

# Contents

<b>Preface</b>	<b>xxix</b>
----------------	-------------

<b>1 Introduction</b>	<b>1</b>
-----------------------	----------

## I Fundamentals 3

<b>2 Probability</b>	<b>5</b>
2.1 Introduction	5
2.1.1 Basic rules of probability	5
2.1.2 Exchangeability and de Finetti's theorem	8
2.2 Some common probability distributions	9
2.2.1 Discrete distributions	9
2.2.2 Continuous distributions on $\mathbb{R}$	11
2.2.3 Continuous distributions on $\mathbb{R}^+$	13
2.2.4 Continuous distributions on $[0, 1]$	17
2.2.5 The multivariate Gaussian (normal) distribution	18
2.2.6 Linear Gaussian systems	24
2.2.7 A general calculus for linear Gaussian systems	26
2.2.8 Some other multivariate continuous distributions	29
2.3 The exponential family	34
2.3.1 Definition	34
2.3.2 Examples	35
2.3.3 Log partition function is cumulant generating function	40
2.3.4 Canonical (natural) vs mean (moment) parameters	41
2.3.5 MLE for the exponential family	42
2.3.6 Exponential dispersion family	43
2.3.7 Maximum entropy derivation of the exponential family	43
2.4 Fisher information matrix (FIM)	44
2.4.1 Definition	44
2.4.2 Equivalence between the FIM and the Hessian of the NLL	45
2.4.3 Examples	46
2.4.4 Approximating KL divergence using FIM	47
2.4.5 Fisher information matrix for exponential family	48
2.5 Transformations of random variables	49
2.5.1 Invertible transformations (bijections)	49
2.5.2 Monte Carlo approximation	50
2.5.3 Probability integral transform	50
2.6 Markov chains	52
2.6.1 Parameterization	52

2.6.2	Application: language modeling	54
2.6.3	Parameter estimation	55
2.6.4	Stationary distribution of a Markov chain	57
2.7	Divergence measures between probability distributions	60
2.7.1	$f$ -divergence	61
2.7.2	Integral probability metrics	63
2.7.3	Maximum mean discrepancy (MMD)	63
2.7.4	Total variation distance	66
2.7.5	Density ratio estimation using binary classifiers	66
<b>3</b>	<b>Statistics</b>	<b>69</b>
3.1	Introduction	69
3.1.1	Frequentist statistics	69
3.1.2	Bayesian statistics	69
3.1.3	Arguments for and against the Bayesian approach	70
3.1.4	Why not just use MAP estimation?	70
3.2	Bayesian concept learning	75
3.2.1	Learning a discrete concept: the number game	76
3.2.2	Learning a continuous concept: the healthy levels game	81
3.3	Conjugate priors for simple models	85
3.3.1	The binomial model	85
3.3.2	The multinomial model	86
3.3.3	The univariate Gaussian model	88
3.4	Conjugate priors for the multivariate Gaussian	93
3.4.1	Posterior of $\mu$ given $\Sigma$	93
3.4.2	Posterior of $\Sigma$ given $\mu$	93
3.4.3	Posterior of $\Sigma$ and $\mu$	95
3.5	Conjugate priors for the exponential family	99
3.6	Beyond conjugate priors	101
3.6.1	Robust (heavy-tailed) priors	102
3.6.2	Priors for variance parameters	102
3.7	Noninformative priors	103
3.7.1	Maximum entropy priors	104
3.7.2	Jeffreys priors	105
3.7.3	Invariant priors	108
3.7.4	Reference priors	109
3.8	Hierarchical priors	109
3.8.1	A hierarchical binomial model	110
3.8.2	A hierarchical Gaussian model	112
3.8.3	Hierarchical conditional models	115
3.9	Empirical Bayes	116
3.9.1	EB for the hierarchical binomial model	116
3.9.2	EB for the hierarchical Gaussian model	117
3.9.3	EB for Markov models (n-gram smoothing)	118
3.10	Model selection	120
3.10.1	Bayesian model selection	121
3.10.2	Bayes model averaging	121
3.10.3	Estimating the marginal likelihood	121
3.10.4	Connection between cross validation and marginal likelihood	122
3.10.5	Conditional marginal likelihood	124
3.10.6	Bayesian leave-one-out (LOO) estimate	124
3.10.7	Information criteria	126
3.11	Model checking	128
3.11.1	Posterior predictive checks	128
3.11.2	Bayesian p-values	129
3.12	Hypothesis testing	131

1			
2	3.12.1	Frequentist approach	131
3	3.12.2	Bayesian model comparison approach	133
4	3.12.3	Bayesian estimation approach	134
5	3.12.4	Approximating nonparametric tests using the rank transform	137
6	3.12.5	Common statistical tests correspond to inference in linear models	138
7	<b>4</b>	<b>Graphical models</b>	<b>143</b>
8	4.1	Introduction	143
9	4.2	Directed graphical models (Bayes nets)	143
10	4.2.1	Representing the joint distribution	143
11	4.2.2	Examples	144
12	4.2.3	Gaussian Bayes nets	148
13	4.2.4	Conditional independence properties	149
14	4.2.5	Generation (sampling)	154
15	4.2.6	Inference	155
16	4.2.7	Learning	155
17	4.2.8	Plate notation	161
18	4.3	Undirected graphical models (Markov random fields)	164
19	4.3.1	Representing the joint distribution	165
20	4.3.2	Fully visible MRFs (Ising, Potts, Hopfield, etc)	166
21	4.3.3	MRFs with latent variables (Boltzmann machines, etc)	172
22	4.3.4	Maximum entropy models	175
23	4.3.5	Gaussian MRFs	177
24	4.3.6	Conditional independence properties	179
25	4.3.7	Generation (sampling)	181
26	4.3.8	Inference	181
27	4.3.9	Learning	182
28	4.4	Conditional random fields (CRFs)	185
29	4.4.1	1d CRFs	186
30	4.4.2	2d CRFs	189
31	4.4.3	Parameter estimation	192
32	4.4.4	Other approaches to structured prediction	193
33	4.5	Comparing directed and undirected PGMs	193
34	4.5.1	CI properties	193
35	4.5.2	Converting between a directed and undirected model	195
36	4.5.3	Conditional directed vs undirected PGMs and the label bias problem	196
37	4.5.4	Combining directed and undirected graphs	197
38	4.5.5	Comparing directed and undirected Gaussian PGMs	199
39	4.6	PGM extensions	201
40	4.6.1	Factor graphs	201
41	4.6.2	Probabilistic circuits	204
42	4.6.3	Directed relational PGMs	205
43	4.6.4	Undirected relational PGMs	207
44	4.6.5	Open-universe probability models	210
45	4.6.6	Programs as probability models	210
46	4.7	Structural causal models	211
47	4.7.1	Example: causal impact of education on wealth	212
	4.7.2	Structural equation models	213
	4.7.3	Do operator and augmented DAGs	213
	4.7.4	Counterfactuals	214
	<b>5</b>	<b>Information theory</b>	<b>217</b>
	5.1	KL divergence	217
	5.1.1	Desiderata	218
	5.1.2	The KL divergence uniquely satisfies the desiderata	219
	5.1.3	Thinking about KL	222
	5.1.4	Properties of KL	225

