FOP: Factorizing Optimal Joint Policy of Maximum-Entropy Multi-Agent Reinforcement Learning

Tianhao Zhang [*1] **Yueheng Li** [*1] **Chen Wang** [1] **Guangming Xie** [1] **Zongqing Lu** [1]

Abstract

Value decomposition recently injects vigorous vitality into multi-agent actor-critic methods. However, existing decomposed actor-critic methods cannot guarantee the convergence of global optimum. In this paper, we present a novel multi-agent actor-critic method, FOP, which can factorize the optimal joint policy induced by maximum-entropy multi-agent reinforcement learning (MARL) into individual policies. Theoretically, we prove that factorized individual policies of FOP converge to the global optimum. Empirically, in the well-known matrix game and differential game, we verify that FOP can converge to the global optimum for both discrete and continuous action spaces. We also evaluate FOP on a set of StarCraft II micromanagement tasks, and demonstrate that FOP substantially outperforms state-of-the-art decomposed value-based and actor-critic methods.

1. Introduction

Cooperative multi-agent reinforcement learning (MARL) has recently made great progress in many aspects by solving social dilemma (Hughes et al., 2018; Eccles et al., 2019), characterizing influence between agents (Jaques et al., 2019), optimizing both efficiency and fairness (Jiang & Lu, 2019), incorporating agent communication (Sukhbaatar et al., 2016; Das et al., 2019; Ding et al., 2020), considering relation between agents (Jiang et al., 2020) or their underlying network (Zhang et al., 2018; Qu et al., 2019).

Among these, fully cooperative MARL that maximizes a reward shared by all agents particularly has attracted much attention. Centralized training with decentralized execution (CTDE) is usually adopted as the learning paradigm for both value-based and actor-critic MARL methods, where global information can be accessed during centralized training and learned policies are executed with only local information in a decentralized way (Oliehoek et al., 2008; Kraemer & Banerjee, 2016). CTDE can resolve the non-stationarity under partial observability and has demonstrated great potential to address complex real-world problems, such as traffic signal control (Zhang et al., 2020; Xu et al., 2021) and autonomous driving cars (Zhou et al., 2020). However, CTDE suffers from the joint action-value function whose complexity grows exponentially with the number of agents, which restricts the performance of related MARL algorithms.

Value decomposition (Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Rashid et al., 2020; Yang et al., 2020; Wang et al., 2021) has witnessed success in handling the joint action-value function to effectively enable CTDE in value-based MARL methods. It aims to express the joint action-value function by individual action-value functions under the constraint of Individual-Global-Max (IGM) for the optimal consistency of joint and individual actions. However, existing decomposed value-based MARL methods limit to discrete actions.

When facing tasks with continuous actions, multi-agent actor-critic methods (Lowe et al., 2017; Foerster et al., 2018; Wei et al., 2018; Iqbal & Sha, 2019) are often taken into consideration. However, the exploration or sub-optimal behavior of one agent could negatively affect other agents' policy learning through the centralized critic (Wang et al., 2020). To address this problem, some studies (Wang et al., 2020; de Witt et al., 2020; Su et al., 2020) decompose the centralized critic to reduce the influence among agents during policy improvement. Although these decomposed actor-critic methods lead to improved performance, they *cannot* guarantee the learned individual policies lead to the optimal joint behavior.

In this paper, we aim to successfully enable agents to perform *globally* optimal behavior by simply executing the individual policy in *any* factorizable task, no matter discrete or continuous action space. We introduce a more general condition than IGM, Individual-Global-Optimal (IGO), which extends the factorizable task to both discrete and con-

[*]Equal contribution, the listing order is randomly determined. [1]Peking University. Correspondence to: Guangming Xie <xiegming@pku.edu.cn>, Zongqing Lu <zongqing.lu@pku.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

tinuous action spaces by the optimal consistency between joint policy and individual policies. With the IGO condition, we propose a *novel* decomposed multi-agent actor-critic method, FOP, which factorizes the joint policy induced by maximum-entropy MARL into individual policies. Theoretically, we show that such factorized individual policies converge to the global optimum. Empirically, in the well-known matrix game (Son et al., 2019) and differential game (Wei et al., 2018), we verify that FOP indeed fully converges to the global optimum for both discrete and continuous action spaces, while existing decomposed actor-critic methods do not. We also evaluate FOP on a set of StarCraft II micromanagement tasks (Samvelyan et al., 2019), and show that FOP substantially outperforms state-of-the-art decomposed value-based and actor-critic methods.

Related Work

Value decomposition (Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Yang et al., 2020; Rashid et al., 2020; Wang et al., 2021) has been increasingly popular in value-based MARL. These methods express the joint action-value function conditioned on global information as a function of individual action-value functions conditioned on local information, to satisfy the IGM consistency. Additivity and monotonicity are respectively considered in VDN (Sunehag et al., 2018) and QMIX (Rashid et al., 2018) which is further enhanced to Weighted QMIX (Rashid et al., 2020) by weighting joint actions. Qatten (Yang et al., 2020) decomposes the joint action-value function by both linearity and monotonicity. QTRAN (Son et al., 2019) and QPLEX (Wang et al., 2021) realize the full expressive capability of value decomposition by transforming IGM into optimization constraints and by duplex dueling architecture, respectively. Although witnessed success in some complex tasks, such as StarCraft II (Samvelyan et al., 2019), these decomposed value-based methods limit to discrete action space.

To handle continuous action space, decomposed actor-critic methods (de Witt et al., 2020; Su et al., 2020; Wang et al., 2020) are proposed to compute policy gradients using the decomposed critic instead of the centralized critic in classical multi-agent actor-critic methods (Lowe et al., 2017; Foerster et al., 2018; Iqbal & Sha, 2019). Su et al. (2020) utilized the VDN and QMIX structure to decompose the joint state-value function, while FacMADDPG (de Witt et al., 2020) learns the joint action-value function using QMIX. DOP (Wang et al., 2020) uses a decomposition structure similar to Qatten (Yang et al., 2020) to calculate policy gradients for off-policy tree backup and on-policy TD(λ). Although DOP proves that individual policies can converge to local optimal even if the decomposed critic has limited expressive capability, these decomposed actor-critic methods cannot guarantee the convergence to global optima. In this paper, we study how to factorize the joint policy into individual policies with the optimal consistency.

