Contents

Preface **xxxiii**

List of Figures **xxxv**

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 $\mathbb{R}^+$ 9
2.2.4 Continuous distributions on $[0, 1]$ 12
2.2.5 The multivariate Gaussian (normal) distribution 13
2.2.6 Other multivariate distributions 15
2.3 The exponential family 16
2.3.1 Definition 17
2.3.2 Examples 18
2.3.3 Log partition function is cumulant generating function 22
2.3.4 Canonical (natural) vs mean (moment) parameters 24
2.3.5 MLE for the exponential family 25
2.3.6 Exponential dispersion family 26
2.3.7 Maximum entropy derivation of the exponential family 26
2.4 Fisher information matrix (FIM) 27
2.4.1 Definition 27
2.4.2 Equivalence between the FIM and the Hessian of the NLL 27
2.4.3 Examples 29
2.4.4 Approximating KL divergence using FIM 30
2.4.5 Fisher information matrix for exponential family 30

2.5 Transformations of random variables 31
2.5.1 Invertible transformations (bijections) 32
2.5.2 Monte Carlo approximation 32
2.5.3 Probability integral transform 33
2.6 Markov chains 33
2.6.1 Parameterization 34
2.6.2 Application: Language modeling 37
2.6.3 Parameter estimation 37
2.6.4 Stationary distribution of a Markov chain 39
2.7 Divergence measures between probability distributions 43
2.7.1 f-divergence 43
2.7.2 Integral probability metrics 45
2.7.3 Maximum mean discrepancy (MMD) 46
2.7.4 Total variation distance 48
2.7.5 Comparing distributions using binary classifiers 48

3 Bayesian statistics 51
3.1 Introduction 51
3.1.1 Frequentist statistics 51
3.1.2 Bayesian statistics 51
3.1.3 Arguments for the Bayesian approach 52
3.1.4 Arguments against the Bayesian approach 53
3.1.5 Why not just use MAP estimation? 53
3.2 Closed-form analysis using conjugate priors 58
3.2.1 The binomial model 58
3.2.2 The multinomial model 59
3.2.3 The univariate Gaussian model 60
3.2.4 The multivariate Gaussian model 65
3.2.5 Conjugate-exponential models 65
3.3 Beyond conjugate priors 68
3.3.1 Robust (heavy-tailed) priors 68
3.3.2 Priors for variance parameters 69
3.4 Noninformative priors 70
3.4.1 Maximum entropy priors 70
3.4.2 Jeffreys priors 71
3.4.3 Invariant priors 74
3.4.4 Reference priors 75
3.5 Hierarchical priors 76
3.5.1 A hierarchical binomial model 76
3.5.2 A hierarchical Gaussian model 78
3.6 Empirical Bayes 81
3.6.1 A hierarchical binomial model 82
3.6.2 A hierarchical Gaussian model 83
3.6.3 Hierarchical Bayes for n-gram smoothing 84
3.7 Model selection and evaluation 86

3.7.1 Bayesian model selection 86
3.7.2 Estimating the marginal likelihood 87
3.7.3 Connection between cross validation and marginal likelihood 88
3.7.4 Pareto-Smoothed Importance Sampling LOO estimate 89
3.7.5 Information criteria 90
3.7.6 Posterior predictive checks 92
3.7.7 Bayesian p-values 93
3.8 Bayesian decision theory 95
3.8.1 Basics 96
3.8.2 Example: classification 96
3.8.3 Regression 97
3.8.4 Structured prediction 97
3.8.5 COVID-19 99
3.8.6 Multi-stage decision problems 100

4 Probabilistic graphical models 101
4.1 Introduction 101
4.2 Directed graphical models (Bayes nets) 101
4.2.1 Representing the joint distribution 101
4.2.2 Examples 102
4.2.3 Gaussian Bayes nets 106
4.2.4 Conditional independence properties 107
4.2.5 Generation (sampling) 112
4.2.6 Inference 112
4.2.7 Learning 114
4.2.8 Plate notation 119
4.3 Undirected graphical models (Markov random fields) 122
4.3.1 Representing the joint distribution 122
4.3.2 Fully visible MRFs (Ising, Potts, Hopfield, etc) 124
4.3.3 MRFs with latent variables (Boltzmann machines, etc) 130
4.3.4 Maximum entropy models 132
4.3.5 Gaussian MRFs 134
4.3.6 Conditional independence properties 137
4.3.7 Generation (sampling) 139
4.3.8 Inference 139
4.3.9 Learning 140
4.4 Conditional random fields (CRFs) 144
4.4.1 1d CRFs 144
4.4.2 2d CRFs 147
4.4.3 Parameter estimation 150
4.4.4 Other approaches to structured prediction 151
4.5 Comparing directed and undirected PGMs 151
4.5.1 CI properties 151
4.5.2 Converting between a directed and undirected model 153
4.5.3 Conditional directed vs undirected PGMs and the label bias problem 154

4.5.4 Combining directed and undirected graphs 155
4.5.5 Comparing directed and undirected Gaussian PGMs 157
4.6 PGM extensions 159
4.6.1 Factor graphs 159
4.6.2 Probabilistic circuits 162
4.6.3 Directed relational PGMs 162
4.6.4 Undirected relational PGMs 164
4.6.5 Open-universe probability models 167
4.6.6 Programs as probability models 169
4.7 Structural causal models 169
4.7.1 Example: causal impact of education on wealth 170
4.7.2 Structural equation models 171
4.7.3 Do operator and augmented DAGs 171
4.7.4 Estimating average treatment effect using path analysis 173
4.7.5 Counterfactuals 174

