Journal of Artificial Intelligence Research 77 (2023) 295-354 Submitted 11/2022; published 05/2023

On Centralized Critics in Multi-Agent Reinforcement Learning

Xueguang Lyu LU.XUE@NORTHEASTERN.EDU
Andrea Baisero BAISERO.A@NORTHEASTERN.EDU
Yuchen Xiao XIAO.YUCH@NORTHEASTERN.EDU
Brett Daley DALEY.BR@NORTHEASTERN.EDU
Christopher Amato C.AMATO@NORTHEASTERN.EDU
Northeastern University, Khoury College of Computer Sciences,
360 Huntington Avenue, Boston, MA 02115 USA

Abstract

Centralized Training for Decentralized Execution, where agents are trained offline in a centralized fashion and execute online in a decentralized manner, has become a popular approach in Multi-Agent Reinforcement Learning (MARL). In particular, it has become popular to develop actor-critic methods that train decentralized actors with a centralized critic where the centralized critic is allowed access global information of the entire system, including the true system state. Such centralized critics are possible given offline information and are not used for online execution. While these methods perform well in a number of domains and have become a de facto standard in MARL, using a centralized critic in this context has yet to be sufficiently analyzed theoretically or empirically. In this paper, we therefore formally analyze centralized and decentralized critic approaches, and analyze the effect of using state-based critics in partially observable environments. We derive theories contrary to the common intuition: critic centralization is not strictly beneficial, and using state values can be harmful. We further prove that, in particular, state-based critics can introduce unexpected bias and variance compared to history-based critics. Finally, we demonstrate how the theory applies in practice by comparing different forms of critics on a wide range of common multi-agent benchmarks. The experiments show practical issues such as the difficulty of representation learning with partial observability, which highlights why the theoretical problems are often overlooked in the literature.

1. Introduction

Centralized Training for Decentralized Execution (CTDE) (Oliehoek, Spaan, & Vlassis, 2008), where agents are trained offline in a centralized manner but execute in a decentralized manner with only local information, has been widely adopted in multi-agent reinforcement learning (MARL). Compared to independent learning, CTDE has great potential for more stable and optimal learning since agents can coordinate offline on how they will behave online. Actor-Critic (AC) methods are popular for CTDE because a centralized critic can be used to train decentralized actors, exploiting the centralized training paradigm; since the critic is only needed to train the actors, it can be discarded once the actors are fully trained without hindering decentralized execution. Because the centralized critic is trained offline in a simulator, it can be trained on the joint observations from all the agents as well as the system state. Using the state is intuitively considered desirable as it is often more concise than the history and provides ground-truth information. This technique of exploiting the system state has become popular after the pioneering centralized critic works of COMA (Foerster, Assael,

De Freitas, & Whiteson, 2016) and MADDPG (Lowe, Wu, Tamar, Harb, Pieter Abbeel, & Mordatch, 2017).

The statements made for the effects of centralized critics are almost entirely positive. For instance, MADDPG (Lowe et al., 2017) notes that it eases learning and helps to learn coordinated behaviors. Later works list similar sentiments, suspecting that critic centralization improves performance (Lee & Lee, 2019), reduces variance (Das, Gervet, Romoff, Batra, Parikh, Rabbat, & Pineau, 2019), stabilizes training (Li, Wu, Cui, Dong, Fang, & Russell, 2019) and is more robust (Simões, Lau, & Reis, 2020). It seems reasonable to make these assumptions because training a centralized value function would solve the cooperation issues (e.g., action shadowing (Claus & Boutilier, 1998)) for a centralized policy, and result in better convergence properties. However, as we will show, a centralized critic does not have the same effect (in solving those problems) on a set of decentralized policies.

The exploitation of the system state is considered one of the significant advantages of the CTDE paradigm. Using the state has also become a selling point for centralized critics (Foerster et al., 2016) since state value functions are usually easier to learn than observation-history value functions. While many works use centralized critics, they often focus on other issues without carefully examining the foundation of the critic centralization techniques that they employ, e.g., improving credit assignment (Wang, Zhang, Kim, & Gu, 2020a; Du, Han, Fang, Dai, Liu, & Tao, 2019), exploration (Zhou, Liu, Sui, Li, & Chung, 2020), emergent tool use (Baker, Kanitscheider, Markov, Wu, Powell, McGrew, & Mordatch, 2020), etc. However, even though centralized critics have become a standard mechanism in recent works, they still lack a thorough theoretical and empirical analysis. In this paper, we give a comprehensive investigation of these claims and show that most gains with critic centralization are questionable, as they often entail hidden trade-offs, both theoretically and empirically.

This paper fills this analytical gap by providing an analysis of critics conditioned on joint observations as well as the system state; we show that the common intuitions stated in most recent works are usually unsound for both cases. In particular,

(1) we show that centralized critics are not theoretically beneficial compared to decentralized critics,

(2) we show that state-based critics may result in bias, making them theoretically inferior to history-based critics,

(3) we show that centralized critics (both history and state-based) result in higher variance,

(4) we advise the use of history-state critics, which use both history and state information as a method to incorporate state without introducing bias, and

(5) we provide an extensive empirical analysis showing the trade-offs of the different critic types.

In our analysis, we prove that critic centralization does not theoretically improve cooperation compared to decentralized critics from a policy learning perspective, even though the values themselves may be easier to learn. For history-based critics, we show in theory that centralized and decentralized critics have the same expected gradient. This implies that

the centralized critic, like decentralized critics, can be used to train decentralized policies in an unbiased way. Yet on the flip side, it also implies that the centralized critic cannot ameliorate the cooperation issues seen with decentralized learners. For state-based critics, we show that when using only state information, with the commonly seen state-based centralized critic, they may incur unbounded bias in the policy gradient compared to their provably correct history-based counterparts. The resulting bias voids any asymptotic convergence properties. Practically, the bias may hinder learning a reasonable policy in many domains. We detail formally and intuitively where the bias originates while analyzing its relationship to the environmental observation model, which is highly related to the bias. Based on our theory and empirical studies, we suggest not using state-based critics in partially observable environments that require active information gathering.

We also compare the *variance* of the policy gradient for different types of critics, where, again, the theoretical result counters common intuition: we show theoretically that the centralized critic adds higher variance to the policy gradient. We also show that even when unbiased, using a state-conditioned critic further exacerbates the policy gradient variance issue. We note that the stability of learning policies from a critic should be a consideration separate from the stability of learning value functions (the critics themselves). We show that in practice, factoring in the effect of value function learning, we face a bias-variance trade-off. We also recommend the usage of a history-state-based critic (Baisero & Amato, 2022), although providing policy gradients with even higher variances in theory, they are unbiased and it is usually a favorable trade-off in terms of empirical performance.

Finally, through toy examples and larger empirical studies, we show that using centralized critics can be, in many tasks, harmful to the overall performance. We compare decentralized history-based critics, centralized history-based critics, centralized state-based critics, and (centralized) history-state-based critics. We also highlight the deficiencies of popular benchmarks, in that they often lack partial observability. Our experiments consist of a wide range of popular benchmarks, but we do not find significant performance gaps on most tasks. We also report performance degradation in environments preferring stable policies, in which decentralized critics (with minimal theoretical policy gradient variance) outperform others. In addition, we find that less partially observable environments are not sensitive to the biases caused by the state-based critics; we point out that those environments are widely used and discuss how to identify those tasks. Overall, critic selection should be a conscious decision dependent on the task. We give general practical advice on how to make trade-offs based on different types of environments.

This paper combines our work on analyzing centralized history-based critics (Lyu, Xiao, Daley, & Amato, 2021) and centralized state-based critics (Lyu, Baisero, Xiao, & Amato, 2022). We unify previous works' assumptions and mathematical notations and propose discounted visitation probabilities for properly analyzing actor-critic algorithms in cooperative multi-agent reinforcement learning, which further formalizes the theoretical results given in previous works while moving our assumptions closer to standard practices. We also emphasize the critical consideration that some tasks are not readily available for centralized training (e.g. when a simulator is not available). In those cases where centralized training is cumbersome or impossible, decentralized critics should be the default choice. In addition, we provide bias and variance analysis for the recommended alternative, the history-state-based critic.

2. Related Work

The CTDE training paradigm is often used in a number of recent deep MARL approaches. Value function-based CTDE approaches such as QMIX and many others focus on how centrally learned value functions can be reasonably decoupled into decentralized ones, and have shown promising results (Son, Kim, Kang, Hostallero, & Yi, 2019; Mahajan, Rashid, Samvelyan, & Whiteson, 2019; Wang, Dong, & Victor Lesser, 2020b; Rashid, Farquhar, Peng, & Whiteson, 2020; Peng, Rashid, Schroeder de Witt, Kamienny, Torr, Böhmer, & Whiteson, 2021; de Witt, Peng, Kamienny, Torr, Böhmer, & Whiteson, 2020; Xiao, Hoffman, Xia, & Amato, 2020; Wang, Wang, Zheng, & Zhang, 2020c; Rashid, Samvelyan, de Witt, Farquhar, Foerster, & Whiteson, 2018; Sunehag, Lever, Gruslys, Czarnecki, Zambaldi, Jaderberg, Lanctot, Sonnerat, Leibo, Tuyls, et al., 2018). On the other hand, CTDE policy gradient methods are almost entirely based on centralized critics.

One of the first methods featuring a centralized critic was COMA (Foerster, Farquhar, Afouras, Nardelli, & Whiteson, 2018). COMA adopted a centralized critic with a counterfactual baseline; note that the critic is state-based, which became the standard in many later approaches. Regarding convergence properties, COMA claims that the overall effect of a centralized critic on the decentralized policy gradient may be reduced to a single-agent actor-critic approach, which ensures convergence under similar assumptions (Konda & Tsitsiklis, 2000); however, the assumption only holds in fully-observable environments and is incorrect for partially observable environments. In this paper, we clarify and expand on the theory of centralized critics by developing convergence properties and bias/variance analysis for centralized and decentralized critics (and their respective policies). Due to their theoretical properties, we provide separate discussions for state-based and history-based critics.

Concurrently with COMA, MADDPG (Lowe et al., 2017) proposed to use a dedicated centralized critic for each agent in semi-competitive domains, demonstrating compelling empirical results in continuous action environments.

Many other agents extend the ideas of COMA and MADDPG. We discuss some of them but many more have been developed. M3DDPG (Li et al., 2019) 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 (Wang et al., 2020a) 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 (Das et al., 2019) and ATOC (Jiang & Lu, 2018). There are also efforts utilizing an attention mechanism addressing scalability problems in MAAC (Iqbal & Sha, 2019). Also, teacher-student style transfer learning LeCTR (Omidshafiei, Kim, Liu, Tesauro, Riemer, Amato, Campbell, & How, 2019) builds on top of centralized critics, which does not assume expert teachers. Other work includes multi-agent credit assignment and exploration in LIIR and LICA (Du et al., 2019; Zhou et al., 2020), goal-conditioned policies with CM3 (Yang, Nakhaei, Isele, Fujimura, & Zha, 2020), and for temporally abstracted policies (Chakravorty, Ward, Roy, Chevalier-Boisvert, Basu, Lupu, & Precup, 2020). Extensive tests based on a centralized critic in a more realistic environment using self-play for hide-and-seek (Baker et al., 2020) have demonstrated impressive results showing emergent tool use. Note that impressive results also use a state-based critic, which is a common practice and used in works

such as SQDDPG (Wang et al., 2020a), LIIR (Du et al., 2019), LICA (Zhou et al., 2020), VDAC-mix (Su, Adams, & Beling, 2021), DOP (Wang, Han, Wang, Dong, & Zhang, 2021) and MACKRL (Schroeder de Witt, Foerster, Farquhar, Torr, Boehmer, & Whiteson, 2019). As mentioned above, these state-of-the-art works use centralized critics, but they do not specifically focus on 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 introduce various forms of value functions and formalize a set of commonly accepted on-policy history (and state) distributions. We also introduce multi-agent actor-critic methods in which the value functions, both centralized and decentralized, are approximated by critic models. In the coming definitions and the rest of this document, we use $\Delta\mathcal{X}$ to denote the set of probability distributions over a set $\mathcal{X}$.

3.1 Dec-POMDPs

Decentralized partially observable Markov decision processes (Dec-POMDPs) (Oliehoek & Amato, 2016) are multi-agent cooperative sequential decision making problems. A Dec-POMDP is a tuple $\langle \mathcal{I}, \mathcal{S}, \boldsymbol{A}, \boldsymbol{\Omega}, \mathcal{T}, \mathcal{O}, R, \gamma \rangle$, composed of a set of agents $\mathcal{I}$, a state space $\mathcal{S}$, with initial state $s_0 \in \mathcal{S}$, a joint action space $\boldsymbol{A} \doteq \bigtimes_{i \in \mathcal{I}} \mathcal{A}_i$, one per agent, a joint observation space $\boldsymbol{\Omega} \doteq \bigtimes_{i \in \mathcal{I}} \Omega_i$, one per agent, a stochastic state transition function $\mathcal{T}\colon \mathcal{S} \times \boldsymbol{A} \to \Delta\mathcal{S}$ that determines state transitions $\Pr(s' \mid s, \boldsymbol{a})$, a stochastic joint observation function $\mathcal{O}\colon \boldsymbol{A} \times \mathcal{S} \to \Delta\boldsymbol{\Omega}$ that determines observation emissions $\Pr(\boldsymbol{o} \mid \boldsymbol{a}, s)$, a joint reward function $R\colon \mathcal{S} \times \boldsymbol{A} \to \mathbb{R}$ that is shared by all agents, and a discount factor $\gamma \in [0, 1)$.

Control in Dec-POMDPs is performed by a set of decentralized agent policies $\boldsymbol{\pi} = \langle \pi_1, \ldots, \pi_{|\mathcal{I}|} \rangle$, each representing a (stochastic) mapping from each agents' individual action-observation history to its next action, $\pi_i\colon \mathcal{H}_i \to \Delta\mathcal{A}_i$, where $\mathcal{H}_i$ is the set of *action-observation* histories for agent i. For instance, agent i's history at timestep t is the sequence of all previous actions and observations $h_{i,t} = \langle o_{i,0}, a_{i,0}, o_{i,1}, \ldots, a_{i,t-1}, o_{i,t} \rangle$. A *joint* history is the set of all agent histories $\boldsymbol{h}_t = \langle h_{1,t}, \ldots, h_{|\mathcal{I}|,t} \rangle$. The set of actions chosen by each agent forms the *joint* action $\boldsymbol{a}_t = \langle a_{1,t}, ..., a_{|\mathcal{I}|,t} \rangle$. As feedback from the system, a scalar reward $R(s_t, \boldsymbol{a}_t)$ is shared by all agents, and each agent receives a local observation $\langle o_{1,t}, \ldots, o_{|\mathcal{I}|,t} \rangle \sim \mathcal{O}(\boldsymbol{a}_{t-1}, s_t)$.

The objective of all agents is to maximize the total performance, i.e., the expected discounted sum of future rewards $J \doteq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$.

3.2 Discounted Visitations: Counts and Distributions

The pairing of a Dec-POMDP with a set of agent policies $\boldsymbol{\pi}$ fully determines the (stochastic) behavior of the system, i.e., the marginal, joint, and conditional probability of the defined random variables (e.g., states and histories). However, before defining core RL concepts like value functions and describing policy gradient variants, it is helpful to define a set of functions related to the likelihood of occurrences of certain states or histories.

3.2.1 Discounted Visitation Counts

First, we define the *discounted visitation counts* function η, which represents a discounted notion of the expected number of times that the system variables take some given value. In the case of the discounted state visitations (also defined in a different but equivalent form for single-agent fully observable control in (Sutton & Barto, 2018)), we define

$$\eta(s) \doteq \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s)\,. \tag{1}$$

Note that η is formally a function of the joint policies $\boldsymbol{\pi}$, and should technically be denoted as $\eta^{\boldsymbol{\pi}}$; however, given the lack of ambiguity, we omit the suffix to simplify notation. The discount factor γ guarantees that η is well-defined and finite for any Dec-POMDP, even those that never terminate or keep visiting the same states ad infinitum—hence the importance of this *discounted* notion of visitations.

We will also use η for the number of times a (joint) history or (joint) history-state pairs are visited. We use the same overloaded symbol η due to similarity with the state visitation counts; the distinction is immediately clear from the inputs and context. In the case of history visitations and history-state visitations, we define

$$\eta(\boldsymbol{h}) \doteq \sum_{t=0}^{\infty} \gamma^t \Pr(\boldsymbol{H}_t = \boldsymbol{h})\,, \qquad \eta(\boldsymbol{h}, s) \doteq \sum_{t=0}^{\infty} \gamma^t \Pr(\boldsymbol{H}_t = \boldsymbol{h}, S_t = s)\,, \tag{2}$$

We note some relevant properties that either relate or are shared by $\eta(s)$, $\eta(\boldsymbol{h})$, and $\eta(\boldsymbol{h})$. First, we note that $\eta(s) = \sum_{\boldsymbol{h}} \eta(\boldsymbol{h}, s)$ and $\eta(\boldsymbol{h}) = \sum_{s} \eta(\boldsymbol{h}, s)$. Then, we note that all of these discounted counts ultimately add up to the same value, as a direct consequence of the geometric series component based on γ, i.e., $\sum_s \eta(s) = \sum_{\boldsymbol{h}} \eta(\boldsymbol{h}) = \sum_{\boldsymbol{h},s} \eta(\boldsymbol{h}, s) = (1-\gamma)^{-1}$. In the case of $\eta(\boldsymbol{h})$ and $\eta(\boldsymbol{h}, s)$, we also note that the sum over timesteps t is technically superfluous, since there is only one timestep where the joint history $\boldsymbol{h}$ can possibly be collectively seen by the agents; therefore, $\eta(\boldsymbol{h})$ and $\eta(\boldsymbol{h}, s)$ are equivalently written as

$$\eta(\boldsymbol{h}) = \gamma^t \Pr(\boldsymbol{H}_t = \boldsymbol{h})\big|_{t=|\boldsymbol{h}|}\,, \qquad \eta(\boldsymbol{h}, s) = \gamma^t \Pr(\boldsymbol{H}_t = \boldsymbol{h}, S_t = s)\big|_{t=|\boldsymbol{h}|}\,. \tag{3}$$

Finally, we further consider visitation variants that include action counts,

$$\eta(s, \boldsymbol{a}) \doteq \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s, \boldsymbol{A}_t = \boldsymbol{a})\,, \tag{4}$$

$$\eta(\boldsymbol{h}, \boldsymbol{a}) \doteq \sum_{t=0}^{\infty} \gamma^t \Pr(\boldsymbol{H}_t = \boldsymbol{h}, \boldsymbol{A}_t = \boldsymbol{a})\,, \tag{5}$$

$$\eta(\boldsymbol{h}, s, \boldsymbol{a}) \doteq \sum_{t=0}^{\infty} \gamma^t \Pr(\boldsymbol{H}_t = \boldsymbol{h}, S_t = s, \boldsymbol{A}_t = \boldsymbol{a})\,, \tag{6}$$

which share similar properties to their action-less counterparts. Importantly, counts $\eta(\boldsymbol{h}, \boldsymbol{a})$ and $\eta(\boldsymbol{h}, s, \boldsymbol{a})$ are respectively related to $\eta(\boldsymbol{h})$ and $\eta(\boldsymbol{h}, s)$ by the policy, according to $\eta(\boldsymbol{h}, \boldsymbol{a}) = \eta(\boldsymbol{h})\boldsymbol{\pi}(\boldsymbol{a}; \boldsymbol{h})$ and $\eta(\boldsymbol{h}, s, \boldsymbol{a}) = \eta(\boldsymbol{h}, s)\boldsymbol{\pi}(\boldsymbol{a}; \boldsymbol{h})$.

3.2.2 Discounted Visitation Probabilities

Next, we define the *discounted visitation probability* functions ρ as normalized versions of the corresponding counts η. Given that η adds up to $(1-\gamma)^{-1}$, this results in

$$\begin{aligned} \rho(s) &\doteq (1-\gamma)\eta(s)\,, & \rho(\boldsymbol{h}) &\doteq (1-\gamma)\eta(\boldsymbol{h})\,, & \rho(\boldsymbol{h},s) &\doteq (1-\gamma)\eta(\boldsymbol{h},s)\,, \\ \rho(s,\boldsymbol{a}) &\doteq (1-\gamma)\eta(s,\boldsymbol{a})\,, & \rho(\boldsymbol{h},\boldsymbol{a}) &\doteq (1-\gamma)\eta(\boldsymbol{h},\boldsymbol{a})\,, & \rho(\boldsymbol{h},s,\boldsymbol{a}) &\doteq (1-\gamma)\eta(\boldsymbol{h},s,\boldsymbol{a})\,. \end{aligned} \tag{7}$$

Each ρ now represents a probability distribution over the space of its inputs. Further, common marginalization properties hold, e.g., $\rho(s) = \sum_{\boldsymbol{h}} \rho(\boldsymbol{h},s)$, and $\rho(\boldsymbol{h}) = \sum_{s} \rho(\boldsymbol{h},s)$. We can also further overload ρ to encompass a notion of conditional probability, e.g.,

$$\begin{aligned} \rho(s \mid \boldsymbol{h}) &\doteq \frac{\eta(\boldsymbol{h},s)}{\eta(\boldsymbol{h})}\,, & \rho(\boldsymbol{h} \mid s) &\doteq \frac{\eta(\boldsymbol{h},s)}{\eta(s)}\,, \\ \rho(\boldsymbol{a} \mid \boldsymbol{h}) &\doteq \frac{\eta(\boldsymbol{h},\boldsymbol{a})}{\eta(\boldsymbol{h})}\,, & \rho(\boldsymbol{a} \mid s) &\doteq \frac{\eta(s,\boldsymbol{a})}{\eta(s)}\,, \end{aligned} \tag{8}$$

for which common conditional properties also hold, e.g., $\rho(s \mid \boldsymbol{h}) = \rho(\boldsymbol{h},s) \,/\, \rho(\boldsymbol{h})$, $\rho(\boldsymbol{h} \mid s) = \rho(\boldsymbol{h},s) \,/\, \rho(s)$, $\rho(\boldsymbol{a} \mid \boldsymbol{h}) = \rho(\boldsymbol{h},\boldsymbol{a}) \,/\, \rho(\boldsymbol{h})$, and $\rho(\boldsymbol{a} \mid s) = \rho(s,\boldsymbol{a}) \,/\, \rho(s)$. We also note that $\rho(\boldsymbol{a} \mid \boldsymbol{h}) = \rho(\boldsymbol{a} \mid \boldsymbol{h},s) = \boldsymbol{\pi}(\boldsymbol{a};\boldsymbol{h})$. Finally, we note that some of these ρ-function outputs are equivalent to direct probabilities induced by the Dec-POMDP's graphical model. Most notably, $\rho(s \mid \boldsymbol{h})$ is equivalent to the conditional state probability given the joint history $\Pr(s \mid \boldsymbol{h})$. For others, however, there is no such corresponding probability, e.g., $\rho(\boldsymbol{h} \mid s)$ is mathematically well-defined, while $\Pr(\boldsymbol{h} \mid s)$ is ill-defined without assuming a particular timestep for $\boldsymbol{h}$ (Baisero & Amato, 2022). We use ρ to primarily simplify the notation associated with policy gradients (Section 3.4), although we will also exploit the conditional-ρ functions to resolve a formal issue that appears in our definition of a specific form of centralized value function (Section 3.3.3).

