Contents

Preface xi

1 Introduction 1

I Fundamentals 3

2 Probability 5
   2.1 Introduction 5
   2.2 Some common univariate distributions 5
      2.2.1 Some common discrete distributions 5
      2.2.2 Some common continuous distributions 8
      2.2.3 Pareto distribution 14
   2.3 The multivariate Gaussian (normal) distribution 16
      2.3.1 Definition 16
      2.3.2 Moment form and canonical form 17
      2.3.3 Marginals and conditionals of a MVN 17
      2.3.4 Bayes’ rule for Gaussians 18
      2.3.5 Example: sensor fusion with known measurement noise 19
      2.3.6 Handling missing data 20
      2.3.7 A calculus for linear Gaussian models 20
   2.4 Some other multivariate continuous distributions 23
      2.4.1 Multivariate Student distribution 23
      2.4.2 Circular normal (von Mises Fisher) distribution 23
      2.4.3 Matrix-variate Gaussian (MVG) distribution 24
      2.4.4 Wishart distribution 24
      2.4.5 Dirichlet distribution 27
   2.5 The exponential family 28
      2.5.1 Definition 29
      2.5.2 Examples 30
      2.5.3 Log partition function is cumulant generating function 34
      2.5.4 Canonical (natural) vs mean (moment) parameters 36
2.5.5 MLE for the exponential family 37
2.5.6 Exponential dispersion family 38
2.5.7 Maximum entropy derivation of the exponential family 38

2.6 Fisher information matrix (FIM) 39
2.6.1 Definition 39
2.6.2 Equivalence between the FIM and the Hessian of the NLL 39
2.6.3 Examples 41
2.6.4 Approximating KL divergence using FIM 42
2.6.5 Fisher information matrix for exponential family 42

2.7 Transformations of random variables 44
2.7.1 Invertible transformations (bijectons) 44
2.7.2 Monte Carlo approximation 44
2.7.3 Probability integral transform 44

2.8 Markov chains 46
2.8.1 Parameterization 46
2.8.2 Application: Language modeling 48
2.8.3 Parameter estimation 49
2.8.4 Stationary distribution of a Markov chain 51

2.9 Divergence measures between probability distributions 54
2.9.1 f-divergence 55
2.9.2 Integral probability metrics 56
2.9.3 Maximum mean discrepancy (MMD) 57
2.9.4 Total variation distance 60
2.9.5 Comparing distributions using binary classifiers 60

3 Statistics 63
3.1 Introduction 63
3.1.1 Frequentist statistics 63
3.1.2 Bayesian statistics 63
3.1.3 Arguments for the Bayesian approach 64
3.1.4 Arguments against the Bayesian approach 64
3.1.5 Why not just use MAP estimation? 65

3.2 Closed-form analysis using conjugate priors 69
3.2.1 The binomial model 69
3.2.2 The multinomial model 77
3.2.3 The univariate Gaussian model 79
3.2.4 The multivariate Gaussian model 84
3.2.5 Conjugate-exponential models 90

3.3 Beyond conjugate priors 92
3.3.1 Robust (heavy-tailed) priors 92
3.3.2 Priors for variance parameters 93

3.4 Noninformative priors 94
3.4.1 Maximum entropy priors 94
3.4.2 Jeffreys priors 95
3.4.3 Invariant priors 98
### CONTENTS

3.4.4 Reference priors  
3.5 Hierarchical priors  
3.5.1 A hierarchical binomial model  
3.5.2 A hierarchical Gaussian model  
3.6 Empirical Bayes  
3.6.1 A hierarchical binomial model  
3.6.2 A hierarchical Gaussian model  
3.6.3 Hierarchical Bayes for n-gram smoothing  
3.7 Model selection and evaluation  
3.7.1 Bayesian model selection  
3.7.2 Estimating the marginal likelihood  
3.7.3 Connection between cross validation and marginal likelihood  
3.7.4 Pareto-Smoothed Importance Sampling LOO estimate  
3.7.5 Information criteria  
3.7.6 Posterior predictive checks  
3.7.7 Bayesian p-values  
3.8 Bayesian decision theory  
3.8.1 Basics  
3.8.2 Example: COVID-19  
3.8.3 One-shot decision problems  
3.8.4 Multi-stage decision problems  
4 Probabilistic graphical models  
4.1 Introduction  
4.2 Directed graphical models (Bayes nets)  
4.2.1 Representing the joint distribution  
4.2.2 Examples  
4.2.3 Conditional independence properties  
4.2.4 Generation (sampling)  
4.2.5 Inference  
4.2.6 Learning  
4.2.7 Plate notation  
4.3 Undirected graphical models (Markov random fields)  
4.3.1 Representing the joint distribution  
4.3.2 Examples  
4.3.3 Conditional independence properties  
4.3.4 Generation (sampling)  
4.3.5 Inference  
4.3.6 Learning  
4.4 Comparing directed and undirected PGMs  
4.4.1 CI properties  
4.4.2 Converting between a directed and undirected model  
4.4.3 Combining directed and undirected graphs  
4.4.4 Comparing directed and undirected Gaussian PGMs  
4.4.5 Factor graphs
4.5 Extensions of Bayes nets
4.5.1 Probabilistic circuits
4.5.2 Relational probability models
4.5.3 Open-universe probability models
4.5.4 Programs as probability models

