UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Differential Privacy in Multi-agent Reinforcement Learning

by Zishuo Cheng

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Tianqing Zhu and Dr. Bo Liu

University of Technology Sydney
Faculty of Engineering and Information Technology

June 2022

Certificate of Original Authorship Template

Graduate research students are required to make a declaration of original authorship when they submit the thesis for examination and in the final bound copies. Please note, the Research Training Program (RTP) statement is for all students. The Certificate of Original Authorship must be placed within the thesis, immediately after the thesis title page.

Required wording for the certificate of original authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Zishuo Cheng, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 6 March 2023

Differential Privacy in Multi-agent Reinforcement Learning

by

Zishuo Cheng

A thesis submitted in fulfilment of the requirements for the
degree of Doctor of Philosophy

Abstract

Due to the increasing demand for autonomous systems, e.g., multi-robot systems and autonomous driving systems, multi-agent reinforcement learning, technology attracts considerable attention. In the field of multi-agent reinforcement learning, agent advising by allowing agents to ask for or give advice to others has become an increasingly important topic as it can significantly promote agents' learning speed with negligible computation overheads. However, there are some critical challenging issues in agent advising learning, particularly performance and privacy issues. Differential privacy is a promising privacy-preserving model with several valuable properties. Its privacy-preserving property provides a provable guarantee to privacy protection while its randomisation property helps to resist the inference in machine learning schemes. Therefore, this thesis aims to explore the feasibility of adopting differential privacy mechanisms to resolve the two research challenges in multi-agent reinforcement learning. In summary, this thesis consists of the following four contributions:

- A differential advising method is proposed, which allows agents to use a piece of advice in various states. This method is implemented by using differential privacy technology to mask the difference in states. In this way, agents' learning performance can be remarkably improved while their communication overheads can be significantly reduced.
- A differential knowledge transfer method is proposed, which stimulates the learning performance in a homogeneous multi-agent reinforcement learning system. This method jointly utilises the randomisation property of differential privacy and relevance weight to mitigate

the interference of negative transfer in a multi-agent reinforcement learning environment. By acquiring a greater number of relevant sample sets, agents' learning rates can be largely improved.

- A novel time-drive and privacy-preserving navigation learning scheme is proposed for multi-agent vehicular communication. This learning scheme, which is particularly suitable for multi-agent deep reinforcement learning systems, consists of three parts: 1) a function for estimating the average traffic flow of a segment; 2) a function for calculating a valid route; 3) confidence weights to estimate the accuracy of the estimated traffic flow. Second, we adopt a customised ϵ-differentially private mechanism for the *TDPP* model. To the best of our knowledge, this is the first navigation system for vehicle-to-vehicle systems to provably guarantee the location privacy of vehicles. Third, we theoretically prove that *TDPP* satisfies the definition of ϵ-differential privacy, accompanied by extensive experiments examining its performance.

- A novel multi-agent reinforcement learning model that jointly adopts deep reinforcement learning and differential privacy is proposed for evolutionary game theory, which promotes cultivating more cooperators while protecting agents' sensitive information. First, this is the first evolutionary cooperation method which takes the privacy of agents into account. By adopting differential privacy, *NNDP* protects agents' private information while still encouraging cooperation with negligible performance reduction. Additionally, compared to encryption methods, differentially private mechanisms use much less time and computational resources. Second, *NNDP* pioneers the use of deep learning to promote cooperation. By adopting a deep reinforcement learning algorithm, the *NNDP* offers a more adaptive, generalised framework and a framework with greater stability in dynamic situations. Third, we theoretically prove that *NNDP* satisfies the definition of differential privacy, accompanied by extensive experiments to examine its performance.

Acknowledgements

I would like to show my deepest gratitude to my supervisor, Prof. Tianqing Zhu, Prof. Wanlei Zhou and Dr. Dayong Ye. They give me a lot of help during the course of this course. Especially Prof. Tianqing Zhu, She gave me a lot of support in my study. I am also very grateful to my parents for their support over the past three years. Finally, I would like to express my heartfelt gratitude once again to those who have helped me.