2. Background

2.1. Model

We consider a fully cooperative multi-agent task as a decentralized partially observable MDP (Dec-POMDP), defined by a tuple $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{U}_i\}, \mathcal{P}, r, \{\mathcal{Z}_i\}, \mathcal{O}, \gamma \rangle$. $\mathcal{N}$ is the set of agents with $|\mathcal{N}| = N$, and $\mathcal{S}$ is the set of states. At each time step, each agent $i \in \mathcal{N}$ chooses an action u_i from its action set $\mathcal{U}_i$, all agents together forming a joint action $\boldsymbol{u} \in \boldsymbol{\mathcal{U}} : \times_{i\in\mathcal{N}}\mathcal{U}_i$. The state $s \in \mathcal{S}$ transitions to the next state s' upon $\boldsymbol{u}$, according to the transition function $\mathcal{P}(s'|s, \boldsymbol{u}) : \mathcal{S} \times \boldsymbol{\mathcal{U}} \times \mathcal{S} \to [0, 1]$ and all agents receive a shared reward $r(s, \boldsymbol{u}) : \mathcal{S} \times \boldsymbol{\mathcal{U}} \to \mathbb{R}$. Moreover, each agent only obtains a partial observation $z_i \in \mathcal{Z}_i$ according to the observation function $\mathcal{O}(s, i) : \mathcal{S} \times \mathcal{N} \to \mathcal{Z}_i$ and learns an individual policy $\pi_i(u_i|\tau_i) : \mathcal{T}_i \times \mathcal{U}_i \to [0, 1]$ conditioned on the trajectory $\tau_i \in \mathcal{T}_i : (\mathcal{Z}_i \times \mathcal{U}_i)^*$. The objective of all agents is to maximize the cumulative return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in [0, 1]$ is the discount factor. Further, from a centralized prospective, we denote the joint trajectory of all agents as $\boldsymbol{\tau} \in \boldsymbol{\mathcal{T}} : \times_{i\in\mathcal{N}}\mathcal{T}_i$, and define a joint policy $\pi_{\text{jt}}(\boldsymbol{u}|\boldsymbol{\tau}) : \boldsymbol{\mathcal{T}} \times \boldsymbol{\mathcal{U}} \to [0, 1]$ that has a joint Q-function $Q_{\text{jt}}^{\pi_{\text{jt}}}(\boldsymbol{\tau}, \boldsymbol{u}) = \mathbb{E}_{\boldsymbol{\tau}_{t+1:\infty}, \boldsymbol{u}_{t+1:\infty}}[\sum_{k=0}^{\infty} \gamma^k r_{t+k}|\boldsymbol{\tau}_t, \boldsymbol{u}_t]$, where we drop the superscript π_{jt} if there is no confusion for simplicity.

2.2. Value Decomposition

Value decomposition (Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Yang et al., 2020; Rashid et al., 2020; Wang et al., 2021) expresses the joint Q-function $Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u})$ as a function of individual Q-functions $[Q_i(\tau_i, u_i)]_{i=1}^N$ to ease the difficulty of learning the joint Q-funciton. An important concept is Individual-Global-Maximum (IGM), which guarantees the consistency between individual optimal actions and optimal joint action, *i.e.*,

$$\arg\max_{\boldsymbol{u}} Q_{\text{jt}} = \big(\arg\max_{u_1} Q_1, \ldots, \arg\max_{u_N} Q_N\big).$$

VDN (Sunehag et al., 2018) and QMIX (Rashid et al., 2018) give sufficient conditions for IGM by additivity and monotonicity, respectively, as

$$\text{(VDN)} \quad Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i=1}^{N} Q_i(\tau_i, u_i),$$

$$\text{(QMIX)} \quad \frac{\partial Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u})}{\partial Q_i(\tau_i, u_i)} > 0, \ \forall i \in \mathcal{N}.$$

QTRAN (Son et al., 2019) transforms IGM into constraints which however make the optimization computationally intractable. The relaxation of these constraints makes QTRAN perform poorly in complex tasks (Mahajan et al., 2019).

QPLEX (Wang et al., 2021) provides IGM consistency by taking advantage of duplex dueling architecture,

$$Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i=1}^{N} Q_i(\boldsymbol{\tau}, u_i) + \sum_{i=1}^{N} (\lambda_i(\boldsymbol{\tau}, \boldsymbol{u}) - 1) A_i(\boldsymbol{\tau}, u_i),$$

where $A_i(\boldsymbol{\tau}, u_i) = w_i(\boldsymbol{\tau})(Q_i(\tau_i, u_i) - V_i(\tau_i))$, $V_i(\tau_i) = \max_{u_i} Q_i(\tau_i, u_i)$, and $w_i(\boldsymbol{\tau})$ is a positive weight. Although QPLEX guarantees the IGM consistency, the operator $\max_{u_i} Q_i$ still limits it to only discrete action space.

Naturally, we can also learn a decomposed critic under multi-agent actor-critic framework and update actors by the gradient computed using the decomposed critic (Wang et al., 2020; de Witt et al., 2020; Su et al., 2020). Among these, DOP (Wang et al., 2020) exploits a linear decomposition to estimate the centralized critic,

$$Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i} k_i(\boldsymbol{\tau}) Q_i(\boldsymbol{\tau}, u_i) + b(\boldsymbol{\tau}). \tag{1}$$

Then, each agent can learn a stochastic policy π_i by

$$g = \mathbb{E}_{\boldsymbol{\pi}}\big[\sum_{i} k_i(\boldsymbol{\tau}) \nabla \log \pi_i(u_i|\tau_i) Q_i(\boldsymbol{\tau}, u_i)\big], \tag{2}$$

or a deterministic policy μ_i by

$$g = \mathbb{E}_{\boldsymbol{\tau}}\big[\sum_{i} k_i(\boldsymbol{\tau}) \nabla \mu_i(\tau_i) \nabla_{u_i} Q_i(\boldsymbol{\tau}, u_i)|_{u_i=\mu_i(\tau_i)}\big]. \tag{3}$$

However, existing decomposed actor-critic methods cannot guarantee the convergence of global optimum. We will discuss how to address this in next section.

3. Analysis

DOP (Wang et al., 2020) analyzes that classic multi-agent actor-critic methods suffer from the exploration or suboptimality of other agents' policies, causing a large variance of policy gradients. On the other hand, decomposed policy gradients have a smaller variance, leading to improved performance. DOP proves that learning individual policies through (2) or (3) can converge to local optima with the limited expressive capability of decomposition structure. However, how to reach the global optimum remains *unknown*. Moreover, empirical results (Section 5.1) show that DOP and FacMADDPG (de Witt et al., 2020) always converge to local optima in both non-monotonic matrix game with discrete action space and differential game with continuous action space. Therefore, we can conclude that the decentralized policies (we use *decentralized policies* or *individual policies* exchangeably in the paper) in existing decomposed multi-agent actor-critic methods cannot promote the optimal joint behavior in some factorizable tasks.

