Contents

Preface **xxxiii**

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 $\mathbb{R}$ 6
2.2.3 Continuous distributions on $\mathbb{R}^+$ 7
2.2.4 Continuous distributions on $[0, 1]$ 8
2.2.5 The multivariate Gaussian (normal) distribution 9
2.2.6 Other multivariate distributions 11
2.3 The exponential family 12
2.3.1 Definition 12
2.3.2 Examples 13
2.3.3 Log partition function is cumulant generating function 17
2.3.4 Canonical (natural) vs mean (moment) parameters 19
2.3.5 MLE for the exponential family 20
2.3.6 Exponential dispersion family 21
2.3.7 Maximum entropy derivation of the exponential family 21
2.4 Fisher information matrix (FIM) 22
2.4.1 Definition 22
2.4.2 Equivalence between the FIM and the Hessian of the NLL 22
2.4.3 Examples 24
2.4.4 Approximating KL divergence using FIM 25
2.4.5 Fisher information matrix for exponential family 25
2.5 Transformations of random variables 26
2.5.1 Invertible transformations (bijections) 27

2.5.2 Monte Carlo approximation 27
2.5.3 Probability integral transform 28
2.6 Markov chains 28
2.6.1 Parameterization 29
2.6.2 Application: Language modeling 32
2.6.3 Parameter estimation 32
2.6.4 Stationary distribution of a Markov chain 34
2.7 Divergence measures between probability distributions 38
2.7.1 f-divergence 38
2.7.2 Integral probability metrics 40
2.7.3 Maximum mean discrepancy (MMD) 41
2.7.4 Total variation distance 43
2.7.5 Comparing distributions using binary classifiers 43

3 Bayesian statistics 47
3.1 Introduction 47
3.1.1 Frequentist statistics 47
3.1.2 Bayesian statistics 47
3.1.3 Arguments for the Bayesian approach 48
3.1.4 Arguments against the Bayesian approach 49
3.1.5 Why not just use MAP estimation? 49
3.2 Closed-form analysis using conjugate priors 54
3.2.1 The binomial model 54
3.2.2 The multinomial model 54
3.2.3 The univariate Gaussian model 55
3.2.4 The multivariate Gaussian model 60
3.2.5 Conjugate-exponential models 60
3.3 Beyond conjugate priors 63
3.3.1 Robust (heavy-tailed) priors 63
3.3.2 Priors for variance parameters 64
3.4 Noninformative priors 65
3.4.1 Maximum entropy priors 65
3.4.2 Jeffreys priors 66
3.4.3 Invariant priors 69
3.4.4 Reference priors 70
3.5 Hierarchical priors 71
3.5.1 A hierarchical binomial model 71
3.5.2 A hierarchical Gaussian model 73
3.6 Empirical Bayes 76
3.6.1 A hierarchical binomial model 77
3.6.2 A hierarchical Gaussian model 78
3.6.3 Hierarchical Bayes for n-gram smoothing 79
3.7 Model selection and evaluation 81
3.7.1 Bayesian model selection 81
3.7.2 Estimating the marginal likelihood 82

3.7.3 Connection between cross validation and marginal likelihood 83
3.7.4 Pareto-Smoothed Importance Sampling LOO estimate 84
3.7.5 Information criteria 85
3.7.6 Posterior predictive checks 87
3.7.7 Bayesian p-values 88
3.8 Bayesian decision theory 90
3.8.1 Basics 91
3.8.2 Example: COVID-19 91
3.8.3 One-shot decision problems 92
3.8.4 Multi-stage decision problems 93

4 Probabilistic graphical models 95

4.1 Introduction 95
4.2 Directed graphical models (Bayes nets) 95
4.2.1 Representing the joint distribution 95
4.2.2 Examples 96
4.2.3 Gaussian Bayes nets 100
4.2.4 Conditional independence properties 101
4.2.5 Generation (sampling) 106
4.2.6 Inference 106
4.2.7 Learning 108
4.2.8 Plate notation 113
4.3 Undirected graphical models (Markov random fields) 116
4.3.1 Representing the joint distribution 116
4.3.2 Fully visible MRFs (Ising, Potts, Hopfield, etc) 118
4.3.3 MRFs with latent variables (Boltzmann machines, etc) 124
4.3.4 Maximum entropy models 126
4.3.5 Gaussian MRFs 128
4.3.6 Conditional independence properties 130
4.3.7 Generation (sampling) 132
4.3.8 Inference 132
4.3.9 Learning 133
4.4 Conditional random fields (CRFs) 137
4.4.1 1d CRFs 137
4.4.2 2d CRFs 140
4.4.3 Parameter estimation 143
4.4.4 Other approaches to structured prediction 144
4.5 Comparing directed and undirected PGMs 144
4.5.1 CI properties 144
4.5.2 Converting between a directed and undirected model 146
4.5.3 Conditional directed vs undirected PGMs and the label bias problem 147
4.5.4 Combining directed and undirected graphs 148
4.5.5 Comparing directed and undirected Gaussian PGMs 150
4.6 PGM extensions 152
4.6.1 Factor graphs 152

4.6.2 Probabilistic circuits 155
4.6.3 Directed relational PGMs 155
4.6.4 Undirected relational PGMs 157
4.6.5 Open-universe probability models 160
4.6.6 Programs as probability models 162
4.7 Structural causal models 162
4.7.1 Example: causal impact of education on wealth 163
4.7.2 Structural equation models 164
4.7.3 Do operator and augmented DAGs 164
4.7.4 Estimating average treatment effect using path analysis 166
4.7.5 Counterfactuals 167

