DPMAC: Differentially Private Communication for Cooperative Multi-Agent Reinforcement Learning

Canzhe Zhao[1] , **Yanjie Ze**[2] , **Jing Dong**[3] , **Baoxiang Wang**[3] and **Shuai Li**[1,†]
[1]John Hopcroft Center for Computer Science, Shanghai Jiao Tong University
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[3]School of Data Science, The Chinese University of Hong Kong, Shenzhen
{canzhezhao, zeyanjie, shuaili8}@sjtu.edu.cn, jingdong@link.cuhk.edu.cn, bxiangwang@cuhk.edu.cn

Abstract

Communication lays the foundation for cooperation in human society and in multi-agent reinforcement learning (MARL). Humans also desire to maintain their privacy when communicating with others, yet such privacy concern has not been considered in existing works in MARL. To this end, we propose the *differentially private multi-agent communication* (DPMAC) algorithm, which protects the sensitive information of individual agents by equipping each agent with a local message sender with rigorous (ϵ, δ)-differential privacy (DP) guarantee. In contrast to directly perturbing the messages with predefined DP noise as commonly done in privacy-preserving scenarios, we adopt a stochastic message sender for each agent respectively and incorporate the DP requirement into the sender, which automatically adjusts the learned message distribution to alleviate the instability caused by DP noise. Further, we prove the existence of a Nash equilibrium in cooperative MARL with privacy-preserving communication, which suggests that this problem is game-theoretically learnable. Extensive experiments demonstrate a clear advantage of DPMAC over baseline methods in privacy-preserving scenarios.

1 Introduction

Multi-agent reinforcement learning (MARL) has shown remarkable achievements in many real-world applications such as sensor networks [Zhang and Lesser, 2011], autonomous driving [Shalev-Shwartz *et al.*, 2016], and traffic control [Wei *et al.*, 2019]. To mitigate non-stationarity when training the multi-agent system, centralized training and decentralized execution (CTDE) paradigm is proposed. The CTDE paradigm yet faces the hardness to enable complex cooperation and coordination for agents during execution due to the inherent partial observability in multi-agent scenarios [Wang *et al.*, 2020b]. To make agents cooperate more efficiently in complex partial observable environments, communication between agents has been considered. Numerous works proposed differentiable communication methods between agents, which can be trained in an end-to-end manner, for more efficient cooperation among agents [Foerster *et al.*, 2016; Jiang and Lu, 2018; Das *et al.*, 2019; Ding *et al.*, 2020; Kim *et al.*, 2021; Wang *et al.*, 2020b].

However, the advantages of communication, resulting from full information sharing, come with the possible privacy leakage of individual agents for both broadcasted and one-to-one messages. Therefore, in practice, one agent may be unwilling to fully share its private information with other agents even though in *cooperative* scenarios. For instance, if we train and deploy an MARL-based autonomous driving system, each autonomous vehicle involved in this system could be regarded as an agent and all vehicles work together to improve the safety and efficiency of the system. Hence, this can be regarded as a cooperative MARL scenario [Shalev-Shwartz *et al.*, 2016; Yang *et al.*, 2020]. However, owners of autonomous vehicles may not allow their vehicles to send private information to other vehicles without any desensitization since this may divulge their private information such as their personal life routines [Hassan *et al.*, 2020]. Hence, a natural question arises:

Can the MARL algorithm with communication under the CTDE framework be endowed with both the rigorous privacy guarantee and the empirical efficiency?

To answer this question, we start with a simple motivating example called *single round binary sums*, where several players attempt to guess the bits possessed by others and they can share their own information by communication. In Section 4, we show that a local message sender using the randomized response mechanism allows an privacy-aware receiver to correctly calculate the binary sum in a privacy-preserving way. From the example, we gain two insights: 1) The information is not supposed to be aggregated likewise in previous communication methods in MARL [Das *et al.*, 2019; Ding *et al.*, 2020], as a trusted data curator is not available in general. On the contrary, privacy is supposed to be achieved locally for every agent; 2) Once the agents know a priori, that certain privacy constraint exists, they could adjust their inference on the noised messages. These two insights indicate the principles of our privacy-preserving communication scheme that we desire *a privacy-preserving local sender* and *a privacy-aware receiver*.

[†] Corresponding author.

Our algorithm, *differentially private multi-agent communication* (DPMAC), instantiates the described principles. More specifically, for the sender part, each agent is equipped with a *local* sender which ensures differential privacy (DP) [Dwork, 2006] by performing an additive Gaussian noise. The message sender in DPMAC is local in the sense that each agent is equipped with its own message sender, which is only used to send its own messages. Equipped with this local sender, DPMAC is able to not only protect the privacy of communications between agents but also satisfy different privacy levels required from different agents. In addition, the sender adopts the Gaussian distribution to represent the message space and sample stochastic messages from the learned distribution. However, it is known that the DP noise may impede the original learning process [Dwork *et al.*, 2014; Alvim *et al.*, 2011], resulting in unstable or even divergent algorithms, especially for deep learning-based methods [Abadi *et al.*, 2016; Chen *et al.*, 2020]. To cope with this issue, we incorporate the noise variance into the representation of the message distribution, so that the agents could learn to adjust the message distribution automatically according to varying noise scales. For the receiver part, due to the gradient chain between the sender and the receiver, our receiver naturally utilizes the privacy-relevant information hidden in the gradients. This implements the privacy-aware receiver described in the motivating example.

When protecting privacy in communication is required in a cooperative game, the game is *not* purely cooperative anymore since each player involved will face a trade-off between the team utility and its personal privacy. To analyze the convergence of cooperative games with privacy-preserving communication, we first define a single-step game, namely the *collaborative game with privacy* (CGP). We prove that under some mild assumptions of the players' value functions, CGP could be transformed into a potential game [Monderer and Shapley, 1996], subsequently leading to the existence of a Nash equilibrium (NE). With this property, NE could also be proven to exist in the single round binary sums. Furthermore, we extend the single round binary sums into a multi-step game called *multiple round sums* using the notion of Markov potential game (MPG) [Leonardos *et al.*, 2021]. Inspired by Macua *et al.* (2018) and modeling the privacy-preserving communication as part of the agent action, we prove the existence of NE, which indicates that the multi-step game with privacy-preserving communication could be learnable.