To address this problem, the individual optimal behaviors should be consistent with the optimal joint behavior. Considering a sequential decision-making task that is amenable to factorization in centralized training, the optimal joint behavior can be generated by the *optimal joint policy* $\pi^*_{\text{jt}}(\boldsymbol{u}|\boldsymbol{\tau})$. We first define Individual-Global-Optimal (**IGO**), *i.e.*, the constraint of optimal policy consistency:

Definition 1 (IGO). *For an optimal joint policy $\pi^*_{jt}(\boldsymbol{u}|\boldsymbol{\tau}) : \boldsymbol{\mathcal{T}} \times \boldsymbol{\mathcal{U}} \to [0, 1]$, where $\boldsymbol{\tau} \in \boldsymbol{\mathcal{T}}$ is a joint trajectory, if there exist individual optimal policies $[\pi^*_i(u_i|\tau_i) : \mathcal{T} \times \mathcal{U} \to [0, 1]]_{i=1}^N$, such that the following holds:*

$$\pi^*_{jt}(\boldsymbol{u}|\boldsymbol{\tau}) = \prod_{i=1}^{N} \pi^*_i(u_i|\tau_i), \tag{4}$$

*then, we say that $[\pi_i]$ satisfy **IGO** for π_{jt} under $\boldsymbol{\tau}$.*

IGO implies the optimal consistency between the joint policy conditioned on the global information and individual policies conditioned on the local information. From IGO, we can see that if we locally move individual policies towards individual optimal policies (*e.g.*, by minimizing KL-divergence), the distance between the joint policy and optimal joint policy also decreases. In this way, we can obtain the joint policy improvement by individual policy improvement (see Appendix A.1 for details), which motivates this work.

We say a task is *factorizable*, which means that the global optimal solution of the task in centralized training can be achieved locally by individuals in decentralized execution. By describing the factorization from the policy perspective, the factorizable task is free from value functions or greedy policies. Therefore, IGO is more general than IGM to describe the factorizability of the task, since IGM can be seen as a special case of IGO if we specialize in the greedy policy (see Appendix A.2).

4. Learning to Factorize Optimal Joint Policy

In this section, we propose a novel multi-agent actor-critic method, **FOP**, which **F**actorizes the **O**ptimal joint **P**olicy induced by maximum-entropy MARL under the IGO constraint and achieves the global optimum through factorized individual policies.

4.1. Factorized Maximum-Entropy MARL

Adopting the paradigm of CTDE, the maximum-entropy objective (Ziebart, 2010) for multi-agent settings can be naturally defined as:

$$J(\pi_{\text{jt}}) = \sum_{t} \mathbb{E}_{\pi_{\text{jt}}}[r_t + \alpha \mathcal{H}(\pi_{\text{jt}}(\cdot|\boldsymbol{\tau}_t))],$$

where α denotes the *team temperature* parameter that determines the relative importance of the entropy term versus the

team reward. The joint policy π_{jt} is the Boltzmann policy:

$$\pi_{\text{jt}}(\boldsymbol{u}|\boldsymbol{\tau}) = \frac{\exp(\frac{Q_{\text{jt}}(\boldsymbol{\tau},\boldsymbol{u})}{\alpha})}{\sum_{\tilde{\boldsymbol{u}}}\exp(\frac{Q_{\text{jt}}(\boldsymbol{\tau},\tilde{\boldsymbol{u}})}{\alpha})}.$$

Further, the optimal joint soft-Q-function can be defined as:

$$\begin{aligned} Q^*_{\text{jt}}(\boldsymbol{\tau}_t, \boldsymbol{u}_t) =& r(\boldsymbol{\tau}_t, \boldsymbol{u}_t) + \\ & \mathbb{E}_{\boldsymbol{\tau}_{t+1},\dots}[\sum_{k=1}^{\infty}\gamma^k\big(r_{t+k} + \alpha\mathcal{H}(\pi^*_{\text{jt}}(\cdot|\boldsymbol{\tau}_{t+k}))\big)], \end{aligned}$$

where π^*_{jt} denotes the optimal joint soft policy and is as:

$$\begin{aligned} &\pi^*_{\text{jt}}(\boldsymbol{u}|\boldsymbol{\tau}) = \exp\big(\frac{1}{\alpha}(Q^*_{\text{jt}}(\boldsymbol{\tau},\boldsymbol{u}) - V^*_{\text{jt}}(\boldsymbol{\tau}))\big), \\ &\text{where } V^*_{\text{jt}}(\boldsymbol{\tau}) := \alpha\int_{\boldsymbol{\mathcal{U}}}\exp(\frac{1}{\alpha}Q^*_{\text{jt}}(\boldsymbol{\tau},\boldsymbol{u}))d\boldsymbol{u}. \end{aligned} \tag{5}$$

Similarly, for each agent i, the individual optimal soft policy conditions only on its own trajectory τ_i and can be defined as:

$$\begin{aligned} &\pi^*_i(u_i|\tau_i) = \exp\big(\frac{1}{\alpha_i}(Q^*_i(\tau_i,u_i) - V^*_i(\tau_i))\big), \\ &\text{where } V^*_i(\tau_i) := \alpha_i\log\int_{\mathcal{U}_i}\exp(\frac{1}{\alpha_i}Q^*_i(\tau_i,u))du, \end{aligned} \tag{6}$$

and α_i denotes *individual temperature* parameter.

To achieve the global optimum by individual policies, the individual optimal soft policies should satisfy IGO. By plugging (5) and (6) into (4), the optimal joint soft-Q-function Q_{jt} and the individual optimal soft-Q-functions $[Q_i]_{i=1}^N$ should satisfy the following:

$$Q^*_{\text{jt}}(\boldsymbol{\tau},\boldsymbol{u}) = \sum_{i=1}^{N}\frac{\alpha}{\alpha_i}\big[Q^*_i(\tau_i,u_i) - V^*_i(\tau_i)\big] + V^*_{\text{jt}}(\boldsymbol{\tau}), \tag{7}$$

where $V^*_{\text{jt}}(\boldsymbol{\tau})$ and $V^*_i(\tau_i)$ are the same as in (5) and (6), respectively. Therefore, the value decomposition is obtained according to the optimal consistency between joint policy and individual policies. This is also one of the reasons for using maximum-entropy RL framework, as its policy is directly tied to the value function.

