Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning

Xueguang Lyu
Northeastern University
lu.xue@northeastern.edu

Yuchen Xiao
Northeastern University
xiao.yuch@northeastern.edu

Brett Daley
Northeastern University
b.daley@northeastern.edu

Christopher Amato
Northeastern University
c.amato@northeastern.edu

ABSTRACT

Centralized Training for Decentralized Execution, where agents are trained offline using centralized information but execute in a decentralized manner online, has gained popularity in the multi-agent reinforcement learning community. In particular, actor-critic methods with a centralized critic and decentralized actors are a common instance of this idea. However, the implications of using a centralized critic in this context are not fully discussed and understood even though it is the standard choice of many algorithms. We therefore formally analyze centralized and decentralized critic approaches, providing a deeper understanding of the implications of critic choice. Because our theory makes unrealistic assumptions, we also empirically compare the centralized and decentralized critic methods over a wide set of environments to validate our theories and to provide practical advice. We show that there exist misconceptions regarding centralized critics in the current literature and show that the centralized critic design is not strictly beneficial, but rather both centralized and decentralized critics have different pros and cons that should be taken into account by algorithm designers.

ACM Reference Format:
Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. 2021. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning. In *Proc. of the 20th International Conference on Autonomous Agents and multi-agent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

Centralized Training for Decentralized Execution (CTDE), where agents are trained offline using centralized information but execute in a decentralized manner online, has seen widespread adoption in multi-agent reinforcement learning (MARL) [9, 15, 28]. In particular, actor-critic methods with centralized critics have become popular after being proposed by Foerster et al. [10] and Lowe et al. [20], since the critic can be discarded once the individual actors are trained. Despite the popularity of centralized critics, the choice is not discussed extensively and its implications for learning remain largely unknown.

One reason for this lack of analysis is that recent state-of-the-art works built on top of a centralized critic focus on other issues such as multi-agent credit assignment [10, 43], multi-agent exploration [8], teaching [29] or emergent tool use [1]. However, state-of-the-art methods with a centralized critic do not compare with decentralized critic versions of their methods. Therefore, without precise theory or tests, previous works relied on intuitions and educated guesses. For example, one of the pioneering works on centralized critics, MADDPG [20], mentions that providing more information to critics eases learning and makes learning coordinated behavior easier. Later works echoed similar viewpoints, suspecting that a centralized critic might speed up training [49], or that it reduces variance [6], is more robust [36], improves performance [18] or stabilizes training [19].

In short, previous works generally give the impression that a centralized critic is an obvious choice without compromises under CTDE, and there are mainly two advertised benefits: *a*) a centralized critic fosters "cooperative behavior", *b*) a centralized critic also stabilizes (or speeds up) training. It makes intuitive sense because training a global value function on its own (i.e., joint learning [5]) would help with cooperation issues and has much better convergence guarantees due to the stationary learning targets. However, these intuitions have never been formally proven or empirically tested; since most related works focus on additional improvements on top of a centralized critic, the centralized critic is usually seen as part of a basic framework rather than an optional hyperparameter choice.[1] In this paper, we look into these unvalidated claims and point out that common intuitions turn out to be inaccurate.

First, we show theoretically that a centralized critic does not necessarily improve cooperation compared to a set of decentralized critics. We prove that the two types of critics provide the decentralized policies with precisely the same gradients in expectation. We validate this theory on classical cooperation games and more realistic domains and report results supporting our theory.

Second, we show theoretically that the centralized critic results in higher variance updates of the decentralized actors assuming converged on-policy value functions. Therefore, we emphasize that stability of *value function learning* does not directly translate to a reduced variance in *policy learning*. We also discuss that, in practice, this results in a bias-variance trade-off. We analyze straightforward examples and empirical evaluations, confirming our theory and showing that the centralized critic often makes the policy learning less stable, contrary to the common intuition.

Proc. of the 20th International Conference on Autonomous Agents and multi-agent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

[1] Methods are also typically implemented with state-based critics [8, 10, 43, 49] instead of the history-based critics we use in this paper, which might be another reason for performance differences. However, we only consider history-based critics to fairly compare the use of centralized and decentralized critics with the same type of information.

Finally, we test standard implementations of the methods over a wide range of popular domains and discuss our empirical findings where decentralized critics often outperform a centralized critic. We further analyze the results and discuss possible reasons for these performance differences. We therefore demonstrate room for improvement with current methods while laying theoretical groundwork for future work.

2 RELATED WORK

Recent deep MARL works often use the CTDE training paradigm. Value function based CTDE approaches [7, 23, 33, 34, 38, 39, 44, 45, 48] focus on how centrally learned value functions can be reasonably decoupled into decentralized ones and have shown promising results. Policy gradient methods on CTDE, on the other hand, have heavily relied on centralized critics. One of the first works utilizing a centralized critic was COMA [10], a framework adopting a centralized critic with a counterfactual baseline. For convergence properties, COMA establishes that the overall effect on decentralized policy gradient with a centralized critic can be reduced to a single-agent actor-critic approach, which ensures convergence under similar assumptions [17]. In this paper, we take the theory one step further and show convergence properties for centralized and decentralized critics as well as their respective policies, while giving a detailed bias-variance analysis.

Concurrently with COMA, MADDPG [20] proposed to use a dedicated centralized critic for each agent in semi-competitive domains, demonstrating compelling empirical results in continuous action environments. M3DDPG [19] focuses on the competitive case and extends MADDPG to learn robust policies against altering adversarial policies by optimizing a minimax objective. On the cooperative side, SQDDPG [43] borrows the counterfactual baseline idea from COMA and extends MADDPG to achieve credit assignment in fully cooperative domains by reasoning over each agent's marginal contribution. Other researchers also use critic centralization for emergent communication with decentralized execution in TarMAC [6] and ATOC [13]. There are also efforts utilizing an attention mechanism addressing scalability problems in MAAC [12]. Besides, teacher-student style transfer learning LeCTR [29] also builds on top of centralized critics, which does not assume expert teachers. Other focuses include multi-agent exploration and credit assignment in LIIR [8], goal-conditioned policies with CM3 [49], and for temporally abstracted policies [4]. Based on a centralized critic, extensive tests on a more realistic environment using self-play for hide-and-seek [1] have demonstrated impressive results showing emergent tool use. However, as mentioned before, these works use centralized critics, but none of them specifically investigate the effectiveness of centralized critics, which is the main focus of this paper.

3 BACKGROUND

This section introduces the formal problem definition of cooperative MARL with decentralized execution and partial observability. We also introduce the single-agent actor-critic method and its most straight-forward multi-agent extensions.

