

Book draft

# **Mathematical Foundations of Reinforcement Learning**

Shiyu Zhao

August, 2023

# Contents

<b>Contents</b>	<b>5</b>
<b>Preface</b>	<b>6</b>
<b>Overview of this Book</b>	<b>8</b>
<b>1 Basic Concepts</b>	<b>13</b>
1.1 A grid world example . . . . .	13
1.2 State and action . . . . .	14
1.3 State transition . . . . .	15
1.4 Policy . . . . .	16
1.5 Reward . . . . .	18
1.6 Trajectories, returns, and episodes . . . . .	20
1.7 Markov decision processes . . . . .	23
1.8 Summary . . . . .	24
1.9 Q&A . . . . .	24
<b>2 State Values and Bellman Equation</b>	<b>26</b>
2.1 Motivating example 1: Why are returns important? . . . . .	27
2.2 Motivating example 2: How to calculate returns? . . . . .	28
2.3 State values . . . . .	30
2.4 Bellman equation . . . . .	31
2.5 Examples for illustrating the Bellman equation . . . . .	33
2.6 Matrix-vector form of the Bellman equation . . . . .	36
2.7 Solving state values from the Bellman equation . . . . .	38
2.7.1 Closed-form solution . . . . .	38
2.7.2 Iterative solution . . . . .	39
2.7.3 Illustrative examples . . . . .	39
2.8 From state value to action value . . . . .	41
2.8.1 Illustrative examples . . . . .	42
2.8.2 The Bellman equation in terms of action values . . . . .	43
2.9 Summary . . . . .	44

2.10	Q&A . . . . .	44
<b>3</b>	<b>Optimal State Values and Bellman Optimality Equation</b>	<b>46</b>
3.1	Motivating example: How to improve policies? . . . . .	47
3.2	Optimal state values and optimal policies . . . . .	48
3.3	Bellman optimality equation . . . . .	49
3.3.1	Maximization of the right-hand side of the BOE . . . . .	50
3.3.2	Matrix-vector form of the BOE . . . . .	51
3.3.3	Contraction mapping theorem . . . . .	51
3.3.4	Contraction property of the right-hand side of the BOE . . . . .	55
3.4	Solving an optimal policy from the BOE . . . . .	57
3.5	Factors that influence optimal policies . . . . .	60
3.6	Summary . . . . .	64
3.7	Q&A . . . . .	65
<b>4</b>	<b>Value Iteration and Policy Iteration</b>	<b>67</b>
4.1	Value iteration . . . . .	68
4.1.1	Elementwise form and implementation . . . . .	68
4.1.2	Illustrative examples . . . . .	69
4.2	Policy iteration . . . . .	72
4.2.1	Algorithm analysis . . . . .	72
4.2.2	Elementwise form and implementation . . . . .	75
4.2.3	Illustrative examples . . . . .	77
4.3	Truncated policy iteration . . . . .	80
4.3.1	Comparing value iteration and policy iteration . . . . .	80
4.3.2	Truncated policy iteration algorithm . . . . .	82
4.4	Summary . . . . .	84
4.5	Q&A . . . . .	84
<b>5</b>	<b>Monte Carlo Methods</b>	<b>87</b>
5.1	Motivating example: Mean estimation . . . . .	88
5.2	MC Basic: The simplest MC-based algorithm . . . . .	90
5.2.1	Converting policy iteration to be model-free . . . . .	90
5.2.2	The MC Basic algorithm . . . . .	91
5.2.3	Illustrative examples . . . . .	93
5.3	MC Exploring Starts . . . . .	96
5.3.1	Utilizing samples more efficiently . . . . .	96
5.3.2	Updating policies more efficiently . . . . .	97
5.3.3	Algorithm description . . . . .	98
5.4	MC $\epsilon$ -Greedy: Learning without exploring starts . . . . .	99
5.4.1	$\epsilon$ -greedy policies . . . . .	99

