A Deeper Understanding of State-Based Critics in Multi-Agent Reinforcement Learning

Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Christopher Amato

Northeastern University
{lu.xue, baisero.a, xiao.yuch, c.amato}@northeastern.edu

Abstract

Centralized Training for Decentralized Execution, where training is done in a centralized offline fashion, has become a popular solution paradigm in Multi-Agent Reinforcement Learning. Many such methods take the form of actor-critic with state-based critics, since centralized training allows access to the true system state, which can be useful during training despite not being available at execution time. State-based critics have become a common empirical choice, albeit one which has had limited theoretical justification or analysis. In this paper, we show that state-based critics can introduce bias in the policy gradient estimates, potentially undermining the asymptotic guarantees of the algorithm. We also show that, even if the state-based critics do not introduce any bias, they can still result in a larger gradient variance, contrary to the common intuition. Finally, we show the effects of the theories in practice by comparing different forms of centralized critics on a wide range of common benchmarks, and detail how various environmental properties are related to the effectiveness of different types of critics.

1 Introduction

The rising popularity of Centralized Critics in Multi-Agent Reinforcement Learning (MARL) has led to the usage of state information becoming common practice. The rationale behind state-based critics is straightforward—centralized critics train in a centralized offline manner and usually have access to the environment state, which appears desirable compared to local observations. Because critics can be discarded after the training of actors, they do not hinder independent execution by each agent. As a result, state-based critics have become a convenient and popular design decision for centralized training with actor-critic methods (Foerster et al. 2018; Wang et al. 2020; Du et al. 2019; Zhou et al. 2020; Su, Adams, and Beling 2021; Du et al. 2021; Schroeder de Witt et al. 2019; Baker et al. 2020; Wang et al. 2021). However, state-based critics have received little formal analysis, making the state-based critic's benefits and downsides insufficiently understood in the field. In this paper, we point out that there exists misconceptions in the field towards state-based critics. Our primary contribution is to fill this knowledge gap by providing both intuition and analysis of theoretical properties of state-based critics, complimented with empirical findings and suggestions.

First, we show how the state-based critic is not entirely sound in theory. We provide bias analysis to conclude that the state-based critic may incur unbounded bias on the policy gradient compared to the asymptotically unbiased history-based critic. Second, even assuming that the state-based critic is unbiased, we analyze policy gradient variance and show that the policy gradient variance with the state-based critic cannot be less than that of the history-based critic. We also give trivial and intuitive toy examples that highlight the essence of our arguments and provide an intuitive understanding of the theories.

We also compare the empirical performance of state-based and history-based critics. Supported by a wide array of experiments, we also discuss the implications of our theories in practice. We highlight particular circumstances regarding reactive policies, information gathering, and representation learning. We demonstrate how the effectiveness of different types of critics depends on observation models of the tasks and highlight the deficiencies of popular benchmarks. We also test an alternative critic that combines state and history information, which demonstrates reliability across tasks. Through linking empirical results with theory, we establish a deeper understanding of where the state-based critics work and where they fall short.

2 Related Work

Centralized Training for Decentralized Execution (CTDE) (Oliehoek, Spaan, and Vlassis 2008; Foerster et al. 2016) has shown significant benefits for learning decentralized policies by addressing environmental non-stationarity that emerges in independent learning methods. CTDE provides access to global information during training while having decentralized execution. Learning decentralized actors with a centralized critic has thus arisen as a direct and widespread use of the CTDE framework. There are many variants of the information used by the centralized critic, such as the environment state, joint observation, joint action-observation history, or even mixed combinations.

COMA (Foerster et al. 2018), as the first deep multi-agent actor-critic-based algorithm with CTDE, learns a centralized critic to provide joint Q-value estimations with explicit ac-

cess to ground truth state information. MADDPG (Lowe et al. 2017), the other pioneer, also uses centralized critics to update decentralized actors. In MADDPG, each critic accesses the joint-observation, which implicitly results in it being a state-based critic as the considered multi-agent domains are fully observable. We include a deeper dive into the implicit assumptions and the scope of the work of COMA and MADDPG in Appendix A.

The impressive performance achieved by COMA in SMAC and MADDPG in OpenAI Particle environments caused many future methods to use state-based critics without further study. For example, numerous other popular works such as SQDDPG (Wang et al. 2020), LIIR (Du et al. 2019), LICA (Zhou et al. 2020),VDAC-mix (Su, Adams, and Beling 2021), DOP (Wang et al. 2021) and MACKRL (Schroeder de Witt et al. 2019) also utilize state-based centralized critics. Centralized state-based critic is also used in larger scale environments for emergence tool use (Baker et al. 2020).

The decision of whether critics condition on states or observations is often considered not of algorithmic importance; thereby, recent state-of-the-art methods treat state-based critics as an engineering decision rather than an algorithmic design (e.g., in DOP (Wang et al. 2021) which focuses on factorization). However, it is unclear whether state-based critics are strictly beneficial since there is little theoretical analysis about how a state-based centralized critic impacts decentralized policy optimization. Ablation comparisons over different centralized critic designs are necessary but missing in the current literature. In this paper, we fill this gap by providing a theoretical and practical analysis of state-based critics.

3 Preliminaries

3.1 Dec-POMDPs

Decentralized partially observable Markov decision processes (Dec-POMDPs) (Oliehoek and Amato 2016) are multi-agent cooperative sequential decision making problems. A Dec-POMDP is 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 $\mathbf{A} \doteq \times_{i\in\mathcal{I}} \mathcal{A}_i$, one per agent, a joint observation space $\mathbf{\Omega} \doteq \times_{i\in\mathcal{I}} \Omega_i$, one per agent, a state transition function $\mathcal{T}\colon \mathcal{S} \times \mathbf{A} \to \Delta\mathcal{S}$, a joint observation function $\mathcal{O}\colon \mathcal{S} \times \mathbf{A} \to \Delta\mathbf{\Omega}$, a joint reward function $\mathcal{R}\colon \mathcal{S} \times \mathbf{A} \times \mathcal{S} \to \mathbb{R}$, and a discount factor $\gamma \in [0, 1]$.

Control in Dec-POMDPs is performed by a set of policies $\boldsymbol{\pi} = \langle \pi_1, \ldots, \pi_{|\mathcal{I}|} \rangle$, one per agent, each representing a (possibly stochastic) mapping from the agent's past observations to its next action, $\pi_i\colon \mathcal{H}_i \to \Delta\mathcal{A}_i$. At each timestep t, a joint action $\boldsymbol{a} = \langle a_{1,t}, ..., a_{|\mathcal{I}|,t} \rangle$ is taken by the agents, each using their own individual history. As feedback from the system, a scalar reward $r_t = \mathcal{R}(\mathbf{s}_t, \boldsymbol{a}_t, \mathbf{s}_{t+1})$ is shared by all agents, and each agent receives a local observation $\langle o_{1,t}, \ldots, o_{|\mathcal{I}|,t} \rangle \sim \mathcal{O}(s_t, \boldsymbol{a}_t)$. The collective objective of all agents is that of maximizing the expected performance $J \doteq \mathbb{E}[G_0]$, where $G_t \doteq \sum_k \gamma^k r_{t+k}$ are the discounted sum of future rewards, also called returns.