5 Information theory 171
5.1 KL divergence 171
5.1.1 Desiderata 172
5.1.2 The KL divergence uniquely satisfies the desiderata 173
5.1.3 Thinking about KL 176
5.1.4 Properties of KL 178
5.1.5 KL divergence and MLE 180
5.1.6 KL divergence and Bayesian Inference 181
5.1.7 KL divergence and Exponential Families 182
5.2 Entropy 183
5.2.1 Definition 183
5.2.2 Differential entropy for continuous random variables 184
5.2.3 Typical sets 185
5.2.4 Cross entropy and perplexity 187
5.3 Mutual information 188
5.3.1 Definition 188
5.3.2 Interpretation 188
5.3.3 Data processing inequality 189
5.3.4 Sufficient Statistics 190
5.3.5 Multivariate mutual information 190
5.3.6 Variational bounds on mutual information 193
5.4 Data compression (source coding) 196
5.4.1 Lossless compression 196
5.4.2 Lossy compression and the rate-distortion tradeoff 196
5.4.3 Bits back coding 199
5.5 Error-correcting codes (channel coding) 199
5.6 The information bottleneck 201
5.6.1 Vanilla IB 201
5.6.2 Variational IB 202
5.6.3 Conditional entropy bottleneck 203

6 Optimization 207
6.1 Introduction 207

6.2 Automatic differentiation 207
6.2.1 Differentiation in functional form 207
6.2.2 Differentiating chains, circuits, and programs 212
6.3 Stochastic gradient descent 217
6.4 Natural gradient descent 218
6.4.1 Defining the natural gradient 218
6.4.2 Interpretations of NGD 219
6.4.3 Benefits of NGD 220
6.4.4 Approximating the natural gradient 221
6.4.5 Natural gradients for the exponential family 222
6.5 Gradients of stochastic functions 224
6.5.1 Minibatch approximation to finite-sum objectives 225
6.5.2 Optimizing parameters of a distribution 225
6.5.3 Score function estimator (likelihood ratio trick) 226
6.5.4 Reparameterization trick 227
6.5.5 The delta method 229
6.5.6 Gumbel softmax trick 229
6.5.7 Stochastic computation graphs 230
6.5.8 Straight-through estimator 230
6.6 Bound optimization (MM) algorithms 231
6.6.1 The general algorithm 231
6.6.2 Example: logistic regression 232
6.6.3 The EM algorithm 234
6.6.4 Example: EM for an MVN with missing data 236
6.6.5 Example: robust linear regression using Student-*t* likelihood 238
6.6.6 Extensions to EM 239
6.7 The Bayesian learning rule 241
6.7.1 Deriving inference algorithms from BLR 242
6.7.2 Deriving optimization algorithms from BLR 244
6.7.3 Variational optimization 247
6.8 Bayesian optimization 248
6.8.1 Sequential model-based optimization 248
6.8.2 Surrogate functions 250
6.8.3 Acquisition functions 251
6.8.4 Other issues 254
6.9 Derivative free optimization 255
6.9.1 Local search 255
6.9.2 Evolutionary algorithms 258
6.9.3 Estimation of distribution (EDA) algorithms 260
6.9.4 Cross-entropy method 262
6.9.5 Evolutionary strategies 263
6.10 Optimal Transport 263
6.10.1 Warm-up: Matching optimally two families of points 263
6.10.2 From Optimal Matchings to Kantorovich and Monge formulations 265
6.10.3 Solving optimal transport 267

6.11 Submodular optimization 272
6.11.1 Intuition, Examples, and Background 273
6.11.2 Submodular Basic Definitions 275
6.11.3 Example Submodular Functions 276
6.11.4 Submodular Optimization 279
6.11.5 Applications of Submodularity in Machine Learning and AI 283
6.11.6 Sketching, CoreSets, Distillation, and Data Subset & Feature Selection 283
6.11.7 Combinatorial Information Functions 287
6.11.8 Clustering, Data Partitioning, and Parallel Machine Learning 288
6.11.9 Active and Semi-Supervised Learning 288
6.11.10 Probabilistic Modeling 289
6.11.11 Structured Norms and Loss Functions 291
6.11.12 Conclusions 291

II Inference 293

7 Inference algorithms: an overview 295
7.1 Introduction 295
7.2 Common inference patterns 295
7.2.1 Global latents 296
7.2.2 Local latents 296
7.2.3 Global and local latents 297
7.3 Exact inference algorithms 297
7.4 Approximate inference algorithms 298
7.4.1 MAP estimation 298
7.4.2 Grid approximation 298
7.4.3 Laplace (quadratic) approximation 299
7.4.4 Variational inference 300
7.4.5 Markov Chain Monte Carlo (MCMC) 302
7.4.6 Sequential Monte Carlo 303
7.5 Evaluating approximate inference algorithms 304

8 Message passing inference 307
8.1 Introduction 307
8.2 Belief propagation for discrete chains 308
8.2.1 Example: casino HMM 308
8.2.2 Forwards filtering 309
8.2.3 Backwards smoothing 312
8.2.4 The Viterbi algorithm 316
8.2.5 Forwards filtering, backwards sampling 319
8.3 Belief propagation for Gaussian chains 320
8.3.1 Example: tracking SSM 321
8.3.2 The Kalman filter 322
8.3.3 The Kalman (RTS) smoother 326