3.3 Value Functions

In this section, we formally define various types of value functions that are used in different forms of multi-agent policy gradient: the joint history value function $Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})$, the individual-history value function $Q_i^{\boldsymbol{\pi}}(h,a)$, the state value function $Q^{\boldsymbol{\pi}}(s,\boldsymbol{a})$, and the joint history-state value function $Q^{\boldsymbol{\pi}}(\boldsymbol{h},s,\boldsymbol{a})$. These value functions all represent some notion of expected (discounted) performance obtained by the entire team of agents, that is given and indicated as a suffix (even for the individual history case $Q_i^{\boldsymbol{\pi}}$), and they differ exclusively in terms of the information that is available to determine the expected team performance.

3.3.1 Joint History Value Function $Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})$

The joint history value function $Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})$ is a form of centralized value function, and the unique solution to the following *joint history* Bellman equality,

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a}) = R(\boldsymbol{h},\boldsymbol{a}) + \gamma\, \mathbb{E}_{o|\boldsymbol{h},\boldsymbol{a}} \left[\sum_{\boldsymbol{a}'} \boldsymbol{\pi}(\boldsymbol{a}';\boldsymbol{h}\boldsymbol{a}\boldsymbol{o}) Q^{\boldsymbol{\pi}}(\boldsymbol{h}\boldsymbol{a}\boldsymbol{o},\boldsymbol{a}') \right], \tag{9}$$

where $R(\boldsymbol{h},\boldsymbol{a}) \doteq \mathbb{E}_{s|\boldsymbol{h}}\left[R(s,\boldsymbol{a})\right]$ is the *joint history* reward function. $Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})$ is the expected long-term performance of the team of agents when each individual agent policy π_i has observed the individual history h_i and has opted to perform a first action a_i.

3.3.2 Individual History Value Function $Q_i^\pi(h, a)$

The individual history value function $Q_i^\pi(h, a)$ is a form of decentralized value function from the singular perspective of the i-th agent, and the unique solution to the following *individual history* Bellman equality,

$$Q_i^\pi(h, a) = R_i^\pi(h, a) + \gamma \mathbb{E}_{o|h,a}\left[\sum_{a'} \pi_i(a'; hao) Q_i^\pi(hao, a')\right], \tag{10}$$

where $R_i^\pi(h, a) \doteq \mathbb{E}_{s, a_{\setminus i}|h_i=h}[R(s, \boldsymbol{a})]$ is the *individual history* reward function, which integrates out the behavior of all other agents and the resulting state. $Q_i^\pi(h, a)$ is the expected long-term performance of the team of agents when the i-th agent policy π_i has observed the individual history h and has opted to perform a first action a.

3.3.3 State Value Function $Q^\pi(s, \boldsymbol{a})$

The state value function $Q^\pi(s, \boldsymbol{a})$ is a form of centralized value function that *attempts* to measure the expected long-term performance of the team of agents when the system state happens to be s, and each individual agent policy π_i has opted to perform a first action a_i. A straightforward (but naïve, as we will see) formalization of this notion is based on defining the state value function as the unique solution to the following *state* Bellman equality,

$$Q^\pi(s, \boldsymbol{a}) = R(s, \boldsymbol{a}) + \gamma \mathbb{E}_{s'|s,\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'} \Pr(\boldsymbol{a}' \mid s') Q^\pi(s', \boldsymbol{a}')\right]. \tag{11}$$

Although this notion of state value seems reasonable at the surface level, it suffers from a subtle formality issue that causes $\Pr(\boldsymbol{a} \mid s)$, and consequently $Q^\pi(s, \boldsymbol{a})$ itself, to be not guaranteeably well-defined for generic control problems and teams of agents; this is an issue intrinsic to partial observability that was already analyzed for the single-agent control case in (Baisero & Amato, 2022), and for the multi-agent control case in (Lyu et al., 2022).

To keep this introductory section brief and compact, we redirect a more thorough discussion on the issues with $\Pr(\boldsymbol{a} \mid s)$ and Equation (11) to Appendix A. Broadly speaking, the issue is related to the fact that $\Pr(\boldsymbol{a} \mid s)$ denotes a time-invariant relationship between variables that is conventionally time-variant, and is therefore undefined when a time index is not available. In fact, we note that there is no issue with a *timed* variant of the state value function $Q_t^\pi(s, \boldsymbol{a})$ defined as the solution to the following Bellman equality

$$Q_t^\pi(s, \boldsymbol{a}) = R(s, \boldsymbol{a}) + \gamma \mathbb{E}_{s'|s,\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'} \Pr(\boldsymbol{A}_{t+1} = \boldsymbol{a}' \mid S_{t+1} = s') Q_{t+1}^\pi(s', \boldsymbol{a}')\right]. \tag{12}$$

However, employing timed value functions is an unsatisfactory solution as they do not generalize well across different (and potentially many) timesteps, and is not a common practice in mainstream RL. We thus consider the following untimed alternative definition for state values $Q^\pi(s, \boldsymbol{a})$ as the unique solution to the following *state* Bellman equality,

$$Q^\pi(s, \boldsymbol{a}) = R(s, \boldsymbol{a}) + \gamma \mathbb{E}_{s'|s,\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'} \rho(\boldsymbol{a}' \mid s') Q^\pi(s', \boldsymbol{a}')\right]. \tag{13}$$

where $\Pr(\boldsymbol{a}' \mid \boldsymbol{s}')$ has been replaced with the discounted visitation conditional probability $\rho(\boldsymbol{a}' \mid \boldsymbol{s}')$ defined in Section 3.2.2. We note that $\rho(\boldsymbol{a}' \mid \boldsymbol{s}')$ is inherently defined in terms of geometric series of timed probabilities, and therefore does not suffer from the same issue as $\Pr(\boldsymbol{a}' \mid \boldsymbol{s}')$. For more details on why we employ ρ rather than other possible notions of the conditional relationship between $\boldsymbol{s}$ and $\boldsymbol{a}$, see Appendix A. Finally, we also note that $\rho(\boldsymbol{a} \mid \boldsymbol{s}) = \sum_{\boldsymbol{h}} \rho(\boldsymbol{h} \mid \boldsymbol{s})\boldsymbol{\pi}(\boldsymbol{a}; \boldsymbol{h})$ satisfies the sum-product rule, which solidifies an intuitive understanding of ρ as a reasonable concrete substitute for the concept of $\Pr(\boldsymbol{a} \mid \boldsymbol{s})$.

3.3.4 Joint History-State Value Function $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{a})$

The joint history-state value function $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{a})$ is a form of centralized value function. $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{a})$ is the expected long-term performance of the team of agents when the unobserved environment state happens to be $\boldsymbol{s}$, each individual policy agent π_i has observed the individual history h_i, and has opted to perform a first action a_i. It is the unique solution to the following *joint history-state* Bellman equation,

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{a}) = R(\boldsymbol{s}, \boldsymbol{a}) + \gamma \, \mathbb{E}_{\boldsymbol{s}', o \mid \boldsymbol{s}, \boldsymbol{a}} \left[\sum_{\boldsymbol{a}'} \boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao}) Q^{\boldsymbol{\pi}}(\boldsymbol{hao}, \boldsymbol{s}', \boldsymbol{a}') \right] . \tag{14}$$

Because this history-state value function has access to both history and state information, it does not suffer from the same issue as the state-only value function; $\Pr(\boldsymbol{a}' \mid \boldsymbol{hao}, \boldsymbol{s}')$ is not only well defined, but can also be trivially reduced to the joint policy probability $\Pr(\boldsymbol{a}' \mid \boldsymbol{hao}, \boldsymbol{s}') = \Pr(\boldsymbol{a}' \mid \boldsymbol{hao}) = \boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao})$ due to the conditional independence between actions and states given histories,

3.4 Multi-Agent Actor-Critic Methods

Actor-Critic methods (AC) (Konda & Tsitsiklis, 2000; Sutton, McAllester, Singh, & Mansour, 2000) are variants of Policy Gradient (PG) approaches that involve the training of policy and critic models. In this section, we define a number of standard centralized and decentralized methods. We primarily consider centralized training of decentralized policies (Lowe et al., 2017; Bono, Dibangoye, Matignon, Pereyron, & Simonin, 2018; Lyu et al., 2021), where there is one policy model per agent (each separately parameterized by θ_i), and a single centralized critic model (parameterized by ϕ). We will omit the model parameterization when clear from the context. To clearly distinguish critic models from the value functions that they are trained to model, we denote them with a hat, e.g., $\hat{V}$ is a critic model trained to model $V^{\boldsymbol{\pi}}$.

Policy gradients are typically expressed in a form that depends on one of the previously defined action-value functions $Q^{\boldsymbol{\pi}}$ (or the respective advantage function $A^{\boldsymbol{\pi}}$). Such values can be estimated in a number of ways; in actor-critic, it is common to use one-step returns and the critic model to estimate $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})$ as $r + \gamma \hat{V}(\boldsymbol{hao})$, and $Q^{\boldsymbol{\pi}}(\boldsymbol{s}, \boldsymbol{a})$ as $r + \gamma \hat{V}(\boldsymbol{s}')$. In advantage actor-critic, the critic model is further used as a baseline for variance reduction,

$$A^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) \doteq Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) - V^{\boldsymbol{\pi}}(\boldsymbol{h}) \approx r + \gamma \hat{V}(\boldsymbol{hao}) - \hat{V}(\boldsymbol{h}) \, , \tag{15}$$

$$A^{\boldsymbol{\pi}}(\boldsymbol{s}, \boldsymbol{a}) \doteq Q^{\boldsymbol{\pi}}(\boldsymbol{s}, \boldsymbol{a}) - V^{\boldsymbol{\pi}}(\boldsymbol{s}) \approx r + \gamma \hat{V}(\boldsymbol{s}') - \hat{V}(\boldsymbol{s}) \, . \tag{16}$$

3.4.1 Centralized Policy

In our evaluation, we also consider the case of a fully centralized policy, i.e., a policy that does not necessarily satisfy conditional independence $\boldsymbol{\pi}(\boldsymbol{a};\boldsymbol{h}) \neq \prod_i \pi_i(a_i;h_i)$, and that fundamentally treats all agents as a single entity that shares all observations and actions. Although this represents a profoundly different control setting from decentralized control, it is a relevant and essential baseline that also represents an upper bound on the performance achievable by decentralized control. The fully centralized policy case is formalized as Joint Actor-Critic (JAC) (Bono et al., 2018; Wang, Hao, Wang, & Taylor, 2019), which employs a centralized history critic $\hat{Q}(\boldsymbol{h},\boldsymbol{a};\phi)$ modeled after $Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})$ to update the fully centralized policy $\boldsymbol{\pi}$ jointly parameterized by θ. The gradient associated with JAC is

$$\nabla_\theta J = (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}\sim\rho(\boldsymbol{h},\boldsymbol{a})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})\nabla_\theta \log \boldsymbol{\pi}(\boldsymbol{a};\boldsymbol{h},\theta)\right] . \tag{17}$$

3.4.2 Decentralized Policy and Critic

Independent Actor-Critic (IAC) Among the decentralized policy gradient variants, we first consider Independent Actor-Critic (IAC) (Peshkin, Kim, Meuleau, & Kaelbling, 2000; Foerster et al., 2018), which trains decentralized policy $\pi_i(a;h)$ and critic $Q_i^{\pi}(h,a)$ models for each agent. Pseudocode for IAC is given in Algorithm 1. In IAC, the gradient associated with the policy parameters θ_i can be derived as

$$\nabla_{\theta_i} J = (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}\sim\rho(\boldsymbol{h},\boldsymbol{a})}\left[Q_i^{\boldsymbol{\pi}}(h_i,a_i)\nabla_{\theta_i} \log \pi_i(a_i;h_i,\theta_i)\right] . \tag{18}$$

3.4.3 Decentralized Policy and Centralized Critic

Next, we consider gradient approximations that make use of centralized values and critics, which is the focus of this work; in some cases, the approximations will be perfect and equivalent to $\nabla_{\theta_i} J(\theta_i)$, while in others they will differ. To distinguish them more clearly from the formally correct gradient $\nabla_{\theta_i} J(\theta_i)$, we use g to denote these approximations.

Independent Actor with Centralized History Critic (IACC-H) IACC-H is a class of centralized critic methods where the joint history critic $\hat{Q}(\boldsymbol{h},\boldsymbol{a};\phi)$ modeled after $Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})$ is used to update each decentralized policy π_i (Foerster et al., 2018; Bono et al., 2018). Pseudocode for IACC-H is given in Algorithm 2. The gradient approximation associated with IACC-H is

$$g_{\boldsymbol{h}} \doteq (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}\sim\rho(\boldsymbol{h},\boldsymbol{a})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})\nabla_{\theta_i} \log \pi_i(a_i;h_i,\theta_i)\right] . \tag{19}$$

The IACC-H approach can be found as one of the variants of COMA (Foerster et al., 2018), which also employs a variance reduction baseline.

Independent Actor with Centralized State Critic (IACC-S) IACC-S employs a centralized state-based critic $\hat{Q}(\boldsymbol{s},\boldsymbol{a};\phi)$ modeled after $Q^{\boldsymbol{\pi}}(\boldsymbol{s},\boldsymbol{a})$ to update each decentralized policy π_i. Pseudocode for IACC-S is given in Algorithm 3. The gradient approximation associated with IACC-S is

$$g_{\boldsymbol{s}} \doteq (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{s},\boldsymbol{a}\sim\rho(\boldsymbol{h},\boldsymbol{s},\boldsymbol{a})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{s},\boldsymbol{a})\nabla_{\theta_i} \log \pi_i(a_i;h_i,\theta_i)\right] . \tag{20}$$

The IACC-S approach can be found both as a variant of COMA (Foerster et al., 2018) and in MADDPG (Lowe et al., 2017).

Abbreviation	Actor	Critic
JAC	Joint Actors	Joint Critic $Q^{\pi}(\boldsymbol{h}, \boldsymbol{a})$
IAC	Independent Actors	Decentralized Critics $Q_i^{\pi}(h_i, a_i)$
IACC-H	Independent Actors	Centralized Critic, history-based $Q^{\pi}(\boldsymbol{h}, \boldsymbol{a})$
IACC-S	Independent Actors	Centralized Critic, state-based $Q^{\pi}(s, \boldsymbol{a})$
IACC-HS	Independent Actors	Centralized Critic, history-state-based $Q^{\pi}(\boldsymbol{h}, s, \boldsymbol{a})$

Table 1: Summary of methods.

Independent Actor with Centralized History-State Critic (IACC-HS) IACC-HS employs a centralized history-state-based critic $\hat{Q}(\boldsymbol{h}, s, \boldsymbol{a}; \phi)$ modeled after $Q^{\pi}(\boldsymbol{h}, s, \boldsymbol{a})$ to update each decentralized policy π_i. Pseudocode for IACC-HS is given in Algorithm 3. The gradient approximation associated with IACC-HS is

$$g_{h,s} \doteq (1-\gamma)\, \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}\sim\rho(\boldsymbol{h},s,\boldsymbol{a})} \left[Q^{\pi}(\boldsymbol{h}, s, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i) \right] . \tag{21}$$

The IACC-HS approach has been used in some value decomposition methods such as VDAC-mix (Su et al., 2021).

In this work, we show that $\nabla_{\theta_i} J$, g_h, and $g_{h,s}$ are all equivalent in expectation (i.e., unbiased), while g_s differs in expectation (i.e., biased). We further show they all have different variance properties. We analyze these gradients via their single-sample Monte Carlo estimates,

$$\hat{g}_i(\boldsymbol{h}, \boldsymbol{a}) \doteq (1-\gamma) Q_i^{\pi}(h_i, a_i) \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i) , \tag{22}$$

$$\hat{g}_h(\boldsymbol{h}, \boldsymbol{a}) \doteq (1-\gamma) Q^{\pi}(\boldsymbol{h}, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i) , \tag{23}$$

$$\hat{g}_s(\boldsymbol{h}, s, \boldsymbol{a}) \doteq (1-\gamma) Q^{\pi}(s, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i) , \tag{24}$$

$$\hat{g}_{h,s}(\boldsymbol{h}, s, \boldsymbol{a}) \doteq (1-\gamma) Q^{\pi}(\boldsymbol{h}, s, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i) , \tag{25}$$

such that

$$\nabla_{\theta_i} J = \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}\sim\rho(\boldsymbol{h},\boldsymbol{a})} \left[\hat{g}_i(\boldsymbol{h}, \boldsymbol{a}) \right] , \tag{26}$$

$$g_h = \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}\sim\rho(\boldsymbol{h},\boldsymbol{a})} \left[\hat{g}_h(\boldsymbol{h}, \boldsymbol{a}) \right] , \tag{27}$$

$$g_s = \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}\sim\rho(\boldsymbol{h},s,\boldsymbol{a})} \left[\hat{g}_s(\boldsymbol{h}, s, \boldsymbol{a}) \right] , \tag{28}$$

$$g_{h,s} = \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}\sim\rho(\boldsymbol{h},s,\boldsymbol{a})} \left[\hat{g}_{h,s}(\boldsymbol{h}, s, \boldsymbol{a}) \right] . \tag{29}$$

To simplify notation, we will occasionally refer to these sample gradients simply as $\hat{g}_i$ $\hat{g}_h$, $\hat{g}_s$, and $\hat{g}_{h,s}$, with their inputs omitted and implicitly sampled from the appropriate distributions. Finally, all the methods described above are summarized in Table 1 for convenience.

Notes on Implementation Details We briefly note that most practical implementations of policy gradients deviate from these theoretical gradient derivations in two ways: firstly, the factor $(1-\gamma)$ can be ignored in practice because optimization practices primarily care about the gradient direction rather than the gradient scale. Modern optimizers also tend to have adaptive step-size mechanisms that in some ways make the scaling factor irrelevant. Secondly, practical implementations estimate the gradient expectations using empirical on-policy experience that does not exactly represent a sample from ρ, since it does not take

into account the discount factor present in the definition of ρ. In practice, this leads to some bias. However, it is likely to be a minor bias for most control problems. In this work, we are interested in analyzing the theoretical properties of the correct gradient definitions. We will therefore consider the discrepancies between the formally correct gradients and common practices as implementation details to be ignored in the theoretical analysis.

Note on Value Function Assumption Our following theories are based on the mathematical definitions of value functions rather than direct application of critic models; this is equivalent to assuming that the critic models have been trained sufficiently and accurately until optimal convergence.

Assumption (Accurate Critics)**.** *We assume critic models that are trained to accurately represent the respective values, i.e.,*

$$\hat{Q}(\boldsymbol{h}, \boldsymbol{a}) = Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) \,, \tag{30}$$

$$\hat{Q}_i(h_i, a_i) = Q^{\boldsymbol{\pi}}(h_i, a_i) \,, \tag{31}$$

$$\hat{Q}(\boldsymbol{s}, \boldsymbol{a}) = Q^{\boldsymbol{\pi}}(\boldsymbol{s}, \boldsymbol{a}) \,, \tag{32}$$

$$\hat{Q}(\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{a}) = Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{a}) \,. \tag{33}$$

This assumption often exists in the form of infinitesimal step sizes for the actors (Konda & Tsitsiklis, 2000; Singh, Kearns, & Mansour, 2000; Zhang & Lesser, 2010; Bowling & Veloso, 2001; Foerster et al., 2018) 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 7.4.

4. Bias and Variance of History-Based Critics

In this section, we establish the convergence of the two history-based critic methods: Independent Actor-Critic (IAC) and Independent Actor with Centralized History Critic (IACC-H). We prove that the decentralized policies receive the same expected gradient from the centralized critic and decentralized critics. We thus 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 (Matignon, Laurent, & Le Fort-Piat, 2012) seen in decentralized learning. It is reassuring that the centralized critic will not encourage a decentralized policy to pursue a centralized policy that is only achievable in a centralized manner but also calls into question the benefits of a centralized critic.

4.1 Bias Analysis of Individual History-Based Gradients

We now show that the gradient updates for the decentralized policies in IAC and IACC-H are the same in expectation. Our work can be thought to extend an early (but perhaps underappreciated) analysis, the difference is that we consider more types of value functions and we use the AC framework and investigate on-policy return values (potentially learned by a critic), rather than Monte Carlo returns;

We now establish a relationship between the values learned by centralized and decentralized critics.

		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 (Claus & Boutilier, 1998).

Lemma 1. *Value functions $Q_i^\pi(h_i, a_i)$ and $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ are related by*

$$Q_i^\pi(h_i, a_i) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right] . \tag{34}$$

(Proof in Appendix B.2). Lemma 1 implies that after the critics have converged to their respective on-policy values, the policy gradients for IACC-H and IAC are equal in expectation (and hence mutually unbiased).

Theorem 1. *The IACC-H sample gradient is an unbiased estimate of the IAC sample gradient, i.e., $g_h = \mathbb{E}\left[\hat{g}_h\right] = \mathbb{E}\left[\hat{g}_i\right] = \nabla_{\theta_i} J$.*

Proof. From Lemma 1, the decentralized value function becomes a marginal expectation of the centralized value function after convergence. Under the expectation over joint histories and joint actions conditioned on (h_i, a_i), these two fixed points are identically equal. Thus, when we invoke Equation (34) in the definition of g_h in Equation (18), we exactly recover the definition of $\nabla_{\theta_i} J$ in Equation (19):

$$\begin{aligned} g_h &= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i)\right] \\ &= (1-\gamma)\,\mathbb{E}_{h_i,a_i}\left[\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right] \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i)\right] \\ &= (1-\gamma)\,\mathbb{E}_{h_i,a_i}\left[Q^\pi(h_i, a_i) \nabla_{\theta_i} \log \pi_i(a_i; h_i, \theta_i)\right] \\ &= \nabla_{\theta_i} J . \end{aligned} \tag{35}$$

□

Theorem 1 establishes the equivalence between the policy gradients given by centralized and decentralized history-based critics. Intuitively, the proof suggests that the marginalization of other agents' histories $\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right]$ occurs in different forms with both centralized and decentralized critics. For decentralized critics, the marginalization is already done implicitly during value learning, as shown in Lemma 1 (Equation (34)). For centralized critics, this marginalization happens explicitly at the policy gradient step, as shown in Equation (35). The implication is significant but may be counter-intuitive. That is, since the expected gradients given by both centralized and decentralized critics are equal, we must conclude that a centralized critic does not lead to better coordination and cooperation in policy learning. In other words, in theory, we cannot expect to obtain a better policy with critic centralization.

4.1.1 Climb Game Example

We use the Climb Game (Claus & Boutilier, 1998), a classic matrix game, to intuitively highlight that IAC and IACC-H give the same policy gradient in expectation. We use a to represent an agent's action variable, and u to represent the concrete values that it can take, e.g., a_1 is the first agent's action, while u_1 is the first action in the action set. The Climb Game is a state-less matrix game (see Table 2) in which agents must cooperate to achieve the optimal reward of 11 by taking the joint action $\langle a_1 = u_1, a_2 = u_1 \rangle$, while overcoming the risk of being punished by worse rewards (-30 or 0) when the agents fail to coordinate. It is notoriously difficult for independent learners to converge to the optimal actions due to the low expected return an agent will receive when the other agent's policy is not already choosing the optimal action.

This cooperation issue arises when a potentially good action u has low on-policy values and high returns are conditional on other agents' policies. In that case, both agents would have to take the lower-valued action (multiple times) to learn to choose it. This creates a dilemma where agents are less prone to take the optimal action u, locking themselves into other local optima. In the Climb Game, the value of u_1 is shadowed because it 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* (Fulda & Ventura, 2007; Panait, Tuyls, & Luke, 2008), a known difficulty in independent learning which usually requires an additional coordination mechanism.

