Contents

Preface xxix

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 31
2.3.2 Examples 32
2.3.3 Log partition function is cumulant generating function 36
2.3.4 Canonical (natural) vs mean (moment) parameters 38
2.3.5 MLE for the exponential family 39
2.3.6 Exponential dispersion family 40
2.3.7 Maximum entropy derivation of the exponential family 40
2.4 Fisher information matrix (FIM) 41
2.4.1 Definition 41
2.4.2 Equivalence between the FIM and the Hessian of the NLL 41
2.4.3 Examples 43
2.4.4 Approximating KL divergence using FIM 44
2.4.5 Fisher information matrix for exponential family 44
2.5 Transformations of random variables 45
2.5.1 Invertible transformations (bijections) 46
2.5.2 Monte Carlo approximation 46
2.5.3 Probability integral transform 47
2.6 Markov chains 47
2.6.1 Parameterization 48
2.6.2 Application: Language modeling 51
2.6.3 Parameter estimation 51

2.6.4 Stationary distribution of a Markov chain 53
2.7 Divergence measures between probability distributions 57
2.7.1 f-divergence 57
2.7.2 Integral probability metrics 59
2.7.3 Maximum mean discrepancy (MMD) 60
2.7.4 Total variation distance 62
2.7.5 Comparing distributions using binary classifiers 62

3 Bayesian statistics 65
3.1 Introduction 65
3.1.1 Frequentist statistics 65
3.1.2 Bayesian statistics 65
3.1.3 Arguments for the Bayesian approach 66
3.1.4 Arguments against the Bayesian approach 67
3.1.5 Why not just use MAP estimation? 67
3.2 Closed-form analysis using conjugate priors 72
3.2.1 The binomial model 72
3.2.2 The multinomial model 73
3.2.3 The univariate Gaussian model 75
3.3 Conjugate Bayesian analysis for the multivariate Gaussian 80
3.3.1 Posterior of $\boldsymbol{\mu}$ given $\boldsymbol{\Sigma}$ 80
3.3.2 Posterior of $\boldsymbol{\Sigma}$ given $\boldsymbol{\mu}$ 80
3.3.3 Posterior of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ 82
3.3.4 Conjugate-exponential models 86
3.4 Beyond conjugate priors 88
3.4.1 Robust (heavy-tailed) priors 88
3.4.2 Priors for variance parameters 89
3.5 Noninformative priors 91
3.5.1 Maximum entropy priors 91
3.5.2 Jeffreys priors 92
3.5.3 Invariant priors 95
3.5.4 Reference priors 96
3.6 Hierarchical priors 96
3.6.1 A hierarchical binomial model 97
3.6.2 A hierarchical Gaussian model 99
3.7 Empirical Bayes 102
3.7.1 A hierarchical binomial model 103
3.7.2 A hierarchical Gaussian model 104
3.7.3 Hierarchical Bayes for n-gram smoothing 105
3.8 Model selection and evaluation 107
3.8.1 Bayesian model selection 107
3.8.2 Estimating the marginal likelihood 108
3.8.3 Connection between cross validation and marginal likelihood 109
3.8.4 Pareto-Smoothed Importance Sampling LOO estimate 110
3.8.5 Information criteria 111
3.8.6 Posterior predictive checks 113
3.8.7 Bayesian p-values 114

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) 128
4.2.6 Inference 128

4.2.7 Learning 130
4.2.8 Plate notation 135
4.3 Undirected graphical models (Markov random fields) 138
4.3.1 Representing the joint distribution 138
4.3.2 Fully visible MRFs (Ising, Potts, Hopfield, etc) 140
4.3.3 MRFs with latent variables (Boltzmann machines, etc) 146
4.3.4 Maximum entropy models 148
4.3.5 Gaussian MRFs 150
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 160
4.4.2 2d CRFs 163
4.4.3 Parameter estimation 166
4.4.4 Other approaches to structured prediction 167
4.5 Comparing directed and undirected PGMs 167
4.5.1 CI properties 167
4.5.2 Converting between a directed and undirected model 169
4.5.3 Conditional directed vs undirected PGMs and the label bias problem 170
4.5.4 Combining directed and undirected graphs 171
4.5.5 Comparing directed and undirected Gaussian PGMs 173
4.6 PGM extensions 175
4.6.1 Factor graphs 175
4.6.2 Probabilistic circuits 178
4.6.3 Directed relational PGMs 178
4.6.4 Undirected relational PGMs 180
4.6.5 Open-universe probability models 183
4.6.6 Programs as probability models 184
4.7 Structural causal models 184
4.7.1 Example: causal impact of education on wealth 185
4.7.2 Structural equation models 186
4.7.3 Do operator and augmented DAGs 187
4.7.4 Estimating average treatment effect using path analysis 188
4.7.5 Counterfactuals 189

5 Information theory 193

5.1 KL divergence 193
5.1.1 Desiderata 194
5.1.2 The KL divergence uniquely satisfies the desiderata 195
5.1.3 Thinking about KL 198
5.1.4 Properties of KL 200
5.1.5 KL divergence and MLE 202
5.1.6 KL divergence and Bayesian Inference 203
5.1.7 KL divergence and Exponential Families 204
5.1.8 Bregman divergence 205
5.2 Entropy 206
5.2.1 Definition 206
5.2.2 Differential entropy for continuous random variables 207
5.2.3 Typical sets 208
5.2.4 Cross entropy and perplexity 209
5.3 Mutual information 210
5.3.1 Definition 210
5.3.2 Interpretation 210
5.3.3 Data processing inequality 211

