

Table of Contents

Preface	xvii
List of Authors	xix
PART 1. MDPs: MODELS AND METHODS	1
Chapter 1. Markov Decision Processes	3
Frédéric GARCIA and Emmanuel RACHELSON	
1.1. Introduction	3
1.2. Markov decision problems	4
1.2.1. Markov decision processes	4
1.2.2. Action policies	7
1.2.3. Performance criterion	8
1.3. Value functions	9
1.3.1. The finite criterion	10
1.3.2. The γ -discounted criterion	10
1.3.3. The total reward criterion	11
1.3.4. The average reward criterion	11
1.4. Markov policies	12
1.4.1. Equivalence of history-dependent and Markov policies	12
1.4.2. Markov policies and valued Markov chains	13
1.5. Characterization of optimal policies	14
1.5.1. The finite criterion	14
1.5.1.1. Optimality equations	14
1.5.1.2. Evaluation of a deterministic Markov policy	15
1.5.2. The discounted criterion	16
1.5.2.1. Evaluation of a stationary Markov policy	16
1.5.2.2. Optimality equations	17
1.5.3. The total reward criterion	22

1.5.4. The average reward criterion	24
1.5.4.1. Evaluation of a stationary Markov policy	24
1.5.4.2. Optimality equations	27
1.6. Optimization algorithms for MDPs	28
1.6.1. The finite criterion	28
1.6.2. The discounted criterion	28
1.6.2.1. Linear programming	28
1.6.2.2. The value iteration algorithm	29
1.6.2.3. The policy iteration algorithm	30
1.6.3. The total reward criterion	34
1.6.3.1. Positive MDPs	34
1.6.3.2. Negative MDPs	34
1.6.4. The average criterion	35
1.6.4.1. Relative value iteration algorithm	35
1.6.4.2. Modified policy iteration algorithm	36
1.7. Conclusion and outlook	37
1.8. Bibliography	37
Chapter 2. Reinforcement Learning	39
Olivier SIGAUD and Frédéric GARCIA	
2.1. Introduction	39
2.1.1. Historical overview	39
2.2. Reinforcement learning: a global view	40
2.2.1. Reinforcement learning as approximate dynamic programming	41
2.2.2. Temporal, non-supervised and trial-and-error based learning	42
2.2.3. Exploration versus exploitation	42
2.2.4. General preliminaries on estimation methods	44
2.3. Monte Carlo methods	45
2.4. From Monte Carlo to temporal difference methods	45
2.5. Temporal difference methods	46
2.5.1. The TD(0) algorithm	47
2.5.2. The SARSA algorithm	48
2.5.3. The Q-learning algorithm	49
2.5.4. The TD(λ), SARSA(λ) and Q(λ) algorithms	51
2.5.5. Eligibility traces and TD(λ)	52
2.5.6. From TD(λ) to SARSA(λ)	55
2.5.7. Q(λ)	56
2.5.8. The R-learning algorithm	58
2.6. Model-based methods: learning a model	59
2.6.1. Dyna architectures	60
2.6.2. The E^3 algorithm	61
2.6.3. The R_{\max} algorithm	62

2.7. Conclusion	63
2.8. Bibliography	63
Chapter 3. Approximate Dynamic Programming	67
Rémi MUNOS	
3.1. Introduction	68
3.2. Approximate value iteration (AVI)	70
3.2.1. Sample-based implementation and supervised learning	71
3.2.2. Analysis of the AVI algorithm	73
3.2.3. Numerical illustration	74
3.3. Approximate policy iteration (API)	77
3.3.1. Analysis in L^∞ -norm of the API algorithm	77
3.3.2. Approximate policy evaluation	79
3.3.3. Linear approximation and least-squares methods	80
3.3.3.1. TD(λ)	81
3.3.3.2. Least-squares methods	82
3.3.3.3. Linear approximation of the state-action value function	85
3.4. Direct minimization of the Bellman residual	87
3.5. Towards an analysis of dynamic programming in L^p -norm	88
3.5.1. Intuition of an L^p analysis in dynamic programming	89
3.5.2. PAC bounds for RL algorithms	91
3.6. Conclusions	93
3.7. Bibliography	93
Chapter 4. Factored Markov Decision Processes	99
Thomas DEGRIS and Olivier SIGAUD	
4.1. Introduction	99
4.2. Modeling a problem with an FMDP	100
4.2.1. Representing the state space	100
4.2.2. The <i>Coffee Robot</i> example	100
4.2.3. Decomposition and function-specific independence	101
4.2.3.1. Transition function decomposition	102
4.2.3.2. Dynamic Bayesian networks in FMDPs	103
4.2.3.3. Factored model of the transition function in an FMDP	104
4.2.3.4. Factored model of the reward function	105
4.2.4. Context-specific independence	107
4.3. Planning with FMDPs	108
4.3.1. Structured policy iteration and structured value iteration	108
4.3.1.1. Decision trees	108
4.3.1.2. Representation of the transition function	108
4.3.1.3. Representation of the reward function	110
4.3.1.4. Representation of a policy	110
4.3.1.5. Representation of the value function	111
4.3.1.6. Algorithms	112

