Contents

Preface to the Second Edition **xiii**

Preface to the First Edition **xvii**

Summary of Notation **xix**

1 Introduction **1**
1.1 Reinforcement Learning 1
1.2 Examples 4
1.3 Elements of Reinforcement Learning 6
1.4 Limitations and Scope 7
1.5 An Extended Example: Tic-Tac-Toe 8
1.6 Summary 13
1.7 Early History of Reinforcement Learning 13

I Tabular Solution Methods **23**

2 Multi-armed Bandits **25**
2.1 A k-armed Bandit Problem 25
2.2 Action-value Methods 27
2.3 The 10-armed Testbed 28
2.4 Incremental Implementation 30
2.5 Tracking a Nonstationary Problem 32
2.6 Optimistic Initial Values 34
2.7 Upper-Confidence-Bound Action Selection 35
2.8 Gradient Bandit Algorithms 37
2.9 Associative Search (Contextual Bandits) 41
2.10 Summary 42

3 Finite Markov Decision Processes **47**
3.1 The Agent–Environment Interface . . . 47
3.2 Goals and Rewards . . . 53
3.3 Returns and Episodes . . . 54
3.4 Unified Notation for Episodic and Continuing Tasks . . . 57
3.5 Policies and Value Functions . . . 58
3.6 Optimal Policies and Optimal Value Functions . . . 62
3.7 Optimality and Approximation . . . 67
3.8 Summary . . . 68

4 Dynamic Programming **73**
4.1 Policy Evaluation (Prediction) . . . 74
4.2 Policy Improvement . . . 76
4.3 Policy Iteration . . . 80
4.4 Value Iteration . . . 82
4.5 Asynchronous Dynamic Programming . . . 85
4.6 Generalized Policy Iteration . . . 86
4.7 Efficiency of Dynamic Programming . . . 87
4.8 Summary . . . 88

5 Monte Carlo Methods **91**
5.1 Monte Carlo Prediction . . . 92
5.2 Monte Carlo Estimation of Action Values . . . 96
5.3 Monte Carlo Control . . . 97
5.4 Monte Carlo Control without Exploring Starts . . . 100
5.5 Off-policy Prediction via Importance Sampling . . . 103
5.6 Incremental Implementation . . . 109
5.7 Off-policy Monte Carlo Control . . . 110
5.8 *Discounting-aware Importance Sampling . . . 112
5.9 *Per-decision Importance Sampling . . . 114
5.10 Summary . . . 115

6 Temporal-Difference Learning **119**
6.1 TD Prediction . . . 119
6.2 Advantages of TD Prediction Methods . . . 124
6.3 Optimality of TD(0) . . . 126
6.4 Sarsa: On-policy TD Control . . . 129
6.5 Q-learning: Off-policy TD Control . . . 131
6.6 Expected Sarsa . . . 133
6.7 Maximization Bias and Double Learning . . . 134
6.8 Games, Afterstates, and Other Special Cases . . . 136
6.9 Summary . . . 138

7 *n*-step Bootstrapping 141
7.1 n-step TD Prediction . . . 142
7.2 n-step Sarsa . . . 145
7.3 n-step Off-policy Learning . . . 148
7.4 *Per-decision Methods with Control Variates . . . 150
7.5 Off-policy Learning Without Importance Sampling: The n-step Tree Backup Algorithm . . . 152
7.6 *A Unifying Algorithm: n-step $Q(\sigma)$ . . . 154
7.7 Summary . . . 157

8 Planning and Learning with Tabular Methods 159
8.1 Models and Planning . . . 159
8.2 Dyna: Integrated Planning, Acting, and Learning . . . 161
8.3 When the Model Is Wrong . . . 166
8.4 Prioritized Sweeping . . . 168
8.5 Expected vs. Sample Updates . . . 172
8.6 Trajectory Sampling . . . 174
8.7 Real-time Dynamic Programming . . . 177
8.8 Planning at Decision Time . . . 180
8.9 Heuristic Search . . . 181
8.10 Rollout Algorithms . . . 183
8.11 Monte Carlo Tree Search . . . 185
8.12 Summary of the Chapter . . . 188
8.13 Summary of Part I: Dimensions . . . 189

II Approximate Solution Methods 195

9 On-policy Prediction with Approximation 197
9.1 Value-function Approximation . . . 198
9.2 The Prediction Objective ($\overline{\text{VE}}$) . . . 199
9.3 Stochastic-gradient and Semi-gradient Methods . . . 200
9.4 Linear Methods . . . 204
9.5 Feature Construction for Linear Methods . . . 210
9.5.1 Polynomials . . . 210
9.5.2 Fourier Basis . . . 211
9.5.3 Coarse Coding . . . 215
9.5.4 Tile Coding . . . 217
9.5.5 Radial Basis Functions . . . 221
9.6 Selecting Step-Size Parameters Manually . . . 222
9.7 Nonlinear Function Approximation: Artificial Neural Networks . . . 223
9.8 Least-Squares TD . . . 228