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

5 Information theory
5.1 KL divergence
5.1.1 Desiderata
5.1.2 The KL divergence uniquely satisfies the desiderata
5.1.3 Thinking about KL
5.1.4 Properties of KL
5.1.5 KL divergence and MLE
5.1.6 KL divergence and Bayesian Inference
5.1.7 KL divergence and Exponential Families

5.2 Entropy
5.2.1 Definition
5.2.2 Differential entropy for continuous random variables
5.2.3 Typical sets
5.2.4 Cross entropy and perplexity

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

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

5.5 Error-correcting codes (channel coding)

5.6 The information bottleneck
5.6.1 Vanilla IB
5.6.2 Variational IB
5.6.3 Conditional entropy bottleneck

6 Optimization
6.1 Introduction
# Contents

6.2 Stochastic gradient descent 219

6.3 Natural gradient descent 220

6.3.1 Defining the natural gradient 221

6.3.2 Interpretations of NGD 221

6.3.3 Benefits of NGD 222

6.3.4 Approximating the natural gradient 223

6.3.5 Natural gradients for the exponential family 224

6.4 Mirror descent 227

6.4.1 Bregman divergence 227

6.4.2 Proximal point method 228

6.4.3 PPM using Bregman divergence 229

6.5 Gradients of stochastic functions 229

6.5.1 Minibatch approximation to finite-sum objectives 229

6.5.2 Optimizing parameters of a distribution 230

6.5.3 Score function estimator (likelihood ratio trick) 230

6.5.4 Reparameterization trick 232

6.5.5 The delta method 233

6.5.6 Gumbel softmax trick 234

6.5.7 Stochastic computation graphs 234

6.5.8 Straight-through estimator 234

6.6 Bound optimization (MM) algorithms 236

6.6.1 The general algorithm 236

6.6.2 Example: logistic regression 237

6.6.3 The EM algorithm 238

6.6.4 Example: EM for an MVN with missing data 240

6.6.5 Example: robust linear regression using Student-t likelihood 242

6.6.6 Extensions to EM 244

6.7 The Bayesian learning rule 246

6.7.1 Deriving inference algorithms from BLR 247

6.7.2 Deriving optimization algorithms from BLR 249

6.7.3 Variational optimization 252

6.8 Bayesian optimization 253

6.8.1 Sequential model-based optimization 253

6.8.2 Surrogate functions 255

6.8.3 Acquisition functions 256

6.8.4 Other issues 259

6.9 Optimal Transport 260

6.9.1 Warm-up: Matching optimally two families of points 260

6.9.2 From Optimal Matchings to Kantorovich and Monge formulations 261

6.9.3 Solving optimal transport 263

6.10 Submodular optimization 268

6.10.1 Intuition, Examples, and Background 269

6.10.2 Submodular Basic Definitions 271

6.10.3 Example Submodular Functions 272

6.10.4 Submodular Optimization 275
6.10.5 Applications of Submodularity in Machine Learning and AI  279
6.10.6 Sketching, CoreSets, Distillation, and Data Subset & Feature Selection  279
6.10.7 Combinatorial Information Functions  283
6.10.8 Clustering, Data Partitioning, and Parallel Machine Learning  284
6.10.9 Active and Semi-Supervised Learning  284
6.10.10 Probabilistic Modeling  285
6.10.11 Structured Norms and Loss Functions  287
6.10.12 Conclusions  287
6.11 Derivative free optimization  288

II Inference  289

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

8 State-space inference  301
8.1 Introduction  301
  8.1.1 State space models  301
  8.1.2 Example: casino HMM  303
  8.1.3 Example: linear-Gaussian SSM for tracking in 2d  304
  8.1.4 Inferential goals  304
8.2 Bayesian filtering and smoothing  307
  8.2.1 The filtering equations  308
  8.2.2 The smoothing equations  308
8.3 Inference for discrete SSMs  309
  8.3.1 Forwards filtering  309
  8.3.2 Backwards smoothing  311
  8.3.3 The forwards-backwards algorithm  311
  8.3.4 Two-slice smoothed marginals  313
  8.3.5 Time and space complexity  314
  8.3.6 The Viterbi algorithm  315
Draft of “Probabilistic Machine Learning: Advanced Topics” by Kevin Murphy. February 27, 2022
| 10.7.1 | Minimizing forwards vs reverse KL | 431 |
| 10.7.2 | EP as generalized ADF | 433 |
| 10.7.3 | Algorithm | 433 |
| 10.7.4 | Example | 434 |
| 10.7.5 | Optimization issues | 434 |
| 10.7.6 | Power EP and $\alpha$-divergence | 435 |
| 10.7.7 | Stochastic EP | 435 |
| 10.7.8 | Applications | 436 |