4.2. Factorized Soft Policy Iteration

In this subsection, we introduce *factorized soft policy iteration* for the factorized maximum-entropy MARL, which is an extension of SAC (Haarnoja et al., 2018a) for multi-agent settings. We mathematically prove that factorized individual policies converge to the global optimum.

In policy evaluation of factorized soft policy iteration, we update the joint soft-Q-function Q_{jt} of π_{jt} by repeatedly applying soft Bellman operator $\Gamma_{\pi_{\text{jt}}}$ as

$$\begin{aligned} &\Gamma_{\pi_{\text{jt}}}Q_{\text{jt}}(\boldsymbol{\tau}_t,\boldsymbol{u}_t) := r_t + \gamma\mathbb{E}_{\boldsymbol{\tau}_{t+1}}[V_{\text{jt}}(\boldsymbol{\tau}_{t+1})], \\ &\text{where } V_{\text{jt}}(\boldsymbol{\tau}_t) = \mathbb{E}_{\pi_{\text{jt}}}[Q_{\text{jt}}(\boldsymbol{\tau}_t,\boldsymbol{u}_t) - \alpha\log\pi_{\text{jt}}(\boldsymbol{u}_t|\boldsymbol{\tau}_t)]. \end{aligned}$$

By this, we can obtain the joint soft-Q-function for any joint soft policy π_{jt}.

Lemma 1 (Joint Soft Policy Evaluation). *Consider the soft Bellman operator $\Gamma_{\pi_{jt}}$ and a mapping $Q^0_{jt}: \boldsymbol{\mathcal{T}}\times\boldsymbol{\mathcal{U}}\to\mathbb{R}$ with $|\boldsymbol{\mathcal{U}}| < \infty$, and define $Q^{k+1}_{jt} = \Gamma_{\pi_{jt}}Q^k_{jt}$. Then, the sequence Q^k_{jt} will converge to the joint soft-Q-function of π_{jt} as $k\to\infty$.*

Proof. See Appendix B.1. □

In policy improvement, the joint soft policy π_{jt} is updated based on *individual soft policies* $[\pi_i]_{i=1}^N$, where the individual soft policies are updated towards the exponential of the new individual soft-Q-functions.

We restrict the individual policy π_i of each agent i to some set of policies Π_i and update the individual policy according to:

$$\begin{aligned} \pi^{\text{new}}_i =& \arg\min_{\pi'_i\in\Pi_i} D_{\text{KL}}\Bigg(\pi'_i(\cdot|\tau_i)\Big|\Big| \\ & \exp\Big(\frac{1}{\alpha_i}\big(Q^{\pi^{\text{old}}_i}_i(\tau_i,\cdot) - V^{\pi^{\text{old}}_i}_i(\tau_i)\big)\Big)\Bigg). \end{aligned} \tag{8}$$

Based on this individual soft policy improvement, we will show that the newly projected joint soft policy has a higher value than the old joint soft policy with respect to the maximum-entropy RL objective.

Lemma 2 (Individual Soft Policy Improvement). *Let $\pi^{old}_i\in\Pi_i$ and π^{new}_i be the optimizer of the minimization problem defined in (8). Then, we have $Q^{\pi^{new}_{jt}}_{jt}(\boldsymbol{\tau}_t,\boldsymbol{u}_t) \geq Q^{\pi^{old}_{jt}}_{jt}(\boldsymbol{\tau}_t,\boldsymbol{u}_t)$ for all $(\boldsymbol{\tau}_t,\boldsymbol{u}_t)\in\boldsymbol{\mathcal{T}}\times\boldsymbol{\mathcal{U}}$ with $|\boldsymbol{\mathcal{U}}| < \infty$, where $\pi^{old}_{jt} = \prod_{i=1}^N\pi^{old}_i$ and $\pi^{new}_{jt} = \prod_{i=1}^N\pi^{new}_i$.*

Proof. See Appendix B.2. □

Factorized soft policy iteration alternates between joint soft policy evaluation and individual soft policy improvement, and provably converges to the global optimum among the policies in $[\Pi_i]_{i=1}^N$.

Theorem 1 (Factorized Soft Policy Iteration). *Considering joint soft policy can be factorized as $\pi_{jt} = \prod_{i=1}^N\pi_i$, repeated application of joint soft policy evaluation and individual soft policy improvement from $\pi_i\in\Pi_i, \forall i\in\mathcal{N}$ converges to a policy π^*_{jt} such that $Q^{\pi^*_{jt}}_{jt}(\boldsymbol{\tau},\boldsymbol{u}) \geq Q^{\pi_{jt}}_{jt}(\boldsymbol{\tau},\boldsymbol{u})$ for all $[\pi_i\in\Pi_i]_{i=1}^N$ and $(\boldsymbol{\tau},\boldsymbol{u})\in\boldsymbol{\mathcal{T}}\times\boldsymbol{\mathcal{U}}$, assuming $|\boldsymbol{\mathcal{U}}| < \infty$.*

Proof. See Appendix B.3. □

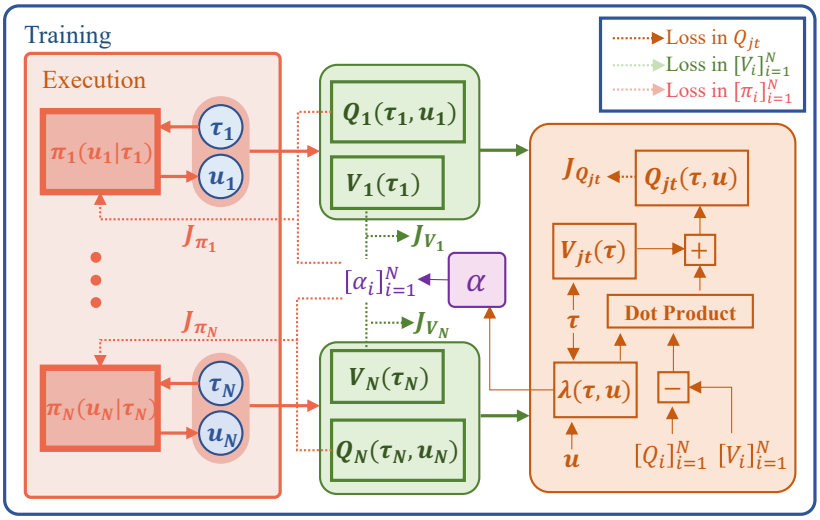


Figure 1. FOP architecture.

With factorized soft policy iteration, we can have decentralized policies with guaranteed convergence of global optimum. Next, we will show how to learn such policies using neural networks.