3.1 Dec-POMDPs

A decentralized partially observable Markov decision process (Dec-POMDP) is an extension of an MDP in decentralized multi-agent settings with partial observability [27]. A Dec-POMDP is formally defined by the tuple $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}, \mathcal{T}, \mathcal{R}, \{\Omega_i\}, \mathcal{O} \rangle$, in which

- $\mathcal{I}$ is the set of agents,
- $\mathcal{S}$ is the set of states including initial state s_0,
- $\mathbf{A} = \times_i \mathcal{A}_i$ is the set of joint actions,
- $\mathcal{T} : \mathcal{S} \times \mathbf{A} \to \mathcal{S}$ is the transition dynamics,
- $\mathcal{R} : \mathcal{S} \times \mathbf{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function,
- $\Omega = \times_i \Omega_i$ is the set of observations for each agent,
- $\mathcal{O} : \mathcal{S} \times \mathbf{A} \to \Omega$ is the observation probabilities.

At each timestep t, a joint action $\boldsymbol{a} = \langle a_{1,t}, ..., a_{|\mathcal{I}|,t} \rangle$ is taken, each agent receives its corresponding local observation $\langle o_{1,t}, \ldots, o_{|\mathcal{I}|,t} \rangle \sim \mathcal{O}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and a global reward $r_t = \mathcal{R}(\mathbf{s}_t, \boldsymbol{a}_t, \mathbf{s}_{t+1})$. The joint expected discounted return is $G = \sum_{t=1}^{T} \gamma^t r_t$ where γ is a discount factor. Agent i's *action-observation history* for timestep t is defined as $\langle o_{i,1}, a_{i,1}, o_{i,2}, \ldots, a_{i,t-1}, o_{i,t} \rangle$, and we define history recursively as $h_{i,t} = \langle h_{i,t-1}, a_{i,t-1}, o_{i,t} \rangle$, likewise, a joint history is $\boldsymbol{h}_t = \langle \boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1}, \boldsymbol{o}_t \rangle$.[2] To solve a Dec-POMDP is to find a set of policies $\boldsymbol{\pi} = \langle \pi_1, \ldots, \pi_{|\mathcal{I}|} \rangle$ where $\pi_i : h_{i,t} \to a_{i,t}$ such that the joint expected discounted return G is maximized.

Notation. For notational readability, we denote the parameterized decentralized policy π_{θ_i} as π_i, on-policy G estimate Q_i^{π} as Q_i, and $Q^{\boldsymbol{\pi}}$ as Q. We denote the objective (discounted expected return) for any agent with a decentralized critic as J_d and the objective with a centralized critic as J_c. In addition, a timestep t is implied for s, $\boldsymbol{h}$, h_i, $\boldsymbol{a}$ or a_i.

3.2 Actor Critic method (AC)

Actor critic (AC) [17] is a widely used policy gradient (PG) architecture and is the basis for many single-agent policy gradient approaches. Directly optimizing a policy, policy gradient (PG) algorithms perturb policy parameters θ in the direction of the gradient of the expected return $\nabla_\theta \mathbb{E}[G_{\pi_\theta}]$, which is conveniently given by the *policy gradient theorem* [17, 41]:

$$\nabla_\theta \mathbb{E}[G_{\pi_\theta}] = \mathbb{E}_{h,a}[\nabla_\theta \log \pi_\theta(a \mid h) Q^{\pi_\theta}(h, a)] \tag{1}$$

Actor-critic (AC) methods [17] directly implement the policy gradient theorem by learning the value function Q^{π_θ}, i.e., the critic, commonly through TD learning [40]. The policy gradient for updating the policy (i.e., the actor) then follows the return estimates given by the critic (Equation 1).

3.3 Multi-Agent Actor Critic methods

We introduce three extensions of single-agent Actor Critic methods to multi-agent settings, which are highlighted in Table 1.

The first AC multi-agent extension, Joint Actor Critic (JAC) [2, 46], treats the multi-agent environment as a single-agent environment and learns in the joint observation-action space; JAC learns a centralized actor, $\boldsymbol{\pi}(\boldsymbol{a} \mid \boldsymbol{h}; \theta)$, and a centralized value function

[2] For Proposition 1 and all following results, we employ a fixed-memory history where the history only consists of the past k observations and actions. Formally, a history at timestep t with memory length k is defined as $\boldsymbol{h}_{t,k} = \langle \boldsymbol{h}_{t-1,k-1}, \boldsymbol{a}_{t-1}, \boldsymbol{o}_t \rangle$ when $k > 0$ and $t > 0$, otherwise $\varnothing$.

Method	Critic	Actor
JAC [2, 46]	Centralized	Centralized
IAC [10, 42]	Decentralized	Decentralized
IACC [10, 20]	Centralized	Decentralized

Table 1: Our Multi-agent Actor Critic Naming Scheme.

(critic), $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}; \phi)$. The policy gradient for JAC follows that of single-agent actor critic:

$$\nabla J(\theta) = \mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}[\nabla \log \boldsymbol{\pi}(\boldsymbol{a} \mid \boldsymbol{h}; \theta) Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}; \phi)]. \tag{2}$$

The second AC multi-agent extension, called Independent Actor Critic (IAC) [10, 42], learns a decentralized policy and critic $\langle \pi_i(a_i \mid h_i; \theta_i), Q_i(h_i, a_i; \phi_i) \rangle$ for each of the agents locally. At every timestep t, a local experience $\langle h_{i,t}, a_{i,t} \rangle$ is generated for agent i. The policy gradient learning for agent i is defined as

$$\nabla_{\theta_i} J_d(\theta_i) = \mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}[\nabla \log \pi_i(a_i \mid h_i; \theta_i) Q_i(h_i, a_i; \phi_i)]. \tag{3}$$

Finally, we define Independent Actor with Centralized Critic (IACC), a class of centralized critic methods where a joint value function $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}; \phi)$ is used to update each decentralized policy π_{θ_i} [2, 10].[3] Naturally, the policy gradient for decentralized policies with a centralized critic is defined as

$$\nabla_{\theta_i} J_c(\theta_i) = \mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}[\nabla \log \pi_i(a_i \mid h_i; \theta_i) Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}; \phi)]. \tag{4}$$

At any timestep, the joint expected return estimate $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}; \phi)$ is used to update the decentralized policy $\pi(a_i \mid h_i; \theta_i)$. Notice that the centralized critic $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}; \phi)$ estimates the return based on joint information (all agents' action-histories) that differs from the decentralized case in Eq. 3. In the following section, we shall show that from a local viewpoint of agent i, for each joint action-history $\langle \boldsymbol{h}_t, \boldsymbol{a}_t \rangle$, $Q^{\boldsymbol{\pi}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \phi)$ is a sample from the return distribution given local action-histories $\Pr(G_{t:T} \mid h_{i,t}, a_{i,t})$, while the decentralized critic $Q_i(h_{i,t}, a_{i,t})$ provides an expectation.