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

**12 Markov Chain Monte Carlo inference** | 453 |
<p>| 12.1 | Introduction | 453 |
| 12.2 | Metropolis Hastings algorithm | 453 |
| 12.2.1 | Basic idea | 454 |
| 12.2.2 | Why MH works | 455 |
| 12.2.3 | Proposal distributions | 456 |
| 12.2.4 | Initialization | 459 |
| 12.2.5 | Simulated annealing | 459 |
| 12.3 | Gibbs sampling | 461 |
| 12.3.1 | Basic idea | 462 |
| 12.3.2 | Gibbs sampling is a special case of MH | 462 |
| 12.3.3 | Example: Gibbs sampling for Ising models | 463 |</p>
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.3.4 Example: Gibbs sampling for Potts models</td>
<td>464</td>
</tr>
<tr>
<td>12.3.5 Example: Gibbs sampling for GMMs</td>
<td>465</td>
</tr>
<tr>
<td>12.3.6 Sampling from the full conditionals</td>
<td>466</td>
</tr>
<tr>
<td>12.3.7 Blocked Gibbs sampling</td>
<td>467</td>
</tr>
<tr>
<td>12.3.8 Collapsed Gibbs sampling</td>
<td>468</td>
</tr>
<tr>
<td>12.4 Auxiliary variable MCMC</td>
<td>470</td>
</tr>
<tr>
<td>12.4.1 Slice sampling</td>
<td>471</td>
</tr>
<tr>
<td>12.4.2 Swendsen Wang</td>
<td>472</td>
</tr>
<tr>
<td>12.5 Hamiltonian Monte Carlo (HMC)</td>
<td>474</td>
</tr>
<tr>
<td>12.5.1 Hamiltonian mechanics</td>
<td>474</td>
</tr>
<tr>
<td>12.5.2 Integrating Hamilton’s equations</td>
<td>475</td>
</tr>
<tr>
<td>12.5.3 The HMC algorithm</td>
<td>476</td>
</tr>
<tr>
<td>12.5.4 Tuning HMC</td>
<td>477</td>
</tr>
<tr>
<td>12.5.5 Riemann Manifold HMC</td>
<td>478</td>
</tr>
<tr>
<td>12.5.6 Langevin Monte Carlo (MALA)</td>
<td>479</td>
</tr>
<tr>
<td>12.5.7 Connection between SGD and Langevin sampling</td>
<td>480</td>
</tr>
<tr>
<td>12.5.8 Applying HMC to constrained parameters</td>
<td>482</td>
</tr>
<tr>
<td>12.5.9 Speeding up HMC</td>
<td>482</td>
</tr>
<tr>
<td>12.6 MCMC convergence</td>
<td>483</td>
</tr>
<tr>
<td>12.6.1 Mixing rates of Markov chains</td>
<td>484</td>
</tr>
<tr>
<td>12.6.2 Practical convergence diagnostics</td>
<td>484</td>
</tr>
<tr>
<td>12.6.3 Improving speed of convergence</td>
<td>491</td>
</tr>
<tr>
<td>12.6.4 Non-centered parameterizations and Neal’s funnel</td>
<td>492</td>
</tr>
<tr>
<td>12.7 Stochastic gradient MCMC</td>
<td>493</td>
</tr>
<tr>
<td>12.7.1 Stochastic Gradient Langevin Dynamics (SGLD)</td>
<td>494</td>
</tr>
<tr>
<td>12.7.2 Preconditioning</td>
<td>494</td>
</tr>
<tr>
<td>12.7.3 Reducing the variance of the gradient estimate</td>
<td>495</td>
</tr>
<tr>
<td>12.7.4 SG-HMC</td>
<td>496</td>
</tr>
<tr>
<td>12.7.5 Underdamped Langevin Dynamics</td>
<td>497</td>
</tr>
<tr>
<td>12.8 Reversible jump (trans-dimensional) MCMC</td>
<td>498</td>
</tr>
<tr>
<td>12.8.1 Basic idea</td>
<td>498</td>
</tr>
<tr>
<td>12.8.2 Example</td>
<td>500</td>
</tr>
<tr>
<td>12.8.3 Discussion</td>
<td>501</td>
</tr>
<tr>
<td>12.9 Annealing methods</td>
<td>501</td>
</tr>
<tr>
<td>12.9.1 Parallel tempering</td>
<td>501</td>
</tr>
<tr>
<td>13 Sequential Monte Carlo inference</td>
<td>503</td>
</tr>
<tr>
<td>13.1 Introduction</td>
<td>503</td>
</tr>
<tr>
<td>13.1.1 Problem statement</td>
<td>503</td>
</tr>
<tr>
<td>13.1.2 Particle filtering for state-space models</td>
<td>503</td>
</tr>
<tr>
<td>13.1.3 SMC samplers for static parameter estimation</td>
<td>505</td>
</tr>
<tr>
<td>13.2 Basics of SMC</td>
<td>505</td>
</tr>
<tr>
<td>13.2.1 Importance sampling</td>
<td>505</td>
</tr>
<tr>
<td>13.2.2 Sequential importance sampling</td>
<td>506</td>
</tr>
<tr>
<td>13.2.3 Sequential importance sampling with resampling</td>
<td>507</td>
</tr>
</tbody>
</table>