4.3. FOP Architecture

The overall architecture of FOP is illustrated in Figure 1. To factorize the optimal joint policy, we need to satisfy (7) and thus the architecture of FOP complies this principle.

For each agent i, there are an individual soft Q-network $Q_i^{\theta_i}(u_i, \tau_i)$, an individual soft V-network $V_i^{\phi_i}(t_i)$, and an individual soft policy $\pi_i^{\psi_i}(u_i|\tau_i)$, parameterized by θ_i, ϕ_i, and ψ_i, respectively. Two centralized components are introduced to compose Q_i and V_i into Q_{jt} according to (7). First, we can easily have a joint soft V-network $V_{\text{jt}}^{\Phi}(\boldsymbol{\tau})$, parameterized by Φ. However, there are some difficulties to handle the temperature parameters. Choosing the temperature is non-trivial since the entropy can vary unpredictably during training as the policy becomes better, SAC (Haarnoja et al., 2018b) introduces automating entropy adjustment by considering it as a constrained optimization problem. The team temperature α can be adjusted in the same way as SAC since the team objective is clear and team reward is available. For individual temperature α_i, as each agent's specific contribution to the team is unknown, we introduce a weight network $\lambda^{\Psi}(\boldsymbol{\tau}, \boldsymbol{u})$, parameterized by Ψ, to obtain $[\alpha_i]_{i=1}^N$ by

$$\alpha_i = \alpha \frac{\mathbb{E}_{u_i \sim \pi_i}[Q_i(\tau_i, u_i) - V_i(\tau_i)]}{\mathbb{E}_{\boldsymbol{u} \sim \pi_{\text{jt}}}[\lambda_i^{\Psi}(\boldsymbol{\tau}, \boldsymbol{u})\big(Q_i(\tau_i, u_i) - V_i(\tau_i)\big)]}.$$

Theoretical analysis on the convergence of individual policies with $\lambda^{\Psi}(\boldsymbol{\tau}, \boldsymbol{u})$ to the global optimum is provided in Appendix C.

Therefore, the factorization structure of FOP can be represented as:

$$Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i=1}^{N} \lambda^{\Psi}(\boldsymbol{\tau}, \boldsymbol{u}) \big[Q_i^{\theta_i}(\tau_i, u_i) - V_i^{\phi_i}(\tau_i)\big] + V_{\text{jt}}^{\Phi}(\boldsymbol{\tau}).$$

Although the factorization structure intuitively looks similar to QPLEX (Wang et al., 2021), they are different in nature. Q_{jt} in FOP is derived from the maximum-entropy RL objective, while Q_{jt} in QPLEX is based on the general RL objective. Besides, V_i in FOP is the soft V-function while V_i in QPLEX is $\max_{u_i} Q_i$. Thus, FOP is not only suitable for discrete action space but also for continuous action space, while QPLEX is only for discrete action space.

FOP is trained in a centralized manner, and during execution each agent uses its own individual soft policy $\pi_i^{\psi_i}(u_i|\tau_i)$ to take action u_i in a decentralized manner. We train individual soft Q-networks, joint soft V-network, and the weight network by minimizing the TD error:

$$\mathcal{L}([\theta_i]_{i=1}^N, \Phi, \Psi) = \mathbb{E}_{(\boldsymbol{\tau}, \boldsymbol{u}, r, \boldsymbol{\tau}', \boldsymbol{u}') \sim \mathcal{D}} \Bigg[\Big(Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - \Big(r + \gamma \big(\hat{Q}_{\text{jt}}(\boldsymbol{\tau}', \boldsymbol{u}') - \alpha \log \pi_{\text{jt}}(\boldsymbol{u}'|\boldsymbol{\tau}') \big) \Big) \Big)^2 \Bigg], \tag{9}$$

where $\mathcal{D}$ is the replay buffer, $\hat{Q}_{\text{jt}}$ is computed using target networks, and $\pi_{\text{jt}}(\boldsymbol{u}'|\boldsymbol{\tau}') = \prod_{i=1}^{N} \pi_i^{\psi_i}(u_i'|\tau_i')\}$. Noted that the gradients of Q_{jt} will not backpropagate to individual V-networks and policy networks. Each individual V-network is learned by minimizing:

$$\mathcal{L}(\phi_i) = \mathbb{E}_{\tau_i \sim \mathcal{D}} \Bigg[\Big(\mathbb{E}_{u_i} \big[Q_i(\tau_i, u_i) - \alpha_i \log \pi_i(u_i|\tau_i) \big] - V_i(\tau_i) \Big)^2 \Bigg]. \tag{10}$$

Algorithm 1 FOP

for $episode = 1$ to $max\_training\_episode$ **do**
  Initialize the environment
  for $t = 1$ to $max\_episode\_length$ **do**
    For each agent i, get τ_i, take action $u_i \sim \pi_i^{\psi_i}(\cdot|\tau_i)$
    Execute joint action $\boldsymbol{u}$, observe reward r, and each agent gets τ_i'
    Store $(\boldsymbol{\tau}, \boldsymbol{u}, r, \boldsymbol{\tau}')$ in replay buffer $\mathcal{D}$
    if $time\_to\_update$ **then**
      Sample a random minibatch of $\mathcal{K}$ samples from $\mathcal{D}$: $\{(\boldsymbol{\tau}_k, \boldsymbol{u}_k, r_k, \boldsymbol{\tau}_k')\}_{\mathcal{K}}$
      Update $[\theta_i]_{i=1}^N, \Phi, \Psi$ using (9)
      Update $[\phi]_{i=1}^N$ using (10)
      Update $[\psi_i]_{i=1}^N$ using (11)
      Update temperature parameters α and $[\alpha_i]_{i=1}^N$
      Update target networks: $[\hat{\theta}_i]_{i=1}^N, \hat{\Phi}, \hat{\Psi}$
    end if
  end for
end for

Each individual soft policy can be learned by directly minimizing the expected KL-divergence in (8). As the log partition function V_i does not affect the optimization, the objective of individual soft policy network can be written as

$$J_{\pi_i}(\psi_i) = \mathbb{E}_{\substack{\tau_i \sim \mathcal{D} \\ u_i \sim \pi_i}} [\alpha_i \log \pi_i(u_i|\tau_i) - Q_i(\tau_i, u_i)]. \quad (11)$$

As the objective does not rely on other agents' behaviors, the gradient has a smaller variance than that of methods based on the centralized critic. Moreover, under the factorized soft policy iteration theorem, theses individual soft policies converge to the global optimum. For completeness, we summarize the training of FOP in Algorithm 1.