Consider solving the Climb Game with IAC, assuming the agents start with uniformly random policies. The independent value function Q_i^π used in IAC results in action values,

$$
\begin{aligned}
&Q_1^\pi(u_1) = \frac{11-30}{3} = -\frac{19}{3}, \quad Q_1^\pi(u_2) = \frac{-30+7}{3} = -\frac{23}{3}, \qquad Q_1^\pi(u_3) = \frac{6+5}{3} = \frac{11}{3}, \\
&Q_2^\pi(u_1) = \frac{11-30}{3} = -\frac{19}{3}, \quad Q_2^\pi(u_2) = \frac{-30+7+6}{3} = -\frac{17}{3}, \quad Q_2^\pi(u_3) = \frac{5}{3},
\end{aligned}
$$

making u_3 the most attractive action to both agents. Note that we only compare the action values Q^π to determine which one is favored, because in all cases the gradient term $\nabla_{\theta_i} \log \pi_i(a_i)$ is the same (Equation (18)). In this case, both agents will update towards favoring u_3, leading them to never favor u_1 and never converge to the optimal value $Q_i^*(u_1) = 11$.

Consider instead solving the Climb Game with IACC-H, again assuming the agents start with uniformly random policies. In this case, the centralized value function $Q^\pi(\boldsymbol{a})$ is fully capable of representing the individual joint values, mirroring the full reward matrix from Table 2. However, despite this, the policy gradients are still computed with respect to individual agents, and averaged over other agents, therefore suffering from the same

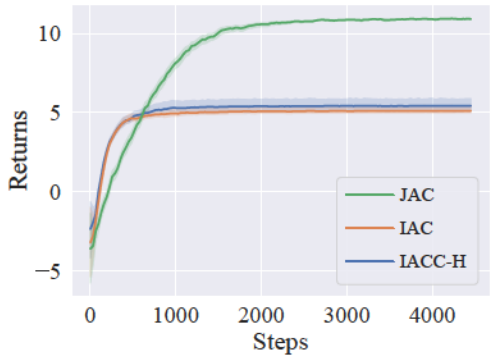


Figure 1: Climb Game empirical returns showing both decentralized and centralized critic methods succumb to the *shadowed equilibrium* problem (showing mean and standard deviation over 50 runs per method).

shadowing issue, which we show next for agent 1,

$$\begin{aligned} g_h &= (1-\gamma)\,\mathbb{E}_{a_1,a_2}\left[Q^{\pi}(a_1,a_2)\nabla_{\theta_i}\log\pi_1(a_1)\right] \\ &= (1-\gamma)\frac{1}{9}\left(Q^{\pi}(u_1,u_1)+Q^{\pi}(u_1,u_2)+Q^{\pi}(u_1,u_3)\right)\nabla_{\theta_i}\log\pi_1(u_1) \\ &\quad+(1-\gamma)\frac{1}{9}\left(Q^{\pi}(u_2,u_1)+Q^{\pi}(u_2,u_2)+Q^{\pi}(u_2,u_3)\right)\nabla_{\theta_i}\log\pi_1(u_2) \\ &\quad+(1-\gamma)\frac{1}{9}\left(Q^{\pi}(u_3,u_1)+Q^{\pi}(u_3,u_2)+Q^{\pi}(u_3,u_3)\right)\nabla_{\theta_i}\log\pi_1(u_3) \\ &= (1-\gamma)\left(-\frac{19}{9}\nabla_{\theta_i}\log\pi_1(u_1)-\frac{23}{9}\nabla_{\theta_i}\log\pi_1(u_2)+\frac{11}{9}\nabla_{\theta_i}\log\pi_1(u_3)\right), \end{aligned} \tag{36}$$

which overall still ends up preferring and promoting the sub-optimal action u_3 over the others. In fact, this should be of no surprise considering that the IAC and AICC-H gradients are the same in expectation, according to Theorem 1, and are therefore subject to the same theoretical cooperation issues in expectation.

Empirical evaluation on the Climb Game (shown in Figure 1) confirms our analysis, showing both methods converge to the suboptimal solution $\langle u_3, u_3\rangle$. At the same time, also unsurprisingly, a centralized controller always gives the optimal solution $\langle u_1, u_1\rangle$. 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 by 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.

4.2 Variance Analysis of Individual History-Based Gradients

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

Theorem 2. *The IACC-H sample gradient has variance greater or equal than that of the IAC sample gradient, i.e.,* $\mathbb{V}\mathrm{ar}\left[\hat{g}_h\right] \geq \mathbb{V}\mathrm{ar}\left[\hat{g}_i\right]$.

Proof. Let $\mu = \mathbb{E}\left[\hat{g}_h\right] = \mathbb{E}\left[\hat{g}_i\right]$ denote the equal expectation of the IAC-HS and IAC gradients (Theorem 1), and $S = \nabla_{\theta_i} \log \pi_i(a_i; h_i) \nabla_{\theta_i} \log \pi_i(a_i; h_i)^\top$ denote the outer product of the policy's score-function vector with itself. Next, we compare the covariance matrices of the two sample gradients,

$$
\begin{aligned}
&\mathbb{C}\mathrm{ov}\left[\hat{g}_h\right] - \mathbb{C}\mathrm{ov}\left[\hat{g}_i\right] \\
&= \mathbb{E}\left[\hat{g}_h \hat{g}_h^\top\right] - \mu\mu^\top - \left(\mathbb{E}\left[\hat{g}_i \hat{g}_i^\top\right] - \mu\mu^\top\right) \\
&= \mathbb{E}\left[\hat{g}_h \hat{g}_h^\top\right] - \mathbb{E}\left[\hat{g}_i \hat{g}_i^\top\right] \\
&= (1-\gamma)\, \mathbb{E}_{h,a}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})^2 S\right] - (1-\gamma)\, \mathbb{E}_{h_i,a_i}\left[Q_i^\pi(h_i, a_i)^2 S\right] \\
&= (1-\gamma)\, \mathbb{E}_{h_i,a_i}\left[\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})^2\right] S\right] - (1-\gamma)\, \mathbb{E}_{h_i,a_i}\left[\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right]^2 S\right] \\
&= (1-\gamma)\, \mathbb{E}_{h_i,a_i}\left[\left(\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})^2\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right]^2\right) S\right] \\
&= (1-\gamma)\, \mathbb{E}_{h_i,a_i}\left[\mathbb{V}\mathrm{ar}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right] S\right] \ . \qquad (37)
\end{aligned}
$$

Because S is positive semi-definite and the variance of any quantity is non-negative, it follows that the matrix in Equation (37) has non-negative diagonal components, which in turn means that $\mathrm{diag}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_h\right]\right) \succeq \mathrm{diag}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_i\right]\right)$ element-wise. Further, $\mathbb{V}\mathrm{ar}\left[\hat{g}_h\right] = \mathrm{trace}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_h\right]\right) \geq \mathrm{trace}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_i\right]\right) = \mathbb{V}\mathrm{ar}\left[\hat{g}_i\right]$. So, not only is the total variance of $\hat{g}_h$ greater than that of $\hat{g}_i$, but the individual variances in any of their dimensions also have the same relationship. □

From the perspective of an agent i trained using IACC-H, the value associated with taking action a_i in history h_i is a random variable taking values $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ depending on other agents' histories $h_{\setminus i}$ and actions $a_{\setminus i}$. The variance of this random variable is dictated by how such values change depending on the other agents' histories and actions, as determined by $\Pr(\boldsymbol{h}, \boldsymbol{a} \mid h_i, a_i)$. In the following subsections, we analyze this total variance increase by examining two sources: The "Multi-Action Variance" (MAV) induced by the other agents' policies, and the "Multi-Observation Variance" (MOV) induced by uncertainty over the other agents' histories. In essence, MAV is influenced by the uncertainty over $\boldsymbol{a}_{\setminus i}$, while MOV by the uncertainty over $\boldsymbol{h}_{\setminus i}$ in Lemma 1.

4.2.1 Multi-Action Variance (MAV)

The Multi-Action Variance is a component of the total value variance dictated by how other agents choose their own actions, according to their own (stochastic) policies, which randomly influences the values experienced by other agents, contributing only to their value variance $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$, but to the variance of the IACC-H sample gradient $\hat{g}_h$ variance as well. In contrast, this variance does not exist for the IAC sample gradient $\hat{g}_i$, because its associated value function $Q_i^\pi(h_i, a_i)$ intrinsically integrates out other agents' actions, per Lemma 1.

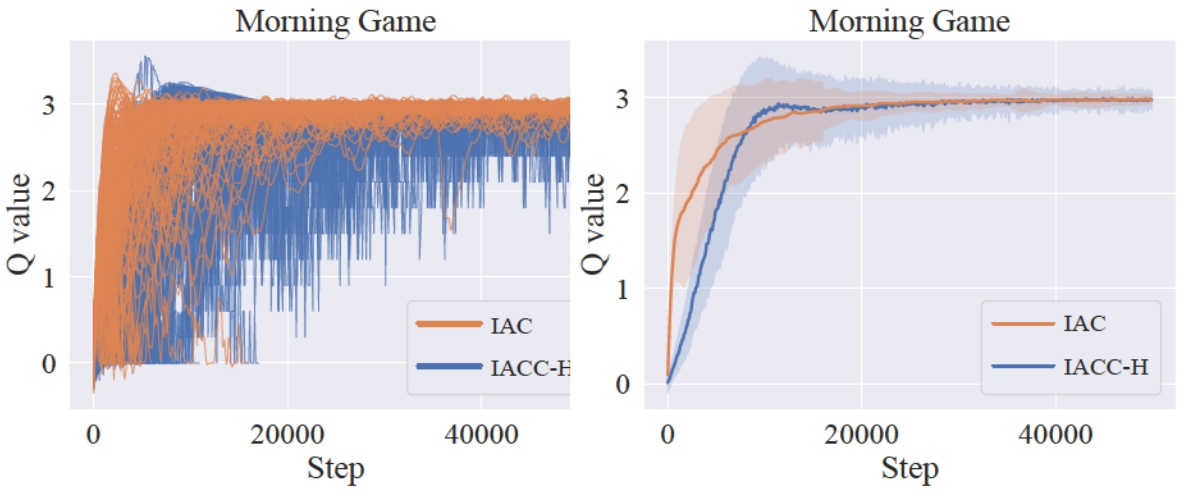
		agent 1	
		pickles	*cereal*
agent 2	*vodka*	1	0
	milk	0	3

Table 3: Return values for the Morning Game.

Morning Game Example The Morning Game shown in Table 3 is a matrix game inspired by previous work (Peshkin et al., 2000) and consists of two agents collaborating to make breakfast by choosing liquid and solid ingredients, of which the best combination is $\langle cereal, milk \rangle$.

A common misleading intuition is that the centralized critic results in lower learning variance. This could be often justified by two separate reasoning: first, the value learning is more stable for a centralized critic; second, the centralized critic have more information and therefore is able to provide a more accurate and specific value estimate. We agree with the former reasoning, due to the lack of environmental stochasticity in this example, the centralized critic can robustly learn accurate values after a few samples, while decentralized critics need more training to correctly average learned values over the unobserved teammate actions. However, when we move on from critic-learning to policy-learning, assume learned or matured (or simply assuming the value function is used) to learn policies, we find the second argument is in fact misleading. In support of this misleading intuitive argument of precise estimates translates to less variance, one might reason that for a centralized critic $\hat{Q}(\boldsymbol{a})$, the joint action *cereal* with *milk* always returns an environmental reward of 3, while *cereal* with *vodka* always returns an environmental reward of 0. On the other hand, for the decentralized critic $\hat{Q}_1(a_1)$, action $a_1 = cereal$ returns an environmental reward of 3 or 0 stochastically, depending on the second agent's action, making it a harder learning task with higher variance.

Contrary to the above intuition, a centralized critic actually results in *higher* variance for the learning agent. What the intuition is missing, is the fact that at the moment when the critic models are used to inform the policy gradients, they may result in additional variance intrinsic to the random variables that are given as input. Suppose agents employ uniform random policies. Then, the IAC value is deterministically $Q_1^\pi(cereal) = \pi_2(milk)Q^\pi(cereal, milk) + \pi_2(vodka)Q^\pi(cereal, vodka) = 1.5$ for $\pi_1(cereal)$ updates, and $Q_1^\pi(pickles) = \pi_2(milk)Q^\pi(pickles, milk) + \pi_2(vodka)Q^\pi(pickles, vodka) = 0.5$ for $\pi_1(pickles)$ updates. In each case, the value associated with the agent's action is deterministic. On the other hand, the IACC-H value is either $Q^\pi(cereal, milk) = 3$ or $Q^\pi(cereal, vodka) = 0$ for $\pi_1(cereal)$ updates, and $Q^\pi(pickles, milk) = 0$ or $Q^\pi(pickles, vodka) = 1$ for $\pi_1(pickles)$ updates, depending on the actual value of the second agent's random action. In essence, decentralized values already represent the expectation over other agents' actions, therefore removing the MAV. Therefore, under both methods, π_1 updates towards *cereal*, but the decentralized-critic updates have less variance and result in a more deterministic improvement in favor of *cereal*.



(a) Individual runs, showing 40 runs per method. (b) Mean and standard deviation.

Figure 2: Q value for updating $\pi(\textit{cereal})$ over time. Showing a larger variance in the values for IACC-H than for IAC, which does not reduce in the long term.

Empirically, Figures 2a and 2b show how the Q-values evolve for the optimal action a_2 in both methods for the Morning Game. First, observe that both types of critics converge to the correct value, around 3, which confirms our convergence analysis for both methods and that the resulting gradients are unbiased. Second, the two methods' Q-value variance appears to differ significantly. Notice that both types of value variances exhibit approximation errors across different seeds, which are also viewed as error bands in the figure. This type of variance caused by approximation error is the only source of variance for decentralized critics. On the other hand, centralized critics have the additional variance that comes from joint actions involving the action-in-question and potentially different actions from other agents, e.g., producing a high value when a teammate chooses *milk* and a low value when a teammate chooses *vodka*. Having to average over teammate actions is, therefore, the essence of MAV.

4.2.2 Multi-Observation Variance (MOV)

For history-based critics in the partially observable case, another source of variance in local value $\mathbb{V}\text{ar}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})\right]$ comes from factored observations. More concretely, for an agent having observed a particular local trajectory h_i, other agents' trajectories $\boldsymbol{h}_{\setminus i}$ may vary, and the decentralized critic must average over this observation variance and provide a single expected value for each local trajectory $Q_i^{\boldsymbol{\pi}}(h_i,a_i)$. The centralized critic, on the other hand, is able to distinguish each combination of trajectories h_i and $\boldsymbol{h}_{\setminus i}$, but when used for a decentralized policy at h_i, teammate history $\boldsymbol{h}_{\setminus i}$ is effectively a random variable sampled from $\Pr(\boldsymbol{h}_{\setminus i} \mid h_i)$, and we expect the mean estimated return during the update process to be $\mathbb{E}_{\boldsymbol{h}|h_i}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})\right]$.

Guess Game Example Consider a one-step two-agent task with binary uniform random observations and binary actions. The agents are rewarded with 10 if both their actions match the other agent's observation, -10 if neither does, and 0 otherwise. In practice, each agent can only guess randomly which action will lead to the positive reward. More broadly,

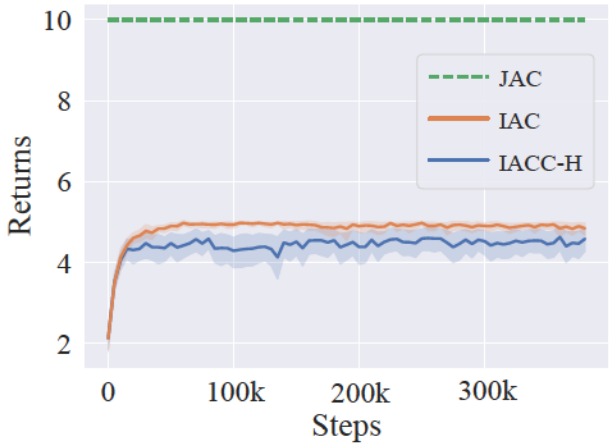


Figure 3: Performance (mean test return) comparison in Guess Game, plotting mean and standard deviation aggregating 40 runs per method; showing centralized critic cannot bias the actors towards the global optimum in the simplest situation.

the decentralized value function takes a value of $Q_i^\pi(h_i, a_i) = 0$ with zero variance for any combination of agent policies. On the other hand, a centralized value function with global scope can distinguish between the various returns, but will return different values depending on the stochastic values of $\boldsymbol{h}$ and $\boldsymbol{a}$, and hence estimates $Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = 10$ with probability 25%, $Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = -10$ with probability 25%, and $Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = 0$ with probability 50%, resulting in a total variance of $\frac{1}{2}10^2 = 50$. In this example, a centralized critic produces returns estimates with higher variance when agents have varying observations.

5. Bias and Variance of State-Based Critics

Having established the convergence of IAC and IACC-H, we will use them as a basis to analyze the convergence of the two state-based critic methods: IACC-S and IACC-HS. We provide a theoretical bias analysis complemented by simple intuitive examples. In this section, we extend the above analysis to commonly used state-based critics, which are often assumed to be strictly beneficial. However, we emphasize that using a state-based critic to train history-based policies is theoretically incorrect, as we show that it may result in bias in the policy gradient. Even with CTDE, where we may access the states during training, we are still aiming to solve a Dec-POMDP problem where the desired output is history-based policies (that can execute independently using local information); that is, state information is not available during execution. Therefore, history-based policies would need to infer, directly or indirectly, the underlying state distribution based on the observed histories. For example, the state distribution may be high in entropy when little information can be derived from observations; hence in some tasks, agents ought to collect information to gain insight into the state distribution before making important decisions. Intuitively, providing the underlying state to the history-based policies during offline training can hinder the policies' ability to reason about uncertainty because when the state is known, efforts such as information collection are unnecessary. This information mismatch between critic (state-based) and actor (history-based) is, thus, intuitively where the bias we discuss in this

section originates. In this section, we show that the state-based critic may introduce bias into the policy gradient compared to the provably convergent history-based critics. That is, the gradient given by the history-based critics $\nabla_{\theta_i} J$ (and g_h) leads to convergence to a local optimum, which is guaranteed to be a Nash equilibrium (Peshkin et al., 2000); however, as we will show next, the gradient given by the state-based critic is biased, so the policy is not guaranteed to also converge to the same local solution. Since we have shown (centralized and decentralized) history-based critics give the same expected policy gradient, we will use centralized history-based critics for the comparison below.

5.1 Bias Analysis of State-Based Gradients

The gradient from a state-based critic is given by:

$$\begin{aligned} g_s &= (1-\gamma)\, \mathbb{E}_{h,s\sim\rho(h,s),a\sim\pi(h)} \left[Q^\pi(s,a) \nabla_{\theta_i} \log \pi_i(a_i; h_i) \right] \\ &= (1-\gamma)\, \mathbb{E}_{h\sim\rho(h),a\sim\pi(h)} \left[\mathbb{E}_{s\sim\rho(s|h)} \left[Q^\pi(s,a) \right] \nabla_{\theta_i} \log \pi_i(a_i; h_i) \right] . \end{aligned} \tag{38}$$

Using the history-based gradient (Equation (19)), we see the state-value-based-gradient g_s is also unbiased iff

$$Q^\pi(h,a) = \mathbb{E}_{s\sim\rho(s|h)} \left[Q^\pi(s,a) \right] . \tag{39}$$

Therefore, the bias of g_s can be analyzed indirectly through $Q^\pi(s,a)$. We adopt the methodology employed by prior work for the single-agent control case (Baisero & Amato, 2022), and will treat $Q^\pi(s,a)$ as an estimator of $Q^\pi(h,a)$. Consequently, if $Q^\pi(s,a)$ is unbiased, then g_s is also unbiased.

We emphasize that expectation over state values cannot be assumed to have the information necessary to produce the correct history values. Essentially, we highlight that history values is an expectation of values over different histories-state pairs, $Q^\pi(h,a) = \mathbb{E}_{s\sim\rho(s|h)} \left[Q^\pi(h,s,a) \right]$, which is different to Equation (39). Of course, when $Q^\pi(h,s,a) = Q^\pi(s,a)$ for all h and s, we can justify Equation (39), yet this equality cannot assume to hold in partially observable settings. We illustrate this point in more detail by comparing history and state value tables and their marginalizations, as well as more intuition and example, in the following paragraphs.

Consider the tabular form of the history-state function under a specific action a, $Q^\pi_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{H}|}$, such that Q table is noted as $Q^\pi_{a,ij} = Q^\pi(h_i, s_j, a)$ where i and j are used to index the joint history h and a state s (we will use this convention for the remainder of this section); if both the state and observation spaces are finite, Q^π_a will be a finite matrix. Also consider the tabular form of the normalized discounted visitations $P^\pi \in [0,1]^{|\mathcal{S}|\times|\mathcal{H}|}$, such that $P^\pi_{ij} = \rho(h,s)$. Note that recovering $\rho(h \mid s)$ and $\rho(s \mid h)$ from P^π requires renormalizing the values in a specific row or column,

$$\rho(h \mid s) = \frac{\rho(h,s)}{\sum_{s'} \rho(s',h)} = \frac{P^\pi_{ij}}{\sum_{j'} P^\pi_{ij'}} , \tag{40}$$

$$\rho(s \mid h) = \frac{\rho(h,s)}{\sum_{h'} \rho(h,s')} = \frac{P^\pi_{ij}}{\sum_{i'} P^\pi_{i'j}} . \tag{41}$$

The matrices Q_a^π and P^π contain the necessary information to determine all marginal history $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ and state values $Q^\pi(s, \boldsymbol{a})$ as normalized dot products:

$$Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \sum_s \rho(s \mid \boldsymbol{h}) Q^\pi(\boldsymbol{h}, s, \boldsymbol{a}) = \sum_{j'} \frac{P^\pi_{ij'}}{\sum_{j'} P^\pi_{ij'}} Q^\pi_{\boldsymbol{a},ij'}\,, \tag{42}$$

$$Q^\pi(s, \boldsymbol{a}) = \sum_{\boldsymbol{h}} \rho(\boldsymbol{h} \mid s) Q^\pi(\boldsymbol{h}, s, \boldsymbol{a}) = \sum_{i'} \frac{P^\pi_{i'j}}{\sum_{i''} P^\pi_{i''j}} Q^\pi_{\boldsymbol{a},i'j}\,. \tag{43}$$

On the other hand, the correct expected state value as listed in Equation (39) is

$$\begin{aligned}
\mathbb{E}_{s\sim\rho(s|h)}\left[Q^\pi(s, \boldsymbol{a})\right] &= \sum_s \rho(s \mid \boldsymbol{h}) Q^\pi(s, \boldsymbol{a}) \\
&= \sum_s \rho(s \mid \boldsymbol{h}) \sum_{\boldsymbol{h}'} \rho(\boldsymbol{h}' \mid s) Q^\pi(\boldsymbol{h}', s, \boldsymbol{a}) \\
&= \sum_{j'} \frac{P^\pi_{ij'}}{\sum_{j''} P^\pi_{ij''}} \sum_{i'} \frac{P^\pi_{i'j'}}{\sum_{i''} P^\pi_{i''j'}} Q^\pi_{\boldsymbol{a},i'j'}\,.
\end{aligned} \tag{44}$$

Note that $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ in Equation (42) only involves the elements on a specific row of both Q_a^π and P^π, while the expected state value in Equation (44) involves all elements of both Q_a^π and P^π. While this provides an intuitive reason for the fact that the two values are not necessarily the same, the inequality proof is still incomplete; in fact, the values in Q_a^π and P^π are not arbitrary, but are related by problem dynamics, policies, and history-state Bellman equations, which may still result in Equations (42) and (44) being numerically equivalent. However, we prove that this is not the case.