We focus on finite horizon problems for theoretical analysis; note, however, that the finite horizon can be arbitrarily large enough to approximate episodic and infinite-horizon problems as well. We will consider value functions over histories, states and state-history pairs. We first introduce the *state-history value* functions (Bono et al. 2018) $Q^{\pi}(s, \boldsymbol{h}, \boldsymbol{a})$, it is later served as a link between state values and history values; it is defined as the expected return given the agents being in history $\boldsymbol{h}$ and state s and taking action $\boldsymbol{a}$:

$$Q^{\boldsymbol{\pi}}(s, \boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}\left[G \mid s, \boldsymbol{h}, \boldsymbol{a}\right] . \tag{1}$$

The history and state values are related to the state-history values via marginalization over the conditional on-policy distributions $\rho(s \mid \boldsymbol{h})$ and $\rho(\boldsymbol{h} \mid s)$, respectively (Sutton, Precup, and Singh 1999) (see Appendix B):

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) = \mathop{\mathbb{E}}_{s\sim\rho(s|\boldsymbol{h})}\left[G \mid s, \boldsymbol{h}, \boldsymbol{a}\right] = \mathop{\mathbb{E}}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{h}, \boldsymbol{a})\right] , \tag{2}$$

$$Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}|s)}\left[G \mid s, \boldsymbol{h}, \boldsymbol{a}\right] = \mathop{\mathbb{E}}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}|s)}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{h}, \boldsymbol{a})\right] . \tag{3}$$

3.2 Multi-Agent Actor Critic Methods

Actor Critic (AC) (Konda and Tsitsiklis 2000; Sutton et al. 2000) is a popular Policy Gradient (PG) method which involves the training of policy and critic models; we consider the centralized training case (Lowe et al. 2017; Bono et al. 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 often omit the model parameterization, when implicitly clear from context. To distinguish the critic models from the value functions that they are trained to model, we will denote them as $\hat{V}$. The policy gradient theorem states that the policy gradient follows the expected returns provided by the joint history values $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})$:

$$\nabla_i J_{\boldsymbol{h}} = \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\nabla_{\theta_i} \log \pi_i(a_i; h_i)\right] . \tag{4}$$

Similarly, COMA (Foerster et al. 2018) and MADDPG (Lowe et al. 2017) introduced policy gradient variants which employed state values $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$:

$$\nabla_i J_s = \mathop{\mathbb{E}}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[\mathop{\mathbb{E}}_{s\sim\rho(s|\boldsymbol{h})} Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\nabla_{\theta_i} \log \pi_i(a_i; h_i)\right] . \tag{5}$$

In either case, the values $Q^{\boldsymbol{\pi}}$ 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 $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ as $r + \gamma\hat{V}(s')$. In advantage actor critic, the critic model is further used as baseline for variance reduction, resulting in the following estimates:

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) - V^{\boldsymbol{\pi}}(\boldsymbol{h}) \approx r + \gamma\hat{V}(\boldsymbol{hao}) - \hat{V}(\boldsymbol{h}) , \tag{6}$$

$$Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) - V^{\boldsymbol{\pi}}(s) \approx r + \gamma\hat{V}(s') - \hat{V}(s) . \tag{7}$$

4 Bias and Variance of State Values and Gradients

In this section, we analyze the bias and variance resulting from the use of state value functions $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$. Without loss

of generality, we consider single-sample Monte Carlo estimates of Equations (4) and (5),

$$\hat{\nabla}_i J_{\boldsymbol{h}} = Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) \nabla \log \pi_i(a_i; h_i)\,, \tag{8}$$

$$\hat{\nabla}_i J_s = Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \nabla \log \pi_i(a_i; h_i)\,, \tag{9}$$

where $\boldsymbol{h}, s \sim \rho(\boldsymbol{h}, s)$ and $\boldsymbol{a} \sim \boldsymbol{\pi}(\boldsymbol{h})$. In Section 4.1, we show that the state-based gradient estimate $\hat{\nabla}_i J_s$ can be biased. In Section 4.2, we show that even if the state-based gradient estimate $\hat{\nabla}_i J_s$ is unbiased, it will have a variance equal to or higher than that of $\hat{\nabla}_i J_{\boldsymbol{h}}$.

In a related single-agent work (Baisero and Amato 2021), state values are found to be undefined due to the potential non-existent of history distributions in infinite-horizon case. Our analysis, on the other hand, assumes finite-horizon, so that history distributions and state values can be properly defined; and we show how even when the state values are properly defined, they are still not unbiased. In addition to bias, we will also provide comparisons on gradient variance. Note that the Q values for our analysis is the analytical return values, not the results of learned models. Representation learning is a separate problem discussed in the experimental section.

4.1 Bias Analysis

We begin by noting that the gradient given by the history-based critic $\hat{\nabla}_i J_{\boldsymbol{h}}$ (Equation (4)) has already been proven to be unbiased (Foerster et al. 2016; Lyu et al. 2021) due to the joint histories are Markov states for the history-MDP (Oliehoek and Amato 2016) in which the policy gradient theorem holds (Sutton et al. 2000). On the other hand, the gradient given by the state-based critic is:

$$\begin{aligned}\nabla_i J_s &= \mathbb{E}_{\boldsymbol{h}, s \sim \rho(\boldsymbol{h}, s), \boldsymbol{a} \sim \boldsymbol{\pi}(\boldsymbol{h})} \left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \nabla \log \pi_i(a_i; h_i) \right] \\ &= \mathbb{E}_{\boldsymbol{h} \sim \rho(\boldsymbol{h}), \boldsymbol{a} \sim \boldsymbol{\pi}(\boldsymbol{h})} \left[\mathbb{E}_{s \sim \rho(s \mid \boldsymbol{h})} \left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \right] \nabla \log \pi_i(a_i; h_i) \right] .\end{aligned} \tag{10}$$

Together with history-based gradient (Equation (4)), we see the state-value-based-gradient $\hat{\nabla}_i J_s$ is also unbiased iff

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s \sim \rho(s \mid \boldsymbol{h})} \left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \right] . \tag{11}$$

Therefore, the analysis of the bias of $\hat{\nabla}_i J_s$ can be performed indirectly through $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$. We adopt the methodology employed by prior work for the single-agent control case (Baisero and Amato 2021), and will treat $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ as an estimator of $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})$. Consequently, if $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ is unbiased, then $\hat{\nabla}_i J_s$ is also unbiased.

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

$$\rho(\boldsymbol{h} \mid s) = \frac{\rho(s, \boldsymbol{h})}{\sum_{s'} \rho(s', \boldsymbol{h})} = \frac{P^{\boldsymbol{\pi}}_{ij}}{\sum_{j'} P^{\boldsymbol{\pi}}_{ij'}} \tag{12}$$

$$\rho(s \mid \boldsymbol{h}) = \frac{\rho(s, \boldsymbol{h})}{\sum_{\boldsymbol{h}'} \rho(s, \boldsymbol{h}')} = \frac{P^{\boldsymbol{\pi}}_{ij}}{\sum_{i'} P^{\boldsymbol{\pi}}_{i'j}} \tag{13}$$

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

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) = \sum_s \rho(s \mid \boldsymbol{h}) Q^{\boldsymbol{\pi}}(s, \boldsymbol{h}, \boldsymbol{a}) = \sum_{j'} \frac{P^{\boldsymbol{\pi}}_{ij'}}{\sum_{j'} P^{\boldsymbol{\pi}}_{ij'}} Q^{\boldsymbol{\pi}}_{\boldsymbol{a}, ij'}\,, \tag{14}$$

$$Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = \sum_{\boldsymbol{h}} \rho(\boldsymbol{h} \mid s) Q^{\boldsymbol{\pi}}(s, \boldsymbol{h}, \boldsymbol{a}) = \sum_{i'} \frac{P^{\boldsymbol{\pi}}_{i'j}}{\sum_{i'} P^{\boldsymbol{\pi}}_{i'j}} Q^{\boldsymbol{\pi}}_{\boldsymbol{a}, i'j}\,. \tag{15}$$

On the other hand, the correct expected state value as listed in Equation (11) is obtained as

$$\begin{aligned}&\mathbb{E}_{s \sim \rho(s \mid h)} \left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \right] \\ &= \sum_s \rho(s \mid \boldsymbol{h}) Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \\ &= \sum_s \rho(s \mid \boldsymbol{h}) \sum_{\boldsymbol{h}'} \rho(\boldsymbol{h}' \mid s) Q^{\boldsymbol{\pi}}(s, \boldsymbol{h}', \boldsymbol{a}) \\ &= \sum_{j'} \frac{P^{\boldsymbol{\pi}}_{ij'}}{\sum_{j''} P^{\boldsymbol{\pi}}_{ij''}} \sum_{i'} \frac{P^{\boldsymbol{\pi}}_{i'j'}}{\sum_{i''} P^{\boldsymbol{\pi}}_{i''j'}} Q^{\boldsymbol{\pi}}_{\boldsymbol{a}, i'j'}\,.\end{aligned} \tag{16}$$

Note that $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})$ in Equation (14) only involves the elements on a specific row of both $Q^{\boldsymbol{\pi}}_{\boldsymbol{a}}$ and $P^{\boldsymbol{\pi}}$, while the expected state value in Equation (16) involves all elements of both $Q^{\boldsymbol{\pi}}_{\boldsymbol{a}}$ and $P^{\boldsymbol{\pi}}$. 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 $Q^{\boldsymbol{\pi}}_{\boldsymbol{a}}$ and $P^{\boldsymbol{\pi}}$ are not arbitrary, but are related by problem dynamics, policies, and state-history Bellman equations, which may still result in Equations (14) and (16) being numerically equivalent. However, we prove that this is not the case.

