1 Sequential decision problems

1.1 The audience
1.2 The communities of sequential decision problems
1.3 Our universal modeling framework
1.4 Designing policies for sequential decision problems
  1.4.1 Policy search
  1.4.2 Policies based on lookahead approximations
  1.4.3 Mixing and matching
  1.4.4 Optimality of the four classes
  1.4.5 Pulling it all together
1.5 Learning
1.6 Themes
  1.6.1 Blending learning and optimization
  1.6.2 Bridging machine learning to sequential decisions
  1.6.3 From deterministic to stochastic optimization
  1.6.4 From single to multiple agents
1.7 Our modeling approach
1.8 How to read this book
## 1.8 Bibliographic notes

### 1.8.1 Organization of topics

- 21

### 1.8.2 How to read each chapter

- 25

### 1.8.3 Organization of exercises

- 25

### 1.9 Bibliographic notes

- 26

### Exercises

- 27

## 2 Canonical Problems and Applications

### 2.1 Canonical problems

- 31

#### 2.1.1 Stochastic search - derivative-based and derivative-free

- 32

#### 2.1.2 Decision trees

- 35

#### 2.1.3 Markov decision processes

- 36

#### 2.1.4 Optimal control

- 37

#### 2.1.5 Approximate dynamic programming

- 39

#### 2.1.6 Reinforcement learning

- 40

#### 2.1.7 Optimal stopping

- 42

#### 2.1.8 Stochastic programming

- 44

#### 2.1.9 The multiarmed bandit problem

- 45

#### 2.1.10 Simulation optimization

- 48

#### 2.1.11 Active learning

- 48

#### 2.1.12 Chance constrained programming

- 49

#### 2.1.13 Model predictive control

- 49

#### 2.1.14 Robust optimization

- 50

### 2.2 A universal modeling framework for sequential decision problems

- 51

#### 2.2.1 Our universal model for sequential decision problems

- 51

#### 2.2.2 A compact modeling presentation

- 54

#### 2.2.3 MDP/RL vs. optimal control modeling frameworks

- 54

### 2.3 Applications

- 55

#### 2.3.1 The newsvendor problems

- 55

#### 2.3.2 Inventory/storage problems

- 57

#### 2.3.3 Shortest path problems

- 60

#### 2.3.4 Some fleet management problems

- 62

#### 2.3.5 Pricing

- 63

#### 2.3.6 Medical decision making

- 64

#### 2.3.7 Scientific exploration

- 65

#### 2.3.8 Machine learning vs. sequential decision problems

- 65

### 2.4 Bibliographic notes

- 66

### Exercises

- 71

## 3 Online learning

### 3.1 Machine learning for sequential decisions

- 84

#### 3.1.1 Observations and data in stochastic optimization

- 84
3.1.2 Indexing input $x^n$ and response $y^{n+1}$ .......................... 85
3.1.3 Functions we are learning ................................................. 85
3.1.4 Sequential learning: from very little data to . . . more data .... 87
3.1.5 Approximation strategies ................................................. 87
3.1.6 From data analytics to decision analytics ......................... 89
3.1.7 Batch vs. online learning ................................................. 90
3.2 Adaptive learning using exponential smoothing ....................... 90
3.3 Lookup tables with frequentist updating ................................ 91
3.4 Lookup tables with Bayesian updating ................................... 92
3.4.1 The updating equations for independent beliefs .................... 92
3.4.2 Updating for correlated beliefs .......................................... 93
3.4.3 Gaussian process regression .............................................. 96
3.5 Computing bias and variance* .............................................. 97
3.6 Lookup tables and aggregation* ............................................ 99
3.6.1 Hierarchical aggregation .................................................. 100
3.6.2 Estimates of different levels of aggregation ......................... 102
3.6.3 Combining multiple levels of aggregation ......................... 106
3.7 Linear parametric models .................................................... 107
3.7.1 Linear regression review ................................................... 108
3.7.2 Sparse additive models and Lasso ....................................... 110
3.8 Recursive least squares for linear models .............................. 111
3.8.1 Recursive least squares for stationary data ......................... 112
3.8.2 Recursive least squares for nonstationary data* .................... 113
3.8.3 Recursive estimation using multiple observations* ............... 115
3.9 Nonlinear parametric models .............................................. 115
3.9.1 Maximum likelihood estimation ......................................... 115
3.9.2 Sampled belief models ..................................................... 116
3.9.3 Neural networks - parametric* ........................................... 117
3.9.4 Limitations of neural networks ......................................... 121
3.10 Nonparametric models* ..................................................... 122
3.10.1 K-nearest neighbor ......................................................... 123
3.10.2 Kernel regression ........................................................... 124
3.10.3 Local polynomial regression ............................................. 126
3.10.4 Deep neural networks ...................................................... 126
3.10.5 Support vector machines .................................................. 127
3.10.6 Indexed functions, tree structures and clustering ............... 128
3.10.7 Comments on nonparametric models ................................. 129
3.11 Nonstationary learning* ..................................................... 130
3.11.1 Nonstationary learning I - Martingale truth ....................... 130
3.11.2 Nonstationary learning II - Transient truth ....................... 131
3.11.3 Learning processes ....................................................... 132
3.12 The curse of dimensionality .............................................. 133
CONTENTS