Contents

Declaration of Authorship **i**

Abstract **ii**

Acknowledgements **iv**

List of Figures **x**

List of Tables **xii**

1 Introduction **1**
1.1 Multi-agent Reinforcement Learning 3
1.1.1 Advising in Multi-agent Reinforcement Learning 3
1.1.2 Knowledge Transfer in Multi-agent Reinforcement Learning 4
1.1.3 Vehicular Communication in Multi-agent Reinforcement Learning 5
1.1.4 Evolutionary Games in Multi-agent Reinforcement Learning 5
1.2 Differential Privacy in Multi-agent Reinforcement Learrning 6
1.3 Research Challenges and Contributions 7
1.4 Thesis Outline 9

2 Background and Related Work **12**
2.1 Preliminaries 12
2.1.1 Differential Privacy 12
2.1.1.1 Sensitivity 13
2.1.1.2 Privacy Budget 13
2.1.1.3 Principle Differential Privacy Mechanisms 13
2.1.2 Reinforcement Learning 14
2.1.3 Multi-agent Reinforcement Learning 15
2.2 Literature Review 15
2.2.1 Differentially Private Advising in Reinforcement Learning 15
2.2.1.1 Transfer Learning Frameworks 16
2.2.1.2 Policy/Sample Transfer 17
2.2.1.3 Advice Transfer 17

2.2.1.4 Other Studies 18
2.2.2 Differentially Private Knowledge Transfer in Reinforcement Learning . . 19
2.2.2.1 Transfer Learning Frameworks 19
2.2.2.2 Policy/Sample Transfer 19
2.2.2.3 Advice Transfer 20
2.2.2.4 Integration of Differential Privacy and AI 20
2.2.3 Differenntial Privacy in Multi-agent Vehicular Communication 21
2.2.3.1 Vehicular Navigation 21
2.2.3.2 Privacy Violation and Cyber Risk in Multi-agent Vehicular Communication 22
2.2.4 Differential Privacy in Multi-agent Game Theoretical Reinforcement Learning 23
2.2.4.1 Agent Performance in Multi-agent Game Theory 23
2.2.4.2 Privacy Violation in Multi-agent game Theory 23
2.3 Summary 24

3 Differentially Private Multi-agent Advising 25
3.1 Introduction 25
3.2 Problem Definition 27
3.2.1 Differentially private advising 27
3.3 The Differentially private Advising Method 29
3.3.1 Overview of the method 30
3.3.2 Deciding whether to ask for advice 33
3.3.3 Deciding whether to give advice 34
3.3.4 How to use advice 35
3.4 Theoretical Analysis 37
3.4.1 Differentially private advising analysis 37
3.4.2 Average reward analysis 38
3.4.3 Convergence analysis 41
3.5 Experiments 42
3.5.1 Experimental setup 42
3.5.1.1 Scenario 1 42
3.5.1.2 Scenario 2 43
3.5.1.3 The methods for comparison 44
3.5.2 Experimental results of scenario 1 45
3.5.2.1 The first setting: agents achieve targets 45
3.5.2.2 The second setting: agents achieve dynamically generated targets 46
3.5.2.3 The third setting: agents achieve dynamically moving targets . 48
3.5.3 Experimental results of scenario 2 49
3.5.4 Discussion and summary 52
3.5.4.1 The metric used in the second setting of the first scenario . . . 52
3.5.4.2 Performance of the three methods 53
3.6 Summary 53