To validate the effectiveness of DPMAC, extensive experiments are conducted in multi-agent particle environment (MPE) [Lowe *et al.*, 2017], including cooperative navigation, cooperative communication and navigation, and predator-prey tasks. Specifically, in privacy-preserving scenarios, DPMAC significantly outperforms baselines. Moreover, even without any privacy constraints, DPMAC could also gain competitive performance against baselines.

To sum up, the contributions of this work are threefold:

- To the best of our knowledge, we make the first attempt to develop a framework for private communication in MARL, named DPMAC, with provable (ϵ, δ)-DP guarantee.
- We prove the existence of the Nash equilibrium for cooperative games with privacy-preserving communication, showing that these games are game-theoretically learnable.
- Extensive experiments show that DPMAC clearly outperforms baselines in privacy-preserving scenarios and gains competitive performance in non-private scenarios.

2 Related Work

2.1 Learning to Communicate in MARL

Learning communication protocols in MARL by backpropagation and end-to-end training has achieved great advances in recent years [Sukhbaatar *et al.*, 2016; Foerster *et al.*, 2016; Jiang and Lu, 2018; Das *et al.*, 2019; Wang *et al.*, 2020b; Ding *et al.*, 2020; Kim *et al.*, 2021; Rangwala and Williams, 2020; Zhang *et al.*, 2019; Singh *et al.*, 2019; Zhang *et al.*, 2020; Zhang *et al.*, 2021; Lin *et al.*, 2021; Peng *et al.*, 2017]. Amongst these works, Sukhbaatar *et al.* (2016) propose CommNet as the first differentiable communication framework for MARL. Further, TarMAC [Das *et al.*, 2019] and ATOC [Jiang and Lu, 2018] utilize the attention mechanism to extract useful information as messages. I2C [Ding *et al.*, 2020] makes the first attempt to enable agents to learn one-to-one communication via causal inference. Wang *et al.* (2020b) propose NDQ, which learns nearly decomposable value functions to reduce the communication overhead. Kim *et al.* (2021) consider sharing an imagined trajectory as an intention for effectiveness. Besides, to communicate in the scenarios with limited bandwidth, some works consider learning to send compact and informative messages in MARL via minimizing the entropy of messages between agents using information bottleneck methods [Wang *et al.*, 2020a; Tucker *et al.*, 2022; Tian *et al.*, 2021; Li *et al.*, 2021]. While learning effective communication in MARL has been extensively investigated, existing communication algorithms potentially leave the privacy of each agent vulnerable to information attacks.

2.2 Privacy Preserving in RL

With wide attention on reinforcement learning (RL) algorithms and applications in recent years, so have concerns about their privacy. Sakuma *et al.* (2008) consider privacy in the distributed RL problem and utilize cryptographic tools to protect the private state-action-state triples. Algorithmically, Balle *et al.* (2016) make the first attempt to establish a policy evaluation algorithm with DP guarantee, where the Monte-Carlo estimates are perturbed with Gaussian noises. Wang and Hegde (2019) generalize the results to Q-learning, where functional noises are added to protect the reward functions. Theoretically, Garcelon *et al.* (2021) study regret minimization of finite-horizon Markov decision processes (MDPs) with DP guarantee in the tabular case. In a large or continuous state space where function approximation is required, Liao *et al.* (2021) and Zhou (2022) subsequently take the first step to establish the sublinear regret in linear mixture MDPs. Meanwhile, a large number of works focus on preserving privacy in multi-armed bandits [Tao *et al.*, 2022; Tenenbaum *et al.*, 2021; Dubey, 2021; Zheng *et al.*, 2020; Dubey and Pentland, 2020; Tossou and Dimitrakakis, 2017].

Privacy is also studied in recent literature on MARL and multi-agent system. Ye *et al.* (2020) study differential advising for value-based agents, which share action values as the advice, largely differing in both the communication framework and the CTDE framework. Dong *et al.* (2020) propose an average consensus algorithm with a DP guarantee in the multi-agent system.

3 Preliminaries

We consider a fully cooperative MARL problem where N agents work collaboratively to maximize the joint rewards. The underlying environment can be captured by a decentralized partially observable Markov decision process (Dec-POMDP), denoted by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. Specifically, $\mathcal{S}$ is the global state space, $\mathcal{A} = \prod_{i=1}^{N} \mathcal{A}_i$ is the joint action space, $\mathcal{O} = \prod_{i=1}^{N} \mathcal{O}_i$ is the joint observation space, $\mathcal{P}\left(s' \mid s, \boldsymbol{a}\right) := \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0,1]$ determines the state transition dynamics, $\mathcal{R}(s, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, and $\gamma \in [0,1)$ is the discount factor. Given a joint policy $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^{N}$, the joint action-value function at time t is $Q^{\boldsymbol{\pi}}\left(s^t, \boldsymbol{a}^t\right) = \mathbb{E}\left[G^t \mid s^t, \boldsymbol{a}^t, \boldsymbol{\pi}\right]$, where $G^t = \sum_{i=0}^{\infty} \gamma^i \mathcal{R}^{t+i}$ is the cumulative reward, and $\boldsymbol{a}^t = \{a_i^t\}_{i=1}^{N}$ is the joint action. The ultimate goal of the agents is to find an optimal policy $\boldsymbol{\pi}^*$ which maximizes $Q^{\boldsymbol{\pi}}\left(s^t, \boldsymbol{a}^t\right)$.

Under the aforementioned cooperative setting, we study the case where agents are allowed to communicate with a joint message space $\mathcal{M} = \prod_{i=1}^{N} \mathcal{M}_i$. When the communication is unrestricted, the problem is reduced to a single-agent RL problem, which effectively solves the challenge posed by partially observable states, but puts the individual agent's privacy at risk. To overcome the challenges of privacy and partial observable states simultaneously, we investigate algorithms that maximize the cumulative rewards while satisfying DP, given in the following definition.