3.13 Designing approximation architectures in adaptive learning 135
3.14 Why does it work?** 136
  3.14.1 Derivation of the recursive estimation equations 136
  3.14.2 The Sherman-Morrison updating formula 137
  3.14.3 Correlations in hierarchical estimation 138
3.15 Bibliographic notes 142
Exercises 144

4 Introduction to stochastic search 151
  4.1 Illustrations of the basic stochastic optimization problem 153
  4.2 Deterministic methods 155
    4.2.1 A “stochastic” shortest path problem 156
    4.2.2 A newsvendor problem with known distribution 156
    4.2.3 Chance-constrained optimization 157
    4.2.4 Optimal control 157
    4.2.5 Discrete Markov decision processes 158
    4.2.6 Remarks 159
  4.3 Sampled models 159
    4.3.1 Formulating a sampled model 160
    4.3.2 Convergence 162
    4.3.3 Creating a sampled model 164
    4.3.4 Decomposition strategies* 165
  4.4 Adaptive learning algorithms 166
    4.4.1 Modeling adaptive learning problems 167
    4.4.2 Online vs. offline applications 168
    4.4.3 Objective functions for learning 169
    4.4.4 Designing policies 172
  4.5 Closing remarks 172
  4.6 Bibliographic notes 173
Exercises 174

5 Derivative-Based Stochastic Search 183
  5.1 Some sample applications 185
  5.2 Modeling uncertainty 187
    5.2.1 Training uncertainty $W^1, \ldots, W^N$ 187
    5.2.2 Model uncertainty $S^0$ 188
    5.2.3 Testing uncertainty 189
    5.2.4 Policy evaluation 189
    5.2.5 Closing notes 190
  5.3 Stochastic gradient methods 190
5.3.1 A stochastic gradient algorithm 191
5.3.2 Introduction to stepsizes 191
5.3.3 Evaluating a stochastic gradient algorithm 193
5.3.4 A note on notation 194

5.4 Styles of gradients 194
5.4.1 Gradient smoothing 194
5.4.2 Second-order methods 195
5.4.3 Finite differences 195
5.4.4 SPSA 197
5.4.5 Constrained problems 198

5.5 Parameter optimization for neural networks* 199
5.5.1 Computing the gradient 199
5.5.2 The stochastic gradient algorithm 201

5.6 Stochastic gradient algorithm as a sequential decision problem 202

5.7 Empirical issues 204
5.8 Transient problems* 204
5.9 Theoretical performance* 205

5.10 Why does it work? 205
5.10.1 Some probabilistic preliminaries 206
5.10.2 An older proof* 207
5.10.3 A more modern proof** 210

5.11 Bibliographic notes 215
Exercises 216

6 Stepsize policies 223
6.1 Deterministic stepsize policies 225
6.1.1 Properties for convergence 226
6.1.2 A collection of deterministic policies 227

6.2 Adaptive stepsize policies 231
6.2.1 The case for adaptive stepsizes 231
6.2.2 Convergence conditions 232
6.2.3 A collection of stochastic policies 233
6.2.4 Experimental notes 236

6.3 Optimal stepsize policies* 236
6.3.1 Optimal stepsizes for stationary data 237
6.3.2 Optimal stepsizes for nonstationary data - I 239
6.3.3 Optimal stepsizes for nonstationary data - II 240