Theorem 3. *$Q^\pi(s, \boldsymbol{a})$ may be a biased estimate of $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$, i.e., the equality*

$$Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|h}\left[Q^\pi(s, \boldsymbol{a})\right] \tag{45}$$

cannot be assumed to hold.

Proof. We prove Theorem 3 with an argument analogous to that made in (Baisero & Amato, 2022), which is applicable to our novel always-well-defined state value function $Q^\pi(s, \boldsymbol{a})$. For a generic environment, consider an arbitrary joint action $\boldsymbol{a}$, and two different joint histories $\boldsymbol{h}_A \neq \boldsymbol{h}_B$ associated with the same discounted conditional state visitations distribution $\rho(s \mid \boldsymbol{h}_A) = \rho(s \mid \boldsymbol{h}_B)$ (a reasonably common occurrence in partially observable environments). Because the joint histories are different, there are many history-based policies $\boldsymbol{\pi}$ which result in different future behaviors starting from $\boldsymbol{h}_A$ and $\boldsymbol{h}_B$, which results in different history values,

$$Q^\pi(\boldsymbol{h}_A, \boldsymbol{a}) \neq Q^\pi(\boldsymbol{h}_B, \boldsymbol{a})\,. \tag{46}$$

On the other hand, the conditional state distributions $\rho(s \mid \boldsymbol{h}_A) = \rho(s \mid \boldsymbol{h}_B)$ are equal, which implies that the expected state values are also equal,

$$\mathbb{E}_{s|h_A}\left[Q^\pi(s, \boldsymbol{a})\right] = \mathbb{E}_{s|h_B}\left[Q^\pi(s, \boldsymbol{a})\right]\,. \tag{47}$$

Assuming that the equality $Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|h}\left[Q^\pi(s, \boldsymbol{a})\right]$ holds leads to a contradiction with Equations (46) and (47),

$$\mathbb{E}_{s|h_A}\left[Q^\pi(s, \boldsymbol{a})\right] = Q^\pi(\boldsymbol{h}_A, \boldsymbol{a}) \neq Q^\pi(\boldsymbol{h}_B, \boldsymbol{a}) = \mathbb{E}_{s|h_B}\left[Q^\pi(s, \boldsymbol{a})\right] . \tag{48}$$

Therefore, $Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|h}\left[Q^\pi(s, \boldsymbol{a})\right]$ cannot assume to hold in general. □

Intuitively, recall that the state values alias over all the histories that visit the state. It is often the case that state spaces are considerably more compact than the history space. As a result, the aliasing compresses values over history spaces into values over state spaces in a way that loses important information regarding *uncertainty*. For a particular state, agents may or may not have gathered information through their observations, and this information usually determines their optimal next step and expected returns. Unfortunately, those distinctions cannot be captured by the state value, hence the expected state value is biased for history-based policies. Compared to history values, which are already established to provide the correct gradient, state values become questionable as they are not equivalent to history values in expectation, as suggested by the above Theorem 3. Indeed, as we will show in the next section, the biased expected values for state-based critics will lead to biased policy gradients. We now use an example to illustrate this issue intuitively.

5.1.1 Dec-Tiger Example

This example also serves as alternative proof for Theorem 3. Dec-Tiger (Nair, Tambe, Yokoo, Pynadath, & Marsella, 2003) is a classic Dec-POMDP domain where two agents face a left door and a right door, behind one of which lies a tiger. The agents can individually *listen*, *open-left* or *open-right*. The *listen* action is used to detect the tiger location and results in a reward of -2, and observations *hear-left* or *hear-right* which indicate the correct tiger location with probability 85%. Opening doors terminate the episode immediately and result in rewards of -50 if both agents open the tiger door, 20 if both agents open the non-tiger door, -100 if the agents open different doors, -101 if only one agent opens a door, and it is the tiger's door, and 9 if only one agent opens a door, and it is not the tiger's door.

Consider the pair of agent policies that will listen in the first two timesteps, and open the most promising door in any other timestep based on their individual observations (opening randomly if neither door is more promising than the other), i.e.,

$$\pi_i(h_i) \sim \begin{cases} \delta_{\textit{listen}} & \text{if } |h_i| < 2\,, \\ \delta_{\textit{open-left}} & \text{if } h_i \text{ contains more } \textit{hear-right} \text{ than } \textit{hear-left}\,, \\ \delta_{\textit{open-right}} & \text{if } h_i \text{ contains mroe } \textit{hear-left} \text{ than } \textit{hear-right}\,, \\ \mathcal{U}(\textit{open-left}, \textit{open-right}) & \text{otherwise} \end{cases} \tag{49}$$

In Appendix C, we compute many values associated with these agent policies, including discounted counts, probabilities, state values, history-state values, and history values. Among those, we have computed the state values associated with the *listen* actions $Q^\pi(s, (\textit{listen}, \textit{listen})) = -18.175$, and the history value associated with the empty history $Q^\pi(\boldsymbol{h} = \epsilon, (\textit{listen}, \textit{listen})) = -16.175$. Note that state values associated with the joint listening action are invariant to the particular state, meaning that history value can never be

recovered from the state value. That is, any weighted average over possible states results in the same value,

$$\begin{aligned}\mathbb{E}_{s|\boldsymbol{h}=\epsilon}\left[Q^{\boldsymbol{\pi}}(s,(\textit{listen},\textit{listen}))\right] &= \frac{1}{2}Q^{\boldsymbol{\pi}}(\textit{tiger-left},(\textit{listen},\textit{listen})) \\ &\quad + \frac{1}{2}Q^{\boldsymbol{\pi}}(\textit{tiger-right},(\textit{listen},\textit{listen})) \\ &= -18.175\,, \end{aligned} \tag{50}$$

which is clearly different from $Q^{\boldsymbol{\pi}}(\boldsymbol{h}=\epsilon,(\textit{listen},\textit{listen}) = -16.175$. In this case, the two values are different but still fairly similar. However, this need not be the case in general, and large differences may also exist, as we show next.

Consider the joint history value associated with both agents hearing the tiger on the left twice, denoted as $\boldsymbol{h}_{\text{L}} = ((\textit{hear-left},\textit{hear-left}),(\textit{hear-left},\textit{hear-left}))$, and joint listening action $(\textit{listen},\textit{listen})$. Even though the policies on their own would not listen on that history, this is still a well-defined counterfactual value. Note that the policies will inevitably choose *open-right* next, regardless of which new observation is received. Therefore, $Q^{\boldsymbol{\pi}}(\boldsymbol{h}_{\text{L}},(\textit{listen},\textit{listen}))$ is the expected reward obtained when opening the right door,

$$\begin{aligned} Q^{\boldsymbol{\pi}}(\boldsymbol{h}_{\text{L}},(\textit{listen},\textit{listen})) &= \mathbb{E}_{s|\boldsymbol{h}_{\text{L}}}\left[R(s,(\textit{open-right},\textit{open-right}))\right] \\ &= 0.999\cdot 20 + 0.001\cdot(-50) \\ &= 19.93\,. \end{aligned} \tag{51}$$

On the other hand, as mentioned previously, any weighted average of $Q^{\boldsymbol{\pi}}(s,(\textit{listen},\textit{listen}))$ over state probabilities results in the same value,

$$\begin{aligned}\mathbb{E}_{s|\boldsymbol{h}_{\text{L}}}\left[Q^{\boldsymbol{\pi}}(s,(\textit{listen},\textit{listen}))\right] &= 0.001Q^{\boldsymbol{\pi}}(\textit{tiger-left},(\textit{listen},\textit{listen})) \\ &\quad + 0.999Q^{\boldsymbol{\pi}}(\textit{tiger-right},(\textit{listen},\textit{listen})) \\ &= -18.175\,, \end{aligned} \tag{52}$$

a much more substantial difference in expected values, and in turn, as we will show next, in expected gradient. Practically, state values can lead to incorrect estimations of the (history-based) policy values and then poor policy learning.

5.1.2 Biased Gradient

We note that state values $Q^{\boldsymbol{\pi}}(s,\boldsymbol{a})$ being biased estimates of history values $Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})$ (per Theorem 3) does not formally imply that state gradients $\hat{g}_s$ are biased estimates of history gradients $\hat{g}_{\boldsymbol{h}}$; because an average of biased estimates could technically result in an unbiased estimate, it is still technically possible for the state gradients to be unbiased after all. Here, we provide a definitive proof that this is not the case.

Theorem 4. *The IACC-S and IACC-H gradients may differ, $g_s \neq g_{\boldsymbol{h}}$.*

Proof. See Appendix B.3. □

To summarize the proof, it uses the fact that state values cannot always be used to obtain history values as in Theorem 3, and for at least some environments and policy parameterizations, this bias in value translates to bias in the policy gradient.

The implication of Theorem 4 is that state-based critics are inherently flawed in training history-based policies, for which history values are established to be correct. This may lead to arbitrarily worse performance for state-based critics; as shown in the Dec-Tiger example above, the state-based critic provides no information as to how the policy should be updated—both states have the same state value, causing all actions to be updated equally. This is very different from history values where observed information affects the values and aids policy changes accordingly. However, states can be compact representations or potentially ease learning. After discussing the variance in the state-based cased, we analyze the case of using both state and history as an alternative that is both unbiased and potentially easy to learn.

5.2 Variance Analysis of State-Based Gradients

In this subsection, we discuss the effect of state values $Q^\pi(s, a)$ on the variance of the policy gradient estimate $\hat{g}_s$. Like our bias analysis, we discuss the oracle values Q^π instead of its estimated counterpart $\hat{Q}$. The policy gradient theorem for Dec-POMDPs explicitly requires the history value $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ to be the value used to weigh the policy's score function (Lyu et al., 2021; Bono et al., 2018). In that capacity, $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ is a specific scalar associated with the history and has no variance. On the other hand, using state value $Q^\pi(s, \boldsymbol{a})$ as an estimator of $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ introduces variance, as s is sampled from the history's associated belief $b(\boldsymbol{h})$. However, it does not necessarily imply that the corresponding gradient $\hat{g}_s$ also has a higher variance than $\hat{g}_h$ in general. Instead, we show that if $Q^\pi(s, \boldsymbol{a})$ is unbiased for a given policy and a Dec-POMDP, then $\hat{g}_s$ has a variance greater than or equal to that of $\hat{g}_h$.

Theorem 5. *In the special cases where state values $Q^\pi(s, \boldsymbol{a})$ are unbiased, the IACC-S sample gradient has variance greater or equal than that of the IACC-H sample gradient, i.e.,*

$$Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s \sim \rho(s|h)} \left[Q^\pi(s, \boldsymbol{a}) \right] \implies \mathbb{V}\mathrm{ar} \left[\hat{g}_s \right] \geq \mathbb{V}\mathrm{ar} \left[\hat{g}_h \right] . \tag{53}$$

Proof. We prove the theorem by assuming that the value equality $Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|h} \left[Q^\pi(s, \boldsymbol{a}) \right]$ holds and showing that the variance inequality also holds. Let $\mu = \mathbb{E} \left[\hat{g}_s \right] = \mathbb{E} \left[\hat{g}_h \right]$ denote the assumed equal expectation of the IACC-S and the IACC-H gradients, and $S = \nabla_{\theta_i} \log \pi_i(a_i; h_i) \nabla_{\theta_i} \log \pi_i(a_i; h_i)^\top$ denote the outer product of the policy's score-function vector with itself. Next, we compare the covariance matrices of the two sample gradients,

$$\begin{aligned}
&\mathbb{C}\mathrm{ov} \left[\hat{g}_s \right] - \mathbb{C}\mathrm{ov} \left[\hat{g}_h \right] \\
&= \mathbb{E} \left[\hat{g}_s \hat{g}_s^\top \right] - \mu\mu^\top - \left(\mathbb{E} \left[\hat{g}_h \hat{g}_h^\top \right] - \mu\mu^\top \right) \\
&= \mathbb{E} \left[\hat{g}_s \hat{g}_s^\top \right] - \mathbb{E} \left[\hat{g}_h \hat{g}_h^\top \right] \\
&= (1 - \gamma) \, \mathbb{E}_{h,s,a} \left[Q^\pi(s, \boldsymbol{a})^2 S \right] - (1 - \gamma) \, \mathbb{E}_{h,a} \left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})^2 S \right] \\
&= (1 - \gamma) \, \mathbb{E}_{h,a} \left[\mathbb{E}_{s|h} \left[Q^\pi(s, \boldsymbol{a})^2 \right] S \right] - (1 - \gamma) \, \mathbb{E}_{h,a} \left[\mathbb{E}_{s|h} \left[Q^\pi(s, \boldsymbol{a}) \right]^2 S \right] \\
&= (1 - \gamma) \, \mathbb{E}_{h,a} \left[\left(\mathbb{E}_{s|h} \left[Q^\pi(s, \boldsymbol{a})^2 \right] - \mathbb{E}_{s|h} \left[Q^\pi(s, \boldsymbol{a}) \right]^2 \right) S \right] \\
&= (1 - \gamma) \, \mathbb{E}_{h,a} \left[\mathbb{V}\mathrm{ar}_{s|h} \left[Q^\pi(s, \boldsymbol{a}) \right] S \right] .
\end{aligned} \tag{54}$$

Because S is positive semi-definite and the variance of any quantity is non-negative, it follows that the matrix in Equation (54) has non-negative diagonal components, which in turn means

that $\operatorname{diag}(\mathbb{C}\mathrm{ov}[\hat{g}_s]) \succeq \operatorname{diag}(\mathbb{C}\mathrm{ov}[\hat{g}_h])$ element-wise. Further, $\mathbb{V}\mathrm{ar}[\hat{g}_s] = \operatorname{trace}(\mathbb{C}\mathrm{ov}[\hat{g}_s]) \geq \operatorname{trace}(\mathbb{C}\mathrm{ov}[\hat{g}_h]) = \mathbb{V}\mathrm{ar}[\hat{g}_h]$. So, when also unbiased, not only is the total variance of $\hat{g}_s$ greater than that of $\hat{g}_h$, but the individual variances in any of their dimensions also have the same relationship.

□

Although Theorem 5 alone contains a result that is conditional to an equality which does not necessarily hold for a generic Dec-POMDP, combining Theorems 3 and 5 results in a broader statement about the overall quality of state-based policy gradient estimates.

Corollary 5.1. *IACC-S never has strictly better bias/variance properties than IACC-H, i.e., either its bias is higher (or equal), or its variance is higher (or equal), or neither is lower (or equal).*

Proof. Follows directly from Theorems 4 and 5. □

Beverage Example Consider a (single-agent) *beverage* domain, in which the agent is a barista who serves *coffee* or *tea* to a client. The client, who prefers *coffee* or *tea*, represents the randomly sampled initial state. The agent does not observe the client's preference (we denote this as $h = \varepsilon$) and receives a reward of 1 if it chooses to serve the correct beverage and -1 if it chooses the wrong beverage. In either case, the episode ends. Suppose the agent chooses to serve *tea*. Then,

$$Q^\pi(s = \textit{coffee}, a = \textit{tea}) = -1\,, \tag{55}$$
$$Q^\pi(s = \textit{tea}, a = \textit{tea}) = 1\,, \tag{56}$$
$$Q^\pi(h = \varepsilon, a = \textit{tea}) = 0\,. \tag{57}$$

While $Q^\pi(h = \varepsilon, a = \textit{tea})$ is a constant with zero variance, the random variable $Q^\pi(s, a = \textit{tea})$ conditioned on $h = \varepsilon$ has strictly positive variance,

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{s|h=\varepsilon}[Q^\pi(s, a = \textit{tea})] &= \mathbb{E}_{s|h=\varepsilon}\left[Q^\pi(s, a = \textit{tea})^2\right] - \mathbb{E}_{s|h=\varepsilon}\left[Q^\pi(s, a = \textit{tea})\right]^2 \\ &= \mathbb{E}_{s|h=\varepsilon}[1] - 0^2 \\ &= 1\,. \end{aligned} \tag{58}$$

Therefore, the policy gradient estimates have a higher variance with the state-based critic.

6. Bias and Variance of History-State-Based Critics

In order to avoid the potential bias issue of state-based critics yet also use the centralized state information, we analyze the History-State-based Critic seen in the Unbiased Asymmetric Actor-Critic method for the single agent case (Baisero & Amato, 2022), which we term IACC-HS. Many CTDE works use some form of history-state value function, such as QMIX (Rashid et al., 2018) and its following works, in which the states are used for constructing the mixing hypernetwork; as well as seen in MAPPO (Yu, Velu, Vinitsky, Gao, Wang, Bayen, & Wu, 2022), where the value estimation as a whole takes both history and state as input.

6.1 Bias Analysis of History-State-Based Gradients

As in the single-agent case, history-state values are unbiased estimators of history values, i.e., they are in expectation equivalent.

Lemma 2. *$Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})$ is an unbiased estimate of $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$, i.e., the equality*

$$Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right] \tag{59}$$

holds.

Proof. See Appendix B.2. □

We conclude that the joint history-state gradient $g_{\boldsymbol{h},s}$ as a whole is unbiased, i.e., equivalent in expectation to the individual history gradient $g_{\boldsymbol{h}}$.

Theorem 6. *The IACC-HS and IACC-H gradients are equal, $g_{\boldsymbol{h},s} = g_{\boldsymbol{h}}$.*

Proof.

$$\begin{aligned}
g_{\boldsymbol{h},s} &= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a}) \nabla \log \pi_i(a_i; h_i)\right] \\
&= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right] \nabla \log \pi_i(a_i; h_i)\right] \\
&= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a}) \nabla \log \pi_i(a_i; h_i)\right] \\
&= g_{\boldsymbol{h}}\,.
\end{aligned} \tag{60}$$

□

The equivalence in expected gradients follows directly from the equivalence in expected value functions (Lemma 2). Therefore, unlike state-based critics, a history-state-based critic provides a policy gradient with similar convergence guarantees as a history-based critic.

6.2 Variance Analysis of History-State-Based Gradients

History-state-based critics give policy gradients with equal or more variance than history-based critics (IACC-H) because, for the same joint observation history, the system could be in a range of different states which may exhibit different history-state values. Since history-state-based critics are not biased, the variance analysis is analogous to Theorem 2.

Theorem 7. *The IACC-HS sample gradient has variance greater or equal to that of the IACC-H sample gradient, i.e., $\mathbb{V}\mathrm{ar}\left[\hat{g}_{\boldsymbol{h},s}\right] \geq \mathbb{V}\mathrm{ar}\left[\hat{g}_{\boldsymbol{h}}\right]$.*

Proof. Let $\mu = \mathbb{E}\left[\hat{g}_{\boldsymbol{h},s}\right] = \mathbb{E}\left[\hat{g}_{\boldsymbol{h}}\right]$ denote the equal expectation of the IACC-HS and the IACC-H gradients (Theorem 6), and $S = \nabla_{\theta_i} \log \pi_i(a_i; h_i) \nabla_{\theta_i} \log \pi_i(a_i; h_i)^\top$ denote the outer product of the policy's score-function vector with itself. Next, we compare the covariance

matrices of the two sample gradients,

$$
\begin{aligned}
&\mathbb{C}\mathrm{ov}\left[\hat{g}_{\boldsymbol{h},s}\right] - \mathbb{C}\mathrm{ov}\left[\hat{g}_{\boldsymbol{h}}\right] \\
&= \mathbb{E}\left[\hat{g}_{\boldsymbol{h},s}\hat{g}_{\boldsymbol{h},s}^{\top}\right] - \mu\mu^{\top} - \left(\mathbb{E}\left[\hat{g}_{\boldsymbol{h}}\hat{g}_{\boldsymbol{h}}^{\top}\right] - \mu\mu^{\top}\right) \\
&= \mathbb{E}\left[\hat{g}_{\boldsymbol{h},s}\hat{g}_{\boldsymbol{h},s}^{\top}\right] - \mathbb{E}\left[\hat{g}_{\boldsymbol{h}}\hat{g}_{\boldsymbol{h}}^{\top}\right] \\
&= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},s,\boldsymbol{a})^2 S\right] - (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},\boldsymbol{a})^2 S\right] \\
&= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},s,\boldsymbol{a})^2\right] S\right] - (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},s,\boldsymbol{a})\right]^2 S\right] \\
&= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\left(\mathbb{E}_{s|\boldsymbol{h}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},s,\boldsymbol{a})^2\right] - \mathbb{E}_{s|\boldsymbol{h}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},s,\boldsymbol{a})\right]^2\right) S\right] \\
&= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{V}\mathrm{ar}_{s|\boldsymbol{h}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h},s,\boldsymbol{a})\right] S\right]\,. \qquad (61)
\end{aligned}
$$

Because S is positive semi-definite and the variance of any quantity is non-negative, it follows that the matrix in Equation (61) has non-negative diagonal components, which in turn means that $\mathrm{diag}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_{\boldsymbol{h},s}\right]\right) \succeq \mathrm{diag}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_{\boldsymbol{h}}\right]\right)$ element-wise. Further, $\mathbb{V}\mathrm{ar}\left[\hat{g}_{\boldsymbol{h},s}\right] = \mathrm{trace}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_{\boldsymbol{h},s}\right]\right) \geq \mathrm{trace}\left(\mathbb{C}\mathrm{ov}\left[\hat{g}_{\boldsymbol{h}}\right]\right) = \mathbb{V}\mathrm{ar}\left[\hat{g}_{\boldsymbol{h}}\right]$. So, not only is the total variance of $\hat{g}_{\boldsymbol{h},s}$ greater than that of $\hat{g}_{\boldsymbol{h}}$, but the individual variances in any of their dimensions also have the same relationship. □

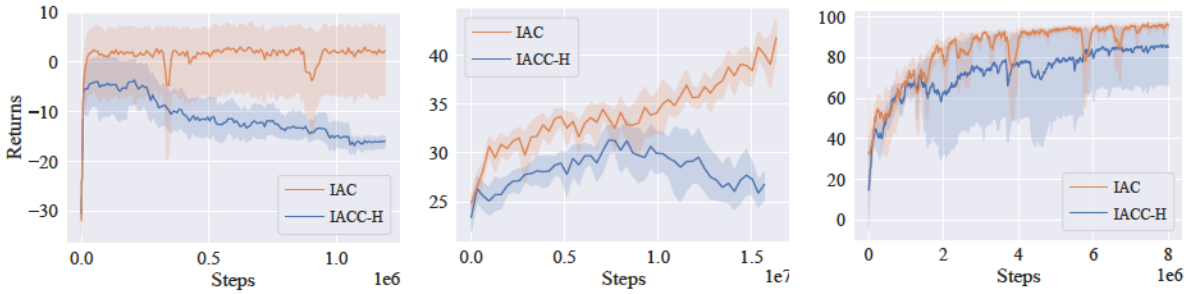
Theorem 7 reports higher policy gradient variance when using a history-state-based critic compared to a history-based critic. This is because, for a particular history, the underlying states may vary; hence for updating that particular history, different corresponding history-state pairs are used, for which the history-state value function can learn different values. Combined with Theorem 2, we get the following corollary relating the variances of the three history-based methods.