1				
2		5.1.5	KL divergence and MLE	228
3		5.1.6	KL divergence and Bayesian inference	228
4		5.1.7	KL divergence and exponential families	230
5	5.2	5.1.8	Bregman divergence	230
6		Entropy	232	
7		5.2.1	Definition	232
8		5.2.2	Differential entropy for continuous random variables	232
9		5.2.3	Typical sets	234
10		5.2.4	Cross entropy and perplexity	235
11	5.3	Mutual information	236	
12		5.3.1	Definition	236
13		5.3.2	Interpretation	236
14		5.3.3	Data processing inequality	237
15		5.3.4	Sufficient statistics	238
16		5.3.5	Multivariate mutual information	238
17		5.3.6	Variational bounds on mutual information	241
18		5.3.7	Relevance networks	244
19	5.4	Data compression (source coding)	245	
20		5.4.1	Lossless compression	245
21		5.4.2	Lossy compression and the rate-distortion tradeoff	245
22		5.4.3	Bits back coding	248
23	5.5	Error-correcting codes (channel coding)	249	
24	5.6	The information bottleneck	250	
25		5.6.1	Vanilla IB	250
26		5.6.2	Variational IB	251
27		5.6.3	Conditional entropy bottleneck	252
28	<b>6</b>	<b>Optimization</b>	<b>255</b>	
29	6.1	Introduction	255	
30	6.2	Automatic differentiation	255	
31		6.2.1	Differentiation in functional form	255
32		6.2.2	Differentiating chains, circuits, and programs	260
33	6.3	Stochastic gradient descent	265	
34	6.4	Natural gradient descent	266	
35		6.4.1	Defining the natural gradient	266
36		6.4.2	Interpretations of NGD	267
37		6.4.3	Benefits of NGD	268
38		6.4.4	Approximating the natural gradient	269
39		6.4.5	Natural gradients for the exponential family	270
40	6.5	Gradients of stochastic functions	272	
41		6.5.1	Minibatch approximation to finite-sum objectives	273
42		6.5.2	Optimizing parameters of a distribution	273
43		6.5.3	Score function estimator (likelihood ratio trick)	274
44		6.5.4	Reparameterization trick	275
45		6.5.5	The delta method	277
46		6.5.6	Gumbel softmax trick	277
47		6.5.7	Stochastic computation graphs	278
		6.5.8	Straight-through estimator	278
	6.6	Bound optimization (MM) algorithms	279	
		6.6.1	The general algorithm	279
		6.6.2	Example: logistic regression	280
		6.6.3	The EM algorithm	282
		6.6.4	Example: EM for an MVN with missing data	284
		6.6.5	Example: robust linear regression using Student likelihood	286
		6.6.6	Extensions to EM	287
	6.7	The Bayesian learning rule	289	

1			
2	6.7.1	Deriving inference algorithms from BLR	290
3	6.7.2	Deriving optimization algorithms from BLR	292
4	6.7.3	Variational optimization	296
5	6.8	Bayesian optimization	296
6	6.8.1	Sequential model-based optimization	297
7	6.8.2	Surrogate functions	298
8	6.8.3	Acquisition functions	299
9	6.8.4	Other issues	302
10	6.9	Derivative-free optimization	303
11	6.9.1	Local search	303
12	6.9.2	Simulated annealing	306
13	6.9.3	Evolutionary algorithms	306
14	6.9.4	Estimation of distribution (EDA) algorithms	309
15	6.9.5	Cross-entropy method	311
16	6.9.6	Evolutionary strategies	311
17	6.10	Optimal transport	312
18	6.10.1	Warm-up: matching optimally two families of points	313
19	6.10.2	From optimal matchings to Kantorovich and Monge formulations	313
20	6.10.3	Solving optimal transport	316
21	6.11	Submodular optimization	321
22	6.11.1	Intuition, examples, and background	321
23	6.11.2	Submodular basic definitions	323
24	6.11.3	Example submodular functions	325
25	6.11.4	Submodular optimization	327
26	6.11.5	Applications of submodularity in machine learning and AI	332
27	6.11.6	Sketching, coresets, distillation, and data subset and feature Selection	332
28	6.11.7	Combinatorial information functions	336
29	6.11.8	Clustering, data partitioning, and parallel machine learning	337
30	6.11.9	Active and semi-supervised learning	337
31	6.11.10	Probabilistic modeling	338
32	6.11.11	Structured norms and loss functions	340
33	6.11.12	Conclusions	340

## II Inference 341

30	7	Inference algorithms: an overview	343
31	7.1	Introduction	343
32	7.2	Common inference patterns	343
33	7.2.1	Global latents	344
34	7.2.2	Local latents	344
35	7.2.3	Global and local latents	345
36	7.3	Exact inference algorithms	345
37	7.4	Approximate inference algorithms	346
38	7.4.1	MAP estimation	346
39	7.4.2	Grid approximation	346
40	7.4.3	Laplace (quadratic) approximation	347
41	7.4.4	Variational inference	348
42	7.4.5	Markov chain Monte Carlo (MCMC)	350
43	7.4.6	Sequential Monte Carlo	351
44	7.4.7	Challenging posteriors	352
45	7.5	Evaluating approximate inference algorithms	352
46	8	Gaussian filtering and smoothing	355
47	8.1	Introduction	355
	8.1.1	Inferential goals	355
	8.1.2	Bayesian filtering equations	357