4 Differentially Private Multi-agent Knowledge Transfer 55
4.1 Introduction . . . 55
4.1.1 Application Scenarios . . . 57
4.2 Problem Definition . . . 59
4.2.1 Knowledge Transfer with Differentially Private Noise . . . 59
4.3 Differentially Private Knowledge Transfer with Relevance Weight (*DTW*) in Reinforcement Learning . . . 59
4.3.1 Overview of the Multi-agent Reinforcement Learning Model with DTW . 59
4.3.2 Description of DTW . . . 61
4.3.3 Differentially Private Noise . . . 64
4.3.4 Relevance Weights . . . 66
4.3.5 Summary of *DTW* in Multi-agent Reinforcement Learning . . . 67
4.4 Theoretical Analysis . . . 67
4.4.1 Differentially Private Knowledge Transfer Analysis . . . 67
4.4.2 Convergence Analysis . . . 68
4.4.3 Knowledge Set Volume Analysis . . . 69
4.5 Experiments of Grid-World . . . 70
4.5.1 Grid-World Experiment Setup . . . 70
4.5.2 Experimental Results . . . 72
4.5.2.1 Performance in Different Domain Sizes . . . 72
4.5.2.2 Performance in Different Domain Complexity . . . 74
4.5.2.3 Performance of *DTW* with Different Privacy Budgets . . . 77
4.5.3 Discussion . . . 78
4.6 Additional Experimental Results with Non-Negative Rewards . . . 79
4.7 Experiments of Half-Field Offense . . . 79
4.7.1 Half-Field Offense Experiment Setup . . . 79
4.7.2 Experiment Results . . . 81
4.7.2.1 Performance in Different Observation Complexity . . . 81
4.7.2.2 Performance of *DTW* with Different Privacy Budgets . . . 82
4.7.3 Discussion . . . 83
4.8 Additional Experimental Results of Grid-World . . . 83
4.9 Summary . . . 84

5 Differentially Private and Time-driven Navigation for Multi-agent Vehicular Communication 85
5.1 Introduction . . . 85
5.2 Problem Definition . . . 87
5.2.1 Multi-agent Deep Reinforcement Learning . . . 87
5.3 Time-driven and Privacy-preserving Navigation Framework . . . 87
5.3.1 Time-driven Navigation Learning Scheme . . . 89
5.3.1.1 Average Estimated Segment Traffic Flow . . . 89
5.3.1.2 Valid Route Set . . . 90
5.3.1.3 Confidence Weights . . . 91
5.3.2 Privacy-preserving System . . . 91

5.3.2.1 Differentially Private Mechanism 91
5.3.2.2 Parameter Analysis 93
5.4 Theoretical Analysis 93
5.4.1 ϵ-Differential Privacy Analysis 93
5.4.2 Q-value analysis 94
5.5 Experiments 96
5.5.1 Experimental Setup 96
5.5.2 Results Analysis 97
5.5.2.1 Adaptivity 97
5.5.2.2 Stability 98
5.6 Summary 99

6 Privacy and Evolutionary Cooperation in Multi-agent Game Theory 100
6.1 Introduction 100
6.1.1 Application Scenarios 103
6.2 Problem Definition 104
6.2.1 Matrix Games and Social Dilemmas 104
6.2.2 Sequential Social Dilemmas 105
6.3 Neural Network with Differential Privacy 106
6.3.1 Overview 106
6.3.2 Neural Network Designs 108
6.3.2.1 A neural network for static environments 109
6.3.2.2 A neural network for dynamic environments 111
6.3.3 Differentially Private Mechanism Designs 113
6.3.3.1 Differential privacy mechanism in static environments 114
6.3.3.2 Differential privacy mechanism in dynamic environments . . . 115
6.3.3.3 Parameter analysis 116
6.4 Theoretical Analysis 117
6.5 Experiments 118
6.5.1 Experimental Setup 118
6.5.2 Games with Static Environments 120
6.5.2.1 Experiments for stability 120
6.5.2.2 Experiments for adaptivity 121
6.5.3 Games with Dynamic Environments 123
6.5.3.1 Experiments for stability 123
6.5.3.2 Experiments for adaptivity 123
6.5.4 Experiment for Observability 124
6.5.4.1 Experiments for stability 125
6.5.4.2 Experiments for Adaptivity 125
6.5.5 Discussion 127
6.6 Summary 128

