

Contents

Preface	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 Some other multivariate continuous distributions	24
2.3 The exponential family	29
2.3.1 Definition	30
2.3.2 Examples	31
2.3.3 Log partition function is cumulant generating function	35
2.3.4 Canonical (natural) vs mean (moment) parameters	37
2.3.5 MLE for the exponential family	38
2.3.6 Exponential dispersion family	39
2.3.7 Maximum entropy derivation of the exponential family	39
2.4 Fisher information matrix (FIM)	40
2.4.1 Definition	40
2.4.2 Equivalence between the FIM and the Hessian of the NLL	40
2.4.3 Examples	42
2.4.4 Approximating KL divergence using FIM	43
2.4.5 Fisher information matrix for exponential family	43
2.5 Transformations of random variables	44
2.5.1 Invertible transformations (bijections)	45
2.5.2 Monte Carlo approximation	45
2.5.3 Probability integral transform	46
2.6 Markov chains	46
2.6.1 Parameterization	47
2.6.2 Application: Language modeling	50
2.6.3 Parameter estimation	50
2.6.4 Stationary distribution of a Markov chain	52
2.7 Divergence measures between probability distributions	56

2.7.1	f-divergence	56	
2.7.2	Integral probability metrics	58	
2.7.3	Maximum mean discrepancy (MMD)	59	
2.7.4	Total variation distance	61	
2.7.5	Comparing distributions using binary classifiers	61	
3	Bayesian statistics	65	
3.1	Introduction	65	
3.1.1	Frequentist statistics	65	
3.1.2	Bayesian statistics	65	
3.1.3	Arguments for the Bayesian approach	66	
3.1.4	Arguments against the Bayesian approach	67	
3.1.5	Why not just use MAP estimation?	67	
3.2	Closed-form analysis using conjugate priors	72	
3.2.1	The binomial model	72	
3.2.2	The multinomial model	73	
3.2.3	The univariate Gaussian model	74	
3.3	Conjugate Bayesian analysis for the multivariate Gaussian	79	
3.3.1	Posterior of μ given Σ	79	
3.3.2	Posterior of Σ given μ	80	
3.3.3	Posterior of Σ and μ	81	
3.3.4	Conjugate-exponential models	85	
3.4	Beyond conjugate priors	87	
3.4.1	Robust (heavy-tailed) priors	88	
3.4.2	Priors for variance parameters	88	
3.5	Noninformative priors	89	
3.5.1	Maximum entropy priors	90	
3.5.2	Jeffreys priors	91	
3.5.3	Invariant priors	94	
3.5.4	Reference priors	95	
3.6	Hierarchical priors	95	
3.6.1	A hierarchical binomial model	96	
3.6.2	A hierarchical Gaussian model	98	
3.7	Empirical Bayes	101	
3.7.1	A hierarchical binomial model	102	
3.7.2	A hierarchical Gaussian model	103	
3.7.3	Hierarchical Bayes for n-gram smoothing	104	
3.8	Model selection and evaluation	106	
3.8.1	Bayesian model selection	106	
3.8.2	Estimating the marginal likelihood	107	
3.8.3	Connection between cross validation and marginal likelihood	108	
3.8.4	Pareto-Smoothed Importance Sampling LOO estimate	109	
3.8.5	Information criteria	110	
3.8.6	Posterior predictive checks	112	
3.8.7	Bayesian p-values	113	
4	Probabilistic graphical models	117	
4.1	Introduction	117	
4.2	Directed graphical models (Bayes nets)	117	
4.2.1	Representing the joint distribution	117	
4.2.2	Examples	118	
4.2.3	Gaussian Bayes nets	122	
4.2.4	Conditional independence properties	123	
4.2.5	Generation (sampling)	128	
4.2.6	Inference	128	
4.2.7	Learning	130	
4.2.8	Plate notation	135	

1			
2	4.3	Undirected graphical models (Markov random fields)	138
3	4.3.1	Representing the joint distribution	138
4	4.3.2	Fully visible MRFs (Ising, Potts, Hopfield, etc)	140
5	4.3.3	MRFs with latent variables (Boltzmann machines, etc)	146
6	4.3.4	Maximum entropy models	148
7	4.3.5	Gaussian MRFs	150
8	4.3.6	Conditional independence properties	153
9	4.3.7	Generation (sampling)	155
10	4.3.8	Inference	155
11	4.3.9	Learning	156
12	4.4	Conditional random fields (CRFs)	160
13	4.4.1	1d CRFs	160
14	4.4.2	2d CRFs	163
15	4.4.3	Parameter estimation	166
16	4.4.4	Other approaches to structured prediction	167
17	4.5	Comparing directed and undirected PGMs	167
18	4.5.1	CI properties	167
19	4.5.2	Converting between a directed and undirected model	169
20	4.5.3	Conditional directed vs undirected PGMs and the label bias problem	170
21	4.5.4	Combining directed and undirected graphs	171
22	4.5.5	Comparing directed and undirected Gaussian PGMs	173
23	4.6	PGM extensions	175
24	4.6.1	Factor graphs	175
25	4.6.2	Probabilistic circuits	178
26	4.6.3	Directed relational PGMs	178
27	4.6.4	Undirected relational PGMs	180
28	4.6.5	Open-universe probability models	183
29	4.6.6	Programs as probability models	184
30	4.7	Structural causal models	184
31	4.7.1	Example: causal impact of education on wealth	185
32	4.7.2	Structural equation models	186
33	4.7.3	Do operator and augmented DAGs	187
34	4.7.4	Estimating average treatment effect using path analysis	188
35	4.7.5	Counterfactuals	189
36	5	Information theory	193
37	5.1	KL divergence	193
38	5.1.1	Desiderata	194
39	5.1.2	The KL divergence uniquely satisfies the desiderata	195
40	5.1.3	Thinking about KL	198
41	5.1.4	Properties of KL	200
42	5.1.5	KL divergence and MLE	202
43	5.1.6	KL divergence and Bayesian Inference	203
44	5.1.7	KL divergence and Exponential Families	204
45	5.1.8	Bregman divergence	205
46	5.2	Entropy	206
47	5.2.1	Definition	206
	5.2.2	Differential entropy for continuous random variables	207
	5.2.3	Typical sets	208
	5.2.4	Cross entropy and perplexity	209
	5.3	Mutual information	210
	5.3.1	Definition	210
	5.3.2	Interpretation	210
	5.3.3	Data processing inequality	211
	5.3.4	Sufficient Statistics	212
	5.3.5	Multivariate mutual information	212