Theorem 1. *$Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ may be a biased estimate of $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})$ (proof in Appendix C.1).*

Corollary 1.1. *$\nabla_i J_s$ may be biased (proof in Appendix C.2).*

The example below serves as prove-by-example for Theorem 1. See Appendix C.1 for an alternative proof.

Example In classic Dec-POMDP domain Dec-Tiger (Nair et al. 2003), two agents face two doors, left and right; a tiger is randomly initialized behind one of the doors. The agents can *listen*, *open-left* or *open-right*. The *listen* action is used to detect the tiger location and produces either *hear-left* or *hear-right* which indicates the correct tiger location with probability 85%. The cost for *listen* is -2; the episode ends with a -50 penalty if the tiger door is opened by both agents, and $+20$ for the other door. If two different doors are opened by the two agents, the episode ends with a penalty of -100. If only one agent opens a door (the

other agent listens), the episode ends with -101 if the state is tiger and $+9$ otherwise. For the purpose of this example, we use a finite horizon of 3 for calculating the values. We use the optimal policy in which the agents listen twice and pick the most promising door. We focus on the history $h = \{(\textit{listen}, \textit{listen}), (\textit{hear\_right}, \textit{hear\_right})\}$ and actions for the agents $a = (\textit{listen}, \textit{listen})$ as our example. By solving analytically (see supplementary material), we come to the follow history and state values:

$$Q(h, a) \approx 13.8859 \tag{17}$$

$$Q(tiger\_left, a) = -16.175 \tag{18}$$

$$Q(tiger\_right, a) = -16.175 \tag{19}$$

That is, after each agent listens once and both hear the tiger on the right, listening again by both agents is has an estimated value of 13.8859. In contrast, the state-based estimates just represent the value achieved after visiting the state and taking the corresponding action. It is obvious that no matter with what probability our state values in Equations (18) and (19) are combined, we can only get a value of -16.175 which is a biased estimator of our history value in Equation (17). Note that our h contains a promising observation $(\textit{hear\_right}, \textit{hear\_right})$ which brings our confidence that the tiger is behind the right to very high (96.98%). However, as illustrated in Equation (16), the state values are averaged over different histories, good *and* bad. The reason why $Q(\textit{tiger\_right}, a)$ is negative is because it has to consider situations where agents get less promising observations (like $(\textit{hear\_left}, \textit{hear\_right})$) or no observations (at the beginning of the episode) while being in the corresponding state. That is, *history critics represent the true value of the history-based policy while state critics are averaged over all the histories that visit the state*. As a result, we see that in Dec-Tiger, the state values are biased estimators of history values.

4.2 Variance Analysis

In this section, we discuss the effect of state values $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ on the variance of the policy gradient estimate $\hat{\nabla}_i J_s$. Like our bias analysis, we discuss the oracle values Q^{π} instead of its estimated counterpart $\hat{Q}^{\pi}$. 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 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 $\hat{\nabla}_i J_s$ also has higher variance than $\hat{\nabla}_i J_{\boldsymbol{h}}$ in general. Instead, we show that, if $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ is unbiased for a given policy and a Dec-POMDP, then $\hat{\nabla}_i J_s$ has a variance greater or equal than that of $\hat{\nabla}_i J_{\boldsymbol{h}}$.

Theorem 2. *When the state value function $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ is unbiased, the state-based policy gradient estimates $\hat{\nabla}_i J_s$ have a variance greater or equal than that of the history-based policy gradient estimates $\hat{\nabla}_i J_{\boldsymbol{h}}$, i.e.,*

$$\begin{aligned} &Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s \sim \rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right] \\ &\implies \operatorname{Var}\left[\hat{\nabla}_i J_s\right] \geq \operatorname{Var}\left[\hat{\nabla}_i J_{\boldsymbol{h}}\right]. \end{aligned} \tag{20}$$

(proof in Appendix C.3).

Although Theorem 2 alone contains a result which is conditional to an equality which does not necessarily hold for a generic Dec-POMDP, combining Theorems 1 and 2 results in a broader statement about the overall quality of state-based policy gradient estimates, i.e.,

Corollary 2.1. *$\hat{\nabla}_i J_s$ cannot be guaranteed to have strictly better bias/variance properties than $\hat{\nabla}_i J_{\boldsymbol{h}}$, i.e., either its bias is higher (or equal), or its variance is higher (or equal), or neither is lower (or equal) (Follows directly from Corollary 1.1 and theorem 2).*

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 either 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,

$$Q^{\pi}(s = \textit{coffee}, a = \textit{tea}) = -1 \tag{21}$$

$$Q^{\pi}(s = \textit{tea}, a = \textit{tea}) = 1 \tag{22}$$

$$Q^{\pi}(h = \varepsilon, a = \textit{tea}) = 0 \tag{23}$$

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,

$$\begin{aligned} &\operatorname{Var}_{s|h=\varepsilon}\left[Q^{\pi}(s, a = \textit{tea})\right] \\ &= \mathbb{E}_{s|h=\varepsilon}\left[Q^{\pi}(s, a = \textit{tea})^2\right] - \mathbb{E}_{s|h=\varepsilon}\left[Q^{\pi}(s, a = \textit{tea})\right]^2 \\ &= \mathbb{E}_{s|h=\varepsilon}\left[1\right] - 0^2 \\ &= 1. \end{aligned} \tag{24}$$

Therefore, the policy gradient estimates will have a higher variance with the state-based critic. For more intuition on the variance of state-based critic, refer to the Dec-Tiger example in Appendix E.

5 Experiments

To understand the performance of centralized critics in practice, we test state-based critics and history-based critics using vanilla Advantage Actor-Critic with a centralized critic. We implement state or history value functions $V(s), V(\boldsymbol{h})$ and $V(\boldsymbol{h}, s)$ instead of Q functions; and the advantages are calculated using one-step differences as mentioned in Section 3.2. We highlight some of the interesting results and discuss the potential reasons behind performance differences. Furthermore, we introduce and test State-History-based Critics (SHC), where the concatenation of state and history is used as the input of the critic, analogous to Unbiased Asymmetric Actor-Critic (Baisero and Amato 2021) in single-agent settings. SHC uses $V(\boldsymbol{h}, s)$ directly as the critic, which can be shown to have no bias but larger variance in theory compared to the history-based critic (Appendix D).

Experiment Setup The figures shown are mean-aggregation of 20 runs per method; standard deviation is drawn as shaded bands around the lines. The experiments were conducted on compute clusters with nodes equipped with Dual Intel Xeon E5-2650 CPUs and 128GB of RAM. Hyperparameters are individually tuned while fixing other hyperparameters.

5.1 Observation Information Sufficiency

Some environments give partial yet "sufficient" local information in the sense that the optimal policy does not depend on the entire history. That is, the history-MDP (or even observation-MDP) is (close to) value-equivalent to the true underlying MDP; in some cases, reactive policies (policies that only condition on the last observation) can achieve optimal performance (Figures 1 and 7). For example, in the Meeting-in-a-Grid domains (Bernstein, Hansen, and Zilberstein 2005; Amato, Dibangoye, and 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 agent only observes local information, the optimal policy does not require remembering history because policies do not benefit from additional action-observation history. In these environments in Figure 1, 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.

5.2 Reactive Policies

We note that using reactive policies is not sufficient 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 (30)). Therefore, the state-based critic is unbiased for a reactive policy only when there is also a deterministic observation model. That is, for a specific agent, a given state-action pair can only produce a certain deterministic observation. This restriction ensures at most one non-null entry in each column of matrix $Q^{\pi}_{\boldsymbol{a}}$, thus warranting unbiased expected state-values $Q(o,a) = \mathbb{E}_{s\sim\phi(s|h)}Q(s,a)$. It is precisely the reason why MADDPG is not biased in their particle environments (Lowe et al. 2017). This situation also is not uncommon, which we see in numerous benchmarks including the classic task Recycling as well as environments with radius-based observation models such as StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al. 2019) which we discuss in Section 5.4. We hence note that the benchmarks used in recent state-of-the-art works usually has the aforementioned deterministic observation model, which is in fact a special case of imperfect information.