8.3.4 Extensions to the nonlinear case 328
8.3.5 Extensions to the non-Gaussian case 328
8.4 Belief propagation on trees 328
8.4.1 BP for polytrees 329
8.4.2 BP for undirected graphs with pairwise potentials 332
8.4.3 Max product belief propagation 333
8.5 Loopy belief propagation 335
8.5.1 Loopy BP for factor graphs 336
8.5.2 Gaussian belief propagation 337
8.5.3 Convergence 338
8.5.4 Accuracy 341
8.5.5 Connection with variational inference 341
8.5.6 Generalized belief propagation 341
8.5.7 Application: error correcting codes 342
8.5.8 Application: Affinity propagation 343
8.5.9 Emulating BP with graph neural nets 344
8.6 The variable elimination (VE) algorithm 345
8.6.1 Derivation of the algorithm 345
8.6.2 Computational complexity of VE 347
8.6.3 Computational complexity of exact inference 349
8.6.4 Drawbacks of VE 350
8.7 The junction tree algorithm (JTA) 351
8.7.1 Creating a junction tree 351
8.7.2 Message passing on a junction tree 355
8.7.3 Gaussian message passing 358
8.7.4 The generalized distributive law 361
8.8 Inference as backpropagation 361

9 Variational inference 365
9.1 Introduction 365
9.1.1 Variational free energy 365
9.1.2 Evidence lower bound (ELBO) 366
9.2 Mean field VI 367
9.2.1 Coordinate ascent variational inference (CAVI) 367
9.2.2 Example: CAVI for the Ising model 368
9.2.3 Variational Bayes 370
9.2.4 Example: VB for a univariate Gaussian 371
9.2.5 Variational Bayes EM 374
9.2.6 Example: VBEM for a GMM 375
9.2.7 Variational message passing (VMP) 381
9.2.8 Autoconj 382
9.3 Fixed-form VI 382
9.3.1 Stochastic variational inference 382
9.3.2 Black-box variational inference 383
9.3.3 Reparameterization VI 385

9.3.4 Full-rank Gaussian VI 386
9.3.5 Low-rank Gaussian VI 387
9.3.6 Example: Full-rank vs diagonal GVI on 1d linear regression 388
9.3.7 Sparse Gaussian VI 390
9.3.8 Automatic differentiation VI 392
9.3.9 Non-Gaussian reparameterized VI 394
9.3.10 Amortized inference 396
9.3.11 Online variational inference 398
9.4 More accurate variational posteriors 401
9.4.1 Structured mean field 401
9.4.2 Hierarchical (auxiliary variable) posteriors 401
9.4.3 Normalizing flow posteriors 402
9.4.4 Implicit posteriors 403
9.4.5 Combining VI with MCMC inference 403
9.5 Lower bounds 404
9.5.1 Multi-sample ELBO (IWAE bound) 404
9.5.2 The thermodynamic variational objective (TVO) 405
9.6 Upper bounds 406
9.6.1 Minimizing the χ-divergence upper bound 407
9.6.2 Minimizing the evidence upper bound 408
9.7 Expectation propagation (EP) 408
9.7.1 Minimizing forwards vs reverse KL 408
9.7.2 EP as generalized ADF 410
9.7.3 Algorithm 410
9.7.4 Example 412
9.7.5 Optimization issues 412
9.7.6 Power EP and α-divergence 413
9.7.7 Stochastic EP 413
9.7.8 Applications 414

10 Monte Carlo inference 415
10.1 Introduction 415
10.2 Monte Carlo integration 415
10.2.1 Example: estimating π by Monte Carlo integration 416
10.2.2 Accuracy of Monte Carlo integration 416
10.3 Generating random samples from simple distributions 418
10.3.1 Sampling using the inverse cdf 418
10.3.2 Sampling from a Gaussian (Box-Muller method) 419
10.4 Rejection sampling 419
10.4.1 Basic idea 420
10.4.2 Example 421
10.4.3 Adaptive rejection sampling 421
10.4.4 Rejection sampling in high dimensions 422
10.5 Importance sampling 422
10.5.1 Direct importance sampling 423

10.5.2 Self-normalized importance sampling 423
10.5.3 Choosing the proposal 424
10.5.4 Annealed importance sampling (AIS) 424
10.6 Controlling Monte Carlo variance 426
10.6.1 Common random numbers 426
10.6.2 Rao-Blackwellisation 426
10.6.3 Control variates 427
10.6.4 Antithetic sampling 428
10.6.5 Quasi Monte Carlo (QMC) 429

11 Markov Chain Monte Carlo inference 431
11.1 Introduction 431
11.2 Metropolis Hastings algorithm 432
11.2.1 Basic idea 432
11.2.2 Why MH works 433
11.2.3 Proposal distributions 434
11.2.4 Initialization 437
11.2.5 Simulated annealing 437
11.3 Gibbs sampling 439
11.3.1 Basic idea 440
11.3.2 Gibbs sampling is a special case of MH 440
11.3.3 Example: Gibbs sampling for Ising models 441
11.3.4 Example: Gibbs sampling for Potts models 442
11.3.5 Example: Gibbs sampling for GMMs 443
11.3.6 Sampling from the full conditionals 445
11.3.7 Blocked Gibbs sampling 446
11.3.8 Collapsed Gibbs sampling 446
11.4 Auxiliary variable MCMC 449
11.4.1 Slice sampling 449
11.4.2 Swendsen Wang 451
11.5 Hamiltonian Monte Carlo (HMC) 452
11.5.1 Hamiltonian mechanics 453
11.5.2 Integrating Hamilton's equations 453
11.5.3 The HMC algorithm 455
11.5.4 Tuning HMC 455
11.5.5 Riemann Manifold HMC 457
11.5.6 Langevin Monte Carlo (MALA) 458
11.5.7 Connection between SGD and Langevin sampling 459
11.5.8 Applying HMC to constrained parameters 460
11.5.9 Speeding up HMC 461
11.6 MCMC convergence 461
11.6.1 Mixing rates of Markov chains 462
11.6.2 Practical convergence diagnostics 463
11.6.3 Improving speed of convergence 470
11.6.4 Non-centered parameterizations and Neal's funnel 471