Draft of “Probabilistic Machine Learning: Advanced Topics” by Kevin Murphy. February 27, 2022
III Prediction 537

14 Prediction models: an overview 539

14.1 Introduction 539

14.1.1 Types of model 539

14.1.2 Model fitting using ERM, MLE and MAP 540

14.1.3 Model fitting using Bayes, VI and generalized Bayes 541

14.2 Evaluating predictive models 542

14.2.1 Proper scoring rules 542

14.2.2 Calibration 542

14.2.3 Beyond evaluating marginal probabilities 546

14.3 Conformal prediction 549

14.3.1 Conformalizing classification 550

14.3.2 Conformalizing regression 551
14.3.3 Conformalizing Bayes 552
14.3.4 What do we do if we don’t have a calibration set? 552

15 Generalized linear models 553

15.1 Introduction 553
15.1.1 Examples 553
15.1.2 GLMs with non-canonical link functions 556
15.1.3 Maximum likelihood estimation 556
15.1.4 Bayesian inference 557

15.2 Linear regression 558
15.2.1 Conjugate priors 558
15.2.2 Uninformative priors 560
15.2.3 Informative priors 562
15.2.4 Spike and slab prior 564
15.2.5 Laplace prior (Bayesian lasso) 565
15.2.6 Horseshoe prior 566
15.2.7 Automatic relevancy determination 567

15.3 Logistic regression 569
15.3.1 Binary logistic regression 570
15.3.2 Multinomial logistic regression 570
15.3.3 Priors 571
15.3.4 Posteriors 572
15.3.5 Laplace approximation 572
15.3.6 MCMC inference 575
15.3.7 Variational inference 576

15.4 Probit regression 576
15.4.1 Latent variable interpretation 576
15.4.2 Maximum likelihood estimation 577
15.4.3 Bayesian inference 578
15.4.4 Ordinal probit regression 579
15.4.5 Multinomial probit models 580

15.5 Multi-level GLMs 580
15.5.1 Generalized linear mixed models (GLMMs) 580
15.5.2 Model fitting 581
15.5.3 Example: radon regression 581