4.3.2. SPUDD: stochastic planning using decision diagrams	112
4.3.2.1. Representing the functions of an FMDP with ADDs	113
4.3.2.2. Algorithm	114
4.3.3. Approximate linear programming in FMDPs	115
4.3.3.1. Representations	116
4.3.3.2. Representation of the transition function	116
4.3.3.3. Representation of the reward function	117
4.3.3.4. Policy representation	118
4.3.3.5. Representation of the value function	119
4.3.3.6. Algorithms	122
4.4. Perspectives and conclusion	122
4.5. Bibliography	123
Chapter 5. Policy-Gradient Algorithms	127
Olivier BUFFET	
5.1. Reminder about the notion of gradient	128
5.1.1. Gradient of a function	128
5.1.2. Gradient descent	129
5.2. Optimizing a parameterized policy with a gradient algorithm	130
5.2.1. Application to MDPs: overview	130
5.2.1.1. First attempt to define a parameterized policy	131
5.2.1.2. Example of definition of a parameterized policy	131
5.2.2. Finite difference method	132
5.2.3. Gradient estimation of f in an MDP, case of the finite time horizon	133
5.2.3.1. Time horizon 1	133
5.2.3.2. Time horizon T	134
5.2.4. Extension to the infinite time horizon: discounted criterion, average criterion	137
5.2.4.1. Case of a regenerative process	138
5.2.4.2. Using a moving window	138
5.2.4.3. Using a discount factor	139
5.2.5. Partially observable case	139
5.2.6. Continuous action spaces	141
5.3. Actor-critic methods	143
5.3.1. A gradient estimator using the Q -values	143
5.3.2. Compatibility with an approximate value function	144
5.3.2.1. Approximating Q^θ	144
5.3.2.2. Compatibility of the approximators	145
5.3.3. An actor-critic algorithm	146
5.4. Complements	147
5.5. Conclusion	150
5.6. Bibliography	150

Chapter 6. Online Resolution Techniques	153
Laurent PÉRET and Frédéric GARCIA	
6.1. Introduction	153
6.1.1. Exploiting time online	153
6.1.2. Online search by simulation	154
6.2. Online algorithms for solving an MDP	155
6.2.1. Offline algorithms, online algorithms	155
6.2.2. Problem formalization	155
6.2.2.1. Forward search over a reasoning horizon	156
6.2.2.2. Tree or graph?	157
6.2.2.3. Complexity and efficiency of the forward search	157
6.2.3. Online heuristic search algorithms for MDPs	158
6.2.3.1. General principles	158
6.2.3.2. The RTDP algorithm	159
6.2.3.3. The <i>LAO*</i> algorithm	160
6.2.4. Kearns, Mansour and Ng’s simulation-based algorithm	163
6.2.4.1. Complexity and convergence of Kearns <i>et al.</i> ’s algorithm	165
6.2.4.2. Efficiency and practical considerations	165
6.2.5. Tesauro and Galperin’s rollout algorithm	165
6.2.5.1. Complexity and convergence of the rollout algorithm	166
6.2.5.2. Efficiency of the rollout algorithm	166
6.3. Controlling the search	167
6.3.1. Error bounds and pathology of the forward search	168
6.3.1.1. Pathology of the simulation-based forward search	168
6.3.1.2. Searching for a good compromise between depth and width	170
6.3.2. Iterative allocation of simulations	170
6.3.2.1. Multi-armed bandits and exploration in MDPs	171
6.3.3. Focused reinforcement learning	173
6.3.3.1. Global sampling error estimation	174
6.3.3.2. Organizing the search in successive trajectories	175
6.3.3.3. Convergence and practical considerations	176
6.3.4. Controlled rollout	178
6.3.4.1. Choice of the horizon	178
6.3.4.2. Iterative allocation of simulations	179
6.4. Conclusion	180
6.5. Bibliography	180
PART 2. BEYOND MDPs	185
Chapter 7. Partially Observable Markov Decision Processes	187
Alain DUTECH and Bruno SCHERRER	
7.1. Formal definitions for POMDPs	188
7.1.1. Definition of a POMDP	188

