

Contents

Part I Introductory Part

1	Reinforcement Learning and Markov Decision Processes	3
	<i>Martijn van Otterlo, Marco Wiering</i>	
1.1	Introduction	3
1.2	Learning Sequential Decision Making	5
1.3	A Formal Framework	10
1.3.1	Markov Decision Processes	10
1.3.2	Policies	13
1.3.3	Optimality Criteria and Discounting	13
1.4	Value Functions and Bellman Equations	15
1.5	Solving Markov Decision Processes	17
1.6	Dynamic Programming: Model-Based Solution Techniques	19
1.6.1	Fundamental DP Algorithms	20
1.6.2	Efficient DP Algorithms	24
1.7	Reinforcement Learning: Model-Free Solution Techniques	27
1.7.1	Temporal Difference Learning	29
1.7.2	Monte Carlo Methods	33
1.7.3	Efficient Exploration and Value Updating	34
1.8	Conclusions	39
	References	39

Part II Efficient Solution Frameworks

2	Batch Reinforcement Learning	45
	<i>Sascha Lange, Thomas Gabel, Martin Riedmiller</i>	
2.1	Introduction	45
2.2	The Batch Reinforcement Learning Problem	46
2.2.1	The Batch Learning Problem	46
2.2.2	The Growing Batch Learning Problem	48
2.3	Foundations of Batch RL Algorithms	49

2.4	Batch RL Algorithms	52
2.4.1	Kernel-Based Approximate Dynamic Programming	53
2.4.2	Fitted Q Iteration	55
2.4.3	Least-Squares Policy Iteration	57
2.4.4	Identifying Batch Algorithms	58
2.5	Theory of Batch RL	60
2.6	Batch RL in Practice	61
2.6.1	Neural Fitted Q Iteration (NFQ)	61
2.6.2	NFQ in Control Applications	63
2.6.3	Batch RL for Learning in Multi-agent Systems	65
2.6.4	Deep Fitted Q Iteration	67
2.6.5	Applications/ Further References	69
2.7	Summary	70
	References	71
3	Least-Squares Methods for Policy Iteration	75
	<i>Lucian Buşoniu, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, Robert Babuška, Bart De Schutter</i>	
3.1	Introduction	76
3.2	Preliminaries: Classical Policy Iteration	77
3.3	Least-Squares Methods for Approximate Policy Evaluation	79
3.3.1	Main Principles and Taxonomy	79
3.3.2	The Linear Case and Matrix Form of the Equations	81
3.3.3	Model-Free Implementations	85
3.3.4	Bibliographical Notes	89
3.4	Online Least-Squares Policy Iteration	89
3.5	Example: Car on the Hill	91
3.6	Performance Guarantees	94
3.6.1	Asymptotic Convergence and Guarantees	95
3.6.2	Finite-Sample Guarantees	98
3.7	Further Reading	104
	References	106
4	Learning and Using Models	111
	<i>Todd Hester, Peter Stone</i>	
4.1	Introduction	112
4.2	What Is a Model?	113
4.3	Planning	115
4.3.1	Monte Carlo Methods	115
4.4	Combining Models and Planning	118
4.5	Sample Complexity	120
4.6	Factored Domains	122
4.7	Exploration	126
4.8	Continuous Domains	130
4.9	Empirical Comparisons	133
4.10	Scaling Up	135

4.11	Conclusion	137
	References	138
5	Transfer in Reinforcement Learning: A Framework and a Survey ...	143
	<i>Alessandro Lazaric</i>	
5.1	Introduction	143
5.2	A Framework and a Taxonomy for Transfer in Reinforcement Learning	145
5.2.1	Transfer Framework	145
5.2.2	Taxonomy	148
5.3	Methods for Transfer from Source to Target with a Fixed State-Action Space	155
5.3.1	Problem Formulation	155
5.3.2	Representation Transfer	156
5.3.3	Parameter Transfer	158
5.4	Methods for Transfer across Tasks with a Fixed State-Action Space	159
5.4.1	Problem Formulation	159
5.4.2	Instance Transfer	160
5.4.3	Representation Transfer	161
5.4.4	Parameter Transfer	162
5.5	Methods for Transfer from Source to Target Tasks with a Different State-Action Spaces	164
5.5.1	Problem Formulation	164
5.5.2	Instance Transfer	166
5.5.3	Representation Transfer	166
5.5.4	Parameter Transfer	167
5.6	Conclusions and Open Questions	168
	References	169
6	Sample Complexity Bounds of Exploration	175
	<i>Lihong Li</i>	
6.1	Introduction	175
6.2	Preliminaries	176
6.3	Formalizing Exploration Efficiency	178
6.3.1	Sample Complexity of Exploration and PAC-MDP	178
6.3.2	Regret Minimization	180
6.3.3	Average Loss	182
6.3.4	Bayesian Framework	183
6.4	A Generic PAC-MDP Theorem	184
6.5	Model-Based Approaches	186
6.5.1	Rmax	186
6.5.2	A Generalization of Rmax	188
6.6	Model-Free Approaches	196
6.7	Concluding Remarks	199
	References	200