5 Information theory 177
5.1 KL divergence 177
5.1.1 Desiderata 178
5.1.2 The KL divergence uniquely satisfies the desiderata 179
5.1.3 Thinking about KL 182
5.1.4 Properties of KL 184
5.1.5 KL divergence and MLE 186
5.1.6 KL divergence and Bayesian Inference 187
5.1.7 KL divergence and Exponential Families 188
5.1.8 Bregman divergence 189
5.2 Entropy 190
5.2.1 Definition 190
5.2.2 Differential entropy for continuous random variables 191
5.2.3 Typical sets 192
5.2.4 Cross entropy and perplexity 193
5.3 Mutual information 194
5.3.1 Definition 194
5.3.2 Interpretation 194
5.3.3 Data processing inequality 195
5.3.4 Sufficient Statistics 196
5.3.5 Multivariate mutual information 196
5.3.6 Variational bounds on mutual information 199
5.4 Data compression (source coding) 201
5.4.1 Lossless compression 202
5.4.2 Lossy compression and the rate-distortion tradeoff 202
5.4.3 Bits back coding 204
5.5 Error-correcting codes (channel coding) 205
5.6 The information bottleneck 206
5.6.1 Vanilla IB 206

5.6.2 Variational IB 207
5.6.3 Conditional entropy bottleneck 209

6 Optimization 211

6.1 Introduction 211
6.2 Automatic differentiation 211
6.2.1 Differentiation in functional form 211
6.2.2 Differentiating chains, circuits, and programs 216
6.3 Stochastic gradient descent 221
6.4 Natural gradient descent 222
6.4.1 Defining the natural gradient 222
6.4.2 Interpretations of NGD 223
6.4.3 Benefits of NGD 224
6.4.4 Approximating the natural gradient 225
6.4.5 Natural gradients for the exponential family 226
6.5 Gradients of stochastic functions 228
6.5.1 Minibatch approximation to finite-sum objectives 229
6.5.2 Optimizing parameters of a distribution 229
6.5.3 Score function estimator (likelihood ratio trick) 230
6.5.4 Reparameterization trick 231
6.5.5 The delta method 233
6.5.6 Gumbel softmax trick 233
6.5.7 Stochastic computation graphs 234
6.5.8 Straight-through estimator 234
6.6 Bound optimization (MM) algorithms 235
6.6.1 The general algorithm 235
6.6.2 Example: logistic regression 236
6.6.3 The EM algorithm 238
6.6.4 Example: EM for an MVN with missing data 240
6.6.5 Example: robust linear regression using Student-t likelihood 242
6.6.6 Extensions to EM 243
6.7 The Bayesian learning rule 245
6.7.1 Deriving inference algorithms from BLR 246
6.7.2 Deriving optimization algorithms from BLR 248
6.7.3 Variational optimization 251
6.8 Bayesian optimization 252
6.8.1 Sequential model-based optimization 253
6.8.2 Surrogate functions 253
6.8.3 Acquisition functions 255
6.8.4 Other issues 258
6.9 Derivative free optimization 259
6.9.1 Local search 259
6.9.2 Simulated annealing 262
6.9.3 Evolutionary algorithms 265
6.9.4 Estimation of distribution (EDA) algorithms 266

Do not distribute without permission from Kevin P. Murphy and MIT Press.

6.9.5 Cross-entropy method 269
6.9.6 Evolutionary strategies 269
6.10 Optimal Transport 270
6.10.1 Warm-up: Matching optimally two families of points 271
6.10.2 From Optimal Matchings to Kantorovich and Monge formulations 271
6.10.3 Solving optimal transport 274
6.11 Submodular optimization 279
6.11.1 Intuition, Examples, and Background 279
6.11.2 Submodular Basic Definitions 282
6.11.3 Example Submodular Functions 283
6.11.4 Submodular Optimization 285
6.11.5 Applications of Submodularity in Machine Learning and AI 290
6.11.6 Sketching, CoreSets, Distillation, and Data Subset & Feature Selection 290
6.11.7 Combinatorial Information Functions 294
6.11.8 Clustering, Data Partitioning, and Parallel Machine Learning 295
6.11.9 Active and Semi-Supervised Learning 295
6.11.10 Probabilistic Modeling 296
6.11.11 Structured Norms and Loss Functions 298
6.11.12 Conclusions 298

II Inference 299

7 Inference algorithms: an overview 301
7.1 Introduction 301
7.2 Common inference patterns 301
7.2.1 Global latents 302
7.2.2 Local latents 302
7.2.3 Global and local latents 303
7.3 Exact inference algorithms 303
7.4 Approximate inference algorithms 304
7.4.1 MAP estimation 304
7.4.2 Grid approximation 304
7.4.3 Laplace (quadratic) approximation 305
7.4.4 Variational inference 306
7.4.5 Markov Chain Monte Carlo (MCMC) 308
7.4.6 Sequential Monte Carlo 309
7.4.7 Challenging posteriors 310
7.5 Evaluating approximate inference algorithms 310

8 Message passing inference 313
8.1 Introduction 313
8.2 Belief propagation for discrete chains 314
8.2.1 Example: casino HMM 314
8.2.2 Forwards filtering 315