4 BIAS ANALYSIS

In this section, we prove policies have the same expected gradient whether using centralized or decentralized critics. We prove that the centralized critic provides unbiased and correct on-policy return estimates, but at the same time makes the agents suffer from the same action shadowing problem [24] seen in decentralized learning. It is reassuring that the centralized critic will not encourage a decentralized policy to pursue a joint policy that is only achievable in a centralized manner, but also calls into question the benefits of a centralized critic. We provide a theoretical proof on bias equivalence, then analyze a classic example for intuitive understanding.

4.1 Equivalence in Expected Gradient

We first show the gradient updates for IAC and IACC are the same in expectation. We assume the existence of a limiting distribution $\Pr(\boldsymbol{h})$ over fixed-length histories, analogous to a steady-state assumption. That is, we treat fixed-memory trajectories as nodes in a Markov chain for which the policies induce a stationary distribution.

Proposition 1. *Suppose that agent histories are truncated to an arbitrary but finite length. A stationary distribution on the set of all possible histories of this length is guaranteed to exist under any collection of agent policies.*

We provide formal justification for Proposition 1 in Appendix A.1 [22], necessary to derive Bellman equations in our following theoretical results. With this, we begin by establishing novel convergence results for the centralized and decentralized critics in Lemmas 1 and 2 below, which culminate in Theorem 1 regarding the expected policy gradient.[4]

Lemma 1. *Given the existence of a steady-state history distribution (Proposition 1), training of the* ***centralized*** *critic is characterized by the Bellman operator B_c which admits a unique fixed point $Q^{\boldsymbol{\pi}}(h_i, h_j, a_i, a_j)$ where $Q^{\boldsymbol{\pi}}$ is the true expected return under the joint policy $\boldsymbol{\pi}$.*

Lemma 2. *Given the existence of a steady-state history distribution (Proposition 1), training of the i-th* ***decentralized*** *critic is characterized by a Bellman operator B_d which admits a unique fixed point $\mathbb{E}_{h_j,a_j}\left[Q^{\boldsymbol{\pi}}(h_i, h_j, a_i, a_j)\right]$ where $Q^{\boldsymbol{\pi}}$ is the true expected return under the joint policy $\boldsymbol{\pi}$.*

Theorem 1. *After convergence of the critics' value functions, the expected policy gradients for the centralized actor and the decentralized actors are equal. That is,*

$$\mathbb{E}[\nabla_\theta J_c(\theta)] = \mathbb{E}[\nabla_\theta J_d(\theta)] \tag{5}$$

where J_c and J_d are the respective objective functions for the central and decentralized actors, and the expectation is taken over all joint histories and joint actions. All actors are assumed to have the same policy parameterization.

Proof sketch: We derive Bellman equations for the centralized and decentralized critics and express them as Q-function operators. We show that these operators are contraction mappings and admit fixed points at $Q^{\boldsymbol{\pi}}(h_i, h_j, a_i, a_j)$ and $\mathbb{E}_{h_j,a_j}\left[Q^{\boldsymbol{\pi}}(h_i, h_j, a_i, a_j)\right]$, respectively. These convergence results reveal that the decentralized critic becomes the marginal expectation of the centralized critic after training for an infinite amount of time. Under the total expectation over joint histories and joint actions, these fixed points are identically equal, implying that gradients computed for the centralized and decentralized actors are the same in expectation and therefore unbiased. The full proofs for Lemmas 1 and 2 and Theorem 1 are respectively provided in Appendices A.2, A.3, and A.4.

Our theory assumes that the critics are trained sufficiently to converge to their true on-policy values. This assumption often exists in the form of infinitesimal step sizes for the actors [3, 10, 17, 37, 50] for convergence arguments of AC, since the critics are on-policy return estimates and the actors need an unbiased and up-to-date critic. Although this assumption is in line with previous theoretical works, it is nevertheless unrealistic; we discuss the practical implications of relaxing this assumption in Section 6.1.

[3] For example, COMA [10] is then considered as an IACC approach with a variance reduction baseline.

[4] While it appears that we analyze the case with two agents, agent i and j, the result holds for arbitrarily many agents by letting j represent all agents except i.

4.2 Climb Game as an Intuitive Example

We use a classic matrix game as an example to intuitively highlight that IAC and IACC give the same policy gradient in expectation. The Climb Game [5], whose reward function is shown in Table 2, is a matrix game (a state-less game) in which agents are supposed to cooperatively try to achieve the highest reward of 11 by taking the joint action $\langle u_1, u_1 \rangle$, facing the risk of being punished by -30 or 0 when agents miscoordinate. It is difficult for independent learners to converge onto the optimal $\langle u_1, u_1 \rangle$ actions due to low expected return for agent 1 to take u_1 when agent 2's policy is not already favoring u_1 and vise versa.

This cooperation issue arises when some potentially good action a has low ("shadowed") on-policy values because yielding a high return depends on other agents' cooperating policies, but frequently taking action a is essential for other agents to learn to adjust their policies accordingly, creating a dilemma where agents are unwilling to frequently take the low-value action and are therefore stuck in a local optimum. In the case of the Climb Game, the value of u_1 is often shadowed, because u_1 does not produce a satisfactory return unless the other agent also takes u_1 frequently enough. This commonly occurring multi-agent local optimum is called a *shadowed equilibrium* [11, 31], a known difficulty in independent learning which usually requires an additional cooperative mechanism (e.g., some form of centralization) to overcome.

Solving the Climb Game independently with IAC, assume the agents start with uniformly random policies. In expectation, IAC's decentralized critic would estimate $Q_1(u_1)$ at $(1/3) \times 11 + (1/3) \times (-30) + 0 \approx -6.3$, making u_3 (with $Q_1(u_3) \approx 3.7$) a much more attractive action, and the same applies for agent 2. Naturally, both agents will update towards favoring u_3; continuing this path, agents would never favor u_1 and never discover the optimal value of u_1: $Q_i^*(u_1) = 11$.