5.3.4 Sufficient Statistics 212
5.3.5 Multivariate mutual information 212
5.3.6 Variational bounds on mutual information 215
5.3.7 Relevance networks 217
5.4 Data compression (source coding) 218
5.4.1 Lossless compression 219
5.4.2 Lossy compression and the rate-distortion tradeoff 219
5.4.3 Bits back coding 221
5.5 Error-correcting codes (channel coding) 222
5.6 The information bottleneck 224
5.6.1 Vanilla IB 224
5.6.2 Variational IB 225
5.6.3 Conditional entropy bottleneck 226

6 Optimization 229

6.1 Introduction 229
6.2 Automatic differentiation 229
6.2.1 Differentiation in functional form 229
6.2.2 Differentiating chains, circuits, and programs 234
6.3 Stochastic gradient descent 239
6.4 Natural gradient descent 240
6.4.1 Defining the natural gradient 240
6.4.2 Interpretations of NGD 241
6.4.3 Benefits of NGD 242
6.4.4 Approximating the natural gradient 243
6.4.5 Natural gradients for the exponential family 244
6.5 Gradients of stochastic functions 246
6.5.1 Minibatch approximation to finite-sum objectives 247
6.5.2 Optimizing parameters of a distribution 247
6.5.3 Score function estimator (likelihood ratio trick) 248
6.5.4 Reparameterization trick 249
6.5.5 The delta method 251
6.5.6 Gumbel softmax trick 251
6.5.7 Stochastic computation graphs 252
6.5.8 Straight-through estimator 252
6.6 Bound optimization (MM) algorithms 253
6.6.1 The general algorithm 253
6.6.2 Example: logistic regression 254
6.6.3 The EM algorithm 256
6.6.4 Example: EM for an MVN with missing data 258
6.6.5 Example: robust linear regression using Student-t likelihood 260
6.6.6 Extensions to EM 261
6.7 The Bayesian learning rule 263
6.7.1 Deriving inference algorithms from BLR 264
6.7.2 Deriving optimization algorithms from BLR 266
6.7.3 Variational optimization 270
6.8 Bayesian optimization 270
6.8.1 Sequential model-based optimization 271
6.8.2 Surrogate functions 272
6.8.3 Acquisition functions 273
6.8.4 Other issues 276
6.9 Derivative free optimization 277
6.9.1 Local search 277
6.9.2 Simulated annealing 280
6.9.3 Evolutionary algorithms 283
6.9.4 Estimation of distribution (EDA) algorithms 284

6.9.5 Cross-entropy method 287
6.9.6 Evolutionary strategies 287
6.10 Optimal Transport 289
6.10.1 Warm-up: Matching optimally two families of points 289
6.10.2 From Optimal Matchings to Kantorovich and Monge formulations 289
6.10.3 Solving optimal transport 292
6.11 Submodular optimization 297
6.11.1 Intuition, Examples, and Background 297
6.11.2 Submodular Basic Definitions 300
6.11.3 Example Submodular Functions 301
6.11.4 Submodular Optimization 303
6.11.5 Applications of Submodularity in Machine Learning and AI 308
6.11.6 Sketching, CoreSets, Distillation, and Data Subset & Feature Selection 308
6.11.7 Combinatorial Information Functions 312
6.11.8 Clustering, Data Partitioning, and Parallel Machine Learning 313
6.11.9 Active and Semi-Supervised Learning 313
6.11.10 Probabilistic Modeling 314
6.11.11 Structured Norms and Loss Functions 316
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 353
8.4 Inference for non-linear and/or non-Gaussian SSMs 355

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) 360
8.5.5 Examples 361
8.6 Inference based on the unscented transform 363
8.6.1 The unscented transform 364
8.6.2 The unscented Kalman filter (UKF) 365
8.6.3 The unscented Kalman smoother (UKS) 366
8.6.4 Examples 366
8.7 Inference based on moment matching 367
8.7.1 Gaussian moment matching 367
8.7.2 The general Gaussian filter 367
8.7.3 Implementation using numerical integration 368
8.8 Inference based on statistical linearization 368
8.8.1 Statistical linear regression 368
8.8.2 Prior linearization filter 369
8.8.3 Iterated posterior linearization filter 370
8.8.4 Iterated posterior linearization smoother 370
8.8.5 Beyond additive Gaussian noise 371
8.9 Other variants of the Kalman filter 373
8.9.1 Ensemble Kalman filter 373
8.9.2 Robust Kalman filters 375
8.9.3 Dual EKF 375
8.10 Assumed density filtering 375
8.10.1 Connection with Gaussian filtering 377
8.10.2 ADF for SLDS (Gaussian sum filter) 378
8.10.3 ADF for logistic regression 379
8.10.4 ADF for DNNs (“probabilistic backpropagation”) 382