8.2.3 Backwards smoothing 317
8.2.4 The forwards-backwards algorithm 319
8.2.5 Time and space complexity 320
8.2.6 The Viterbi algorithm 321
8.2.7 Forwards filtering, backwards sampling 324
8.3 Belief propagation for Gaussian chains 325
8.3.1 Example: tracking SSM 325
8.3.2 The Kalman filter 327
8.3.3 The Kalman (RTS) smoother 330
8.4 Belief propagation for general chain-structured PGMs 331
8.4.1 Extended and unscented inference 331
8.4.2 Inference based on discretization 331
8.4.3 Particle filtering 332
8.4.4 Assumed density filtering 333
8.5 Belief propagation on trees 334
8.5.1 BP for polytrees 334
8.5.2 BP for undirected graphs with pairwise potentials 337
8.5.3 Max product belief propagation 338
8.6 Loopy belief propagation 340
8.6.1 Loopy BP for factor graphs 340
8.6.2 Gaussian belief propagation 341
8.6.3 Convergence 342
8.6.4 Accuracy 345
8.6.5 Connection with variational inference 345
8.6.6 Generalized belief propagation 346
8.6.7 Application: error correcting codes 346
8.6.8 Application: Affinity propagation 347
8.6.9 Emulating BP with graph neural nets 349
8.7 The variable elimination (VE) algorithm 350
8.7.1 Derivation of the algorithm 350
8.7.2 Computational complexity of VE 351
8.7.3 Computational complexity of exact inference 353
8.7.4 Drawbacks of VE 354
8.8 The junction tree algorithm (JTA) 355
8.8.1 Creating a junction tree 355
8.8.2 Message passing on a junction tree 360
8.8.3 Gaussian message passing 363
8.8.4 The generalized distributive law 365
8.9 Inference as optimization 366
8.9.1 Inference as backpropagation 367
8.9.2 Perturb and MAP 369

9 Variational inference 371
9.1 Introduction 371
9.1.1 Variational free energy 371

9.1.2 Evidence lower bound (ELBO) 372
9.2 Mean field VI 373
9.2.1 Coordinate ascent variational inference (CAVI) 373
9.2.2 Example: CAVI for the Ising model 374
9.2.3 Variational Bayes 376
9.2.4 Example: VB for a univariate Gaussian 377
9.2.5 Variational Bayes EM 380
9.2.6 Example: VBEM for a GMM 381
9.2.7 Variational message passing (VMP) 387
9.2.8 Autoconj 388
9.3 Fixed-form VI 388
9.3.1 Stochastic variational inference 388
9.3.2 Black-box variational inference 389
9.3.3 Reparameterization VI 391
9.3.4 Full-rank Gaussian VI 392
9.3.5 Low-rank Gaussian VI 393
9.3.6 Example: Full-rank vs diagonal GVI on 1d linear regression 394
9.3.7 Sparse Gaussian VI 396
9.3.8 Automatic differentiation VI 398
9.3.9 Non-Gaussian reparameterized VI 400
9.3.10 Amortized inference 402
9.3.11 Online variational inference 404
9.4 More accurate variational posteriors 407
9.4.1 Structured mean field 407
9.4.2 Hierarchical (auxiliary variable) posteriors 407
9.4.3 Normalizing flow posteriors 408
9.4.4 Implicit posteriors 409
9.4.5 Combining VI with MCMC inference 409
9.5 Lower bounds 410
9.5.1 Multi-sample ELBO (IWAE bound) 410
9.5.2 The thermodynamic variational objective (TVO) 411
9.6 Upper bounds 412
9.6.1 Minimizing the χ-divergence upper bound 413
9.6.2 Minimizing the evidence upper bound 414
9.7 Expectation propagation (EP) 414
9.7.1 Minimizing forwards vs reverse KL 414
9.7.2 EP as generalized ADF 416
9.7.3 Algorithm 416
9.7.4 Example 418
9.7.5 Optimization issues 418
9.7.6 Power EP and α-divergence 419
9.7.7 Stochastic EP 419
9.7.8 Applications 420

10 Monte Carlo inference 421

10.1 Introduction 421
10.2 Monte Carlo integration 421
10.2.1 Example: estimating π by Monte Carlo integration 422
10.2.2 Accuracy of Monte Carlo integration 422
10.3 Generating random samples from simple distributions 424
10.3.1 Sampling using the inverse cdf 424
10.3.2 Sampling from a Gaussian (Box-Muller method) 425
10.4 Rejection sampling 425
10.4.1 Basic idea 426
10.4.2 Example 427
10.4.3 Adaptive rejection sampling 427
10.4.4 Rejection sampling in high dimensions 428
10.5 Importance sampling 428
10.5.1 Direct importance sampling 429
10.5.2 Self-normalized importance sampling 429
10.5.3 Choosing the proposal 430
10.5.4 Annealed importance sampling (AIS) 431
10.6 Controlling Monte Carlo variance 432
10.6.1 Common random numbers 432
10.6.2 Rao-Blackwellisation 432
10.6.3 Control variates 433
10.6.4 Antithetic sampling 434
10.6.5 Quasi Monte Carlo (QMC) 435

11 Markov Chain Monte Carlo inference 437

11.1 Introduction 437
11.2 Metropolis Hastings algorithm 438
11.2.1 Basic idea 438
11.2.2 Why MH works 439
11.2.3 Proposal distributions 440
11.2.4 Initialization 443
11.3 Gibbs sampling 443
11.3.1 Basic idea 443
11.3.2 Gibbs sampling is a special case of MH 444
11.3.3 Example: Gibbs sampling for Ising models 444
11.3.4 Example: Gibbs sampling for Potts models 446
11.3.5 Example: Gibbs sampling for GMMs 446
11.3.6 Sampling from the full conditionals 448
11.3.7 Blocked Gibbs sampling 449
11.3.8 Collapsed Gibbs sampling 450
11.4 Auxiliary variable MCMC 452
11.4.1 Slice sampling 453
11.4.2 Swendsen Wang 454
11.5 Hamiltonian Monte Carlo (HMC) 456
11.5.1 Hamiltonian mechanics 456

