Reinforcement Learning

Notes

Jingye Wang

wangjy5@shanghaitech.edu.cn

Spring 2020

Contents

1 Introduction **3**

2 Review of Basic Probability **5**

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

3 Bandit Algorithms **14**

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

4 Markov Chains **20**

4.1 Markov Model 20

4.2 Basic Computations 20

4.3 Classifications 21

4.4 Stationary Distribution . . . 22
4.5 Reversibility . . . 22
4.6 Markov Chain Monte Carlo . . . 23

5 Markov Decision Process 25
5.1 Markov Reward Process . . . 25
5.2 Markov Decision Process . . . 26
5.3 Dynamic Programming . . . 28

6 Model-Free Prediction 33
6.1 Monte-Carlo Policy Evaluation . . . 33
6.2 Temporal-Difference Learning . . . 35

7 Model-Free Control 38
7.1 On Policy Monte-Carlo Control . . . 38
7.2 On Policy Temporal-Difference Control: Sarsa . . . 40
7.3 Off-Policy Temporal-Difference Control: Q-Learning . . . 41

8 Value Function Approximation 42
8.1 Introduction on Function Approximation . . . 42
8.2 Incremental Method . . . 42
8.3 Batch Methods . . . 44
8.4 Deep Q-Learning . . . 45

9 Policy Optimization 46
9.1 Policy Optimization . . . 47
9.2 Monte-Carlo Policy Gradient . . . 48
9.3 Actor-Critic Policy Gradient . . . 51
9.4 Extension of Policy Gradient . . . 52