6.4 Optimal stepsizes for approximate value iteration* 243

6.5 Convergence 245

6.6 Guidelines for choosing stepsize formulas 246

6.7 Why does it work* 248
6.7.1 Proof of BAKF stepsize 248
7 Derivative-Free Stochastic Search

7.1 Overview of derivative-free stochastic search
7.1.1 Applications and time scales
7.1.2 The communities of derivative-free stochastic search
7.1.3 The multiarmed bandit story
7.1.4 From passive learning to active learning to bandit problems
7.2 Modeling derivative-free stochastic search
7.2.1 The universal model
7.2.2 Illustration: optimizing a manufacturing process
7.2.3 Major problem classes
7.3 Designing policies
7.4 Policy function approximations
7.5 Cost function approximations
7.6 VFA-based policies
7.6.1 An optimal policy
7.6.2 Beta-Bernoulli belief model
7.6.3 Backward approximate dynamic programming
7.6.4 Gittins indices for learning in steady state
7.7 Direct lookahead policies
7.7.1 When do we need lookahead policies?
7.7.2 Single period lookahead policies
7.7.3 Restricted multiperiod lookahead
7.7.4 Multiperiod deterministic lookahead
7.7.5 Multiperiod stochastic lookahead policies
7.7.6 Hybrid direct lookahead
7.8 The knowledge gradient (continued)*
7.8.1 The belief model
7.8.2 The knowledge gradient for maximizing final reward
7.8.3 The knowledge gradient for maximizing cumulative reward
7.8.4 The knowledge gradient for sampled belief model*
7.8.5 Knowledge gradient for correlated beliefs
7.9 Learning in batches
7.10 Simulation optimization*
7.10.1 An indifference zone algorithm
7.10.2 Optimal computing budget allocation
7.11 Evaluating policies
7.11.1 Alternative performance metrics*
7.11.2 Perspectives of optimality*
7.12 Designing policies
CONTENTS

7.12.1 Characteristics of a policy 323
7.12.2 The effect of scaling 324
7.12.3 Tuning 325
7.13 Extensions* 325
7.13.1 Learning in nonstationary settings 326
7.13.2 Strategies for designing time-dependent policies 326
7.13.3 A transient learning model 327
7.13.4 The knowledge gradient for transient problems 328
7.13.5 Learning with large or continuous choice sets 329
7.13.6 Learning with exogenous state information - the contextual bandit problem 331
7.13.7 State-dependent vs. state-independent problems 333
7.14 Bibliographic notes 334
Exercises 336

8 State-dependent problems 353
8.1 Graph problems 355
8.1.1 A stochastic shortest path problem 355
8.1.2 The nomadic trucker 356
8.1.3 The transformer replacement problem 357
8.1.4 Asset valuation 358
8.2 Inventory problems 360
8.2.1 A basic inventory problem 360
8.2.2 The inventory problem - II 361
8.2.3 The lagged asset acquisition problem 362
8.2.4 The batch replenishment problem 363
8.3 Complex resource allocation problems 365
8.3.1 The dynamic assignment problem 365
8.3.2 The blood management problem 368
8.4 State-dependent learning problems 373
8.4.1 Medical decision making 373
8.4.2 Laboratory experimentation 374
8.4.3 Bidding for ad-clicks 375
8.4.4 An information-collecting shortest path problem 375
8.5 A sequence of problem classes 375
8.6 Bibliographic notes 377
Exercises 377