1			
2	5.3.6	Variational bounds on mutual information	215
3	5.3.7	Relevance networks	217
4	5.4	Data compression (source coding)	218
5	5.4.1	Lossless compression	219
6	5.4.2	Lossy compression and the rate-distortion tradeoff	219
7	5.4.3	Bits back coding	221
8	5.5	Error-correcting codes (channel coding)	222
9	5.6	The information bottleneck	224
10	5.6.1	Vanilla IB	224
11	5.6.2	Variational IB	225
12	5.6.3	Conditional entropy bottleneck	226
13	6	Optimization	229
14	6.1	Introduction	229
15	6.2	Automatic differentiation	229
16	6.2.1	Differentiation in functional form	229
17	6.2.2	Differentiating chains, circuits, and programs	234
18	6.3	Stochastic gradient descent	239
19	6.4	Natural gradient descent	240
20	6.4.1	Defining the natural gradient	240
21	6.4.2	Interpretations of NGD	241
22	6.4.3	Benefits of NGD	242
23	6.4.4	Approximating the natural gradient	243
24	6.4.5	Natural gradients for the exponential family	244
25	6.5	Gradients of stochastic functions	246
26	6.5.1	Minibatch approximation to finite-sum objectives	247
27	6.5.2	Optimizing parameters of a distribution	247
28	6.5.3	Score function estimator (likelihood ratio trick)	248
29	6.5.4	Reparameterization trick	249
30	6.5.5	The delta method	251
31	6.5.6	Gumbel softmax trick	251
32	6.5.7	Stochastic computation graphs	252
33	6.5.8	Straight-through estimator	252
34	6.6	Bound optimization (MM) algorithms	253
35	6.6.1	The general algorithm	253
36	6.6.2	Example: logistic regression	254
37	6.6.3	The EM algorithm	256
38	6.6.4	Example: EM for an MVN with missing data	258
39	6.6.5	Example: robust linear regression using Student- <i>t</i> likelihood	260
40	6.6.6	Extensions to EM	261
41	6.7	The Bayesian learning rule	263
42	6.7.1	Deriving inference algorithms from BLR	264
43	6.7.2	Deriving optimization algorithms from BLR	266
44	6.7.3	Variational optimization	270
45	6.8	Bayesian optimization	270
46	6.8.1	Sequential model-based optimization	271
47	6.8.2	Surrogate functions	272
	6.8.3	Acquisition functions	273
	6.8.4	Other issues	276
	6.9	Derivative free optimization	277
	6.9.1	Local search	277
	6.9.2	Simulated annealing	280
	6.9.3	Evolutionary algorithms	283
	6.9.4	Estimation of distribution (EDA) algorithms	284
	6.9.5	Cross-entropy method	287
	6.9.6	Evolutionary strategies	287

1			
2	6.10	Optimal Transport	289
3	6.10.1	Warm-up: Matching optimally two families of points	289
4	6.10.2	From Optimal Matchings to Kantorovich and Monge formulations	289
5	6.10.3	Solving optimal transport	292
6	6.11	Submodular optimization	297
7	6.11.1	Intuition, Examples, and Background	297
8	6.11.2	Submodular Basic Definitions	300
9	6.11.3	Example Submodular Functions	301
10	6.11.4	Submodular Optimization	303
11	6.11.5	Applications of Submodularity in Machine Learning and AI	308
12	6.11.6	Sketching, CoreSets, Distillation, and Data Subset & Feature Selection	308
13	6.11.7	Combinatorial Information Functions	312
14	6.11.8	Clustering, Data Partitioning, and Parallel Machine Learning	313
15	6.11.9	Active and Semi-Supervised Learning	313
16	6.11.10	Probabilistic Modeling	314
17	6.11.11	Structured Norms and Loss Functions	316
18	6.11.12	Conclusions	316
19	II	Inference	317
20	7	Inference algorithms: an overview	319
21	7.1	Introduction	319
22	7.2	Common inference patterns	319
23	7.2.1	Global latents	320
24	7.2.2	Local latents	320
25	7.2.3	Global and local latents	321
26	7.3	Exact inference algorithms	321
27	7.4	Approximate inference algorithms	322
28	7.4.1	MAP estimation	322
29	7.4.2	Grid approximation	322
30	7.4.3	Laplace (quadratic) approximation	323
31	7.4.4	Variational inference	324
32	7.4.5	Markov Chain Monte Carlo (MCMC)	326
33	7.4.6	Sequential Monte Carlo	327
34	7.4.7	Challenging posteriors	328
35	7.5	Evaluating approximate inference algorithms	328
36	8	Inference for state-space models	331
37	8.1	Introduction	331
38	8.2	Inference for discrete chains	331
39	8.2.1	Example: casino HMM	332
40	8.2.2	Forwards filtering	333
41	8.2.3	Backwards smoothing	335
42	8.2.4	The forwards-backwards algorithm	337
43	8.2.5	Numerically stable implementation	338
44	8.2.6	Time and space complexity	340
45	8.2.7	The Viterbi algorithm	340
46	8.2.8	Forwards filtering, backwards sampling	343
47	8.3	Inference for linear-Gaussian chains	344
	8.3.1	Examples	344
	8.3.2	The Kalman filter	347
	8.3.3	The Kalman (RTS) smoother	349
	8.3.4	Information form filtering and smoothing	351
	8.4	Inference for non-linear and/or non-Gaussian chains	351
	8.5	Inference based on local linearization	351
	8.5.1	Taylor series expansion	353

