

Reinforcement Learning

Summary

Fabian Damken

July 24, 2022



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Contents

1	Introduction	9
1.1	Recent and Not-So-Recent Successes	9
1.2	“Artificial Intelligence”	9
1.3	Reinforcement Learning Formulation	9
1.3.1	Flavors	9
1.3.2	Components	9
1.4	Wrap-Up	9
2	Preliminaries	10
2.1	Functional Analysis	10
2.1.1	Normed Vector Spaces	10
2.1.2	Contractions	10
2.1.3	Fixed Point (Theorem)	10
2.2	Statistics	10
2.2.1	Stochastic Processes	10
2.2.2	Monte-Carlo Estimation	10
2.2.3	Bias-Variance Trade-Off	10
2.2.4	Important Sampling	10
2.2.5	Linear Function Approximation	10
2.2.6	Fisher Information Matrix	11
2.2.7	Entropy and Relative Entropy	11
2.2.8	Reparametrization Trick	11
2.3	Miscellaneous	11
2.3.1	Useful Integrals	11
2.3.2	Conjugate Gradient	11
3	Markov Decision Processes and Policies	12
3.1	Markov Decision Processes	12
3.1.1	Continuous State-Action-Space	12
3.1.2	Example	12
3.2	Markov Reward Processes	12
3.2.1	Return and Discount	12
3.2.2	Value Function	12
3.2.3	Example	12
3.3	Markov Decision Processes	12
3.3.1	Policies	12
3.3.2	Example	12
3.4	Wrap-Up	12

4	Dynamic Programming	13
4.1	Finite Horizon DP	13
4.2	Policy Iteration	13
4.2.1	Policy Evaluation	13
4.2.2	Policy Improvement	13
4.2.3	Using the Action-Value Function	13
4.2.4	Examples	13
4.3	Value Iteration	13
4.3.1	Principle of Optimality	13
4.3.2	Convergence	13
4.3.3	Example	13
4.4	Policy vs. Value Iteration	13
4.5	Efficiency	13
5	Monte-Carlo Algorithms	14
5.1	Policy Evaluation	14
5.2	Example	14
6	Temporal Difference Learning	15
6.1	Temporal Differences vs. Monte-Carlo	15
6.1.1	Bias-Variance Trade-Off	15
6.1.2	Markov Property	15
6.1.3	Backup	15
6.2	Bootstrapping and Sampling	15
6.3	TD(λ)	15
6.3.1	n -Step Return	15
6.3.2	λ -Return	15
6.3.3	Eligibility Traces	15
6.4	Example	15
6.5	Wrap-Up	15
7	Tabular Reinforcement Learning	16
7.0.1	Monte-Carlo Methods	16
7.0.2	TD-Learning: SARSA	16
7.1	Off-Policy Methods	16
7.1.1	Monte-Carlo	16
7.1.2	TD-Learning	16
7.2	Remarks	17
7.3	Wrap-Up	17
8	Function Approximation	18
8.1	On-Policy Methods	18
8.1.1	Stochastic Gradient Descent	18
8.1.2	Gradient Monte-Carlo	18
8.1.3	Semi-Gradient Methods	18
8.1.4	Least-Squares TD	18
8.2	Off-Policy Methods	18
8.2.1	Semi-Gradient TD	18