16 Deep neural networks 585

16.1 Introduction 585
16.2 Building blocks of differentiable circuits 585
16.2.1 Linear layers 586
16.2.2 Non-linearities 586
16.2.3 Convolutional layers 587
16.2.4 Residual (skip) connections 588
16.2.5 Normalization layers 589
16.2.6 Dropout layers 589
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.2.7</td>
<td>Attention layers</td>
<td>590</td>
</tr>
<tr>
<td>16.2.8</td>
<td>Recurrent layers</td>
<td>592</td>
</tr>
<tr>
<td>16.2.9</td>
<td>Multiplicative layers</td>
<td>593</td>
</tr>
<tr>
<td>16.2.10</td>
<td>Implicit layers</td>
<td>594</td>
</tr>
<tr>
<td>16.3</td>
<td>Canonical examples of neural networks</td>
<td>594</td>
</tr>
<tr>
<td>16.3.1</td>
<td>Multi-layer perceptrons (MLP)</td>
<td>594</td>
</tr>
<tr>
<td>16.3.2</td>
<td>Convolutional neural networks (CNN)</td>
<td>595</td>
</tr>
<tr>
<td>16.3.3</td>
<td>Recurrent neural networks (RNN)</td>
<td>595</td>
</tr>
<tr>
<td>16.3.4</td>
<td>Transformers</td>
<td>597</td>
</tr>
<tr>
<td>16.3.5</td>
<td>Graph neural networks (GNNs)</td>
<td>601</td>
</tr>
<tr>
<td>16.4</td>
<td>Automatic differentiation</td>
<td>604</td>
</tr>
<tr>
<td>16.4.1</td>
<td>Differentiation in functional form</td>
<td>605</td>
</tr>
<tr>
<td></td>
<td>Differentiating chains, circuits, and programs</td>
<td>609</td>
</tr>
<tr>
<td>17</td>
<td>Bayesian neural networks</td>
<td>617</td>
</tr>
<tr>
<td>17.1</td>
<td>Introduction</td>
<td>617</td>
</tr>
<tr>
<td>17.2</td>
<td>Priors for BNNs</td>
<td>617</td>
</tr>
<tr>
<td></td>
<td>Gaussian priors</td>
<td>618</td>
</tr>
<tr>
<td></td>
<td>Sparsity-promoting priors</td>
<td>619</td>
</tr>
<tr>
<td></td>
<td>Learning the prior</td>
<td>619</td>
</tr>
<tr>
<td></td>
<td>Priors in function space</td>
<td>620</td>
</tr>
<tr>
<td></td>
<td>Architectural priors</td>
<td>620</td>
</tr>
<tr>
<td>17.3</td>
<td>Likelihoods for BNNs</td>
<td>621</td>
</tr>
<tr>
<td>17.4</td>
<td>Posteriors for BNNs</td>
<td>622</td>
</tr>
<tr>
<td></td>
<td>Laplace approximation</td>
<td>622</td>
</tr>
<tr>
<td></td>
<td>Variational inference</td>
<td>623</td>
</tr>
<tr>
<td></td>
<td>Expectation propagation</td>
<td>624</td>
</tr>
<tr>
<td></td>
<td>Last layer methods</td>
<td>624</td>
</tr>
<tr>
<td></td>
<td>Dropout</td>
<td>624</td>
</tr>
<tr>
<td></td>
<td>MCMC methods</td>
<td>625</td>
</tr>
<tr>
<td></td>
<td>Methods based on the SGD trajectory</td>
<td>625</td>
</tr>
<tr>
<td></td>
<td>Deep ensembles</td>
<td>626</td>
</tr>
<tr>
<td></td>
<td>Approximating the posterior predictive distribution</td>
<td>629</td>
</tr>
<tr>
<td>17.5</td>
<td>Generalization in Bayesian deep learning</td>
<td>630</td>
</tr>
<tr>
<td></td>
<td>Sharp vs flat minima</td>
<td>631</td>
</tr>
<tr>
<td></td>
<td>Effective dimensionality of a model</td>
<td>632</td>
</tr>
<tr>
<td></td>
<td>The hypothesis space of DNNs</td>
<td>633</td>
</tr>
<tr>
<td></td>
<td>Double descent</td>
<td>634</td>
</tr>
<tr>
<td></td>
<td>A Bayesian Resolution to Double Descent</td>
<td>637</td>
</tr>
<tr>
<td></td>
<td>PAC-Bayes</td>
<td>638</td>
</tr>
<tr>
<td></td>
<td>Out-of-Distribution Generalization for BNNs</td>
<td>639</td>
</tr>
<tr>
<td>17.6</td>
<td>Online inference</td>
<td>641</td>
</tr>
<tr>
<td></td>
<td>Extended Kalman Filtering for DNNs</td>
<td>642</td>
</tr>
<tr>
<td></td>
<td>Assumed Density Filtering for DNNs</td>
<td>644</td>
</tr>
<tr>
<td></td>
<td>Sequential Laplace for DNNs</td>
<td>645</td>
</tr>
</tbody>
</table>
17.6.4 Variational methods 646

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

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

19 Structured prediction 709
  19.1 Introduction 709
  19.2 Conditional random fields (CRFs) 709
    19.2.1 1d CRFs 709
| 1 | 19.2.2 2d CRFs | 713 |
| 2 | 19.2.3 Parameter estimation | 715 |
| 3 | 19.3 Time series forecasting | 716 |
| 4 | 19.3.1 Structural time series models | 716 |
| 5 | 19.3.2 Prophet | 722 |
| 6 | 19.3.3 Gaussian processes for timeseries forecasting | 723 |
| 7 | 19.3.4 Neural forecasting methods | 725 |
| 8 | 19.3.5 Causal impact of a time series intervention | 726 |