In the case of an IACC method, unsurprisingly, the centralized critic can quickly learn the correct values, including the value of the optimal joint action $Q(u_1, u_1) = 11$ since there is no environmental stochasticity. However, consider at timestep t agent 1 takes optimal action u_1, the (centralized) Q value estimate used in policy gradient $\nabla Q(u_1, a_{2,t})\pi_1(u_1)$ actually depends on what action agent 2 chooses to take according to its policy $a_{2,t} \sim \pi_2$. Again assuming uniform policies, consider a rollout where action $a_{2,t}$ is sampled from the policy of agent 2; then with sufficient sampling, we expect the mean policy gradient (given by centralized critic) for updating $\pi_1(u_1)$ would be

$$
\begin{aligned}
&\mathbb{E}_{\pi_2} \nabla J_C(\theta \mid a_{1,t} = u_1) \\
&= \pi_2(u_1)\nabla Q(u_1, u_1)\pi_1(u_1) + \pi_2(u_2)\nabla Q(u_1, u_2)\pi_1(u_1) \\
&\quad + \pi_2(u_3)\nabla Q(u_1, u_3)\pi_1(u_1) \qquad (6) \\
&= (1/3)\nabla 11\pi_1(u_1) + (1/3)\nabla - 30\pi_1(u_1) + (1/3)\nabla 0\pi_1(u_1) \\
&\approx -6.3 \cdot \nabla \pi_1(u_1)
\end{aligned}
$$

That is, with probabilities $\pi_2(u_1)$, $\pi_2(u_2)$ and $\pi_2(u_3)$, the joint Q-values 11,-30 and 0 are sampled for generating the policy gradient of action 1 for agent 1, $\pi_1(u_1)$. It implies that the sample-average gradient for $\pi_1(u_1)$ is high only when agent 2 takes action u_1 frequently. If agent 2 has a uniform policy, the sample-average gradient $-6.3\nabla\pi_1(u_1)$ cannot compete with the gradient for u_3 at $1.7\nabla\pi_1(u_3)$. Therefore, in the IACC case with a centralized critic, we see the rise of an almost identical action shadowing problem we described for IAC, even though the centralized critic trained jointly and has the correct estimate of the optimal joint action.

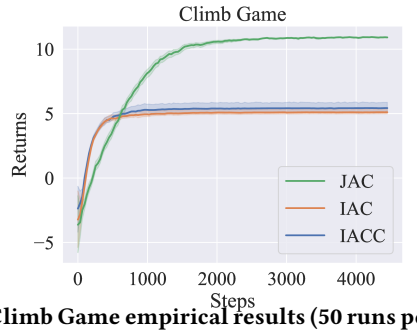


Figure 1: Climb Game empirical results (50 runs per method) showing both decentralized and centralized critic methods succumb to the *shadowed equilibrium* problem.

		agent 1		
		u_1	u_2	u_3
agent 2	u_1	11	-30	0
	u_2	-30	7	6
	u_3	0	0	5

Table 2: Return values for Climb Game [5].

Empirical evaluation on the Climb Game (shown in Figure 1) conforms to our analysis, showing both methods converge to the suboptimal solution $\langle a_3, a_3 \rangle$. At the same time, unsurprisingly, a centralized controller always gives the optimal solution $\langle a_1, a_1 \rangle$. In general, we observe that the centralized critic has the information of the optimal solution, information that is only obtainable in a centralized fashion and is valuable for agents to break out of their cooperative local optima. However, this information is unable to be effectively utilized on the individual actors to form a cooperative policy. Therefore, contrary to the common intuition, in its current form, the centralized critic is unable to foster cooperative behavior more easily than the decentralized critics.

5 VARIANCE ANALYSIS

In this section, we first show that with true on-policy value functions, the centralized critic formulation can increase policy gradient variance. More precisely, we prove that the policy gradient variance using a centralized critic is at least as large as the policy gradient variance with decentralized critics. We again assume that the critics have converged under fixed policies, thus ignoring the variance due to value function learning; we discuss the relaxation of this assumption in Section 6.

We begin by comparing the policy gradient variance between centralized and decentralized critics:

Theorem 2. *Assume that all agents have the same policy parameterization. After convergence of the value functions, the variance of the policy gradient using a centralized critic is at least as large as that of a decentralized critic along every dimension.*

Proof sketch: As with our proof of Theorem 1, Lemmas 1 and 2 show that a decentralized critic's value estimate is equal to the

marginal expectation of the central critic's Q-function after convergence. This implies that the decentralized critics have already averaged out the randomness caused by the other agents' decisions. Since each agent has the same policy parameterization, their policy gradient covariance matrices are equal up to the scale factors induced by the critics' value estimates. By Jensen's inequality, we show that the additional stochasticity of the central critic can increase (but not decrease) these scale factors compared to a decentralized critic; hence, $\mathrm{Var}(J_c(\theta)) \geq \mathrm{Var}(J_d(\theta))$ element-wise.

In the following subsections, we define and analyze this variance increase by examining its two independent sources: the "Multi-Action Variance" (MAV) induced by the other actors' policies, and the "Multi-Observation Variance" (MOV) induced by uncertainty regarding the other agents' histories from the local perspective of a single agent. We introduce these concepts along with concrete examples to illustrate how they affect learning.

5.1 Multi-Action Variance (MAV)

We discuss MAV in the fully observable case and will address the partially observable case in MOV. Intuitively, from the local perspective of agent i, when taking an action a^i at state s, MAV is the variance $\mathrm{Var}[G(s, a_i)]$ in return estimates due to the fact that teammates might take different actions according to their own stochastic policies. With decentralized critics, MAV is averaged into the value function (Lemma 2) which is an expectation incorporating teammates actions a_j; thus, at given timestep t, $Q_i(s, a_i)$ has no variance. On the other hand, a centralized critic $Q(s, a_i, a_j)$ distinguishes between all action combinations $\langle a_i, a_j \rangle$ (Lemma 1), but a_j is sampled by agent j during execution: $a_j \sim \pi_j(s)$; therefore, the value of $Q(s, a_i, a_j)$ varies during policy updates depending on a_j. We propose a simple domain to clarify this variance's cause and effect and show that a centralized critic transfers MAV directly to policy updates.

5.1.1 The Morning Game and the MAV. The Morning Game, inspired by Peshkin et al. [32], shown in Table 3, consists of two agents collaborating on making breakfast in which the most desired combination is $\langle cereal, milk \rangle$. Since there is no environmental stochasticity, a centralized critic can robustly learn all the values correctly after only a few samples. In contrast, the decentralized critics need to average over unobserved teammate actions. Take a closer look at action *cereal*: for a centralized critic, cereal with milk returns 3 and cereal with vodka returns 0; meanwhile, a decentralized critic receives stochastic targets (3 or 0) for taking action *cereal*, and only when agent 2 favors *milk*, a return of 3 comes more often, which then would make $Q_1(cereal)$ a higher estimate. Therefore, the centralized critic has a lower variance (zero in this case), and the decentralized critic has a large variance on the update target.