9 Modeling sequential decision problems 383
9.1 A simple modeling illustration 387
9.2 Notational style 390
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.3</td>
<td>Modeling time</td>
<td>392</td>
</tr>
<tr>
<td>9.4</td>
<td>The states of our system</td>
<td>394</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Defining the state variable</td>
<td>394</td>
</tr>
<tr>
<td>9.4.2</td>
<td>The three states of our system</td>
<td>397</td>
</tr>
<tr>
<td>9.4.3</td>
<td>Initial state $S_0$ vs. subsequent states $S_t$, $t &gt; 0$</td>
<td>400</td>
</tr>
<tr>
<td>9.4.4</td>
<td>Lagged state variables*</td>
<td>401</td>
</tr>
<tr>
<td>9.4.5</td>
<td>The post-decision state variable*</td>
<td>402</td>
</tr>
<tr>
<td>9.4.6</td>
<td>A shortest path illustration</td>
<td>404</td>
</tr>
<tr>
<td>9.4.7</td>
<td>Belief states*</td>
<td>406</td>
</tr>
<tr>
<td>9.4.8</td>
<td>Latent variables*</td>
<td>407</td>
</tr>
<tr>
<td>9.4.9</td>
<td>Rolling forecasts*</td>
<td>408</td>
</tr>
<tr>
<td>9.4.10</td>
<td>Flat vs. factored state representations*</td>
<td>408</td>
</tr>
<tr>
<td>9.4.11</td>
<td>A programmer’s perspective of state variables</td>
<td>409</td>
</tr>
<tr>
<td>9.5</td>
<td>Modeling decisions</td>
<td>410</td>
</tr>
<tr>
<td>9.5.1</td>
<td>Types of decisions</td>
<td>411</td>
</tr>
<tr>
<td>9.5.2</td>
<td>Initial decision $x_0$ vs. subsequent decisions $x_t$, $t &gt; 0$</td>
<td>412</td>
</tr>
<tr>
<td>9.5.3</td>
<td>Strategic, tactical and execution decisions</td>
<td>412</td>
</tr>
<tr>
<td>9.5.4</td>
<td>Constraints</td>
<td>413</td>
</tr>
<tr>
<td>9.5.5</td>
<td>Introducing policies</td>
<td>414</td>
</tr>
<tr>
<td>9.6</td>
<td>The exogenous information process</td>
<td>414</td>
</tr>
<tr>
<td>9.6.1</td>
<td>Basic notation for information processes</td>
<td>414</td>
</tr>
<tr>
<td>9.6.2</td>
<td>Outcomes and scenarios</td>
<td>416</td>
</tr>
<tr>
<td>9.6.3</td>
<td>Lagged information processes*</td>
<td>417</td>
</tr>
<tr>
<td>9.6.4</td>
<td>Models of information processes*</td>
<td>418</td>
</tr>
<tr>
<td>9.6.5</td>
<td>Supervisory processes*</td>
<td>420</td>
</tr>
<tr>
<td>9.7</td>
<td>The transition function</td>
<td>421</td>
</tr>
<tr>
<td>9.7.1</td>
<td>A general model</td>
<td>421</td>
</tr>
<tr>
<td>9.7.2</td>
<td>Model-free dynamic programming</td>
<td>422</td>
</tr>
<tr>
<td>9.7.3</td>
<td>Exogenous transitions</td>
<td>423</td>
</tr>
<tr>
<td>9.8</td>
<td>The objective function</td>
<td>424</td>
</tr>
<tr>
<td>9.8.1</td>
<td>The performance metric</td>
<td>424</td>
</tr>
<tr>
<td>9.8.2</td>
<td>Optimizing the policy</td>
<td>425</td>
</tr>
<tr>
<td>9.8.3</td>
<td>Dependence of optimal policy on $S_0$</td>
<td>425</td>
</tr>
<tr>
<td>9.8.4</td>
<td>State-dependent variations</td>
<td>426</td>
</tr>
<tr>
<td>9.8.5</td>
<td>Uncertainty operators</td>
<td>427</td>
</tr>
<tr>
<td>9.9</td>
<td>Illustration: An energy storage model</td>
<td>428</td>
</tr>
<tr>
<td>9.9.1</td>
<td>With a time-series price model</td>
<td>429</td>
</tr>
<tr>
<td>9.9.2</td>
<td>With passive learning</td>
<td>430</td>
</tr>
<tr>
<td>9.9.3</td>
<td>With active learning</td>
<td>430</td>
</tr>
<tr>
<td>9.9.4</td>
<td>With rolling forecasts</td>
<td>431</td>
</tr>
<tr>
<td>9.10</td>
<td>Base models and lookahead models</td>
<td>431</td>
</tr>
<tr>
<td>9.11</td>
<td>A classification of problems*</td>
<td>432</td>
</tr>
</tbody>
</table>
CONTENTS

9.12 Policy evaluation* 435
9.13 Advanced probabilistic modeling concepts** 437
  9.13.1 A measure-theoretic view of information** 438
  9.13.2 Policies and measurability 440
9.14 Looking forward 442
9.15 Bibliographic notes 443
Exercises 445

10 Uncertainty modeling 459

10.1 Sources of uncertainty 460
  10.1.1 Observational errors 461
  10.1.2 Exogenous uncertainty 463
  10.1.3 Prognostic uncertainty 463
  10.1.4 Inferential (or diagnostic) uncertainty 465
  10.1.5 Experimental variability 467
  10.1.6 Model uncertainty 467
  10.1.7 Transitional uncertainty 469
  10.1.8 Control/implementation uncertainty 469
  10.1.9 Communication errors and biases 470
  10.1.10 Algorithmic instability 470
  10.1.11 Goal uncertainty 471
  10.1.12 Political/regulatory uncertainty 471
  10.1.13 Discussion 472