11.5.2 Integrating Hamilton's equations 457
11.5.3 The HMC algorithm 458
11.5.4 Tuning HMC 459
11.5.5 Riemann Manifold HMC 460
11.5.6 Langevin Monte Carlo (MALA) 461
11.5.7 Connection between SGD and Langevin sampling 462
11.5.8 Applying HMC to constrained parameters 464
11.5.9 Speeding up HMC 465
11.6 MCMC convergence 465
11.6.1 Mixing rates of Markov chains 466
11.6.2 Practical convergence diagnostics 466
11.6.3 Improving speed of convergence 474
11.6.4 Non-centered parameterizations and Neal's funnel 474
11.7 Stochastic gradient MCMC 476
11.7.1 Stochastic Gradient Langevin Dynamics (SGLD) 476
11.7.2 Preconditionining 477
11.7.3 Reducing the variance of the gradient estimate 477
11.7.4 SG-HMC 478
11.7.5 Underdamped Langevin Dynamics 479
11.8 Reversible jump (trans-dimensional) MCMC 480
11.8.1 Basic idea 481
11.8.2 Example 482
11.8.3 Discussion 483
11.9 Annealing methods 484
11.9.1 Parallel tempering 484

12 Sequential Monte Carlo inference 485
12.1 Introduction 485
12.1.1 Problem statement 485
12.1.2 Particle filtering for state-space models 485
12.1.3 SMC samplers for static parameter estimation 487
12.2 Particle filtering 487
12.2.1 Importance sampling 487
12.2.2 Sequential importance sampling 488
12.2.3 Sequential importance sampling with resampling 489
12.2.4 Resampling methods 492
12.2.5 Adaptive resampling 494
12.2.6 Proposal distributions 495
12.2.7 Example: Robot localization 495
12.3 SMC samplers 496
12.3.1 Ingredients of an SMC sampler 497
12.3.2 Likelihood tempering (geometric path) 498
12.3.3 Data tempering 501
12.3.4 Sampling rare events and extrema 502
12.3.5 SMC-ABC and likelihood-free inference 503

12.3.6 SMC^2 503
12.4 Particle MCMC methods 504
12.4.1 Particle Marginal Metropolis Hastings 504
12.4.2 Particle Independent Metropolis Hastings 505
12.4.3 Particle Gibbs 506

III Prediction 507

13 Predictive models: an overview 509
13.1 Introduction 509
13.1.1 Types of model 509
13.1.2 Model fitting using ERM, MLE and MAP 510
13.1.3 Model fitting using Bayes, VI and generalized Bayes 511
13.2 Evaluating predictive models 512
13.2.1 Proper scoring rules 512
13.2.2 Calibration 512
13.2.3 Beyond evaluating marginal probabilities 516
13.3 Conformal prediction 519
13.3.1 Conformalizing classification 520
13.3.2 Conformalizing regression 521
13.3.3 Conformalizing Bayes 522
13.3.4 What do we do if we don't have a calibration set? 523

14 Generalized linear models 525
14.1 Introduction 525
14.1.1 Examples 525
14.1.2 GLMs with non-canonical link functions 528
14.1.3 Maximum likelihood estimation 528
14.1.4 Bayesian inference 529
14.2 Linear regression 530
14.2.1 Conjugate priors 530
14.2.2 Uninformative priors 532
14.2.3 Informative priors 534
14.2.4 Spike and slab prior 536
14.2.5 Laplace prior (Bayesian lasso) 537
14.2.6 Horseshoe prior 538
14.2.7 Automatic relevancy determination 539
14.3 Logistic regression 541
14.3.1 Binary logistic regression 542
14.3.2 Multinomial logistic regression 542
14.3.3 Priors 543
14.3.4 Posteriors 544
14.3.5 Laplace approximation 544
14.3.6 MCMC inference 547

14.3.7 Variational inference 548
14.3.8 Assumed density filtering 548
14.4 Probit regression 551
14.4.1 Latent variable interpretation 551
14.4.2 Maximum likelihood estimation 552
14.4.3 Bayesian inference 554
14.4.4 Ordinal probit regression 554
14.4.5 Multinomial probit models 555
14.5 Multi-level GLMs 555
14.5.1 Generalized linear mixed models (GLMMs) 555
14.5.2 Model fitting 556
14.5.3 Example: radon regression 556

15 Deep neural networks 561

15.1 Introduction 561
15.2 Building blocks of differentiable circuits 561
15.2.1 Linear layers 562
15.2.2 Non-linearities 562
15.2.3 Convolutional layers 563
15.2.4 Residual (skip) connections 564
15.2.5 Normalization layers 565
15.2.6 Dropout layers 565
15.2.7 Attention layers 566
15.2.8 Recurrent layers 569
15.2.9 Multiplicative layers 569
15.2.10 Implicit layers 570
15.3 Canonical examples of neural networks 570
15.3.1 Multi-layer perceptrons (MLP) 571
15.3.2 Convolutional neural networks (CNN) 571
15.3.3 Recurrent neural networks (RNN) 571
15.3.4 Transformers 573
15.3.5 Graph neural networks (GNNs) 576