1			
2	8.1.3	Bayesian smoothing equations	358
3	8.1.4	The Gaussian ansatz	359
4	8.2	Inference for linear-Gaussian SSMs	359
5	8.2.1	Examples	360
6	8.2.2	The Kalman filter	361
7	8.2.3	The Kalman (RTS) smoother	365
8	8.2.4	Information form filtering and smoothing	368
9	8.3	Inference based on local linearization	371
10	8.3.1	Taylor series expansion	371
11	8.3.2	The extended Kalman filter (EKF)	372
12	8.3.3	The extended Kalman smoother (EKS)	375
13	8.4	Inference based on the unscented transform	375
14	8.4.1	The unscented transform	375
15	8.4.2	The unscented Kalman filter (UKF)	378
16	8.4.3	The unscented Kalman smoother (UKS)	378
17	8.5	Other variants of the Kalman filter	378
18	8.5.1	General Gaussian filtering	378
19	8.5.2	Conditional moment Gaussian filtering	381
20	8.5.3	Iterated filters and smoothers	382
21	8.5.4	Ensemble Kalman filter	384
22	8.5.5	Robust Kalman filters	385
23	8.5.6	Dual EKF	385
24	8.6	Assumed density filtering	385
25	8.6.1	Connection with Gaussian filtering	387
26	8.6.2	ADF for SLDS (Gaussian sum filter)	388
27	8.6.3	ADF for online logistic regression	389
28	8.6.4	ADF for online DNNs	392
29	8.7	Other inference methods for SSMs	392
30	8.7.1	Grid-based approximations	392
31	8.7.2	Expectation propagation	393
32	8.7.3	Variational inference	394
33	8.7.4	MCMC	394
34	8.7.5	Particle filtering	394
35	<b>9</b>	<b>Message passing algorithms</b>	<b>397</b>
36	9.1	Introduction	397
37	9.2	Belief propagation on chains	397
38	9.2.1	Hidden Markov Models	398
39	9.2.2	The forwards algorithm	399
40	9.2.3	The forwards-backwards algorithm	400
41	9.2.4	Forwards filtering backwards smoothing	403
42	9.2.5	Time and space complexity	404
43	9.2.6	The Viterbi algorithm	405
44	9.2.7	Forwards filtering backwards sampling	408
45	9.3	Belief propagation on trees	408
46	9.3.1	Directed vs undirected trees	408
47	9.3.2	Sum-product algorithm	410
	9.3.3	Max-product algorithm	411
	9.4	Loopy belief propagation	413
	9.4.1	Loopy BP for pairwise undirected graphs	414
	9.4.2	Loopy BP for factor graphs	414
	9.4.3	Gaussian belief propagation	415
	9.4.4	Convergence	417
	9.4.5	Accuracy	419
	9.4.6	Generalized belief propagation	420
	9.4.7	Convex BP	420

1			
2		9.4.8 Application: error correcting codes	420
3		9.4.9 Application: affinity propagation	422
4		9.4.10 Emulating BP with graph neural nets	423
5	9.5	The variable elimination (VE) algorithm	424
6		9.5.1 Derivation of the algorithm	424
7		9.5.2 Computational complexity of VE	426
8		9.5.3 Picking a good elimination order	428
9		9.5.4 Computational complexity of exact inference	428
10		9.5.5 Drawbacks of VE	429
11	9.6	The junction tree algorithm (JTA)	430
12	9.7	Inference as optimization	431
13		9.7.1 Inference as backpropagation	431
14		9.7.2 Perturb and MAP	432
15	<b>10 Variational inference</b>	<b>435</b>	
16	10.1	Introduction	435
17		10.1.1 Variational free energy	435
18		10.1.2 Evidence lower bound (ELBO)	436
19	10.2	Mean field VI	437
20		10.2.1 Coordinate ascent variational inference (CAVI)	437
21		10.2.2 Example: CAVI for the Ising model	439
22		10.2.3 Variational Bayes	441
23		10.2.4 Example: VB for a univariate Gaussian	442
24		10.2.5 Variational Bayes EM	445
25		10.2.6 Example: VBEM for a GMM	446
26		10.2.7 Variational message passing (VMP)	452
27		10.2.8 Autoconj	453
28	10.3	Fixed-form VI	453
29		10.3.1 Stochastic variational inference	453
30		10.3.2 Blackbox variational inference	454
31		10.3.3 Reparameterization VI	456
32		10.3.4 Gaussian VI	459
33		10.3.5 Automatic differentiation VI	460
34		10.3.6 Amortized inference	463
35	10.4	More accurate variational posteriors	464
36		10.4.1 Structured mean field	464
37		10.4.2 Hierarchical (auxiliary variable) posteriors	464
38		10.4.3 Normalizing flow posteriors	465
39		10.4.4 Implicit posteriors	465
40		10.4.5 Combining VI with MCMC inference	465
41	10.5	Tighter bounds	466
42		10.5.1 Multi-sample ELBO (IWAE bound)	466
43		10.5.2 The thermodynamic variational objective (TVO)	467
44		10.5.3 Minimizing the evidence upper bound	467
45	10.6	Wake-sleep algorithm	468
46		10.6.1 Wake phase	468
47		10.6.2 Sleep phase	469
		10.6.3 Daydream phase	470
		10.6.4 Summary of algorithm	470
	10.7	Expectation propagation (EP)	470
		10.7.1 Algorithm	471
		10.7.2 Example	473
		10.7.3 EP as generalized ADF	473
		10.7.4 Optimization issues	473
		10.7.5 Power EP and $\alpha$ -divergence	474
		10.7.6 Stochastic EP	474

1			
2	<b>11 Monte Carlo methods</b>	<b>477</b>	
3	11.1	Introduction	477
4	11.2	Monte Carlo integration	477
5		11.2.1 Example: estimating $\pi$ by Monte Carlo integration	478
6		11.2.2 Accuracy of Monte Carlo integration	478
7	11.3	Generating random samples from simple distributions	480
8		11.3.1 Sampling using the inverse cdf	480
9		11.3.2 Sampling from a Gaussian (Box-Muller method)	481
10	11.4	Rejection sampling	481
11		11.4.1 Basic idea	482
12		11.4.2 Example	483
13		11.4.3 Adaptive rejection sampling	483
14		11.4.4 Rejection sampling in high dimensions	484
15	11.5	Importance sampling	484
16		11.5.1 Direct importance sampling	485
17		11.5.2 Self-normalized importance sampling	485
18		11.5.3 Choosing the proposal	486
19		11.5.4 Annealed importance sampling (AIS)	486
20	11.6	Controlling Monte Carlo variance	488
21		11.6.1 Common random numbers	488
22		11.6.2 Rao-Blackwellization	488
23		11.6.3 Control variates	489
24		11.6.4 Antithetic sampling	490
25		11.6.5 Quasi-Monte Carlo (QMC)	491
26	<b>12 Markov chain Monte Carlo</b>	<b>493</b>	
27	12.1	Introduction	493
28	12.2	Metropolis-Hastings algorithm	494
29		12.2.1 Basic idea	494
30		12.2.2 Why MH works	495
31		12.2.3 Proposal distributions	496
32		12.2.4 Initialization	498
33	12.3	Gibbs sampling	499
34		12.3.1 Basic idea	499
35		12.3.2 Gibbs sampling is a special case of MH	499
36		12.3.3 Example: Gibbs sampling for Ising models	500
37		12.3.4 Example: Gibbs sampling for Potts models	502
38		12.3.5 Example: Gibbs sampling for GMMs	502
39		12.3.6 Metropolis within Gibbs	504
40		12.3.7 Blocked Gibbs sampling	504
41		12.3.8 Collapsed Gibbs sampling	505
42	12.4	Auxiliary variable MCMC	507
43		12.4.1 Slice sampling	507
44		12.4.2 Swendsen-Wang	509
45	12.5	Hamiltonian Monte Carlo (HMC)	510
46		12.5.1 Hamiltonian mechanics	511
47		12.5.2 Integrating Hamilton's equations	511
		12.5.3 The HMC algorithm	513
		12.5.4 Tuning HMC	514
		12.5.5 Riemann manifold HMC	515
		12.5.6 Langevin Monte Carlo (MALA)	515
		12.5.7 Connection between SGD and Langevin sampling	516
		12.5.8 Applying HMC to constrained parameters	517
		12.5.9 Speeding up HMC	518
	12.6	MCMC convergence	518
		12.6.1 Mixing rates of Markov chains	519