1			
2		8.5.2 The extended Kalman filter (EKF)	355
3		8.5.3 The extended Kalman smoother	358
4	8.6	Inference based on the unscented transform	358
5		8.6.1 The unscented transform	358
6		8.6.2 The unscented Kalman filter (UKF)	360
7		8.6.3 The unscented Kalman smoother	362
8	8.7	Other variants of the Kalman filter	362
9		8.7.1 Ensemble Kalman filter	362
10		8.7.2 Robust Kalman filters	364
11		8.7.3 Gaussian filtering	364
12	8.8	Assumed density filtering	366
13		8.8.1 Gaussian sum filter	367
14		8.8.2 ADF for logistic regression	369
15		8.8.3 ADF for DNNs	372
16	9	Inference for graphical models	373
17	9.1	Introduction	373
18	9.2	Belief propagation on trees	374
19		9.2.1 BP for undirected graphs with pairwise potentials	374
20		9.2.2 Max product belief propagation	375
21	9.3	Loopy belief propagation	377
22		9.3.1 Loopy BP for factor graphs	377
23		9.3.2 Gaussian belief propagation	378
24		9.3.3 Convergence	380
25		9.3.4 Accuracy	382
26		9.3.5 Generalized belief propagation	383
27		9.3.6 Convex BP	383
28		9.3.7 Application: error correcting codes	383
29		9.3.8 Application: Affinity propagation	385
30		9.3.9 Emulating BP with graph neural nets	386
31	9.4	The variable elimination (VE) algorithm	387
32		9.4.1 Derivation of the algorithm	387
33		9.4.2 Computational complexity of VE	388
34		9.4.3 Computational complexity of exact inference	390
35		9.4.4 Drawbacks of VE	391
36	9.5	The junction tree algorithm (JTA)	392
37	9.6	Inference as optimization	392
38		9.6.1 Inference as backpropagation	393
39		9.6.2 Perturb and MAP	395
40	10	Variational inference	397
41	10.1	Introduction	397
42		10.1.1 Variational free energy	397
43		10.1.2 Evidence lower bound (ELBO)	398
44	10.2	Mean field VI	399
45		10.2.1 Coordinate ascent variational inference (CAVI)	399
46		10.2.2 Example: CAVI for the Ising model	400
47		10.2.3 Variational Bayes	402
48		10.2.4 Example: VB for a univariate Gaussian	403
49		10.2.5 Variational Bayes EM	406
50		10.2.6 Example: VBEM for a GMM	407
51		10.2.7 Variational message passing (VMP)	413
52		10.2.8 Autoconj	414
53	10.3	Fixed-form VI	414
54		10.3.1 Stochastic variational inference	414
55		10.3.2 Black-box variational inference	415
56		10.3.3 Reparameterization VI	417

1			
2		10.3.4 Full-rank Gaussian VI	418
3		10.3.5 Low-rank Gaussian VI	418
4		10.3.6 Automatic differentiation VI	420
5		10.3.7 Sparse Gaussian VI	424
6		10.3.8 Non-Gaussian reparameterized VI	426
7		10.3.9 Amortized inference	427
8	10.4	More accurate variational posteriors	429
9		10.4.1 Structured mean field	429
10		10.4.2 Hierarchical (auxiliary variable) posteriors	429
11		10.4.3 Normalizing flow posteriors	430
12		10.4.4 Implicit posteriors	432
13		10.4.5 Combining VI with MCMC inference	432
14	10.5	Lower bounds	432
15		10.5.1 Multi-sample ELBO (IWAE bound)	432
16		10.5.2 The thermodynamic variational objective (TVO)	433
17	10.6	Upper bounds	434
18		10.6.1 Minimizing the χ -divergence upper bound	435
19		10.6.2 Minimizing the evidence upper bound	436
20	10.7	Expectation propagation (EP)	436
21		10.7.1 Minimizing forwards vs reverse KL	436
22		10.7.2 EP as generalized ADF	438
23		10.7.3 Algorithm	439
24		10.7.4 Example	440
25		10.7.5 Optimization issues	441
26		10.7.6 Power EP and α -divergence	441
27		10.7.7 Stochastic EP	441
28		10.7.8 Applications	442
29	11	Monte Carlo inference	443
30		11.1 Introduction	443
31		11.2 Monte Carlo integration	443
32		11.2.1 Example: estimating π by Monte Carlo integration	444
33		11.2.2 Accuracy of Monte Carlo integration	444
34	11.3	Generating random samples from simple distributions	446
35		11.3.1 Sampling using the inverse cdf	446
36		11.3.2 Sampling from a Gaussian (Box-Muller method)	447
37	11.4	Rejection sampling	447
38		11.4.1 Basic idea	448
39		11.4.2 Example	449
40		11.4.3 Adaptive rejection sampling	449
41		11.4.4 Rejection sampling in high dimensions	450
42	11.5	Importance sampling	450
43		11.5.1 Direct importance sampling	451
44		11.5.2 Self-normalized importance sampling	451
45		11.5.3 Choosing the proposal	452
46		11.5.4 Annealed importance sampling (AIS)	453
47	11.6	Controlling Monte Carlo variance	454
		11.6.1 Common random numbers	454
		11.6.2 Rao-Blackwellisation	454
		11.6.3 Control variates	455
		11.6.4 Antithetic sampling	456
		11.6.5 Quasi Monte Carlo (QMC)	457
	12	Markov Chain Monte Carlo inference	459
		12.1 Introduction	459
		12.2 Metropolis Hastings algorithm	460
		12.2.1 Basic idea	460

1			
2	12.2.2	Why MH works	461
3	12.2.3	Proposal distributions	462
4	12.2.4	Initialization	465
5	12.3	Gibbs sampling	465
6	12.3.1	Basic idea	465
7	12.3.2	Gibbs sampling is a special case of MH	466
8	12.3.3	Example: Gibbs sampling for Ising models	466
9	12.3.4	Example: Gibbs sampling for Potts models	468
10	12.3.5	Example: Gibbs sampling for GMMs	468
11	12.3.6	Sampling from the full conditionals	470
12	12.3.7	Blocked Gibbs sampling	471
13	12.3.8	Collapsed Gibbs sampling	472
14	12.4	Auxiliary variable MCMC	474
15	12.4.1	Slice sampling	475
16	12.4.2	Swendsen Wang	476
17	12.5	Hamiltonian Monte Carlo (HMC)	478
18	12.5.1	Hamiltonian mechanics	478
19	12.5.2	Integrating Hamilton's equations	479
20	12.5.3	The HMC algorithm	480
21	12.5.4	Tuning HMC	481
22	12.5.5	Riemann Manifold HMC	482
23	12.5.6	Langevin Monte Carlo (MALA)	483
24	12.5.7	Connection between SGD and Langevin sampling	484
25	12.5.8	Applying HMC to constrained parameters	486
26	12.5.9	Speeding up HMC	487
27	12.6	MCMC convergence	487
28	12.6.1	Mixing rates of Markov chains	488
29	12.6.2	Practical convergence diagnostics	488
30	12.6.3	Improving speed of convergence	496
31	12.6.4	Non-centered parameterizations and Neal's funnel	496
32	12.7	Stochastic gradient MCMC	497
33	12.7.1	Stochastic Gradient Langevin Dynamics (SGLD)	498
34	12.7.2	Preconditioning	498
35	12.7.3	Reducing the variance of the gradient estimate	499
36	12.7.4	SG-HMC	500
37	12.7.5	Underdamped Langevin Dynamics	501
38	12.8	Reversible jump (trans-dimensional) MCMC	502
39	12.8.1	Basic idea	502
40	12.8.2	Example	504
41	12.8.3	Discussion	505
42	12.9	Annealing methods	505
43	12.9.1	Parallel tempering	506
44	13	Sequential Monte Carlo inference	507
45	13.1	Introduction	507
46	13.1.1	Problem statement	507
47	13.1.2	Particle filtering for state-space models	507
	13.1.3	SMC samplers for static parameter estimation	509
	13.2	Particle filtering	509
	13.2.1	Importance sampling	509
	13.2.2	Sequential importance sampling	510
	13.2.3	Sequential importance sampling with resampling	511
	13.2.4	Resampling methods	514
	13.2.5	Adaptive resampling	516
	13.3	Proposal distributions	517
	13.3.1	Locally optimal proposal	517

