creativity 875

V Discovery 877

27 Discovery methods: an overview 879
27.1 Introduction 879
27.2 Overview of Part V 880

28 Latent factor models 881
28.1 Introduction 881
28.2 Mixture models 881
28.2.1 Gaussian mixture models (GMMs) 882
28.2.2 Bernoulli mixture models 884
28.2.3 Gaussian scale mixtures (GSMs) 884
28.2.4 Using GMMs as a prior for inverse imaging problems 886
28.2.5 Using mixture models for classification problems 889
28.3 Factor analysis 891
28.3.1 Factor analysis: the basics 891
28.3.2 Probabilistic PCA 896
28.3.3 Mixture of factor analysers 898
28.3.4 Factor analysis models for paired data 904
28.3.5 Factor analysis with exponential family likelihoods 907
28.3.6 Factor analysis with DNN likelihoods (VAEs) 908
28.3.7 Factor analysis with GP likelihoods (GP-LVM) 909
28.4 LFMs with non-Gaussian priors 910
28.4.1 Non-negative matrix factorization (NMF) 911
28.4.2 Multinomial PCA 913
28.5 Topic models 914
28.5.1 Latent Dirichlet Allocation (LDA) 914
28.5.2 Correlated topic model 918
28.5.3 Dynamic topic model 918
28.5.4 LDA-HMM 919
28.6 Independent components analysis (ICA) 922
28.6.1 Noiseless ICA model 923
28.6.2 The need for non-Gaussian priors 924
28.6.3 Maximum likelihood estimation 925
28.6.4 Alternatives to MLE 926
28.6.5 Sparse coding 928
28.6.6 Nonlinear ICA 929

29 State-space model 931
29.1 Introduction 931
29.2 Hidden Markov models (HMMs) 932
29.2.1 Conditional independence properties 932
29.2.2 State transition model 932
29.2.3 Discrete likelihoods 933
29.2.4 Gaussian likelihoods 934
29.2.5 Autoregressive likelihoods 934
29.3 HMMs: Applications 936
29.3.1 Time series segmentation 936
29.3.2 Protein sequence alignment 938
29.3.3 Spelling correction 940
29.4 HMMs: parameter learning 942
29.4.1 The Baum-Welch (EM) algorithm 942
29.4.2 Parameter estimation using SGD 946
29.4.3 Parameter estimation using spectral methods 947
29.4.4 Bayesian HMMs 947
29.5 HMMs: Generalizations 949
29.5.1 Hidden semi-Markov model (HSMM) 949

29.5.2 Hierarchical HMMs 951
29.5.3 Factorial HMMs 953
29.5.4 Coupled HMMs 954
29.5.5 Dynamic Bayes nets (DBN) 955
29.5.6 Changepoint detection 955
29.6 Linear dynamical systems (LDS) 958
29.6.1 Conditional independence properties 958
29.6.2 Parameterization 958
29.7 LDS: Applications 959
29.7.1 Object tracking and state estimation 959
29.7.2 Online Bayesian linear regression (recursive least squares) 960
29.7.3 Adaptive filtering 962
29.7.4 Timeseries forecasting 962
29.8 LDS: parameter learning 962
29.8.1 EM for LDS 962
29.8.2 Subspace identification methods 964
29.8.3 Ensuring stability of the dynamical system 965
29.8.4 Bayesian LDS 965
29.9 Switching linear dynamical systems (SLDS) 966
29.9.1 Parameterization 966
29.9.2 Posterior inference 967
29.9.3 Application: Multi-target tracking 968
29.10 Nonlinear SSMs 971
29.10.1 Example: object tracking and state estimation 971
29.10.2 Posterior inference 972
29.11 Non-Gaussian SSMs 972
29.11.1 Example: Spike train modeling 972
29.11.2 Example: Stochastic volatility models 973
29.11.3 Posterior inference 974
29.12 Structural time series models 974
29.12.1 Basics 974
29.12.2 Details 975
29.12.3 Example: forecasting CO_2 levels from Mauna Loa 977
29.12.4 Example: forecasting (real-valued) electricity usage 978
29.12.5 Example: forecasting (integer valued) sales 979
29.12.6 Example: hierarchical SSM for electoral panel data 981
29.12.7 Causal impact of a time series intervention 981
29.12.8 Prophet 984
29.12.9 Neural forecasting methods 986
29.13 Deep SSMs 987
29.13.1 Deep Markov models 987
29.13.2 Recurrent SSM 989
29.13.3 Improving multi-step predictions 989
29.13.4 Variational RNNs 990