11.7 Stochastic gradient MCMC 472
11.7.1 Stochastic Gradient Langevin Dynamics (SGLD) 473
11.7.2 Preconditionining 473
11.7.3 Reducing the variance of the gradient estimate 474
11.7.4 SG-HMC 475
11.7.5 Underdamped Langevin Dynamics 476
11.8 Reversible jump (trans-dimensional) MCMC 477
11.8.1 Basic idea 477
11.8.2 Example 479
11.8.3 Discussion 480
11.9 Annealing methods 480
11.9.1 Parallel tempering 480

12 Sequential Monte Carlo inference 481
12.1 Introduction 481
12.1.1 Problem statement 481
12.1.2 Particle filtering for state-space models 481
12.1.3 SMC samplers for static parameter estimation 483
12.2 Basics of SMC 483
12.2.1 Importance sampling 483
12.2.2 Sequential importance sampling 484
12.2.3 Sequential importance sampling with resampling 485
12.2.4 Resampling methods 488
12.2.5 Adaptive resampling 490
12.3 Some applications of particle filtering 491
12.3.1 1d pendulum model with outliers 491
12.3.2 Visual object tracking 493
12.3.3 Robot localization 493
12.3.4 Online parameter estimation 495
12.4 Proposal distributions 495
12.4.1 Locally optimal proposal 496
12.4.2 Proposals based on the Laplace approximation 496
12.4.3 Proposals based on the extended and unscented Kalman filter 498
12.4.4 Proposals based on SMC 498
12.4.5 Neural adaptive SMC 499
12.4.6 Amortized adaptive SMC 499
12.4.7 Variational SMC 500
12.5 Rao-Blackwellised particle filtering (RBPF) 501
12.5.1 Mixture of Kalman filters 501
12.5.2 FastSLAM 505
12.6 SMC samplers 506
12.6.1 Ingredients of an SMC sampler 507
12.6.2 Likelihood tempering (geometric path) 508
12.6.3 Data tempering 510
12.6.4 Sampling rare events and extrema 512

12.6.5 SMC-ABC and likelihood-free inference 513
12.6.6 SMC^2 513
12.7 Particle MCMC methods 513
12.7.1 Particle Marginal Metropolis Hastings 514
12.7.2 Particle Independent Metropolis Hastings 515
12.7.3 Particle Gibbs 516

III Prediction 517

13 Predictive models: an overview 519
13.1 Introduction 519
13.1.1 Types of model 519
13.1.2 Model fitting using ERM, MLE and MAP 520
13.1.3 Model fitting using Bayes, VI and generalized Bayes 521
13.2 Evaluating predictive models 522
13.2.1 Proper scoring rules 522
13.2.2 Calibration 522
13.2.3 Beyond evaluating marginal probabilities 526
13.3 Conformal prediction 529
13.3.1 Conformalizing classification 530
13.3.2 Conformalizing regression 531
13.3.3 Conformalizing Bayes 532
13.3.4 What do we do if we don’t have a calibration set? 533

14 Generalized linear models 535
14.1 Introduction 535
14.1.1 Examples 535
14.1.2 GLMs with non-canonical link functions 538
14.1.3 Maximum likelihood estimation 538
14.1.4 Bayesian inference 539
14.2 Linear regression 540
14.2.1 Conjugate priors 540
14.2.2 Uninformative priors 542
14.2.3 Informative priors 544
14.2.4 Spike and slab prior 546
14.2.5 Laplace prior (Bayesian lasso) 547
14.2.6 Horseshoe prior 548
14.2.7 Automatic relevancy determination 549
14.3 Logistic regression 551
14.3.1 Binary logistic regression 552
14.3.2 Multinomial logistic regression 552
14.3.3 Priors 553
14.3.4 Posteriors 554
14.3.5 Laplace approximation 554

14.3.6 MCMC inference 557
14.3.7 Variational inference 558
14.3.8 Assumed density filtering 558
14.4 Probit regression 562
14.4.1 Latent variable interpretation 563
14.4.2 Maximum likelihood estimation 563
14.4.3 Bayesian inference 565
14.4.4 Ordinal probit regression 565
14.4.5 Multinomial probit models 566
14.5 Multi-level GLMs 566
14.5.1 Generalized linear mixed models (GLMMs) 566
14.5.2 Model fitting 567
14.5.3 Example: radon regression 567

15 Deep neural networks 571

15.1 Introduction 571
15.2 Building blocks of differentiable circuits 571
15.2.1 Linear layers 572
15.2.2 Non-linearities 572
15.2.3 Convolutional layers 573
15.2.4 Residual (skip) connections 574
15.2.5 Normalization layers 575
15.2.6 Dropout layers 575
15.2.7 Attention layers 576
15.2.8 Recurrent layers 579
15.2.9 Multiplicative layers 579
15.2.10 Implicit layers 580
15.3 Canonical examples of neural networks 580
15.3.1 Multi-layer perceptrons (MLP) 581
15.3.2 Convolutional neural networks (CNN) 581
15.3.3 Recurrent neural networks (RNN) 581
15.3.4 Transformers 583
15.3.5 Graph neural networks (GNNs) 586

