

# Contents

## 1. Exact Dynamic Programming

1.1. Deterministic Dynamic Programming . . . . .	p. 2
1.1.1. Deterministic Problems . . . . .	p. 2
1.1.2. The Dynamic Programming Algorithm . . . . .	p. 7
1.1.3. Approximation in Value Space . . . . .	p. 12
1.2. Stochastic Dynamic Programming . . . . .	p. 14
1.3. Examples, Variations, and Simplifications . . . . .	p. 18
1.3.1. Deterministic Shortest Path Problems . . . . .	p. 19
1.3.2. Discrete Deterministic Optimization . . . . .	p. 21
1.3.3. Problems with a Termination State . . . . .	p. 25
1.3.4. Forecasts . . . . .	p. 26
1.3.5. Problems with Uncontrollable State Components . . . . .	p. 29
1.3.6. Partial State Information and Belief States . . . . .	p. 34
1.3.7. Linear Quadratic Optimal Control . . . . .	p. 38
1.3.8. Systems with Unknown Parameters - Adaptive Control . . . . .	p. 40
1.4. Reinforcement Learning and Optimal Control - Some Terminology . . . . .	p. 43
1.5. Notes and Sources . . . . .	p. 45

## 2. Approximation in Value Space

2.1. Approximation Approaches in Reinforcement Learning . . . . .	p. 50
2.1.1. General Issues of Approximation in Value Space . . . . .	p. 54
2.1.2. Off-Line and On-Line Methods . . . . .	p. 56
2.1.3. Model-Based Simplification of the Lookahead Minimization . . . . .	p. 57
2.1.4. Model-Free Q-Factor Approximation in Value Space . . . . .	p. 58
2.1.5. Approximation in Policy Space on Top of Approximation in Value Space . . . . .	p. 61
2.1.6. When is Approximation in Value Space Effective? . . . . .	p. 62
2.2. Multistep Lookahead . . . . .	p. 64



2.2.1. Multistep Lookahead and Rolling Horizon . . . . .	p. 65
2.2.2. Multistep Lookahead and Deterministic Problems . . . . .	p. 67
2.3. Problem Approximation . . . . .	p. 69
2.3.1. Enforced Decomposition . . . . .	p. 69
2.3.2. Probabilistic Approximation - Certainty Equivalent Control . . . . .	p. 76
2.4. Rollout and the Policy Improvement Principle . . . . .	p. 83
2.4.1. On-Line Rollout for Deterministic Discrete Optimization . . . . .	p. 84
2.4.2. Stochastic Rollout and Monte Carlo Tree Search . . . . .	p. 95
2.4.3. Rollout with an Expert . . . . .	p. 104
2.5. On-Line Rollout for Deterministic Infinite-Spaces Problems - Optimization Heuristics . . . . .	p. 106
2.5.1. Model Predictive Control . . . . .	p. 108
2.5.2. Target Tubes and the Constrained Controllability Condition . . . . .	p. 115
2.5.3. Variants of Model Predictive Control . . . . .	p. 118
2.6. Notes and Sources . . . . .	p. 120
<b>3. Parametric Approximation</b>	
3.1. Approximation Architectures . . . . .	p. 126
3.1.1. Linear and Nonlinear Feature-Based Architectures . . . . .	p. 126
3.1.2. Training of Linear and Nonlinear Architectures . . . . .	p. 134
3.1.3. Incremental Gradient and Newton Methods . . . . .	p. 135
3.2. Neural Networks . . . . .	p. 149
3.2.1. Training of Neural Networks . . . . .	p. 153
3.2.2. Multilayer and Deep Neural Networks . . . . .	p. 157
3.3. Sequential Dynamic Programming Approximation . . . . .	p. 161
3.4. Q-Factor Parametric Approximation . . . . .	p. 162
3.5. Parametric Approximation in Policy Space by Classification . . . . .	p. 165
3.6. Notes and Sources . . . . .	p. 171
<b>4. Infinite Horizon Dynamic Programming</b>	
4.1. An Overview of Infinite Horizon Problems . . . . .	p. 174
4.2. Stochastic Shortest Path Problems . . . . .	p. 177
4.3. Discounted Problems . . . . .	p. 187
4.4. Semi-Markov Discounted Problems . . . . .	p. 192
4.5. Asynchronous Distributed Value Iteration . . . . .	p. 197
4.6. Policy Iteration . . . . .	p. 200
4.6.1. Exact Policy Iteration . . . . .	p. 200
4.6.2. Optimistic and Multistep Lookahead Policy Iteration . . . . .	p. 205
4.6.3. Policy Iteration for Q-factors . . . . .	p. 208