### 20 Beyond the iid assumption | 731

#### 20.1 Introduction | 731

#### 20.2 Distribution shift | 731

##### 20.2.1 Motivating examples | 731

##### 20.2.2 A causal view of distribution shift | 733

##### 20.2.3 Covariate shift | 734

##### 20.2.4 Domain shift | 734

##### 20.2.5 Label / prior shift | 735

##### 20.2.6 Concept shift | 735

##### 20.2.7 Manifestation shift | 735

##### 20.2.8 Selection bias | 735

#### 20.3 Training-time techniques for distribution shift | 736

##### 20.3.1 Importance weighting for covariate shift | 736

##### 20.3.2 Domain adaptation | 738

##### 20.3.3 Domain randomization | 738

##### 20.3.4 Data augmentation | 739

##### 20.3.5 Unsupervised label shift estimation | 739

##### 20.3.6 Distributionally robust optimization | 739

#### 20.4 Test-time techniques for distribution shift | 740

##### 20.4.1 Detecting shifts using two-sample testing | 740

##### 20.4.2 Detecting single out-of-distribution (OOD) inputs | 740

##### 20.4.3 Selective prediction | 743

##### 20.4.4 Open world recognition | 745

##### 20.4.5 Online adaptation | 745

#### 20.5 Learning from multiple distributions | 745

##### 20.5.1 Transfer learning | 746

##### 20.5.2 Few-shot learning | 747

##### 20.5.3 Multi-task learning | 748

##### 20.5.4 Domain generalization | 748

##### 20.5.5 Invariant risk minimization | 750

#### 20.6 Meta-learning | 751

##### 20.6.1 Meta-learning as probabilistic inference for prediction | 751

##### 20.6.2 Gradient-based meta-learning | 753

##### 20.6.3 Metric-based few-shot learning | 753

##### 20.6.4 VERSA | 753

##### 20.6.5 Neural processes | 754
CONTENTS

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

23 Auto-regressive models 827
  23.1 Introduction 827
  23.2 Neural autoregressive density estimators (NADE) 828
  23.3 Causal CNNs 828
    23.3.1 1d causal CNN (Convolutional Markov models) 829
    23.3.2 2d causal CNN (PixelCNN) 829
  23.4 Transformer decoders 830
    23.4.1 Text generation (GPT) 831
    23.4.2 Music generation 831
Draft of “Probabilistic Machine Learning: Advanced Topics” by Kevin Murphy. February 27, 2022
23.4.3 Text-to-image generation (DALL-E) 832

24 Normalizing Flows 835

24.1 Introduction 835
24.1.1 Preliminaries 835
24.1.2 Example 837
24.1.3 How to train a flow model 838

24.2 Constructing Flows 839
24.2.1 Affine flows 839
24.2.2 Elementwise flows 840
24.2.3 Coupling flows 842
24.2.4 Autoregressive flows 844
24.2.5 Residual flows 848
24.2.6 Continuous-time flows 851

24.3 Applications 852
24.3.1 Density estimation 852
24.3.2 Generative Modeling 853
24.3.3 Inference 853

25 Energy-based models 855

25.1 Introduction 855
25.1.1 Example: Products of experts (PoE) 856
25.1.2 Computational difficulties 856

25.2 Maximum Likelihood Training 857
25.2.1 Gradient-based MCMC methods 858
25.2.2 Contrastive divergence 858

25.3 Score Matching (SM) 861
25.3.1 Basic score matching 862
25.3.2 Denoising Score Matching (DSM) 862
25.3.3 Sliced Score Matching (SSM) 864
25.3.4 Connection to Contrastive Divergence 865
25.3.5 Score-Based Generative Models 866

25.4 Noise Contrastive Estimation 868
25.4.1 Connection to Score Matching 870

25.5 Other Methods 871
25.5.1 Minimizing Differences/Derivatives of KL Divergences 871
25.5.2 Minimizing the Stein Discrepancy 871
25.5.3 Adversarial Training 872

26 Denoising diffusion models 875

26.1 Model definition 875
26.2 Examples 877
26.3 Model training 878
26.4 Connections with other generative models 880
26.4.1 Connection with score matching 880
26.4.2 Connection with VAEs 881
26.4.3 Connection with flow models 881

27 Generative adversarial networks 883

27.1 Introduction 883
27.2 Learning by Comparison 884
  27.2.1 Guiding principles 885
  27.2.2 Class probability estimation 886
  27.2.3 Bounds on $f$-divergences 889
  27.2.4 Integral probability metrics 890
  27.2.5 Moment matching 892
  27.2.6 On density ratios and differences 893
27.3 Generative Adversarial Networks 894
  27.3.1 From learning principles to loss functions 895
  27.3.2 Gradient Descent 896
  27.3.3 Challenges with GAN training 897
  27.3.4 Improving GAN optimization 899
  27.3.5 Convergence of GAN training 899
27.4 Conditional GANs 902
27.5 Inference with GANs 904
27.6 Neural architectures in GANs 904
  27.6.1 The importance of discriminator architectures 905
  27.6.2 Architectural inductive biases 905
  27.6.3 Attention in GANs 905
  27.6.4 Progressive generation 906
  27.6.5 Regularization 907
  27.6.6 Scaling up GAN models 908
27.7 Applications 908
  27.7.1 GANs for image generation 909
  27.7.2 Video generation 911
  27.7.3 Audio generation 912
  27.7.4 Text generation 912
  27.7.5 Imitation Learning 913
  27.7.6 Domain Adaptation 914
  27.7.7 Design, Art and Creativity 914
29.2.1 Gaussian mixture models (GMMs) 920
29.2.2 Bernoulli mixture models 922
29.2.3 Gaussian scale mixtures 922
29.2.4 Using GMMs as a prior for inverse imaging problems 924

