
Contents

<i>Preface</i>	<i>page xi</i>
1 Introduction	1
1.1 What You Can Find in Here	1
1.2 What's Missing?	4
1.3 Resources	5
Part I Fundamentals without Noise	7
2 Control Crash Course	9
2.1 You Have a Control Problem	9
2.2 What to Do about It?	11
2.3 State Space Models	12
2.4 Stability and Performance	17
2.5 A Glance Ahead: From Control Theory to RL	29
2.6 How Can We Ignore Noise?	32
2.7 Examples	32
2.8 Exercises	43
2.9 Notes	49
3 Optimal Control	51
3.1 Value Function for Total Cost	51
3.2 Bellman Equation	52
3.3 Variations	59
3.4 Inverse Dynamic Programming	63
3.5 Bellman Equation Is a Linear Program	64
3.6 Linear Quadratic Regulator	65
3.7 A Second Glance Ahead	67
3.8 Optimal Control in Continuous Time*	68
3.9 Examples	70
3.10 Exercises	78
3.11 Notes	83
4 ODE Methods for Algorithm Design	84
4.1 Ordinary Differential Equations	84

4.2	A Brief Return to Reality	87
4.3	Newton–Raphson Flow	88
4.4	Optimization	90
4.5	Quasistochastic Approximation	97
4.6	Gradient-Free Optimization	113
4.7	Quasi Policy Gradient Algorithms	118
4.8	Stability of ODEs*	123
4.9	Convergence Theory for QSA*	131
4.10	Exercises	149
4.11	Notes	154
5	Value Function Approximations	159
5.1	Function Approximation Architectures	160
5.2	Exploration and ODE Approximations	168
5.3	TD-Learning and Linear Regression	171
5.4	Projected Bellman Equations and TD Algorithms	176
5.5	Convex Q-Learning	186
5.6	Q-Learning in Continuous Time*	191
5.7	Duality*	193
5.8	Exercises	196
5.9	Notes	199
Part II	Reinforcement Learning and Stochastic Control	203
6	Markov Chains	205
6.1	Markov Models Are State Space Models	205
6.2	Simple Examples	208
6.3	Spectra and Ergodicity	211
6.4	A Random Glance Ahead	215
6.5	Poisson’s Equation	216
6.6	Lyapunov Functions	218
6.7	Simulation: Confidence Bounds and Control Variates	222
6.8	Sensitivity and Actor-Only Methods	230
6.9	Ergodic Theory for General Markov Chains*	233
6.10	Exercises	236
6.11	Notes	243
7	Stochastic Control	244
7.1	MDPs: A Quick Introduction	244
7.2	Fluid Models for Approximation	248
7.3	Queues	251
7.4	Speed Scaling	253
7.5	LQG	257
7.6	A Queueing Game	261
7.7	Controlling Rover with Partial Information	263

Contents

ix

7.8	Bandits	266
7.9	Exercises	271
7.10	Notes	278
8	Stochastic Approximation	280
8.1	Asymptotic Covariance	281
8.2	Themes and Roadmaps	283
8.3	Examples	292
8.4	Algorithm Design Example	297
8.5	Zap Stochastic Approximation	300
8.6	Buyer Beware	304
8.7	Some Theory*	307
8.8	Exercises	314
8.9	Notes	315
9	Temporal Difference Methods	318
9.1	Policy Improvement	319
9.2	Function Approximation and Smoothing	323
9.3	Loss Functions	325
9.4	TD(λ) Learning	327
9.5	Return to the Q-Function	330
9.6	Watkins's Q-Learning	337
9.7	Relative Q-Learning	344
9.8	GQ and Zap	348
9.9	Technical Proofs*	353
9.10	Exercises	357
9.11	Notes	359
10	Setting the Stage, Return of the Actors	362
10.1	The Stage, Projection, and Adjoint	363
10.2	Advantage and Innovation	367
10.3	Regeneration	369
10.4	Average Cost and Every Other Criterion	371
10.5	Gather the Actors	376
10.6	SGD without Bias	380
10.7	Advantage and Control Variates	382
10.8	Natural Gradient and Zap	384
10.9	Technical Proofs*	385
10.10	Notes	389
	Appendices	393
A	Mathematical Background	395
A.1	Notation and Math Background	395
A.2	Probability and Markovian Background	397

B	Markov Decision Processes	401
B.1	Total Cost and Every Other Criterion	401
B.2	Computational Aspects of MDPs	403
C	Partial Observations and Belief States	409
C.1	POMDP Model	409
C.2	A Fully Observed MDP	410
C.3	Belief State Dynamics	413
	<i>References</i>	415
	<i>Glossary of Symbols and Acronyms</i>	431
	<i>Index</i>	433