Corollary 7.1. *The IACC-HS, IACC-H, and IAC sample gradients have non-monotonically decreasing variance, i.e.,* $\mathbb{V}\mathrm{ar}\left[\hat{g}_{\boldsymbol{h},s}\right] \geq \mathbb{V}\mathrm{ar}\left[\hat{g}_{\boldsymbol{h}}\right] \geq \mathbb{V}\mathrm{ar}\left[\hat{g}_{i}\right]$.

The history-state-based critic is associated with the largest policy gradient variance, whereas the decentralized history-based critic is associated with the lowest policy gradient variance. Loosely speaking, the more aligned a critic is to the policy, the less variance there is to its policy gradient. This theoretical result may be counter-intuitive since a history-state-based critic value function describes a more specific situation, which usually varies less. However, because the policies are decentralized-history-based, using more specific value functions only increases the variation of values when considering a particular local history. However, as we shall see in the next section, our theory only tells half of the story. The three theoretically unbiased critics exhibit different levels of approximation error, where history-state-based critics are often preferable, despite the most policy gradient variance in theory.

7. Empirical Findings and Discussions

In this section, we present experimental results comparing different types of critics. We test on a variety of popular research domains including, but not limited to, classical matrix games, the



(a) Dec-Tiger. (b) Cleaner (c) Find Treasure.

Figure 4: Performances of IAC and IACC-H in different domains (showing mean and standard deviation over 20 runs per method).

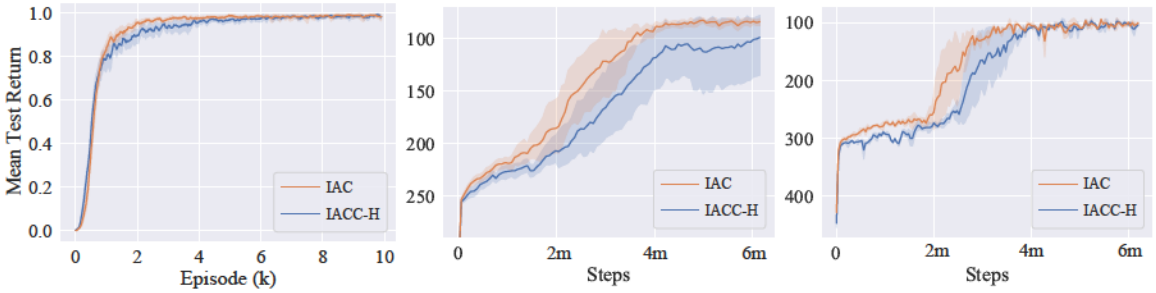
StarCraft Multi-Agent Challenge (SMAC) (Samvelyan, Rashid, de Witt, Farquhar, Nardelli, Rudner, Hung, Torr, Foerster, & Whiteson, 2019), the Multi-agent Particle Environments (MPE) (Mordatch & Abbeel, 2018), and the MARL Environments Compilation (Jiang, 2019). We provide open-source implementations[1] used for our experiments. We first discuss, for history-based critics, how higher policy gradient variance in centralized critics affects performance. We identify several problems caused by the higher policy gradient variance, including exploration and scalability issues. Second, we show empirical evidence that the unbiased centralized history-based critic is subject to the same independent learning pathologies seen with decentralized critics. Third, we highlight the bias and some practical issues with state-based critics. We also discuss insufficiencies of popular benchmarks, which we identify as a sub-class of Dec-POMDPs.

Finally, we report overall preferable performance for history-state-based critics. From Corollary 7.1, we have seen that the history-state-based critic has the most variance. However, in the following experiments, we will see that the approach performs exceptionally well due to its unbiased nature (compared to state-based critics) and low approximation error compared to the history-based critic. We consider the approximation error as an approximation bias, which will become an essential practical trade-off we discuss in this section. We explain why despite having the highest theoretical policy gradient variance, the history-state-based critic may often be favorable due to being unbiased and typically having low approximation bias in practice. In other words, a state-history-based critic is not only theoretically sound but also potentially easier to learn than other critics.

7.1 Policy Gradient Variance May Affect Performance

We first discuss the empirical performance issues of history-based centralized critics (IACC-H) caused by higher policy gradient variance by comparing their performance to decentralized history-based critics (IAC). Although their performance is often identical in most environments, the higher variance in policy learning signal could lead to unstable policies, which may hinder policy learning for one particular agent, or also affect its teammates as shifts in policy cause non-stationarity of the environment from a local perspective. In this section, we

1. `https://github.com/lyu-xg/on-centralized-critics-in-marl`



(a) SMAC 3m. (b) Cross. (c) Antipodal.

Figure 5: Performance comparison in SMAC 3m domain and cooperative navigation domains. In these domains, agents navigate to designated target locations for reward, and are penalized for collisions (showing mean and standard deviation over 20 runs per method).

discuss, with real benchmark examples, some of those impacts on performance and how the policy could become unstable due to additional policy gradient variance.

We report that IAC (independent actor-critic) and IACC-H (an independent actor with centralized history-based critic) usually perform identically on most tasks (Lyu et al., 2021), including Go Together (Jiang, 2019), Merge (Yang et al., 2020), Predator and Prey (Lowe et al., 2017), Capture Target (Omidshafiei, Pazis, Amato, How, & Vian, 2017; Xiao, Hoffman, & Amato, 2019), Small Box Pushing and SMAC (Samvelyan et al., 2019) tasks. IACC-H (with a centralized history-based critic) is less stable in only a few domains shown in Figures 4 and 5. Since the performance of the two history-based critics is similar in most results, we expect that it is because both are unbiased asymptotically (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. Recent investigations of independent learners such as IPPO (Yu et al., 2022) also suggest similar results where decentralized critics show strong performance.

In cooperative navigation domains Antipodal, Cross (Mordatch & Abbeel, 2018; Yang et al., 2020), and Find Treasure (Jiang, 2019), we observe a slightly more pronounced performance difference among runs. These cooperative navigation domains, shown in Figure 5, have 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. Those tasks appear easy to solve, but the observation space is continuous, thus causing large MOV (observation-based variance) in the gradient updates for IACC-H. Observe that some IACC-H 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 drag on the value estimates, which, in turn, hinders improving the policies themselves.

We observe similar policy degradation, more specifically the lazy-agent problem, in the Cleaner domain (Jiang, 2019), a grid-world maze in which agents are rewarded for stepping onto novel locations in the maze. The performance is shown in Figure 4b. The optimal policy is to have the agents split up and cover as much ground as possible. The maze has

two non-colliding paths 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-H), both agents start to take the longer path with more locations to "clean." 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-H with a centralized critic, due to the high variance, which we will discuss in Section 7.1.2, the safer option is favored, resulting in both agents completely ignoring the other path (performance shown in Figure 4b). Overall, high variance in the policy gradient (in the case of the centralized critic) makes the policy more volatile and can indeed result in poor coordination performance in environments that require coordinated series of actions to discover the optimal solution. The same phenomenon is seen in IPPO (Yu et al., 2022) where centralized critics hinder decentralized policy learning.

7.1.1 Hindering Exploration

We test on the classic yet challenging domain Dec-Tiger (Nair et al., 2003). Recall that in Dec-Tiger, 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 a 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 to gain maximum return. Conversely, for simultaneous undesirable actions, 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 4a that IACC-H (with a centralized critic and higher policy gradient variance) does not perform as well as IAC. In the end, the IACC-H 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. That is, IACC-H 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. IAC, on the other hand, has comparatively reduced variance in policy learning, so the environment (including other agents) from the local perspective is more stable, and the agent can thus find a good response policy for adequate performance and cooperation.

7.1.2 Scalability

Another critical consideration for critic design is the scale of the task. A centralized history-based critic's feature representation needs to scale linearly (in the best case) or exponentially (in the worse case) with the number of agents. In contrast, decentralized history-based critics' features can remain constant, and in homogeneous-agent systems, decentralized history-based critics can even share parameters to leverage the problem symmetry. Also, some environments may only require an agent to reason a small amount about the other agents. For example, in environments where agents' decisions rarely depend on other agents' trajectories, the gain of

learning joint value functions is likely minimal. Therefore, we expect decentralized critics to perform better while having better sample efficiency in those domains.

To empirically test our hypothesis, we find environments where we can artificially increase the observation space to highlight the scalability issue. In Capture Target (Lyu & Amato, 2020) where agents are rewarded by simultaneously catching a moving target in a grid world (Figure 6), by increasing the grid size from 6×6 to 12×12, we see a notable comparative drop in overall performance for the centralized critic approach, IACC-H (Figure 6). Since an increase in observation space leads to an increase in Multi-Observation Variance (MOV), it indicates that here the policies of IACC-H do not handle MOV as well as the decentralized critics in IAC. The result may generalize that, for large environments with neural networks, decentralized critics scale better in the face of MOV because they do not involve MOV in policy learning.

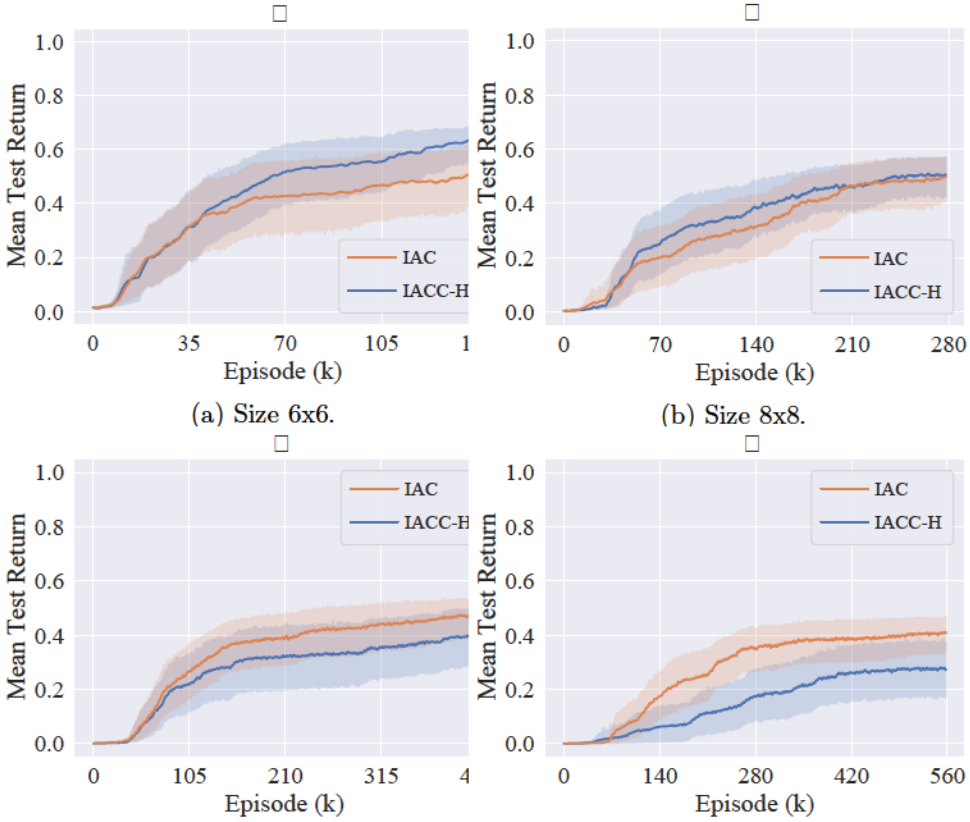
The impact of variance will also change as the number of agents increases. In particular, when learning stochastic policies with a centralized critic in IACC-H, the potential variance in the policy gradient also scales with the number of agents combined with exponentially growing action and observation spaces. On the other hand, IAC's decentralized critics potentially have less stable learning targets in critic bootstrapping as the number of agents increases, but the policy updates still have low variance. Therefore, scalability remains an issue for history-based critics, and the actual performance is likely to depend on the domain, function approximation setup, and other factors. Therefore, IAC should be a better starting point due to more stable policy updates and potentially shared parameters. For state-based critics, although without an explosive feature space, the variance issue similar to that of IACC-H remains (and the issue of bias may cause additional problems as discussed below).

7.2 Cooperation Pathologies

We now investigate how cooperation problems arise with different critic choices. We point out that centralized history-based critic is subjected to the same cooperation issues as their decentralized counterparts.

Move Box (Jiang, 2019) is another commonly used domain where grid-world agents are rewarded by pushing a heavy box (requiring both agents) onto any of two target locations. The farther destination gives a +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 9(d) that both methods (IAC and IACC-H) fall for the *shadowed equilibrium* (mentioned in Section 4.1.1), 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, the on-policy return of the optimal option is inadequate due to an uncooperative teammate policy; thus, the optimal actions are rarely sampled when updating the policies. The same applies to both agents, so the system reaches a suboptimal equilibrium, even though IACC-H is trained in a centralized manner. As a result, same as in Climb Game, Cleaner, and Dec-Tiger, we would not expect centralized critics to help in learning cooperative policies; this connects with Theorem 1 where IAC and IACC-H have the same expected policy gradient.



(a) Size 6x6. (b) Size 8x8.

(c) Size 10x10. (d) Size 12x12.

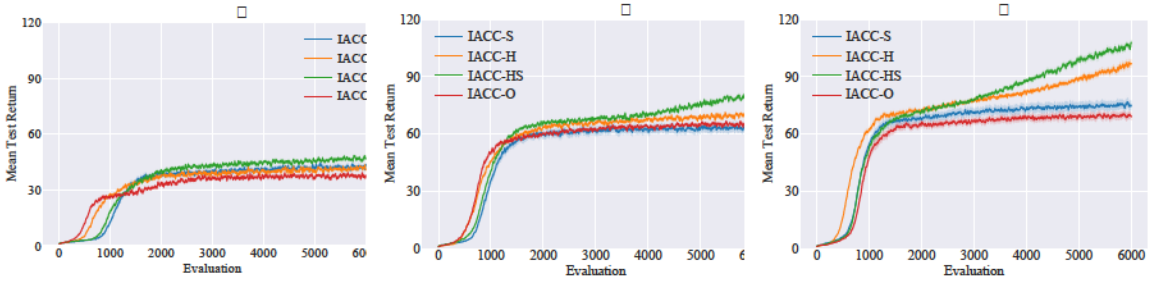
Figure 6: Evaluation on Capture Target domain of IAC and IACC-H (showing mean and standard deviation over 20 runs per method).

7.3 Analysis of State-Based Critics

We now discuss how the bias introduced by state-based critics does not, as one may hope, bias the policies toward a higher-quality solution but often pushes the policies toward suboptimality with limited examples of improved performance.

We show that it cannot facilitate cooperative behavior any better than IAC in our experiments. For example, we consider a multi-agent recycling task with results shown in Figure 9. In the multi-agent recycling task, we expect the agents to work together while maintaining battery levels. The task contains recyclable small and large targets, and the large ones require two agents to act simultaneously. The agents only observe their own battery levels. For policies that recycle large targets to be competitive against small-target-policies (reactive), agents must estimate their teammate's battery level based on their observation history (non-reactive).

Figure 9 shows performance for Recycling agents with various critics, in which none of the methods learns to recycle large targets. Note that a reactive policy cannot achieve optimal performance, which suggests that the methods all fail to cooperate (more discussion in Section 7.3.1). Learning to cooperate in this case is especially difficult because the agents



(a) Observation radius 4. (b) Observation radius 8. (c) Observation radius 16.

Figure 7: Performance comparison in Predator and Prey, showing higher performance of history-state-based critic; that is, highlighting deficiencies of using only state or history-based critic. The results also suggest that the state-based critic can be asymptotically biased thus hindering the convergence of decentralized policies (showing mean and standard deviation over 20 runs per method).

suffer the problem of shadowed equilibrium (Matignon et al., 2012) despite the usage of a centralized critic (as we mentioned in the previous subsection and in Section 4.1.1). Therefore, we observe that state-based critics cannot bias policies in a significant and specific way such that it learns to cooperate. As a result, using state-based critic methods are still subjected to shadowed equilibrium and other cooperation issues.

Interestingly, in the Move Box domain (Figure 9), the state-based critic, being potentially biased, managed to escape the action shadowing Nash equilibrium and managed to obtain higher rewards by moving the box to the farther destination (Pareto optimal). It suggests that in practice, bias and variance may, in special cases, lead to improved performance, such as helping in breaking out of suboptimal local solutions.

7.3.1 Special Cases of Dec-POMDPs

We now investigate the usage of state-based critics, which can be biased in theory (Section 5) in a sub-class of Dec-POMDPs. In particular, we identify a special sub-class of Dec-POMDPs, which is surprisingly common in practice, in which the state-based critic is not critically (or at all) biased. Some environments give partial yet "sufficient" local information, in other words, a decent policy does not depend on the entire history and is not required to retain information gathered from the past. In that case, the observation value and history values are mostly (or exactly) value-equivalent. That is, more information from the full observation history does not add to the value expectation over using just the last observation; in some tasks, reactive policies (policies that only condition on the last observation) can achieve optimal performance (see reactive policies' performance in Figure 9). For example, in the Meeting-in-a-Grid domains (Bernstein, Hansen, & Zilberstein, 2005; Amato, Dibangoye, & Zilberstein, 2009), agents observe their own location but not the teammate while trying to meet in a grid-world. Optimally, agents would navigate to a predetermined spot (e.g., the center) and wait. Another example is Find Treasure (Jiang, 2019) in which one agent has to step onto a trigger location to open a door while the other agent goes through the door to find treasure. Again, even though the task is clearly partially observable, and the

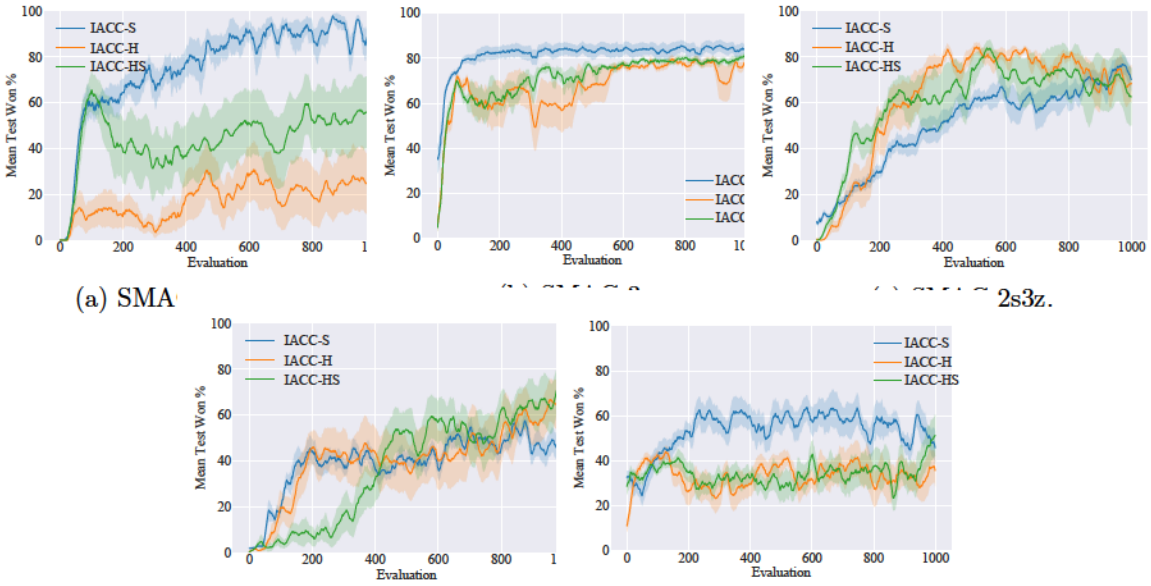


Figure 8: Performance comparison of COMA with different critics in multiple SMAC scenarios (showing mean and standard deviation over 20 runs per variant).

agents only observe local information, the optimal policy does not require remembering history information because policies do not benefit (in expected returns) from the additional observation history. In these environments, one may safely assume that a given state will produce similar return distributions with different histories because the history information does not meaningfully affect the policy and the return distributions since the policy conditions on the last observation, which is produced by the state. However, we note that in harder and more partially-observable tasks, we expect that the optimal policies are not reactive. Environments that require more than a reactive policy require the agent to collect information and retain information over time. If such information gathering is needed, then we can conclude that the state-based critics are, in theory, biased, even assuming no approximation error.

A Subclass of Observation Models Depending on the type of observation model in the task, we also identify acceptable or preferred tasks to use state-based critics, which we call an Agent-Centered Observation Model (ACOM). ACOM can usually indicate where reactive policies may be optimal or near-optimal, thus making state-based critics unbiased or nearly unbiased. Therefore, we also see tasks where it is acceptable or even preferred to use state-based critics. A typical example is the (version 1) StarCraft Multi-Agent Challenge (SMAC) benchmark (Samvelyan et al., 2019). As seen in Figure 8, most scenarios show that state-based critics exhibit the best overall performance. In SMAC, each agent controls a combat unit to engage enemy units, and the state includes the status of friendly agents and opponent units. In contrast, the observations for a given agent are the status of itself and the units whose distances are within a specific range. It uses a deterministic observation model, and in some SMAC tasks, we find that information about far-away units rarely

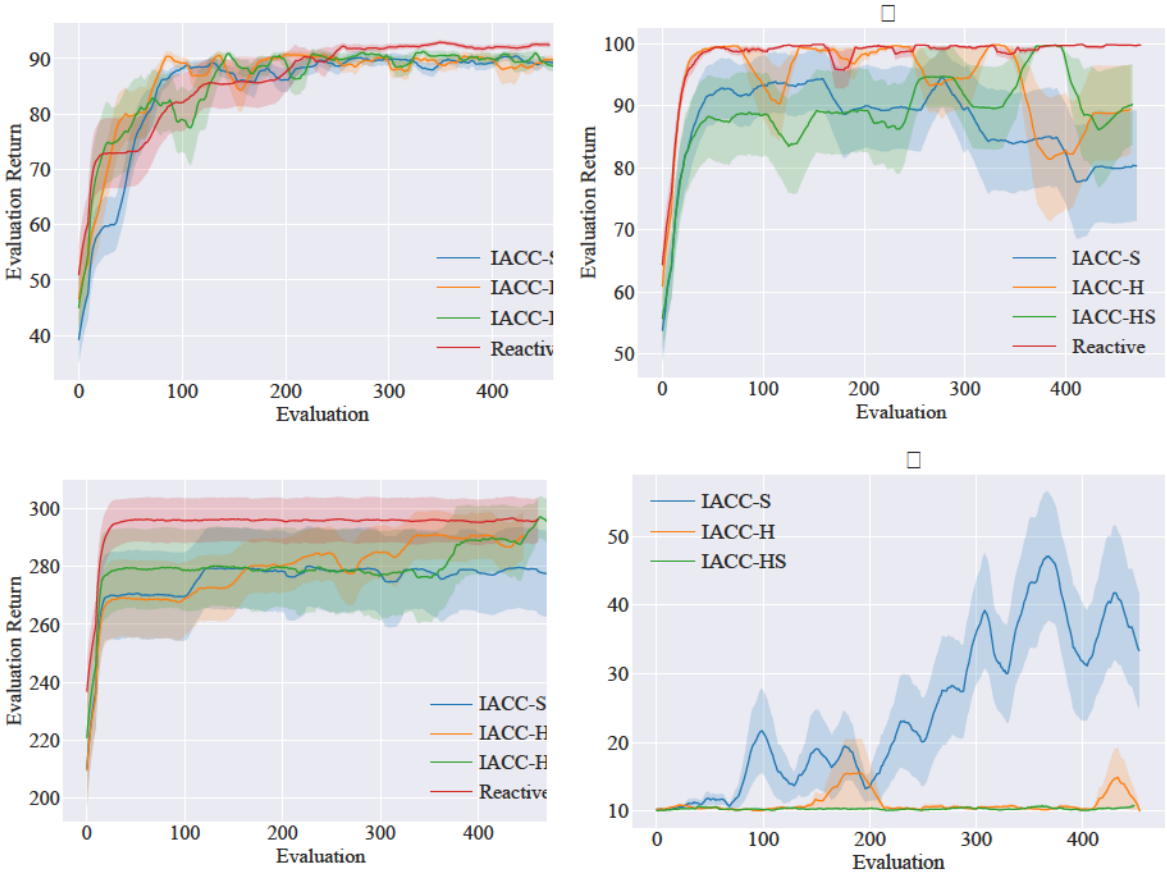
affects an agent's decisions (in a value-equivalent sense). In other words, the occluded information contributes little to the expected values. Thus, the history values associated with a state are usually value-equivalent; as a result, using a state-based critic does not introduce significant bias in this case. In specific scenarios of SMAC, shown in Figure 8, we report that state-based critics perform better than history-based ones. Since both the state and the observation have similar and concise representations, while the history will contain redundant information, it is reasonable to expect that state-based critics may give a more reliable estimate for policy training. In theory, the agents will learn to ignore the redundant history information, which may be time-consuming or difficult to do in practice. Figure 8 uses the original COMA (Foerster et al., 2018) paired with a different type of critics; the high-performing state-based critic shows why subsequent centralized critics literature tends to inherit state-based critics from COMA. Compared to COMA, vanilla multi-agent actor-critic methods also show similar trends (Lyu et al., 2021).