1			
2	12.6.2	Practical convergence diagnostics	520
3	12.6.3	Effective sample size	523
4	12.6.4	Improving speed of convergence	525
5	12.6.5	Non-centered parameterizations and Neal's funnel	526
6	12.7	Stochastic gradient MCMC	527
7	12.7.1	Stochastic gradient Langevin dynamics (SGLD)	527
8	12.7.2	Preconditioning	528
9	12.7.3	Reducing the variance of the gradient estimate	528
10	12.7.4	SG-HMC	529
11	12.7.5	Underdamped Langevin dynamics	530
12	12.8	Reversible jump (transdimensional) MCMC	530
13	12.8.1	Basic idea	530
14	12.8.2	Example	531
15	12.8.3	Discussion	533
16	12.9	Annealing methods	534
17	12.9.1	Simulated annealing	534
18	12.9.2	Parallel tempering	536
19	<b>13</b>	<b>Sequential Monte Carlo</b>	<b>537</b>
20	13.1	Introduction	537
21	13.1.1	Problem statement	537
22	13.1.2	Particle filtering for state-space models	537
23	13.1.3	SMC samplers for static parameter estimation	539
24	13.2	Particle filtering	539
25	13.2.1	Importance sampling	539
26	13.2.2	Sequential importance sampling	541
27	13.2.3	Sequential importance sampling with resampling	542
28	13.2.4	Resampling methods	545
29	13.2.5	Adaptive resampling	547
30	13.3	Proposal distributions	548
31	13.3.1	Locally optimal proposal	548
32	13.3.2	Proposals based on the extended and unscented Kalman filter	549
33	13.3.3	Proposals based on the Laplace approximation	549
34	13.3.4	Proposals based on SMC (nested SMC)	551
35	13.4	Rao-Blackwellized particle filtering (RBPF)	551
36	13.4.1	Mixture of Kalman filters	551
37	13.4.2	Example: tracking a maneuvering object	553
38	13.4.3	Example: FastSLAM	554
39	13.5	Extensions of the particle filter	557
40	13.6	SMC samplers	558
41	13.6.1	Ingredients of an SMC sampler	558
42	13.6.2	Likelihood tempering (geometric path)	559
43	13.6.3	Data tempering	561
44	13.6.4	Sampling rare events and extrema	562
45	13.6.5	SMC-ABC and likelihood-free inference	563
46	13.6.6	SMC <sup>2</sup>	564
47	13.6.7	Variational filtering SMC	564
	13.6.8	Variational smoothing SMC	564

### III Prediction 567

43	<b>14</b>	<b>Predictive models: an overview</b>	<b>569</b>
44	14.1	Introduction	569
45	14.1.1	Types of model	569
46	14.1.2	Model fitting using ERM, MLE, and MAP	570

1			
2		14.1.3 Model fitting using Bayes, VI, and generalized Bayes	571
3	14.2	Evaluating predictive models	572
4	14.2.1	Proper scoring rules	572
5	14.2.2	Calibration	572
6	14.2.3	Beyond evaluating marginal probabilities	576
7	14.3	Conformal prediction	579
8	14.3.1	Conformalizing classification	581
9	14.3.2	Conformalizing regression	581
10	<b>15</b>	<b>Generalized linear models</b>	<b>583</b>
11	15.1	Introduction	583
12	15.1.1	Some popular GLMs	583
13	15.1.2	GLMs with noncanonical link functions	586
14	15.1.3	Maximum likelihood estimation	587
15	15.1.4	Bayesian inference	587
16	15.2	Linear regression	588
17	15.2.1	Conjugate priors	588
18	15.2.2	Uninformative priors	590
19	15.2.3	Informative priors	592
20	15.2.4	Spike and slab prior	594
21	15.2.5	Laplace prior (Bayesian lasso)	595
22	15.2.6	Horseshoe prior	596
23	15.2.7	Automatic relevancy determination	597
24	15.2.8	Multivariate linear regression	600
25	15.3	Logistic regression	601
26	15.3.1	Binary logistic regression	602
27	15.3.2	Multinomial logistic regression	602
28	15.3.3	Dealing with class imbalance and the long tail	603
29	15.3.4	Parameter priors	604
30	15.3.5	Laplace approximation to the posterior	605
31	15.3.6	Approximating the posterior predictive distribution	605
32	15.3.7	MCMC inference	607
33	15.3.8	Other approximate inference methods	609
34	15.3.9	Case study: is Berkeley admissions biased against women?	609
35	15.4	Probit regression	612
36	15.4.1	Latent variable interpretation	612
37	15.4.2	Maximum likelihood estimation	613
38	15.4.3	Bayesian inference	615
39	15.4.4	Ordinal probit regression	615
40	15.4.5	Multinomial probit models	616
41	15.5	Multilevel (hierarchical) GLMs	616
42	15.5.1	Generalized linear mixed models (GLMMs)	617
43	15.5.2	Example: radon regression	618
44	<b>16</b>	<b>Deep neural networks</b>	<b>621</b>
45	16.1	Introduction	621
46	16.2	Building blocks of differentiable circuits	621
47	16.2.1	Linear layers	622
	16.2.2	Nonlinearities	622
	16.2.3	Convolutional layers	623
	16.2.4	Residual (skip) connections	624
	16.2.5	Normalization layers	625
	16.2.6	Dropout layers	625
	16.2.7	Attention layers	626
	16.2.8	Recurrent layers	628
	16.2.9	Multiplicative layers	629
	16.2.10	Implicit layers	630