5. Experiments

In this section, first we empirically study the optimality of FOP by two didactic games for discrete and continuous action spaces, compared with existing decomposed actor-critic methods. By ablation studies, we verify the robustness of FOP to temperature parameter, and the importance of the IGO constraint and soft policy. Then, in StarCraft II, we demonstrate that FOP outperforms state-of-the-art baselines including both decomposed value-based and actor-critic methods. Note that in the experiments, all the learning curves are plotted based on five training runs with difference random seeds using mean and standard deviation with confidence internal 95%.

5.1. Matrix Game and Differential Game

In both discrete matrix game and continuous differential game, we investigate whether FOP can converge to optimal

Table 1. The non-monotonic cooperative matrix game. Boldface means the optimal action from individual policies.

(a) Payoff of matrix game

u_1 \ u_2	A	B	C
A	**8**	-12	-12
B	-12	0	0
C	-12	0	0

(b) FOP: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**3.3(A)**	0.1(B)	0.1(C)
4.7(A)	**8.0**	-12.0	-12.0
-0.1(B)	-12.0	0.0	0.0
-0.1(C)	-12.0	0.0	0.0

(c) FacM: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	-0.8(A)	**-0.1(B)**	-0.2(C)
-0.6(A)	-8.5	-5.4	-5.8
-0.1(B)	-6.0	-1.5	-2.1
-0.0(C)	-5.3	**-0.4**	-1.0

(d) DOP: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	-2.5(A)	-1.3(B)	**-0.0(C)**
-2.2(A)	-7.8	-6.0	-4.2
-1.0(B)	-6.1	-4.4	-2.6
-0.3(C)	-4.2	-2.4	**-0.7**

compared with existing decomposed actor-critic methods including FacMADDPG (abbreviated as FacM) (de Witt et al., 2020) and DOP (Wang et al., 2020). The two games have one common characteristic: some destructive penalties are around with the optimal solution, making the sub-optimal solution have a higher expected return than that of the optimal solution (Wei et al., 2018). Thus, these tasks pose a dramatic challenge to general actor-critic methods (Lowe et al., 2017; Iqbal & Sha, 2019) because the policy gradient tends to converge to sub-optima.

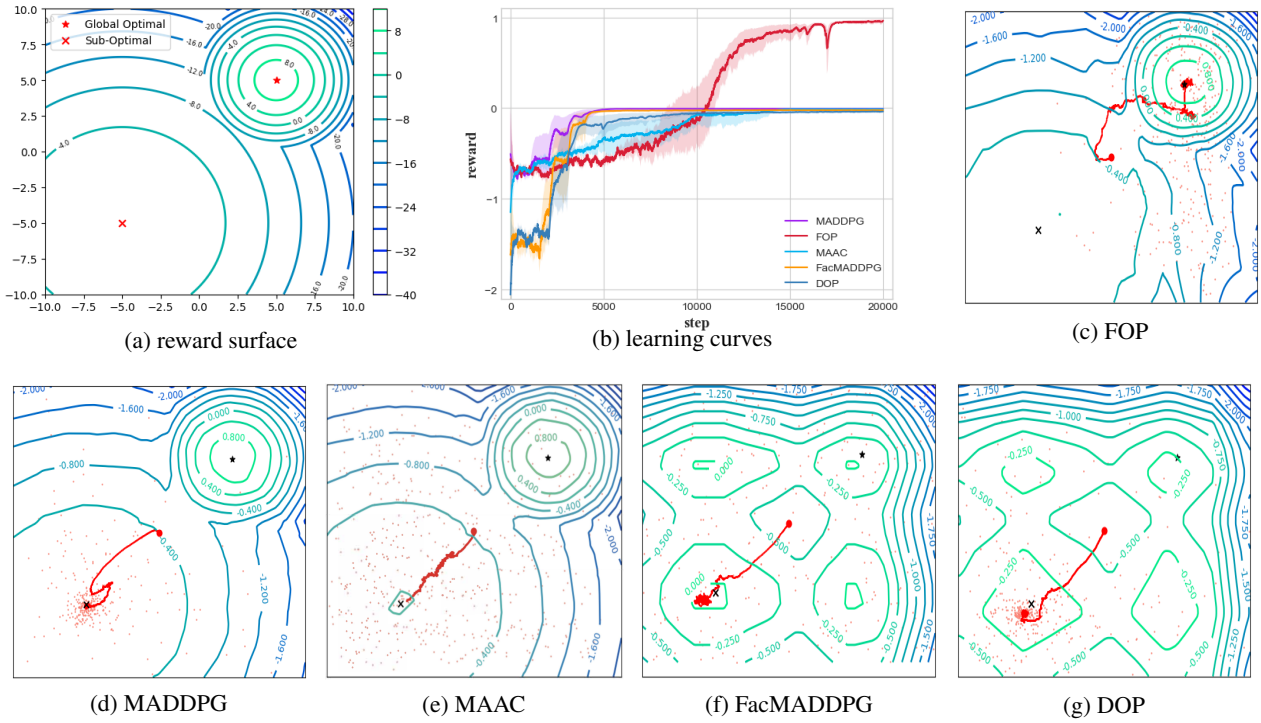
Matrix Game

The matrix game proposed by QTRAN (Son et al., 2019) is as illustrated in Table 1a. Such a non-monotonic matrix game consists of two agents with three actions and a shared reward. We show the results of FOP, FacMADDPG, and DOP over $10k$ learning steps, as in Table 1b, 1c, and 1d, respectively. FOP achieves the optimum, while FacMADDPG and DOP fall into the sub-optimum induced by miscoordination penalties. Note that FOP, QPLEX (Wang et al., 2021), and QTRAN (Son et al., 2019) are the only three algorithms that can successfully converge to the optimum, and FOP is the *first* decomposed actor-critic method. More details about the experiments on the matrix game are included in Appendix D.1.

Differential Game

We adopt the differential game, the Max of Two Quadratic (MTQ) Game, from Panait et al. (2007) and Wei et al. (2018). The MTQ game consists of two agents, where each has one-dimensional bounded continuous action space with a shared reward function:

$$\begin{cases} f_1 = 0.8 \times \left[-(\frac{u_1+5}{3})^2 - (\frac{u_2+5}{3})^2 \right] \\ f_2 = 1 \times \left[-(\frac{u_1-5}{1})^2 - (\frac{u_2-5}{1})^2 \right] + 10 \\ r(u_1, u_2) = \max(f_1, f_2). \end{cases}$$



(a) reward surface (b) learning curves (c) FOP

(d) MADDPG (e) MAAC (f) FacMADDPG (g) DOP

Figure 2. The Max of Two Quadratic Game: (a) reward surface, (b) the learning curves of all the methods; and (c)-(g) their learning paths.