1			
2	13.3.2	Proposals based on the Laplace approximation	518
3	13.3.3	Proposals based on the extended and unscented Kalman filter	518
4	13.3.4	Proposals based on SMC	520
5	13.3.5	Learned (“neural”) proposals (Unfinished)	520
6	13.4	Rao-Blackwellised particle filtering (RBPF)	521
7	13.4.1	Mixture of Kalman filters	521
8	13.4.2	Example: tracking a maneuvering object	523
9	13.4.3	Example: Simultaneous localization and mapping (SLAM)	523
10	13.5	Extensions of the particle filter	527
11	13.6	SMC samplers	528
12	13.6.1	Ingredients of an SMC sampler	528
13	13.6.2	Likelihood tempering (geometric path)	529
14	13.6.3	Data tempering	532
15	13.6.4	Sampling rare events and extrema	533
16	13.6.5	SMC-ABC and likelihood-free inference	534
17	13.6.6	SMC ²	534
18			
19	III	Prediction	537
20	14	Predictive models: an overview	539
21	14.1	Introduction	539
22	14.1.1	Types of model	539
23	14.1.2	Model fitting using ERM, MLE and MAP	540
24	14.1.3	Model fitting using Bayes, VI and generalized Bayes	541
25	14.2	Evaluating predictive models	542
26	14.2.1	Proper scoring rules	542
27	14.2.2	Calibration	542
28	14.2.3	Beyond evaluating marginal probabilities	546
29	14.3	Conformal prediction	549
30	14.3.1	Conformalizing classification	550
31	14.3.2	Conformalizing regression	551
32	14.3.3	Conformalizing Bayes	552
33	14.3.4	What do we do if we don’t have a calibration set?	553
34	15	Generalized linear models	555
35	15.1	Introduction	555
36	15.1.1	Examples	555
37	15.1.2	GLMs with non-canonical link functions	558
38	15.1.3	Maximum likelihood estimation	558
39	15.1.4	Bayesian inference	559
40	15.2	Linear regression	560
41	15.2.1	Conjugate priors	560
42	15.2.2	Uninformative priors	562
43	15.2.3	Informative priors	564
44	15.2.4	Spike and slab prior	566
45	15.2.5	Laplace prior (Bayesian lasso)	567
46	15.2.6	Horseshoe prior	568
47	15.2.7	Automatic relevancy determination	569
	15.3	Logistic regression	571
	15.3.1	Binary logistic regression	572
	15.3.2	Multinomial logistic regression	572
	15.3.3	Priors	573
	15.3.4	Posteriors	574
	15.3.5	Laplace approximation	574
	15.3.6	MCMC inference	577
	15.3.7	Variational inference	578

1			
2		15.3.8 Assumed density filtering	578
3	15.4	Probit regression	578
4		15.4.1 Latent variable interpretation	578
5		15.4.2 Maximum likelihood estimation	579
6		15.4.3 Bayesian inference	581
7		15.4.4 Ordinal probit regression	581
8		15.4.5 Multinomial probit models	582
9	15.5	Multi-level GLMs	582
10		15.5.1 Generalized linear mixed models (GLMMs)	583
11		15.5.2 Model fitting	583
12		15.5.3 Example: radon regression	583
13	16	Deep neural networks	587
14	16.1	Introduction	587
15	16.2	Building blocks of differentiable circuits	587
16		16.2.1 Linear layers	588
17		16.2.2 Non-linearities	588
18		16.2.3 Convolutional layers	589
19		16.2.4 Residual (skip) connections	590
20		16.2.5 Normalization layers	591
21		16.2.6 Dropout layers	591
22		16.2.7 Attention layers	592
23		16.2.8 Recurrent layers	595
24		16.2.9 Multiplicative layers	595
25		16.2.10 Implicit layers	596
26	16.3	Canonical examples of neural networks	596
27		16.3.1 Multi-layer perceptrons (MLP)	597
28		16.3.2 Convolutional neural networks (CNN)	597
29		16.3.3 Recurrent neural networks (RNN)	597
30		16.3.4 Transformers	599
31		16.3.5 Graph neural networks (GNNs)	603
32	17	Bayesian neural networks	609
33	17.1	Introduction	609
34	17.2	Priors for BNNs	609
35		17.2.1 Gaussian priors	609
36		17.2.2 Sparsity-promoting priors	612
37		17.2.3 Learning the prior	612
38		17.2.4 Priors in function space	612
39		17.2.5 Architectural priors	612
40	17.3	Likelihoods for BNNs	613
41	17.4	Posteriors for BNNs	614
42		17.4.1 Laplace approximation	614
43		17.4.2 Variational inference	615
44		17.4.3 Expectation propagation	616
45		17.4.4 Last layer methods	616
46		17.4.5 Dropout	617
47		17.4.6 MCMC methods	617
		17.4.7 Methods based on the SGD trajectory	617
		17.4.8 Deep ensembles	619
		17.4.9 Approximating the posterior predictive distribution	623
	17.5	Generalization in Bayesian deep learning	624
		17.5.1 Sharp vs flat minima	624
		17.5.2 Effective dimensionality of a model	626
		17.5.3 The hypothesis space of DNNs	626
		17.5.4 Double descent	627
		17.5.5 A Bayesian Resolution to Double Descent	630