9.9 Memory-based Function Approximation . . . 230
9.10 Kernel-based Function Approximation . . . 232
9.11 Looking Deeper at On-policy Learning: Interest and Emphasis . . . 234
9.12 Summary . . . 236

10 On-policy Control with Approximation 243
10.1 Episodic Semi-gradient Control . . . 243
10.2 Semi-gradient n-step Sarsa . . . 247
10.3 Average Reward: A New Problem Setting for Continuing Tasks . . . 249
10.4 Deprecating the Discounted Setting . . . 253
10.5 Differential Semi-gradient n-step Sarsa . . . 255
10.6 Summary . . . 256

11 *Off-policy Methods with Approximation 257
11.1 Semi-gradient Methods . . . 258
11.2 Examples of Off-policy Divergence . . . 260
11.3 The Deadly Triad . . . 264
11.4 Linear Value-function Geometry . . . 266
11.5 Gradient Descent in the Bellman Error . . . 269
11.6 The Bellman Error is Not Learnable . . . 274
11.7 Gradient-TD Methods . . . 278
11.8 Emphatic-TD Methods . . . 281
11.9 Reducing Variance . . . 283
11.10 Summary . . . 284

12 Eligibility Traces 287
12.1 The λ-return . . . 288
12.2 TD(λ) . . . 292
12.3 n-step Truncated λ-return Methods . . . 295
12.4 Redoing Updates: Online λ-return Algorithm . . . 297
12.5 True Online TD(λ) . . . 299
12.6 *Dutch Traces in Monte Carlo Learning . . . 301
12.7 Sarsa(λ) . . . 303
12.8 Variable λ and γ . . . 307
12.9 *Off-policy Traces with Control Variates . . . 309
12.10 Watkins's Q(λ) to Tree-Backup(λ) . . . 312
12.11 Stable Off-policy Methods with Traces . . . 314
12.12 Implementation Issues . . . 316
12.13 Conclusions . . . 317

13 Policy Gradient Methods **321**
13.1 Policy Approximation and its Advantages . . . 322
13.2 The Policy Gradient Theorem . . . 324
13.3 REINFORCE: Monte Carlo Policy Gradient . . . 326
13.4 REINFORCE with Baseline . . . 329
13.5 Actor–Critic Methods . . . 331
13.6 Policy Gradient for Continuing Problems . . . 333
13.7 Policy Parameterization for Continuous Actions . . . 335
13.8 Summary . . . 337

III Looking Deeper **339**

14 Psychology **341**
14.1 Prediction and Control . . . 342
14.2 Classical Conditioning . . . 343
14.2.1 Blocking and Higher-order Conditioning . . . 345
14.2.2 The Rescorla–Wagner Model . . . 346
14.2.3 The TD Model . . . 349
14.2.4 TD Model Simulations . . . 350
14.3 Instrumental Conditioning . . . 357
14.4 Delayed Reinforcement . . . 361
14.5 Cognitive Maps . . . 363
14.6 Habitual and Goal-directed Behavior . . . 364
14.7 Summary . . . 368

15 Neuroscience **377**
15.1 Neuroscience Basics . . . 378
15.2 Reward Signals, Reinforcement Signals, Values, and Prediction Errors . . . 380
15.3 The Reward Prediction Error Hypothesis . . . 381
15.4 Dopamine . . . 383
15.5 Experimental Support for the Reward Prediction Error Hypothesis . . . 387
15.6 TD Error/Dopamine Correspondence . . . 390
15.7 Neural Actor–Critic . . . 395
15.8 Actor and Critic Learning Rules . . . 398
15.9 Hedonistic Neurons . . . 402
15.10 Collective Reinforcement Learning . . . 404
15.11 Model-based Methods in the Brain . . . 407
15.12 Addiction . . . 409
15.13 Summary . . . 410

16 Applications and Case Studies **421**
16.1 TD-Gammon . . . 421
16.2 Samuel's Checkers Player . . . 426
16.3 Watson's Daily-Double Wagering . . . 429
16.4 Optimizing Memory Control . . . 432
16.5 Human-level Video Game Play . . . 436
16.6 Mastering the Game of Go . . . 441
16.6.1 AlphaGo . . . 444
16.6.2 AlphaGo Zero . . . 447
16.7 Personalized Web Services . . . 450
16.8 Thermal Soaring . . . 453

17 Frontiers **459**
17.1 General Value Functions and Auxiliary Tasks . . . 459
17.2 Temporal Abstraction via Options . . . 461
17.3 Observations and State . . . 464
17.4 Designing Reward Signals . . . 469
17.5 Remaining Issues . . . 472
17.6 The Future of Artificial Intelligence . . . 475

References **481**

Index **519**