1			
2	16.3	Canonical examples of neural networks	630
3	16.3.1	Multilayer perceptrons (MLPs)	630
4	16.3.2	Convolutional neural networks (CNNs)	631
5	16.3.3	Autoencoders	632
6	16.3.4	Recurrent neural networks (RNNs)	634
7	16.3.5	Transformers	634
8	16.3.6	Graph neural networks (GNNs)	635
9	<b>17</b>	<b>Bayesian neural networks</b>	<b>637</b>
10	17.1	Introduction	637
11	17.2	Priors for BNNs	637
12	17.2.1	Gaussian priors	638
13	17.2.2	Sparsity-promoting priors	640
14	17.2.3	Learning the prior	640
15	17.2.4	Priors in function space	640
16	17.2.5	Architectural priors	641
17	17.3	Posteriors for BNNs	641
18	17.3.1	Monte Carlo dropout	641
19	17.3.2	Laplace approximation	642
20	17.3.3	Variational inference	643
21	17.3.4	Expectation propagation	644
22	17.3.5	Last layer methods	644
23	17.3.6	SNGP	645
24	17.3.7	MCMC methods	645
25	17.3.8	Methods based on the SGD trajectory	646
26	17.3.9	Deep ensembles	647
27	17.3.10	Approximating the posterior predictive distribution	651
28	17.3.11	Tempered and cold posteriors	654
29	17.4	Generalization in Bayesian deep learning	655
30	17.4.1	Sharp vs flat minima	655
31	17.4.2	Mode connectivity and the loss landscape	656
32	17.4.3	Effective dimensionality of a model	657
33	17.4.4	The hypothesis space of DNNs	657
34	17.4.5	PAC-Bayes	658
35	17.4.6	Out-of-distribution generalization for BNNs	659
36	17.4.7	Model selection for BNNs	662
37	17.5	Online inference	662
38	17.5.1	Sequential Laplace for DNNs	662
39	17.5.2	Extended Kalman filtering for DNNs	663
40	17.5.3	Assumed density filtering for DNNs	665
41	17.5.4	Online variational inference for DNNs	666
42	17.6	Hierarchical Bayesian neural networks	667
43	17.6.1	Example: multimoons classification	667
44	<b>18</b>	<b>Gaussian processes</b>	<b>671</b>
45	18.1	Introduction	671
46	18.1.1	GPs: what and why?	671
47	18.2	Mercer kernels	673
	18.2.1	Stationary kernels	674
	18.2.2	Nonstationary kernels	679
	18.2.3	Kernels for nonvectorial (structured) inputs	680
	18.2.4	Making new kernels from old	680
	18.2.5	Mercer's theorem	681
	18.2.6	Approximating kernels with random features	682
	18.3	GPs with Gaussian likelihoods	683
	18.3.1	Predictions using noise-free observations	683
	18.3.2	Predictions using noisy observations	684

1				
2	18.3.3	Weight space vs function space	685	
3	18.3.4	Semiparametric GPs	686	
4	18.3.5	Marginal likelihood	687	
5	18.3.6	Computational and numerical issues	687	
6	18.3.7	Kernel ridge regression	688	
7	18.4	GPs with non-Gaussian likelihoods	691	
8	18.4.1	Binary classification	692	
9	18.4.2	Multiclass classification	693	
10	18.4.3	GPs for Poisson regression (Cox process)	694	
11	18.4.4	Other likelihoods	694	
12	18.5	Scaling GP inference to large datasets	695	
13	18.5.1	Subset of data	695	
14	18.5.2	Nyström approximation	696	
15	18.5.3	Inducing point methods	697	
16	18.5.4	Sparse variational methods	700	
17	18.5.5	Exploiting parallelization and structure via kernel matrix multiplies	704	
18	18.5.6	Converting a GP to an SSM	706	
19	18.6	Learning the kernel	707	
20	18.6.1	Empirical Bayes for the kernel parameters	707	
21	18.6.2	Bayesian inference for the kernel parameters	710	
22	18.6.3	Multiple kernel learning for additive kernels	711	
23	18.6.4	Automatic search for compositional kernels	712	
24	18.6.5	Spectral mixture kernel learning	715	
25	18.6.6	Deep kernel learning	716	
26	18.7	GPs and DNNs	718	
27	18.7.1	Kernels derived from infinitely wide DNNs (NN-GP)	719	
28	18.7.2	Neural tangent kernel (NTK)	721	
29	18.7.3	Deep GPs	721	
30	18.8	Gaussian processes for time series forecasting	722	
31	18.8.1	Example: Mauna Loa	722	
32	<b>19</b>	<b>Beyond the iid assumption</b>	<b>725</b>	
33	19.1	Introduction	725	
34	19.2	Distribution shift	725	
35	19.2.1	Motivating examples	725	
36	19.2.2	A causal view of distribution shift	727	
37	19.2.3	The four main types of distribution shift	728	
38	19.2.4	Selection bias	730	
39	19.3	Detecting distribution shifts	730	
40	19.3.1	Detecting shifts using two-sample testing	731	
41	19.3.2	Detecting single out-of-distribution (OOD) inputs	731	
42	19.3.3	Selective prediction	734	
43	19.3.4	Open set and open world recognition	735	
44	19.4	Robustness to distribution shifts	735	
45	19.4.1	Data augmentation	736	
46	19.4.2	Distributionally robust optimization	736	
47	19.5	Adapting to distribution shifts	736	
	19.5.1	Supervised adaptation using transfer learning	736	
	19.5.2	Weighted ERM for covariate shift	738	
	19.5.3	Unsupervised domain adaptation for covariate shift	739	
	19.5.4	Unsupervised techniques for label shift	740	
	19.5.5	Test-time adaptation	740	
	19.6	Learning from multiple distributions	741	
	19.6.1	Multitask learning	741	
	19.6.2	Domain generalization	742	
	19.6.3	Invariant risk minimization	744	

1			
2		19.6.4	Meta learning 745
3	19.7	Continual learning	748
4		19.7.1	Domain drift 748
5		19.7.2	Concept drift 749
6		19.7.3	Task incremental learning 750
7		19.7.4	Catastrophic forgetting 751
8		19.7.5	Online learning 753
9	19.8	Adversarial examples	754
10		19.8.1	Whitebox (gradient-based) attacks 756
11		19.8.2	Blackbox (gradient-free) attacks 757
12		19.8.3	Real world adversarial attacks 758
13		19.8.4	Defenses based on robust optimization 758
14		19.8.5	Why models have adversarial examples 759