Often neglected is that using a centralized critic has higher variance when it comes to updating decentralized policies. Suppose agents employ uniform random policies for both IAC and IACC, in which case agent 1's local expected return for *cereal* would be $Q_1(cereal) = \pi_2(milk) \cdot 3 + \pi_2(vodka) \cdot 0 = 1.5$. Assuming converged value functions, then in IACC, a centralized critic would uniformly give either $Q(cereal, milk) = 3$ or $Q(cereal, vodka) = 0$ for $\pi_1(cereal)$ updates, and $Q(pickles, milk) = 0$ or $Q(pickles, vodka) =$ 1 for $\pi_1(pickles)$ updates. With IAC, a decentralized critic always gives $Q(cereal) = 1.5$ for $\pi_1(cereal)$, and $Q(pickle) = 0.5$ for $\pi_2(pickle)$. Obviously, under both methods, π_1 converges towards *cereal*, but the decentralized critic makes the update direction less variable and much more deterministic in favor of *cereal*.

		agent 1	
		pickles	*cereal*
agent 2	*vodka*	1	0
	milk	0	3

Table 3: Return values for the proposed Morning Game.

5.2 Multi-Observation Variance (MOV)

In the partially observable case, another source of variance in local value $\mathrm{Var}[G(h_i, a_i)]$ comes from factored observations. More concretely, for an agent in a particular local trajectory h_i, other agents' experiences $h_j \in \mathcal{H}_j$ may vary, over which the decentralized agent has to average. A decentralized critic is designed to average over this observation variance and provide and single expected value for each local trajectory $Q_i(h_i, a_i)$. The centralized critic, on the other hand, is able to distinguish each combination of trajectories $\langle h_i, h_j \rangle$, but when used for a decentralized policy at h_i, teammate history h_j can be considered to be sampled from $\Pr(h_j \mid h_i)$, and we expect the mean estimated return during the update process to be $\mathbb{E}_{h_j} Q(h_i, h_j, \boldsymbol{a})$.

We use a thought experiment[5] as an example. Consider a one-step task where two agents have binary actions and are individually rewarded r if the action matches the other's randomly given binary observations and $-r$ for a mismatch; that is, $R(h_i, h_j, a_i) = r$ when $h_j = a_i$ and $-r$ otherwise. With any policy, assuming converged value functions, a decentralized critic would estimate $Q(h_i, a_i) = 0$ with zero variance. On the other hand, a centralized critic with a global scope is able to recognize whether the current situation $\langle h_i, h_j, a_i \rangle$ would result in a return of 1 or 0, hence estimates r with probability 0.5 (when $h_j = a_i$ by definition) and $-r$ with probability 0.5 (when $h_j \neq a_i$), resulting in a variance of r^2. In this example, we see that a centralized critic produces returns estimates with more significant variance when agents have varying observations.

6 DISCUSSION

In this section, we discuss the anticipated trade-off in practice. We look at the practical aspects of a centralized critic and address the unrealistic true value function assumption. We note that, although both types of critics have the same expected gradient for policy updates, they have different amount of actual bias in practice. We also discuss how and why the way of handling variance is different for IAC and IACC. We conclude that we do not expect one method to dominate the other in terms of performance in general.

6.1 Value Function Learning

So far, in terms of theory, we have only considered value functions that are assumed to be correct. However, in practice, value functions

[5] We also propose a toy domain called Guess Game based on this thought experiment, which we elaborate and show empirical results in Appendix C.

are difficult to learn. We argue that it is generally more so with decentralized value functions.

In MARL, the on-policy value function is non-stationary since the return distributions heavily depend on the current joint policy. When policies update and change, the value function is partially obsolete and is biased towards the historical policy-induced return distribution. Bootstrapping from outdated values creates bias. Compounding bias can cause learning instability depending on the learning rate and how drastically the policy changes.

The non-stationarity applies for both types of critics since they are both on-policy estimates. However, centralized value function learning is generally better equipped in the face of non-stationarity because it has no variance in the update targets. Therefore, the bootstrapping may be more stable in the case of a centralized critic. As a result, in a cooperative environment with a moderate number of agents (as discussed later), we expect a centralized critic would learn more stably and be less biased, perhaps counteracting the effect of having larger variance in the policy gradient.

6.2 Handling MAV and MOV

Learning a decentralized policy requires reasoning about local optimality. That is, given local information h_i, agent i needs to explicitly or implicitly consider the distribution of global information: other agents' experiences and actions $\Pr(a_j, h_j \mid h_i)$, through which agent i needs to take the expectation of global action-history values. Abstractly, the process of learning via sampling over the joint space of a_j and h_j thus generates MAV and MOV for agent i's policy gradient. Interestingly, this process is inevitable but takes place in different forms: during IAC and IACC training, this averaging process is done by different entities. In IAC, those expectations are implicitly taken by the decentralized critic and produce a single expected value for the local history; this is precisely why decentralized critic learning has unstable learning targets as discussed in Section 6.1. On the other hand, in IACC, the expectation takes place directly in the policy learning. Different samples of global value estimates are used in policy updates for a local trajectory, hence the higher policy gradient variance we discussed in Section 5. Thus, for decentralized policy learning purposes, we expect decentralized critics to give estimates with more bias and less variance, and the centralized critic to give estimates with less bias and more variance. Consequently, the trade-off largely depends on the domain; as we shall see in the next section, certain domains favor a more stable policy while others favor a more accurate critic.

6.3 Scalability

Another important consideration is the scale of the task. A centralized critic's feature representation needs to scale linearly (in the best case) or exponentially (in the worse case) with the system's number of agents. In contrast, a decentralized critic's number of features can remain constant, and in homogeneous-agent systems, decentralized critics can even share parameters. Also, some environments may not require much reasoning from other agents. For example, in environments where agents' decisions rarely depend on other agents' trajectories, the gain of learning value functions jointly is likely to be minimal, and we expect decentralized critics to perform better while having better sample efficiency in those domains. We show this empirically in Section 7.4.

The impact of variance will also change as the number of agents increases. In particular, when learning stochastic policies with a centralized critic in IACC, the maximum potential variance in the policy gradient also scales with the number of agents (see Theorem 2). On the other hand, IAC's decentralized critics potentially have less stable learning targets in critic bootstrapping with increasing numbers of agents, but the policy updates still have low variance. Therefore, scalability may be an issue for both methods, and the actual performance is likely to depend on the domain, function approximation setups, and other factors. However, we expect that IAC should be a better starting point due to more stable policy updates and potentially shared parameters.