We note that using reactive policies is insufficient to make state-based critics unbiased in the general case because at any time step, the observation produced by the state may still cause the policy to behave differently (Equation (46)). However, if the environment is almost fully observable, where reactive policies can achieve high performance, then the bias from state-based critic would likely be minimal. It is precisely why MADDPG is not biased in their particle environments (Lowe et al., 2017). This situation is also not uncommon, which we see in numerous benchmarks, including the classic task Recycling and environments with radius-based observation models such as SMAC. We hence note that the benchmarks used in recent state-of-the-art works usually have the aforementioned deterministic observation model with limited partial observability, which is, in fact, a special case of imperfect information.

7.4 The Empirical Bias-Variance Trade-Off

Recall that our theoretical analysis is based on the assumption of real value functions, or converged and accurate critic models, and the theory does not strictly hold for learned critics, as those models may inherently carry arbitrary bias, which we call approximation bias. However, we still expect the theorems to provide valuable insights that broadly apply to learned critics as well. In those cases, the approximation bias associated with the real-life learned critics should be considered an additional form of bias on top of the biases (if any) already discussed for the pure value functions in our theories. In practice, as we have seen, value functions are difficult to learn. It is generally more difficult to learn history value functions, especially decentralized history value functions, compared to state or history-state value functions. Therefore, we come to an important observation that, the methods that are more preferable for policy learning in theory have more approximation bias. For example, IAC enjoys the best theoretical policy gradient properties but its decentralized history-based values are the most difficult to learn. This is why some of our theories may seem counter-intuitive, and it is a necessary trade-off to consider when designing and selecting an algorithm for different tasks.

In MARL, the on-policy value is non-stationary since the return distributions depend heavily on the current set of policies. When policies update and change, the learned critic becomes at least partially (if not entirely) obsolete, and biased toward the historical policy-induced return distribution. Bootstrapping from outdated values creates additional bias on



(c) Recycling (Amato et al., 2007).

(d) Move Box (Jiang, 2019).

Figure 9: Performance comparison of IACC-S, IACC-H and IACC-HS in cooperative Dec-POMDPs (showing mean and standard deviation over 20 runs per method).

top of the modeling bias inherent to the model itself. Compounding biases may lead to the divergence of the decentralized policies depending on the learning rate and how drastically the policy changes.

These non-stationarity issues affect all critic approaches since they are all on-policy estimates. However, centralized value function learning is generally better equipped, and state values are generally easier to learn due to their commonly simple representations (compared to histories). Therefore, bootstrapping can be more stable in the case of a centralized critic, especially a centralized state-based critic. As a result, in a cooperative environment with a moderate number of agents, we expect a centralized critic would learn more stably and be less biased (in the actual policy gradient), perhaps counteracting the effect of having a larger variance in the policy gradient, or the case state-based critics, some inherent bias.

Combining our discussions of the variance problems from Section 7.1, 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 the variance of policy updates and the bias of the critic model: a centralized history critic should have a lower bias because it will

have more stable 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; an easy-to-learn state-based critic may have even less practical bias, especially in certain ACOM environments; a history-state-based critic can further reduce the policy gradient bias in practice but at the cost of higher policy gradient variance (Corollary 7.1). In short, the more policy gradient bias we reduce in practice, the more we pay in variance. Regardless, we note that the centralized critic likely faces more severe scalability issues in not only critic learning but also in policy gradient variance. As a result, we cannot expect one method will always dominate the other in terms of performance.

We also identify scalability issues with both centralized and decentralized critics. We observe that the exponential action and history spaces that come with a scaling set of agents significantly increase the centralized critics' policy gradient variance, not to mention that larger models are needed for the factorized spaces. The latter problem can be mitigated by learning a state-based critic if, as identified earlier, we can avoid its bias for the given environment. On the other hand, for decentralized critics, a larger set of agents adds difficulty in value learning due to larger learning target variance for the decentralized critics themselves, a trade-off requiring consideration for each task.

7.4.1 The Safe Option: History-State-Based Critics

We test IACC-HS by concatenating the state and history without changing other aspects of implementation. Unlike IACC-S, IACC-HS is unbiased (Theorem 6) even with large amounts of partial observability. However, IACC-HS is an extreme on the bias-variance trade-off scale in the sense that it should have the least policy gradient bias in practice but the most variance (Corollary 7.1) in theory. We report that for most environments, we obtain similar performance compared to simple history or state-based critics, but for tasks that are more complicated or partially observable, IACC-HS usually outperforms others. Therefore, it is the critic option most widely adaptable to different environments and ought to be the default choice for a task whose nature is unclear. For example, in the Predator and Prey domain in Figure 7, we see that the history-state-based critics have a clear advantage over critics that use state or history information alone, especially in the later stages of training. It suggests that there exist situations where the history-state-based critic can benefit from the advantages of both the state as well as the history as discussed in Section 7.3.1. In general, both Box Pushing and Cleaner are grid-world tasks with observations local to the cells around the agent. However, the agent needs to estimate their teammates' locations or paths to act optimally; this makes a purely state-based critic biased as it cannot encourage information gathering. On the other hand, by incorporating histories, the history-state-based critic is not biased while still having the state to help with value learning.

8. Conclusions

This paper summarizes, standardizes, and analyzes an exhaustive set of centralized critics in multi-agent reinforcement learning. We provide a theoretical basis with bias and variance analysis for each critic variant. Specifically, we prove convergence and policy gradient equivalence between centralized and decentralized history-based critics. We show that history-based centralized critics have higher policy gradient variance in theory than their decentralized

counterparts. The implication of this theoretical result is crucial in understanding that the centralized critics not only cannot meaningfully mitigate cooperation issues but also may harm performance due to its higher policy gradient variance.

For the commonly used state-based critic, we prove that it may produce biased policy gradients compared to the history-based critics. As a result, state-based critics are not theoretically sound and may lead to poor performance. We also give an analysis of specific environment features, which may remove or reduce said biases and explains impressive performance on certain benchmark domains.

We also provide comprehensive practical insights and advice on top of the theoretical results. We argue that the most prominent trade-off is between policy gradient bias and variance; specifically, the higher the policy gradient variance, the less policy gradient bias occurs practically during training, making none of the methods a clear winner on all tasks. However, in general, we recommend the robust IACC-HS (history-state-based critic) as the first choice for a generic or unfamiliar task. The history-state-based critic has the highest theoretical policy gradient variance but often yields favorable performance due to its ease of value learning and unbiased policy gradient.

9. Acknowledgements

This research is supported by the U.S. Office of Naval Research award N00014-19-1-2131, Army Research Office award W911NF20-1-0265 and NSF awards 1816382 and 2044993.

Appendix A. Problems and Solutions of $\Pr(\boldsymbol{a} \mid s)$ and $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$

In Section 3.3.3, we showed a tentative definition for the state value function $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ as the solution to a Bellman equality which we repeat here for clarity,

$$Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = R(s, \boldsymbol{a}) + \gamma \mathbb{E}_{s' \mid s, \boldsymbol{a}} \left[\sum_{\boldsymbol{a}'} \Pr(\boldsymbol{a}' \mid s') Q^{\boldsymbol{\pi}}(s', \boldsymbol{a}') \right], \tag{62}$$

and then argued that this definition is improper due to issues with the definition of the $\Pr(\boldsymbol{a} \mid s)$ term, which conceptually represents the statistical dependency of the actions taken from the system state, but is actually subtly ill-defined. In this section, we first provide a deeper explanation of the problems with $\Pr(\boldsymbol{a} \mid s)$ and why it is not generally well-defined. The issue is often overlooked by practical methods and implementations that attempt to learn centralized state value functions in the form of critic models, but was recently identified and exposed for the single-agent control case (Baisero & Amato, 2022) and for the finite-horizon multi-agent control case (Lyu et al., 2022); here, we extend this prior work to the infinite-horizon multi-agent control case. We then consider a few options to "patch" the issue, providing an alternative definition for $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ which is mathematically sound.

A.1 Problems

At a high level, the issue with the naive definition of state values $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ is that it relies upon a *time-invariant* notion of conditional probabilities $\Pr(\boldsymbol{a} \mid s)$, which turns out to be

generally ill-defined because probabilities $\Pr(\boldsymbol{A}_t = \boldsymbol{a} \mid S_t = s)$ are actually *time-variant*. More formally, the issue is that the term $\Pr(\boldsymbol{a} \mid s)$ is ambiguous and does not adequately identify the random variables $\boldsymbol{A}$ and S onto which the values $\boldsymbol{a}$ and s are assigned. The graphical model associated with the Dec-POMDP defines a joint distribution over *timed* random variables ($S_0, S_1, S_2, \ldots$, $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, $\boldsymbol{A}_2, \ldots$, and others), but not over *time-less* random variables ($S, \boldsymbol{A}$), so probabilities are formally defined only for the timed variables. This issue is explained in more detail for the single-agent control case in (Baisero & Amato, 2022).

A.2 Solutions

Having determined this fundamental flaw in the tentative definition of $Q^\pi(s, \boldsymbol{a})$, we must either resign to the idea that state values are an intrinsically flawed concept, or we can try to determine a different viable definition for state values. In this section, we explore a few ways to fix the issue with the definition of $Q^\pi(s, \boldsymbol{a})$ by replacing $\Pr(\boldsymbol{a} \mid s)$ with a separate expression $p(\boldsymbol{a} \mid s)$. We identify 3 candidates (although more can also be formulated), of which only one is viable.

A.2.1 Limiting Distribution

The limiting distribution p_L is defined as the limit for $t \to \infty$ of the timed distribution,

$$p_L(s) \doteq \lim_{t \to \infty} \Pr(S_t = s)\,, \tag{63}$$

$$p_L(s, \boldsymbol{a}) \doteq \lim_{t \to \infty} \Pr(S_t = s, \boldsymbol{A}_t = \boldsymbol{a})\,, \tag{64}$$

$$p_L(\boldsymbol{a} \mid s) \doteq \frac{p_L(s, \boldsymbol{a})}{p_L(s)}\,. \tag{65}$$

However, the theory of Markov chains tells us that limiting distributions are not guaranteed to converge. Since we are looking for a notion of value function which is well-defined for all possible Dec-POMDPs and policies, the limiting distribution does not satisfy our needs.

A.2.2 Total Visitations Distribution

Another option is the total visitation distribution p_{TV}, defined as the limit for $N \to \infty$ of the average distribution over the first N timesteps,

$$p_{TV}(s) \doteq \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \Pr(S_t = s)\,, \tag{66}$$

$$p_{TV}(s, \boldsymbol{a}) \doteq \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \Pr(S_t = s, \boldsymbol{A}_t = a)\,, \tag{67}$$

$$p_{TV}(\boldsymbol{a} \mid s) \doteq \frac{p_{TV}(s, \boldsymbol{a})}{p_{TV}(s)}\,. \tag{68}$$

Unfortunately, this limit is also not guaranteed to exist for all Dec-POMDPs and policies, and therefore inadequate to satisfy our needs. We show this with a simple counter-example.

Counter-Example of Total Visitations Distribution Consider a Dec-POMDP with a singleton state space $\mathcal{S} = \{\zeta\}$, binary actions $\mathcal{A}_i = \{0, 1\}$, and a singleton observation space $\mathcal{O}_i = \{\omega\}$. We note that the agent policies are intrinsically time-variant due to their dependence on their respective history and history lengths; therefore, even though each agent can only ever receive the same observation over and over, the behaviors of the agents can vary arbitrarily with time. Further, policies in Dec-POMDP are exclusively dependent on time $\pi_i(a; t)$, and deterministic policies can be represented directly as sequences of actions taken at each timestep, e.g., $\pi_i \equiv (0, 1, 0, 1, \ldots)$ denotes a policy that takes first action 0, second action 1, third action 0, fourth action 1, and so on.

Consider the deterministic agent policies $\pi_i \equiv (0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, \ldots)$, which alternates the emission of actions in sequences of ever-increasing length (i.e., one 0, one 1, one 0, one 1, two 0s, two 1s, four 0s, four 1s, eight 0s, eight 1s, sixteen 0s, sixteen 1s, etc). In this case, the average in Equation (67) as a function of time oscillates forever, never converging, e.g., $\frac{1}{N}\sum_{t=0}^{N-1} \Pr(S_t = \zeta, \boldsymbol{A}_t = \mathbf{1}) = \frac{1}{2}$ when $N = 2, 4, 8, 16, 32, \ldots$, and $\frac{1}{N}\sum_{t=0}^{N-1} \Pr(S_t = \zeta, \boldsymbol{A}_t = \mathbf{1}) = \frac{1}{3}$ when $N = 3, 6, 12, 24, 48, \ldots$. Interestingly, since the counter-example uses a singleton state space, the issue with p_{TV} is not even related to state uncertainty, but rather to the general notion of computing non-discounted action counts for time-indefinite horizon sequences of actions produced by time-variant policies.

A.2.3 Discounted Visitations Distribution

Finally, consider the discounted visitation distribution p_{DV}, defined as the average discounted distribution over all timesteps,

$$p_{DV}(s) \doteq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s), \tag{69}$$

$$p_{DV}(s, a) \doteq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s, \boldsymbol{A}_t = \boldsymbol{a}), \tag{70}$$

$$p_{DV}(\boldsymbol{a} \mid s) \doteq \frac{p_{DV}(s, \boldsymbol{a})}{p_{DV}(s)}. \tag{71}$$

In this case, the geometric factor γ^t guarantees convergence in all cases, even in the above counter-example. This is a notion closely related to actual visitation counts, albeit slightly "warped" by the weighting, which makes later visitations count less than earlier visitations. The effect of this "warping" is obviously larger for larger discounting (smaller γ), and smaller for smaller discounting (large γ). Despite this warping, we have finally achieve a somewhat reasonable notion of state-conditioned action distribution, with which to define the state value function as the unique solution to the following *state* Bellman equality,

$$Q^{\pi}(s, \boldsymbol{a}) = R(s, \boldsymbol{a}) + \gamma \mathbb{E}_{s' \mid s, \boldsymbol{a}} \left[\sum_{\boldsymbol{a}'} p_{DV}(\boldsymbol{a}' \mid s') Q^{\pi}(s', \boldsymbol{a}') \right]. \tag{72}$$

Final Notes about Notation and Consistency We conclude by noting that p_{DV} is the same as the ρ defined in Section 3.2.2, which is the preferred notation in all other sections of this paper. Finally, we note that this "patch" is in some sense still consistent with the

definitions of the other value functions, because in the cases where the conditional action distribution also takes some history into account, then $\rho(\boldsymbol{a} \mid \cdot)$ is equivalent to $\boldsymbol{\pi}(\boldsymbol{a} \mid \cdot)$, e.g., $\rho(\boldsymbol{a} \mid \boldsymbol{h}) = \boldsymbol{\pi}(\boldsymbol{a}; \boldsymbol{h})$ and $\rho(\boldsymbol{a} \mid \boldsymbol{h}, s) = \boldsymbol{\pi}(\boldsymbol{a}; \boldsymbol{h})$.

Appendix B. Proofs

B.1 Uniqueness of Bellman Equation Solutions

Here, we prove that the Bellman equations in Equations (9), (10), (13) and (14) each have a unique solution. To simplify the following analysis, we represent Q-functions and the reward function as vectors, which allows us to prove the uniqueness of the solution for a wide class of Bellman-like equations simultaneously.

Let $Q \in \mathbb{R}^n$ be an arbitrary Q-function and let $R \in \mathbb{R}^m$ be a reward function. Equations (9), (10), (13) and (14) can be written in the general form

$$Q = P_0 R + \gamma P_1 Q \,, \tag{73}$$

where $P_0 \in \mathbb{R}^{n \times m}$ and $P_1 \in \mathbb{R}^{n \times n}$ are stochastic matrices. (To see why, notice that each element of $P_0 R$ and $P_1 Q$ is an expectation over the values of R or Q, since the rows of P_0 and P_1 represent probability distributions.) Consequently, we have $0 \leq P_k \leq 1$ elementwise and $P_k e = e$, $\forall\, k \in \{0, 1\}$, where e is the vector of ones with appropriate dimensions.

Proposition 1. *Define the mapping* $B \colon Q \mapsto P_0 R + \gamma P_1 Q$. *When* $\gamma < 1$, *the equation*

$$Q = BQ \tag{74}$$

has a unique solution $Q^* \doteq \sum_{k=0}^{\infty} (\gamma P_1)^k P_0 R$.

Proof. Let $\|Q\|$ denote the maximum norm of Q. Since the infinity norm is submultiplicative (and coincides with the maximum norm for vectors), then for every $Q, Q' \in \mathbb{R}^n$,

$$\begin{aligned} \left\| BQ - BQ' \right\| &= \left\| P_0 R + \gamma P_1 Q - P_0 R - \gamma P_1 Q' \right\| \\ &= \left\| \gamma P_1 (Q - Q') \right\| \\ &\leq \gamma \left\| P_1 \right\|_\infty \left\| Q - Q' \right\| . \end{aligned} \tag{75}$$

The facts that $P_1 \geq 0$ and $P_1 e = e$ imply that $\left\| P_1 \right\|_\infty = 1$. Therefore,

$$\left\| BQ_1 - BQ_2 \right\| \leq \gamma \left\| Q_1 - Q_2 \right\| , \tag{76}$$

and B is a contraction mapping with respect to the maximum norm. We conclude that B admits a unique fixed point Q^*; if B had two distinct fixed points $Q_1^* \neq Q_2^*$, then $\|BQ_1^* - BQ_2^*\| = \|Q_1^* - Q_2^*\|$, which would contradict Equation (76) for $\gamma < 1$.

To complete the proof, we must identify Q^*. Expanding Equation (74) gives

$$Q = P_0 R + \gamma P_1 Q \,,$$

and rearranging the terms yields the solution

$$Q = (I - \gamma P_1)^{-1} P_0 R \,.$$

From the infinite geometric series, we have $(I - \gamma P_1)^{-1} = \sum_{k=0}^{\infty} (\gamma P_1)^k$, and therefore $Q^* = \sum_{k=0}^{\infty} (\gamma P_1)^k P_0 R$. □

B.2 Value Function Identities

Lemma 1. *Value functions $Q_i^\pi(h_i, a_i)$ and $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ are related by*

$$Q_i^\pi(h_i, a_i) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right] . \tag{34}$$

Proof. $Q_i^\pi(h_i, a_i)$ is defined as the unique solution to Equation (10). Next, we show that $\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right]$ (as a function of h_i and a_i) is also a solution to Equation (10); since the solution is unique by Proposition 1, the two functions must be identically equal and Equation (34) must hold.

We first relate the two history-based reward functions:

$$\begin{aligned}
R_i(h_i, a_i) &= \mathbb{E}_{s,a_{\setminus i}|h_i}\left[R(s, \boldsymbol{a})\right] \\
&= \mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[R(s, \boldsymbol{a})\right] \\
&= \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}|h_i,a_i}\left[R(s, \boldsymbol{a})\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[R(s, \boldsymbol{a})\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[R(\boldsymbol{h}, \boldsymbol{a})\right] .
\end{aligned} \tag{77}$$

Next, we show the following identity:

$$\begin{aligned}
&\mathbb{E}_{o_i|h_i,a_i}\left[\sum_{a_i'}\pi_i(a_i'; h_i a_i o_i)\,\mathbb{E}_{\boldsymbol{hao},\boldsymbol{a}'|h_i a_i o_i,a_i'}\left[Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right]\right] \\
&= \mathbb{E}_{o_i|h_i,a_i}\left[\mathbb{E}_{a_i'|h_i a_i o_i}\left[\mathbb{E}_{\boldsymbol{hao},\boldsymbol{a}'|h_i a_i o_i,a_i'}\left[Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right]\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a},o,\boldsymbol{a}'|h_i,a_i}\left[Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[\mathbb{E}_{o|\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{\boldsymbol{a}'|\boldsymbol{hao}}\left[Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right]\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[\mathbb{E}_{o|\boldsymbol{h},\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'}\boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao})Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right]\right] .
\end{aligned} \tag{78}$$

Therefore, when we substitute $\mathbb{E}_{\boldsymbol{hao},\boldsymbol{a}'|h_i a_i o_i,a_i'}\left[Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right]$ (analogous to $\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right]$) into Equation (10), we obtain

$$\begin{aligned}
&R_i(h_i, a_i) + \gamma\,\mathbb{E}_{o_i|h_i,a_i}\left[\sum_{a_i'}\pi_i(a_i'; h_i a_i o_i)\,\mathbb{E}_{\boldsymbol{hao},\boldsymbol{a}'|h_i a_i o_i,a_i'}\left[Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[R(\boldsymbol{h}, \boldsymbol{a}) + \gamma\,\mathbb{E}_{o|\boldsymbol{h},\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'}\boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao})Q^\pi(\boldsymbol{hao}, \boldsymbol{a}')\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right] ,
\end{aligned} \tag{79}$$

where the final equality follows from Equation (9). Because the original quantity $\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right]$ is unchanged after the substitution into Equation (10), we conclude that it must be the unique fixed point guaranteed by Proposition 1, and hence $Q_i^\pi(h_i, a_i) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a})\right]$. □

Lemma 2. *$Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})$ is an unbiased estimate of $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$, i.e., the equality*

$$Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right] \tag{59}$$

holds.

