

# **Mathematical Foundations of Reinforcement Learning**

Shiyu Zhao

# Contents

<b>Contents</b>	<b>v</b>
<b>Preface</b>	<b>vii</b>
<b>Overview of this Book</b>	<b>ix</b>
<b>1 Basic Concepts</b>	<b>1</b>
1.1 A grid world example . . . . .	1
1.2 State and action . . . . .	2
1.3 State transition . . . . .	3
1.4 Policy . . . . .	4
1.5 Reward . . . . .	6
1.6 Trajectories, returns, and episodes . . . . .	8
1.7 Markov decision processes . . . . .	11
1.8 Summary . . . . .	12
1.9 Q&A . . . . .	12
<b>2 State Values and Bellman Equation</b>	<b>15</b>
2.1 Motivating example 1: Why are returns important? . . . . .	16
2.2 Motivating example 2: How to calculate returns? . . . . .	17
2.3 State values . . . . .	19
2.4 Bellman equation . . . . .	20
2.5 Examples for illustrating the Bellman equation . . . . .	22
2.6 Matrix-vector form of the Bellman equation . . . . .	25
2.7 Solving state values from the Bellman equation . . . . .	27
2.7.1 Closed-form solution . . . . .	27
2.7.2 Iterative solution . . . . .	28
2.7.3 Illustrative examples . . . . .	28
2.8 From state value to action value . . . . .	30
2.8.1 Illustrative examples . . . . .	31
2.8.2 The Bellman equation in terms of action values . . . . .	32
2.9 Summary . . . . .	33

---

2.10	Q&A . . . . .	33
<b>3</b>	<b>Optimal State Values and Bellman Optimality Equation</b>	<b>35</b>
3.1	Motivating example: How to improve policies? . . . . .	36
3.2	Optimal state values and optimal policies . . . . .	37
3.3	Bellman optimality equation . . . . .	38
3.3.1	Maximization of the right-hand side of the BOE . . . . .	39
3.3.2	Matrix-vector form of the BOE . . . . .	40
3.3.3	Contraction mapping theorem . . . . .	40
3.3.4	Contraction property of the right-hand side of the BOE . . . . .	44
3.4	Solving an optimal policy from the BOE . . . . .	46
3.5	Factors that influence optimal policies . . . . .	49
3.6	Summary . . . . .	53
3.7	Q&A . . . . .	54
<b>4</b>	<b>Value Iteration and Policy Iteration</b>	<b>57</b>
4.1	Value iteration . . . . .	58
4.1.1	Elementwise form and implementation . . . . .	58
4.1.2	Illustrative examples . . . . .	59
4.2	Policy iteration . . . . .	62
4.2.1	Algorithm analysis . . . . .	62
4.2.2	Elementwise form and implementation . . . . .	65
4.2.3	Illustrative examples . . . . .	67
4.3	Truncated policy iteration . . . . .	70
4.3.1	Comparing value iteration and policy iteration . . . . .	70
4.3.2	Truncated policy iteration algorithm . . . . .	72
4.4	Summary . . . . .	74
4.5	Q&A . . . . .	74
<b>5</b>	<b>Monte Carlo Methods</b>	<b>77</b>
5.1	Motivating example: Mean estimation . . . . .	78
5.2	MC Basic: The simplest MC-based algorithm . . . . .	80
5.2.1	Converting policy iteration to be model-free . . . . .	80
5.2.2	The MC Basic algorithm . . . . .	81
5.2.3	Illustrative examples . . . . .	83
5.3	MC Exploring Starts . . . . .	86
5.3.1	Utilizing samples more efficiently . . . . .	86
5.3.2	Updating policies more efficiently . . . . .	87
5.3.3	Algorithm description . . . . .	88
5.4	MC $\epsilon$ -Greedy: Learning without exploring starts . . . . .	89
5.4.1	$\epsilon$ -greedy policies . . . . .	89