16 Bayesian neural networks 583

16.1 Introduction 583
16.2 Priors for BNNs 583
16.2.1 Gaussian priors 583
16.2.2 Sparsity-promoting priors 586
16.2.3 Learning the prior 586
16.2.4 Priors in function space 586
16.2.5 Architectural priors 586
16.3 Likelihoods for BNNs 587
16.4 Posteriors for BNNs 588
16.4.1 Laplace approximation 588
16.4.2 Variational inference 589

16.4.3 Expectation propagation 590
16.4.4 Last layer methods 590
16.4.5 Dropout 591
16.4.6 MCMC methods 591
16.4.7 Methods based on the SGD trajectory 591
16.4.8 Deep ensembles 592
16.4.9 Approximating the posterior predictive distibution 597
16.5 Generalization in Bayesian deep learning 598
16.5.1 Sharp vs flat minima 598
16.5.2 Effective dimensionality of a model 599
16.5.3 The hypothesis space of DNNs 601
16.5.4 Double descent 602
16.5.5 A Bayesian Resolution to Double Descent 603
16.5.6 PAC-Bayes 605
16.5.7 Out-of-Distribution Generalization for BNNs 606
16.6 Online inference 609
16.6.1 Extended Kalman Filtering for DNNs 609
16.6.2 Assumed Density Filtering for DNNs 611
16.6.3 Sequential Laplace for DNNs 613
16.6.4 Variational methods 613
16.7 Hierarchical Bayesian neural networks 613
16.7.1 Solving multiple related classification problems 614

17 Gaussian processes 617
17.1 Introduction 617
17.1.1 GPs: What and why? 617
17.2 Mercer kernels 619
17.2.1 Some popular Mercer kernels 620
17.2.2 Mercer's theorem 626
17.2.3 Kernels from Spectral Densities 627
17.3 GPs with Gaussian likelihoods 628
17.3.1 Predictions using noise-free observations 628
17.3.2 Predictions using noisy observations 630
17.3.3 Weight space vs function space 631
17.3.4 Semi-parametric GPs 631
17.3.5 Marginal likelihood 632
17.3.6 Computational and numerical issues 633
17.3.7 Kernel ridge regression 633
17.4 GPs with non-Gaussian likelihoods 636
17.4.1 Binary classification 637
17.4.2 Multi-class classification 638
17.4.3 GPs for Poisson regression (Cox process) 639
17.4.4 Other likelihoods 639
17.5 Scaling GP inference to large datasets 640
17.5.1 Subset of data 640

17.5.2 Nyström approximation 641
17.5.3 Inducing point methods 642
17.5.4 Sparse variational methods 645
17.5.5 Exploiting parallelization and structure via kernel matrix multiplies 649
17.5.6 Converting a GP to a SSM 651
17.6 Learning the kernel 651
17.6.1 Empirical Bayes for the kernel parameters 652
17.6.2 Bayesian inference for the kernel parameters 654
17.6.3 Multiple kernel learning for additive kernels 655
17.6.4 Automatic search for compositional kernels 657
17.6.5 Spectral mixture kernel learning 659
17.6.6 Deep kernel learning 661
17.6.7 Functional kernel learning 663
17.7 GPs and DNNs 663
17.7.1 Kernels derived from random DNNs (NN-GP) 663
17.7.2 Kernels derived from trained DNNs (neural tangent kernel) 667
17.7.3 Deep GPs 669

18 Beyond the iid assumption 675
18.1 Introduction 675
18.2 Distribution shift 675
18.2.1 Motivating examples 675
18.2.2 A causal view of distribution shift 677
18.2.3 Covariate shift 677
18.2.4 Domain shift 678
18.2.5 Label / prior shift 679
18.2.6 Concept shift 679
18.2.7 Manifestation shift 679
18.2.8 Selection bias 680
18.3 Training-time techniques for distribution shift 680
18.3.1 Importance weighting for covariate shift 681
18.3.2 Domain adaptation 682
18.3.3 Domain randomization 682
18.3.4 Data augmentation 683
18.3.5 Unsupervised label shift estimation 683
18.3.6 Distributionally robust optimization 684
18.4 Test-time techniques for distribution shift 684
18.4.1 Detecting shifts using two-sample testing 684
18.4.2 Detecting single out-of-distribution (OOD) inputs 684
18.4.3 Selective prediction 687
18.4.4 Open world recognition 688
18.4.5 Online adaptation 689
18.5 Learning from multiple distributions 690
18.5.1 Transfer learning 691
18.5.2 Few-shot learning 692

18.5.3 Prompt tuning 692
18.5.4 Zero-shot learning 692
18.5.5 Multi-task learning 693
18.5.6 Domain generalization 694
18.5.7 Invariant risk minimization 695
18.6 Meta-learning 696
18.6.1 Meta-learning as probabilistic inference for prediction 697
18.6.2 Gradient-based meta-learning 698
18.6.3 Metric-based few-shot learning 698
18.6.4 VERSA 698
18.6.5 Neural processes 699
18.7 Continual learning 699
18.7.1 Domain drift 699
18.7.2 Concept drift 699
18.7.3 Task incremental learning 701
18.7.4 Catastrophic forgetting 702
18.7.5 Online learning 704
18.8 Adversarial examples 705
18.8.1 Whitebox (gradient-based) attacks 707
18.8.2 Blackbox (gradient-free) attacks 707
18.8.3 Real world adversarial attacks 709
18.8.4 Defenses based on robust optimization 709
18.8.5 Why models have adversarial examples 710