30 Graph learning 993

30.1 Introduction 993
30.2 Latent variable models for graphs 993
30.2.1 Stochastic block model 993
30.2.2 Mixed membership stochastic block model 995
30.2.3 Infinite relational model 997
30.3 Graphical model structure learning 999
30.3.1 Applications 999
30.3.2 Methods 1001

31 Non-parametric Bayesian models 1003

31.1 Introduction 1003

31.2 Dirichlet processes 1003
31.2.1 Definition of a DP 1004
31.2.2 Stick breaking construction of the DP 1006
31.2.3 The Chinese restaurant process (CRP) 1007
31.3 Dirichlet process mixture models 1008
31.3.1 Model definition 1010
31.3.2 Fitting using collapsed Gibbs sampling 1010
31.3.3 Fitting using variational Bayes 1012
31.3.4 Other fitting algorithms 1014
31.3.5 Choosing the hyper-parameters 1015
31.4 Generalizations of the Dirichlet process 1015
31.4.1 Pitman-Yor process 1016
31.4.2 Dependent random probability measures 1017
31.5 The Indian buffet process and the Beta process 1019
31.6 Small-variance asymptotics 1022
31.7 Completely random measures 1025
31.8 Lévy processes 1026
31.9 Point processes with repulsion and reinforcement 1028
31.9.1 Poisson process 1028
31.9.2 Renewal process 1029
31.9.3 Hawkes process 1030
31.9.4 Gibbs point process 1032
31.9.5 Determinantal point process 1033

32 Representation learning 1037

32.1 Introduction 1037
32.2 Evaluating and comparing learned representations 1037
32.2.1 Downstream performance 1038
32.2.2 Representational similarity 1040
32.3 Approaches for learning representations 1044
32.3.1 Supervised Representation Learning and Transfer 1045
32.3.2 Generative Representation Learning 1047
32.3.3 Self-Supervised Representation Learning 1049
32.3.4 Multiview representation learning 1052
32.4 Theory of Representation Learning 1057
32.4.1 Identifiability 1057
32.4.2 Information maximization 1058

33 Interpretability 1061

33.1 Introduction 1061
33.1.1 The Role of Interpretability: Unknowns and Under-Specifications 1062
33.1.2 Terminology and Framework 1063
33.2 Methods for Interpretable Machine Learning 1067
33.2.1 Inherently Interpretable Models: The Model is its Explanation 1067
33.2.2 Semi-Inherently Interpretable Models: Example-Based Methods 1069
33.2.3 Post-hoc or Joint training: The Explanation gives a Partial View of the Model 1069
33.2.4 Transparency and Visualization 1073
33.3 Properties: The Abstraction Between Context and Method 1074
33.3.1 Properties of Explanations from Interpretable Machine Learning 1075
33.3.2 Properties of Explanations from Cognitive Science 1077
33.4 Evaluation of Interpretable Machine Learning Models 1078
33.4.1 Computational Evaluation: Does the Method have Desired Properties? 1079
33.4.2 User Study-based Evaluation: Does the Method Help a User Perform a Target Task? 1082
33.5 Discussion: How to Think about Interpretable Machine Learning 1086