1			
2	17.5.6	PAC-Bayes	632
3	17.5.7	Out-of-Distribution Generalization for BNNs	632
4	17.6	Online inference	635
5	17.6.1	Extended Kalman Filtering for DNNs	635
6	17.6.2	Assumed Density Filtering for DNNs	638
7	17.6.3	Sequential Laplace for DNNs	639
8	17.6.4	Variational methods	640
9	17.7	Hierarchical Bayesian neural networks	640
10	17.7.1	Solving multiple related classification problems	641
11	18	Gaussian processes	645
12	18.1	Introduction	645
13	18.1.1	GPs: What and why?	645
14	18.2	Mercer kernels	647
15	18.2.1	Some popular Mercer kernels	648
16	18.2.2	Mercer's theorem	654
17	18.2.3	Kernels from Spectral Densities	655
18	18.3	GPs with Gaussian likelihoods	656
19	18.3.1	Predictions using noise-free observations	656
20	18.3.2	Predictions using noisy observations	658
21	18.3.3	Weight space vs function space	659
22	18.3.4	Semi-parametric GPs	659
23	18.3.5	Marginal likelihood	660
24	18.3.6	Computational and numerical issues	661
25	18.3.7	Kernel ridge regression	661
26	18.4	GPs with non-Gaussian likelihoods	664
27	18.4.1	Binary classification	665
28	18.4.2	Multi-class classification	666
29	18.4.3	GPs for Poisson regression (Cox process)	667
30	18.4.4	Other likelihoods	667
31	18.5	Scaling GP inference to large datasets	668
32	18.5.1	Subset of data	668
33	18.5.2	Nyström approximation	669
34	18.5.3	Inducing point methods	670
35	18.5.4	Sparse variational methods	673
36	18.5.5	Exploiting parallelization and structure via kernel matrix multiplies	677
37	18.5.6	Converting a GP to a SSM	679
38	18.6	Learning the kernel	679
39	18.6.1	Empirical Bayes for the kernel parameters	680
40	18.6.2	Bayesian inference for the kernel parameters	682
41	18.6.3	Multiple kernel learning for additive kernels	683
42	18.6.4	Automatic search for compositional kernels	685
43	18.6.5	Spectral mixture kernel learning	687
44	18.6.6	Deep kernel learning	690
45	18.7	GPs and DNNs	691
46	18.7.1	Kernels derived from random DNNs (NN-GP)	691
47	18.7.2	Kernels derived from trained DNNs (neural tangent kernel)	694
	18.7.3	Deep GPs	696
	18.8	Gaussian processes for timeseries forecasting	701
	18.8.1	Example: Mauna Loa	701
	19	Beyond the iid assumption	703
	19.1	Introduction	703
	19.2	Distribution shift	703
	19.2.1	Motivating examples	703
	19.2.2	A causal view of distribution shift	705
	19.2.3	Covariate shift	705

1				
2	19.2.4	Domain shift	706	
3	19.2.5	Label / prior shift	707	
4	19.2.6	Concept shift	707	
5	19.2.7	Manifestation shift	707	
6	19.2.8	Selection bias	708	
7	19.3	Training-time techniques for distribution shift	708	
8	19.3.1	Importance weighting for covariate shift	709	
9	19.3.2	Domain adaptation	710	
10	19.3.3	Domain randomization	710	
11	19.3.4	Data augmentation	711	
12	19.3.5	Unsupervised label shift estimation	711	
13	19.3.6	Distributionally robust optimization	712	
14	19.4	Test-time techniques for distribution shift	712	
15	19.4.1	Detecting shifts using two-sample testing	712	
16	19.4.2	Detecting single out-of-distribution (OOD) inputs	712	
17	19.4.3	Selective prediction	715	
18	19.4.4	Open world recognition	716	
19	19.4.5	Online adaptation	717	
20	19.5	Learning from multiple distributions	718	
21	19.5.1	Transfer learning	719	
22	19.5.2	Few-shot learning	720	
23	19.5.3	Prompt tuning	720	
24	19.5.4	Zero-shot learning	720	
25	19.5.5	Multi-task learning	721	
26	19.5.6	Domain generalization	722	
27	19.5.7	Invariant risk minimization	723	
28	19.6	Meta-learning	724	
29	19.6.1	Meta-learning as probabilistic inference for prediction	725	
30	19.6.2	Gradient-based meta-learning	726	
31	19.6.3	Metric-based few-shot learning	726	
32	19.6.4	VERSA	726	
33	19.6.5	Neural processes	727	
34	19.7	Continual learning	727	
35	19.7.1	Domain drift	727	
36	19.7.2	Concept drift	727	
37	19.7.3	Task incremental learning	729	
38	19.7.4	Catastrophic forgetting	730	
39	19.7.5	Online learning	732	
40	19.8	Adversarial examples	733	
41	19.8.1	Whitebox (gradient-based) attacks	735	
42	19.8.2	Blackbox (gradient-free) attacks	735	
43	19.8.3	Real world adversarial attacks	737	
44	19.8.4	Defenses based on robust optimization	737	
45	19.8.5	Why models have adversarial examples	738	

IV Generation 741

40	20	Generative models: an overview	743	
41	20.1	Introduction	743	
42	20.2	Types of generative model	743	
43	20.3	Goals of generative modeling	745	
44	20.3.1	Generating data	745	
45	20.3.2	Density estimation	746	
46	20.3.3	Imputation	747	
47	20.3.4	Structure discovery	748	