Part III Constructive-Representational Directions

7 Reinforcement Learning in Continuous State and Action Spaces 207
Hado van Hasselt

7.1 Introduction 207

7.1.1 Markov Decision Processes in Continuous Spaces 208

7.1.2 Methodologies to Solve a Continuous MDP 211

7.2 Function Approximation 212

7.2.1 Linear Function Approximation 213

7.2.2 Non-linear Function Approximation 217

7.2.3 Updating Parameters 218

7.3 Approximate Reinforcement Learning 223

7.3.1 Value Approximation 223

7.3.2 Policy Approximation 229

7.4 An Experiment on a Double-Pole Cart Pole 238

7.5 Conclusion 242

References 243

8 Solving Relational and First-Order Logical Markov Decision Processes: A Survey 253
Martijn van Otterlo

8.1 Introduction to Sequential Decisions in Relational Worlds 253

8.1.1 MDPs: Representation and Generalization 254

8.1.2 Short History and Connections to Other Fields 256

8.2 Extending MDPs with Objects and Relations 257

8.2.1 Relational Representations and Logical Generalization . . 257

8.2.2 Relational Markov Decision Processes 258

8.2.3 Abstract Problems and Solutions 259

8.3 Model-Based Solution Techniques 261

8.3.1 The Structure of Bellman Backups 262

8.3.2 Exact Model-Based Algorithms 263

8.3.3 Approximate Model-Based Algorithms 266

8.4 Model-Free Solutions 268

8.4.1 Value-Function Learning with Fixed Generalization 269

8.4.2 Value Functions with Adaptive Generalization 270

8.4.3 Policy-Based Solution Techniques 274

8.5 Models, Hierarchies, and Bias 276

8.6 Current Developments 280

8.7 Conclusions and Outlook 283

References 283

9 Hierarchical Approaches 293
Bernhard Hengst

9.1 Introduction 293

9.2 Background 296

9.2.1 Abstract Actions 297

9.2.2	Semi-Markov Decision Problems	297
9.2.3	Structure	300
9.2.4	State Abstraction	301
9.2.5	Value-Function Decomposition	303
9.2.6	Optimality	303
9.3	Approaches to Hierarchical Reinforcement Learning (HRL)	305
9.3.1	Options	306
9.3.2	HAMQ-Learning	307
9.3.3	MAXQ	309
9.4	Learning Structure	313
9.4.1	HEXQ	315
9.5	Related Work and Ongoing Research	317
9.6	Summary	319
	References	319
10	Evolutionary Computation for Reinforcement Learning	325
	<i>Shimon Whiteson</i>	
10.1	Introduction	325
10.2	Neuroevolution	328
10.3	TWEANNs	330
10.3.1	Challenges	332
10.3.2	NEAT	333
10.4	Hybrids	334
10.4.1	Evolutionary Function Approximation	335
10.4.2	XCS	336
10.5	Coevolution	339
10.5.1	Cooperative Coevolution	339
10.5.2	Competitive Coevolution	342
10.6	Generative and Developmental Systems	343
10.7	On-Line Methods	345
10.7.1	Model-Based Methods	345
10.7.2	On-Line Evolutionary Computation	346
10.8	Conclusion	347
	References	348
Part IV Probabilistic Models of Self and Others		
11	Bayesian Reinforcement Learning	359
	<i>Nikos Vlassis, Mohammad Ghavamzadeh, Shie Mannor, Pascal Poupart</i>	
11.1	Introduction	359
11.2	Model-Free Bayesian Reinforcement Learning	361
11.2.1	Value-Function Based Algorithms	361
11.2.2	Policy Gradient Algorithms	365
11.2.3	Actor-Critic Algorithms	369
11.3	Model-Based Bayesian Reinforcement Learning	372
11.3.1	POMDP Formulation of Bayesian RL	372