16 Bayesian neural networks 593

16.1 Introduction 593
16.2 Priors for BNNs 593
16.2.1 Gaussian priors 593
16.2.2 Sparsity-promoting priors 596
16.2.3 Learning the prior 596
16.2.4 Priors in function space 596
16.2.5 Architectural priors 596
16.3 Likelihoods for BNNs 597
16.4 Posteriors for BNNs 598
16.4.1 Laplace approximation 598

16.4.2 Variational inference 599
16.4.3 Expectation propagation 600
16.4.4 Last layer methods 600
16.4.5 Dropout 601
16.4.6 MCMC methods 601
16.4.7 Methods based on the SGD trajectory 601
16.4.8 Deep ensembles 602
16.4.9 Approximating the posterior predictive distibution 607
16.5 Generalization in Bayesian deep learning 608
16.5.1 Sharp vs flat minima 608
16.5.2 Effective dimensionality of a model 609
16.5.3 The hypothesis space of DNNs 611
16.5.4 Double descent 612
16.5.5 A Bayesian Resolution to Double Descent 613
16.5.6 PAC-Bayes 615
16.5.7 Out-of-Distribution Generalization for BNNs 616
16.6 Online inference 619
16.6.1 Extended Kalman Filtering for DNNs 619
16.6.2 Assumed Density Filtering for DNNs 621
16.6.3 Sequential Laplace for DNNs 623
16.6.4 Variational methods 623
16.7 Hierarchical Bayesian neural networks 623
16.7.1 Solving multiple related classification problems 624

17 Gaussian processes 627
17.1 Introduction 627
17.1.1 GPs: What and why? 627
17.2 Mercer kernels 629
17.2.1 Some popular Mercer kernels 630
17.2.2 Mercer's theorem 636
17.2.3 Kernels from Spectral Densities 637
17.3 GPs with Gaussian likelihoods 638
17.3.1 Predictions using noise-free observations 638
17.3.2 Predictions using noisy observations 639
17.3.3 Weight space vs function space 640
17.3.4 Semi-parametric GPs 641
17.3.5 Marginal likelihood 642
17.3.6 Computational and numerical issues 642
17.3.7 Kernel ridge regression 643
17.4 GPs with non-Gaussian likelihoods 646
17.4.1 Binary classification 647
17.4.2 Multi-class classification 648
17.4.3 GPs for Poisson regression (Cox process) 649
17.4.4 Other likelihoods 649
17.5 Scaling GP inference to large datasets 650

17.5.1 Subset of data 650
17.5.2 Nyström approximation 651
17.5.3 Inducing point methods 652
17.5.4 Sparse variational methods 655
17.5.5 Exploiting parallelization and structure via kernel matrix multiplies 659
17.5.6 Converting a GP to a SSM 661
17.6 Learning the kernel 661
17.6.1 Empirical Bayes for the kernel parameters 662
17.6.2 Bayesian inference for the kernel parameters 664
17.6.3 Multiple kernel learning for additive kernels 665
17.6.4 Automatic search for compositional kernels 667
17.6.5 Spectral mixture kernel learning 669
17.6.6 Deep kernel learning 671
17.6.7 Functional kernel learning 673
17.7 GPs and DNNs 673
17.7.1 Kernels derived from random DNNs (NN-GP) 673
17.7.2 Kernels derived from trained DNNs (neural tangent kernel) 677
17.7.3 Deep GPs 679

18 Beyond the iid assumption 685
18.1 Introduction 685
18.2 Distribution shift 685
18.2.1 Motivating examples 685
18.2.2 A causal view of distribution shift 687
18.2.3 Covariate shift 688
18.2.4 Domain shift 688
18.2.5 Label / prior shift 689
18.2.6 Concept shift 689
18.2.7 Manifestation shift 689
18.2.8 Selection bias 690
18.3 Training-time techniques for distribution shift 690
18.3.1 Importance weighting for covariate shift 691
18.3.2 Domain adaptation 692
18.3.3 Domain randomization 692
18.3.4 Data augmentation 693
18.3.5 Unsupervised label shift estimation 693
18.3.6 Distributionally robust optimization 694
18.4 Test-time techniques for distribution shift 694
18.4.1 Detecting shifts using two-sample testing 694
18.4.2 Detecting single out-of-distribution (OOD) inputs 694
18.4.3 Selective prediction 697
18.4.4 Open world recognition 698
18.4.5 Online adaptation 699
18.5 Learning from multiple distributions 700
18.5.1 Transfer learning 701

18.5.2 Few-shot learning 702
18.5.3 Prompt tuning 702
18.5.4 Zero-shot learning 702
18.5.5 Multi-task learning 703
18.5.6 Domain generalization 704
18.5.7 Invariant risk minimization 705
18.6 Meta-learning 706
18.6.1 Meta-learning as probabilistic inference for prediction 707
18.6.2 Gradient-based meta-learning 708
18.6.3 Metric-based few-shot learning 708
18.6.4 VERSA 708
18.6.5 Neural processes 709
18.7 Continual learning 709
18.7.1 Domain drift 709
18.7.2 Concept drift 709
18.7.3 Task incremental learning 711
18.7.4 Catastrophic forgetting 712
18.7.5 Online learning 714
18.8 Adversarial examples 715
18.8.1 Whitebox (gradient-based) attacks 717
18.8.2 Blackbox (gradient-free) attacks 717
18.8.3 Real world adversarial attacks 719
18.8.4 Defenses based on robust optimization 719
18.8.5 Why models have adversarial examples 720