5.4.2	Algorithm description . . . . .	100
5.4.3	Illustrative examples . . . . .	101
5.5	Exploration and exploitation of $\epsilon$ -greedy policies . . . . .	102
5.6	Summary . . . . .	107
5.7	Q&A . . . . .	107
<b>6</b>	<b>Stochastic Approximation</b>	<b>110</b>
6.1	Motivating example: Mean estimation . . . . .	111
6.2	Robbins-Monro algorithm . . . . .	112
6.2.1	Convergence properties . . . . .	114
6.2.2	Application to mean estimation . . . . .	117
6.3	Dvoretzky's convergence theorem . . . . .	118
6.3.1	Proof of Dvoretzky's theorem . . . . .	119
6.3.2	Application to mean estimation . . . . .	121
6.3.3	Application to the Robbins-Monro theorem . . . . .	121
6.3.4	An extension of Dvoretzky's theorem . . . . .	122
6.4	Stochastic gradient descent . . . . .	123
6.4.1	Application to mean estimation . . . . .	125
6.4.2	Convergence pattern of SGD . . . . .	125
6.4.3	A deterministic formulation of SGD . . . . .	127
6.4.4	BGD, SGD, and mini-batch GD . . . . .	128
6.4.5	Convergence of SGD . . . . .	130
6.5	Summary . . . . .	132
6.6	Q&A . . . . .	132
<b>7</b>	<b>Temporal-Difference Methods</b>	<b>134</b>
7.1	TD learning of state values . . . . .	135
7.1.1	Algorithm description . . . . .	135
7.1.2	Property analysis . . . . .	137
7.1.3	Convergence analysis . . . . .	139
7.2	TD learning of action values: Sarsa . . . . .	142
7.2.1	Algorithm description . . . . .	142
7.2.2	Optimal policy learning via Sarsa . . . . .	143
7.3	TD learning of action values: $n$ -step Sarsa . . . . .	147
7.4	TD learning of optimal action values: Q-learning . . . . .	149
7.4.1	Algorithm description . . . . .	149
7.4.2	Off-policy vs on-policy . . . . .	150
7.4.3	Implementation . . . . .	153
7.4.4	Illustrative examples . . . . .	153
7.5	A unified viewpoint . . . . .	154
7.6	Summary . . . . .	157

7.7	Q&A . . . . .	158
<b>8</b>	<b>Value Function Approximation</b>	<b>160</b>
8.1	Value representation: From table to function . . . . .	161
8.2	TD learning of state values based on function approximation . . . . .	164
8.2.1	Objective function . . . . .	165
8.2.2	Optimization algorithms . . . . .	170
8.2.3	Selection of function approximators . . . . .	171
8.2.4	Illustrative examples . . . . .	173
8.2.5	Theoretical analysis . . . . .	176
8.3	TD learning of action values based on function approximation . . . . .	188
8.3.1	Sarsa with function approximation . . . . .	188
8.3.2	Q-learning with function approximation . . . . .	189
8.4	Deep Q-learning . . . . .	190
8.4.1	Algorithm description . . . . .	191
8.4.2	Illustrative examples . . . . .	193
8.5	Summary . . . . .	195
8.6	Q&A . . . . .	196
<b>9</b>	<b>Policy Gradient Methods</b>	<b>199</b>
9.1	Policy representation: From table to function . . . . .	200
9.2	Metrics for defining optimal policies . . . . .	201
9.3	Gradients of the metrics . . . . .	206
9.3.1	Derivation of the gradients in the discounted case . . . . .	208
9.3.2	Derivation of the gradients in the undiscounted case . . . . .	213
9.4	Monte Carlo policy gradient (REINFORCE) . . . . .	218
9.5	Summary . . . . .	221
9.6	Q&A . . . . .	221
<b>10</b>	<b>Actor-Critic Methods</b>	<b>223</b>
10.1	The simplest actor-critic algorithm (QAC) . . . . .	224
10.2	Advantage actor-critic (A2C) . . . . .	225
10.2.1	Baseline invariance . . . . .	225
10.2.2	Algorithm description . . . . .	228
10.3	Off-policy actor-critic . . . . .	229
10.3.1	Importance sampling . . . . .	229
10.3.2	The off-policy policy gradient theorem . . . . .	232
10.3.3	Algorithm description . . . . .	234
10.4	Deterministic actor-critic . . . . .	235
10.4.1	The deterministic policy gradient theorem . . . . .	235
10.4.2	Algorithm description . . . . .	242

10.5 Summary . . . . .	243
10.6 Q&A . . . . .	244
<b>A Preliminaries for Probability Theory</b>	<b>245</b>
<b>B Measure-Theoretic Probability Theory</b>	<b>250</b>
<b>C Convergence of Sequences</b>	<b>257</b>
C.1 Convergence of deterministic sequences . . . . .	257
C.2 Convergence of stochastic sequences . . . . .	260
<b>D Preliminaries for Gradient Descent</b>	<b>264</b>
<b>Bibliography</b>	<b>275</b>
<b>Symbols</b>	<b>276</b>
<b>Index</b>	<b>278</b>