10.2 A modeling case study: the COVID pandemic 472

10.3 Stochastic modeling 472
  10.3.1 Sampling exogenous information 472
  10.3.2 Types of distributions 474
  10.3.3 Modeling sample paths 475
  10.3.4 State/action dependent processes 475
  10.3.5 Modeling correlations 476

10.4 Monte Carlo simulation 477
  10.4.1 Generating uniform $[0, 1]$ random variables 477
  10.4.2 Uniform and normal random variable 478
  10.4.3 Generating random variables from inverse cumulative distributions 480
  10.4.4 Inverse cumulative from quantile distributions 480
  10.4.5 Distributions with uncertain parameters 481

10.5 Case study: modeling electricity prices 483
  10.5.1 Mean reversion 483
  10.5.2 Jump-diffusion models 484
  10.5.3 Quantile distributions 485
  10.5.4 Regime shifting 486
## 10.5.5 Crossing times 486

10.6 Sampling vs. sampled models 488
10.6.1 Iterative sampling: A stochastic gradient algorithm 488
10.6.2 Static sampling: Solving a sampled model 488
10.6.3 Sampled representation with Bayesian updating 489

10.7 Closing notes 489
10.8 Bibliographic notes 490

### Exercises 490

## 11 Designing policies 495

11.1 From optimization to machine learning to sequential decision problems 497
11.2 The classes of policies 498
11.3 Policy function approximations 501
11.4 Cost function approximations 503
11.5 Value function approximations 504
11.6 Direct lookahead approximations 505
11.6.1 The basic idea 505
11.6.2 Modeling the lookahead problem 507
11.6.3 The policy-within-a-policy 508

11.7 Hybrid strategies 509
11.8 Randomized policies 513
11.9 Illustration: An energy storage model revisited 514
11.9.1 Policy function approximation 515
11.9.2 Cost function approximation 515
11.9.3 Value function approximation 515
11.9.4 Deterministic lookahead 515
11.9.5 Hybrid lookahead-cost function approximation 516
11.9.6 Experimental testing 516

11.10 Choosing the policy class 517
11.10.1 The policy classes 517
11.10.2 Policy complexity-computational tradeoffs 521
11.10.3 Screening questions 522

11.11 Policy evaluation 525
11.12 Parameter tuning 526
11.12.1 The soft issues 527
11.12.2 Searching across policy classes 528

11.13 Bibliographic notes 529

### Exercises 529

## 12 Policy function approximations and policy search 537

12.1 Policy search as a sequential decision problem 539
12.2 Classes of policy function approximations 540
  12.2.1 Lookup table policies 540
  12.2.2 Boltzmann policies for discrete actions 541
  12.2.3 Linear decision rules 541
  12.2.4 Monotone policies 542
  12.2.5 Nonlinear policies 543
  12.2.6 Nonparametric/locally linear policies 544
  12.2.7 Contextual policies 545
12.3 Problem characteristics 545
12.4 Flavors of policy search 546
12.5 Policy search with numerical derivatives 548
12.6 Derivative-free methods for policy search 550
  12.6.1 Belief models 550
  12.6.2 Learning through perturbed PFAs 551
  12.6.3 Learning CFAs 553
  12.6.4 DLA using the knowledge gradient 555
  12.6.5 Comments 555
12.7 Exact derivatives for continuous sequential problems* 555
12.8 Exact derivatives for discrete dynamic programs** 558
  12.8.1 A stochastic policy 558
  12.8.2 The objective function 559
  12.8.3 The policy gradient theorem 560
  12.8.4 Computing the policy gradient 560
12.9 Supervised learning 562
12.10 Why does it work? 563
  12.10.1 Derivation of the policy gradient theorem 563
12.11 Bibliographic notes 565
  Exercises 566