Proof. $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ is defined as the unique solution to Equation (9). Next, we show that $\mathbb{E}_{s|\boldsymbol{h},\boldsymbol{a}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right]$ (as a function of $\boldsymbol{h}$ and $\boldsymbol{a}$) is also a solution to Equation (9); since the solution is unique by Proposition 1, the two functions must be identically equal and Equation (59) must hold. We first recall the definition of the joint history reward function, $R(\boldsymbol{h}, \boldsymbol{a}) \doteq \mathbb{E}_{s|\boldsymbol{h}}\left[R(s, \boldsymbol{a})\right]$. Next, we show the following identity:

$$\begin{aligned}
&\mathbb{E}_{o|\boldsymbol{h},\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'} \boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao})\, \mathbb{E}_{s'|\boldsymbol{hao}}\left[Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right]\right] \\
&= \mathbb{E}_{o|\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{\boldsymbol{a}'|\boldsymbol{hao}}\left[\mathbb{E}_{s'|\boldsymbol{hao}}\left[Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right]\right]\right] \\
&= \mathbb{E}_{o,s',\boldsymbol{a}'|\boldsymbol{h},\boldsymbol{a}}\left[Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right] \\
&= \mathbb{E}_{s,o,s',\boldsymbol{a}'|\boldsymbol{h},\boldsymbol{a}}\left[Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right] \\
&= \mathbb{E}_{s|\boldsymbol{h}}\left[\mathbb{E}_{s',o|s,\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'} \boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao})\, \mathbb{E}_{s'|\boldsymbol{hao}}\left[Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right]\right]\right] .
\end{aligned} \tag{80}$$

Therefore, when we substitute $\mathbb{E}_{s'|\boldsymbol{hao}}\left[Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right]$ (analogous to $\mathbb{E}_{s|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right]$) into Equation (9), we obtain

$$\begin{aligned}
&R(\boldsymbol{h}, \boldsymbol{a}) + \gamma\, \mathbb{E}_{o|\boldsymbol{ha}}\left[\sum_{\boldsymbol{a}'} \boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao})\, \mathbb{E}_{s'|\boldsymbol{hao}}\left[Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right]\right] \\
&= \mathbb{E}_{s|\boldsymbol{h}}\left[R(s, \boldsymbol{a}) + \gamma\, \mathbb{E}_{s',o|s,\boldsymbol{a}}\left[\sum_{\boldsymbol{a}'} \boldsymbol{\pi}(\boldsymbol{a}'; \boldsymbol{hao}) Q^\pi(\boldsymbol{hao}, s', \boldsymbol{a}')\right]\right] \\
&= \mathbb{E}_{s|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right] ,
\end{aligned} \tag{81}$$

where the final equality follows from Equation (14). Because the original quantity $\mathbb{E}_{s|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right]$ is unchanged after the substitution into Equation (9), we conclude that it must be the unique fixed point guaranteed by Proposition 1, and hence $Q^\pi(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})\right]$.

□

B.3 Biased State-Based Critic

Lemma 3. *$Q^\pi(s, \boldsymbol{a})$ may be a biased estimate of $Q_i^\pi(h_i, a_i)$, i.e., the equality*

$$Q_i^\pi(h_i, a_i) = \mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s, \boldsymbol{a})\right] \tag{82}$$

does not necessarily hold.

Proof. We note that Theorem 3 does not imply this theorem, however, we prove this theorem using an analogous argument by contradiction, which is applicable to our novel always-well-defined state value function $Q^\pi(s, \boldsymbol{a})$. For a generic environment, consider an arbitrary action

a_i, and two different individual histories $h_{(i,A)} \neq h_{(i,B)}$ associated with the same state and joint action distributions $\rho(s, \boldsymbol{a} \mid h_{(i,A)}, a_i) = \rho(s, \boldsymbol{a} \mid h_{(i,B)}, a_i)$ (a plausible occurrence in partially observable environments). Because the individual histories are different, there are many history-based policies π_i which result in different future behaviors starting from $h_{(i,A)}$ and $h_{(i,B)}$, which result in different history values,

$$Q_i^\pi(h_{(i,A)}, a_i) \neq Q_i^\pi(h_{(i,B)}, a_i) \,. \tag{83}$$

On the other hand, the conditional state and joint action distributions $\rho(s, \boldsymbol{a} \mid h_{(i,A)}, \boldsymbol{a}) = \rho(s, \boldsymbol{a} \mid h_{(i,B)}, \boldsymbol{a})$ are equal, which implies that the expected state values are also equal,

$$\mathbb{E}_{s,\boldsymbol{a}|h_{(i,A)},a_i}\left[Q^\pi(s, \boldsymbol{a})\right] = \mathbb{E}_{s,\boldsymbol{a}|h_{(i,B)},a_i}\left[Q^\pi(s, \boldsymbol{a})\right] \,. \tag{84}$$

Assuming that the equality $Q_i^\pi(h_i, a_i) = \mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s, \boldsymbol{a})\right]$ holds leads to a contradiction with Equations (83) and (84),

$$\mathbb{E}_{s,\boldsymbol{a}|h_{(i,A)},a_i}\left[Q^\pi(s, \boldsymbol{a})\right] = Q_i^\pi(h_{(i,A)}, a_i) \neq Q_i^\pi(h_{(i,B)}, a_i) = \mathbb{E}_{s,\boldsymbol{a}|h_{(i,B)},a_i}\left[Q^\pi(s, \boldsymbol{a})\right] \,. \tag{85}$$

Therefore, $Q_i^\pi(h_i, a_i) = \mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s, \boldsymbol{a})\right]$ does not hold in general. □

Theorem 4. *The IACC-S and IACC-H gradients may differ, $g_s \neq g_h$.*

Proof. To cover different angles and intuitions, we provide two proofs; one by contradiction, and one by example. Both proofs are based on a simple policy parameterization for which the statements of this theorem and Theorem 3 imply each other. Since Theorem 3 is already proven to hold in a way that is invariant to the policy parameterization, then this theorem must also hold. The policy parameterization is the tabular one where $\pi_i(a; h) = \theta_{i,h,a}$, which can likewise be written as a linear model $\pi_i(a; h) = \theta_i^\top \mathbb{1}_{h,a}$ where $\mathbb{1}_{h,a}$ is an indicator vector of 0s with a 1 only in the dimension associated with the h, a pair; this parameterization requires constraints $\theta_{i,h,a} \geq 0$ and $\sum_a \theta_{i,h,a} = 1$, but this does not influence the conclusion.

Proof by Contradiction For tabular policies, we have

$$\begin{aligned}
g_s &= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}}\left[Q^\pi(s, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i)\right] \\
&= (1-\gamma)\,\mathbb{E}_{h_i,a_i}\left[\mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i)\right]\right] \\
&= (1-\gamma)\,\mathbb{E}_{h_i,a_i}\left[\mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s, \boldsymbol{a})\right] \frac{\mathbb{1}_{h_i,a_i}}{\pi_i(a_i; h_i)}\right] \\
&= (1-\gamma) \sum_{h_i,a_i} \rho(h_i, a_i)\, \mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s, \boldsymbol{a})\right] \frac{\mathbb{1}_{h_i,a_i}}{\pi_i(a_i; h_i)} \,,
\end{aligned} \tag{86}$$

$$\begin{aligned}
g_h &= (1-\gamma)\,\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i)\right] \\
&= (1-\gamma)\,\mathbb{E}_{h_i,a_i}\left[\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i}\left[Q^\pi(\boldsymbol{h}, \boldsymbol{a}) \nabla_{\theta_i} \log \pi_i(a_i; h_i)\right]\right] \\
&= (1-\gamma)\,\mathbb{E}_{h_i,a_i}\left[Q_i^\pi(h_i, a_i) \frac{\mathbb{1}_{h_i,a_i}}{\pi_i(a_i; h_i)}\right] \\
&= (1-\gamma) \sum_{h_i,a_i} \rho(h_i, a_i) Q_i^\pi(h_i, a_i) \frac{\mathbb{1}_{h_i,a_i}}{\pi_i(a_i; h_i)} \,.
\end{aligned} \tag{87}$$

More specifically, the dimensions of g_s and g_h associated with h_i, a_i take on values

$$[g_s]_{h_i,a_i} = (1-\gamma)\rho(h_i,a_i)\,\mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s,\boldsymbol{a})\right]\frac{1}{\pi_i(a_i;h_i)}\,, \tag{88}$$

$$[g_h]_{h_i,a_i} = (1-\gamma)\rho(h_i,a_i)Q_i^\pi(h_i,a_i)\frac{1}{\pi_i(a_i;h_i)}\,, \tag{89}$$

which are equal if and only if $Q_i^\pi(a_i,h_i) = \mathbb{E}_{s,\boldsymbol{a}|h_i,a_i}\left[Q^\pi(s,\boldsymbol{a})\right]$. Lemma 3 generally disproves such an equality. Therefore, the gradient equality $g_s = g_h$ does not necessarily hold.

Proof by Example Continuing from the Dec-Tiger example in Section 5.1, we compare the IACC-H and IACC-S gradients g_h and g_s for agent 1 at one specific dimension and show that they are different. We focus on the values associated with $h_1 = (\textit{listen}, \textit{hear-left}, \textit{listen}, \textit{hear-left})$, which is intrinsically associated with action $a_1 = \textit{open-right}$. We denote this dimension of the two gradient as $[g_h]_{h_1,a_1}$ and $[g_s]_{h_1,a_1}$. To compute these values, we need to reference all the joint histories that are compatible with the individual history h_1; we denote these using labels that reference the second agent's observations,

$$\boldsymbol{h}_{LL} = ((\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-left}),(\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-left}))\,, \tag{90}$$

$$\boldsymbol{h}_{LR} = ((\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-left}),(\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-right}))\,, \tag{91}$$

$$\boldsymbol{h}_{RL} = ((\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-right}),(\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-left}))\,, \tag{92}$$

$$\boldsymbol{h}_{RR} = ((\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-right}),(\textit{listen},\textit{listen}),(\textit{hear-left},\textit{hear-right}))\,. \tag{93}$$

We also need to reference all the joint actions that are compatible with the individual action a_1; we denote these using labels that reference the second agent's action, $\boldsymbol{a}_L = (\textit{open-right}, \textit{open-left})$ and $\boldsymbol{a}_R = (\textit{open-right}, \textit{open-right})$. We do not consider other joint histories and actions because they do not contribute to $[g_h]_{h_1,a_1}$ or $[g_s]_{h_1,a_1}$. Given the details of the Dec-Tiger problem, we can compute the conditional probabilities of these joint histories and actions $\Pr(\boldsymbol{h},\boldsymbol{a} \mid h_1,a_1)$, contained in Table 7. To start, we have

$$\begin{aligned}\rho(h_1,a_1) &= \rho(h_1)\pi_1(a_1;h_1)\\ &= \sum_s \rho(h_1 \mid s)\rho(s)\pi_1(a_1;h_1)\\ &= 0.85^2\cdot\frac{1}{2}\cdot 1 + 0.15^2\cdot\frac{1}{2}\cdot 1\\ &= 0.373\,.\end{aligned} \tag{94}$$

Then, using probabilities from Tables 7 and 8 and values from Appendix C,

$$\begin{aligned}
[g_h]_{h_1,a_1} &= (1-\gamma)\rho(h_1,a_1)\,\mathbb{E}_{h,a|h_1,a_1}\left[Q^\pi(\boldsymbol{h},\boldsymbol{a})\right]\frac{1}{\pi_1(a_1;h_1)} \\
&= (1-\gamma)\cdot 0.373\cdot(-0.854)\cdot\frac{1}{1} \\
&= -0.319(1-\gamma)\,. \qquad (95) \\
[g_s]_{h_1,a_1} &= (1-\gamma)\rho(h_1,a_1)\,\mathbb{E}_{s,a|h_1,a_1}\left[Q^\pi(s,\boldsymbol{a})\right]\frac{1}{\pi_1(a_1;h_1)} \\
&= (1-\gamma)\cdot 0.373\cdot(4.465)\cdot\frac{1}{1} \\
&= 1.665(1-\gamma)\,. \qquad (96)
\end{aligned}$$

Comparing values, we have $[g_s]_{h_1,a_1} \neq [g_h]_{h_1,a_1}$, and therefore $g_s \neq g_h$. □

Appendix C. Dec-Tiger Values and Gradients

Here, we manually calculate the state values $Q^\pi(s,\boldsymbol{a})$ for the Dec-Tiger problem. We consider the case of two agents who always listen twice, then open the most promising door: left door, if the tiger has been heard more on the right; right door, if the tiger has been heard more on the right; otherwise, randomly. We proceed by progressively computing all values of $\eta(s)$, $\eta(s,\boldsymbol{a})$, $\rho(\boldsymbol{a}\mid s)$, and $Q^\pi(s,\boldsymbol{a})$ for state `L` and the viable joint actions; by symmetry, analogous calculations can be made for state `R`.

To simplify notation, we denote state *tiger-left* as `L`, state *tiger-right* as `R`, observation *hear-left* as `L`, observation *hear-right* as `R`, action *listen* as `Li`, action *open-left* as `L`, and action *open-right* as `R`.

C.1 Discounted Counts and Probabilities

We start by programmatically computing relevant probabilities, in Tables 4 to 8 [2]. Because Dec-Tiger is episodic, and our policies guarantee that the episode terminates after timestep 3, we get the discounted state counts

$$\begin{aligned}
\eta(s=\texttt{L}) &= \Pr(S_0=\texttt{L}) + \gamma\Pr(S_1=\texttt{L}) + \gamma^2\Pr(S_2=\texttt{L}) \\
&= \frac{1+\gamma+\gamma^2}{2}\,. \qquad (97)
\end{aligned}$$

2. See `https://github.com/lyu-xg/on-centralized-critics-in-marl/tree/main/dectiger`

Table 4: $\Pr(\boldsymbol{A}_2 = \boldsymbol{a} \mid \boldsymbol{H}_2 = \boldsymbol{h}) = \boldsymbol{\pi}(\boldsymbol{a}; \boldsymbol{h})$. In this table, a joint observation $\boldsymbol{h} = ((A, B), (C, D))$ means that the two agents have received observations A and B on the first timestep, and C and D on the second timestep.

$\boldsymbol{a}$	(L, L)	(L, R)	(R, L)	(R, R)
$\boldsymbol{h}$				
((L, L), (L, L))				1.0
((L, L), (L, R))			0.5	0.5
((L, L), (R, L))		0.5		0.5
((L, L), (R, R))	0.25	0.25	0.25	0.25
((L, R), (L, L))			0.5	0.5
((L, R), (L, R))			1.0	
((L, R), (R, L))	0.25	0.25	0.25	0.25
((L, R), (R, R))	0.5		0.5	
((R, L), (L, L))		0.5		0.5
((R, L), (L, R))	0.25	0.25	0.25	0.25
((R, L), (R, L))		1.0		
((R, L), (R, R))	0.5	0.5		
((R, R), (L, L))	0.25	0.25	0.25	0.25
((R, R), (L, R))	0.5		0.5	
((R, R), (R, L))	0.5	0.5		
((R, R), (R, R))	1.0			

Similarly, we get the discounted state-action counts

$$\begin{aligned}
\eta(\boldsymbol{s} = \mathtt{L}, \boldsymbol{a} = (\mathtt{Li}, \mathtt{Li})) &= \Pr(S_0 = \mathtt{L}, \boldsymbol{A}_0 = (\mathtt{Li}, \mathtt{Li})) \\
&\quad + \gamma \Pr(S_1 = \mathtt{L}, \boldsymbol{A}_1 = (\mathtt{Li}, \mathtt{Li})) \\
&\quad + \gamma^2 \Pr(S_2 = \mathtt{L}, \boldsymbol{A}_2 = (\mathtt{Li}, \mathtt{Li})) \\
&= \frac{1+\gamma}{2}, && (98) \\
\eta(\boldsymbol{s} = \mathtt{L}, \boldsymbol{a} = (\mathtt{L}, \mathtt{L})) &= \gamma^2 \Pr(S_2 = \mathtt{L}, \boldsymbol{A}_2 = (\mathtt{L}, \mathtt{L})) \\
&= \gamma^2 \Pr(S_2 = \mathtt{L}) \Pr(\boldsymbol{A}_2 = (\mathtt{L}, \mathtt{L}) \mid S_2 = \mathtt{L}) \\
&= \gamma^2 \frac{1}{2} 0.0225, && (99) \\
\eta(\boldsymbol{s} = \mathtt{L}, \boldsymbol{a} = (\mathtt{L}, \mathtt{R})) &= \gamma^2 \Pr(S_2 = \mathtt{L}, \boldsymbol{A}_2 = (\mathtt{L}, \mathtt{R})) \\
&= \gamma^2 \Pr(S_2 = \mathtt{L}) \Pr(\boldsymbol{A}_2 = (\mathtt{L}, \mathtt{R}) \mid S_2 = \mathtt{L}) \\
&= \gamma^2 \frac{1}{2} 0.1275, && (100) \\
\eta(\boldsymbol{s} = \mathtt{L}, \boldsymbol{a} = (\mathtt{R}, \mathtt{L})) &= \gamma^2 \Pr(S_2 = \mathtt{L}, \boldsymbol{A}_2 = (\mathtt{R}, \mathtt{L})) \\
&= \gamma^2 \Pr(S_2 = \mathtt{L}) \Pr(\boldsymbol{A}_2 = (\mathtt{R}, \mathtt{L}) \mid S_2 = \mathtt{L}) \\
&= \gamma^2 \frac{1}{2} 0.1275, && (101) \\
\eta(\boldsymbol{s} = \mathtt{L}, \boldsymbol{a} = (\mathtt{R}, \mathtt{R})) &= \gamma^2 \Pr(S_2 = \mathtt{L}, \boldsymbol{A}_2 = (\mathtt{R}, \mathtt{R})) \\
&= \gamma^2 \Pr(S_2 = \mathtt{L}) \Pr(\boldsymbol{A}_2 = (\mathtt{R}, \mathtt{R}) \mid S_2 = \mathtt{L}) \\
&= \gamma^2 \frac{1}{2} \overset{341}{0.7225}, && (102)
\end{aligned}$$

Table 5: $\Pr(\boldsymbol{H}_2 = \boldsymbol{h} \mid S_2 = s) = 0.85^n \cdot 0.15^{2-n}$, where n is the number of times the tiger was heard through the correct door. In this table, a joint observation $\boldsymbol{h} = ((A, B), (C, D))$ means that the two agents have received observations A and B on the first timestep, and C and D on the second timestep.

s	L	R
$\boldsymbol{h}$		
((L, L), (L, L))	0.522	0.001
((L, L), (L, R))	0.092	0.003
((L, L), (R, L))	0.092	0.003
((L, L), (R, R))	0.016	0.016
((L, R), (L, L))	0.092	0.003
((L, R), (L, R))	0.016	0.016
((L, R), (R, L))	0.016	0.016
((L, R), (R, R))	0.003	0.092
((R, L), (L, L))	0.092	0.003
((R, L), (L, R))	0.016	0.016
((R, L), (R, L))	0.016	0.016
((R, L), (R, R))	0.003	0.092
((R, R), (L, L))	0.016	0.016
((R, R), (L, R))	0.003	0.092
((R, R), (R, L))	0.003	0.092
((R, R), (R, R))	0.001	0.552

Table 6: $\Pr(\boldsymbol{A}_2 = \boldsymbol{a} \mid S_2 = s) = \sum_{\boldsymbol{h}} \Pr(\boldsymbol{A}_2 = \boldsymbol{a} \mid \boldsymbol{H}_2 = \boldsymbol{h}) \Pr(\boldsymbol{H}_2 = \boldsymbol{h} \mid S_2 = s)$.

s	L	R
$\boldsymbol{a}$		
(L, L)	0.0225	0.7225
(L, R)	0.1275	0.1275
(R, L)	0.1275	0.1275
(R, R)	0.7225	0.0225

Table 7: $\Pr(\boldsymbol{h}, \boldsymbol{a} \mid h_1, a_1)$. In this table, a joint observation $\boldsymbol{h} = ((A, B), (C, D))$ means that the two agents have received observations A and B on the first timestep, and C and D on the second timestep.

$\boldsymbol{a}$	(R, L)	(R, R)
$\boldsymbol{h}$		
((L, L), (L, L))	0.0	0.7014
((L, L), (L, R))	0.0638	0.0638
((L, R), (L, L))	0.0638	0.0638
((L, R), (L, R))	0.0436	0.0

Table 8: $\Pr(s, \boldsymbol{a} \mid h_1, a_1)$.

s	$\boldsymbol{a}$	$\Pr(s, \boldsymbol{a} \mid h_1, a_1)$
L	(R, L)	0.145
L	(R, R)	0.824
R	(R, L)	0.026
R	(R, R)	0.005

and the discounted distributions

$$\rho(\boldsymbol{a} = (\texttt{Li}, \texttt{Li}) \mid s = L) = \frac{1+\gamma}{1+\gamma+\gamma^2} \tag{103}$$

$$\rho(\boldsymbol{a} = (\texttt{L}, \texttt{L}) \mid s = L) = \frac{0.0225\gamma^2}{1+\gamma+\gamma^2} \tag{104}$$

$$\rho(\boldsymbol{a} = (\texttt{L}, \texttt{R}) \mid s = L) = \frac{0.1275\gamma^2}{1+\gamma+\gamma^2} \tag{105}$$

$$\rho(\boldsymbol{a} = (\texttt{R}, \texttt{L}) \mid s = L) = \frac{0.1275\gamma^2}{1+\gamma+\gamma^2} \tag{106}$$

$$\rho(\boldsymbol{a} = (\texttt{R}, \texttt{R}) \mid s = L) = \frac{0.7225\gamma^2}{1+\gamma+\gamma^2} \tag{107}$$

C.2 State Values

The state values of actions L and R are straightforward due to the episode terminating immediately, and are simply the respective rewards,

$$Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{L}, \texttt{L})) = -50\,, \tag{108}$$

$$Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{L}, \texttt{R})) = -100\,, \tag{109}$$

$$Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{R}, \texttt{L})) = -100\,, \tag{110}$$

$$Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{R}, \texttt{R})) = 20\,. \tag{111}$$

On the other hand, the state value associated with both agents listening is obtained using all of the previously calculated probabilities, and a recursive relation,

$$\begin{aligned} Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) &= -2 + \gamma \sum_{\boldsymbol{a}'} \rho(\boldsymbol{a}' \mid s' = \texttt{L}) Q^\pi(s' = \texttt{L}, \boldsymbol{a}') \\ &= -2 + \frac{0.0225\gamma^3(-50)}{1+\gamma+\gamma^2} + \frac{0.1275\gamma^3(-100)}{1+\gamma+\gamma^2} \\ &\quad + \frac{0.1275\gamma^3(-100)}{1+\gamma+\gamma^2} + \frac{0.7225\gamma^3(20)}{1+\gamma+\gamma^2} \\ &\quad + \frac{\gamma+\gamma^2}{1+\gamma+\gamma^2} Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) \\ &= -2 - 12.175 \frac{\gamma^3}{1+\gamma+\gamma^2} \\ &\quad + \frac{\gamma+\gamma^2}{1+\gamma+\gamma^2} Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li}))\,, \end{aligned} \tag{112}$$

then, fixing $\gamma = 1.0$,

$$Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) = -\frac{18.175}{3} + \frac{2}{3} Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li}))\,, \tag{113}$$

which results in

$$Q^\pi(s = \texttt{L}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) = -18.175\,. \tag{114}$$

By the symmetries of the Dec-Tiger problem and the given policies, we similarly get

$$Q^\pi(s = \texttt{R}, \boldsymbol{a} = (\texttt{L}, \texttt{L})) = 20\,, \tag{115}$$
$$Q^\pi(s = \texttt{R}, \boldsymbol{a} = (\texttt{L}, \texttt{R})) = -100\,, \tag{116}$$
$$Q^\pi(s = \texttt{R}, \boldsymbol{a} = (\texttt{R}, \texttt{L})) = -100\,, \tag{117}$$
$$Q^\pi(s = \texttt{R}, \boldsymbol{a} = (\texttt{R}, \texttt{R})) = -50\,. \tag{118}$$
$$Q^\pi(s = \texttt{R}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) = -18.175\,. \tag{119}$$