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 408

10 Variational inference 411

10.1 Introduction 411
10.1.1 Variational free energy 411
10.1.2 Evidence lower bound (ELBO) 412
10.2 Mean field VI 413
10.2.1 Coordinate ascent variational inference (CAVI) 413
10.2.2 Example: CAVI for the Ising model 414
10.2.3 Variational Bayes 416
10.2.4 Example: VB for a univariate Gaussian 417
10.2.5 Variational Bayes EM 420
10.2.6 Example: VBEM for a GMM 421
10.2.7 Variational message passing (VMP) 427
10.2.8 Autoconj 428
10.3 Fixed-form VI 428
10.3.1 Stochastic variational inference 428
10.3.2 Black-box variational inference 429
10.3.3 Reparameterization VI 431
10.3.4 Full-rank Gaussian VI 432
10.3.5 Low-rank Gaussian VI 432
10.3.6 Automatic differentiation VI 434
10.3.7 Sparse Gaussian VI 438
10.3.8 Non-Gaussian reparameterized VI 440
10.3.9 Amortized inference 441
10.4 More accurate variational posteriors 443
10.4.1 Structured mean field 443
10.4.2 Hierarchical (auxiliary variable) posteriors 443
10.4.3 Normalizing flow posteriors 444
10.4.4 Implicit posteriors 446
10.4.5 Combining VI with MCMC inference 446
10.5 Lower bounds 446
10.5.1 Multi-sample ELBO (IWAE bound) 446
10.5.2 The thermodynamic variational objective (TVO) 447
10.6 Upper bounds 448
10.6.1 Minimizing the χ-divergence upper bound 449
10.6.2 Minimizing the evidence upper bound 450
10.7 Expectation propagation (EP) 450
10.7.1 Minimizing forwards vs reverse KL 450
10.7.2 EP as generalized ADF 452
10.7.3 Algorithm 453
10.7.4 Example 454
10.7.5 Optimization issues 455
10.7.6 Power EP and α-divergence 455
10.7.7 Stochastic EP 455
10.7.8 Applications 456

11 Monte Carlo inference 457

11.1 Introduction 457
11.2 Monte Carlo integration 457
11.2.1 Example: estimating π by Monte Carlo integration 458
11.2.2 Accuracy of Monte Carlo integration 458
11.3 Generating random samples from simple distributions 460
11.3.1 Sampling using the inverse cdf 460
11.3.2 Sampling from a Gaussian (Box-Muller method) 461
11.4 Rejection sampling 461

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

11.4.1 Basic idea 462
11.4.2 Example 463
11.4.3 Adaptive rejection sampling 463
11.4.4 Rejection sampling in high dimensions 464
11.5 Importance sampling 464
11.5.1 Direct importance sampling 465
11.5.2 Self-normalized importance sampling 465
11.5.3 Choosing the proposal 466
11.5.4 Annealed importance sampling (AIS) 467
11.6 Controlling Monte Carlo variance 468
11.6.1 Common random numbers 468
11.6.2 Rao-Blackwellisation 468
11.6.3 Control variates 469
11.6.4 Antithetic sampling 470
11.6.5 Quasi Monte Carlo (QMC) 471

12 Markov Chain Monte Carlo inference 473

12.1 Introduction 473
12.2 Metropolis Hastings algorithm 474
12.2.1 Basic idea 474
12.2.2 Why MH works 475
12.2.3 Proposal distributions 476
12.2.4 Initialization 479
12.3 Gibbs sampling 479
12.3.1 Basic idea 479
12.3.2 Gibbs sampling is a special case of MH 480
12.3.3 Example: Gibbs sampling for Ising models 480
12.3.4 Example: Gibbs sampling for Potts models 482
12.3.5 Example: Gibbs sampling for GMMs 482
12.3.6 Sampling from the full conditionals 484
12.3.7 Blocked Gibbs sampling 485
12.3.8 Collapsed Gibbs sampling 486
12.4 Auxiliary variable MCMC 488
12.4.1 Slice sampling 489
12.4.2 Swendsen Wang 490
12.5 Hamiltonian Monte Carlo (HMC) 492
12.5.1 Hamiltonian mechanics 492
12.5.2 Integrating Hamilton's equations 493
12.5.3 The HMC algorithm 494
12.5.4 Tuning HMC 495
12.5.5 Riemann Manifold HMC 496
12.5.6 Langevin Monte Carlo (MALA) 497
12.5.7 Connection between SGD and Langevin sampling 498
12.5.8 Applying HMC to constrained parameters 500
12.5.9 Speeding up HMC 501
12.6 MCMC convergence 501
12.6.1 Mixing rates of Markov chains 502
12.6.2 Practical convergence diagnostics 502
12.6.3 Improving speed of convergence 510
12.6.4 Non-centered parameterizations and Neal's funnel 510
12.7 Stochastic gradient MCMC 511
12.7.1 Stochastic Gradient Langevin Dynamics (SGLD) 512
12.7.2 Preconditionining 512
12.7.3 Reducing the variance of the gradient estimate 513
12.7.4 SG-HMC 514
12.7.5 Underdamped Langevin Dynamics 515

12.8 Reversible jump (trans-dimensional) MCMC 516
12.8.1 Basic idea 516
12.8.2 Example 518
12.8.3 Discussion 519
12.9 Annealing methods 519
12.9.1 Parallel tempering 520