## IV Generation 761

### 20 Generative models: an overview 763

16	20.1	Introduction	763
17	20.2	Types of generative model	763
18	20.3	Goals of generative modeling	765
19		20.3.1	Generating data 765
20		20.3.2	Density estimation 767
21		20.3.3	Imputation 768
22		20.3.4	Structure discovery 769
23		20.3.5	Latent space interpolation 769
24		20.3.6	Latent space arithmetic 771
25		20.3.7	Generative design 772
26		20.3.8	Model-based reinforcement learning 772
27		20.3.9	Representation learning 772
28		20.3.10	Data compression 772
29	20.4	Evaluating generative models	772
30		20.4.1	Likelihood-based evaluation 773
31		20.4.2	Distances and divergences in feature space 774
32		20.4.3	Precision and recall metrics 775
33		20.4.4	Statistical tests 776
34		20.4.5	Challenges with using pretrained classifiers 777
35		20.4.6	Using model samples to train classifiers 777
36		20.4.7	Assessing overfitting 777
37		20.4.8	Human evaluation 778

### 21 Variational autoencoders 779

36	21.1	Introduction	779
37	21.2	VAE basics	779
38		21.2.1	Modeling assumptions 780
39		21.2.2	Evidence lower bound (ELBO) 781
40		21.2.3	Evaluating the ELBO 782
41		21.2.4	Optimizing the ELBO 782
42		21.2.5	Using the reparameterization trick to compute ELBO gradients 783
43		21.2.6	Comparison of VAEs and autoencoders 786
44		21.2.7	VAEs optimize in an augmented space 787
45	21.3	VAE generalizations	789
46		21.3.1	$\beta$ -VAE 790
47		21.3.2	InfoVAE 791
		21.3.3	Multimodal VAEs 793
		21.3.4	VAEs with missing data 795

1				
2		21.3.5	Semisupervised VAEs	797
3		21.3.6	VAEs with sequential encoders/decoders	798
4	21.4	Avoiding posterior collapse	801	
5		21.4.1	KL annealing	802
6		21.4.2	Lower bounding the rate	802
7		21.4.3	Free bits	802
8		21.4.4	Adding skip connections	802
9		21.4.5	Improved variational inference	803
10		21.4.6	Alternative objectives	803
11	21.5	VAEs with hierarchical structure	804	
12		21.5.1	Bottom-up vs top-down inference	805
13		21.5.2	Example: very deep VAE	806
14		21.5.3	Connection with autoregressive models	807
15		21.5.4	Variational pruning	808
16		21.5.5	Other optimization difficulties	809
17	21.6	Vector quantization VAE	809	
18		21.6.1	Autoencoder with binary code	809
19		21.6.2	VQ-VAE model	810
20		21.6.3	Learning the prior	812
21		21.6.4	Hierarchical extension (VQ-VAE-2)	812
22		21.6.5	Discrete VAE	813
23		21.6.6	VQ-GAN	813
24	<b>22</b>	<b>Autoregressive models</b>	<b>815</b>	
25		22.1	Introduction	815
26		22.2	Neural autoregressive density estimators (NADE)	816
27		22.3	Causal CNNs	816
28		22.3.1	1d causal CNN (convolutional Markov models)	817
29		22.3.2	2d causal CNN (PixelCNN)	817
30	22.4	Transformers	818	
31		22.4.1	Text generation (GPT, etc.)	819
32		22.4.2	Image generation (DALL-E, etc.)	820
33		22.4.3	Other applications	821
34	<b>23</b>	<b>Normalizing flows</b>	<b>823</b>	
35		23.1	Introduction	823
36		23.1.1	Preliminaries	823
37		23.1.2	How to train a flow model	825
38	23.2	Constructing flows	826	
39		23.2.1	Affine flows	826
40		23.2.2	Elementwise flows	826
41		23.2.3	Coupling flows	829
42		23.2.4	Autoregressive flows	830
43		23.2.5	Residual flows	836
44		23.2.6	Continuous-time flows	838
45	23.3	Applications	840	
46		23.3.1	Density estimation	840
47		23.3.2	Generative modeling	840
		23.3.3	Inference	841
	<b>24</b>	<b>Energy-based models</b>	<b>843</b>	
		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

1			
2	24.3	Score matching (SM)	850
3	24.3.1	Basic score matching	850
4	24.3.2	Denoising score matching (DSM)	851
5	24.3.3	Sliced score matching (SSM)	852
6	24.3.4	Connection to contrastive divergence	853
7	24.3.5	Score-based generative models	854
8	24.4	Noise contrastive estimation	854
9	24.4.1	Connection to score matching	856
10	24.5	Other methods	856
11	24.5.1	Minimizing Differences/Derivatives of KL Divergences	857
12	24.5.2	Minimizing the Stein discrepancy	857
13	24.5.3	Adversarial training	858
14	<b>25</b>	<b>Diffusion models</b>	<b>861</b>
15	25.1	Introduction	861
16	25.2	Denoising diffusion probabilistic models (DDPMs)	861
17	25.2.1	Encoder (forwards diffusion)	862
18	25.2.2	Decoder (reverse diffusion)	863
19	25.2.3	Model fitting	864
20	25.2.4	Learning the noise schedule	865
21	25.2.5	Example: image generation	867
22	25.3	Score-based generative models (SGMs)	868
23	25.3.1	Example	868
24	25.3.2	Adding noise at multiple scales	868
25	25.3.3	Equivalence to DDPM	870
26	25.4	Continuous time models using differential equations	871
27	25.4.1	Forwards diffusion SDE	871
28	25.4.2	Forwards diffusion ODE	872
29	25.4.3	Reverse diffusion SDE	873
30	25.4.4	Reverse diffusion ODE	874
31	25.4.5	Comparison of the SDE and ODE approach	875
32	25.4.6	Example	875
33	25.5	Speeding up diffusion models	875
34	25.5.1	DDIM sampler	876
35	25.5.2	Non-Gaussian decoder networks	876
36	25.5.3	Distillation	877
37	25.5.4	Latent space diffusion	878
38	25.6	Conditional generation	879
39	25.6.1	Conditional diffusion model	879
40	25.6.2	Classifier guidance	879
41	25.6.3	Classifier-free guidance	880
42	25.6.4	Generating high resolution images	880
43	25.7	Diffusion for discrete state spaces	881
44	25.7.1	Discrete Denoising Diffusion Probabilistic Models	881
45	25.7.2	Choice of Markov transition matrices for the forward processes	882
46	25.7.3	Parameterization of the reverse process	883
47	25.7.4	Noise schedules	884
	25.7.5	Connections to other probabilistic models for discrete sequences	884
	<b>26</b>	<b>Generative adversarial networks</b>	<b>887</b>
	26.1	Introduction	887
	26.2	Learning by comparison	888
	26.2.1	Guiding principles	889
	26.2.2	Density ratio estimation using binary classifiers	890
	26.2.3	Bounds on $f$ -divergences	892
	26.2.4	Integral probability metrics	894
	26.2.5	Moment matching	896