7 Conclusion 129
7.1 Contributions 129

7.2 Future Works 131

Bibliography **134**

List of Figures

3.1 A multi-robot search and rescue example 26

3.2 The overview of agent i asking for advice from agent j 31

3.3 Performance of the three methods in different sizes of environments in the first setting 45

3.4 Performance of the three methods as time progresses in the first setting 46

3.5 Performance of the three methods in different sizes of environments in the second setting 47

3.6 Performance of the three methods as time progresses in the second setting 48

3.7 Performance of the three methods in different sizes of environments in the third setting 48

3.8 Performance of the three methods as time progresses in the third setting 49

3.9 Performance of the three methods 50

3.10 Performance of the three methods as learning progresses 52

4.1 An example multi-agent coordination 57

4.2 Agent 1 requesting knowledge from neighbours 61

4.3 *DTW* algorithm processing different types of knowledge 65

4.4 TotalHits of the four methods in *Cleaning Rubbish* (left), *Rescuing Victims* (central) and *Collecting Samples* (right) with five domain sizes at the 600th epoch . . 72

4.5 StepCost of the four methods in *Cleaning Rubbish* (left), *Rescuing Victims* (central) and *Collecting Samples* (right) with five domain sizes at the 600th epoch . . 72

4.6 AverageHits of the four methods on *Cleaning Rubbish* (left), *Rescuing Victims* (central) and *Collecting Samples* (right) with domain size 24×16 over 600 epochs 74

4.7 StepCost of the four methods on *Cleaning Rubbish* (left), *Rescuing Victims* (central) and *Collecting Samples* (right) with domain size 24×16 over 600 epochs . 74

4.8 Performance of *DTW* in three scenarios with domain size 24×16 over 600 epochs 76

4.9 Convergence of the *DTW* method with non-negative reward in grid-world and HFO domains 79

4.10 Goal rate of the four methods with 2 (left), 3 (central) and 4 (right)-feature observations over 3000 epochs 82

4.11 Goal rate of *DTW* with 4-feature observations over 3000 epochs 83

4.12 Performance of the four methods in *Cleaning Rubbish* with five domain sizes at the 600th epoch ($N_{goal} = 20, N_{agent} = 2, N_{obstacle} = 15$) 84

4.13 Performance of the four methods in *Rescuing Victims* with five domain sizes at the 600th epoch ($N_{goal} = 20, N_{agent} = 2, N_{obstacle} = 15$) 84

4.14 Performance of the four methods in *Collecting Samples* with five domain sizes at the 600th epoch ($N_{goal} = 20, N_{agent} = 2, N_{obstacle} = 15$) 84

5.1 Overview of the *TDPP* navigation system 89
5.2 Navigation example 91
5.3 Map of the CARLA environment 97
5.4 Distribution of T_{Ave} of four methods with 120 and 180 vehicles in 100 runs . . . 97
5.5 Distribution of T_{Ave} of *TDPP* with different privacy budgets and with 120 and 180 vehicles in 100 runs 98

6.1 Workflow of the *NNDP* multi-agent system. The model is a homogeneous multi-agent system where all the agents share a joint action set. An agent's state includes its observations of the environment plus those made by its single-hop neighbours. Observations about multi-hop neighbours are treated as sensitive information and therefore protected with differential privacy. 107
6.2 The neural network architecture for static environments. The architecture is a deep Q-learning neural network containing $n+1$ features in the input layer. The features include the observation made by the information requester and the observations made by the information requester's n single-hop neighbours. The observations made by each agent are the actions both they and their single-hop neighbours took. 109
6.3 An example of a network consisting of 6 agents. Agent A's single-hop neighbours are Agents B and E. Agent B's single-hop neighbours are Agents A, C and F. Agent C's single-hop neighbours are Agents A and F. After Once Agent A's single-hop neighbours B and C make and pass along the observations on their own single-hop neighbours, Agent A is informed of the actions taken by its multi-hop neighbours, Agents C and F. However, it still does not know the action taken by the multi-hop neighbour, Agent D. 110
6.4 The neural network architecture for dynamic environments. This architecture is a deep recurrent neural network based on Q-learning modified such that a recurrent LSTM (of the same size) replaces the last hidden layer of the DQN. Like the neural network architecture for static environments, the input layer contains $n + 1$ features, which include an observation made by the information requester, and the observations made by the information requester's n single-hop neighbours. However, the observation format is different from that in static environments. The observation in a dynamic network is the number of agents who took each action in action set A. 112
6.5 Cooperation Levels of Different Mechanisms in Static Environments (Initial Cooperation Level = 0.5, ϵ = 10) 120
6.6 Final Cooperation Levels of Different Mechanisms in Static Environments (ϵ = 10) 120
6.7 Cooperation Levels of Different Mechanisms in Dynamic Environments (Initial Cooperation Level = 0.5, ϵ = 10) 122
6.8 Final Cooperation Levels of Different Mechanisms in Dynamic Environments (ϵ = 10) 123