1			
2		20.3.5 Latent space interpolation	749
3		20.3.6 Representation learning	750
4	20.4	Evaluating generative models	750
5		20.4.1 Likelihood	751
6		20.4.2 Distances and divergences in feature space	752
7		20.4.3 Precision and recall metrics	753
8		20.4.4 Statistical tests	754
9		20.4.5 Challenges with using pretrained classifiers	755
10		20.4.6 Using model samples to train classifiers	755
11		20.4.7 Assessing overfitting	755
12		20.4.8 Human evaluation	755
13	21	Variational autoencoders	757
14	21.1	Introduction	757
15	21.2	VAE basics	757
16		21.2.1 Modeling assumptions	758
17		21.2.2 Evidence lower bound	759
18		21.2.3 Optimization	760
19		21.2.4 The reparameterization trick	760
20		21.2.5 Computing the reparameterized ELBO	762
21		21.2.6 Comparison of VAEs and autoencoders	763
22		21.2.7 VAEs optimize in an augmented space	764
23	21.3	VAE generalizations	767
24		21.3.1 σ -VAE	767
25		21.3.2 β -VAE	769
26		21.3.3 InfoVAE	771
27		21.3.4 Multi-modal VAEs	774
28		21.3.5 VAEs with missing data	775
29		21.3.6 Semi-supervised VAEs	777
30		21.3.7 VAEs with sequential encoders/decoders	778
31	21.4	Avoiding posterior collapse	781
32		21.4.1 KL annealing	782
33		21.4.2 Lower bounding the rate	782
34		21.4.3 Free bits	783
35		21.4.4 Adding skip connections	783
36		21.4.5 Improved variational inference	783
37		21.4.6 Alternative objectives	784
38		21.4.7 Enforcing identifiability	784
39	21.5	VAEs with hierarchical structure	784
40		21.5.1 Bottom-up vs top-down inference	785
41		21.5.2 Example: Very deep VAE	786
42		21.5.3 Connection with autoregressive models	788
43		21.5.4 Variational pruning	789
44		21.5.5 Other optimization difficulties	790
45	21.6	Vector quantization VAE	791
46		21.6.1 Autoencoder with binary code	791
47		21.6.2 VQ-VAE model	791
		21.6.3 Learning the prior	793
		21.6.4 Hierarchical extension (VQ-VAE-2)	793
		21.6.5 Discrete VAE	794
		21.6.6 VQ-GAN	795
	21.7	Wake-sleep algorithm	796
		21.7.1 Wake phase	797
		21.7.2 Sleep phase	797
		21.7.3 Daydream phase	798
		21.7.4 Summary of algorithm	799

1			
2	22 Auto-regressive models	801	
3	22.1 Introduction	801	
4	22.2 Neural autoregressive density estimators (NADE)	802	
5	22.3 Causal CNNs	802	
6	22.3.1 1d causal CNN (Convolutional Markov models)	803	
7	22.3.2 2d causal CNN (PixelCNN)	803	
8	22.4 Transformer decoders	804	
9	22.4.1 Text generation (GPT)	805	
10	22.4.2 Music generation	805	
11	22.4.3 Text-to-image generation (DALL-E)	806	
12	23 Normalizing Flows	809	
13	23.1 Introduction	809	
14	23.1.1 Preliminaries	809	
15	23.1.2 Example	811	
16	23.1.3 How to train a flow model	812	
17	23.2 Constructing Flows	813	
18	23.2.1 Affine flows	813	
19	23.2.2 Elementwise flows	814	
20	23.2.3 Coupling flows	816	
21	23.2.4 Autoregressive flows	818	
22	23.2.5 Residual flows	823	
23	23.2.6 Continuous-time flows	825	
24	23.3 Applications	827	
25	23.3.1 Density estimation	827	
26	23.3.2 Generative Modeling	828	
27	23.3.3 Inference	828	
28	24 Energy-based models	831	
29	24.1 Introduction	831	
30	24.1.1 Example: Products of experts (PoE)	832	
31	24.1.2 Computational difficulties	832	
32	24.2 Maximum Likelihood Training	833	
33	24.2.1 Gradient-based MCMC methods	834	
34	24.2.2 Contrastive divergence	834	
35	24.3 Score Matching (SM)	838	
36	24.3.1 Basic score matching	838	
37	24.3.2 Denoising Score Matching (DSM)	839	
38	24.3.3 Sliced Score Matching (SSM)	840	
39	24.3.4 Connection to Contrastive Divergence	841	
40	24.3.5 Score-Based Generative Models	842	
41	24.4 Noise Contrastive Estimation	845	
42	24.4.1 Connection to Score Matching	846	
43	24.5 Other Methods	847	
44	24.5.1 Minimizing Differences/Derivatives of KL Divergences	847	
45	24.5.2 Minimizing the Stein Discrepancy	848	
46	24.5.3 Adversarial Training	848	
47	25 Diffusion models	851	
	25.1 Variational diffusion models	851	
	25.1.1 Encoder	851	
	25.1.2 Decoder	853	
	25.1.3 Model fitting	855	
	25.1.4 Connection to DDPM	858	
	25.1.5 2d Example	858	
	25.1.6 Application to image generation	859	
	25.2 Conditional diffusion models	859	

1			
2	25.2.1	Classifier guidance	860
3	25.2.2	Classifier-free guidance	860
4	25.2.3	Conditional image generation	861
5	25.2.4	Other forms of conditional generation	861
6	25.3	Speeding up the generation process	861
7	26	Generative adversarial networks	865
8	26.1	Introduction	865
9	26.2	Learning by Comparison	866
10	26.2.1	Guiding principles	867
11	26.2.2	Class probability estimation	868
12	26.2.3	Bounds on f -divergences	871
13	26.2.4	Integral probability metrics	872
14	26.2.5	Moment matching	874
15	26.2.6	On density ratios and differences	875
16	26.3	Generative Adversarial Networks	876
17	26.3.1	From learning principles to loss functions	877
18	26.3.2	Gradient Descent	878
19	26.3.3	Challenges with GAN training	879
20	26.3.4	Improving GAN optimization	881
21	26.3.5	Convergence of GAN training	881
22	26.4	Conditional GANs	885
23	26.5	Inference with GANs	886
24	26.6	Neural architectures in GANs	887
25	26.6.1	The importance of discriminator architectures	887
26	26.6.2	Architectural inductive biases	887
27	26.6.3	Attention in GANs	888
28	26.6.4	Progressive generation	889
29	26.6.5	Regularization	890
30	26.6.6	Scaling up GAN models	891
31	26.7	Applications	891
32	26.7.1	GANs for image generation	891
33	26.7.2	Video generation	893
34	26.7.3	Audio generation	894
35	26.7.4	Text generation	895
36	26.7.5	Imitation Learning	896
37	26.7.6	Domain Adaptation	896
38	26.7.7	Design, Art and Creativity	897
39	V	Discovery	899
40	27	Discovery methods: an overview	901
41	27.1	Introduction	901
42	27.2	Overview of Part V	902
43	28	Latent factor models	903
44	28.1	Introduction	903
45	28.2	Mixture models	903
46	28.2.1	Gaussian mixture models (GMMs)	904
47	28.2.2	Bernoulli mixture models	906
	28.2.3	Gaussian scale mixtures (GSMs)	906
	28.2.4	Using GMMs as a prior for inverse imaging problems	908
	28.2.5	Using mixture models for classification problems	911
	28.3	Factor analysis	913
	28.3.1	Factor analysis: the basics	913
	28.3.2	Probabilistic PCA	917