7.1.2. Performance criteria	189
7.1.3. Information state	190
7.1.3.1. Definition	190
7.1.3.2. Complete information state	191
7.1.3.3. Sufficient statistics	191
7.1.3.4. Belief states	191
7.1.4. Policy	193
7.1.5. Value function	195
7.2. Non-Markovian problems: incomplete information	196
7.2.1. Adapted policies	196
7.2.2. Discounted reward	197
7.2.2.1. Adapted stochastic policies	197
7.2.2.2. Adapted value function	198
7.2.2.3. Convergence of adapted algorithms	199
7.2.3. Adapted algorithms and adapted average reward criterion	201
7.3. Computation of an exact policy on information states	202
7.3.1. The general case	202
7.3.1.1. Finite horizon	202
7.3.1.2. Infinite horizon	203
7.3.2. Belief states and piecewise linear value function	203
7.3.2.1. Choice of the θ vectors	206
7.3.2.2. Infinite horizon	207
7.4. Exact value iteration algorithms	207
7.4.1. Steps for the dynamic programming operator	207
7.4.2. A parsimonious representation of V	210
7.4.2.1. Region	210
7.4.2.2. Parsimonious representation	212
7.4.2.3. Pruning of dominated vectors	213
7.4.2.4. Pruning	213
7.4.2.5. Choice of a vector for a belief state	215
7.4.3. The WITNESS algorithm	216
7.4.3.1. Neighborhood of a vector	216
7.4.3.2. The algorithm	218
7.4.4. Iterative pruning	219
7.4.4.1. Complete enumeration	219
7.4.4.2. Incremental enumeration	220
7.5. Policy iteration algorithms	222
7.6. Conclusion and perspectives	223
7.7. Bibliography	225
Chapter 8. Stochastic Games	229
Andriy BURKOV, Laëtitia MATIGNON and Brahim CHAIB-DRAA	
8.1. Introduction	229

8.2. Background on game theory	230
8.2.1. Some basic definitions	230
8.2.1.1. Criteria distinguishing different forms of games	230
8.2.2. Static games of complete information	232
8.2.2.1. Games in strategic form, pure strategy, mixed strategy	232
8.2.2.2. Zero-sum game and minimax	234
8.2.2.3. Equilibrium in dominating strategies	236
8.2.2.4. Nash equilibrium	238
8.2.3. Dynamic games of complete information	240
8.2.3.1. Games in extensive form with perfect information	241
8.2.3.2. Repeated games	243
8.3. Stochastic games	245
8.3.1. Definition and equilibrium of a stochastic game	246
8.3.2. Solving stochastic games	248
8.3.2.1. Game model available	249
8.3.2.2. Game model unavailable, \mathbf{A}_{-i} observed, equilibrium learners	252
8.3.2.3. Game model unavailable, \mathbf{A}_{-i} observed, opponent modeling	254
8.3.2.4. Game model unavailable, \mathbf{A}_{-i} not observed	256
8.3.3. Complexity and scalability of multi-agent learning algorithms	262
8.3.4. Beyond the search for an equilibrium	264
8.3.4.1. Efficient play	264
8.3.4.2. Regret minimization	265
8.3.4.3. Metastrategies	267
8.3.5. Discussion	267
8.4. Conclusion and outlook	269
8.5. Bibliography	270
Chapter 9. DEC-MDP/POMDP	277
Aurélie BEYNIER, François CHARPILLET, Daniel SZER and Abdel-Ilah MOUADDIB	
9.1. Introduction	277
9.2. Preliminaries	278
9.3. Multiagent Markov decision processes	279
9.4. Decentralized control and local observability	280
9.4.1. Decentralized Markov decision processes	281
9.4.2. Multiagent team decision problems	283
9.4.3. Complexity	285
9.5. Sub-classes of DEC-POMDPs	285
9.5.1. Transition and observation independence	285
9.5.2. Goal oriented DEC-POMDPs	287
9.5.3. DEC-MDPs with constraints	288