Definition 3.1 ((ϵ, δ)-DP, [Dwork, 2006]). *A randomized mechanism $f : \mathcal{D} \rightarrow \mathcal{Y}$ satisfies (ϵ, δ)-differential privacy if for any neighbouring datasets $D, D' \in \mathcal{D}$ and $S \subset \mathcal{Y}$, it holds that* $\Pr[f(D) \in S] \leq e^{\epsilon} \Pr\left[f\left(D'\right) \in S\right] + \delta$.

DP offers a mathematically rigorous way to quantify the privacy of an algorithm [Dwork, 2006]. An algorithm is said to be "privatized" under the notion of DP if it is statistically hard to infer the presence of an individual data point in the dataset by observing the output of the algorithm. More intuitively, an algorithm satisfies DP if it provides nearly the same outputs given the neighbouring input datasets (*i.e.*, $\Pr[f(D) \in S] \approx \Pr\left[f\left(D'\right) \in S\right]$), which hence protects the sensitive information from the curious attacker.

With DP, each agent i is assigned with a privacy budget ϵ_i, which is negatively correlated to the level of privacy protection. Then we have $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^{N}$ as the set of all privacy budgets. In addition to maximizing the joint rewards as usually required in cooperative MARL, the messages sent from agent i are also required to satisfy the privacy budget ϵ_i with probability at least $1 - \delta$.

4 Motivating Example

Before introducing our communication framework, we first investigate a motivating example, which is a *cooperative* game and inspires the design principles of private communication mechanisms in MARL. The motivating example is a simple yet interesting game, called *single round binary sums*. The game is extended from the example provided in [Cheu, 2021] for analyzing the shuffle model, while we illustrate the game from the perspective of multi-agent systems. We note that though this game is one-step, which is different from the sequential decision process like MDP, it is illustrative enough to show how the communication protocol works as a tool to achieve a better trade-off between privacy and utility.

Assume that there are N agents involved in this game. Each agent $i \in [N]$ has a bit $b_i \in \{0,1\}$ and can tell other agents the information about its bit by communication. The objective of the game is for every agent to guess $\sum_i b_i$, the sum of the bits of all agents. Namely, each agent i makes a guess g_i and the utility of the agent is to maximize $r_i = -|\sum_j b_j - \mathbb{E}[g_i]|$. The (global) reward of this game is the sum of the utility over all agents, *i.e.*, $\sum_i r_i$.

Without loss of generality, we write the guess g_i into $g_i = \sum_{j \neq i} y_{ij} + b_i$, where y_{ij} is the guessed bit of agent j by agent i. If all agents share their bits without covering up, the guessed bit y_{ij} will obviously be equal to b_j and all agents attain an optimal return. Hence this game is fully cooperative under no privacy constraints. However, the optimal strategy is under the assumption that *everyone is altruistic to share their own bits*.

To preserve the privacy in communication, the message (*i.e.*, the sent bit) could be randomized using *randomized response*, which perturbs the bit b_i with probability p, as shown below:

$$x_i = \mathcal{R}_{\mathrm{RR}}\left(b_i\right) := \begin{cases} \mathrm{Ber}(1/2) & \text{with probability } p \\ b_i & \text{otherwise} \end{cases},$$

where x_i is the random message and $\mathrm{Ber}(\cdot)$ indicates the Bernoulli distribution. Under our context, $\mathcal{R}_{\mathrm{RR}}$ is a *privacy-preserving message sender*, whose privacy guarantee is guaranteed by the following proposition.

Proposition 4.1 ([Beimel *et al.*, 2008]). *Setting $p = \frac{2}{e^{\epsilon}+1}$ in $\mathcal{R}_{RR}$ suffices for $(\epsilon, 0)$-differential privacy.*

When each agent is equipped with such a privacy-preserving sender $\mathcal{R}_{RR}$ while adhering to the originally optimal strategy (*i.e.*, believing what others tell and doing the guess), all agents would make an inaccurate guess. The bias of the guess denoted as err_i caused by $\mathcal{R}_{RR}$ is then

$$\begin{aligned} \mathrm{err}_i &= \mathbb{E}[g_i] - \sum_i b_i = \sum_{j \neq i} \mathbb{E}[x_j - b_j] = p \sum_{j \neq i} (\frac{1}{2} - b_j) \\ &= \frac{p(N-1)}{2} - p \sum_{j \neq i} b_j \, . \end{aligned}$$

Without any prior knowledge, the bias could not be reduced for $(\epsilon, 0)$-DP algorithms. However, if the probability p of perturbation is set as prior common knowledge for all agents before the game starts, things will be different. One could transform the biased guess into

$$g_i^{\mathcal{A}} = \mathcal{A}_{\mathrm{RR}}(\vec{x}_{-i}) := \frac{1}{1-p}\left(\sum_{j\neq i} x_j - (N-1)p/2\right) ,$$

where $\vec{x}_{-i} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N]^\top$ denote the messages received by agent i. Then the estimate will be unbiased as

$$\mathbb{E}\left[g_i^{\mathcal{A}}\right] = \frac{1}{1-p}\left(\mathbb{E}\left[\sum_{j\neq i} x_j\right] - \frac{p(N-1)}{2}\right) + b_i = \sum_i b_i .$$

This example inspires that a communication algorithm could be both privacy-preserving and efficient. From the perspective of privacy, by the post-processing lemma of DP, any post-processing does not affect the original privacy level. From the perspective of utility, we could eliminate the bias err_i if the agent is equipped with the receiver $\mathcal{A}_{RR}$ and the prior knowledge p is given.

In general, our motivating example gives two principles for designing privacy-preserving communication frameworks. First, to prevent sensitive information from being inferred by other curious agents, we equip each agent with a local message sender with certain privacy constraints. Second, given prior knowledge about the privacy requirement of other agents, the receiver could strategically analyze the received noisy messages to statistically reduce errors due to the noisy communication. These two design principles correspond to two parts of our DPMAC framework respectively, *i.e.*, a *privacy-preserving local sender*, and a *privacy-aware receiver*.