29.3 Factor analysis 927
29.3.1 Vanilla factor analysis 927
29.3.2 Probabilistic PCA 931
29.3.3 Factor analysis models for paired data 934
29.3.4 Factor analysis with exponential family likelihoods 937
29.3.5 Factor analysis with DNN likelihoods 938
29.3.6 Factor analysis with GP likelihoods (GP-LVM) 939

29.4 Mixture of factor analysers 941
29.4.1 Model definition 941
29.4.2 Model fitting 942
29.4.3 MixFA for image generation 943

29.5 LVMs with non-Gaussian priors 947
29.5.1 Non-negative matrix factorization (NMF) 947
29.5.2 Multinomial PCA 948
29.5.3 Latent Dirichlet Allocation (LDA) 950

29.6 Independent components analysis (ICA) 951
29.6.1 Noiseless ICA model 952
29.6.2 The need for non-Gaussian priors 952
29.6.3 Maximum likelihood estimation 953
29.6.4 Alternatives to MLE 954
29.6.5 Sparse coding 956
29.6.6 Nonlinear ICA 957

30 Hidden Markov models 959
30.1 Introduction 959
30.2 HMMs: parameterization 959
30.2.1 Transition model 959
30.2.2 Observation model 960

30.3 HMMs: Applications 963
30.3.1 Segmentation of time series data 963
30.3.2 Spelling correction 965
30.3.3 Protein sequence alignment 968

30.4 HMMs: parameter learning 969
30.4.1 The Baum-Welch (EM) algorithm 969
30.4.2 Parameter estimation using SGD 973
30.4.3 Parameter estimation using spectral methods 975
30.4.4 Bayesian parameter inference 976

30.5 HMMs: Generalizations 976
30.5.1 Hidden semi-Markov model (HSMM) 977
30.5.2 HSMMs for changepoint detection 979
30.5.3 Hierarchical HMMs 982
## CONTENTS

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

### 31 State-space models 989

#### 31.1 Introduction 989

#### 31.2 Linear dynamical systems 989

- 31.2.1 Example: Noiseless 1d spring-mass system 990
- 31.2.2 Example: Noisy 2d tracking problem 991
- 31.2.3 Example: Online linear regression 994
- 31.2.4 Example: structural time series forecasting 996
- 31.2.5 Parameter estimation 996

#### 31.3 Non-linear dynamical systems 998

- 31.3.1 Example: nonlinear 2d tracking problem 999
- 31.3.2 Example: Simultaneous localization and mapping (SLAM) 999
- 31.3.3 Example: stochastic volatility models 1001
- 31.3.4 Example: Multi-target tracking 1002

#### 31.4 Other kinds of SSM 1004

- 31.4.1 Exponential family SSM 1004
- 31.4.2 Bayesian SSM 1008
- 31.4.3 GP-SSM (Unfinished) 1008

#### 31.5 Deep state space models (Unfinished) 1008

- 31.5.1 Deep Markov models 1009
- 31.5.2 Recurrent SSM 1010
- 31.5.3 Improving multi-step predictions 1011
- 31.5.4 Variational RNNs 1012
- 31.5.5 Structured State Space Sequence model (S4) 1013

### 32 Graph learning 1017

#### 32.1 Introduction 1017

#### 32.2 Latent variable models for graphs 1017

- 32.2.1 Stochastic block model 1017
- 32.2.2 Mixed membership stochastic block model 1019
- 32.2.3 Infinite relational model 1021

#### 32.3 Graphical model structure learning 1022

- 32.3.1 Applications 1023
- 32.3.2 Relevance networks 1024
- 32.3.3 Learning sparse PGMs 1026