1			
2	28.3.3	Mixture of factor analysers	919
3	28.3.4	Factor analysis models for paired data	926
4	28.3.5	Factor analysis with exponential family likelihoods	929
5	28.3.6	Factor analysis with DNN likelihoods	930
6	28.3.7	Factor analysis with GP likelihoods (GP-LVM)	931
7	28.4	LFMs with non-Gaussian priors	933
8	28.4.1	Non-negative matrix factorization (NMF)	933
9	28.4.2	Multinomial PCA	934
10	28.5	Topic models	936
11	28.5.1	Latent Dirichlet Allocation (LDA)	936
12	28.5.2	Correlated topic model	940
13	28.5.3	Dynamic topic model	940
14	28.5.4	LDA-HMM	941
15	28.6	Independent components analysis (ICA)	945
16	28.6.1	Noiseless ICA model	945
17	28.6.2	The need for non-Gaussian priors	946
18	28.6.3	Maximum likelihood estimation	947
19	28.6.4	Alternatives to MLE	948
20	28.6.5	Sparse coding	950
21	28.6.6	Nonlinear ICA	950
22	29	State-space models	953
23	29.1	Introduction	953
24	29.2	Hidden Markov models (HMMs)	954
25	29.2.1	Conditional independence properties	954
26	29.2.2	State transition model	954
27	29.2.3	Discrete likelihoods	955
28	29.2.4	Gaussian likelihoods	955
29	29.2.5	Autoregressive likelihoods	956
30	29.3	HMMs: Applications	958
31	29.3.1	Time series segmentation	958
32	29.3.2	Protein sequence alignment	960
33	29.3.3	Spelling correction	961
34	29.4	HMMs: parameter learning	964
35	29.4.1	The Baum-Welch (EM) algorithm	964
36	29.4.2	Parameter estimation using SGD	967
37	29.4.3	Parameter estimation using spectral methods	970
38	29.4.4	Bayesian HMM	970
39	29.5	HMMs: Generalizations	971
40	29.5.1	Hidden semi-Markov model (HSMM)	971
41	29.5.2	Hierarchical HMMs	973
42	29.5.3	Factorial HMMs	975
43	29.5.4	Coupled HMMs	976
44	29.5.5	Dynamic Bayes nets (DBN)	977
45	29.5.6	Changepoint detection	977
46	29.6	Linear dynamical systems (LDS)	980
47	29.6.1	Conditional independence properties	980
	29.6.2	Parameterization	980
	29.7	LDS: Applications	981
	29.7.1	Object tracking and state estimation	981
	29.7.2	Timeseries forecasting	982
	29.7.3	Online linear regression	982
	29.8	Structural time series models	982
	29.8.1	Basics	983
	29.8.2	Details	983
	29.8.3	Example: modeling CO ₂ levels from Mauna Loa	986

1			
2	29.8.4	Example: forecasting electricity demand	987
3	29.8.5	Causal impact of a time series intervention	987
4	29.8.6	Prophet	991
5	29.8.7	Neural forecasting methods	993
6	29.9	LDS: parameter learning	994
7	29.9.1	EM for LDS	994
8	29.9.2	Subspace identification methods	996
9	29.9.3	Ensuring stability of the dynamical system	997
10	29.9.4	Bayesian LDS	997
11	29.10	Switching linear dynamical systems (SLDS)	997
12	29.10.1	Parameterization	998
13	29.10.2	Posterior inference	998
14	29.10.3	Application: Multi-target tracking	999
15	29.11	Non-linear dynamical systems	1002
16	29.11.1	Application: Nonlinear state estimation and tracking	1002
17	29.12	Deep SSMS	1003
18	29.12.1	Deep Markov models	1004
19	29.12.2	Recurrent SSM	1005
20	29.12.3	Improving multi-step predictions	1005
21	29.12.4	Variational RNNs	1006
22	30	Graph learning	1009
23	30.1	Introduction	1009
24	30.2	Latent variable models for graphs	1009
25	30.2.1	Stochastic block model	1009
26	30.2.2	Mixed membership stochastic block model	1011
27	30.2.3	Infinite relational model	1013
28	30.3	Graphical model structure learning	1015
29	30.3.1	Applications	1016
30	30.3.2	Methods	1017
31	31	Non-parametric Bayesian models	1019
32	31.1	Introduction	1019
33	31.2	Dirichlet processes	1019
34	31.2.1	Definition of a DP	1020
35	31.2.2	Stick breaking construction of the DP	1022
36	31.2.3	The Chinese restaurant process (CRP)	1023
37	31.3	Dirichlet process mixture models	1024
38	31.3.1	Model definition	1025
39	31.3.2	Fitting using collapsed Gibbs sampling	1026
40	31.3.3	Other fitting algorithms	1028
41	31.3.4	Choosing the hyper-parameters	1029
42	31.4	Generalizations of the Dirichlet process	1030
43	31.4.1	Pitman-Yor process	1031
44	31.4.2	Dependent random probability measures	1032
45	31.5	The Indian buffet process and the Beta process	1034
46	31.6	Small-variance asymptotics	1037
47	31.7	Completely random measures	1040
	31.8	Lévy processes	1041
	31.9	Point processes with repulsion and reinforcement	1043
	31.9.1	Poisson process	1043
	31.9.2	Renewal process	1044
	31.9.3	Hawkes process	1045
	31.9.4	Gibbs point process	1047
	31.9.5	Determinantal point process	1048