1			
2		26.2.6 On density ratios and differences	897
3	26.3	Generative adversarial networks	898
4		26.3.1 From learning principles to loss functions	898
5		26.3.2 Gradient descent	899
6		26.3.3 Challenges with GAN training	901
7		26.3.4 Improving GAN optimization	902
8		26.3.5 Convergence of GAN training	903
9	26.4	Conditional GANs	906
10	26.5	Inference with GANs	907
11	26.6	Neural architectures in GANs	908
12		26.6.1 The importance of discriminator architectures	908
13		26.6.2 Architectural inductive biases	909
14		26.6.3 Attention in GANs	909
15		26.6.4 Progressive generation	910
16		26.6.5 Regularization	912
17		26.6.6 Scaling up GAN models	912
18	26.7	Applications	913
19		26.7.1 GANs for image generation	913
20		26.7.2 Video generation	915
21		26.7.3 Audio generation	916
22		26.7.4 Text generation	917
23		26.7.5 Imitation learning	918
24		26.7.6 Domain adaptation	918
25		26.7.7 Design, art and creativity	918
26			
27	<b>V</b>	<b>Discovery</b>	<b>921</b>
28	<b>27</b>	<b>Discovery methods: an overview</b>	<b>923</b>
29		27.1 Introduction	923
30		27.2 Overview of Part V	924
31	<b>28</b>	<b>Latent factor models</b>	<b>925</b>
32		28.1 Introduction	925
33		28.2 Mixture models	925
34		28.2.1 Gaussian mixture models (GMMs)	926
35		28.2.2 Bernoulli mixture models	928
36		28.2.3 Gaussian scale mixtures (GSMs)	928
37		28.2.4 Using GMMs as a prior for inverse imaging problems	930
38		28.2.5 Using mixture models for classification problems	933
39	28.3	Factor analysis	935
40		28.3.1 Factor analysis: the basics	935
41		28.3.2 Probabilistic PCA	940
42		28.3.3 Mixture of factor analyzers	942
43		28.3.4 Factor analysis models for paired data	949
44		28.3.5 Factor analysis with exponential family likelihoods	951
45		28.3.6 Factor analysis with DNN likelihoods (VAEs)	954
46		28.3.7 Factor analysis with GP likelihoods (GP-LVM)	954
47	28.4	LFMs with non-Gaussian priors	955
		28.4.1 Non-negative matrix factorization (NMF)	955
		28.4.2 Multinomial PCA	956
	28.5	Topic models	959
		28.5.1 Latent Dirichlet allocation (LDA)	959
		28.5.2 Correlated topic model	963
		28.5.3 Dynamic topic model	963
		28.5.4 LDA-HMM	964



1			
2	28.6	Independent components analysis (ICA)	968
3	28.6.1	Noiseless ICA model	968
4	28.6.2	The need for non-Gaussian priors	969
5	28.6.3	Maximum likelihood estimation	970
6	28.6.4	Alternatives to MLE	971
7	28.6.5	Sparse coding	972
8	28.6.6	Nonlinear ICA	973
9	<b>29</b>	<b>State-space models</b>	<b>975</b>
10	29.1	Introduction	975
11	29.2	Hidden Markov models (HMMs)	976
12	29.2.1	Conditional independence properties	976
13	29.2.2	State transition model	976
14	29.2.3	Discrete likelihoods	977
15	29.2.4	Gaussian likelihoods	978
16	29.2.5	Autoregressive likelihoods	978
17	29.2.6	Neural network likelihoods	979
18	29.3	HMMs: applications	980
19	29.3.1	Time series segmentation	980
20	29.3.2	Protein sequence alignment	982
21	29.3.3	Spelling correction	984
22	29.4	HMMs: parameter learning	986
23	29.4.1	The Baum-Welch (EM) algorithm	986
24	29.4.2	Parameter estimation using SGD	989
25	29.4.3	Parameter estimation using spectral methods	991
26	29.4.4	Bayesian HMMs	991
27	29.5	HMMs: generalizations	993
28	29.5.1	Hidden semi-Markov model (HSMM)	993
29	29.5.2	Hierarchical HMMs	996
30	29.5.3	Factorial HMMs	997
31	29.5.4	Coupled HMMs	998
32	29.5.5	Dynamic Bayes nets (DBN)	999
33	29.5.6	Changepoint detection	999
34	29.6	Linear dynamical systems (LDSs)	1002
35	29.6.1	Conditional independence properties	1003
36	29.6.2	Parameterization	1003
37	29.7	LDS: applications	1003
38	29.7.1	Object tracking and state estimation	1003
39	29.7.2	Online Bayesian linear regression (recursive least squares)	1004
40	29.7.3	Adaptive filtering	1006
41	29.7.4	Time series forecasting	1007
42	29.8	LDS: parameter learning	1007
43	29.8.1	EM for LDS	1007
44	29.8.2	Subspace identification methods	1009
45	29.8.3	Ensuring stability of the dynamical system	1010
46	29.8.4	Bayesian LDS	1010
47	29.9	Switching linear dynamical systems (SLDSs)	1011
	29.9.1	Parameterization	1011
	29.9.2	Posterior inference	1011
	29.9.3	Application: Multitarget tracking	1013
	29.10	Nonlinear SSMs	1016
	29.10.1	Example: object tracking and state estimation	1016
	29.10.2	Posterior inference	1017
	29.11	Non-Gaussian SSMs	1017
	29.11.1	Example: spike train modeling	1017
	29.11.2	Example: stochastic volatility models	1018