11.3.2	Bayesian RL via Dynamic Programming	373
11.3.3	Approximate Online Algorithms	376
11.3.4	Bayesian Multi-Task Reinforcement Learning	377
11.3.5	Incorporating Prior Knowledge	379
11.4	Finite Sample Analysis and Complexity Issues	380
11.5	Summary and Discussion	382
	References	382
12	Partially Observable Markov Decision Processes	387
	<i>Matthijs T.J. Spaan</i>	
12.1	Introduction	387
12.2	Decision Making in Partially Observable Environments	389
12.2.1	POMDP Model	389
12.2.2	Continuous and Structured Representations	391
12.2.3	Memory for Optimal Decision Making	391
12.2.4	Policies and Value Functions	394
12.3	Model-Based Techniques	395
12.3.1	Heuristics Based on MDP Solutions	396
12.3.2	Value Iteration for POMDPs	397
12.3.3	Exact Value Iteration	400
12.3.4	Point-Based Value Iteration Methods	401
12.3.5	Other Approximate Methods	403
12.4	Decision Making Without a-Priori Models	404
12.4.1	Memoryless Techniques	405
12.4.2	Learning Internal Memory	405
12.5	Recent Trends	408
	References	409
13	Predictively Defined Representations of State	415
	<i>David Wingate</i>	
13.1	Introduction	416
13.1.1	What Is “State”?	416
13.1.2	Which Representation of State?	418
13.1.3	Why Predictions about the Future?	419
13.2	PSRs	420
13.2.1	Histories and Tests	421
13.2.2	Prediction of a Test	422
13.2.3	The System Dynamics Vector	422
13.2.4	The System Dynamics Matrix	423
13.2.5	Sufficient Statistics	424
13.2.6	State	424
13.2.7	State Update	425
13.2.8	Linear PSRs	425
13.2.9	Relating Linear PSRs to POMDPs	426
13.2.10	Theoretical Results on Linear PSRs	427
13.3	Learning a PSR Model	428

13.3.1	The Discovery Problem	428
13.3.2	The Learning Problem	429
13.3.3	Estimating the System Dynamics Matrix	429
13.4	Planning with PSRs	429
13.5	Extensions of PSRs	431
13.6	Other Models with Predictively Defined State	432
13.6.1	Observable Operator Models	433
13.6.2	The Predictive Linear-Gaussian Model	433
13.6.3	Temporal-Difference Networks	434
13.6.4	Diversity Automaton	435
13.6.5	The Exponential Family PSR	435
13.6.6	Transformed PSRs	436
13.7	Conclusion	436
	References	437
14	Game Theory and Multi-agent Reinforcement Learning	441
	<i>Ann Nowé, Peter Vrancx, Yann-Michaël De Hauwere</i>	
14.1	Introduction	441
14.2	Repeated Games	445
14.2.1	Game Theory	445
14.2.2	Reinforcement Learning in Repeated Games	449
14.3	Sequential Games	454
14.3.1	Markov Games	455
14.3.2	Reinforcement Learning in Markov Games	456
14.4	Sparse Interactions in Multi-agent System	461
14.4.1	Learning on Multiple Levels	461
14.4.2	Learning to Coordinate with Sparse Interactions	462
14.5	Further Reading	467
	References	467
15	Decentralized POMDPs	471
	<i>Frans A. Oliehoek</i>	
15.1	Introduction	471
15.2	The Decentralized POMDP Framework	473
15.3	Histories and Policies	475
15.3.1	Histories	475
15.3.2	Policies	476
15.3.3	Structure in Policies	477
15.3.4	The Quality of Joint Policies	479
15.4	Solution of Finite-Horizon Dec-POMDPs	480
15.4.1	Brute Force Search and Dec-POMDP Complexity	480
15.4.2	Alternating Maximization	481
15.4.3	Optimal Value Functions for Dec-POMDPs	481
15.4.4	Forward Approach: Heuristic Search	485
15.4.5	Backwards Approach: Dynamic Programming	489
15.4.6	Other Finite-Horizon Methods	493