---

5.4.2	Algorithm description . . . . .	90
5.4.3	Illustrative examples . . . . .	91
5.5	Exploration and exploitation of $\epsilon$ -greedy policies . . . . .	92
5.6	Summary . . . . .	97
5.7	Q&A . . . . .	97
<b>6</b>	<b>Stochastic Approximation</b>	<b>101</b>
6.1	Motivating example: Mean estimation . . . . .	102
6.2	Robbins-Monro algorithm . . . . .	103
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 . . . . .	112
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
<b>7</b>	<b>Temporal-Difference Methods</b>	<b>125</b>
7.1	TD learning of state values . . . . .	126
7.1.1	Algorithm description . . . . .	126
7.1.2	Property analysis . . . . .	128
7.1.3	Convergence analysis . . . . .	130
7.2	TD learning of action values: Sarsa . . . . .	133
7.2.1	Algorithm description . . . . .	133
7.2.2	Optimal policy learning via Sarsa . . . . .	134
7.3	TD learning of action values: $n$ -step Sarsa . . . . .	138
7.4	TD learning of optimal action values: Q-learning . . . . .	140
7.4.1	Algorithm description . . . . .	140
7.4.2	Off-policy vs on-policy . . . . .	141
7.4.3	Implementation . . . . .	144
7.4.4	Illustrative examples . . . . .	144
7.5	A unified viewpoint . . . . .	145
7.6	Summary . . . . .	148

---

7.7	Q&A . . . . .	149
<b>8</b>	<b>Value Function Methods</b>	<b>151</b>
8.1	Value representation: From table to function . . . . .	152
8.2	TD learning of state values based on function approximation . . . . .	155
8.2.1	Objective function . . . . .	156
8.2.2	Optimization algorithms . . . . .	161
8.2.3	Selection of function approximators . . . . .	162
8.2.4	Illustrative examples . . . . .	164
8.2.5	Theoretical analysis . . . . .	167
8.3	TD learning of action values based on function approximation . . . . .	179
8.3.1	Sarsa with function approximation . . . . .	179
8.3.2	Q-learning with function approximation . . . . .	180
8.4	Deep Q-learning . . . . .	181
8.4.1	Algorithm description . . . . .	182
8.4.2	Illustrative examples . . . . .	184
8.5	Summary . . . . .	186
8.6	Q&A . . . . .	187
<b>9</b>	<b>Policy Gradient Methods</b>	<b>191</b>
9.1	Policy representation: From table to function . . . . .	192
9.2	Metrics for defining optimal policies . . . . .	193
9.3	Gradients of the metrics . . . . .	198
9.3.1	Derivation of the gradients in the discounted case . . . . .	200
9.3.2	Derivation of the gradients in the undiscounted case . . . . .	205
9.4	Monte Carlo policy gradient (REINFORCE) . . . . .	210
9.5	Summary . . . . .	213
9.6	Q&A . . . . .	213
<b>10</b>	<b>Actor-Critic Methods</b>	<b>215</b>
10.1	The simplest actor-critic algorithm (QAC) . . . . .	216
10.2	Advantage actor-critic (A2C) . . . . .	217
10.2.1	Baseline invariance . . . . .	217
10.2.2	Algorithm description . . . . .	220
10.3	Off-policy actor-critic . . . . .	221
10.3.1	Importance sampling . . . . .	221
10.3.2	The off-policy policy gradient theorem . . . . .	224
10.3.3	Algorithm description . . . . .	226
10.4	Deterministic actor-critic . . . . .	227
10.4.1	The deterministic policy gradient theorem . . . . .	227
10.4.2	Algorithm description . . . . .	234

---

10.5 Summary . . . . .	235
10.6 Q&A . . . . .	236
<b>A Preliminaries for Probability Theory</b>	<b>237</b>
<b>B Measure-Theoretic Probability Theory</b>	<b>243</b>
<b>C Convergence of Sequences</b>	<b>251</b>
C.1 Convergence of deterministic sequences . . . . .	251
C.2 Convergence of stochastic sequences . . . . .	254
<b>D Preliminaries for Gradient Descent</b>	<b>259</b>
<b>Bibliography</b>	<b>270</b>
<b>Symbols</b>	<b>271</b>
<b>Index</b>	<b>273</b>