6.4 The Overall Trade-off

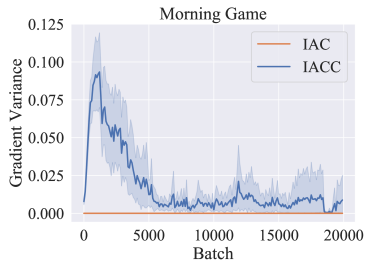
Combining our discussions in Sections 6.1 and 6.2, we conclude that whether to use critic centralization can be essentially considered a bias-variance trade-off decision. More specifically, it is a trade-off between variance in policy updates and bias in the value function: a centralized critic should have a lower bias because it will have more stable Q-values that can be updated straightforwardly when policies change, but higher variance because the policy updates need to be averaged over (potentially many) other agents. In other words, the policies trained by centralized critics avoid more-biased estimates usually produced by decentralized critics, but in return suffer more variance in the training process. The optimal choice is then largely dependent on the environment settings. Regardless, the centralized critic likely faces more severe scalability issues in not only critic learning but also in policy gradient variance. As a result, we do not expect one method will always dominate the other in terms of performance.

7 EXPERIMENTS AND ANALYSIS

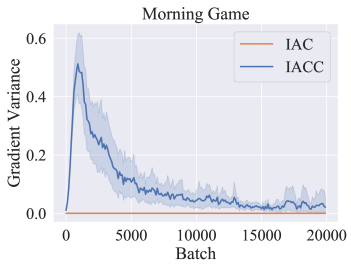
In this section, we present experimental results comparing centralized and decentralized critics. We test on a variety of popular research domains including (but not limited to) classical matrix games, the StarCraft Multi-Agent Challenge [35], the Particle Environments [25], and the MARL Environments Compilation [14]. Our hyperparameter tuning uses grid search. Each figure, if not otherwise specified, shows the aggregation of 20 runs per method.

7.1 Variance in the Morning Game

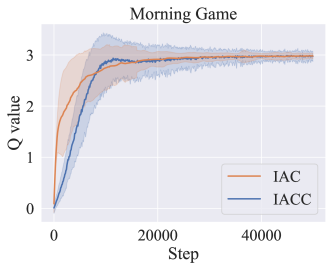
As we can see in Figures 2a and 2b, the variance per rollout in the policy gradient for both actions is zero for IAC and non-zero for IACC, validating our theoretical variance analysis. Figure 2c shows how the Q-values evolve for the optimal action a_2 in both methods. First, observe that both types of critics converged to the correct value 3, which confirms our bias analysis. Second, the two methods' Q-value variance in fact comes from different sources. For the decentralized critic, the variance comes from the critics having different biases across trials. For centralized critics, there is the additional variance that comes from incorporating other agent actions, producing a high value when a teammate chooses *milk* and a low value when a teammate chooses *vodka*.



(a) Per-rollout gradient variance for action *pickles*.

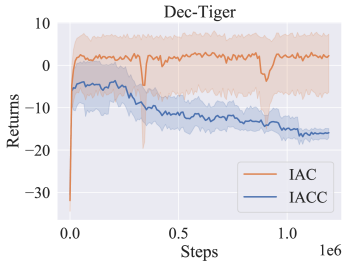


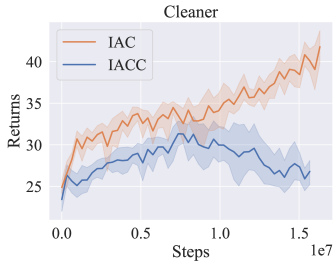
(b) Per-rollout gradient variance for action *cereal*.



(c) Q values used for updating $\pi(cereal)$; see Figure 6 for clearer illustration.

Figure 2: Gradient updates of the Morning Game, shows 200 independent trials for each method.





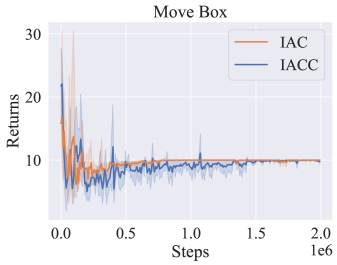


Figure 3: Performance comparison in different domains: (a) Dec-Tiger, (b) Cleaner and (c) Move Box. Dec-Tiger and Cleaner highlight instability resulting from high-variance actor updates employing a centralized critic; Move Box shows the centralized critic is not able to bias the actors towards the joint optimum (at 100).

7.2 Unstable Policies

7.2.1 Dec-Tiger. We test on the classic yet difficult Dec-Tiger domain [26], a multi-agent extension to the Tiger domain [16]. To end an episode, each agent has a high-reward action (opening a door with treasure inside) and a high-punishment action (opening a door with tiger inside). The treasure and tiger are randomly initialized in each episode, hence, a third action (*listen*) gathers noisy information regarding which of the two doors is the rewarding one. The multi-agent extension of Tiger requires two agents to open the correct door simultaneously in order to gain maximum return. Conversely, if the bad action is taken simultaneously, the agents take less punishment. Note that any fast-changing decentralized policies are less likely to coordinate the simultaneous actions with high probability, thus lowering return estimates for the critic and hindering joint policy improvement. As expected, we see in Figure 3a that IACC (with a centralized critic and higher policy gradient variance) does not perform as well as IAC due. In the end, the IACC agent learns to avoid high punishment (agents simultaneously open different doors, −100) by not gathering any information (*listen*) and opening an agreed-upon door on the first timestep. IACC gives up completely on high returns (where both agents *listen* for some timesteps and open the correct door at the same time, +20) because the unstable policies make coordinating a high return of +20 extremely unlikely.

7.2.2 Cleaner. We observe similar policy degradation in the Cleaner domain [14], a grid-world maze in which agents are rewarded for stepping onto novel locations in the maze. The optimal policy is to have the agents split up and cover as much ground as possible. The maze has two non-colliding paths (refer to Appendix D.3 for visualization) so that as soon as the agents split up, they can follow a locally greedy policy to get an optimal return. However, with a centralized critic (IACC), both agents start to take the longer path with more locations to "clean." The performance is shown in Figure 3b. When policy-shifting agents are not completely locally greedy, the issue is that they cannot "clean" enough ground in their paths. Subsequently, they discover that having both agents go for the longer path (the lower path) yields a better return, converging to a suboptimal solution. Again we see that in IACC with centralized critic, due to the high variance we discussed in Section 6.2, the safer option is favored, resulting in both agents completely ignoring the other path (performance shown in Figure 3b). Overall, we see that high variance in the policy gradient (in the case of the centralized critic) makes the policy more volatile and can result in poor coordination performance in environments that require coordinated series of actions to discover the optimal solution.

7.3 Shadowed Equilibrium

Move Box [14] is another commonly used domain, where grid-world agents are rewarded by pushing a heavy box (requiring both agents) onto any of the two destinations (see Appendix D.4 for details). The farther destination gives +100 reward while the nearer destination gives +10. Naturally, the optimal policy is for both agents to go for +100, but if *either* of the agents is unwilling to do so, this optimal option is "shadowed" and *both* agents will have to go for +10. We see in Figure 3c that both methods fall for the *shadowed equilibrium*, favoring the safe but less rewarding option.