Cooperation The situation where reactive policies are not optimal is captured in the multi-agent recycling task for which policies with different types of critics all converged to a suboptimal solution. 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 observation history (non-reactive). Learning to cooperate in this case is especially difficult because the agents suffer the problem of shadowed equilibrium (Matignon, Laurent, and Le Fort-Piat 2012) despite the usage of a centralized critic (Lyu et al. 2021). Figure 1 shows performance for Recycling agents with various critics, in which none of the methods learns to recycle large targets (See Appendix F.6 for performance of reactive policies). Therefore, reactive policies are learned that are not significantly biased with state-based critics. Note that being unbiased means that the shadowed equilibrium also applies for policies trained with state-based critics.

5.3 State-History-Based Critics

We also test state-history-based critics (SHC), which estimate the value of a state and history pair. For experiments, we concatenate the state and history without changing other aspects of implementation. Note that the results shown in Figure 2 are not explicitly tuned but use the set of parameters tuned for the history-based critic (HC). In the Dec-Tiger (Bernstein, Hansen, and Zilberstein 2005), Box Pushing and Cleaner (Jiang 2019) domains in Figure 2 we see that the state-history-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 state-history-based critic can benefit from the advantages of both the state as well as the history as discussed in Section 6.2 and Section 6.1. We refer readers to the appendix for the details on those environments, but generally, both Box Pushing and Cleaner are grid world tasks with observations local to the cells around the agent. The agent needs to estimate their teammates' locations or paths to act optimally. Information gathering is even more crucial in Dec-Tiger, where we see history-based critics already distinctly outperform state-based ones, with SHC achieving the best results. By listening, the agents move from histories with little or no information to histories with more information in the same state; hence history and state-history critics can accurately represent the different values. The state-based critic, on the other hand, regard different histories with the same value, hence not incentivizing the policy to gather information, as we have shown in Section 4.1.

5.4 State Representation

We also see tasks where it is acceptable or even preferred to use state-based critics. A typical example is the StarCraft Multi-Agent Challenge (SMAC) benchmark. As seen in Figure 3, most scenarios show that state-based critics exhibit

Figure 1: Performance evaluation of state-based critics (SC), history-based critics (HC) and state-history-based critics (SHC) in cooperative multi-agent partially observable environments: Meeting-in-a-Grid (Amato, Dibangoye, and Zilberstein 2009), Find Treasure (Jiang 2019), Multi-agent Recycling (Amato, Bernstein, and Zilberstein 2007)

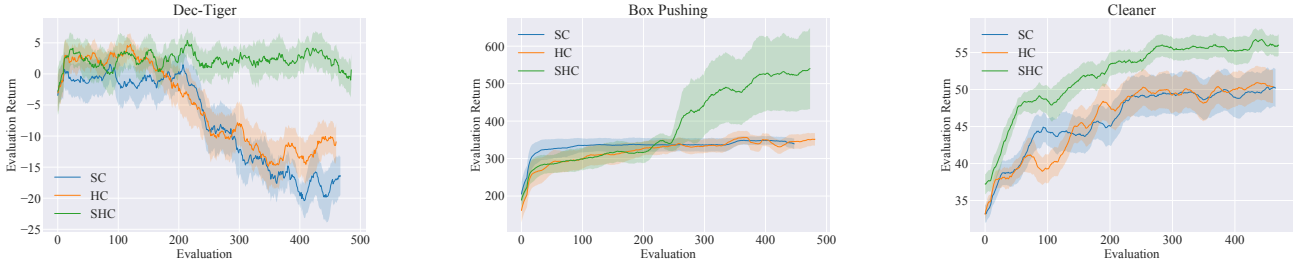

Figure 2: Performance evaluation of SC, HC and SHC on Dec-Tiger (Nair et al. 2003), Box Pushing (Seuken and Zilberstein 2007) and Cleaner (Jiang 2019).

the best overall performance. In SMAC, each agent controls a combat unit to engage enemy units, and the state includes the statuses of friendly agents and opponent units. We find SMAC agents usually have self-centered acting limits (e.g., firing ranges). We find far-away units rarely affect an agent's decisions. This limit causes the occluded information to rarely affect expected values. Thus, the history values associated with a state are usually similar; as a result, using a state-based critic does not introduce significant bias in this case. It is related to results also discussed in Section 5.2. In specific scenarios, we see that state-based critics perform better than history-based ones. Note that the observation for an agent is a masked version of the state, in which the information regarding far-away units is unobservable. Therefore, both the state and the observation have similar and concise representations, while the history will contain redundant and outdated information. Learning to ignore this information can be difficult and take time. Together with a deterministic observation model, and history carrying semi-redundant information, it is reasonable to believe that state-based critics may give a overall more reliable estimate for policy training.

6 Discussion

We discuss the tasks and scenarios where we can see the advantages of each type of critics and the underlying rationale.

6.1 Advantage of History-Based Critics

It is shown that the history-based critic is unbiased in theory, so does it translate to increased performance? Toy examples aside, some of the more realistic environments shown above may also favor history-based critics due to its ability to evaluate observed information—for a given history, the ability to distinguish how much information was gathered by the policy in the form of observation-action histories and reflect the information in terms of values. The information gained from observations helps in reducing uncertainty and with such histories the policy can achieve better values compared to ones with less information, and history values encode this value difference explicitly. At a high level, history-based policies and critics have a simple one-to-one update relationship, but state-based critics have history-based policies that may have a many-to-many update relationship. Often, you will have two or more histories being mapped to the same state-value when they would have different history values. This mapping can cause aliasing in values, which is the source of the bias.

Being able to represent the value of information is the key strength we see in history-based critics and the key weakness of state-based critics. For example, suppose the agents can gather information regarding the distribution of the world state. In that case, histories with more information have an advantage compared to less informed histories (Kaelbling, Littman, and Cassandra 1998). Using the notation in our bias analysis, that is, for a certain state s, its state-history values (a column of Q_a^π) vary depending on the history. Each (s,h) pair, therefore, represents a distinct condition in which the agent can have different behaviors (and thus returns), but share the same state value. Since the state-based value function is the marginalization of the values of these more or less informed history values, state values remain incapable of evaluating how information effects future returns. On the other hand, for a specific state, the history values distinguish histories with different levels of uncertainty, which results in different payoffs; Therefore, in environments where gathering

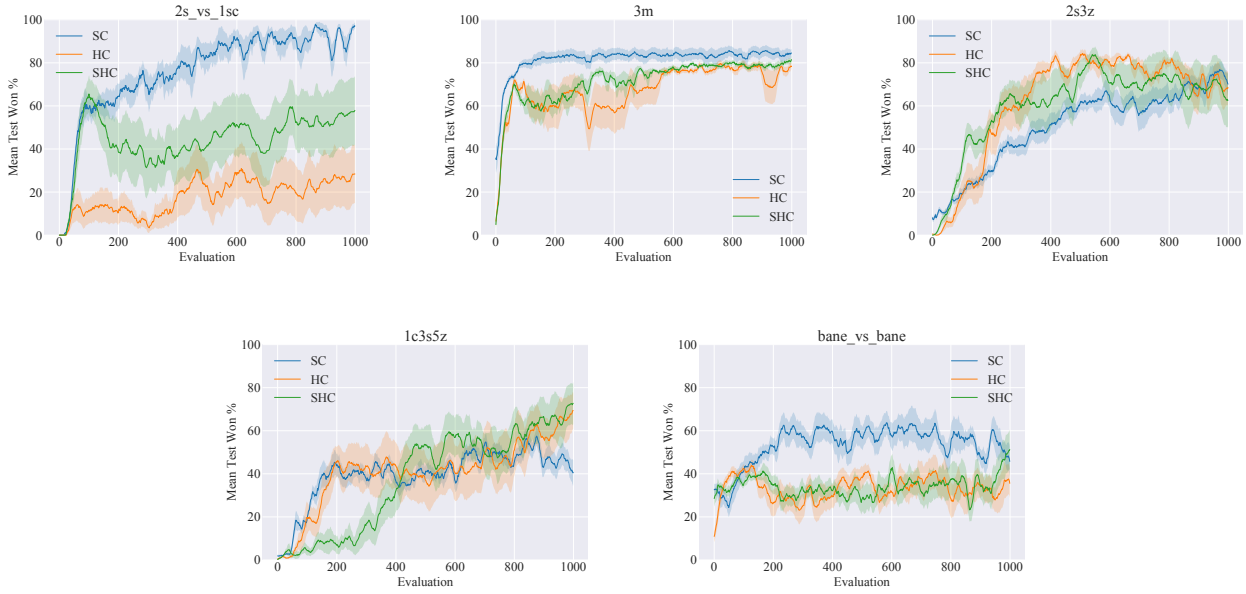

Figure 3: Performance comparison of COMA with different critics in multiple scenarios of SMAC.

information is critical, we suggest using state-history critics that can take information gathering into account and benefit from state information.