1			
2	32 Representation learning (Unfinished)	1051	
3	32.1 CLIP	1051	
4	33 Interpretability	1053	
5	33.1 Introduction	1053	
6	33.1.1 The Role of Interpretability	1054	
7	33.1.2 Terminology and Framework	1055	
8	33.2 Methods for Interpretable Machine Learning	1059	
9	33.2.1 Inherently Interpretable Models: The Model is its Explanation	1059	
10	33.2.2 Semi-Inherently Interpretable Models: Example-Based Methods	1062	
11	33.2.3 Post-hoc or Joint training: The Explanation gives a Partial View of the Model	1062	
12	33.2.4 Transparency and Visualization	1066	
13	33.3 Properties: The Abstraction Between Context and Method	1067	
14	33.3.1 Properties of Explanations from Interpretable Machine Learning	1068	
15	33.3.2 Properties of Explanations from Cognitive Science	1070	
16	33.4 Evaluation of Interpretable Machine Learning Models	1071	
17	33.4.1 Computational Evaluation: Does the Method have Desired Properties?	1072	
18	33.4.2 User Study-based Evaluation: Does the Method Help a User Perform a Task?	1077	
19	33.5 Discussion: How to Think about Interpretable Machine Learning	1080	
20	VI Action	1087	
21	34 Decision making under uncertainty	1089	
22	34.1 Bayesian decision theory	1089	
23	34.1.1 Basics	1089	
24	34.1.2 Classification	1090	
25	34.1.3 Regression	1090	
26	34.1.4 Structured prediction	1091	
27	34.1.5 Fairness	1092	
28	34.2 Decision (influence) diagrams	1092	
29	34.2.1 Example: oil wildcatter	1092	
30	34.2.2 Information arcs	1093	
31	34.2.3 Value of information	1094	
32	34.2.4 Computing the optimal policy	1095	
33	34.3 A/B testing	1095	
34	34.3.1 A Bayesian approach	1096	
35	34.3.2 Example	1099	
36	34.4 Contextual bandits	1100	
37	34.4.1 Types of bandit	1101	
38	34.4.2 Applications	1102	
39	34.4.3 Exploration-exploitation tradeoff	1102	
40	34.4.4 The optimal solution	1103	
41	34.4.5 Upper confidence bounds (UCB)	1104	
42	34.4.6 Thompson sampling	1106	
43	34.4.7 Regret	1107	
44	34.5 Markov decision problems	1109	
45	34.5.1 Basics	1109	
46	34.5.2 Partially observed MDPs	1110	
47	34.5.3 Episodes and returns	1110	
	34.5.4 Value functions	1112	
	34.5.5 Optimal value functions and policies	1112	
	34.6 Planning in an MDP	1113	
	34.6.1 Value iteration	1114	
	34.6.2 Policy iteration	1115	
	34.6.3 Linear programming	1116	

1			
2	34.7	Active learning	1117
3	34.7.1	Relationship to other forms of sequential decision making	1117
4	34.7.2	Common heuristics	1117
5	34.7.3	Batch methods	1118
6	35	Reinforcement learning	1119
7	35.1	Introduction	1119
8	35.1.1	Overview of methods	1119
9	35.1.2	Value based methods	1121
10	35.1.3	Policy search methods	1121
11	35.1.4	Model-based RL	1121
12	35.1.5	Exploration-exploitation tradeoff	1122
13	35.2	Value-based RL	1124
14	35.2.1	Monte Carlo RL	1124
15	35.2.2	Temporal difference (TD) learning	1124
16	35.2.3	TD learning with eligibility traces	1125
17	35.2.4	SARSA: on-policy TD control	1126
18	35.2.5	Q-learning: off-policy TD control	1127
19	35.2.6	Deep Q-network (DQN)	1129
20	35.3	Policy-based RL	1130
21	35.3.1	The policy gradient theorem	1130
22	35.3.2	REINFORCE	1131
23	35.3.3	Actor-critic methods	1132
24	35.3.4	Bound optimization methods	1134
25	35.3.5	Deterministic policy gradient methods	1136
26	35.3.6	Gradient-free methods	1137
27	35.4	Model-based RL	1137
28	35.4.1	Model predictive control (MPC)	1137
29	35.4.2	Combining model-based and model-free	1139
30	35.4.3	MBRL using Gaussian processes	1139
31	35.4.4	MBRL using DNNs	1141
32	35.4.5	MBRL using latent-variable models	1141
33	35.4.6	Robustness to model errors	1144
34	35.5	Off-policy learning	1144
35	35.5.1	Basic techniques	1145
36	35.5.2	The curse of horizon	1148
37	35.5.3	The deadly triad	1149
38	35.6	Control as inference	1150
39	35.6.1	Maximum entropy reinforcement learning	1150
40	35.6.2	Other approaches	1153
41	35.6.3	Imitation learning	1154
42	36	Causality	1157
43	36.1	Introduction	1157
44	36.2	Causal Formalism	1159
45	36.2.1	Structural Causal Models	1159
46	36.2.2	Causal DAGs	1161
47	36.2.3	Identification	1163
	36.2.4	Counterfactuals and the Causal Hierarchy	1164
	36.3	Randomized Control Trials	1166
	36.4	Confounder Adjustment	1167
	36.4.1	Causal Estimand, Statistical Estimand, and Identification	1167
	36.4.2	ATE Estimation with Observed Confounders	1170
	36.4.3	Uncertainty Quantification	1175
	36.4.4	Matching	1175
	36.4.5	Practical Considerations and Procedures	1176
	36.4.6	Summary and Practical Advice	1179

<u>1</u>				
<u>2</u>	36.5	Instrumental Variable Strategies	1181	
<u>3</u>		36.5.1 Additive Unobserved Confounding	1183	
<u>4</u>		36.5.2 Instrument Monotonicity and Local Average Treatment Effect	1184	
<u>5</u>		36.5.3 Two Stage Least Squares	1188	
<u>6</u>	36.6	Difference in Differences	1188	
<u>7</u>		36.6.1 Estimation	1192	
<u>8</u>	36.7	Credibility Checks	1192	
<u>9</u>		36.7.1 Placebo Checks	1193	
<u>10</u>		36.7.2 Sensitivity Analysis to Unobserved Confounding	1193	
<u>11</u>	36.8	The Do Calculus	1201	
<u>12</u>		36.8.1 The three rules	1201	
<u>13</u>		36.8.2 Revisiting Backdoor Adjustment	1202	
<u>14</u>		36.8.3 Frontdoor Adjustment	1203	
<u>15</u>	36.9	Further Reading	1205	
<u>16</u>	Bibliography	1220		

17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47