List of Tables

3.1 Notations of DA-RL 30

4.1 The numbers of agents, goals and obstacles in three experimental domains with five domain sizes 71

4.2 Total number of times that obstacles were hit in three scenarios with domain size 24×16 at the 600th epoch 74

5.1 Notations of TDPP 88

6.1 Social dilemmas 105

6.2 Notations of NNDP 106

6.3 Cooperation Levels of *NNDP* with Different Privacy Budgets in Static Environments (Initial Cooperation Level = 0.5) 119

6.4 Cooperation Levels of *NNDP* with Different Privacy Budgets in Dynamic Environments (Initial Cooperation Level = 0.5) 122

6.5 Final Cooperation Levels of *NNDP* with Different Observability Settings in Static Environments (Initial Cooperation Level = 0.5) 124

6.6 Final Cooperation Levels of *NNDP* with Different Observability Settings in Dynamic Environments (Initial Cooperation Level = 0.5) 125

6.7 Gap in Final Cooperation Levels with Different Observability Levels and Initial Levels of Cooperation in Static Environments ($\epsilon = 10$) 126

6.8 Gap in Final Cooperation Levels with Different Observability Levels and Initial Levels of Cooperation in Dynamic Environments ($\epsilon = 10$) 126

List of Publications

1. **Z. Cheng**, D. Ye, T. Zhu, W. Zhou, P. Yu, ‘Multi-Agent Reinforcement Learning via Knowledge Transfer with Differentially Private Noise’, International Journal of Intelligent Systems, published

2. **Z. Cheng**, D. Ye, T. Zhu, W. Zhou, T. Zhang, ‘Evolutionary Cooperation in Neural Network Games with Differentially Private Mechanisms’, IEEE Trans on System, Man and Cybernetics, in the second round review after major revision.

3. **Z. Cheng**, D. Ye, T. Zhu, C. Zhu, W. Zhou, ‘Time-Driven and Privacy-Preserving Navigation Model for V2V Communications’, IEEE Trans on Vehicular and Technology, under review.

4. D. Ye, T. Zhu, **Z. Cheng**, W. Zhou, P. Yu, ‘Differentially-Private Knowledge Transfer for Reinforcement Learning’, IEEE Trans on Cybernetics, published.

5. T. Zhu, D. Ye, **Z. Cheng**, W. Zhou, ”Learning Games for Defending Advanced Persistent Threats in Cyber Systems”, IEEE Transactions on Systems, Man and Cybernetics: Systems, in the second round review after major revision.

6. T. Zhu, W. Zhou, D. Ye, **Z. Cheng**, J. Li, ‘Resource Allocation in IoT Edge Computing via Concurrent Federated Reinforcement Learning’, IEEE Internet of Things, published.

7. S. Abahussein, **Z. Cheng**, T. Zhu, D. Ye, W. Zhou, ‘Privacy-preserving in Double Deep Q-Network with Differential Privacy in Continuous Spaces’, 2021 International Conference on Web-based Learning (ICWL), published.

8. S. Abahussein, D. Ye, **Z. Cheng**, T. Zhu, W. Zhou, ‘Differential privacy to protect privacy in Dueling Deep-Q-Network’, Concurrency and Computation Practice and Experience, under review.