13 Cost function approximations 575
13.1 General formulation for parametric CFA 577
13.2 Objective-modified CFAs 578
  13.2.1 Linear cost function correction 578
  13.2.2 CFAs for dynamic assignment problems 578
  13.2.3 Dynamic shortest paths 580
  13.2.4 Dynamic trading policy 582
  13.2.5 Discussion 585
13.3 Constraint-modified CFAs 585
  13.3.1 General formulation of constraint-modified CFAs 586
  13.3.2 A blood management problem 587
  13.3.3 An energy storage example with rolling forecasts 588
13.4 Bibliographic notes 595
14 Exact dynamic programming 605
14.1 Discrete dynamic programming 606
14.2 The optimality equations 608
  14.2.1 Bellman’s equations 608
  14.2.2 Computing the transition matrix 612
  14.2.3 Random contributions 612
  14.2.4 Bellman’s equation using operator notation* 613
14.3 Finite horizon problems 613
14.4 Continuous problems with exact solutions 616
  14.4.1 The gambling problem 616
  14.4.2 The continuous budgeting problem 618
14.5 Infinite horizon problems* 619
14.6 Value iteration for infinite horizon problems* 621
  14.6.1 A Gauss-Seidel variation 622
  14.6.2 Relative value iteration 622
  14.6.3 Bounds and rates of convergence 624
14.7 Policy iteration for infinite horizon problems* 626
14.8 Hybrid value-policy iteration* 627
14.9 Average reward dynamic programming* 628
14.10 The linear programming method for dynamic programs** 629
14.11 Linear quadratic regulation 630
14.12 Why does it work?** 632
  14.12.1 The optimality equations 632
  14.12.2 Convergence of value iteration 636
  14.12.3 Monotonicity of value iteration 639
  14.12.4 Bounding the error from value iteration 640
  14.12.5 Randomized policies 641
14.13 Bibliographic notes 643
Exercises 643

15 Backward approximate dynamic programming 655
15.1 Backward approximate dynamic programming for finite horizon problems 656
  15.1.1 Some preliminaries 657
  15.1.2 Backward ADP using lookup tables 658
  15.1.3 Backward ADP algorithm with continuous approximations 660
15.2 Fitted value iteration for infinite horizon problems 662
15.3 Value function approximation strategies 664
  15.3.1 Linear models 664
15.3.2 Monotone functions 665
15.3.3 Other approximation models 667
15.4 Computational observations 667
15.4.1 Experimental benchmarking of backward ADP 667
15.4.2 Computational notes 671
15.5 Bibliographic notes 672
Exercises 672

16 Forward ADP I: The value of a policy 679

16.1 Sampling the value of a policy 680
16.1.1 Direct policy evaluation for finite horizon problems 680
16.1.2 Policy evaluation for infinite horizon problems 681
16.1.3 Temporal difference updates 683
16.1.4 TD(\(\lambda\)) 684
16.1.5 TD(0) and approximate value iteration 685
16.1.6 TD learning for infinite horizon problems 686
16.2 Stochastic approximation methods 689
16.3 Bellman’s equation using a linear model* 690
16.3.1 A matrix-based derivation** 691
16.3.2 A simulation-based implementation 693
16.3.3 Least squares temporal differences (LSTD) 693
16.3.4 Least squares policy evaluation (LSPE) 694
16.4 Analysis of TD(0), LSTD and LSPE using a single state* 695
16.4.1 Recursive least squares and TD(0) 695
16.4.2 LSPE 696
16.4.3 LSTD 696
16.4.4 Discussion 697
16.5 Gradient-based methods for approximate value iteration* 697
16.5.1 Approximate value iteration with linear models** 697
16.5.2 A geometric view of linear models* 701
16.6 Value function approximations based on Bayesian learning* 703
16.6.1 Minimizing bias for infinite horizon problems 703
16.6.2 Lookup tables with correlated beliefs 704
16.6.3 Parametric models 704
16.6.4 Creating the prior 705
16.7 Learning algorithms and stepsizes 705
16.7.1 Least squares temporal differences 706
16.7.2 Least squares policy evaluation 707
16.7.3 Recursive least squares 707
16.7.4 Bounding 1/n convergence for approximate value iteration 708
16.7.5 Discussion 709
16.8 Bibliographic notes 710
17 Forward ADP II: Policy optimization

