（様式２）

学位論文の概要及び要旨

氏　　名　　鯨井　俊宏　　　　　　　印

題　　目　Greedy Action Selection and Pessimistic Q-value Updating in Multi-Agent Reinforcement Learning with Sparse Interaction
（スパースな干渉下での強化学習におけるグリーディな行動選択と悲観的なQ値の更新）

学位論文の概要及び要旨

Although multi-agent reinforcement learning (MARL) is a promising method for learning a collaborative action policy that will enable each agent to accomplish specific tasks, the state-action space increased exponentially resulting in difficulty of applying to real problems.

By assuming sparse interaction between agents, the state-action space can be dramatically reduced. This thesis proposes six multi-agent reinforcement learning methods that exploit the assumption of sparse interaction between agents based on existing CQ-learning method. This thesis demonstrates that the proposed methods improve the performance of learning comparing to existing methods using five maze games and seven patterns of pursuit games.

Chapter 1 briefly describes some characteristics of single-agent systems and multi-agent systems, and sparse interaction in multi-agent systems.

Chapter 2 describes a background of reinforcement learning for a single-agent environment. The chapter starts with defining a MDP (Markov Decision Process) that is a basic problem statement of long term optimization problems and introducing value functions. The chapter then explains three basic methods to efficiently solve MDPs; Dynamic Programming, Monte Carlo methods, and Temporal-Difference (TD) learning. The chapter also describes the detailed algorithm of Q-learning that is the most well-known TD learning method.

In Chapter 3, some characteristic of a multi-agent system are described. The chapter explains some extensions of MDP to deal with some class of propble in a multi-agent system. Maze games are introduced as an example of a multi-agent system. The chapter presents there is a problem of explosion of state-action space in a multi-agent system and demonstarated how the explosion affects the performance of reinforcement learning methos. The chapter then introduces a concept of sparse interaction, which dramatically reduces state-action space, and formulated such a proble as Dec-SIMDP (Decentralized Sparse Interaction MDP). The chapter finally introduces several reinforcement learning methods for Dec-SIMDP including Coordinating Q-learning (CQ-learning), which our proposed methods are based on. CQ-learning effectively reduces the state-action space by having each agent determine when it should consider the states of other agents on the basis of a comparison between the immediate rewards in a single-agent environment and those in a multi-agent environment.

Chapter 4 points out that CQ-learning has at least six issues to be improved; namely (1) how prelearning should be conducted, (2) unnecessary exploration by ϵ-greedily action selection, (3) optimistic Q-value updating, and (4) which Q-values should be used if more than two agents involve in an interference, (5) a problem in a multi-agent environment must be manually converted multiple problems in a single-agent environment, and (6) it only detects a difference of immediate rewards to identify interfered states.

The chapter presents four approaches to solve the issues of (1)-(4). The first approach for prelearning is to set ϵ value of ϵ-greedily action selection in a single-agent environment to 0.8 to ensure that an agent can explore all state-action combination enough. The second approach for avoiding unnecessary exploration is making an agent select its action greedily if it is in an unaugmented state exploiting knowledge learned in a single-agent environment. The third approach for avoiding optimistic Q-value updating is to change Q-value updating equation based on whether an agent is still in an interference state after taking previous action. The last approach for dealing with interference among more than two agents is randomly selecting one agent from agents that are in the interference state.

Evaluation using five maze games demonstrates that if both greedy action selection and changing Q-value updating equation based on wheter an agent is still in an interference state after taking previous action are applied, we call the learning method GPCQ-learning, the performance of CQ-learning is improved substantially.

Chapter 5 points out that in some pursuit games GPCQ-learning fell into a deadlock due to greedy action selection at an unaugmented state and failing to detect the deadlock because there are no difference of instant reward between in a single-agent environment and in a multi-agent environment.

The chapter proposes two approaches to break a deadlock caused by GPCQ-learning. The first approach is directly detecing the deadlock and augmenting the unaugmented state. The second approach is updating Q-values of unaugmented states as well as augmented states.

Evaluation using seven patterns of pursuit games and five maze games demonstrates that the two proposed approaches improved the performance of GPCQ-learning by breaking a deadlock.

Chapter 6 concludes the thesis by restating our contributions and describing some future work in this domain.

Table 5.2 Evaluation of average numer of steps to finish in pursuit games.

Patterns	1		2		3		4		5		6		7	
Minimum Steps	4		4		4		4		3		4		5	
Methods	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.
Independent	5.07	1.56	5.65	1.76	5.02	1.54	6.77	2.56	5.11	1.72	6.40	5.07	8.32	7.27
JSQ	5.23	2.35	5.38	2.82	6.87	8.14	6.69	4.12	5.22	2.47	6.59	4.91	11.5	11.8
JSAQ	6.91	9.65	7.11	10.5	108	110	7.73	14.2	7.47	10.7	8.54	31.2	141	119
CQ	5.15	1.73	6.45	8.88	6.26	13.7	5.87	1.57	18.8	28.2	178	291	30.4	54.0
GCQ	4.00	0.00	4.00	0.00	4.00	0.00	5.12	0.482	–	–	–	–	–	–
PCQ	5.17	1.96	6.69	10.1	6.00	18.5	5.94	1.63	26.4	30.5	56.6	57.8	43.5	109
GPCQ	4.00	0.00	4.00	0.00	4.00	0.00	5.14	0.518	–	–	–	–	–	–
GPCQBD	4.00	0.00	4.00	0.00	4.00	0.00	5.09	0.443	7.64	2.09	5.50	1.05	8.98	2.53
GPCQwU	4.00	0.00	4.00	0.00	4.00	0.00	5.00	0.00	4.00	0.00	5.00	0.00	6.00	0.00

Table 5.3 Evaluation of average number of steps to goal in maze games.

	Tunnel2Goal		ISR		CIT		CMU		TunnelToGoal3	
	mean	std	mean	std	mean	std	mean	std	mean	std
Min Steps	11		6		13		33		12	
CQ	13.5	3.84	31.6	79.9	25.7	26.9	41.7	8.58	23.8	19.2
GCQ	13.2	4.87	51.2	229	32.9	133	33.8	5.01	28.2	16.7
PCQ	12.8	1.59	8.64	5.87	19.1	15.2	39.4	4.56	23.1	17.7
GPCQ	11.3	0.616	7.42	1.68	15.6	7.68	33.3	2.94	30.9	22.6
GPCQBD	11.4	0.720	7.00	1.13	16.8	14.73	33.1	0.37	21.8	11.8
GPCQwU	11.4	1.61	6.03	0.17	24.0	0.00	33.0	0.00	14.5	0.883