8.2.2	Divergence	18
8.3	The Deadly Triad	18
8.4	Offline Methods	18
8.4.1	Batch Reinforcement Learning	18
8.4.2	Least-Squares Policy Iteration	18
8.4.3	Fitted Q-Iteration	18
8.5	Wrap-Up	18
9	Policy Search	19
9.1	Policy Gradient	19
9.1.1	Computing the Gradient	19
9.1.2	REINFORCE	19
9.1.3	GPOMDP	19
9.2	Natural Policy Gradient	19
9.3	The Policy Gradient Theorem	19
9.3.1	Actor-Critic	19
9.3.2	Compatible Function Approximation	19
9.3.3	Advantage Function	19
9.3.4	Episodic Actor-Critic	19
9.4	Wrap-Up	19
10	Deep Reinforcement Learning	20
10.1	Deep Q-Learning: DQN	20
10.1.1	Replay Buffer	20
10.1.2	Target Network	20
10.1.3	Minibatch Updates	20
10.1.4	Reward- and Target-Clipping	20
10.1.5	Examples	20
10.2	DQN Enhancements	20
10.2.1	Overestimation and Double Deep Q-Learning	20
10.2.2	Prioritized Replay Buffer	20
10.2.3	Dueling DQN	20
10.2.4	Noisy DQN	20
10.2.5	Distributional DQN	20
10.2.6	Rainbow	20
10.3	Other DQN-Bases Methods	20
10.3.1	Count-Based Exploration	20
10.3.2	Curiosity-Driven Exploration	20
10.3.3	Ensemble-Driven Exploration	20
10.4	Wrap-Up	20
11	Deep Actor-Critic	21
11.1	Surrogate Loss	21
11.1.1	Kakade-Langford-Lemma	21
11.1.2	Practical Surrogate Loss	21
11.2	Advantage Actor-Critic (A2C)	21
11.3	On-Policy Methods	21
11.3.1	Trust-Region Policy Optimization (TRPO)	21



11.3.2 Proximal Policy Optimization (PPO)	21
11.4 Off-Policy Methods	21
11.4.1 Deep Deterministic Policy Gradient (DDPG)	21
11.4.2 Twin Delayed DDPG (TD3)	21
11.4.3 Soft Actor-Critic (SAC)	21
11.5 Wrap-Up	21
12 Frontiers	22
12.1 Partial Observability	22
12.2 Hierarchical Control	22
12.2.1 The Options Framework	22
12.3 Markov Decision Process Without Reward	22
12.3.1 Intrinsic Motivation	22
12.3.2 Inverse Reinforcement Learning	22
12.4 Model-Based Reinforcement Learning	22
12.5 Wrap-Up	22



List of Figures



List of Tables



List of Algorithms

1 Introduction

1.1 Recent and Not-So-Recent Successes

1.2 “Artificial Intelligence”