IV Generation 723

19 Generative models: an overview 725
19.1 Introduction 725
19.2 Types of generative model 725
19.3 Goals of generative modeling 727
19.3.1 Generating data 727
19.3.2 Density estimation 728
19.3.3 Imputation 728
19.3.4 Structure discovery 730
19.3.5 Latent space interpolation 730
19.3.6 Representation learning 731
19.4 Evaluating generative models 731
19.4.1 Likelihood 732
19.4.2 Distances and divergences in feature space 734
19.4.3 Precision and recall metrics 735
19.4.4 Statistical tests 736
19.4.5 Challenges with using pretrained classifiers 736
19.4.6 Using model samples to train classifiers 737

19.4.7 Assessing overfitting 737
19.4.8 Human evaluation 738

20 Variational autoencoders 739

20.1 Introduction 739
20.2 VAE basics 739
20.2.1 Modeling assumptions 740
20.2.2 Evidence lower bound 741
20.2.3 Optimization 742
20.2.4 The reparameterization trick 742
20.2.5 Computing the reparameterized ELBO 744
20.2.6 Comparison of VAEs and autoencoders 746
20.2.7 VAEs optimize in an augmented space 747
20.3 VAE generalizations 748
20.3.1 σ-VAE 748
20.3.2 β-VAE 750
20.3.3 InfoVAE 752
20.3.4 Multi-modal VAEs 754
20.3.5 VAEs with missing data 757
20.3.6 Semi-supervised VAEs 759
20.3.7 VAEs with sequential encoders/decoders 760
20.4 Avoiding posterior collapse 763
20.4.1 KL annealing 764
20.4.2 Lower bounding the rate 764
20.4.3 Free bits 764
20.4.4 Adding skip connections 765
20.4.5 Improved variational inference 765
20.4.6 Alternative objectives 765
20.4.7 Enforcing identifiability 766
20.5 VAEs with hierarchical structure 767
20.5.1 Bottom-up vs top-down inference 767
20.5.2 Example: Very deep VAE 768
20.5.3 Connection with autoregressive models 769
20.5.4 Variational pruning 771
20.5.5 Other optimization difficulties 772
20.6 Vector quantization VAE 772
20.6.1 Autoencoder with binary code 772
20.6.2 VQ-VAE model 774
20.6.3 Learning the prior 775
20.6.4 Hierarchical extension (VQ-VAE-2) 775
20.6.5 Discrete VAE 776
20.6.6 VQ-GAN 777
20.7 Wake-sleep algorithm 778
20.7.1 Wake phase 779
20.7.2 Sleep phase 780

20.7.3 Daydream phase 780
20.7.4 Summary of algorithm 781

21 Auto-regressive models 783

21.1 Introduction 783
21.2 Neural autoregressive density estimators (NADE) 784
21.3 Causal CNNs 784
21.3.1 1d causal CNN (Convolutional Markov models) 785
21.3.2 2d causal CNN (PixelCNN) 785
21.4 Transformer decoders 786
21.4.1 Text generation (GPT) 787
21.4.2 Music generation 787
21.4.3 Text-to-image generation (DALL-E) 788

22 Normalizing Flows 791

22.1 Introduction 791
22.1.1 Preliminaries 791
22.1.2 Example 793
22.1.3 How to train a flow model 794
22.2 Constructing Flows 795
22.2.1 Affine flows 795
22.2.2 Elementwise flows 796
22.2.3 Coupling flows 798
22.2.4 Autoregressive flows 800
22.2.5 Residual flows 805
22.2.6 Continuous-time flows 807
22.3 Applications 809
22.3.1 Density estimation 809
22.3.2 Generative Modeling 810
22.3.3 Inference 810

23 Energy-based models 813

23.1 Introduction 813
23.1.1 Example: Products of experts (PoE) 814
23.1.2 Computational difficulties 814
23.2 Maximum Likelihood Training 815
23.2.1 Gradient-based MCMC methods 816
23.2.2 Contrastive divergence 816
23.3 Score Matching (SM) 819
23.3.1 Basic score matching 820
23.3.2 Denoising Score Matching (DSM) 821
23.3.3 Sliced Score Matching (SSM) 822
23.3.4 Connection to Contrastive Divergence 823
23.3.5 Score-Based Generative Models 824
23.4 Noise Contrastive Estimation 827

23.4.1 Connection to Score Matching 828
23.5 Other Methods 829
23.5.1 Minimizing Differences/Derivatives of KL Divergences 829
23.5.2 Minimizing the Stein Discrepancy 830
23.5.3 Adversarial Training 830

24 Denoising diffusion models 833
24.1 Model definition 833
24.2 Examples 835
24.3 Model training 836
24.4 Connections with other generative models 838
24.4.1 Connection with score matching 838
24.4.2 Connection with VAEs 839
24.4.3 Connection with flow models 839

25 Generative adversarial networks 841
25.1 Introduction 841
25.2 Learning by Comparison 842
25.2.1 Guiding principles 843
25.2.2 Class probability estimation 844
25.2.3 Bounds on f-divergences 847
25.2.4 Integral probability metrics 848
25.2.5 Moment matching 850
25.2.6 On density ratios and differences 851
25.3 Generative Adversarial Networks 852
25.3.1 From learning principles to loss functions 853
25.3.2 Gradient Descent 854
25.3.3 Challenges with GAN training 855
25.3.4 Improving GAN optimization 857
25.3.5 Convergence of GAN training 858
25.4 Conditional GANs 861
25.5 Inference with GANs 862
25.6 Neural architectures in GANs 863
25.6.1 The importance of discriminator architectures 863
25.6.2 Architectural inductive biases 864
25.6.3 Attention in GANs 864
25.6.4 Progressive generation 865
25.6.5 Regularization 867
25.6.6 Scaling up GAN models 867
25.7 Applications 867
25.7.1 GANs for image generation 867
25.7.2 Video generation 871
25.7.3 Audio generation 871
25.7.4 Text generation 871
25.7.5 Imitation Learning 872