5 Methodology

Based on our design principles, we now introduce our DPMAC framework, as shown in Figure 1. Our framework is general and flexible, which makes it compatible with any CTDE method.

5.1 Privacy-preserving Local Sender with Stochastic Gaussian Messages

In this section, we present the sender's perspective on the privacy guarantee. At time t, for agent i, a message function f_i^s is used to generate a message for communication. f_i^s takes a subset of transitions in local trajectory τ_i^t as input, where the subset is sampled uniformly without replacement from τ_i^t (denote the sampling rate as γ_1). This message is perturbed by the Gaussian mechanism with variance σ_i^2 [Dwork, 2006]. Agent i then samples a subset of other agents to share this message (denote the sampling rate as γ_2). The following theorem guarantees the DP of the sender.

Theorem 5.1 (Privacy guarantee for DPMAC). *Let $\gamma_1, \gamma_2 \in (0,1)$, and C be the ℓ_2 norm of the message functions. For any $\delta > 0$ and privacy budget ϵ_i, the communication of agent i satisfies (ϵ_i, δ)-DP when $\sigma_i^2 = \frac{14\gamma_2\gamma_1^2 NC^2\alpha}{\beta\epsilon_i}$, if we have $\alpha = \frac{\log\delta^{-1}}{\epsilon_i(1-\beta)} + 1 \leq 2\sigma'^2 \log\left(1/\gamma_1\alpha\left(1+\sigma'^2\right)\right)/3 + 1$ with $\beta \in (0,1)$ and $\sigma'^2 = \sigma_i^2/(4C^2) \geq 0.7$.*

With Theorem 5.1, one can directly translate a non-private MARL with a communication algorithm into a private one. However, as we shall see in our experiment section, directly injecting the privacy noise into existing MARL with communication algorithms may lead to serious performance degradation. In fact, the injected noise might jeopardize the useful information incorporated in the messages, or even leads to meaningless messages. To alleviate the negative impacts of the injected privacy noise on the cooperation between agents, we adopt a stochastic message sender in the sense that the messages sent by our sender are sampled from a learned message distribution. This makes DPMAC different from existing works in MARL that communicate through deterministic messages [Sukhbaatar *et al.*, 2016; Foerster *et al.*, 2016; Jiang and Lu, 2018; Das *et al.*, 2019; Ding *et al.*, 2020; Kim *et al.*, 2021].

In the following, we drop the dependency of parameters on t when it is clear from the context. Without loss of generality, let the message distribution be multivariate Gaussian and let p_i be the message sampled from the message distribution $\mathcal{N}(\mu_i, \Sigma_i)$, where $\mu_i = f_i^\mu(o_i, a_i; \theta_i^\mu)$ and $\Sigma_i = f_i^\sigma(o_i, a_i; \theta_i^\sigma)$ are the mean vector and covariance matrix learned by the sender, and θ_i^μ and θ_i^σ are the parameters of the sender's neural networks. Then θ_i^μ and θ_i^σ will be optimized towards making all the agents to send more effective messages to encourage better team cooperation and gain higher team rewards. For notational convenience, let $\theta_i^s = [\theta_i^{\mu\top}, \theta_i^{\sigma\top}]^\top$. Then the sent privatized message $m_i = p_i + u_i$ where $u_i \sim \mathcal{N}(0, \sigma_i^2\mathbf{I}_d)$ is the additive privacy noise. It is clear that $m_i \sim \mathcal{N}(\mu_i, \Sigma_i + \sigma_i^2\mathbf{I}_d)$ since p_i is independent from u_i. Counterfactually, let $m_i' \sim \mathcal{N}(\mu_i', \Sigma_i')$ be the sent message when it was not under any privacy constraint, where $\mu_i' = f_i^\mu(o_i, a_i; \theta_i^{\mu\prime})$ and $\Sigma_i' = f_i^\sigma(o_i, a_i; \theta_i^{\sigma\prime})$.

Let the optimal message distribution be $\mathcal{N}(\mu_i^*, \Sigma_i^*)$. We are interested to characterize $\theta_i^{s\prime}$ and θ_i^s. By the optimality of μ_i^*, Σ_i^*,

$$\begin{aligned}\theta_i^{s\prime} &= \arg\min_\theta D_{\mathrm{KL}}(\mathcal{N}(\mu_i', \Sigma_i') \| \mathcal{N}(\mu_i^*, \Sigma_i^*)) \\ &= \arg\min_\theta \log\frac{|\Sigma_i^*|}{|\Sigma_i'|} + \mathrm{tr}\{\Sigma_i^{*-1}\Sigma_i'\} + \|\mu_i' - \mu_i^*\|^2_{\Sigma_i^{*-1}} .\end{aligned} \tag{1}$$

Then under the privacy constraints, the stochastic sender will learn θ_i^s such that

$$\begin{aligned}\theta_i^s &= \arg\min_\theta D_{\mathrm{KL}}(\mathcal{N}(\mu_i, \Sigma_i + \sigma_i^2\mathbf{I}_d) \| \mathcal{N}(\mu_i^*, \Sigma_i^*)) \\ &= \arg\min_\theta \log\frac{|\Sigma_i^*|}{|\Sigma_i + \sigma_i^2\mathbf{I}_d|} + \mathrm{tr}\{\Sigma_i^{*-1}(\Sigma_i + \sigma_i^2\mathbf{I}_d)\} \\ &\qquad + \|\mu_i - \mu_i^*\|^2_{\Sigma_i^{*-1}} .\end{aligned} \tag{2}$$

Through Equation (2), it is possible to directly incorporate the distribution of privacy noise into the optimization process of the sender to help to learn θ_i^s such that $D_{\mathrm{KL}}(\mathcal{N}(\mu_i, \Sigma_i + \sigma_i^2\mathbf{I}_d) \| \mathcal{N}(\mu_i^*, \Sigma_i^*)) \leq D_{\mathrm{KL}}(\mathcal{N}(\mu_i', \Sigma_i') \| \mathcal{N}(\mu_i^*, \Sigma_i^*))$, which means that the sender could learn to send private message $m_i = p_i + u_i$ that is at least as effective as the non-private message m_i'. In this manner, the performance degradation is expected to be well alleviated.