Using State Value as Baseline The gradient estimates of implementations which use Monte Carlo returns with value functions as baselines, such as in DICG (Li et al. 2021) or MAGIC (Niu, Paleja, and Gombolay 2021), are guaranteed to be unbiased. As a baseline, the bias of the state value function does not affect the bias of the overall policy gradient. However, it may affect its variance: although the history value baselines are widely considered to have good variance reduction properties, it is not clear whether the same can be said of state value baselines due to its potential bias.

6.2 Advantage of State-Based Critics

On the other hand, state-based critics have proven helpful in recent works (Wang et al. 2020; Du et al. 2019, 2021). We speculate that there are three main reasons behind state-based critics being able to provide signals that give good performance despite their theoretical shortcomings. First, observe that in some environments, using the state as input more easily allows the value function to extract meaningful features compared to extracting features from histories. It is especially the case when the state representation is concise or when the history feature extraction requires a considerable amount of training. For example, some grid-world environments' states involve agent locations, while observations stem from on-screen pixels that are of much higher dimension. This happen to be not unusual in current multi-agent reinforcement learning benchmarks (see Sections 5.3 and 5.4).

Second, we observe that the state representation is usually a complete version of the observation information in multi-agent benchmarks used by recent state-of-the-art works. As discussed in Section 5.4, those environments generate observations either by masking the underlying state representation or by outputting state fragments. As a result, the state information is not drastically different from observations, implying that the state-based value functions are well aligned with history-based value functions. This alignment is amplified in situations where the information omitted in observations (e.g., information regarding far-away teammates or opponents) has zero or negligible effect on the expected values, which is applied in popular benchmarks such as SMAC (Samvelyan et al. 2019) and partially observable particle environments (Lowe et al. 2017). Shown in Figures 3, 4 and 6, as long as the information gathering issue mentioned above is either non-existent or not effecting the value estimation, the state-based critics can give good return signals for the purpose of policy learning.

7 Conclusion

In this paper, we take a close look at state-based critics in multi-agent actor-critic methods. We show how state-based critic values may incur bias in training decentralized history-based policies. We also show that, in theory, that state-based critics exhibit more variance in the policy gradient. We also suggest and evaluate an alternative using state information in conjunction with the history information to train critics which have seen reliable results empirically. We discuss task-specific conditions in which the bias will prominently occur, explain the empirical performance of these critics on various environments, summarizing where and why state and history-based critics should be effective. This work fill in the gap of theoretical understanding of state-based critics popular in multi-agent reinforcement learning, providing a principled foundation for future works on centralized critics in multi-agent reinforcement learning.

References

Amato, C.; Bernstein, D. S.; and Zilberstein, S. 2007. Optimizing Memory-Bounded Controllers for Decentralized POMDPs. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

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

Baisero, A.; and Amato, C. 2021. Unbiased Asymmetric Actor-Critic for Partially Observable Reinforcement Learning. arXiv:2105.11674.

Baker, B.; Kanitscheider, I.; Markov, T.; Wu, Y.; Powell, G.; McGrew, B.; and Mordatch, I. 2020. Emergent Tool Use from Multi-Agent Autocurricula. In *Proceedings of the Eighth International Conference on Learning Representations ICLR*.

Bernstein, D. S.; Hansen, E. A.; and Zilberstein, S. 2005. Bounded policy iteration for decentralized POMDPs. In *Proceedings of the nineteenth international joint conference on artificial intelligence (IJCAI)*, 52–57.

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

Du, Y.; Han, L.; Fang, M.; Dai, T.; Liu, J.; and Tao, D. 2019. LIIR: Learning Individual Intrinsic Reward in Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*.

Du, Y.; Liu, B.; Moens, V.; Liu, Z.; Ren, Z.; Wang, J.; Chen, X.; and Zhang, H. 2021. Learning Correlated Communication Topology in Multi-Agent Reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 456–464.

Foerster, J.; Assael, I. A.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, 2137–2145.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual Multi-Agent Policy Gradients. In *AAAI 2018: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Jiang, S. 2019. Multi Agent Reinforcement Learning Environments Compilation.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134.

Konda, V. R.; and Tsitsiklis, J. N. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*, 1008–1014.

Li, S.; Gupta, J. K.; Morales, P.; Allen, R.; and Kochenderfer, M. J. 2021. Deep implicit coordination graphs for multi-agent reinforcement learning. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Neural Information Processing Systems (NIPS)*.

Lyu, X.; Xiao, Y.; Daley, B.; and Amato, C. 2021. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 844–852.

Matignon, L.; Laurent, G. J.; and 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.

Nair, R.; Tambe, M.; Yokoo, M.; Pynadath, D.; and Marsella, S. 2003. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI*, volume 3, 705–711.

Niu, Y.; Paleja, R.; and Gombolay, M. 2021. Multi-Agent Graph-Attention Communication and Teaming. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 964–973.

Oliehoek, F. A.; and Amato, C. 2016. *A Concise Introduction to Decentralized POMDPs*. Springer Publishing Company, Incorporated.

Oliehoek, F. A.; Spaan, M. T. J.; and Vlassis, N. A. 2008. Optimal and Approximate Q-value Functions for Decentralized POMDPs. *J. Artif. Intell. Res.*, 32: 289–353.

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.; and Whiteson, S. 2019. The StarCraft Multi-Agent Challenge. *CoRR*, abs/1902.04043.

Schroeder de Witt, C.; Foerster, J.; Farquhar, G.; Torr, P.; Boehmer, W.; and Whiteson, S. 2019. Multi-agent common knowledge reinforcement learning. *Advances in Neural Information Processing Systems*, 32: 9927–9939.

Seuken, S.; and Zilberstein, S. 2007. Improved memory-bounded dynamic programming for decentralized POMDPs. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, 344–351.

Su, J.; Adams, S.; and Beling, P. A. 2021. Value-Decomposition Multi-Agent Actor-Critics. In *Proceedings of the Thirty-fifth AAAI Conference on Artificial Intelligence*.

Sutton, R.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, 112: 181–211.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057–1063.

Wang, J.; Zhang, Y.; Kim, T.-K.; and Gu, Y. 2020. Shapley Q-Value: A Local Reward Approach to Solve Global Reward Games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7285–7292.

Wang, R. E.; Everett, M.; and How, J. P. 2019. R-MADDPG for Partially Observable Environments and Limited Communication.

Wang, Y.; Han, B.; Wang, T.; Dong, H.; and Zhang, C. 2021. Off-Policy Multi-Agent Decomposed Policy Gradients. *ICLR*.

Zhou, M.; Liu, Z.; Sui, P.; Li, Y.; and Chung, Y. Y. 2020. Learning Implicit Credit Assignment for Cooperative Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*.

A On Assumptions of Previous Works

In this section, we discuss the assumptions or the limited scope of previous work, which may have contributed to the recent neglect of the precise effects of state-based critics.

A.1 MADDPG

The theories given by MADDPG (Lowe et al. 2017) is only correct under certain implicit conditions. In particular, it assumes reactive policies and the environment has a deterministic observation model. In MADDPG, both criteria are met during testing but are not particularly emphasized in the main paper. Therefore their theories should not be considered true without the above restrictions. Without emphasizing the scope, MADDPG also states that it is also applicable to recurrent policies, however, the centralized critic bias depends on the information included in the x in their Eq.5. Understandably, some works make a very straightforward extension of MADDPG such as learning history-based policy with a state-based critic without further investigating the internal interaction between the critic and the actor during updates (Eq.3); For example, RMADDPG (Wang, Everett, and How 2019) tried to investigate the effect of putting a recurrent layer in actor, critic or both, but without any theoretical justification. Issues of this nature are what we aim to resolve with this work with a theoretical and empirical analysis.