Analogous to the Climb Game, even if the centralized critic is able to learn the optimal values due to its unbiased on-policy nature shown in Section 4.1, the on-policy return of the optimal option is extremely low due to an uncooperative teammate policy; thus, the optimal actions are rarely sampled when updating the policies. The

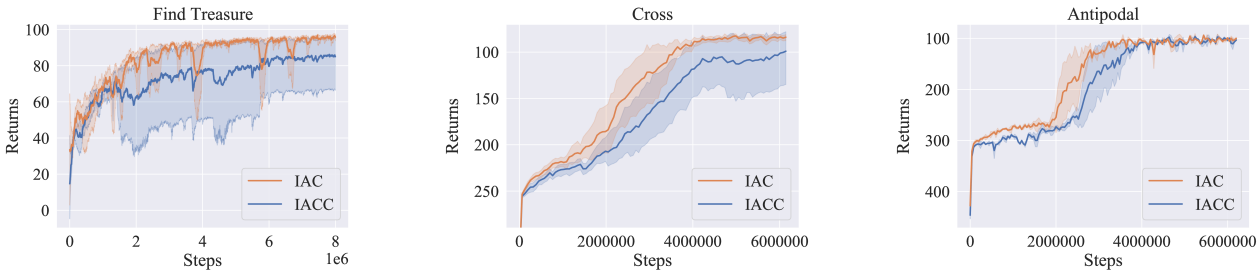


Figure 4: Performance comparison of domains where IAC does better than IACC:, Find Treasure, Cross and Antipodal.

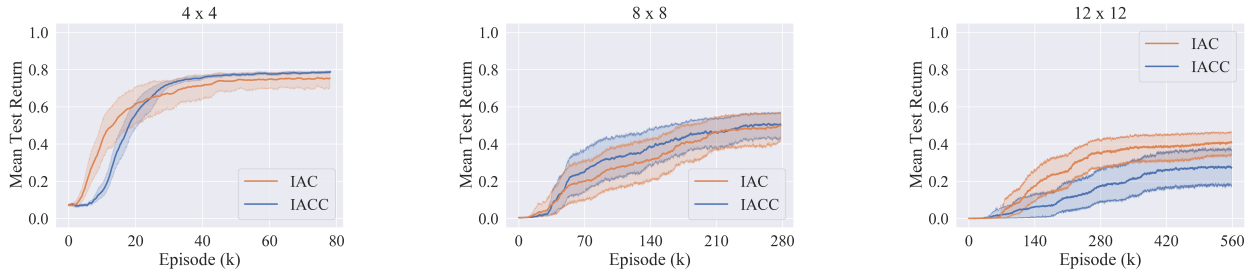


Figure 5: Performance comparison on capture target under various grid world sizes.

same applies to both agents, so the system reaches a suboptimal equilibrium, even though IACC is trained in a centralized manner.

7.4 Robustness in Performance

Through tests on Go Together [14], Merge [49], Predator and Prey [20], Capture Target [30, 47], Small Box Pushing and SMAC [35] tasks (see Appendices D, E, F, G, H, I, and J), we observe that the two types of critics perform similarly in all these domains, with IACC (with a centralized critic) being less stable in only a few other domains shown in Figures 3 and 4. Since the performance of the two critic types is similar in most results, we expect that it is due to the fact that both are unbiased asymptotically (Lemmas 1, 2 and Theorem 1). We observe that, although decentralized critics might be more biased when considering finite training, it does not affect real-world performance in a significant fashion in these domains.

In cooperative navigation domains Antipodal, Cross [25, 49], and Find Treasure [14], we observe a more pronounced performance difference among runs (Figure 4). In these cooperative navigation domains (details in Appendices D.2 and E), there are no suboptimal equilibria that trap the agents, and on most of the timesteps, the optimal action aligns with the locally greedy action. Those tasks only require agents to coordinate their actions for a few timesteps to avoid collisions. It appears that those tasks are easy to solve, but the observation space is continuous, thus causing large MOV in the gradient updates for IACC. Observe that some IACC runs struggle to reach the optimal solution robustly, while IAC robustly converges, conforming to our scalability discussion regarding large MOV. A centralized critic induces higher variance for policy updates, where the shifting policies can become a drag on the value estimates which, in turn, become a hindrance to improving the policies themselves.

The scalability issue can be better highlighted in environments where we can increase the observation space. For example, in Capture Target [21] where agents are rewarded by simultaneously catching a moving target in a grid world (details in Appendix G), by increasing the grid size from 4×4 to 12×12, we see a notable comparative drop in overall performance for IACC (Figure 5). Since an increase in observation space leads to an increase in Multi-Observation Variance (MOV) and nothing else, it indicates that here the policies of IACC do not handle MOV as well as the decentralized critics in IAC. The result might imply that, for large environments, decentralized critics scale better in the face of MOV due to the fact that they do not involve MOV in policy learning.

8 CONCLUSION

In this paper, we present an examination of critic centralization theoretically and empirically. The core takeaways are: 1) in theory, centralized and decentralized critics are the same in expectation for the purpose of updating decentralized policies; 2) in theory, centralized critics will lead to higher variance in policy updates; 3) in practice, there is a bias-variance trade-off due to potentially higher bias with limited samples and less-correct value functions with decentralized critics; 4) in practice, a decentralized critic regularly gives more robust performance since stable policy gradients appear to be more crucial than stable value functions in our domains.

Although IACC uses a centralized critic that is trained in a centralized manner, the method does not produce policies that exploit centralized knowledge. Therefore, future work on IACC may explore feasible ways of biasing the decentralized policies towards a better joint policy by exploiting the centralized information. Reducing variance in policy updates and other methods that make better use of centralized training are promising future directions.

9 ACKNOWLEDGMENTS

We thank the reviewers for their helpful feedback. We also thank Andrea Baisero and Linfeng Zhao for helpful comments and discussions. This research is supported in part by the U. S. Office of Naval Research under award number N00014-19-1-2131, Army Research Office award W911NF-20-1-0265 and an Amazon Research Award.

REFERENCES

[1] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2020. Emergent tool use from multi-agent autocurricula. In *Proceedings of the Eighth Int'l Conf. on Learning Representations*.

[2] Guillaume Bono, Jilles Steeve Dibangoye, Laëtitia Matignon, Florian Pereyron, and Olivier Simonin. 2018. Cooperative multi-agent policy gradient. In *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*. Springer, 459–476.

[3] Michael Bowling and Manuela Veloso. 2001. Convergence of gradient dynamics with a variable learning rate. In *Int'l Conf. on Machine Learning*. 27–34.