1			
2	29.11.3	Posterior inference	1019
3	29.12	Structural time series models	1019
4	29.12.1	Introduction	1019
5	29.12.2	Structural building blocks	1020
6	29.12.3	Model fitting	1022
7	29.12.4	Forecasting	1023
8	29.12.5	Examples	1023
9	29.12.6	Causal impact of a time series intervention	1027
10	29.12.7	Prophet	1031
11	29.12.8	Neural forecasting methods	1031
12	29.13	Deep SSMs	1032
13	29.13.1	Deep Markov models	1032
14	29.13.2	Recurrent SSM	1034
15	29.13.3	Improving multistep predictions	1035
16	29.13.4	Variational RNNs	1036
17	<b>30</b>	<b>Graph learning</b>	<b>1039</b>
18	30.1	Introduction	1039
19	30.2	Latent variable models for graphs	1039
20	30.3	Graphical model structure learning	1039
21	<b>31</b>	<b>Nonparametric Bayesian models</b>	<b>1043</b>
22	31.1	Introduction	1043
23	<b>32</b>	<b>Representation learning</b>	<b>1045</b>
24	32.1	Introduction	1045
25	32.2	Evaluating and comparing learned representations	1045
26	32.2.1	Downstream performance	1046
27	32.2.2	Representational similarity	1048
28	32.3	Approaches for learning representations	1052
29	32.3.1	Supervised representation learning and transfer	1053
30	32.3.2	Generative representation learning	1055
31	32.3.3	Self-supervised representation learning	1057
32	32.3.4	Multiview representation learning	1060
33	32.4	Theory of representation learning	1065
34	32.4.1	Identifiability	1065
35	32.4.2	Information maximization	1066
36	<b>33</b>	<b>Interpretability</b>	<b>1069</b>
37	33.1	Introduction	1069
38	33.1.1	The role of interpretability: unknowns and under-specifications	1070
39	33.1.2	Terminology and framework	1071
40	33.2	Methods for interpretable machine learning	1075
41	33.2.1	Inherently interpretable models: the model is its explanation	1075
42	33.2.2	Semi-inherently interpretable models: example-based methods	1077
43	33.2.3	Post-hoc or joint training: the explanation gives a partial view of the model	1077
44	33.2.4	Transparency and visualization	1081
45	33.3	Properties: the abstraction between context and method	1082
46	33.3.1	Properties of explanations from interpretable machine learning	1082
47	33.3.2	Properties of explanations from cognitive science	1085
48	33.4	Evaluation of interpretable machine learning models	1086
49	33.4.1	Computational evaluation: does the method have desired properties?	1086
50	33.4.2	User study-based evaluation: does the method help a user perform a target task?	1090
51	33.5	Discussion: how to think about interpretable machine learning	1094

## VI Action 1099

### 34 Decision making under uncertainty 1101

34.1	Bayesian decision theory	1101	
34.1.1	Basics	1101	
34.1.2	Classification	1102	
34.1.3	Regression	1102	
34.1.4	Structured prediction	1103	
34.1.5	Fairness	1104	
34.2	Decision (influence) diagrams	1104	
34.2.1	Example: oil wildcatter	1104	
34.2.2	Information arcs	1105	
34.2.3	Value of information	1106	
34.2.4	Computing the optimal policy	1107	
34.3	A/B testing	1107	
34.3.1	A Bayesian approach	1108	
34.3.2	Example	1111	
34.4	Contextual bandits	1112	
34.4.1	Types of bandit	1112	
34.4.2	Applications	1114	
34.4.3	Exploration-exploitation tradeoff	1114	
34.4.4	The optimal solution	1115	
34.4.5	Upper confidence bounds (UCBs)	1116	
34.4.6	Thompson sampling	1118	
34.4.7	Regret	1119	
34.5	Markov decision problems	1121	
34.5.1	Basics	1121	
34.5.2	Partially observed MDPs	1122	
34.5.3	Episodes and returns	1122	
34.5.4	Value functions	1124	
34.5.5	Optimal value functions and policies	1124	
34.6	Planning in an MDP	1125	
34.6.1	Value iteration	1126	
34.6.2	Policy iteration	1127	
34.6.3	Linear programming	1128	
34.7	Active learning	1129	
34.7.1	Active learning scenarios	1129	
34.7.2	Relationship to other forms of sequential decision making	1130	
34.7.3	Acquisition strategies	1131	
34.7.4	Batch active learning	1133	

### 35 Reinforcement learning 1137

35.1	Introduction	1137	
35.1.1	Overview of methods	1137	
35.1.2	Value-based methods	1139	
35.1.3	Policy search methods	1139	
35.1.4	Model-based RL	1139	
35.1.5	Exploration-exploitation tradeoff	1140	
35.2	Value-based RL	1142	
35.2.1	Monte Carlo RL	1142	
35.2.2	Temporal difference (TD) learning	1142	
35.2.3	TD learning with eligibility traces	1143	
35.2.4	SARSA: on-policy TD control	1144	
35.2.5	Q-learning: off-policy TD control	1145	
35.2.6	Deep Q-network (DQN)	1146	
35.3	Policy-based RL	1148	
35.3.1	The policy gradient theorem	1149	

1			
2	35.3.2	REINFORCE	1150
3	35.3.3	Actor-critic methods	1150
4	35.3.4	Bound optimization methods	1152
5	35.3.5	Deterministic policy gradient methods	1154
6	35.3.6	Gradient-free methods	1155
7	35.4	Model-based RL	1155
8	35.4.1	Model predictive control (MPC)	1155
9	35.4.2	Combining model-based and model-free	1157
10	35.4.3	MBRL using Gaussian processes	1158
11	35.4.4	MBRL using DNNs	1159
12	35.4.5	MBRL using latent-variable models	1160
13	35.4.6	Robustness to model errors	1162
14	35.5	Off-policy learning	1162
15	35.5.1	Basic techniques	1163
16	35.5.2	The curse of horizon	1166
17	35.5.3	The deadly triad	1167
18	35.6	Control as inference	1169
19	35.6.1	Maximum entropy reinforcement learning	1169
20	35.6.2	Other approaches	1171
21	35.6.3	Imitation learning	1172
22	<b>36</b>	<b>Causality</b>	<b>1175</b>
23	36.1	Introduction	1175
24	36.2	Causal formalism	1177
25	36.2.1	Structural causal models	1177
26	36.2.2	Causal DAGs	1178
27	36.2.3	Identification	1180
28	36.2.4	Counterfactuals and the causal hierarchy	1182
29	36.3	Randomized control trials	1184
30	36.4	Confounder adjustment	1185
31	36.4.1	Causal estimand, statistical estimand, and identification	1185
32	36.4.2	ATE estimation with observed confounders	1188
33	36.4.3	Uncertainty quantification	1193
34	36.4.4	Matching	1193
35	36.4.5	Practical considerations and procedures	1194
36	36.4.6	Summary and practical advice	1197
37	36.5	Instrumental variable strategies	1199
38	36.5.1	Additive unobserved confounding	1200
39	36.5.2	Instrument monotonicity and local average treatment effect	1202
40	36.5.3	Two stage least squares	1205
41	36.6	Difference in differences	1206
42	36.6.1	Estimation	1209
43	36.7	Credibility checks	1210
44	36.7.1	Placebo checks	1210
45	36.7.2	Sensitivity analysis to unobserved confounding	1211
46	36.8	The do-calculus	1219
47	36.8.1	The three rules	1219
	36.8.2	Revisiting backdoor adjustment	1220
	36.8.3	Frontdoor adjustment	1221
	36.9	Further reading	1222
	<b>Bibliography</b>		<b>1243</b>