A.2 COMA

Similarly with COMA (Foerster et al. 2016), some may regard Eq. 15 as incorrect, or argue that there is an assumption being made implicitly. In the work of COMA, without deeper discussion, Eq. 15 assumes that individual history-based actors can be combined into a centralized state-based policy. Indeed, under such an assumption, it seems using a state-critic with history-based actors won't lead to extra bias. However, this assumption cannot hold in general and cannot hold even with COMA on SMAC: the centralized behavior conditioned on state is highly likely to be different from decentralized behavior based on each agent's local history,

$$\pi(\mathbf{u} \mid s) = \prod_a \pi^a(u^a \mid s) \neq \prod_a \pi^a(u^a \mid \tau^a) \tag{25}$$

Consider the following example: let two local histories lead to a same state, the decentralized/history-based policy can produce two different action distributions, while the centralized state-based policy can only produce one action distribution.

B The On-Policy State-History Distributions

Here, we define the "on-policy" state-history distribution in the context of finite-horizon Dec-POMDPs. These definitions are analogous to those by Sutton and Barto (Sutton and Barto 2018), which we extend to the multi-agent case.

Let $\eta(\boldsymbol{h}, s)$ be the expected number of time-steps that the agents spend in state s while having observed the joint histories $\boldsymbol{h}$. Denoting the marginal distribution of states and joint histories at a specific time-step as $\Pr_t(\boldsymbol{h}, s)$, we have that η satisfies

$$\eta(\boldsymbol{h}, s) = \Pr{}_0(\boldsymbol{h}, s) + \sum_{\bar{\boldsymbol{h}}, \bar{s}} \eta(\bar{\boldsymbol{h}}, \bar{s}) \sum_{\boldsymbol{a}} \boldsymbol{\pi}(\boldsymbol{a} \mid \bar{\boldsymbol{h}}) \Pr(\boldsymbol{h}, s \mid \bar{\boldsymbol{h}}, \bar{s}, \boldsymbol{a}) \,. \tag{26}$$

which is solved by $\eta(\boldsymbol{h}, s) \doteq \sum_t \Pr_t(\boldsymbol{h}, s)$. Note that it is solvable because the set of histories are finite due to the finite-horizon assumption. The on-policy state-history distribution $\rho(\boldsymbol{h}, s)$ is defined by normalizing the visitation counts,

$$\rho(\boldsymbol{h}, s) \doteq \frac{\eta(\boldsymbol{h}, s)}{\sum_{\boldsymbol{h}', s'} \eta(\boldsymbol{h}, s)} \,. \tag{27}$$

From the joint state-history distribution $\rho(\boldsymbol{h}, s)$, we can also define the respective marginal and conditional distributions,

$$\rho(\boldsymbol{h}) \doteq \sum_s \rho(\boldsymbol{h}, s) \,, \qquad \rho(s) \doteq \sum_{\boldsymbol{h}} \rho(\boldsymbol{h}, s) \,, \tag{28}$$

$$\rho(\boldsymbol{h} \mid s) \doteq \frac{\rho(\boldsymbol{h}, s)}{\rho(s)} \,, \qquad \rho(s \mid \boldsymbol{h}) \doteq \frac{\rho(\boldsymbol{h}, s)}{\rho(\boldsymbol{h})} \,. \tag{29}$$

C Proofs

C.1 Bias of State Values

Here we prove Theorem 1 using an argument similar to that made in (Baisero and Amato 2021), with the key differences that our case involves arbitrary finite-horizon environments and policies (not reactive policies), Q values (not V values), and the on-policy history/state distributions (rather than the belief distribution).

Proof. For a generic environment, consider an arbitrary joint action $\boldsymbol{a}$, and two different joint histories $\boldsymbol{h} \neq \boldsymbol{h}'$ which are associated with the same empirical conditional state distribution $\rho(s \mid \boldsymbol{h}) = \rho(s \mid \boldsymbol{h}')$ (a fairly common occurrence in partially observable environments). Because the joint histories are different, the future behaviors of the agents are generally going to differ, generally resulting in different history values,

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) \neq Q^{\boldsymbol{\pi}}(\boldsymbol{h}', \boldsymbol{a})\,. \tag{30}$$

On the other hand, because both the conditional state distributions and the individual state values are the same for every state, the expected state values must be the same,

$$\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right] = \mathbb{E}_{s\sim\rho(s|\boldsymbol{h}')}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right]\,. \tag{31}$$

For these specific histories, the equality $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, a) = \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right]$, combined with eq. (30), leads to a contradiction of eq. (31).

$$\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right] = Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) \neq Q^{\boldsymbol{\pi}}(\boldsymbol{h}', \boldsymbol{a}) = \mathbb{E}_{s\sim\rho(s|\boldsymbol{h}')}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right]\,. \tag{32}$$

□

C.2 Policy Gradient Bias of State-Based Critic

Joint Policies Here we prove Corollary 1.1 with the case of joint policies. We will be considering "tabular" agent policies, i.e., policies which are parameterized by a separate parameter for each possible joint history-action pair, $\pi(\boldsymbol{a}; \boldsymbol{h}) = \theta^\top \mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}$. We first note that any non-tabular policy can be converted into an equivalent tabular policy which behaves the same way and chooses the same actions under the same histories. Therefore, because Theorem 1 says that there is at least one policy for which the state values are biased, then there is also at least one tabular policy for which the state values are biased.

Proof. We prove that the state-based critic's policy gradient is biased if the values are biased for tabular joint policies. We prove by contradiction. We first assume that $\nabla_i J_{\boldsymbol{h}} = \nabla_i J_s$, and will show that under this assumption we have to conclude that the policies' state values are unbiased, which is a contradiction with Theorem 1 (which, as noted above, is also valid on the sub-family of tabular policies).

For tabular policies, we have

$$\nabla_\theta \log \pi(\boldsymbol{a}; \boldsymbol{h}) = \frac{\nabla_\theta \pi(\boldsymbol{a}; \boldsymbol{h})}{\pi(\boldsymbol{a}; \boldsymbol{h})} \tag{33}$$

$$= \frac{\nabla_\theta \left(\theta^\top \mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}\right)}{\pi(\boldsymbol{a}; \boldsymbol{h})} \tag{34}$$

$$= \frac{\mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}}{\pi(\boldsymbol{a}; \boldsymbol{h})}\,. \tag{35}$$

Therefore,

$$\nabla_i J_{\boldsymbol{h}} = \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\nabla_{\theta_i} \log \pi(\boldsymbol{a}; \boldsymbol{h})\right] \tag{36}$$

$$= \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\frac{\mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}}{\pi(\boldsymbol{a}; \boldsymbol{h})}\right]\,, \tag{37}$$

$$\nabla_i J_s = \mathbb{E}_{\boldsymbol{h},s\sim\rho(\boldsymbol{h},s),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\nabla_{\theta_i} \log \pi(\boldsymbol{a}; \boldsymbol{h})\right] \tag{38}$$

$$= \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right]\frac{\mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}}{\pi(\boldsymbol{a}; \boldsymbol{h})}\right]\,. \tag{39}$$

Using the assumption of $\nabla_i J_{\boldsymbol{h}} = \nabla_i J_s$, we have

$$\mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\frac{\mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}}{\pi(\boldsymbol{a}; \boldsymbol{h})}\right] = \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right]\frac{\mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}}{\pi(\boldsymbol{a}; \boldsymbol{h})}\right] \tag{40}$$

$$\mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\frac{1}{\pi(\boldsymbol{a}; \boldsymbol{h})}\right] = \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h}),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\mathbb{1}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})\right]\frac{1}{\pi(\boldsymbol{a}; \boldsymbol{h})}\right] \tag{41}$$

Due to the one-hot vector, we only have one non-zero value from both sides along each $(\boldsymbol{h}, \boldsymbol{a})$ dimension, therefore

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\frac{1}{\pi(\boldsymbol{a}; \boldsymbol{h})} = \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})]\frac{1}{\pi(\boldsymbol{a}; \boldsymbol{h})} \tag{42}$$

$$Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})]\,, \tag{43}$$

which contradicts Theorem 1. □

Independent Policies Here we prove Corollary 1.1 with the case of independent policies. We continue to use "tabular" agent policies, $\pi_i(a_i; h_i) = \theta^\top \mathbb{1}_{h_i,a_i}$ and prove by example.