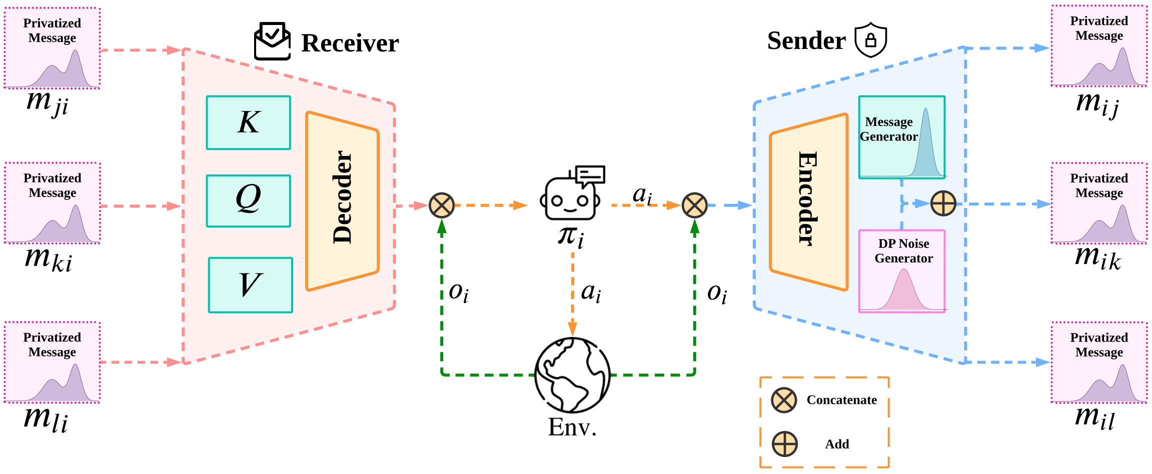


Figure 1: The overall structure of DPMAC. The message receiver of agent i integrates other agents' messages $\{m_{ji}, m_{ki}, m_{li}\}$ with the self-attention mechanism and the integrated message is fed into the policy π_i together with the observation o_i. Agent i interacts with the environment by taking action a_i. Then o_i and a_i are concatenated and encoded by a privacy-preserving message sender and sent to other agents.

5.2 Privacy-aware Message Receiver

As shown in our motivating example, the message receiver with knowledge a priori could statistically reduce the communication error in privacy-preserving scenarios. In the practical design, this motivation could be naturally instantiated with the gradient flow between the message sender and the message receiver.

Specifically, agent i first concatenates all the received privatized messages as $\boldsymbol{m}_{(-i)i} := \{m_{ji}\}_{j=1,j\neq i}^{N}$ and then decodes $\boldsymbol{m}_{(-i)i}$ into an aggregated message $q_i = f_i^r(\boldsymbol{m}_{(-i)i} \mid \theta_i^r)$ with the decoding function f_i^r parameterized by θ_i^r. Then a similar argument to the policy gradient theorem [Sutton *et al.*, 1999] states that the gradient of the receiver is

$$\begin{aligned}\nabla_{\theta_i^r} \mathcal{J}(\theta_i^r) = \mathbb{E}_{\boldsymbol{\tau},\boldsymbol{o},\boldsymbol{a}} \Big[\mathbb{E}_{\pi_i}[\nabla_{\theta_i^r} f_i^r \left(q_i \mid \boldsymbol{m}_{(-i)i}\right) \\ \cdot \nabla_{q_i} \log \pi_i \left(a_i \mid o_i, q_i\right) Q^{\boldsymbol{\pi}}(\boldsymbol{a}, \boldsymbol{o})]\Big] ,\end{aligned}$$

where $\mathcal{J}(\theta_i^r) = \mathbb{E}[G^1 \mid \boldsymbol{\pi}]$ is the cumulative discounted reward from the starting state. In this way, the receiver could utilize the prior knowledge σ_i of the privacy-preserving sender encoded in the gradient during the optimization process. Please refer to Appendix A and Appendix E for the complete pseudo code of DPMAC, and detailed optimization process of the message senders and receivers, respectively.

6 Privacy-preserving Equilibrium Analysis

As aforementioned, when considering the privacy constraints, the "cooperative" multi-agent games will *not* be purely cooperative anymore, due to the appearance of the trade-off between the team utility and each player's personal privacy. As the convergence of MARL algorithms could depend on the existence of NE, we first investigate such existence in privacy-preserving single-step games and then extend the result to privacy-preserving multi-step games.

6.1 Single-step Games

We study a class of two-player collaborative games, denoted as *collaborative game with privacy (CGP)*. The game involves two agents, each equipped with a privacy parameter p_n, $n \in \{1, 2\}$. The value of p_n represents the importance of privacy to agent n, with the larger value referring to greater importance. Let $\mathcal{M}$ be some message mechanism. We denote the privacy loss by $c^{\mathcal{M}}(p_n)$, which measures the quantity of the potential privacy leakage and is formally defined in Definition C.2. Besides, let $b\left(V_n, V_n^{\mathcal{M}}(p_1, p_2)\right)$ be the utility gained by measuring the gap between private value function $V_n^{\mathcal{M}}(p_1, p_2)$ and non-private value function V_n. Then the trade-off between the utility and the privacy is depicted by the total utility function $u_n(p_1, p_2)$ in Equation (3). The formal definition of CGP is given in Definition 6.1. See more details in Appendix C.1.