4.7. Notes and Sources . . . . .	p. 209
4.8. Appendix: Mathematical Analysis . . . . .	p. 211
4.8.1. Proofs for Stochastic Shortest Path Problems . . . . .	p. 212
4.8.2. Proofs for Discounted Problems . . . . .	p. 217
4.8.3. Convergence of Exact and Optimistic Policy Iteration . . . . .	p. 218
<b>5. Infinite Horizon Reinforcement Learning</b>	
5.1. Approximation in Value Space - Performance Bounds . . . . .	p. 222
5.1.1. Limited Lookahead . . . . .	p. 224
5.1.2. Rollout and Approximate Policy Improvement . . . . .	p. 227
5.1.3. Approximate Policy Iteration . . . . .	p. 232
5.2. Fitted Value Iteration . . . . .	p. 235
5.3. Simulation-Based Policy Iteration with Parametric Approximation . . . . .	p. 239
5.3.1. Self-Learning and Actor-Critic Methods . . . . .	p. 239
5.3.2. Model-Based Variant of a Critic-Only Method . . . . .	p. 241
5.3.3. Model-Free Variant of a Critic-Only Method . . . . .	p. 243
5.3.4. Implementation Issues of Parametric Policy Iteration . . . . .	p. 246
5.3.5. Convergence Issues of Parametric Policy Iteration - Oscillations . . . . .	p. 249
5.4. Q-Learning . . . . .	p. 253
5.4.1. Optimistic Policy Iteration with Parametric Q-Factor Approximation - SARSA and DQN . . . . .	p. 255
5.5. Additional Methods - Temporal Differences . . . . .	p. 256
5.6. Exact and Approximate Linear Programming . . . . .	p. 267
5.7. Approximation in Policy Space . . . . .	p. 270
5.7.1. Training by Cost Optimization - Policy Gradient, Cross-Entropy, and Random Search Methods . . . . .	p. 276
5.7.2. Expert-Based Supervised Learning . . . . .	p. 286
5.7.3. Approximate Policy Iteration, Rollout, and Approximation in Policy Space . . . . .	p. 288
5.8. Notes and Sources . . . . .	p. 293
5.9. Appendix: Mathematical Analysis . . . . .	p. 298
5.9.1. Performance Bounds for Multistep Lookahead . . . . .	p. 299
5.9.2. Performance Bounds for Rollout . . . . .	p. 301
5.9.3. Performance Bounds for Approximate Policy Iteration . . . . .	p. 304
<b>6. Aggregation</b>	
6.1. Aggregation with Representative States . . . . .	p. 308
6.1.1. Continuous State and Control Space Discretization . . . . .	p. 314
6.1.2. Continuous State Space - POMDP Discretization . . . . .	p. 315



6.2. Aggregation with Representative Features . . . . . p. 317

6.2.1. Hard Aggregation and Error Bounds . . . . . p. 320

6.2.2. Aggregation Using Features . . . . . p. 322

6.3. Methods for Solving the Aggregate Problem . . . . . p. 328

6.3.1. Simulation-Based Policy Iteration . . . . . p. 328

6.3.2. Simulation-Based Value Iteration and Q-Learning . . . . . p. 331

6.4. Feature-Based Aggregation with a Neural Network . . . . . p. 332

6.5. Biased Aggregation . . . . . p. 334

6.6. Notes and Sources . . . . . p. 337

6.7. Appendix: Mathematical Analysis . . . . . p. 340

**References** . . . . . p. 345

**Index** . . . . . p. 369