1.3 Reinforcement Learning Formulation

1.3.1 Flavors

Full

Partially Observable Markov Decision Problem

Markov Decision Problem

Further Simplifications

1.3.2 Components

1.4 Wrap-Up

2 Preliminaries

2.1 Functional Analysis

2.1.1 Normed Vector Spaces

2.1.2 Contractions

2.1.3 Fixed Point (Theorem)

2.2 Statistics

2.2.1 Stochastic Processes

2.2.2 Monte-Carlo Estimation

2.2.3 Bias-Variance Trade-Off

2.2.4 Important Sampling

2.2.5 Linear Function Approximation

Feature Construction

Polynomial Features

Fourier Basis

Coarse Coding

Tile Coding

Radial Basis Functions

Neural Networks

2.2.6 Fisher Information Matrix

2.2.7 Entropy and Relative Entropy

2.2.8 Reparametrization Trick

2.3 Miscellaneous

2.3.1 Useful Integrals

2.3.2 Conjugate Gradient

3 Markov Decision Processes and Policies

3.1 Markov Decision Processes

3.1.1 Continuous State-Action-Space

3.1.2 Example

3.2 Markov Reward Processes

3.2.1 Return and Discount

3.2.2 Value Function

Bellman Equation

3.2.3 Example

3.3 Markov Decision Processes

3.3.1 Policies

Value Functions

Bellman Expectation Equation

Bellman Operator

Optimality

Bellman Optimality Equation

Bellman Optimality Operator

3.3.2 Example

3.4 Wrap-Up

4 Dynamic Programming

4.1 Finite Horizon DP

4.2 Policy Iteration

4.2.1 Policy Evaluation

4.2.2 Policy Improvement

4.2.3 Using the Action-Value Function

4.2.4 Examples

4.3 Value Iteration

4.3.1 Principle of Optimality

4.3.2 Convergence

4.3.3 Example

4.4 Policy vs. Value Iteration

4.5 Efficiency



5 Monte-Carlo Algorithms

5.1 Policy Evaluation

5.2 Example

6 Temporal Difference Learning

6.1 Temporal Differences vs. Monte-Carlo

6.1.1 Bias-Variance Trade-Off

6.1.2 Markov Property

6.1.3 Backup

6.2 Bootstrapping and Sampling

6.3 $TD(\lambda)$

6.3.1 n -Step Return

6.3.2 λ -Return

6.3.3 Eligibility Traces

6.4 Example

6.5 Wrap-Up

7 Tabular Reinforcement Learning

7.0.1 Monte-Carlo Methods

Generalized Policy Iteration

Greediness and Exploration vs. Exploitation

ϵ -Greedy Exploration and Policy Improvement

Monte-Carlo Policy Iteration and Control

GLIE Monte-Carlo Control

7.0.2 TD-Learning: SARSA

Convergence

n -Step

Eligibility Traces and SARSA(λ)

Example

7.1 Off-Policy Methods

7.1.1 Monte-Carlo

7.1.2 TD-Learning

Importance Sampling

Q-Learning

Convergence

Example

7.2 Remarks

7.3 Wrap-Up

8 Function Approximation

8.1 On-Policy Methods

8.1.1 Stochastic Gradient Descent

8.1.2 Gradient Monte-Carlo

... with Linear Function Approximation

8.1.3 Semi-Gradient Methods

... with Linear Function Approximation

8.1.4 Least-Squares TD

Semi-Gradient SARSA

8.2 Off-Policy Methods

8.2.1 Semi-Gradient TD

8.2.2 Divergence

8.3 The Deadly Triad

8.4 Offline Methods

8.4.1 Batch Reinforcement Learning

8.4.2 Least-Squares Policy Iteration

8.4.3 Fitted Q-Iteration

8.5 Wrap-Up

9 Policy Search

9.1 Policy Gradient

9.1.1 Computing the Gradient

Finite Differences

Least-Squares-Based Finite Differences

Likelihood-Ratio Trick

9.1.2 REINFORCE

Gradient Variance and Baselines

Example

9.1.3 GPOMDP

9.2 Natural Policy Gradient

9.3 The Policy Gradient Theorem

9.3.1 Actor-Critic

9.3.2 Compatible Function Approximation

Example

9.3.3 Advantage Function

9.3.4 Episodic Actor-Critic

9.4 Wrap-Up

10 Deep Reinforcement Learning

10.1 Deep Q-Learning: DQN

10.1.1 Replay Buffer

10.1.2 Target Network

10.1.3 Minibatch Updates

10.1.4 Reward- and Target-Clipping

10.1.5 Examples

10.2 DQN Enhancements

10.2.1 Overestimation and Double Deep Q-Learning

10.2.2 Prioritized Replay Buffer

10.2.3 Dueling DQN

10.2.4 Noisy DQN

10.2.5 Distributional DQN

10.2.6 Rainbow

10.3 Other DQN-Bases Methods

10.3.1 Count-Based Exploration

10.3.2 Curiosity-Driven Exploration

10.3.3 Ensemble-Driven Exploration

10.4 Wrap-Up

11 Deep Actor-Critic

11.1 Surrogate Loss

11.1.1 Kakade-Langford-Lemma

11.1.2 Practical Surrogate Loss

11.2 Advantage Actor-Critic (A2C)

11.3 On-Policy Methods

11.3.1 Trust-Region Policy Optimization (TRPO)

Practical Implementation

11.3.2 Proximal Policy Optimization (PPO)

11.4 Off-Policy Methods

11.4.1 Deep Deterministic Policy Gradient (DDPG)

11.4.2 Twin Delayed DDPG (TD3)

11.4.3 Soft Actor-Critic (SAC)

11.5 Wrap-Up

12 Frontiers

12.1 Partial Observability

12.2 Hierarchical Control

12.2.1 The Options Framework

12.3 Markov Decision Process Without Reward

12.3.1 Intrinsic Motivation

12.3.2 Inverse Reinforcement Learning

12.4 Model-Based Reinforcement Learning

12.5 Wrap-Up