IV Generation 713

19 Generative models: an overview 715

19.1 Introduction 715
19.2 Types of generative model 715
19.3 Goals of generative modeling 717
19.3.1 Generating data 717
19.3.2 Density estimation 718
19.3.3 Imputation 719
19.3.4 Structure discovery 719
19.3.5 Latent space interpolation 720
19.3.6 Representation learning 721
19.4 Evaluating generative models 721
19.4.1 Likelihood 722
19.4.2 Distances and divergences in feature space 724
19.4.3 Precision and recall metrics 725
19.4.4 Statistical tests 726
19.4.5 Challenges with using pretrained classifiers 726
19.4.6 Using model samples to train classifiers 726
19.4.7 Assessing overfitting 727

19.4.8 Human evaluation 727

20 Variational autoencoders 729

20.1 Introduction 729
20.2 VAE basics 729
20.2.1 Modeling assumptions 730
20.2.2 Evidence lower bound 731
20.2.3 Optimization 732
20.2.4 The reparameterization trick 732
20.2.5 Computing the reparameterized ELBO 734
20.2.6 Comparison of VAEs and autoencoders 736
20.2.7 VAEs optimize in an augmented space 737
20.3 VAE generalizations 738
20.3.1 σ-VAE 738
20.3.2 β-VAE 740
20.3.3 InfoVAE 742
20.3.4 Multi-modal VAEs 744
20.3.5 VAEs with missing data 747
20.3.6 Semi-supervised VAEs 749
20.3.7 VAEs with sequential encoders/decoders 750
20.4 Avoiding posterior collapse 753
20.4.1 KL annealing 754
20.4.2 Lower bounding the rate 754
20.4.3 Free bits 754
20.4.4 Adding skip connections 755
20.4.5 Improved variational inference 755
20.4.6 Alternative objectives 755
20.4.7 Enforcing identifiability 756
20.5 VAEs with hierarchical structure 757
20.5.1 Bottom-up vs top-down inference 757
20.5.2 Example: Very deep VAE 758
20.5.3 Connection with autoregressive models 759
20.5.4 Variational pruning 761
20.5.5 Other optimization difficulties 762
20.6 Vector quantization VAE 762
20.6.1 Autoencoder with binary code 762
20.6.2 VQ-VAE model 764
20.6.3 Learning the prior 765
20.6.4 Hierarchical extension (VQ-VAE-2) 765
20.6.5 Discrete VAE 766
20.6.6 VQ-GAN 767
20.7 Wake-sleep algorithm 768
20.7.1 Wake phase 769
20.7.2 Sleep phase 770
20.7.3 Daydream phase 770

20.7.4 Summary of algorithm 771

21 Auto-regressive models 773

21.1 Introduction 773
21.2 Neural autoregressive density estimators (NADE) 774
21.3 Causal CNNs 774
21.3.1 1d causal CNN (Convolutional Markov models) 775
21.3.2 2d causal CNN (PixelCNN) 775
21.4 Transformer decoders 776
21.4.1 Text generation (GPT) 777
21.4.2 Music generation 777
21.4.3 Text-to-image generation (DALL-E) 778

22 Normalizing Flows 781

22.1 Introduction 781
22.1.1 Preliminaries 781
22.1.2 Example 783
22.1.3 How to train a flow model 784
22.2 Constructing Flows 785
22.2.1 Affine flows 785
22.2.2 Elementwise flows 786
22.2.3 Coupling flows 788
22.2.4 Autoregressive flows 790
22.2.5 Residual flows 795
22.2.6 Continuous-time flows 797
22.3 Applications 799
22.3.1 Density estimation 799
22.3.2 Generative Modeling 800
22.3.3 Inference 800

23 Energy-based models 803

23.1 Introduction 803
23.1.1 Example: Products of experts (PoE) 804
23.1.2 Computational difficulties 804
23.2 Maximum Likelihood Training 805
23.2.1 Gradient-based MCMC methods 806
23.2.2 Contrastive divergence 806
23.3 Score Matching (SM) 809
23.3.1 Basic score matching 810
23.3.2 Denoising Score Matching (DSM) 811
23.3.3 Sliced Score Matching (SSM) 812
23.3.4 Connection to Contrastive Divergence 813
23.3.5 Score-Based Generative Models 814
23.4 Noise Contrastive Estimation 817
23.4.1 Connection to Score Matching 818

23.5 Other Methods 819
23.5.1 Minimizing Differences/Derivatives of KL Divergences 819
23.5.2 Minimizing the Stein Discrepancy 820
23.5.3 Adversarial Training 820

24 Denoising diffusion models 823
24.1 Model definition 823
24.2 Examples 825
24.3 Model training 826
24.4 Connections with other generative models 828
24.4.1 Connection with score matching 828
24.4.2 Connection with VAEs 829
24.4.3 Connection with flow models 829

25 Generative adversarial networks 831
25.1 Introduction 831
25.2 Learning by Comparison 832
25.2.1 Guiding principles 833
25.2.2 Class probability estimation 834
25.2.3 Bounds on f-divergences 837
25.2.4 Integral probability metrics 838
25.2.5 Moment matching 840
25.2.6 On density ratios and differences 841
25.3 Generative Adversarial Networks 842
25.3.1 From learning principles to loss functions 843
25.3.2 Gradient Descent 844
25.3.3 Challenges with GAN training 845
25.3.4 Improving GAN optimization 847
25.3.5 Convergence of GAN training 847
25.4 Conditional GANs 851
25.5 Inference with GANs 852
25.6 Neural architectures in GANs 853
25.6.1 The importance of discriminator architectures 853
25.6.2 Architectural inductive biases 853
25.6.3 Attention in GANs 854
25.6.4 Progressive generation 855
25.6.5 Regularization 856
25.6.6 Scaling up GAN models 857
25.7 Applications 857
25.7.1 GANs for image generation 857
25.7.2 Video generation 859
25.7.3 Audio generation 860
25.7.4 Text generation 861
25.7.5 Imitation Learning 862
25.7.6 Domain Adaptation 862