The reward surface is illustrated in Figure 2a, where there is a sub-optimal solution 0 at $(-5, -5)$ and a global optimal solution 10 at $(5, 5)$. In MTQ, we compare FOP with both regular multi-agent actor-critic methods, *i.e.*, MADDPG (Lowe et al., 2017) and MAAC (Iqbal & Sha, 2019), and decomposed actor-critic methods (FacMADDPG and DOP). Both agents scaled their reward by 0.1. The learning curves ($20k$ steps) of all the methods are illustrated in Figure 2b, where agents' policies are initialized at $(0, 0)$, and the learning paths of all the methods are depicted in Figure 2c to 2g.

FOP always converges to the global optimum while all other baselines fall into the sub-optimum on the left. Note that FOP is the *first* value decomposition MARL method that can successfully converge to the optimum in the MTQ. Interestingly, by comparing the learning paths where the scattered red dots are the exploration trails before convergence, we can see that FOP is the only one that can not only estimate Q_{jt} accurately (comparing with the reward surface in Figure 2a), but also converge to the global optimum. In contrast, as depicted in Figure 2f and 2g, FacMADDPG and DOP converge to the sub-optimum and also have limitation to express Q_{jt}. In addition, although MADDPG and MAAC can accurately estimate Q_{jt}, they fail to converge to the optimum. One of the possible reasons is that the gradient of the actors of MADDPG and MAAC has large variance due to the centralized critic (Wang et al., 2020), which is also verified in the experiment. More details are available in Appendix D.1. The pathology of finding a sub-optimal solution is also called *relative overgeneralization* (Wei & Luke, 2016; Castellini et al., 2019), which is discussed further in Appendix D.2.

Table 2. Ablation study of FOP in the non-monotonic cooperative matrix game. Boldface means the optimal action from individual policies.

(a) FOP: $\hat{\alpha} = 1$

Q_1 \ Q_2	**0.3(A)**	0.1(B)	0.1(C)
7.7(A)	**8.0**	-9.4	-4.9
0.0(B)	3.8	0.1	-0.0
-0.1(C)	3.8	-0.1	-0.0

(b) FOP: $\hat{\alpha} = 0.5$

Q_1 \ Q_2	**5.5(A)**	1.2(B)	0.8(C)
2.4(A)	**8.0**	-11.9	-11.8
-0.5(B)	-11.7	-0.1	-0.0
-0.8(C)	-11.8	0.0	-0.0

(c) FOP: $\hat{\alpha} = 0.1$

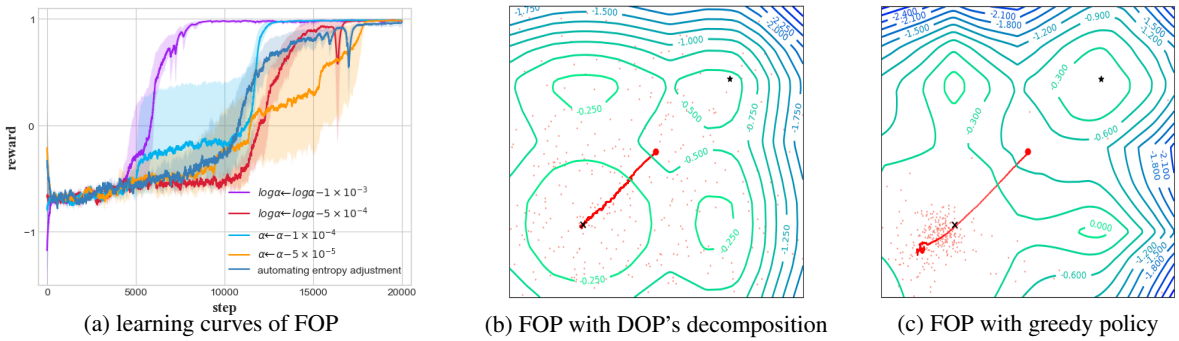
Q_1 \ Q_2	**7.6(A)**	2.3(B)	0.2(C)
0.4(A)	**8.0**	-12.0	-12.0
-0.1(B)	-12.0	-0.0	-0.0
-0.1(C)	-12.0	-0.1	0.1

(d) FOP: $\hat{\alpha} = 0.01$

Q_1 \ Q_2	**3.3(A)**	0.1(B)	0.1(C)
4.7(A)	**8.0**	-12.0	-12.0
-0.1(B)	-12.0	0.0	0.0
-0.1(C)	-12.0	0.0	0.0

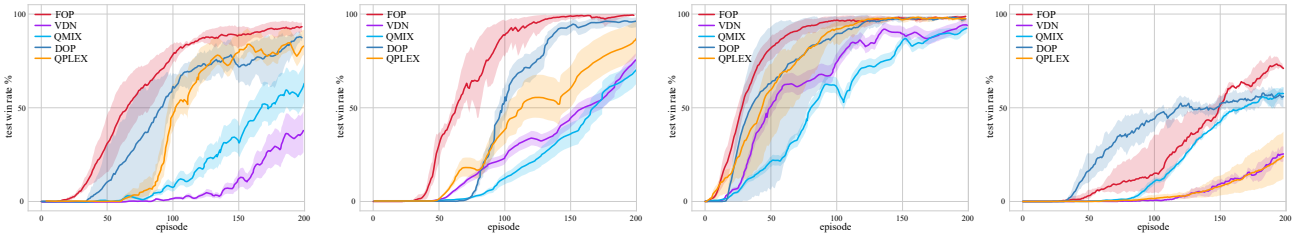
Ablation Studies

We first investigate the robustness of FOP to the team temperature α. Table 2 shows the performance of FOP in the matrix game, where FOP is trained by linearly annealing α from 1 to different $\hat{\alpha}$ over $10k$ steps. We can see that FOP



(a) learning curves of FOP (b) FOP with DOP's decomposition (c) FOP with greedy policy

Figure 3. Ablation study in MTQ: (a) the learning curves of FOP with different annealing approaches; (b) the learning path of FOP with DOP's decomposition; (c) the learning path of FOP with greedy policy.



(a) 2c_vs_64zg (b) 3s_vs_3z (c) MMM (d) MMM2

Figure 4. Learning curves of all the methods in four maps of StarCraft II.

can always find the optimal solution, even if the policy is not greedy ($\hat{\alpha}$ is not 0). We also train FOP in the differential game by different temperature adjustment approaches. The learning curves are shown in Figure 3a. We can see that FOP always converges to the optimal, and the only difference under these approaches is the convergence rate. These experiments demonstrate FOP is robust with the team temperature.