13 Sequential Monte Carlo inference 521
13.1 Introduction 521
13.1.1 Problem statement 521
13.1.2 Particle filtering for state-space models 521
13.1.3 SMC samplers for static parameter estimation 523
13.2 Particle filtering 523
13.2.1 Importance sampling 523
13.2.2 Sequential importance sampling 524
13.2.3 Sequential importance sampling with resampling 525
13.2.4 Resampling methods 528
13.2.5 Adaptive resampling 530
13.3 Proposal distributions 531
13.3.1 Locally optimal proposal 531
13.3.2 Proposals based on the Laplace approximation 532
13.3.3 Proposals based on the extended and unscented Kalman filter 532
13.3.4 Proposals based on SMC 534
13.3.5 Learned ("neural") proposals **(Unfinished)** 534
13.4 Rao-Blackwellised particle filtering (RBPF) 535
13.4.1 Mixture of Kalman filters 535
13.4.2 Example: tracking a maneuvering object 537
13.4.3 Example: Simultaneous localization and mapping (SLAM) 537
13.5 Extensions of the particle filter 541
13.6 SMC samplers 542
13.6.1 Ingredients of an SMC sampler 542
13.6.2 Likelihood tempering (geometric path) 543
13.6.3 Data tempering 546
13.6.4 Sampling rare events and extrema 547
13.6.5 SMC-ABC and likelihood-free inference 548
13.6.6 SMC^2 548

III Prediction 551

14 Predictive models: an overview 553
14.1 Introduction 553
14.1.1 Types of model 553
14.1.2 Model fitting using ERM, MLE and MAP 554
14.1.3 Model fitting using Bayes, VI and generalized Bayes 555
14.2 Evaluating predictive models 556
14.2.1 Proper scoring rules 556
14.2.2 Calibration 556
14.2.3 Beyond evaluating marginal probabilities 560
14.3 Conformal prediction 563
14.3.1 Conformalizing classification 565
14.3.2 Conformalizing regression 565

15 Generalized linear models 567
15.1 Introduction 567
15.1.1 Examples 567
15.1.2 GLMs with non-canonical link functions 570

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

15.1.3 Maximum likelihood estimation 570
15.1.4 Bayesian inference 571
15.2 Linear regression 572
15.2.1 Conjugate priors 572
15.2.2 Uninformative priors 574
15.2.3 Informative priors 576
15.2.4 Spike and slab prior 578
15.2.5 Laplace prior (Bayesian lasso) 579
15.2.6 Horseshoe prior 580
15.2.7 Automatic relevancy determination 581
15.3 Logistic regression 583
15.3.1 Binary logistic regression 584
15.3.2 Multinomial logistic regression 584
15.3.3 Priors 585
15.3.4 Posteriors 586
15.3.5 Laplace approximation 586
15.3.6 MCMC inference 589
15.3.7 Variational inference 590
15.3.8 Assumed density filtering 590
15.4 Probit regression 590
15.4.1 Latent variable interpretation 590
15.4.2 Maximum likelihood estimation 591
15.4.3 Bayesian inference 593
15.4.4 Ordinal probit regression 593
15.4.5 Multinomial probit models 594
15.5 Multi-level GLMs 594
15.5.1 Generalized linear mixed models (GLMMs) 595
15.5.2 Model fitting 595
15.5.3 Example: radon regression 595

16 Deep neural networks 599

16.1 Introduction 599
16.2 Building blocks of differentiable circuits 599
16.2.1 Linear layers 600
16.2.2 Non-linearities 600
16.2.3 Convolutional layers 601
16.2.4 Residual (skip) connections 602
16.2.5 Normalization layers 603
16.2.6 Dropout layers 603
16.2.7 Attention layers 604
16.2.8 Recurrent layers 607
16.2.9 Multiplicative layers 607
16.2.10 Implicit layers 608
16.3 Canonical examples of neural networks 608
16.3.1 Multi-layer perceptrons (MLP) 609
16.3.2 Convolutional neural networks (CNN) 609
16.3.3 Autoencoders 609
16.3.4 Recurrent neural networks (RNN) 611
16.3.5 Transformers 612
16.3.6 Graph neural networks (GNNs) 614

17 Bayesian neural networks 621

17.1 Introduction 621
17.2 Priors for BNNs 621
17.2.1 Gaussian priors 621
17.2.2 Sparsity-promoting priors 624
17.2.3 Learning the prior 624

17.2.4 Priors in function space 624
17.2.5 Architectural priors 624
17.3 Likelihoods for BNNs 625
17.4 Posteriors for BNNs 626
17.4.1 Laplace approximation 626
17.4.2 Variational inference 627
17.4.3 Expectation propagation 628
17.4.4 Last layer methods 628
17.4.5 Dropout 629
17.4.6 MCMC methods 629
17.4.7 Methods based on the SGD trajectory 630
17.4.8 Deep ensembles 631
17.4.9 Approximating the posterior predictive distibution 635
17.5 Generalization in Bayesian deep learning 636
17.5.1 Sharp vs flat minima 636
17.5.2 Effective dimensionality of a model 638
17.5.3 The hypothesis space of DNNs 639
17.5.4 Double descent 640
17.5.5 A Bayesian Resolution to Double Descent 643
17.5.6 PAC-Bayes 644
17.5.7 Out-of-Distribution Generalization for BNNs 645
17.6 Online inference 647
17.6.1 Extended Kalman Filtering for DNNs 647
17.6.2 Assumed Density Filtering for DNNs 649
17.6.3 Sequential Laplace for DNNs 651
17.6.4 Variational methods 651
17.7 Hierarchical Bayesian neural networks 652
17.7.1 Solving multiple related classification problems 652