VI Action 1091

34 Decision making under uncertainty 1093

34.1 Bayesian decision theory 1093
34.1.1 Basics 1093
34.1.2 Classification 1094
34.1.3 Regression 1094
34.1.4 Structured prediction 1095
34.1.5 Fairness 1096
34.2 Decision (influence) diagrams 1096
34.2.1 Example: oil wildcatter 1096
34.2.2 Information arcs 1097
34.2.3 Value of information 1098
34.2.4 Computing the optimal policy 1099
34.3 A/B testing 1099
34.3.1 A Bayesian approach 1100
34.3.2 Example 1103
34.4 Contextual bandits 1104
34.4.1 Types of bandit 1104
34.4.2 Applications 1106
34.4.3 Exploration-exploitation tradeoff 1106
34.4.4 The optimal solution 1106
34.4.5 Upper confidence bounds (UCB) 1108
34.4.6 Thompson sampling 1110
34.4.7 Regret 1111
34.5 Markov decision problems 1112
34.5.1 Basics 1113
34.5.2 Partially observed MDPs 1114
34.5.3 Episodes and returns 1115
34.5.4 Value functions 1115
34.5.5 Optimal value functions and policies 1116
34.6 Planning in an MDP 1117
34.6.1 Value iteration 1118
34.6.2 Policy iteration 1119
34.6.3 Linear programming 1120
34.7 Active learning 1121
34.7.1 Active learning scenarios 1121
34.7.2 Relationship to other forms of sequential decision making 1122
34.7.3 Acquisition strategies 1122
34.7.4 Batch active learning 1125

35 Reinforcement learning 1129

35.1 Introduction 1129
35.1.1 Overview of methods 1129
35.1.2 Value based methods 1131
35.1.3 Policy search methods 1131
35.1.4 Model-based RL 1131
35.1.5 Exploration-exploitation tradeoff 1132
35.2 Value-based RL 1134
35.2.1 Monte Carlo RL 1134
35.2.2 Temporal difference (TD) learning 1134
35.2.3 TD learning with eligibility traces 1135
35.2.4 SARSA: on-policy TD control 1136
35.2.5 Q-learning: off-policy TD control 1137
35.2.6 Deep Q-network (DQN) 1139
35.3 Policy-based RL 1140
35.3.1 The policy gradient theorem 1140

35.3.2 REINFORCE 1141
35.3.3 Actor-critic methods 1142
35.3.4 Bound optimization methods 1144
35.3.5 Deterministic policy gradient methods 1146
35.3.6 Gradient-free methods 1147
35.4 Model-based RL 1147
35.4.1 Model predictive control (MPC) 1147
35.4.2 Combining model-based and model-free 1149
35.4.3 MBRL using Gaussian processes 1149
35.4.4 MBRL using DNNs 1151
35.4.5 MBRL using latent-variable models 1151
35.4.6 Robustness to model errors 1154
35.5 Off-policy learning 1154
35.5.1 Basic techniques 1154
35.5.2 The curse of horizon 1158
35.5.3 The deadly triad 1159
35.6 Control as inference 1160
35.6.1 Maximum entropy reinforcement learning 1160
35.6.2 Other approaches 1163
35.6.3 Imitation learning 1164

36 Causality 1167

36.1 Introduction 1167
36.2 Causal Formalism 1168
36.2.1 Structural Causal Models 1169
36.2.2 Causal DAGs 1170
36.2.3 Identification 1172
36.2.4 Counterfactuals and the Causal Hierarchy 1174
36.3 Randomized Control Trials 1176
36.4 Confounder Adjustment 1177
36.4.1 Causal Estimand, Statistical Estimand, and Identification 1177
36.4.2 ATE Estimation with Observed Confounders 1179
36.4.3 Uncertainity Quantification 1184
36.4.4 Matching 1185
36.4.5 Practical Considerations and Procedures 1186
36.4.6 Summary and Practical Advice 1189
36.5 Instrumental Variable Strategies 1190
36.5.1 Additive Unobserved Confounding 1192
36.5.2 Instrument Monotonicity and Local Average Treatment Effect 1194
36.5.3 Two Stage Least Squares 1197
36.6 Difference in Differences 1198
36.6.1 Estimation 1201
36.7 Credibility Checks 1201
36.7.1 Placebo Checks 1202
36.7.2 Sensitivity Analysis to Unobserved Confounding 1202
36.8 The Do Calculus 1210
36.8.1 The three rules 1210
36.8.2 Revisiting Backdoor Adjustment 1211
36.8.3 Frontdoor Adjustment 1212
36.9 Further Reading 1214

Bibliography 1235