15.5 Further Topics 493

15.5.1 Generalization and Special Cases 493

15.5.2 Infinite-Horizon Dec-POMDPs 495

15.5.3 Reinforcement Learning 496

15.5.4 Communication 497

References 498

Part V Domains and Background

16 Psychological and Neuroscientific Connections with Reinforcement Learning 507

Ashvin Shah

16.1 Introduction 507

16.2 Classical (or Pavlovian) Conditioning 508

16.2.1 Behavior 509

16.2.2 Theory 511

16.2.3 Summary and Additional Considerations 512

16.3 Operant (or Instrumental) Conditioning 513

16.3.1 Behavior 513

16.3.2 Theory 514

16.3.3 Model-Based Versus Model-Free Control 516

16.3.4 Summary and Additional Considerations 517

16.4 Dopamine 518

16.4.1 Dopamine as a Reward Prediction Error 518

16.4.2 Dopamine as a General Reinforcement Signal 520

16.4.3 Summary and Additional Considerations 521

16.5 The Basal Ganglia 521

16.5.1 Overview of the Basal Ganglia 522

16.5.2 Neural Activity in the Striatum 523

16.5.3 Cortico-basal Ganglia-thalamic Loops 524

16.5.4 Summary and Additional Considerations 526

16.6 Chapter Summary 527

References 528

17 Reinforcement Learning in Games 539

István Szita

17.1 Introduction 539

17.1.1 Aims and Structure 540

17.1.2 Scope 541

17.2 A Showcase of Games 541

17.2.1 Backgammon 542

17.2.2 Chess 545

17.2.3 Go 550

17.2.4 Tetris 555

17.2.5 Real-Time Strategy Games 558

17.3 Challenges of Applying Reinforcement Learning to Games 561

17.3.1	Representation Design	561
17.3.2	Exploration	564
17.3.3	Source of Training Data	565
17.3.4	Dealing with Missing Information	566
17.3.5	Opponent Modelling	567
17.4	Using RL in Games	568
17.4.1	Opponents That Maximize Fun	568
17.4.2	Development-Time Learning	570
17.5	Closing Remarks	571
	References	572
18	Reinforcement Learning in Robotics: A Survey	579
	<i>Jens Kober, Jan Peters</i>	
18.1	Introduction	579
18.2	Challenges in Robot Reinforcement Learning	581
18.2.1	Curse of Dimensionality	582
18.2.2	Curse of Real-World Samples	583
18.2.3	Curse of Real-World Interactions	584
18.2.4	Curse of Model Errors	584
18.2.5	Curse of Goal Specification	585
18.3	Foundations of Robot Reinforcement Learning	585
18.3.1	Value Function Approaches	586
18.3.2	Policy Search	588
18.4	Tractability through Representation	589
18.4.1	Smart State-Action Discretization	590
18.4.2	Function Approximation	592
18.4.3	Pre-structured Policies	592
18.5	Tractability through Prior Knowledge	594
18.5.1	Prior Knowledge through Demonstrations	594
18.5.2	Prior Knowledge through Task Structuring	596
18.5.3	Directing Exploration with Prior Knowledge	596
18.6	Tractability through Simulation	596
18.6.1	Role of Models	597
18.6.2	Mental Rehearsal	598
18.6.3	Direct Transfer from Simulated to Real Robots	599
18.7	A Case Study: Ball-in-a-Cup	599
18.7.1	Experimental Setting: Task and Reward	599
18.7.2	Appropriate Policy Representation	601
18.7.3	Generating a Teacher's Demonstration	601
18.7.4	Reinforcement Learning by Policy Search	601
18.7.5	Use of Simulations in Robot Reinforcement Learning ...	603
18.7.6	Alternative Approach with Value Function Methods	603
18.8	Conclusion	603
	References	604

Part VI Closing

19 Conclusions, Future Directions and Outlook 613
Marco Wiering, Martijn van Otterlo

19.1 Looking Back 613

19.1.1 What Has Been Accomplished?..... 613

19.1.2 Which Topics Were Not Included? 614

19.2 Looking into the Future 620

19.2.1 Things That Are Not Yet Known..... 620

19.2.2 Seemingly Impossible Applications for RL 622

19.2.3 Interesting Directions 623

19.2.4 Experts on Future Developments..... 624

References 626

Index 631