18 Gaussian processes 655

18.1 Introduction 655
18.1.1 GPs: What and why? 655
18.2 Mercer kernels 657
18.2.1 Some popular Mercer kernels 658
18.2.2 Mercer’s theorem 664
18.2.3 Kernels from Spectral Densities 665
18.3 GPs with Gaussian likelihoods 666
18.3.1 Predictions using noise-free observations 666
18.3.2 Predictions using noisy observations 668
18.3.3 Weight space vs function space 669
18.3.4 Semi-parametric GPs 669
18.3.5 Marginal likelihood 670
18.3.6 Computational and numerical issues 671
18.3.7 Kernel ridge regression 671
18.4 GPs with non-Gaussian likelihoods 674
18.4.1 Binary classification 675
18.4.2 Multi-class classification 676
18.4.3 GPs for Poisson regression (Cox process) 677
18.4.4 Other likelihoods 677
18.5 Scaling GP inference to large datasets 678
18.5.1 Subset of data 678
18.5.2 Nyström approximation 679
18.5.3 Inducing point methods 680
18.5.4 Sparse variational methods 683
18.5.5 Exploiting parallelization and structure via kernel matrix multiplies 687
18.5.6 Converting a GP to a SSM 689

18.6 Learning the kernel 689
18.6.1 Empirical Bayes for the kernel parameters 690
18.6.2 Bayesian inference for the kernel parameters 692
18.6.3 Multiple kernel learning for additive kernels 693
18.6.4 Automatic search for compositional kernels 695
18.6.5 Spectral mixture kernel learning 697
18.6.6 Deep kernel learning 700
18.7 GPs and DNNs 701
18.7.1 Kernels derived from random DNNs (NN-GP) 701
18.7.2 Kernels derived from trained DNNs (neural tangent kernel) 704
18.7.3 Deep GPs 706
18.8 Gaussian processes for timeseries forecasting 711
18.8.1 Example: Mauna Loa 711

19 Beyond the iid assumption 715

19.1 Introduction 715
19.2 Distribution shift 715
19.2.1 Motivating examples 715
19.2.2 A causal view of distribution shift 717
19.2.3 Covariate shift 717
19.2.4 Domain shift 718
19.2.5 Label / prior shift 719
19.2.6 Concept shift 719
19.2.7 Manifestation shift 719
19.2.8 Selection bias 720
19.3 Training-time techniques for distribution shift 720
19.3.1 Importance weighting for covariate shift 721
19.3.2 Domain adaptation 722
19.3.3 Domain randomization 722
19.3.4 Data augmentation 723
19.3.5 Unsupervised label shift estimation 723
19.3.6 Distributionally robust optimization 724
19.4 Test-time techniques for distribution shift 724
19.4.1 Detecting shifts using two-sample testing 724
19.4.2 Detecting single out-of-distribution (OOD) inputs 724
19.4.3 Selective prediction 727
19.4.4 Open world recognition 728
19.4.5 Online adaptation 729
19.5 Learning from multiple distributions 730
19.5.1 Transfer learning 731
19.5.2 Few-shot learning 732
19.5.3 Prompt tuning 732
19.5.4 Zero-shot learning 732
19.5.5 Multi-task learning 733
19.5.6 Domain generalization 734
19.5.7 Invariant risk minimization 735
19.6 Meta-learning 736
19.6.1 Meta-learning as probabilistic inference for prediction 737
19.6.2 Gradient-based meta-learning 738
19.6.3 Metric-based few-shot learning 738
19.6.4 VERSA 738
19.6.5 Neural processes 739
19.7 Continual learning 739
19.7.1 Domain drift 739
19.7.2 Concept drift 739
19.7.3 Task incremental learning 741

19.7.4 Catastrophic forgetting 742
19.7.5 Online learning 744
19.8 Adversarial examples 745
19.8.1 Whitebox (gradient-based) attacks 747
19.8.2 Blackbox (gradient-free) attacks 747
19.8.3 Real world adversarial attacks 749
19.8.4 Defenses based on robust optimization 749
19.8.5 Why models have adversarial examples 750

IV Generation 753

20 Generative models: an overview 755

20.1 Introduction 755
20.2 Types of generative model 755
20.3 Goals of generative modeling 757
20.3.1 Generating data 757
20.3.2 Density estimation 758
20.3.3 Imputation 759
20.3.4 Structure discovery 760
20.3.5 Latent space interpolation 761
20.3.6 Representation learning 762
20.4 Evaluating generative models 762
20.4.1 Likelihood 763
20.4.2 Distances and divergences in feature space 764
20.4.3 Precision and recall metrics 765
20.4.4 Statistical tests 766
20.4.5 Challenges with using pretrained classifiers 767
20.4.6 Using model samples to train classifiers 767
20.4.7 Assessing overfitting 767
20.4.8 Human evaluation 767

21 Variational autoencoders 769

21.1 Introduction 769
21.2 VAE basics 769
21.2.1 Modeling assumptions 770
21.2.2 Evidence lower bound (ELBO) 771
21.2.3 Evaluating the ELBO 772
21.2.4 Optimizing the ELBO 772
21.2.5 Using the reparameterization trick to compute ELBO gradients 773
21.2.6 Comparison of VAEs and autoencoders 776
21.2.7 VAEs optimize in an augmented space 776
21.3 VAE generalizations 779
21.3.1 σ-VAE 779
21.3.2 β-VAE 781
21.3.3 InfoVAE 783
21.3.4 Multi-modal VAEs 785
21.3.5 VAEs with missing data 787
21.3.6 Semi-supervised VAEs 788
21.3.7 VAEs with sequential encoders/decoders 790
21.4 Avoiding posterior collapse 793
21.4.1 KL annealing 794
21.4.2 Lower bounding the rate 794
21.4.3 Free bits 794
21.4.4 Adding skip connections 795
21.4.5 Improved variational inference 795