25.7.6 Domain Adaptation 873
25.7.7 Design, Art and Creativity 873

V Discovery 875

26 Discovery methods: an overview 877
26.1 Introduction 877
26.2 Overview of Part V 878

27 Latent factor models 879
27.1 Introduction 879
27.2 Mixture models 879
27.2.1 Gaussian mixture models (GMMs) 880
27.2.2 Bernoulli mixture models 882
27.2.3 Gaussian scale mixtures 882
27.2.4 Using GMMs as a prior for inverse imaging problems 884
27.3 Factor analysis 887
27.3.1 Vanilla factor analysis 887
27.3.2 Probabilistic PCA 891
27.3.3 Factor analysis models for paired data 894
27.3.4 Factor analysis with exponential family likelihoods 897
27.3.5 Factor analysis with DNN likelihoods 898
27.3.6 Factor analysis with GP likelihoods (GP-LVM) 899
27.4 Mixture of factor analysers 901
27.4.1 Model definition 901
27.4.2 Model fitting 902
27.4.3 MixFA for image generation 903
27.5 LVMs with non-Gaussian priors 907
27.5.1 Non-negative matrix factorization (NMF) 907
27.5.2 Multinomial PCA 908
27.6 Topic models 911
27.6.1 Latent Dirichlet Allocation (LDA) 911
27.6.2 Correlated topic model 914
27.6.3 Dynamic topic model 915
27.6.4 LDA-HMM 916
27.7 Independent components analysis (ICA) 918
27.7.1 Noiseless ICA model 919
27.7.2 The need for non-Gaussian priors 920
27.7.3 Maximum likelihood estimation 922
27.7.4 Alternatives to MLE 922
27.7.5 Sparse coding 924
27.7.6 Nonlinear ICA 925

28 State-space models 927
28.1 Introduction 927

28.1.1 Models 927
28.1.2 Inferential goals 927
28.2 Hidden Markov models (HMMs) 929
28.2.1 Example models 929
28.2.2 Application: time series segmentation 932
28.2.3 Parameter estimation 935
28.2.4 Model extensions 935
28.3 Linear dynamical systems 935
28.3.1 Example: 2d tracking problem 936
28.3.2 Example: Recursive least squares 938
28.4 Non-linear dynamical systems 940
28.4.1 Example: nonlinear 2d tracking problem 940
28.4.2 Example: Simultaneous localization and mapping (SLAM) 941
28.4.3 Example: stochastic volatility models 943
28.4.4 Example: Multi-target tracking 944
28.5 Other kinds of SSM 945
28.5.1 Exponential family SSMs 946
28.5.2 Bayesian SSMs 949
28.5.3 GP SSMs 949
28.6 Deep SSMs 950
28.6.1 Deep Markov models 950
28.6.2 Recurrent SSM 951
28.6.3 Improving multi-step predictions 952
28.6.4 Variational RNNs 953
28.7 Time series forecasting 954
28.7.1 Structural time series models 954
28.7.2 Causal impact of a time series intervention 961
28.7.3 Prophet 965
28.7.4 Gaussian processes for timeseries forecasting 966
28.7.5 Neural forecasting methods 967

29 Graph learning 969

29.1 Introduction 969
29.2 Latent variable models for graphs 969
29.2.1 Stochastic block model 969
29.2.2 Mixed membership stochastic block model 971
29.2.3 Infinite relational model 973
29.3 Graphical model structure learning 975
29.3.1 Applications 975
29.3.2 Relevance networks 977
29.4 Learning tree structures 978
29.4.1 Directed or undirected tree? 978
29.4.2 Chow-Liu algorithm 979
29.4.3 Finding the MAP forest 980
29.4.4 Mixtures of trees 981

29.5 Learning DAG structures 981
29.5.1 Faithfulness 981
29.5.2 Markov equivalence 982
29.5.3 Bayesian model selection: statistical foundations 983
29.5.4 Bayesian model selection: algorithms 986
29.5.5 Constraint-based approach 989
29.5.6 Methods based on sparse optimization 991
29.5.7 Consistent estimators 992
29.5.8 Handling latent variables 992
29.6 Learning undirected graph structures 1000
29.6.1 Dependency networks 1001
29.6.2 Graphical lasso for GGMs 1002
29.6.3 Graphical lasso for discrete MRFs/CRFs 1004
29.6.4 Bayesian inference for undirected graph structures 1005
29.7 Learning causal DAGs 1006
29.7.1 Learning cause-effect pairs 1007
29.7.2 Learning causal DAGs from interventional data 1010
29.7.3 Learning from low-level inputs 1011

30 Non-parametric Bayesian models 1013

30.1 Introduction 1013
30.2 Dirichlet process 1014
30.2.1 Definition 1014
30.2.2 Stick breaking construction of the DP 1016
30.2.3 The Chinese restaurant process (CRP) 1017
30.2.4 Dirichlet process mixture models 1019
30.3 Generalizations of the Dirichlet process 1024
30.3.1 Pitman-Yor process 1025
30.3.2 Dependent random probability measures 1026
30.4 The Indian buffet process and the Beta process 1028
30.5 Small-variance asymptotics 1031
30.6 Completely random measures 1034
30.7 Lévy processes 1035
30.8 Point processes with repulsion and reinforcement 1037
30.8.1 Poisson process 1037
30.8.2 Renewal process 1038
30.8.3 Hawkes process 1039
30.8.4 Gibbs point process 1041
30.8.5 Determinantal point process 1042

