Book draft

Mathematical Foundation of Reinforcement Learning

Shiyu Zhao

August 2022

Contents

Preface 5

1 Basic Concepts 8

1.1 A grid-world example . . . 8
1.2 State and action . . . 9
1.3 State transition . . . 10
1.4 Policy . . . 11
1.5 Reward . . . 13
1.6 Trajectory, return, and episode . . . 15
1.7 Markov decision process . . . 17
1.8 Summary . . . 19

2 State Value and Bellman Equation 21

2.1 Motivating examples: Why return is important? . . . 21
2.2 Motivating example: How to calculate return? . . . 22
2.3 State value . . . 24
2.4 Bellman equation . . . 25
2.4.1 Illustrative examples . . . 26
2.4.2 Alternative expressions of the Bellman equation . . . 29
2.5 Solving state values from the Bellman equation . . . 30
2.5.1 Matrix-vector form of the Bellman equation . . . 30
2.5.2 Closed-form solution . . . 31
2.5.3 Iterative solution . . . 32
2.5.4 Illustrative examples . . . 32
2.6 From state value to action value . . . 34
2.6.1 Illustrative examples . . . 35
2.6.2 Bellman equation in terms of action values . . . 36
2.7 Summary . . . 37
2.8 Q&A . . . 37

3 Optimal State Value and Bellman Optimality Equation 39

3.1 Motivating example: how to improve policies? . . . 39

3.2 Optimal state value and optimal policy . . . 40
3.3 Bellman optimality equation . . . 41
3.3.1 Maximization on the right-hand side of the BOE . . . 42
3.3.2 Matrix-vector form of the BOE . . . 43
3.3.3 Contraction mapping theorem . . . 44
3.3.4 Contraction property of the right-hand side of the BOE . . . 47
3.4 Solutions of the BOE . . . 49
3.5 Factors that influence optimal policies . . . 52
3.6 Summary . . . 57
3.7 Q&A . . . 57

4 Value Iteration and Policy Iteration **59**
4.1 Value iteration . . . 59
4.1.1 Elementwise form and implementation . . . 60
4.1.2 Illustrative examples . . . 61
4.2 Policy iteration . . . 63
4.2.1 Algorithm analysis . . . 63
4.2.2 Elementwise form and implementation . . . 67
4.2.3 Illustrative examples . . . 67
4.3 Truncated policy iteration . . . 71
4.3.1 Compare value iteration and policy iteration . . . 71
4.3.2 Truncated policy iteration algorithm . . . 73
4.3.3 Illustrative examples . . . 75
4.4 Summary . . . 76
4.5 Q&A . . . 77

5 Monte Carlo Learning **79**
5.1 Motivating example: Monte Carlo estimation . . . 79
5.2 The simplest MC learning algorithm . . . 81
5.2.1 Converting policy iteration to be model-free . . . 81
5.2.2 MC Basic algorithm . . . 82
5.2.3 Illustrative examples . . . 84
5.3 MC Exploring Starts . . . 87
5.3.1 Using samples more efficiently . . . 87
5.3.2 Updating estimate more efficiently . . . 88
5.3.3 Algorithm description . . . 89
5.4 MC learning without exploring starts . . . 89
5.4.1 Soft policies . . . 90
5.4.2 Algorithm description . . . 90
5.4.3 Algorithm convergence . . . 92
5.4.4 Illustrative example . . . 93

5.5 Exploitation vs exploration . . . 93
5.6 Summary . . . 98
5.7 Q&A . . . 99

6 Stochastic Approximation 102
6.1 Motivating example: mean estimation . . . 102
6.2 Robbins-Monro Algorithm . . . 104
6.2.1 Convergence properties . . . 105
6.2.2 Application to mean estimation . . . 108
6.3 Dvoretzky's convergence theorem . . . 109
6.3.1 Proof of Dvoretzky's Theorem . . . 110
6.3.2 Application to mean estimation . . . 111
6.3.3 Application to the Robbins-Monro theorem . . . 112
6.3.4 An extension of Dvoretzky's Theorem . . . 113
6.4 Stochastic gradient descent . . . 114
6.4.1 Application to mean estimation . . . 116
6.4.2 Convergence pattern of SGD . . . 116
6.4.3 A deterministic formulation of SGD . . . 118
6.4.4 BGD, SGD, and Mini-batch GD . . . 119
6.4.5 Convergence of SGD . . . 121
6.5 Summary . . . 123
6.6 Q&A . . . 123

7 Temporal-Difference Learning 125
7.1 Motivating examples: stochastic algorithms . . . 125
7.2 TD learning of state values . . . 127
7.2.1 Algorithm . . . 127
7.2.2 Properties . . . 129
7.2.3 Convergence . . . 130
7.3 TD learning of action values: Sarsa . . . 133
7.3.1 Algorithm . . . 133
7.3.2 Implementation . . . 134
7.4 TD learning of action values: Expected Sarsa . . . 136
7.5 TD learning of action values: n-step Sarsa . . . 138
7.6 TD learning of optimal action values: Q-learning . . . 140
7.6.1 Algorithm . . . 140
7.6.2 Off-policy vs on-policy . . . 141
7.6.3 Q-learning is off-policy . . . 142
7.6.4 Implementation . . . 143
7.6.5 Illustrative examples . . . 144
7.7 A unified viewpoint . . . 147

7.8 Summary 147
7.9 Q&A 149

8 Value Function Approximation 152
8.1 Motivating examples: curve fitting 152
8.2 Objective function 154
8.3 Optimization algorithms 161
8.4 Linear function approximation 162
8.5 Illustrative examples 163
8.6 Theoretical analysis of TD learning 167
8.6.1 Convergence analysis 168
8.6.2 TD learning minimizes the projected Bellman error 173
8.6.3 Least-squares TD 176
8.7 Sarsa with function approximation 178
8.8 Q-learning with function approximation 179
8.9 Deep Q-learning 181
8.9.1 Algorithm description 181
8.9.2 Illustrative examples 183
8.10 Summary 186
8.11 Q&A 186

9 Policy Gradient Methods 188
9.1 Basic idea of policy gradient 188
9.2 Metrics to define optimal policies 189
9.3 Gradients of the metrics 193
9.3.1 Gradients in the discounted case 195
9.3.2 Gradients in the undiscounted case 200
9.4 Policy gradient by Monte Carlo estimation 205
9.5 Summary 207
9.6 Q&A 207

A Preliminaries to Probability Theory 209

B Preliminaries to Gradient Descent 214
B.1 Basics 214
B.2 Gradient-Descent Algorithms 217