21.4.6 Alternative objectives 795
21.5 VAEs with hierarchical structure 796
21.5.1 Bottom-up vs top-down inference 797
21.5.2 Example: Very deep VAE 798
21.5.3 Connection with autoregressive models 800
21.5.4 Variational pruning 801
21.5.5 Other optimization difficulties 801
21.6 Vector quantization VAE 802
21.6.1 Autoencoder with binary code 802
21.6.2 VQ-VAE model 803
21.6.3 Learning the prior 805
21.6.4 Hierarchical extension (VQ-VAE-2) 805
21.6.5 Discrete VAE 805
21.6.6 VQ-GAN 806
21.7 Wake-sleep algorithm 807
21.7.1 Wake phase 808
21.7.2 Sleep phase 809
21.7.3 Daydream phase 810
21.7.4 Summary of algorithm 810

22 Auto-regressive models 813

22.1 Introduction 813
22.2 Neural autoregressive density estimators (NADE) 814
22.3 Causal CNNs 814
22.3.1 1d causal CNN (Convolutional Markov models) 815
22.3.2 2d causal CNN (PixelCNN) 815
22.4 Transformer decoders 816
22.4.1 Text generation (GPT) 817
22.4.2 Music generation 817
22.4.3 Text-to-image generation (DALL-E) 818

23 Normalizing Flows 821

23.1 Introduction 821
23.1.1 Preliminaries 821
23.1.2 Example 823
23.1.3 How to train a flow model 824
23.2 Constructing Flows 825
23.2.1 Affine flows 825
23.2.2 Elementwise flows 826
23.2.3 Coupling flows 828
23.2.4 Autoregressive flows 830
23.2.5 Residual flows 835
23.2.6 Continuous-time flows 837
23.3 Applications 839
23.3.1 Density estimation 839
23.3.2 Generative Modeling 840
23.3.3 Inference 840

24 Energy-based models 843

24.1 Introduction 843
24.1.1 Example: Products of experts (PoE) 844
24.1.2 Computational difficulties 844
24.2 Maximum Likelihood Training 845
24.2.1 Gradient-based MCMC methods 846
24.2.2 Contrastive divergence 846
24.3 Score Matching (SM) 850
24.3.1 Basic score matching 850

24.3.2 Denoising Score Matching (DSM) 851
24.3.3 Sliced Score Matching (SSM) 852
24.3.4 Connection to Contrastive Divergence 853
24.3.5 Score-Based Generative Models 854
24.4 Noise Contrastive Estimation 857
24.4.1 Connection to Score Matching 858
24.5 Other Methods 859
24.5.1 Minimizing Differences/Derivatives of KL Divergences 859
24.5.2 Minimizing the Stein Discrepancy 860
24.5.3 Adversarial Training 860

25 Diffusion models 863

25.1 Variational diffusion models 863
25.1.1 Encoder 863
25.1.2 Decoder 865
25.1.3 Model fitting 867
25.1.4 Connection to DDPM 870
25.1.5 2d Example 870
25.1.6 Application to image generation 871
25.2 Conditional diffusion models 871
25.2.1 Classifier guidance 872
25.2.2 Classifier-free guidance 872
25.2.3 Conditional image generation 873
25.2.4 Other forms of conditional generation 873
25.3 Speeding up the generation process 873

26 Generative adversarial networks 877

26.1 Introduction 877
26.2 Learning by Comparison 878
26.2.1 Guiding principles 879
26.2.2 Class probability estimation 880
26.2.3 Bounds on f-divergences 883
26.2.4 Integral probability metrics 884
26.2.5 Moment matching 886
26.2.6 On density ratios and differences 887
26.3 Generative Adversarial Networks 888
26.3.1 From learning principles to loss functions 889
26.3.2 Gradient Descent 890
26.3.3 Challenges with GAN training 891
26.3.4 Improving GAN optimization 893
26.3.5 Convergence of GAN training 893
26.4 Conditional GANs 897
26.5 Inference with GANs 898
26.6 Neural architectures in GANs 899
26.6.1 The importance of discriminator architectures 899
26.6.2 Architectural inductive biases 899
26.6.3 Attention in GANs 900
26.6.4 Progressive generation 901
26.6.5 Regularization 902
26.6.6 Scaling up GAN models 903
26.7 Applications 903
26.7.1 GANs for image generation 903
26.7.2 Video generation 906
26.7.3 Audio generation 907
26.7.4 Text generation 907
26.7.5 Imitation Learning 908
26.7.6 Domain Adaptation 909

26.7.7 Design, Art and Creativity 909

V Discovery 911

27 Discovery methods: an overview 913

27.1 Introduction 913
27.2 Overview of Part V 914

28 Latent factor models 915

28.1 Introduction 915
28.2 Mixture models 915
28.2.1 Gaussian mixture models (GMMs) 916
28.2.2 Bernoulli mixture models 918
28.2.3 Gaussian scale mixtures (GSMs) 918
28.2.4 Using GMMs as a prior for inverse imaging problems 920
28.2.5 Using mixture models for classification problems 923
28.3 Factor analysis 925
28.3.1 Factor analysis: the basics 925
28.3.2 Probabilistic PCA 930
28.3.3 Mixture of factor analysers 932
28.3.4 Factor analysis models for paired data 939
28.3.5 Factor analysis with exponential family likelihoods 941
28.3.6 Factor analysis with DNN likelihoods 944
28.3.7 Factor analysis with GP likelihoods (GP-LVM) 944
28.4 LFMs with non-Gaussian priors 945
28.4.1 Non-negative matrix factorization (NMF) 945
28.4.2 Multinomial PCA 946
28.5 Topic models 948
28.5.1 Latent Dirichlet Allocation (LDA) 949
28.5.2 Correlated topic model 952
28.5.3 Dynamic topic model 953
28.5.4 LDA-HMM 954
28.6 Independent components analysis (ICA) 956
28.6.1 Noiseless ICA model 957
28.6.2 The need for non-Gaussian priors 958
28.6.3 Maximum likelihood estimation 960
28.6.4 Alternatives to MLE 960
28.6.5 Sparse coding 962
28.6.6 Nonlinear ICA 963