31 Representation learning (Unfinished) 1045

31.1 CLIP 1045

32 Interpretability 1047

32.1 Introduction 1047

32.1.1 The Role of Interpretability 1048
32.1.2 Terminology and Framework 1049
32.2 Methods for Interpretable Machine Learning 1053
32.2.1 Inherently Interpretable Models: The Model is its Explanation 1053
32.2.2 Semi-Inherently Interpretable Models: Example-Based Methods 1055
32.2.3 Post-hoc or Joint training: The Explanation gives a Partial View of the Model 1056
32.2.4 Transparency and Visualization 1060
32.3 Properties: The Abstraction Between Context and Method 1061
32.3.1 Properties of Explanations from Interpretable Machine Learning 1062
32.3.2 Properties of Explanations from Cognitive Science 1064
32.4 Evaluation of Interpretable Machine Learning Models 1065
32.4.1 Computational Evaluation: Does the Method have Desired Properties? 1066
32.4.2 User Study-based Evaluation: Does the Method Help a User Perform a Task? 1071
32.5 Discussion: How to Think about Interpretable Machine Learning 1074

VI Decision making 1081

33 Multi-step decision problems 1083
33.1 Introduction 1083
33.2 Decision (influence) diagrams 1083
33.2.1 Example: oil wildcatter 1083
33.2.2 Information arcs 1084
33.2.3 Value of information 1085
33.2.4 Computing the optimal policy 1086
33.3 A/B testing 1086
33.3.1 A Bayesian approach 1087
33.3.2 Example 1090
33.4 Contextual bandits 1091
33.4.1 Types of bandit 1091
33.4.2 Applications 1093
33.4.3 Exploration-exploitation tradeoff 1093
33.4.4 The optimal solution 1093
33.4.5 Upper confidence bounds (UCB) 1095
33.4.6 Thompson sampling 1097
33.4.7 Regret 1098
33.5 Markov decision problems 1099
33.5.1 Basics 1100
33.5.2 Partially observed MDPs 1101
33.5.3 Episodes and returns 1102
33.5.4 Value functions 1102
33.5.5 Optimal value functions and policies 1103
33.6 Planning in an MDP 1104

33.6.1 Value iteration 1105
33.6.2 Policy iteration 1106
33.6.3 Linear programming 1107

34 Reinforcement learning 1109

34.1 Introduction 1109
34.1.1 Overview of methods 1109
34.1.2 Value based methods 1110
34.1.3 Policy search methods 1111
34.1.4 Model-based RL 1111
34.1.5 Exploration-exploitation tradeoff 1112
34.2 Value-based RL 1114
34.2.1 Monte Carlo RL 1114
34.2.2 Temporal difference (TD) learning 1114
34.2.3 TD learning with eligibility traces 1115
34.2.4 SARSA: on-policy TD control 1116
34.2.5 Q-learning: off-policy TD control 1117
34.2.6 Deep Q-network (DQN) 1119
34.3 Policy-based RL 1120
34.3.1 The policy gradient theorem 1120
34.3.2 REINFORCE 1121
34.3.3 Actor-critic methods 1122
34.3.4 Bound optimization methods 1124
34.3.5 Deterministic policy gradient methods 1126
34.3.6 Gradient-free methods 1127
34.4 Model-based RL 1127
34.4.1 Model predictive control (MPC) 1127
34.4.2 Combining model-based and model-free 1129
34.4.3 MBRL using Gaussian processes 1129
34.4.4 MBRL using DNNs 1131
34.4.5 MBRL using latent-variable models 1131
34.4.6 Robustness to model errors 1134
34.5 Off-policy learning 1134
34.5.1 Basic techniques 1135
34.5.2 The curse of horizon 1138
34.5.3 The deadly triad 1139
34.6 Control as inference 1140
34.6.1 Maximum entropy reinforcement learning 1140
34.6.2 Other approaches 1143
34.6.3 Imitation learning 1144

35 Causality 1147

35.1 Introduction 1147
35.1.1 Why is causality different than other forms of ML? 1147
35.2 Causal Formalism 1149

35.2.1 Structural Causal Models 1149
35.2.2 Causal DAGs 1151
35.2.3 Identification 1153
35.2.4 Counterfactuals and the Causal Hierarchy 1154
35.3 Randomized Control Trials 1156
35.4 Confounder Adjustment 1157
35.4.1 Causal Estimand, Statistical Estimand, and Identification 1157
35.4.2 ATE Estimation with Observed Confounders 1160
35.4.3 Uncertainity Quantification 1165
35.4.4 Matching 1166
35.4.5 Practical Considerations and Procedures 1167
35.4.6 Summary and Practical Advice 1170
35.5 Instrumental Variable Strategies 1171
35.5.1 Additive Unobserved Confounding 1173
35.5.2 Instrument Monotonicity and Local Average Treatment Effect 1174
35.5.3 Two Stage Least Squares 1178
35.6 Difference in Differences 1178
35.6.1 Estimation 1182
35.7 Credibility Checks 1182
35.7.1 Placebo Checks 1183
35.7.2 Sensitivity Analysis to Unobserved Confounding 1183
35.8 The Do Calculus 1191
35.8.1 The three rules 1191
35.8.2 Revisiting Backdoor Adjustment 1192
35.8.3 Frontdoor Adjustment 1193
35.9 Further Reading 1195

Bibliography 1210