C.3 History-State Values

Here, we compute the history-state values $Q^\pi(\boldsymbol{h}, s, \boldsymbol{a})$ for some relevant combinations of history, state and actions. We begin by noting that, because all door-opening combinations lead to the episode terminating, then each respective value is invariant to the given history, and simply equal to the respective reward,

$$Q^\pi(\boldsymbol{h}, s, \boldsymbol{a}) = R(s, \boldsymbol{a}) \qquad \forall \boldsymbol{a} \in \{(\texttt{L}, \texttt{L}), (\texttt{L}, \texttt{R}), (\texttt{R}, \texttt{L}), (\texttt{R}, \texttt{R})\}\,. \tag{120}$$

Then, we compute the values associated with the beginning of the episode, i.e., with the empty joint history $\boldsymbol{h} = \epsilon$. Because the second action for both agents will be to listen again, and since the state will not change, we can quickly compute the initial history-state values

as,

$$\begin{aligned}
&Q^\pi(\boldsymbol{h} = \epsilon, s = \texttt{L}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) \\
&= -2 - 2\gamma + \gamma^2 \sum_{\boldsymbol{h}} \Pr(\boldsymbol{H}_2 = \boldsymbol{h} \mid S_2 = \texttt{L}) \sum_{\boldsymbol{a}} \Pr(\boldsymbol{A}_2 = \boldsymbol{a} \mid \boldsymbol{H}_2 = \boldsymbol{h}) Q^\pi(\boldsymbol{h}, s = \texttt{L}, \boldsymbol{a} = \boldsymbol{a}) \\
&= -2 - 2\gamma + \gamma^2 \sum_{\boldsymbol{a}} \Pr(\boldsymbol{A}_2 = \boldsymbol{a} \mid S_2 = \texttt{L}) R(s = \texttt{L}, \boldsymbol{a} = \boldsymbol{a}) \\
&= -2 - 2\gamma + \gamma^2 \left(0.0225 \cdot (-50) + 0.1275 \cdot (-100) + 0.1275 \cdot (-100) + 0.7225 \cdot 20\right) \\
&= -2 - 2\gamma - 12.175\gamma^2 \,, \qquad (121)
\end{aligned}$$

then, fixing $\gamma = 1.0$,

$$= -16.175 \,. \qquad (122)$$

By the symmetries of the Dec-Tiger problem and the given policies, we similarly get

$$Q^\pi(\boldsymbol{h} = \epsilon, s = \texttt{R}, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) = -16.175 \,. \qquad (123)$$

C.4 History Values

Finally, we use the equivalence from Lemma 2 to compute history values $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$. For door-opening actions, we obtain

$$\begin{aligned}
Q^\pi(\boldsymbol{h}, \boldsymbol{a}) &= \mathbb{E}_{s|\boldsymbol{h}} \left[Q^\pi(\boldsymbol{h}, s, \boldsymbol{a}) \right] \,, \\
&= \mathbb{E}_{s|\boldsymbol{h}} \left[R(s, \boldsymbol{a}) \right] \qquad \forall \boldsymbol{a} \in \{(\texttt{L}, \texttt{L}), (\texttt{L}, \texttt{R}), (\texttt{R}, \texttt{L}), (\texttt{R}, \texttt{R})\} \,. \qquad (124)
\end{aligned}$$

The state probabilities $\Pr(s \mid \boldsymbol{h})$ are computed and shown in Table 9, and the respective history values are shown in Table 10. Finally, we also compute the history value associated with the beginning of the episode, i.e., with the empty joint history $\boldsymbol{h} = \epsilon$, for which trivially $\Pr(s \mid \boldsymbol{h} = \epsilon) = \frac{1}{2}$,

$$\begin{aligned}
Q^\pi(\boldsymbol{h} = \epsilon, \boldsymbol{a} = (\texttt{Li}, \texttt{Li}) &= \frac{1}{2} \sum_{s \in \{\texttt{L}, \texttt{R}\}} Q^\pi(\boldsymbol{h} = \epsilon, s, \boldsymbol{a} = (\texttt{Li}, \texttt{Li})) \\
&= -16.175 \,. \qquad (125)
\end{aligned}$$

Table 9: $\Pr(s \mid \boldsymbol{h}) \propto \Pr(\boldsymbol{h} \mid s) \Pr(s)$, which can be calculated from the values in Table 5. In this table, a joint observation $\boldsymbol{h} = ((A, B), (C, D))$ means that the two agents have received observations A and B on the first timestep, and C and D on the second timestep.

s	L	R
$\boldsymbol{h}$		
$((\mathtt{L}, \mathtt{L}), (\mathtt{L}, \mathtt{L}))$	0.999	0.001
$((\mathtt{L}, \mathtt{L}), (\mathtt{L}, \mathtt{R}))$	0.970	0.030
$((\mathtt{L}, \mathtt{L}), (\mathtt{R}, \mathtt{L}))$	0.970	0.030
$((\mathtt{L}, \mathtt{L}), (\mathtt{R}, \mathtt{R}))$	0.500	0.500
$((\mathtt{L}, \mathtt{R}), (\mathtt{L}, \mathtt{L}))$	0.970	0.030
$((\mathtt{L}, \mathtt{R}), (\mathtt{L}, \mathtt{R}))$	0.500	0.500
$((\mathtt{L}, \mathtt{R}), (\mathtt{R}, \mathtt{L}))$	0.500	0.500
$((\mathtt{L}, \mathtt{R}), (\mathtt{R}, \mathtt{R}))$	0.030	0.970
$((\mathtt{R}, \mathtt{L}), (\mathtt{L}, \mathtt{L}))$	0.970	0.030
$((\mathtt{R}, \mathtt{L}), (\mathtt{L}, \mathtt{R}))$	0.500	0.500
$((\mathtt{R}, \mathtt{L}), (\mathtt{R}, \mathtt{L}))$	0.500	0.500
$((\mathtt{R}, \mathtt{L}), (\mathtt{R}, \mathtt{R}))$	0.030	0.970
$((\mathtt{R}, \mathtt{R}), (\mathtt{L}, \mathtt{L}))$	0.500	0.500
$((\mathtt{R}, \mathtt{R}), (\mathtt{L}, \mathtt{R}))$	0.030	0.970
$((\mathtt{R}, \mathtt{R}), (\mathtt{R}, \mathtt{L}))$	0.030	0.970
$((\mathtt{R}, \mathtt{R}), (\mathtt{R}, \mathtt{R}))$	0.001	0.999

Table 10: $Q^{\pi}(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s|h}\left[R(s, \boldsymbol{a})\right] \qquad \forall \boldsymbol{a} \in \{(\mathtt{L}, \mathtt{L}), (\mathtt{L}, \mathtt{R}), (\mathtt{R}, \mathtt{L}), (\mathtt{R}, \mathtt{R})\}$. In this table, a joint observation $\boldsymbol{h} = ((A, B), (C, D))$ means that the two agents have received observations A and B on the first timestep, and C and D on the second timestep.

$\boldsymbol{a}$	$(\mathtt{L}, \mathtt{L})$	$(\mathtt{L}, \mathtt{R})$	$(\mathtt{R}, \mathtt{L})$	$(\mathtt{R}, \mathtt{R})$
$\boldsymbol{h}$				
$((\mathtt{L}, \mathtt{L}), (\mathtt{L}, \mathtt{L}))$	-49.93	-100.00	-100.00	19.93
$((\mathtt{L}, \mathtt{L}), (\mathtt{L}, \mathtt{R}))$	-47.89	-100.00	-100.00	17.89
$((\mathtt{L}, \mathtt{L}), (\mathtt{R}, \mathtt{L}))$	-47.89	-100.00	-100.00	17.89
$((\mathtt{L}, \mathtt{L}), (\mathtt{R}, \mathtt{R}))$	-15.00	-100.00	-100.00	-15.00
$((\mathtt{L}, \mathtt{R}), (\mathtt{L}, \mathtt{L}))$	-47.89	-100.00	-100.00	17.89
$((\mathtt{L}, \mathtt{R}), (\mathtt{L}, \mathtt{R}))$	-15.00	-100.00	-100.00	-15.00
$((\mathtt{L}, \mathtt{R}), (\mathtt{R}, \mathtt{L}))$	-15.00	-100.00	-100.00	-15.00
$((\mathtt{L}, \mathtt{R}), (\mathtt{R}, \mathtt{R}))$	17.89	-100.00	-100.00	-47.89
$((\mathtt{R}, \mathtt{L}), (\mathtt{L}, \mathtt{L}))$	-47.89	-100.00	-100.00	17.89
$((\mathtt{R}, \mathtt{L}), (\mathtt{L}, \mathtt{R}))$	-15.00	-100.00	-100.00	-15.00
$((\mathtt{R}, \mathtt{L}), (\mathtt{R}, \mathtt{L}))$	-15.00	-100.00	-100.00	-15.00
$((\mathtt{R}, \mathtt{L}), (\mathtt{R}, \mathtt{R}))$	17.89	-100.00	-100.00	-47.89
$((\mathtt{R}, \mathtt{R}), (\mathtt{L}, \mathtt{L}))$	-15.00	-100.00	-100.00	-15.00
$((\mathtt{R}, \mathtt{R}), (\mathtt{L}, \mathtt{R}))$	17.89	-100.00	-100.00	-47.89
$((\mathtt{R}, \mathtt{R}), (\mathtt{R}, \mathtt{L}))$	17.89	-100.00	-100.00	-47.89
$((\mathtt{R}, \mathtt{R}), (\mathtt{R}, \mathtt{R}))$	19.93	-100.00	-100.00	-49.93

Appendix D. Pseudocode

We note that practical implementations of Actor-Critic methods do not model the Q-values Q^π directly using a Q-model $\hat{Q}$, but rather using a bootstrapped V-model $\hat{V}$, e.g., using 1-step returns $\hat{Q}_t \approx r_t + \gamma \hat{V}_{t+1}$ (other generalizations like n-step returns or lambda-returns are also common). Hence, although the theory of Actor-Critic is based on Q-values and models, the pseudocode shown in this section employs V-values and models.

Algorithm 1 IAC

Require: Individual actor models $\pi_i(a; h)$, parameterized by θ_i
Require: Individual critic models $\hat{V}_i(h)$, parameterized by ϑ_i
loop
  Sample environment episode $(\boldsymbol{o}_0, \boldsymbol{a}_0, r_0, \boldsymbol{o}_1, \boldsymbol{a}_1, r_1, \dots, \boldsymbol{o}_{T-1}, \boldsymbol{a}_{T-1}, r_{T-1})$
  Denote individual observed histories as $h_{i,t} \leftarrow (o_{i,0}, a_{i,0}, \dots o_{i,t-1}, a_{i,t-1}, o_{i,t}), \forall i, t$
  for each agent i **do**
    Compute individual advantage estimates $\delta_{i,t} \leftarrow r_t + \gamma \hat{V}_i(h_{i,t+1}) - \hat{V}_i(h_{i,t}), \forall t$
    Compute actor gradient estimate $\sum_{t=0}^{T-1} \gamma^t \delta_{i,t} \nabla \log \pi_i(a_{i,t}; h_{i,t})$
    Update actor parameters θ_i using gradient estimate
    Compute critic gradient estimate $\frac{1}{T} \sum_{t=0}^{T-1} \delta_{i,t} \nabla \hat{V}_i(h_{i,t})$
    Update critic parameters ϑ_i using gradient estimate
  end for
end loop

Algorithm 2 IACC-H

Require: Individual actor models $\pi_i(a; h)$, parameterized by θ_i
Require: Centralized critic model $\hat{V}(\boldsymbol{h})$, parameterized by ϑ
loop
  Sample environment episode $(\boldsymbol{o}_0, \boldsymbol{a}_0, r_0, \boldsymbol{o}_1, \boldsymbol{a}_1, r_1, \dots, \boldsymbol{o}_{T-1}, \boldsymbol{a}_{T-1}, r_{T-1})$
  Denote joint observed histories as $\boldsymbol{h}_t \leftarrow (\boldsymbol{o}_0, \boldsymbol{a}_0, \dots \boldsymbol{o}_{t-1}, \boldsymbol{a}_{t-1}, \boldsymbol{o}_t), \forall t$
  Denote individual observed histories as $h_{i,t} \leftarrow (o_{i,0}, a_{i,0}, \dots o_{i,t-1}, a_{i,t-1}, o_{i,t}), \forall i, t$
  Compute centralized advantage estimates $\delta_t \leftarrow r_t + \gamma \hat{V}(\boldsymbol{h}_{t+1}) - \hat{V}(\boldsymbol{h}_t), \forall t$
  for each agent i **do**
    Compute actor gradient estimate $\sum_{t=0}^{T-1} \gamma^t \delta_t \nabla \log \pi_i(a_{i,t}; h_{i,t})$
    Update actor parameters θ_i using gradient estimate
  end for
  Compute critic gradient estimate $\frac{1}{T} \sum_{t=0}^{T-1} \delta_t \nabla \hat{V}(\boldsymbol{h}_t)$
  Update critic parameters ϑ using gradient estimate
end loop

Algorithm 3 IACC-S

Require: Individual actor models $\pi_i(a; h)$, parameterized by θ_i
Require: Centralized critic model $\hat{V}(\boldsymbol{s})$, parameterized by ϑ
loop
  Sample environment episode $(\boldsymbol{s}_0, \boldsymbol{o}_0, \boldsymbol{a}_0, r_0, \boldsymbol{s}_1, \boldsymbol{o}_1, \boldsymbol{a}_1, r_1, \ldots, \boldsymbol{s}_{T-1}, \boldsymbol{o}_{T-1}, \boldsymbol{a}_{T-1}, r_{T-1})$
  Denote individual observed histories as $h_{i,t} \leftarrow (o_{i,0}, a_{i,0}, \ldots o_{i,t-1}, a_{i,t-1}, o_{i,t}), \forall i, t$
  Compute centralized advantage estimates $\delta_t \leftarrow r_t + \gamma \hat{V}(\boldsymbol{s}_{t+1}) - \hat{V}(\boldsymbol{s}_t), \forall t$
  for each agent i **do**
    Compute actor gradient estimate $\sum_{t=0}^{T-1} \gamma^t \delta_t \nabla \log \pi_i(a_{i,t}; h_{i,t})$
    Update actor parameters θ_i using gradient estimate
  end for
  Compute critic gradient estimate $\frac{1}{T} \sum_{t=0}^{T-1} \delta_t \nabla \hat{V}(\boldsymbol{s}_t)$
  Update critic parameters ϑ using gradient estimate
end loop

Algorithm 4 IACC-HS

Require: Individual actor models $\pi_i(a; h)$, parameterized by θ_i
Require: Centralized critic model $\hat{V}(\boldsymbol{h}, \boldsymbol{s})$, parameterized by ϑ
loop
  Sample environment episode $(\boldsymbol{s}_0, \boldsymbol{o}_0, \boldsymbol{a}_0, r_0, \boldsymbol{s}_1, \boldsymbol{o}_1, \boldsymbol{a}_1, r_1, \ldots, \boldsymbol{s}_{T-1}, \boldsymbol{o}_{T-1}, \boldsymbol{a}_{T-1}, r_{T-1})$
  Denote joint observed histories as $\boldsymbol{h}_t \leftarrow (\boldsymbol{o}_0, \boldsymbol{a}_0, \ldots \boldsymbol{o}_{t-1}, \boldsymbol{a}_{t-1}, \boldsymbol{o}_t), \forall t$
  Denote individual observed histories as $h_{i,t} \leftarrow (o_{i,0}, a_{i,0}, \ldots o_{i,t-1}, a_{i,t-1}, o_{i,t}), \forall i, t$
  Compute centralized advantage estimates $\delta_t \leftarrow r_t + \gamma \hat{V}(\boldsymbol{h}_{t+1}, \boldsymbol{s}_{t+1}) - \hat{V}(\boldsymbol{h}_t, \boldsymbol{s}_t), \forall t$
  for each agent i **do**
    Compute actor gradient estimate $\sum_{t=0}^{T-1} \gamma^t \delta_t \nabla \log \pi_i(a_{i,t}; h_{i,t})$
    Update actor parameters θ_i using gradient estimate
  end for
  Compute critic gradient estimate $\frac{1}{T} \sum_{t=0}^{T-1} \delta_t \nabla \hat{V}(\boldsymbol{h}_t, \boldsymbol{s}_t)$
  Update critic parameters ϑ using gradient estimate
end loop

References

Amato, C., Dibangoye, J. S., & Zilberstein, S. (2009). Incremental policy generation for finite-horizon DEC-POMDPs. In *Proceedings of the International Conference on Automated Planning and Scheduling*, pp. 2–9. AAAI Press.

Amato, C., Bernstein, D. S., & Zilberstein, S. (2007). Optimizing memory-bounded controllers for decentralized POMDPs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Baisero, A., & Amato, C. (2022). Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020). Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*.

Bernstein, D. S., Hansen, E. A., & Zilberstein, S. (2005). Bounded policy iteration for decentralized POMDPs. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 52–57.

Bono, G., Dibangoye, J. S., Matignon, L., Pereyron, F., & Simonin, O. (2018). Cooperative multi-agent policy gradient. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 459–476. Springer.

Bowling, M., & Veloso, M. (2001). Convergence of gradient dynamics with a variable learning rate. In *Proceedings of the International Conference on Machine Learning*, pp. 27–34.

Chakravorty, J., Ward, P. N., Roy, J., Chevalier-Boisvert, M., Basu, S., Lupu, A., & Precup, D. (2020). Option-critic in cooperative multi-agent systems. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1792–1794.

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 746–752.

Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., & Pineau, J. (2019). TarMAC: Targeted multi-agent communication. In *Proceedings of the International Conference on Machine Learning*, pp. 1538–1546.

de Witt, C. S., Peng, B., Kamienny, P.-A., Torr, P., Böhmer, W., & Whiteson, S. (2020). Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709*, *19*.

Du, Y., Han, L., Fang, M., Dai, T., Liu, J., & Tao, D. (2019). LIIR: Learning individual intrinsic reward in multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*.

Foerster, J., Assael, I. A., De Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2137–2145.

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

Fulda, N., & Ventura, D. (2007). Predicting and preventing coordination problems in cooperative Q-learning systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 780–785.

Iqbal, S., & Sha, F. (2019). Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, Vol. 97, pp. 2961–2970.

Jiang, J., & Lu, Z. (2018). Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, pp. 7254–7264.

Jiang, S. (2019). Multi-agent reinforcement learning environments compilation. `https://github.com/Bigpig4396/Multi-Agent-Reinforcement-Learning-Environment`.

Konda, V. R., & Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pp. 1008–1014.

Lee, H.-R., & Lee, T. (2019). Improved cooperative multi-agent reinforcement learning algorithm augmented by mixing demonstrations from centralized policy. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1089–1098.

Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., & Russell, S. (2019). Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 4213–4220.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, Vol. 30.

Lyu, X., & Amato, C. (2020). Likelihood quantile networks for coordinating multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 798–806.

Lyu, X., Baisero, A., Xiao, Y., & Amato, C. (2022). A deeper understanding of state-based critics in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 9396–9404.

Lyu, X., Xiao, Y., Daley, B., & Amato, C. (2021). Contrasting centralized and decentralized critics in multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 844–852.

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

Matignon, L., Laurent, G. J., & Le Fort-Piat, N. (2012). Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, *27*(1), 1–31.

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 1495–1502. AAAI Press.

Nair, R., Tambe, M., Yokoo, M., Pynadath, D., & Marsella, S. (2003). Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Vol. 3, pp. 705–711.

Oliehoek, F. A., & Amato, C. (2016). *A Concise Introduction to Decentralized POMDPs*. Springer.

Oliehoek, F. A., Spaan, M. T., & Vlassis, N. (2008). Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, *32*, 289–353.

Omidshafiei, S., Kim, D.-K., Liu, M., Tesauro, G., Riemer, M., Amato, C., Campbell, M., & How, J. P. (2019). Learning to teach in cooperative multiagent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6128–6136.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., & Vian, J. (2017). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *Proceedings of the International Conference on Machine Learning*, pp. 2681–2690.

Panait, L., Tuyls, K., & Luke, S. (2008). Theoretical advantages of lenient Q-learners: An evolutionary game theoretic perspective. *Journal of Machine Learning Research*, *9*(Mar), 423–457.

Peng, B., Rashid, T., Schroeder de Witt, C., Kamienny, P.-A., Torr, P., Böhmer, W., & Whiteson, S. (2021). Facmac: Factored multi-agent centralised policy gradients. In *Advances in Neural Information Processing Systems*, pp. 12208–12221.

Peshkin, L., Kim, K.-E., Meuleau, N., & Kaelbling, L. P. (2000). Learning to cooperate via policy search. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 489–496.

Rashid, T., Farquhar, G., Peng, B., & Whiteson, S. (2020). Weighted QMIX: Expanding monotonic value function factorisation. In *Advances in Neural Information Processing Systems*.

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.

Samvelyan, M., Rashid, T., de Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J., & Whiteson, S. (2019). The StarCraft multi-agent challenge. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 2148–2150.

Schroeder de Witt, C., Foerster, J., Farquhar, G., Torr, P., Boehmer, W., & Whiteson, S. (2019). Multi-agent common knowledge reinforcement learning. In *Advances in Neural Information Processing Systems*, Vol. 32, pp. 9927–9939.

Simões, D., Lau, N., & Reis, L. P. (2020). Multi-agent actor centralized-critic with communication. *Neurocomputing*, *390*, 40–56.

Singh, S. P., Kearns, M. J., & Mansour, Y. (2000). Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 541–548.

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019). QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896. PMLR.

Su, J., Adams, S., & Beling, P. A. (2021). Value-decomposition multi-agent actor-critics. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. (2018). Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 2085–2087.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second edition). The MIT Press.

Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063.

Wang, J., Zhang, Y., Kim, T.-K., & Gu, Y. (2020a). Shapley Q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, p. 7285–7292.

Wang, T., Dong, H., & Victor Lesser, C. Z. (2020b). ROMA: Multi-agent reinforcement learning with emergent roles. In *Proceedings of the International Conference on Machine Learning*.

Wang, T., Wang, J., Zheng, C., & Zhang, C. (2020c). Learning nearly decomposable value functions via communication minimization. In *International Conference on Learning Representations*.

Wang, W., Hao, J., Wang, Y., & Taylor, M. (2019). Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas. In *Proceedings of the International Conference on Distributed Artificial Intelligence*, pp. 1–7.

Wang, Y., Han, B., Wang, T., Dong, H., & Zhang, C. (2021). Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations*.

Xiao, Y., Hoffman, J., & Amato, C. (2019). Macro-action-based deep multi-agent reinforcement learning. In *Conference on Robot Learning*.

Xiao, Y., Hoffman, J., Xia, T., & Amato, C. (2020). Learning multi-robot decentralized macro-action-based policies via a centralized Q-net. In *Proceedings of the International Conference on Robotics and Automation*.

Yang, J., Nakhaei, A., Isele, D., Fujimura, K., & Zha, H. (2020). CM3: Cooperative multi-goal multi-stage multi-agent reinforcement learning. In *International Conference on Learning Representations*.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24611–24624.

Zhang, C., & Lesser, V. (2010). Multi-agent learning with policy prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24.

Zhou, M., Liu, Z., Sui, P., Li, Y., & Chung, Y. Y. (2020). Learning implicit credit assignment for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*.