29 State-space models 965

29.1 Introduction 965
29.2 Hidden Markov models (HMMs) 966
29.2.1 Conditional independence properties 966
29.2.2 State transition model 966
29.2.3 Discrete likelihoods 967
29.2.4 Gaussian likelihoods 968
29.2.5 Autoregressive likelihoods 968
29.3 HMMs: Applications 970
29.3.1 Time series segmentation 970
29.3.2 Protein sequence alignment 972
29.3.3 Spelling correction 974
29.4 HMMs: parameter learning 976
29.4.1 The Baum-Welch (EM) algorithm 976
29.4.2 Parameter estimation using SGD 980
29.4.3 Parameter estimation using spectral methods 981

29.4.4 Bayesian HMM 981
29.5 HMMs: Generalizations 982
29.5.1 Hidden semi-Markov model (HSMM) 982
29.5.2 Hierarchical HMMs 984
29.5.3 Factorial HMMs 986
29.5.4 Coupled HMMs 987
29.5.5 Dynamic Bayes nets (DBN) 988
29.5.6 Changepoint detection 988
29.6 Linear dynamical systems (LDS) 991
29.6.1 Conditional independence properties 991
29.6.2 Parameterization 991
29.7 LDS: Applications 992
29.7.1 Object tracking and state estimation 992
29.7.2 Online Bayesian linear regression (recursive least squares) 993
29.7.3 Adaptive filtering 995
29.7.4 Timeseries forecasting 995
29.8 LDS: parameter learning 995
29.8.1 EM for LDS 995
29.8.2 Subspace identification methods 997
29.8.3 Ensuring stability of the dynamical system 998
29.8.4 Bayesian LDS 998
29.9 Switching linear dynamical systems (SLDS) 998
29.9.1 Parameterization 999
29.9.2 Posterior inference 999
29.9.3 Application: Multi-target tracking 1000
29.10 Nonlinear SSMs 1003
29.10.1 Example: object tracking and state estimation 1003
29.10.2 Posterior inference 1004
29.11 Non-Gaussian SSMs 1004
29.11.1 Example: Spike train modeling 1005
29.11.2 Example: Stochastic volatility models 1006
29.11.3 Posterior inference 1006
29.12 Structural time series models 1007
29.12.1 Basics 1007
29.12.2 Details 1008
29.12.3 Example: forecasting CO_2 levels from Mauna Loa 1010
29.12.4 Example: forecasting (real-valued) electricity usage 1011
29.12.5 Example: forecasting (integer valued) sales 1011
29.12.6 Example: hierarchical model of electoral panel data 1013
29.12.7 Causal impact of a time series intervention 1014
29.12.8 Prophet 1018
29.12.9 Neural forecasting methods 1019
29.13 Deep SSMs 1020
29.13.1 Deep Markov models 1020
29.13.2 Recurrent SSM 1021
29.13.3 Improving multi-step predictions 1022
29.13.4 Variational RNNs 1023

30 Graph learning 1025

30.1 Introduction 1025
30.2 Latent variable models for graphs 1025
30.2.1 Stochastic block model 1025
30.2.2 Mixed membership stochastic block model 1027
30.2.3 Infinite relational model 1029
30.3 Graphical model structure learning 1031
30.3.1 Applications 1032

30.3.2 Methods 1033

31 Non-parametric Bayesian models 1035

31.1 Introduction 1035
31.2 Dirichlet processes 1035
31.2.1 Definition of a DP 1036
31.2.2 Stick breaking construction of the DP 1038
31.2.3 The Chinese restaurant process (CRP) 1039
31.3 Dirichlet process mixture models 1040
31.3.1 Model definition 1041
31.3.2 Fitting using collapsed Gibbs sampling 1042
31.3.3 Fitting using variational Bayes 1044
31.3.4 Other fitting algorithms 1047
31.3.5 Choosing the hyper-parameters 1048
31.4 Generalizations of the Dirichlet process 1048
31.4.1 Pitman-Yor process 1048
31.4.2 Dependent random probability measures 1049
31.5 The Indian buffet process and the Beta process 1052
31.6 Small-variance asymptotics 1055
31.7 Completely random measures 1058
31.8 Lévy processes 1060
31.9 Point processes with repulsion and reinforcement 1061
31.9.1 Poisson process 1061
31.9.2 Renewal process 1062
31.9.3 Hawkes process 1063
31.9.4 Gibbs point process 1065
31.9.5 Determinantal point process 1066