### 33 Non-parametric Bayesian models 1027

#### 33.1 Introduction 1027

#### 33.2 Dirichlet process 1028

- 33.2.1 Definition 1028
- 33.2.2 Stick breaking construction of the DP 1030
- 33.2.3 The Chinese restaurant process (CRP) 1031
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.4.1</td>
<td>Model predictive control (MPC)</td>
<td>1141</td>
</tr>
<tr>
<td>37.4.2</td>
<td>Combining model-based and model-free</td>
<td>1142</td>
</tr>
<tr>
<td>37.4.3</td>
<td>MBRL using Gaussian processes</td>
<td>1143</td>
</tr>
<tr>
<td>37.4.4</td>
<td>MBRL using DNNs</td>
<td>1144</td>
</tr>
<tr>
<td>37.4.5</td>
<td>MBRL using latent-variable models</td>
<td>1145</td>
</tr>
<tr>
<td>37.4.6</td>
<td>Robustness to model errors</td>
<td>1147</td>
</tr>
<tr>
<td>37.5</td>
<td>Off-policy learning</td>
<td>1147</td>
</tr>
<tr>
<td>37.5.1</td>
<td>Basic techniques</td>
<td>1148</td>
</tr>
<tr>
<td>37.5.2</td>
<td>The curse of horizon</td>
<td>1151</td>
</tr>
<tr>
<td>37.5.3</td>
<td>The deadly triad</td>
<td>1152</td>
</tr>
<tr>
<td>37.6</td>
<td>Control as inference</td>
<td>1154</td>
</tr>
<tr>
<td>37.6.1</td>
<td>Maximum entropy reinforcement learning</td>
<td>1154</td>
</tr>
<tr>
<td>37.6.2</td>
<td>Active inference</td>
<td>1156</td>
</tr>
<tr>
<td>37.6.3</td>
<td>Other approaches</td>
<td>1157</td>
</tr>
<tr>
<td>37.6.4</td>
<td>Imitation learning</td>
<td>1158</td>
</tr>
<tr>
<td>38</td>
<td>Causality</td>
<td>1161</td>
</tr>
<tr>
<td>38.1</td>
<td>Introduction</td>
<td>1161</td>
</tr>
<tr>
<td>38.1.1</td>
<td>Why is causality different than other forms of ML?</td>
<td>1161</td>
</tr>
<tr>
<td>38.2</td>
<td>Causal Formalism</td>
<td>1163</td>
</tr>
<tr>
<td>38.2.1</td>
<td>Structural Causal Models</td>
<td>1163</td>
</tr>
<tr>
<td>38.2.2</td>
<td>Causal DAGs</td>
<td>1165</td>
</tr>
<tr>
<td>38.2.3</td>
<td>Identification</td>
<td>1167</td>
</tr>
<tr>
<td>38.2.4</td>
<td>Counterfactuals and the Causal Hierarchy</td>
<td>1168</td>
</tr>
<tr>
<td>38.3</td>
<td>Randomized Control Trials</td>
<td>1170</td>
</tr>
<tr>
<td>38.4</td>
<td>Confounder Adjustment</td>
<td>1171</td>
</tr>
<tr>
<td>38.4.1</td>
<td>Causal Estimand, Statistical Estimand, and Identification</td>
<td>1171</td>
</tr>
<tr>
<td>38.4.2</td>
<td>ATE Estimation with Observed Confounders</td>
<td>1174</td>
</tr>
<tr>
<td>38.4.3</td>
<td>Uncertainty Quantification</td>
<td>1179</td>
</tr>
<tr>
<td>38.4.4</td>
<td>Matching</td>
<td>1180</td>
</tr>
<tr>
<td>38.4.5</td>
<td>Practical Considerations and Procedures</td>
<td>1181</td>
</tr>
<tr>
<td>38.4.6</td>
<td>Summary and Practical Advice</td>
<td>1184</td>
</tr>
<tr>
<td>38.5</td>
<td>Instrumental Variable Strategies</td>
<td>1185</td>
</tr>
<tr>
<td>38.5.1</td>
<td>Additive Unobserved Confounding</td>
<td>1187</td>
</tr>
<tr>
<td>38.5.2</td>
<td>Instrument Monotonicity and Local Average Treatment Effect</td>
<td>1188</td>
</tr>
<tr>
<td>38.5.3</td>
<td>Two Stage Least Squares</td>
<td>1192</td>
</tr>
<tr>
<td>38.6</td>
<td>Difference in Differences</td>
<td>1192</td>
</tr>
<tr>
<td>38.6.1</td>
<td>Estimation</td>
<td>1196</td>
</tr>
<tr>
<td>38.7</td>
<td>Credibility Checks</td>
<td>1196</td>
</tr>
<tr>
<td>38.7.1</td>
<td>Placebo Checks</td>
<td>1197</td>
</tr>
<tr>
<td>38.7.2</td>
<td>Sensitivity Analysis to Unobserved Confounding</td>
<td>1197</td>
</tr>
<tr>
<td>38.8</td>
<td>The Do Calculus</td>
<td>1205</td>
</tr>
<tr>
<td>38.8.1</td>
<td>The three rules</td>
<td>1205</td>
</tr>
<tr>
<td>38.8.2</td>
<td>Revisiting Backdoor Adjustment</td>
<td>1206</td>
</tr>
<tr>
<td>38.8.3</td>
<td>Frontdoor Adjustment</td>
<td>1207</td>
</tr>
</tbody>
</table>