Next, we change FOP's factorization to DOP's linear decomposition in (1). By comparing the learning path in Figure 3b against Figure 2c, we can see that without the factorization based on IGO, even if the joint action space is well explored, the biased estimate of the joint Q-value leads the individual policies to the sub-optimal, which reveals the significance of the IGO constraint.

Last, we replace the soft policy of FOP to the greedy policy by setting $\alpha = \alpha_i = 0$ during the training. Figure 3c illustrates that the lack of effective exploration biases the estimate of the joint Q-value and leads the policies to the sub-optimal, which reveals the importance of the soft policy. More details of the ablation studies are included in Appendix D.3.

5.2. StarCraft II

We evaluate FOP on the challenging StarCraft Multi-Agent Challenge (SMAC) benchmark (Samvelyan et al., 2019) in four maps including 2c_vs_64zg, 3s_vs_3z, MMM and MMM2. The baselines include the decomposed actor-critic method (stochastic DOP, the best-performing DOP in StarCraft II (Wang et al., 2020)) and the decomposed value-based methods (VDN, QMIX and QPLEX). Results are shown in Figure 4, where each episode has about $10k$ time steps, totally two million time steps. In general, FOP outperforms the baselines in all the scenarios. In 2c_vs_64zg, 3s_vs_3z and MMM, FOP outperforms all the baselines in both convergence speed and final performance. In MMM2, DOP soon falls into sub-optima though it converges fast at first, while FOP can keep exploring for a better policy. More details of the StarCraft II experiments are included in Appendix D.4.

6. Conclusion

We proposed FOP, which learns factorized individual policies in cooperative MARL. Unlike existing decomposed actor-critic methods that are only guaranteed to converge to local optimum, the factorized individual policies of FOP provably converge to the global optimum under the derived factorized soft policy iteration theorem. Empirically, in the well-known matrix game and differential game, we verified FOP can converge to the global optimum for both discrete and continuous action spaces, and it also substantially outperforms the state-of-the-art decomposed value-based and actor-critic methods on a set of StarCraft II tasks.

References

Castellini, J., Oliehoek, F. A., Savani, R., and Whiteson, S. The representational capacity of action-value networks for multi-agent reinforcement learning. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2019.

Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., and Pineau, J. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning (ICML)*, 2019.

de Witt, C. S., Peng, B., Kamienny, P.-A., Torr, P., Böhmer, W., and Whiteson, S. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709*, 2020.

Ding, Z., Huang, T., and Lu, Z. Learning individually inferred communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., and Leibo, J. Z. Learning reciprocity in complex sequential social dilemmas. *arXiv:1903.08082*, 2019.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018a.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018b.

Hughes, E., Leibo, J. Z., Phillips, M. G., Tuyls, K., Duéñez-Guzmán, E. A., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K. R., Koster, R., et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *arXiv preprint arXiv:1803.08884*, 2018.

Iqbal, S. and Sha, F. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

Jiang, J. and Lu, Z. Learning fairness in multi-agent systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jiang, J., Dun, C., Huang, T., and Lu, Z. Graph convolutional reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Kraemer, L. and Banerjee, B. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190:82–94, 2016.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Mahajan, A., Rashid, T., Samvelyan, M., and Whiteson, S. Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Oliehoek, F. A., Spaan, M. T., and Vlassis, N. Optimal and approximate q-value functions for decentralized pomdps. *Journal of Artificial Intelligence Research*, 32:289–353, 2008.

Panait, L., Luke, S., and Wiegand, R. P. Biasing coevolutionary search for optimal multiagent behaviors. *IEEE Transactions on Evolutionary Computation*, 10(6):629–645, 2007.

Qu, C., Mannor, S., Xu, H., Qi, Y., Song, L., and Xiong, J. Value propagation for decentralized networked deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., and Whiteson, S. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Rashid, T., Farquhar, G., Peng, B., and Whiteson, S. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Samvelyan, M., Rashid, T., Schroeder de Witt, C., Farquhar, G., Nardelli, N., Rudner, T. G., Hung, C.-M., Torr, P. H., Foerster, J., and Whiteson, S. The starcraft multi-agent challenge. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2019.

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., and Yi, Y. Qtran: Learning to factorize with transformation for

cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

Su, J., Adams, S., and Beling, P. A. Value-decomposition multi-agent actor-critics. *arXiv preprint arXiv:2007.12306*, 2020.

Sukhbaatar, S., Fergus, R., et al. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V. F., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2018.

Sutton, R. S. and Barto, A. G. Reinforcement learning: An introduction, 2018.

Wang, J., Ren, Z., Liu, T., Yu, Y., and Zhang, C. Qplex: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Wang, Y., Han, B., Wang, T., Dong, H., and Zhang, C. Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations (ICLR)*, 2020.

Wei, E. and Luke, S. Lenient learning in independent-learner stochastic cooperative games. *Journal of Machine Learning Research*, 17(1):2914–2955, 2016.

Wei, E., Wicke, D., Freelan, D., and Luke, S. Multiagent soft q-learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Xu, B., Wang, Y., Wang, Z., Jia, H., and Lu, Z. Hierarchically and cooperatively learning traffic signal control. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Yang, Y., Hao, J., Liao, B., Shao, K., Chen, G., Liu, W., and Tang, H. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv preprint arXiv:2002.03939*, 2020.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar., T. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning (ICML)*, 2018.

Zhang, Z., Yang, J., and Zha, H. Integrating independent and centralized multi-agent reinforcement learning for traffic signal network optimization. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2020.

Zhou, M., Luo, J., Villela, J., Yang, Y., Rusu, D., Miao, J., Zhang, W., Alban, M., Fadakar, I., Chen, Z., Huang, A. C., Wen, Y., Hassanzadeh, K., Graves, D., Chen, D., Zhu, Z., Nguyen, N. M., Elsayed, M., Shao, K., Ahilan, S., Zhang, B., Wu, J., Fu, Z., Rezaee, K., Yadmellat, P., Rohani, M., Nieves, N. P., Ni, Y., Banijamali, S., Cowen-Rivers, A. I., Tian, Z., Palenicek, D., Bou-Ammar, H., Zhang, H., Liu, W., Hao, J., and Wang, J. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. In *Conference on Robot Learning (CoRL)*, 2020.

Ziebart, B. D. Modeling purposeful adaptive behavior with the principle of maximum causal entropy, 2010.