Proof. Continuing from the Dec-Tiger example in Section 4.1. We explicitly calculate the gradient for the case of $h_i = (\textit{listen}, \textit{hear-left}, \textit{listen}, \textit{hear-right}), a_i = \textit{listen}$. Here we use a horizon of 4 such that $\pi(a_i; h_i) = 1$. We denote the relevant joint histories as:

$$\boldsymbol{h}_1 = (\textit{listen}, \textit{listen}), (\textit{hear-left}, \textit{hear-left}), (\textit{listen}, \textit{listen}), (\textit{hear-right}, \textit{hear-left}) \tag{44}$$
$$\boldsymbol{h}_2 = (\textit{listen}, \textit{listen}), (\textit{hear-left}, \textit{hear-left}), (\textit{listen}, \textit{listen}), (\textit{hear-right}, \textit{hear-right}) \tag{45}$$
$$\boldsymbol{h}_3 = (\textit{listen}, \textit{listen}), (\textit{hear-left}, \textit{hear-right}), (\textit{listen}, \textit{listen}), (\textit{hear-right}, \textit{hear-left}) \tag{46}$$
$$\boldsymbol{h}_4 = (\textit{listen}, \textit{listen}), (\textit{hear-left}, \textit{hear-right}), (\textit{listen}, \textit{listen}), (\textit{hear-right}, \textit{hear-right}) \tag{47}$$

We first calculate the gradient provided by the history values:

$$\begin{aligned}
&\nabla J_{\boldsymbol{h}}(h_i, a_i) \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i} Q^\pi(\boldsymbol{h}, \boldsymbol{a}) \frac{\mathbb{1}_{h_i,a_i}}{\pi(a_i; h_i)} \\
&= Pr(\boldsymbol{h}_1 \mid h_i, a_i) \cdot Q^\pi(\boldsymbol{h}_1, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} + Pr(\boldsymbol{h}_2 \mid h_i, a_i) \cdot Q^\pi(\boldsymbol{h}_2, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} \\
&+ Pr(\boldsymbol{h}_3 \mid h_i, a_i) \cdot Q^\pi(\boldsymbol{h}_3, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} + Pr(\boldsymbol{h}_4 \mid h_i, a_i) \cdot Q^\pi(\boldsymbol{h}_4, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} \\
&= (0.3725 \cdot -6.854 + 0.1275 \cdot -18.175 + 0.1275 \cdot -18.175 + 0.3725 \cdot -6.854) \mathbb{1}_{h_i,a_i} \\
&= -9.74 \, \mathbb{1}_{h_i,a_i} \, .
\end{aligned} \tag{48}$$

We then calculate the gradient provided by the state values:

$$\begin{aligned}
&\nabla J_s(h_i, a_i) \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}|h_i,a_i} \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})} Q(s, \boldsymbol{a}) \frac{\mathbb{1}_{h_i,a_i}}{\pi(a_i; h_i)} \\
&= Pr(\boldsymbol{h}_1 \mid h_i, a_i) \cdot \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})} Q(s, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} + Pr(\boldsymbol{h}_2 \mid h_i, a_i) \cdot \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})} Q(s, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} \\
&+ Pr(\boldsymbol{h}_3 \mid h_i, a_i) \cdot \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})} Q(s, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} + Pr(\boldsymbol{h}_4 \mid h_i, a_i) \cdot \mathbb{E}_{s\sim\rho(s|\boldsymbol{h})} Q(s, (\textit{listen}, \textit{listen})) \mathbb{1}_{h_i,a_i} \\
&= (0.3725 \cdot 0.0474 + 0.1275 \cdot 0.0474 + 0.1275 \cdot 0.0474 + 0.3725 \cdot 0.0474) \mathbb{1}_{h_i,a_i} \\
&= 0.0474 \, \mathbb{1}_{h_i,a_i} \, .
\end{aligned} \tag{49}$$

Therefore, $\nabla J_s \neq \nabla J_h$. □

C.3 Variance of State-Based Gradient Estimates

Here, we complete the proof for Theorem 2 which states that, if $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ is unbiased, then the variance of the state-based policy gradient estimates $\hat{\nabla}_i J_s$ are greater or equal than that of the history-based policy gradient estimates $\hat{\nabla}_i J_{\boldsymbol{h}}$.

First, we find that, for any given joint history $\boldsymbol{h}$ and joint action $\boldsymbol{a}$,

$$\begin{aligned}
\mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s \right] &= \mathbb{E}_{s|\boldsymbol{h}} \left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \nabla \log \pi_i(a_i; h_i) \right] \\
&= \mathbb{E}_{s|\boldsymbol{h}} \left[Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \right] \nabla \log \pi_i(a_i; h_i) \\
&= Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a}) \nabla \log \pi_i(a_i; h_i) \\
&= \hat{\nabla}_i J_{\boldsymbol{h}} \, .
\end{aligned} \tag{50}$$

We will use the fact that history-based critic policy gradient is the expectation of the state-based critic policy gradient to establish the variance relationship. Then, using Jensen's inequality, we get,

$$\begin{aligned}
\mathrm{Var}\left[\hat{\nabla}_i J_s \right] &= \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}} \left[\hat{\nabla}_i J_s^\mathsf{T} \hat{\nabla}_i J_s \right] - \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}} \left[\hat{\nabla}_i J_s \right]^\mathsf{T} \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}} \left[\hat{\nabla}_i J_s \right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s^\mathsf{T} \hat{\nabla}_i J_s \right] \right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s \right] \right]^\mathsf{T} \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s \right] \right] \\
&\geq \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s \right]^\mathsf{T} \mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s \right] \right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s \right] \right]^\mathsf{T} \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\mathbb{E}_{s|\boldsymbol{h}} \left[\hat{\nabla}_i J_s \right] \right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\hat{\nabla}_i J_{\boldsymbol{h}}^\mathsf{T} \hat{\nabla}_i J_{\boldsymbol{h}} \right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\hat{\nabla}_i J_{\boldsymbol{h}} \right]^\mathsf{T} \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}} \left[\hat{\nabla}_i J_{\boldsymbol{h}} \right] \\
&= \mathrm{Var}\left[\hat{\nabla}_i J_{\boldsymbol{h}} \right] \, .
\end{aligned} \tag{51}$$

We will discuss the intuition behind this proof in Appendix E.

D State-History Values and Gradient Variances

Let us consider a variant of the policy gradient which uses the state-history value function $Q^{\boldsymbol{\pi}}(\boldsymbol{h}, s)$,

$$\nabla_i J_{\boldsymbol{h}s} = \mathbb{E}_{\boldsymbol{h},s\sim\rho(\boldsymbol{h},s),\boldsymbol{a}\sim\boldsymbol{\pi}(\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, s, \boldsymbol{a})\nabla_{\theta_i}\log\pi_i(a_i; h_i)\right] . \tag{52}$$

As for $\hat{\nabla}_i J_{\boldsymbol{h}}$ and $\hat{\nabla}_i J_s$, we consider the Monte Carlo estimate associated with a single joint-history, state, and joint action sample,

$$\hat{\nabla}_i J_{\boldsymbol{h}s} = Q^{\boldsymbol{\pi}}(\boldsymbol{h}, s, \boldsymbol{a})\nabla\log\pi_i(a_i; h_i) , \tag{53}$$

where $\boldsymbol{h}, s \sim \rho(\boldsymbol{h}, s)$ and $\boldsymbol{a} \sim \boldsymbol{\pi}(\boldsymbol{h})$.

Next, we show that $\hat{\nabla}_i J_{\boldsymbol{h}s}$ is unbiased,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h},s\sim\rho(\boldsymbol{h},s)}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right] &= \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h})}\left[\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h})}\left[\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, s, \boldsymbol{a})\nabla_i\log\pi_i(a_i; h_i)\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h})}\left[\mathbb{E}_{s\sim\rho(s|\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, s, \boldsymbol{a})\right]\nabla_i\log\pi_i(a_i; h_i)\right] \\
&= \mathbb{E}_{\boldsymbol{h}\sim\rho(\boldsymbol{h})}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\nabla_i\log\pi_i(a_i; h_i)\right] \\
&= \nabla_i J_{\boldsymbol{h}} .
\end{aligned} \tag{54}$$

On the other hand, its variance is at least as high as that of $\hat{\nabla}_i J_{\boldsymbol{h}}$. This proof is similar to that in Appendix C.3, except that it does not rely on an assumption of no bias, since no bias is already proven. First, we find that, for any given joint history $\boldsymbol{h}$ and joint action $\boldsymbol{a}$,

$$\begin{aligned}
\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right] &= \mathbb{E}_{s|\boldsymbol{h}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, s, \boldsymbol{a})\nabla\log\pi_i(a_i; h_i)\right] \\
&= \mathbb{E}_{s|\boldsymbol{h}}\left[Q^{\boldsymbol{\pi}}(\boldsymbol{h}, s, \boldsymbol{a})\right]\nabla\log\pi_i(a_i; h_i) \\
&= Q^{\boldsymbol{\pi}}(\boldsymbol{h}, \boldsymbol{a})\nabla\log\pi_i(a_i; h_i) \\
&= \hat{\nabla}_i J_{\boldsymbol{h}}
\end{aligned} \tag{55}$$