25.7.7 Design, Art and Creativity 863

V Discovery 865

26 Discovery methods: an overview 867
26.1 Introduction 867
26.2 Overview of Part V 868

27 Latent factor models 869
27.1 Introduction 869
27.2 Mixture models 869
27.2.1 Gaussian mixture models (GMMs) 870
27.2.2 Bernoulli mixture models 872
27.2.3 Gaussian scale mixtures 872
27.2.4 Using GMMs as a prior for inverse imaging problems 874
27.3 Factor analysis 877
27.3.1 Vanilla factor analysis 877
27.3.2 Probabilistic PCA 881
27.3.3 Factor analysis models for paired data 884
27.3.4 Factor analysis with exponential family likelihoods 887
27.3.5 Factor analysis with DNN likelihoods 888
27.3.6 Factor analysis with GP likelihoods (GP-LVM) 889
27.4 Mixture of factor analysers 891
27.4.1 Model definition 891
27.4.2 Model fitting 892
27.4.3 MixFA for image generation 893
27.5 LFMs with non-Gaussian priors 897
27.5.1 Non-negative matrix factorization (NMF) 897
27.5.2 Multinomial PCA 898
27.6 Topic models 901
27.6.1 Latent Dirichlet Allocation (LDA) 901
27.6.2 Correlated topic model 904
27.6.3 Dynamic topic model 905
27.6.4 LDA-HMM 906
27.7 Independent components analysis (ICA) 908
27.7.1 Noiseless ICA model 909
27.7.2 The need for non-Gaussian priors 910
27.7.3 Maximum likelihood estimation 912
27.7.4 Alternatives to MLE 912
27.7.5 Sparse coding 914
27.7.6 Nonlinear ICA 915

28 State-space models 917
28.1 Introduction 917
28.1.1 Models 917

28.1.2 Inferential goals 918
28.2 Hidden Markov models (HMMs) 919
28.2.1 Parameterization 919
28.2.2 Application: time series segmentation 921
28.2.3 Parameter estimation 923
28.2.4 Model extensions 923
28.3 Linear dynamical systems 923
28.3.1 Example: 2d tracking problem 924
28.3.2 Example: Recursive least squares 926
28.4 Non-linear dynamical systems 929
28.4.1 Example: nonlinear 2d tracking problem 929
28.4.2 Simultaneous localization and mapping (SLAM) 929
28.4.3 Example: stochastic volatility models 932
28.4.4 Switching linear dynamical systems 932
28.4.5 Deep SSMs 934
28.5 Time series forecasting 934
28.5.1 Structural time series models 934
28.5.2 Prophet 935
28.5.3 Gaussian processes for timeseries forecasting 936
28.5.4 Neural forecasting methods 937

29 Graph learning 939
29.1 Introduction 939
29.2 Latent variable models for graphs 939
29.2.1 Stochastic block model 939
29.2.2 Mixed membership stochastic block model 941
29.2.3 Infinite relational model 943
29.3 Graphical model structure learning 945
29.3.1 Applications 945
29.3.2 Relevance networks 947
29.4 Learning tree structures 948
29.4.1 Directed or undirected tree? 948
29.4.2 Chow-Liu algorithm 949
29.4.3 Finding the MAP forest 950
29.4.4 Mixtures of trees 951
29.5 Learning DAG structures 951
29.5.1 Faithfulness 952
29.5.2 Markov equivalence 952
29.5.3 Bayesian model selection: statistical foundations 953
29.5.4 Bayesian model selection: algorithms 957
29.5.5 Constraint-based approach 959
29.5.6 Methods based on sparse optimization 961
29.5.7 Consistent estimators 962
29.5.8 Handling latent variables 962
29.6 Learning undirected graph structures 970

29.6.1 Dependency networks 971
29.6.2 Graphical lasso for GGMs 972
29.6.3 Graphical lasso for discrete MRFs/CRFs 974
29.6.4 Bayesian inference for undirected graph structures 975
29.7 Learning causal DAGs 976
29.7.1 Learning cause-effect pairs 977
29.7.2 Learning causal DAGs from interventional data 980
29.7.3 Learning from low-level inputs 981

30 Non-parametric Bayesian models 983
30.1 Introduction 983
30.2 Dirichlet processes 984
30.3 Dirichlet process 984
30.3.1 Definition 984
30.3.2 Stick breaking construction of the DP 986
30.3.3 The Chinese restaurant process (CRP) 987
30.3.4 Dirichlet process mixture models 989
30.4 Generalizations of the Dirichlet process 994
30.4.1 Pitman-Yor process 995
30.4.2 Dependent random probability measures 996
30.5 The Indian buffet process and the Beta process 998
30.6 Small-variance asymptotics 1001
30.7 Completely random measures 1004
30.8 Lévy processes 1005
30.9 Point processes with repulsion and reinforcement 1007
30.9.1 Poisson process 1007
30.9.2 Renewal process 1008
30.9.3 Hawkes process 1009
30.9.4 Gibbs point process 1011
30.9.5 Determinantal point process 1012

31 Representation learning (Unfinished) 1015
31.1 CLIP 1015

32 Interpretability 1017
32.1 Introduction 1017
32.1.1 The Role of Interpretability 1018
32.1.2 Terminology and Framework 1019
32.2 Methods for Interpretable Machine Learning 1023
32.2.1 Inherently Interpretable Models: The Model is its Explanation 1023
32.2.2 Semi-Inherently Interpretable Models: Example-Based Methods 1025
32.2.3 Post-hoc or Joint training: The Explanation gives a Partial View of the Model 1026
32.2.4 Transparency and Visualization 1030
32.3 Properties: The Abstraction Between Context and Method 1031

Do not distribute without permission from Kevin P. Murphy and MIT Press.