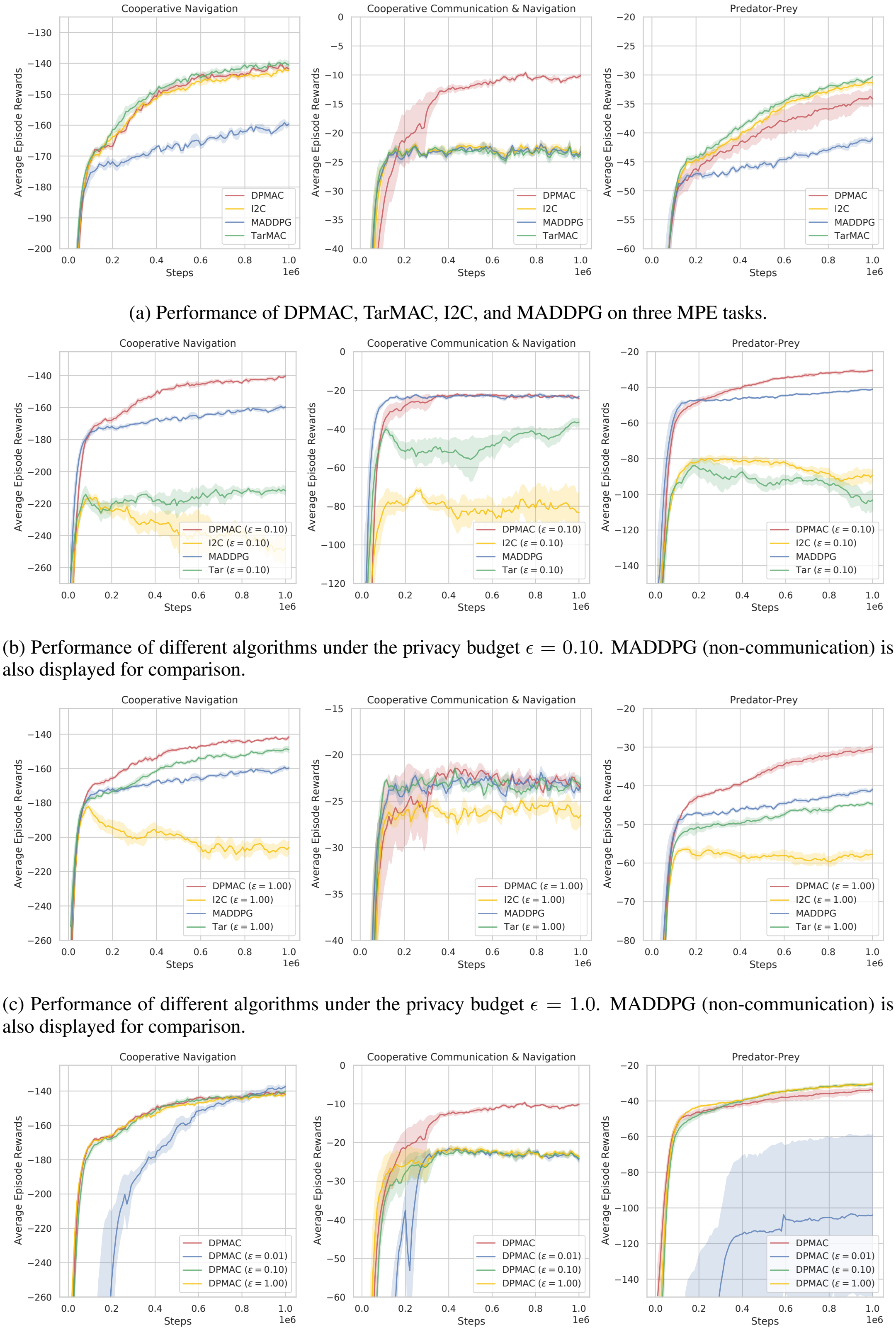
Definition 6.1 (Collaborative game with privacy (CGP))**.** *The collaborative game with privacy is denoted by a tuple $\langle \mathcal{N}, \Sigma, \mathcal{U} \rangle$, where $\mathcal{N} = \{1, 2\}$ is the the set of players, $\Sigma = \{p_1, p_2\}$ is the action set with $p_1, p_2 \in [0, 1]$ representing the privacy level, and $\mathcal{U} = \{u_1, u_2\}$ is the set of utility functions satisfying $\forall n \in \mathcal{N}$,*

$$u_n\left(p_1, p_2\right) = B_n b\left(V_n, V_n^{\mathcal{M}}(p_1, p_2)\right) - C_n^{\mathcal{M}} c^{\mathcal{M}}\left(p_n\right) . \quad (3)$$

Then the following theorem shows that if changes in the value function of each player can be expressed as a change in their own privacy parameter, then CGP is a potential game and a pure NE thereafter exists. The proof is deferred to Appendix C.1.

Theorem 6.1 (CGP's NE guarantee)**.** *The collaborative game with privacy has at least one non-trivial pure-strategy Nash equilibrium if $\partial_{p_1}^i V_1 = \partial_{p_2}^i V_2$, $\forall i \in \{1, 2\}$.*

Equilibrium in single round binary sums. Let us revisit our motivating example. Armed with the CGP framework, it



(a) Performance of DPMAC, TarMAC, I2C, and MADDPG on three MPE tasks.

(b) Performance of different algorithms under the privacy budget $\epsilon = 0.10$. MADDPG (non-communication) is also displayed for comparison.

(c) Performance of different algorithms under the privacy budget $\epsilon = 1.0$. MADDPG (non-communication) is also displayed for comparison.

(d) Performance of DPMAC under different privacy budgets ($\epsilon = 0.01, 0.10, 1.00$).

Figure 2: Performance of DPMAC and baseline algorithms. The curves are averaged over 5 seeds. Shaded areas denote 1 standard deviation.

is immediate that the single round binary sums game guarantees the existence of a NE. This result is formally stated in Theorem C.2 in Appendix C.1.

6.2 Multi-step Games

We now consider an extended version of single round binary sums named *multiple round sums*. Consider an N-player game where player i owns a saving $x_{i,t}$. Rather than sending a binary bit, the agent can choose to give out $b_{i,t}$ at round t. Meanwhile, each player i selects privacy level $p_{i,t}$ and sends messages to each other with a sender f_i^s encoding the information of $b_{i,t}$ with the privacy level $p_{i,t}$. The reward of the agent is designed to find a good trade-off between privacy and utility.

We first transform this game into a Markov potential game (MPG), with the reward of each agent transformed into a combination of the team reward and the individual reward. Then with existing theoretical results from Macua *et al.* (2018), we present the following result while deferring its proof to Appendix C.2.