9.5.3.1. Event-driven DEC-MDPs	289
9.5.3.2. Opportunity-cost DEC-MDPs	289
9.5.3.3. Problem statement	289
9.5.3.4. The OC-DEC-MDP model	290
9.5.4. Communication and DEC-POMDPs	292
9.5.4.1. DEC-POMDPs with communication	293
9.5.4.2. Complexity results	294
9.5.4.3. Discussion	294
9.6. Algorithms for solving DEC-POMDPs	295
9.6.1. Optimal algorithms	296
9.6.1.1. Dynamic programming for DEC-POMDPs	296
9.6.1.2. Heuristic search for DEC-POMDPs	299
9.6.1.3. Optimal algorithms for sub-classes of DEC-POMDPs	303
9.6.2. Approximate algorithms	303
9.6.2.1. Heuristics and approximate dynamic programming	304
9.6.2.2. Bounded memory	305
9.6.2.3. Co-evolutionary algorithms	305
9.6.2.4. Gradient descent for policy search	307
9.6.2.5. Bayesian games	307
9.6.2.6. Heuristics for communicating agents	308
9.6.2.7. Approximate solutions for OC-DEC-MDPs	308
9.7. Applicative scenario: multirobot exploration	310
9.8. Conclusion and outlook	312
9.9. Bibliography	313
Chapter 10. Non-Standard Criteria	319
Matthieu BOUSSARD, Maroua BOUZID, Abdel-illah MOUADDIB, Régis SABBADIN and Paul WENG	
10.1. Introduction	319
10.2. Multicriteria approaches	320
10.2.1. Multicriteria decision-making	321
10.2.2. Multicriteria MDPS	322
10.2.2.1. Operators for multicriteria decision-making	323
10.3. Robustness in MDPS	327
10.4. Possibilistic MDPS	329
10.4.1. Possibilistic counterpart of expected utility	330
10.4.2. Possibilistic dynamic programming	333
10.4.2.1. Finite horizon	333
10.4.2.2. Value iteration	334
10.4.2.3. Policy iteration	338
10.4.3. Extensions of possibilistic MDPS	338
10.4.3.1. Possibilistic reinforcement learning	339
10.4.3.2. Possibilistic partially observable MDPS	340
10.4.3.3. Possibilistic influence diagrams (PID)	342

10.5. Algebraic MDPs	342
10.5.1. Background	343
10.5.1.1. Semirings	343
10.5.1.2. Plausibility measures	344
10.5.1.3. Generalized expected utility	345
10.5.2. Definition of an algebraic MDP	345
10.5.3. Value function of a policy	347
10.5.4. Conditions	348
10.5.5. Examples of AMDPs	349
10.5.5.1. Probabilistic multicriteria AMDP	349
10.5.5.2. Possibilistic multicriteria AMDPs	350
10.5.5.3. AMDPs whose rewards are non-decreasing functions	351
10.6. Conclusion	354
10.7. Bibliography	355
PART 3. APPLICATIONS	361
Chapter 11. Online Learning for Micro-Object Manipulation	363
Guillaume LAURENT	
11.1. Introduction	363
11.2. Manipulation device	364
11.2.1. Micro-positioning by pushing	364
11.2.2. Manipulation device	365
11.2.3. Control loop	366
11.2.4. Representation of the manipulation task as an MDP	366
11.2.4.1. Definition of the state space	366
11.2.4.2. Definition of the action space	367
11.2.4.3. Definition of the reward function	367
11.2.4.4. Definition of an episode	367
11.3. Choice of the reinforcement learning algorithm	367
11.3.1. Characteristics of the MDP	367
11.3.2. A suitable algorithm: <i>STM-Q</i>	368
11.4. Experimental results	370
11.4.1. Experimental setup	370
11.4.2. Results	370
11.5. Conclusion	373
11.6. Bibliography	373
Chapter 12. Conservation of Biodiversity	375
Iadine CHADÈS	
12.1. Introduction	375
12.2. When to protect, survey or surrender cryptic endangered species	376
12.2.1. Surveying and managing the Sumatran tiger	376