[4] Jhelum Chakravorty, Patrick Nadeem Ward, Julien Roy, Maxime Chevalier-Boisvert, Sumana Basu, Andrei Lupu, and Doina Precup. 2020. Option-critic in cooperative multi-agent systems. In *Proceedings of the 19th Int'l Conf. on Autonomous Agents and MultiAgent Systems*. 1792–1794.

[5] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998, 746-752 (1998), 2.

[6] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. Tarmac: targeted multi-agent communication. In *Int'l Conf. on Machine Learning*. 1538–1546.

[7] Christian Schroeder de Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. 2018. Deep multi-Agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709v3* (2018).

[8] Yali Du, Lei Han, Meng Fang, Tianhong Dai, Ji Liu, and Dacheng Tao. 2019. LIIR: learning individual intrinsic reward in multi-agent reinforcement learning.. In *Advances in Neural Information Processing Systems*.

[9] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*. 2137–2145.

[10] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Thirty-second AAAI conference on artificial intelligence*.

[11] Nancy Fulda and Dan Ventura. 2007. Predicting and preventing coordination problems in cooperative Q-learning systems.. In *Int'l Joint Conf.s on Artificial Intelligence*, Vol. 2007. 780–785.

[12] Shariq Iqbal and Fei Sha. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th Int'l Conf. on Machine Learning*, Vol. 97. PMLR, 2961–2970.

[13] Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. In *Advances in neural information processing systems*. 7254–7264.

[14] Shuo Jiang. 2019. Multi agent reinforcement learning environments compilation. https://github.com/Bigpig4396/Multi-Agent-Reinforcement-Learning-Environment

[15] Emilio Jorge, Mikael Kågebäck, Fredrik D Johansson, and Emil Gustavsson. 2016. Learning to play guess who? and inventing a grounded language as a consequence. *arXiv preprint arXiv:1611.03218* (2016).

[16] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[17] Vijay R Konda and John N Tsitsiklis. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*. 1008–1014.

[18] Hyun-Rok Lee and Taesik Lee. 2019. Improved cooperative multi-agent reinforcement learning algorithm augmented by mixing demonstrations from centralized policy. In *Proceedings of the 18th Int'l Conf. on Autonomous Agents and MultiAgent Systems*. 1089–1098.

[19] Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. 2019. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. *Proceedings of the AAAI Conf. on Artificial Intelligence* 33 (2019), 4213–4220. https://doi.org/10.1609/aaai.v33i01.33014213

[20] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*. 6379–6390.

[21] Xueguang Lyu and Christopher Amato. 2020. Likelihood quantile networks for coordinating multi-agent reinforcement learning. In *Proceedings of the 19th Int'l Conf. on Autonomous Agents and MultiAgent Systems*. 798–806.

[22] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. 2021. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning. *arXiv:2102.04402* (2021).

[23] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. 2019. MAVEN: multi-agent variational exploration. In *Proceedings of the Thirty-third Annual Conf. on Neural Information Processing Systems*.

[24] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. 2012. Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems. *Knowledge Engineering Review* 27, 1 (2012), 1–31.

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

[26] Ranjit Nair, Milind Tambe, Makoto Yokoo, David Pynadath, and Stacy Marsella. 2003. Taming decentralized POMDPs: towards efficient policy computation for multiagent settings. In *Int'l Joint Conf.s on Artificial Intelligence*, Vol. 3. 705–711.

[27] Frans A. Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs*. Springer.

[28] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (2008), 289–353.

[29] Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P How. 2019. Learning to teach in cooperative multiagent reinforcement learning. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, Vol. 33. 6128–6136.

[30] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. 2017. Deep decentralized multi-task multi-Agent reinforcement learning under partial observability. In *Int'l Conf. on Machine Learning*. 2681–2690.

[31] Liviu Panait, Karl Tuyls, and Sean Luke. 2008. Theoretical advantages of lenient learners: An evolutionary game theoretic perspective. *Journal of Machine Learning Research* 9, Mar (2008), 423–457.

[32] Leonid Peshkin, Kee-Eung Kim, Nicolas Meuleau, and Leslie Pack Kaelbling. 2000. Learning to cooperate via policy search. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. 489–496.

[33] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. 2018. Weighted QMIX: expanding monotonic value function factorisation. *arXiv preprint arXiv:2006.10800v1* (2018).

[34] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the Thirty-Fifth Int'l Conf. on Machine Learning*.

[35] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft multi-agent challenge. *CoRR* abs/1902.04043 (2019).

[36] David Simões, Nuno Lau, and Luís Paulo Reis. 2020. Multi-agent actor centralized-critic with communication. *Neurocomputing* (2020).

[37] Satinder P Singh, Michael J Kearns, and Yishay Mansour. 2000. Nash convergence of gradient dynamics in general-sum games. In *UAI*. 541–548.

[38] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. 2019. QTRAN: learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th Int'l Conf. on Machine Learning*, Vol. 97. PMLR, 5887–5896.

[39] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2018. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th Int'l Conf. on Autonomous Agents and MultiAgent Systems*. 2085–2087.

[40] Richard S Sutton. 1985. Temporal credit assignment in reinforcement learning. (1985).

[41] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*. 1057–1063.

[42] Ming Tan. 1993. Multi-agent reinforcement learning: independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*. 330–337.

[43] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020. Shapley Q-value: a local reward approach to solve global reward games. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, Vol. 34. 7285–7292.

[44] Tonghan Wang, Heng Dong, and Chongjie Zhang Victor Lesser. 2020. ROMA: multi-agent reinforcement learning with emergent roles. In *Proceedings of the Thirty-Seventh Int'l Conf. on Machine Learning*.

[45] Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. 2020. Learning nearly decomposable value functions via communication minimization. *Int'l Conf. on Learning Representations* (2020).

[46] Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. 2019. Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas. In *Proceedings of the First Int'l Conf. on Distributed Artificial Intelligence*. 1–7.

[47] Yuchen Xiao, Joshua Hoffman, and Christopher Amato. 2019. Macro-action-based deep multi-agent reinforcement learning. In *3rd Annual Conf. on Robot Learning*.

[48] Yuchen Xiao, Joshua Hoffman, Tian Xia, and Christopher Amato. 2020. Learning multi-robot decentralized macro-action-based policies via a centralized Q-net. In *Proceedings of the Int'l Conf. on Robotics and Automation*.

[49] Jiachen Yang, Alireza Nakhaei, David Isele, Kikuo Fujimura, and Hongyuan Zha. 2019. CM3: cooperative multi-goal multi-stage multi-agent reinforcement learning. In *Int'l Conf. on Learning Representations*.

[50] Chongjie Zhang and Victor Lesser. 2010. Multi-agent learning with policy prediction. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, Vol. 24.