Theorem 6.2 (NE guarantee in multiple round sums)**.** *If Assumptions 1, 2, 3, 4 (see Appendix C.2) are satisfied, our MPG has a NE with potential function J defined as,*

$$J(x_t, \pi(x_t)) = \sum_{j\in[N]} \left((1-p_{j,t})b_{j,t} + \alpha x_{j,t} + \beta p_{i,t}\right) . \quad (4)$$

7 Experiments

In this section, we present the experiment results and corresponding experiment analyses. Please see Appendix H for more detailed analyses of the experiment results.

Baselines. We implement our DPMAC upon MADDPG [Lowe *et al.*, 2017] (see Appendix D for concrete implementation details) and evaluate it against TarMAC [Das *et al.*, 2019], I2C [Ding *et al.*, 2020], and MADDPG. All algorithms are tested with and without the privacy requirement except for MADDPG, which involves no communication among agents. Since TarMAC and I2C do not have a local sender and have no DP guarantee, we add Gaussian noise to their receiver according to the noise variance specified in Theorem 5.1 for a fair comparison. Please see Appendix E for more training details.

Environments. We evaluate the algorithms on the multi-agent particle environment (MPE) [Mordatch and Abbeel, 2017], which is with continuous observation and discrete action space. This environment is commonly used among existing MARL literature [Lowe *et al.*, 2017; Jiang and Lu, 2018; Ding *et al.*, 2020; Kim *et al.*, 2021]. We evaluate a wide range of tasks in MPE, including cooperative navigation (CN), cooperative communication and navigation (CCN), and predator prey (PP). More details on the environmental settings are given in Appendix F.

Experiment results without privacy. We first compare DPMAC with TarMAC, I2C, and MADDPG on three MPE tasks without the privacy requirement. As shown in Figure 2a, DPMAC outperforms baselines on CCN task, and has comparable performance on CN and PP tasks. More detailed analyses of the experiment results without privacy are deferred to Appendix H.1.

Experiment results with privacy. We now investigate the performance of algorithms with communication under privacy constraints. In particular, Figure 2b and 2c show the performance under the privacy budget $\epsilon = 0.10, 1.0$ and both with $\delta = 10^{-4}$. We also include MADDPG as a non-communication baseline method. Overall, the privacy constraints impose obvious disturbances to the performance of all algorithms. Specifically, the performance of TarMAC and I2C degenerates significantly and becomes even inferior to the performance of MADDPG. However, in most cases, the performance of DPMAC with privacy constraints only suffers a slight decline and is still superior or comparable to the performance of MADDPG. See Appendix H.2 for more concrete analyses on the experiment results with privacy.

DPMAC under different privacy budgets. In Figure 2d, we further present the comparison between the performance of DPMAC under different privacy budgets. When $\epsilon = 0.01$, DPMAC still gains remarkable performance on CN and CCN tasks, while other baselines' performance suffers serious degeneration, as we have analyzed above. Besides, on the PP task under the privacy constraint with $\epsilon = 0.01$, DPMAC also suffers clear performance degradation. Overall, the experiments of DPMAC under different privacy budgets also show that DPMAC could automatically adjust the variance of the stochastic message sender so that it learns a noise-robust message representation. As shown in Figure 2d, DPMAC gains very close performance when $\epsilon = 0.1$ and $\epsilon = 1.0$, though the privacy requirements of $\epsilon = 0.1$ and $\epsilon = 1.0$ differ by one order of magnitude. However, one can see large performance gaps for the same baseline algorithms under different ϵ from Figure 2b and 2c. For clarity, we also present these performance gaps of TarMAC and I2C algorithms in Figure 3 and 4. Please see Appendix H.3 for more detailed analyses of the performance of DPMAC under different privacy budgets.

8 Conclusion

In this paper, we study the privacy-preserving communication in MARL. Motivated by a simple yet effective example of the binary sums game, we propose DPMAC, a new efficient communicating MARL algorithm that preserves agents' privacy through DP. Our algorithm is justified both theoretically and empirically. Besides, to show that the privacy-preserving communication problem is learnable, we analyze the single-step game and the multi-step game via the notion of MPG and show the existence of the Nash equilibrium. This existence further implies the learnability of several instances of MPG under privacy constraints. Extensive experiments conducted on 3 MPE tasks with varying privacy constraints demonstrate the effectiveness of DPMAC against the baseline methods.

Acknowledgments

The corresponding author Shuai Li is supported by National Natural Science Foundation of China (92270201, 62006151, 62076161) and Shanghai Sailing Program. Baoxiang Wang is partially supported by National Natural Science Foundation of China (62106213, 72150002) and Shenzhen Science and Technology Program (RCBS20210609104356063, JCYJ20210324120011032).

Contribution Statement

Canzhe Zhao and Yanjie Ze contributed equally to this work.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM CCS*, 2016.

[Alvim *et al.*, 2011] Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*. Springer, 2011.

[Balle *et al.*, 2016] Borja Balle, Maziar Gomrokchi, and Doina Precup. Differentially private policy evaluation. In *ICML*. PMLR, 2016.

[Beimel *et al.*, 2008] Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *Advances in Cryptology - CRYPTO 2008*. Springer, 2008.

[Chen *et al.*, 2020] Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. GS-WGAN: A gradient-sanitized approach for learning differentially private generators. In *NeurIPS 33*, 2020.

[Cheu, 2021] Albert Cheu. Differential privacy in the shuffle model: A survey of separations. *arXiv:2107.11839*, 2021.

[Das *et al.*, 2019] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. In *ICML 2019*. PMLR, 2019.

[Ding *et al.*, 2020] Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning individually inferred communication for multi-agent cooperation. In *NeurIPS 33*, 2020.

[Dong *et al.*, 2020] Tao Dong, Xiangyu Bu, and Wenjie Hu. Distributed differentially private average consensus for multi-agent networks by additive functional laplace noise. *Journal of the Franklin Institute*, (6), 2020.

[Dubey and Pentland, 2020] Abhimanyu Dubey and Alex 'Sandy' Pentland. Differentially-private federated linear bandits. In *NeurIPS 33*, 2020.