32.3.1 Properties of Explanations from Interpretable Machine Learning 1032
32.3.2 Properties of Explanations from Cognitive Science 1034
32.4 Evaluation of Interpretable Machine Learning Models 1035
32.4.1 Computational Evaluation: Does the Method have Desired Properties? 1036
32.4.2 User Study-based Evaluation: Does the Method Help a User Perform a Task? 1041
32.5 Discussion: How to Think about Interpretable Machine Learning 1044

VI Decision making 1051

33 Multi-step decision problems 1053
33.1 Introduction 1053
33.2 Decision (influence) diagrams 1053
33.2.1 Example: oil wildcatter 1053
33.2.2 Information arcs 1054
33.2.3 Value of information 1055
33.2.4 Computing the optimal policy 1056
33.3 A/B testing 1056
33.3.1 A Bayesian approach 1057
33.3.2 Example 1060
33.4 Contextual bandits 1061
33.4.1 Types of bandit 1061
33.4.2 Applications 1063
33.4.3 Exploration-exploitation tradeoff 1063
33.4.4 The optimal solution 1063
33.4.5 Upper confidence bounds (UCB) 1065
33.4.6 Thompson sampling 1067
33.4.7 Regret 1068
33.5 Markov decision problems 1069
33.5.1 Basics 1070
33.5.2 Partially observed MDPs 1071
33.5.3 Episodes and returns 1072
33.5.4 Value functions 1072
33.5.5 Optimal value functions and policies 1073
33.6 Planning in an MDP 1074
33.6.1 Value iteration 1075
33.6.2 Policy iteration 1076
33.6.3 Linear programming 1077

34 Reinforcement learning 1079
34.1 Introduction 1079
34.1.1 Overview of methods 1079
34.1.2 Value based methods 1081
34.1.3 Policy search methods 1081

34.1.4 Model-based RL 1081
34.1.5 Exploration-exploitation tradeoff 1082
34.2 Value-based RL 1084
34.2.1 Monte Carlo RL 1084
34.2.2 Temporal difference (TD) learning 1084
34.2.3 TD learning with eligibility traces 1085
34.2.4 SARSA: on-policy TD control 1086
34.2.5 Q-learning: off-policy TD control 1087
34.2.6 Deep Q-network (DQN) 1089
34.3 Policy-based RL 1090
34.3.1 The policy gradient theorem 1090
34.3.2 REINFORCE 1091
34.3.3 Actor-critic methods 1092
34.3.4 Bound optimization methods 1094
34.3.5 Deterministic policy gradient methods 1096
34.3.6 Gradient-free methods 1097
34.4 Model-based RL 1097
34.4.1 Model predictive control (MPC) 1097
34.4.2 Combining model-based and model-free 1099
34.4.3 MBRL using Gaussian processes 1099
34.4.4 MBRL using DNNs 1101
34.4.5 MBRL using latent-variable models 1101
34.4.6 Robustness to model errors 1104
34.5 Off-policy learning 1104
34.5.1 Basic techniques 1105
34.5.2 The curse of horizon 1108
34.5.3 The deadly triad 1109
34.6 Control as inference 1110
34.6.1 Maximum entropy reinforcement learning 1110
34.6.2 Other approaches 1113
34.6.3 Imitation learning 1114

35 Causality 1117

35.1 Introduction 1117
35.1.1 Why is causality different than other forms of ML? 1117
35.2 Causal Formalism 1119
35.2.1 Structural Causal Models 1119
35.2.2 Causal DAGs 1121
35.2.3 Identification 1123
35.2.4 Counterfactuals and the Causal Hierarchy 1124
35.3 Randomized Control Trials 1126
35.4 Confounder Adjustment 1127
35.4.1 Causal Estimand, Statistical Estimand, and Identification 1127
35.4.2 ATE Estimation with Observed Confounders 1130
35.4.3 Uncertainity Quantification 1135

Do not distribute without permission from Kevin P. Murphy and MIT Press.

35.4.4 Matching 1135
35.4.5 Practical Considerations and Procedures 1136
35.4.6 Summary and Practical Advice 1139
35.5 Instrumental Variable Strategies 1141
35.5.1 Additive Unobserved Confounding 1143
35.5.2 Instrument Monotonicity and Local Average Treatment Effect 1144
35.5.3 Two Stage Least Squares 1148
35.6 Difference in Differences 1148
35.6.1 Estimation 1152
35.7 Credibility Checks 1152
35.7.1 Placebo Checks 1153
35.7.2 Sensitivity Analysis to Unobserved Confounding 1153
35.8 The Do Calculus 1161
35.8.1 The three rules 1161
35.8.2 Revisiting Backdoor Adjustment 1162
35.8.3 Frontdoor Adjustment 1163
35.9 Further Reading 1165

Bibliography 1179