Contents

1 Introduction **1**
- 1.1 The dynamic programming and reinforcement learning problem . . 2
- 1.2 Approximation in dynamic programming and reinforcement learning 5
- 1.3 About this book 8

2 An introduction to dynamic programming and reinforcement learning **11**
- 2.1 Introduction 11
- 2.2 Markov decision processes 14
  - 2.2.1 Deterministic setting 14
  - 2.2.2 Stochastic setting 19
- 2.3 Value iteration 23
  - 2.3.1 Model-based value iteration 23
  - 2.3.2 Model-free value iteration and the need for exploration . . . 28
- 2.4 Policy iteration 30
  - 2.4.1 Model-based policy iteration 31
  - 2.4.2 Model-free policy iteration 37
- 2.5 Policy search 38
- 2.6 Summary and discussion 41

3 Dynamic programming and reinforcement learning in large and continuous spaces **43**
- 3.1 Introduction 43
- 3.2 The need for approximation in large and continuous spaces 47
- 3.3 Approximation architectures 49
  - 3.3.1 Parametric approximation 49
  - 3.3.2 Nonparametric approximation 51
  - 3.3.3 Comparison of parametric and nonparametric approximation 53
  - 3.3.4 Remarks 54
- 3.4 Approximate value iteration 54
  - 3.4.1 Model-based value iteration with parametric approximation 55
  - 3.4.2 Model-free value iteration with parametric approximation . 58
  - 3.4.3 Value iteration with nonparametric approximation 62
  - 3.4.4 Convergence and the role of nonexpansive approximation . 63
  - 3.4.5 Example: Approximate Q-iteration for a DC motor 66
- 3.5 Approximate policy iteration 71
  - 3.5.1 Value iteration-like algorithms for approximate policy evaluation 73

3.5.2 Model-free policy evaluation with linearly parameterized approximation 74
3.5.3 Policy evaluation with nonparametric approximation 84
3.5.4 Model-based approximate policy evaluation with rollouts . . 84
3.5.5 Policy improvement and approximate policy iteration 85
3.5.6 Theoretical guarantees 88
3.5.7 Example: Least-squares policy iteration for a DC motor . . 90
3.6 Finding value function approximators automatically 95
3.6.1 Basis function optimization 96
3.6.2 Basis function construction 98
3.6.3 Remarks 100
3.7 Approximate policy search 101
3.7.1 Policy gradient and actor-critic algorithms 102
3.7.2 Gradient-free policy search 107
3.7.3 Example: Gradient-free policy search for a DC motor 109
3.8 Comparison of approximate value iteration, policy iteration, and policy search 113
3.9 Summary and discussion 114

4 Approximate value iteration with a fuzzy representation 117
4.1 Introduction 117
4.2 Fuzzy Q-iteration 119
4.2.1 Approximation and projection mappings of fuzzy Q-iteration 119
4.2.2 Synchronous and asynchronous fuzzy Q-iteration 123
4.3 Analysis of fuzzy Q-iteration 127
4.3.1 Convergence 127
4.3.2 Consistency 135
4.3.3 Computational complexity 140
4.4 Optimizing the membership functions 141
4.4.1 A general approach to membership function optimization . . 141
4.4.2 Cross-entropy optimization 143
4.4.3 Fuzzy Q-iteration with cross-entropy optimization of the membership functions 144
4.5 Experimental study 145
4.5.1 DC motor: Convergence and consistency study 146
4.5.2 Two-link manipulator: Effects of action interpolation, and comparison with fitted Q-iteration 152
4.5.3 Inverted pendulum: Real-time control 157
4.5.4 Car on the hill: Effects of membership function optimization 160
4.6 Summary and discussion 164

5 Approximate policy iteration for online learning and continuous-action control 167
5.1 Introduction 167
5.2 A recapitulation of least-squares policy iteration 168

5.3 Online least-squares policy iteration 170
5.4 Online LSPI with prior knowledge 173
5.4.1 Online LSPI with policy approximation 174
5.4.2 Online LSPI with monotonic policies 175
5.5 LSPI with continuous-action, polynomial approximation 177
5.6 Experimental study 180
5.6.1 Online LSPI for the inverted pendulum 180
5.6.2 Online LSPI for the two-link manipulator 192
5.6.3 Online LSPI with prior knowledge for the DC motor 195
5.6.4 LSPI with continuous-action approximation for the inverted pendulum 198
5.7 Summary and discussion 201

6 Approximate policy search with cross-entropy optimization of basis functions 205
6.1 Introduction 205
6.2 Cross-entropy optimization 207
6.3 Cross-entropy policy search 209
6.3.1 General approach 209
6.3.2 Cross-entropy policy search with radial basis functions . . . 213
6.4 Experimental study 216
6.4.1 Discrete-time double integrator 216
6.4.2 Bicycle balancing 223
6.4.3 Structured treatment interruptions for HIV infection control 229
6.5 Summary and discussion 233

Appendix A Extremely randomized trees 235
A.1 Structure of the approximator 235
A.2 Building and using a tree 236

Appendix B The cross-entropy method 239
B.1 Rare-event simulation using the cross-entropy method 239
B.2 Cross-entropy optimization 242

Symbols and abbreviations 245

Bibliography 249

List of algorithms 267

Index 269