[Dubey, 2021] Abhimanyu Dubey. No-regret algorithms for private gaussian process bandit optimization. In *AISTATS 2021*. PMLR, 2021.

[Dwork *et al.*, 2014] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, (3-4), 2014.

[Dwork, 2006] Cynthia Dwork. Differential privacy. In *ICALP*. Springer, 2006.

[Foerster *et al.*, 2016] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *NeurIPS 29*, 2016.

[Garcelon *et al.*, 2021] Evrard Garcelon, Vianney Perchet, Ciara Pike-Burke, and Matteo Pirotta. Local differential privacy for regret minimization in reinforcement learning. In *NeurIPS 34*, 2021.

[Hassan *et al.*, 2020] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Differential privacy techniques for cyber physical systems: A survey. *IEEE Commun. Surv. Tutorials*, 2020.

[Jiang and Lu, 2018] Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. In *NeurIPS 31*, 2018.

[Kim *et al.*, 2021] Woojun Kim, Jongeui Park, and Youngchul Sung. Communication in multi-agent reinforcement learning: Intention sharing. In *ICLR*, 2021.

[Leonardos *et al.*, 2021] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv:2106.01969*, 2021.

[Li *et al.*, 2021] Ziyan Li, Quan Yuan, Guiyang Luo, and Jinglin Li. Learning effective multi-vehicle cooperation at unsignalized intersection via bandwidth-constrained communication. In *94th IEEE Vehicular Technology Conference, VTC 2021*. IEEE, 2021.

[Liao *et al.*, 2021] Chonghua Liao, Jiafan He, and Quanquan Gu. Locally differentially private reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2110.10133*, 2021.

[Lin *et al.*, 2021] Toru Lin, Jacob Huh, Christopher Stauffer, Ser-Nam Lim, and Phillip Isola. Learning to ground multi-agent communication with autoencoders. In *NeurIPS 34*, 2021.

[Lowe *et al.*, 2017] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS 30*, 2017.

[Macua *et al.*, 2018] Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. In *ICLR*, 2018.

[Monderer and Shapley, 1996] Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 1996.

[Mordatch and Abbeel, 2017] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv:1703.04908*, 2017.

[Peng *et al.*, 2017] Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, and Jun Wang. Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games. *CoRR*, abs/1703.10069, 2017.

[Rangwala and Williams, 2020] Murtaza Rangwala and Ryan Williams. Learning multi-agent communication through structured attentive reasoning. In *NeurIPS 33*, 2020.

[Sakuma *et al.*, 2008] Jun Sakuma, Shigenobu Kobayashi, and Rebecca N Wright. Privacy-preserving reinforcement learning. In *ICML*, 2008.

[Shalev-Shwartz *et al.*, 2016] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *CoRR*, abs/1610.03295, 2016.

[Singh *et al.*, 2019] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *ICLR*, 2019.

[Sukhbaatar *et al.*, 2016] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *NeurIPS 29*, 2016.

[Sutton *et al.*, 1999] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NeurIPS 12*, 1999.

[Tao *et al.*, 2022] Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. In *AISTATS*. PMLR, 2022.

[Tenenbaum *et al.*, 2021] Jay Tenenbaum, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Differentially private multi-armed bandits in the shuffle model. In *NeurIPS 34*, 2021.

[Tian *et al.*, 2021] Qi Tian, Kun Kuang, Baoxiang Wang, Furui Liu, and Fei Wu. Multi-agent communication with graph information bottleneck under limited bandwidth. *CoRR*, abs/2112.10374, 2021.

[Tossou and Dimitrakakis, 2017] Aristide Charles Yedia Tossou and Christos Dimitrakakis. Achieving privacy in the adversarial multi-armed bandit. In *AAAI*, 2017.

[Tucker *et al.*, 2022] Mycal Tucker, Julie Shah, Roger Levy, and Noga Zaslavsky. Towards human-agent communication via the information bottleneck principle. *CoRR*, abs/2207.00088, 2022.

[Wang and Hegde, 2019] Baoxiang Wang and Nidhi Hegde. Privacy-preserving q-learning with functional noise in continuous spaces. *NeurIPS*, 2019.

[Wang *et al.*, 2020a] Rundong Wang, Xu He, Runsheng Yu, Wei Qiu, Bo An, and Zinovi Rabinovich. Learning efficient multi-agent communication: An information bottleneck approach. In *ICML 2020*. PMLR, 2020.

[Wang *et al.*, 2020b] Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly decomposable value functions via communication minimization. In *ICLR*, 2020.

[Wei *et al.*, 2019] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu anda Kai Xu, and Zhenhui Li. Colight: Learning network-level cooperation for traffic signal control. In *ACM CIKM*, 2019.

[Yang *et al.*, 2020] Jiachen Yang, Alireza Nakhaei, David Isele, Kikuo Fujimura, and Hongyuan Zha. CM3: cooperative multi-goal multi-stage multi-agent reinforcement learning. In *ICLR*, 2020.

[Ye *et al.*, 2020] Dayong Ye, Tianqing Zhu, Zishuo Cheng, Wanlei Zhou, and S Yu Philip. Differential advising in multiagent reinforcement learning. *IEEE Transactions on Cybernetics*, 2020.

[Zhang and Lesser, 2011] Chongjie Zhang and Victor R. Lesser. Coordinated multi-agent reinforcement learning in networked distributed pomdps. In *AAAI*, 2011.

[Zhang *et al.*, 2019] Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Efficient communication in multi-agent reinforcement learning via variance based control. In *NeurIPS 32*, 2019.

[Zhang *et al.*, 2020] Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Succinct and robust multi-agent communication with temporal message control. In *NeurIPS 33*, 2020.

[Zhang *et al.*, 2021] Xin Zhang, Zhuqing Liu, Jia Liu, Zhengyuan Zhu, and Songtao Lu. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. In *NeurIPS 34*, 2021.

[Zheng *et al.*, 2020] Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning. In *NeurIPS 33*, 2020.

[Zhou, 2022] Xingyu Zhou. Differentially private reinforcement learning with linear function approximation. *Proc. ACM Meas. Anal. Comput. Syst.*, 2022.