

# Reinforcement Learning

## An Introductory Note

Jingye Wang

✉ wangjy5@shanghaitech.edu.cn

Spring 2020

---

### Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Review of Basic Probability</b>	<b>5</b>
2.1	Interpretation of Probability . . . . .	5
2.2	Transformations . . . . .	5
2.3	Limit Theorem . . . . .	5
2.4	Sampling & Monte Carlo Methods . . . . .	6
2.5	Basic Inequalities . . . . .	8
2.6	Concentration Inequalities . . . . .	10
2.7	Conditional Expectation . . . . .	12
<b>3</b>	<b>Bandit Algorithms</b>	<b>14</b>
3.1	Bandit Models . . . . .	14
3.2	Stochastic Bandits . . . . .	14
3.3	Greedy Algorithms . . . . .	15
3.4	UCB Algorithms . . . . .	16
3.5	Thompson Sampling Algorithms . . . . .	17
3.6	Gradient Bandit Algorithms . . . . .	18
<b>4</b>	<b>Markov Chains</b>	<b>20</b>
4.1	Markov Model . . . . .	20
4.2	Basic Computations . . . . .	20
4.3	Classifications . . . . .	21

CONTENTS	2
4.4 Stationary Distribution . . . . .	22
4.5 Reversibility . . . . .	22
4.6 Markov Chain Monte Carlo . . . . .	23
<b>5 Markov Decision Process</b>	<b>25</b>
5.1 Markov Reward Process . . . . .	25
5.2 Markov Decision Process . . . . .	26
5.3 Dynamic Programming . . . . .	28
<b>6 Model-Free Prediction</b>	<b>33</b>
6.1 Monte-Carlo Policy Evaluation . . . . .	33
6.2 Temporal-Difference Learning . . . . .	35
<b>7 Model-Free Control</b>	<b>37</b>
7.1 On Policy Monte-Carlo Control . . . . .	37
7.2 On Policy Temporal-Difference Control: Sarsa . . . . .	39
7.3 Off-Policy Temporal-Difference Control: Q-Learning . . . . .	40
<b>8 Value Function Approximation</b>	<b>41</b>
8.1 Semi-gradient Method . . . . .	41
8.2 Deep Q-Learning . . . . .	43
<b>9 Policy Optimization</b>	<b>46</b>
9.1 Policy Optimization Theorem . . . . .	46
9.2 REINFORCE: Monte-Carlo Policy Gradient . . . . .	49
9.3 Actor-Critic Policy Gradient . . . . .	51
9.4 Extension of Policy Gradient . . . . .	52