32 Representation learning (Unfinished) 1069

32.1 CLIP 1069

33 Interpretability 1071

33.1 Introduction 1071
33.1.1 The Role of Interpretability: Unknowns and Under-Specifications 1072
33.1.2 Terminology and Framework 1073
33.2 Methods for Interpretable Machine Learning 1077
33.2.1 Inherently Interpretable Models: The Model is its Explanation 1077
33.2.2 Semi-Inherently Interpretable Models: Example-Based Methods 1079
33.2.3 Post-hoc or Joint training: The Explanation gives a Partial View of the Model 1079
33.2.4 Transparency and Visualization 1083
33.3 Properties: The Abstraction Between Context and Method 1084
33.3.1 Properties of Explanations from Interpretable Machine Learning 1085
33.3.2 Properties of Explanations from Cognitive Science 1087
33.4 Evaluation of Interpretable Machine Learning Models 1089
33.4.1 Computational Evaluation: Does the Method have Desired Properties? 1089
33.4.2 User Study-based Evaluation: Does the Method Help a User Perform a Target Task? 1094
33.5 Discussion: How to Think about Interpretable Machine Learning 1098

VI Action 1103

34 Decision making under uncertainty 1105

34.1 Bayesian decision theory 1105
34.1.1 Basics 1105
34.1.2 Classification 1106
34.1.3 Regression 1106
34.1.4 Structured prediction 1107

34.1.5 Fairness 1108
34.2 Decision (influence) diagrams 1108
34.2.1 Example: oil wildcatter 1108
34.2.2 Information arcs 1109
34.2.3 Value of information 1110
34.2.4 Computing the optimal policy 1111
34.3 A/B testing 1111
34.3.1 A Bayesian approach 1112
34.3.2 Example 1115
34.4 Contextual bandits 1116
34.4.1 Types of bandit 1116
34.4.2 Applications 1118
34.4.3 Exploration-exploitation tradeoff 1118
34.4.4 The optimal solution 1119
34.4.5 Upper confidence bounds (UCB) 1120
34.4.6 Thompson sampling 1122
34.4.7 Regret 1123
34.5 Markov decision problems 1125
34.5.1 Basics 1125
34.5.2 Partially observed MDPs 1126
34.5.3 Episodes and returns 1126
34.5.4 Value functions 1128
34.5.5 Optimal value functions and policies 1128
34.6 Planning in an MDP 1129
34.6.1 Value iteration 1130
34.6.2 Policy iteration 1131
34.6.3 Linear programming 1132
34.7 Active learning 1133
34.7.1 Relationship to other forms of sequential decision making 1133
34.7.2 Common heursitics 1133
34.7.3 Batch methods 1134

35 Reinforcement learning 1135

35.1 Introduction 1135
35.1.1 Overview of methods 1135
35.1.2 Value based methods 1137
35.1.3 Policy search methods 1137
35.1.4 Model-based RL 1137
35.1.5 Exploration-exploitation tradeoff 1138
35.2 Value-based RL 1140
35.2.1 Monte Carlo RL 1140
35.2.2 Temporal difference (TD) learning 1140
35.2.3 TD learning with eligibility traces 1141
35.2.4 SARSA: on-policy TD control 1142
35.2.5 Q-learning: off-policy TD control 1143
35.2.6 Deep Q-network (DQN) 1145
35.3 Policy-based RL 1146
35.3.1 The policy gradient theorem 1146
35.3.2 REINFORCE 1147
35.3.3 Actor-critic methods 1148
35.3.4 Bound optimization methods 1150
35.3.5 Deterministic policy gradient methods 1152
35.3.6 Gradient-free methods 1153
35.4 Model-based RL 1153
35.4.1 Model predictive control (MPC) 1153
35.4.2 Combining model-based and model-free 1155

35.4.3 MBRL using Gaussian processes 1155
35.4.4 MBRL using DNNs 1157
35.4.5 MBRL using latent-variable models 1157
35.4.6 Robustness to model errors 1160
35.5 Off-policy learning 1160
35.5.1 Basic techniques 1161
35.5.2 The curse of horizon 1164
35.5.3 The deadly triad 1165
35.6 Control as inference 1166
35.6.1 Maximum entropy reinforcement learning 1166
35.6.2 Other approaches 1169
35.6.3 Imitation learning 1170

36 Causality 1173

36.1 Introduction 1173
36.2 Causal Formalism 1175
36.2.1 Structural Causal Models 1175
36.2.2 Causal DAGs 1177
36.2.3 Identification 1179
36.2.4 Counterfactuals and the Causal Hierarchy 1180
36.3 Randomized Control Trials 1182
36.4 Confounder Adjustment 1183
36.4.1 Causal Estimand, Statistical Estimand, and Identification 1183
36.4.2 ATE Estimation with Observed Confounders 1186
36.4.3 Uncertainity Quantification 1191
36.4.4 Matching 1191
36.4.5 Practical Considerations and Procedures 1192
36.4.6 Summary and Practical Advice 1195
36.5 Instrumental Variable Strategies 1197
36.5.1 Additive Unobserved Confounding 1199
36.5.2 Instrument Monotonicity and Local Average Treatment Effect 1200
36.5.3 Two Stage Least Squares 1204
36.6 Difference in Differences 1204
36.6.1 Estimation 1208
36.7 Credibility Checks 1208
36.7.1 Placebo Checks 1209
36.7.2 Sensitivity Analysis to Unobserved Confounding 1209
36.8 The Do Calculus 1217
36.8.1 The three rules 1217
36.8.2 Revisiting Backdoor Adjustment 1218
36.8.3 Frontdoor Adjustment 1219
36.9 Further Reading 1221

Bibliography 1236