12.2.2. The model	377
12.2.3. Results	377
12.2.4. Extension to more than one population	379
12.3. Can sea otters and abalone co-exist?	381
12.3.1. Abalone and sea otters: two endangered species	381
12.3.2. The models	382
12.3.2.1. Population dynamics of abalone	382
12.3.2.2. Sea otter population model	383
12.3.2.3. States	384
12.3.2.4. Decisions	385
12.3.2.5. Interaction between sea otters and abalone	386
12.3.2.6. Multicriteria objective and reward function	386
12.3.3. Methods	387
12.3.4. Results	387
12.3.4.1. Scenario 1: sea otter reintroduction and anti-poaching enforcement	387
12.3.4.2. Scenario 2: control of sea otters	389
12.3.4.3. Scenario 3: combined action of sea otter control and anti-poaching	389
12.3.5. Conclusion	389
12.4. Other applications in conservation biology and discussions	391
12.5. Bibliography	392
Chapter 13. Autonomous Helicopter Searching for a Landing Area in an Uncertain Environment	395
Patrick FABIANI and Florent TEICHTEIL-KÖNIGSBUCH	
13.1. Introduction	395
13.2. Exploration scenario	397
13.2.1. Planning problem	398
13.2.2. States and actions	399
13.2.3. Uncertainties	400
13.2.4. Optimization criterion	400
13.2.5. Formalization of the decision problem	401
13.3. Embedded control and decision architecture	401
13.3.1. Global view	401
13.3.2. Multi-thread planning triggered by the supervisor	403
13.3.2.1. Policy optimization	403
13.3.2.2. Dialogue with the supervisor	404
13.4. Incremental stochastic dynamic programming	404
13.4.1. Obtaining the initial safe policy quickly	405
13.4.2. Generating the sub-space of reachable states	405
13.4.3. Local policy optimization	406
13.4.4. Launching local replanning processes	407

13.5. Flight tests and return on experience	407
13.6. Conclusion	410
13.7. Bibliography	410
Chapter 14. Resource Consumption Control for an Autonomous Robot . .	413
Simon LE GLOANNEC and Abdel-Ilah MOUADDIB	
14.1. The rover's mission	414
14.2. Progressive processing formalism	415
14.3. MDP/PRU model	416
14.3.1. States	416
14.3.2. Actions	417
14.3.3. Transition function	418
14.3.4. Reward function	418
14.4. Policy calculation	418
14.4.1. Value function	419
14.4.2. Propagation algorithm	419
14.5. How to model a real mission	419
14.6. Extensions	422
14.7. Conclusion	423
14.8. Bibliography	423
Chapter 15. Operations Planning	425
Sylvie THIÉBAUX and Olivier BUFFET	
15.1. Operations planning	425
15.1.1. Intuition	425
15.1.1.1. Problem features	426
15.1.1.2. Plans	427
15.1.2. Formal definitions	428
15.1.2.1. Planning problem, operations	429
15.1.2.2. Execution	431
15.1.2.3. Decision epochs, states, plans	432
15.1.2.4. Objective	432
15.2. MDP value function approaches	433
15.2.1. Formalizations, CoMDP	433
15.2.1.1. States, actions, transitions	433
15.2.1.2. Rewards, costs	434
15.2.2. Algorithms	435
15.2.2.1. (L)RTDP	435
15.2.2.2. Memory management	435
15.2.2.3. Reduction of the number of updates	436
15.2.2.4. Hybrid algorithms	436
15.2.2.5. Algorithms with upper bounds	437

15.2.3. Heuristics	438
15.2.3.1. Basic heuristics	438
15.2.3.2. Heuristics obtained by relaxation of the CoMDP	439
15.2.3.3. Planning graph heuristics	440
15.3. Reinforcement learning: FPG	442
15.3.1. Employing approximate methods	442
15.3.2. Parameterized policy	443
15.3.2.1. Inputs	443
15.3.2.2. Outputs	443
15.3.2.3. Function approximator	444
15.3.3. Gradient methods	445
15.3.3.1. Terminating an execution	445
15.3.3.2. Choice of OL_{pomdp}	445
15.3.3.3. Optimized criterion	445
15.3.4. Improving FPG	446
15.4. Experiments	446
15.5. Conclusion and outlook	448
15.6. Bibliography	450
Index	453