17.1 Overview of algorithmic strategies
17.2 Approximate value iteration and Q-learning using lookup tables
   17.2.1 Value iteration using a pre-decision state variable
   17.2.2 Q-learning
   17.2.3 Value iteration using a post-decision state variable
   17.2.4 Value iteration using a backward pass
17.3 Styles of learning
   17.3.1 Offline learning
   17.3.2 From offline to online
   17.3.3 Evaluating offline and online learning policies
   17.3.4 Lookahead policies
17.4 Approximate value iteration using linear models
17.5 On-policy vs off-policy learning and the exploration-exploitation problem
   17.5.1 Terminology
   17.5.2 Learning with lookup tables
   17.5.3 Learning with generalized belief models
17.6 Applications
   17.6.1 Pricing an American option
   17.6.2 Playing “lose tic-tac-toe”
   17.6.3 Approximate dynamic programming for deterministic problems
17.7 Approximate policy iteration
   17.7.1 Finite horizon problems using lookup tables
   17.7.2 Finite horizon problems using linear models
   17.7.3 LSTD for infinite horizon problems using linear models
17.8 The actor-critic paradigm
17.9 Statistical bias in the max operator*
17.10 The linear programming method using linear models*
17.11 Finite horizon approximations for steady-state applications
17.12 Bibliographic notes

Exercises
18.2 Values versus marginal values 775
18.3 Piecewise linear approximations for scalar functions 776
18.4 Regression methods 779
18.5 Separable piecewise linear approximations 781
18.6 Benders decomposition for nonseparable approximations** 783
  18.6.1 Benders’ decomposition for two-stage problems 783
  18.6.2 Asymptotic analysis of Benders with regularization** 787
  18.6.3 Benders with regularization 790
18.7 Linear approximations for high-dimensional applications 791
18.8 Resource allocation with exogenous information state 792
18.9 Closing notes 793
18.10 Bibliographic notes 793
Exercises 795

19 Direct lookahead policies 805
19.1 Optimal policies using lookahead models 807
19.2 Creating an approximate lookahead model 811
  19.2.1 Modeling the lookahead model 812
  19.2.2 Strategies for approximating the lookahead model 813
19.3 Modified objectives in lookahead models 817
  19.3.1 Managing risk 817
  19.3.2 Utility functions for multiobjective problems 822
  19.3.3 Model discounting 822
19.4 Evaluating DLA policies 823
  19.4.1 Evaluating policies in a simulator 824
  19.4.2 Evaluating risk-adjusted policies 824
  19.4.3 Evaluating policies in the field 826
  19.4.4 Tuning direct lookahead policies 826
19.5 Why use a DLA? 827
19.6 Deterministic lookaheads 828
  19.6.1 A deterministic lookahead: shortest path problems 830
  19.6.2 Parameterized lookaheads 831
19.7 A tour of stochastic lookahead policies 833
  19.7.1 Lookahead PFAs 833
  19.7.2 Lookahead CFAs 834
  19.7.3 Lookahead VFAs for the lookahead model 835
  19.7.4 Lookahead DLAs for the lookahead model 835
  19.7.5 Discussion 836
19.8 Monte Carlo tree search for discrete decisions 836
  19.8.1 Basic idea 837
  19.8.2 The steps of MCTS 837
  19.8.3 Discussion 841
CONTENTS

19.8.4 Optimistic Monte Carlo tree search 842
19.9 Two-stage stochastic programming for vector decisions* 844
  19.9.1 The basic two-stage stochastic program 844
  19.9.2 Two-stage approximation of a sequential problem 845
  19.9.3 Discussion 848
19.10 Observations on DLA policies 848
19.11 Bibliographic notes 849
   Exercises 851

20 Multiagent modeling and learning 859
  20.1 Overview of multiagent systems 860
     20.1.1 Dimensions of a multiagent system 860
     20.1.2 Communication 862
     20.1.3 Modeling a multiagent system 863
     20.1.4 Controlling architectures 866
  20.2 A learning problem - flu mitigation 867
     20.2.1 A static model 867
     20.2.2 Variations of our flu model 868
     20.2.3 Two-agent learning models 871
     20.2.4 Transition functions for two-agent model 874
     20.2.5 Designing policies for the flu problem 875
  20.3 The POMDP perspective* 879
  20.4 The two-agent newsvendor problem 881
  20.5 Multiple independent agents - An HVAC controller model 885
     20.5.1 Model 886
     20.5.2 Designing policies 887
  20.6 Cooperative agents - A spatially distributed blood management problem 888
  20.7 Closing notes 891
  20.8 Why does it work? 891
     20.8.1 Derivation of the POMDP belief transition function 891
  20.9 Bibliographic notes 893
   Exercises 894