Then,

$$\begin{aligned}
\mathrm{Var}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right] &= \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}^{\mathsf{T}}\hat{\nabla}_i J_{\boldsymbol{h}s}\right] - \mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]^{\mathsf{T}}\mathbb{E}_{\boldsymbol{h},s,\boldsymbol{a}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}^{\mathsf{T}}\hat{\nabla}_i J_{\boldsymbol{h}s}\right]\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]\right]^{\mathsf{T}}\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]\right] \\
&\geq \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]^{\mathsf{T}}\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]\right]^{\mathsf{T}}\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\mathbb{E}_{s|\boldsymbol{h}}\left[\hat{\nabla}_i J_{\boldsymbol{h}s}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\hat{\nabla}_i J_{\boldsymbol{h}}^{\mathsf{T}}\hat{\nabla}_i J_{\boldsymbol{h}}\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\hat{\nabla}_i J_{\boldsymbol{h}}\right]^{\mathsf{T}}\mathbb{E}_{\boldsymbol{h},\boldsymbol{a}}\left[\hat{\nabla}_i J_{\boldsymbol{h}}\right] \\
&= \mathrm{Var}\left[\hat{\nabla}_i J_{\boldsymbol{h}}\right] .
\end{aligned} \tag{56}$$

E Variance Intuition

We again use the Dec-Tiger problem (Nair et al. 2003) as a thought exercise to highlight the intuition behind the gradient variance bound of state-based critic. We use the same setting from section 4.1. We denote left door with x and right with y. Consider the following trajectory: $\{o_1 = \varnothing,\ a_1 = l,\ o_2 = x,\ a_2 = l, o_3 = x\}$ which we simplify as $h = lxlx$, where the agent listened twice and observed that the treasure is likely to be inside door x both of the time.

We now look at the true value functions of Dec-Tiger. Note that the history values of opening doors depend on the observations, the more information contained in the history, the more we expect the agent to obtain high return, i.e. $Q(\varnothing, l) \leq Q(lx, l) \leq Q(lxlx, l)$. This table show three different histories' values with the same underlying state, as the agent gathers more information, the history value improves.

h	s	$Q(s, a = l)$	$Q(h, a = l)$
$\varnothing$	x	0.0474	0.0474
lx	x	0.0474	7.23
$lxlx$	x	0.0474	15.9

The values are calculated analytically, see supplementary material for implementation. Let us now consider the variance of the values, which can be rather deceptive. On the surface, by looking at the table, it appears that a single state is related to a

potentially large set of histories that vary in history values; biased or not, it *seems* that state values have much less variance. However, we are interested in the policy gradient for history-based policies. If we employ a state-based critic, what is used to update a history-based policy is a distribution of state values in accordance to $\rho(s \mid h)$. To see such an update for a particular history, we next show values for a particular history in a table — the two tables below each focus on a particular history $h = lxlx$ and $h = \varnothing$ respectively, taken from multiple episodes. Based on the distribution $\rho(s|h)$ ($\rho(x|lxlx) \approx 0.99$, $\rho(y|lxlx) \approx 0.001$), we might find our rollout look like the following table if we only show timesteps for which $h = lxlx$

episode	h	s	$Q(s, a = or)$	$Q(h, a = or)$
1	$l-hl-l-hl$	l	20	13.9
3	$l-hl-l-hl$	l	20	13.9
8	$l-hl-l-hl$	l	20	13.9
29	$l-hl-l-hl$	l	20	13.9
30	$l-hl-l-hl$	l	20	13.9
33	$l-hl-l-hl$	l	20	13.9
60	$l-hl-l-hl$	r	-50	13.9
89	$l-hl-l-hl$	l	20	13.9
91	$l-hl-l-hl$	l	20	13.9

We see that the history-value column is constant but the state-value column now varies. Because each table only looks at a specific history $\varnothing$ and $lxlx$, it is unsurprising that the history value remains constant in each table and thus have no variance. It is essentially the reason why one cannot obtain values of smaller variance than history values, because true history values essentially have no additional variance for a given history in policy update.

F OpenAI Particle Environments with Partial Observability

F.1 Speaker and Listener

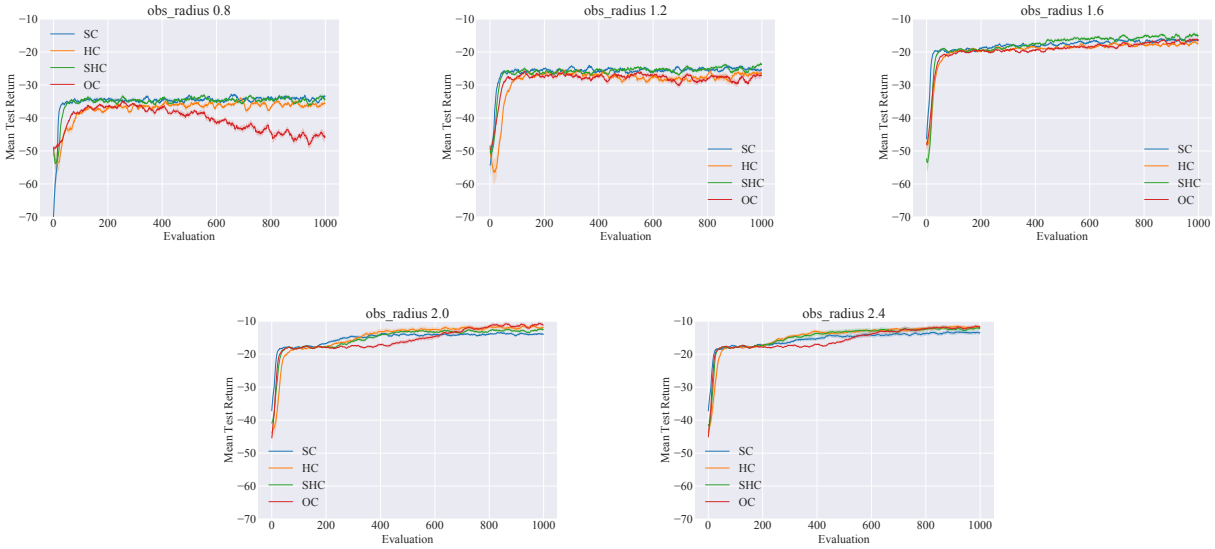

Figure 4: Performance comparison in Speaker and Listener.

F.2 Predator and Prey

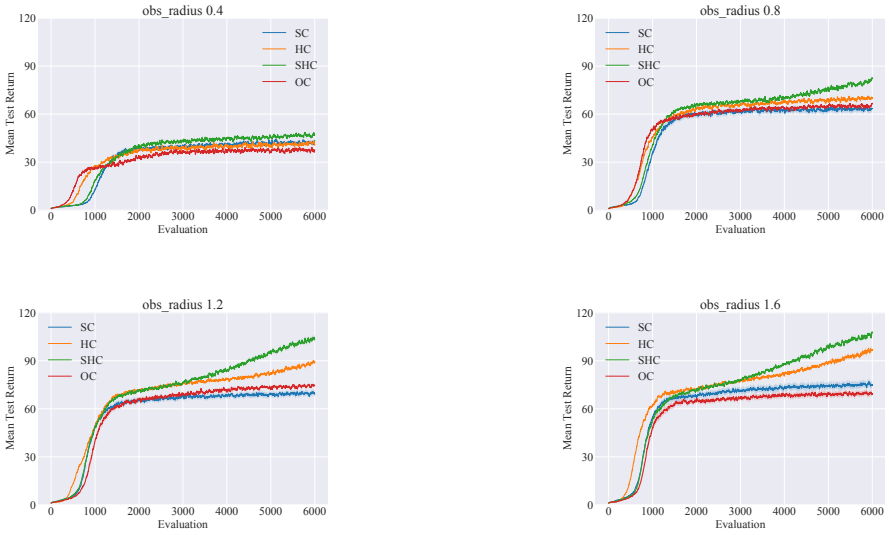

Figure 5: Performance comparison in Predator and Prey.

F.3 Cooperative Navigation

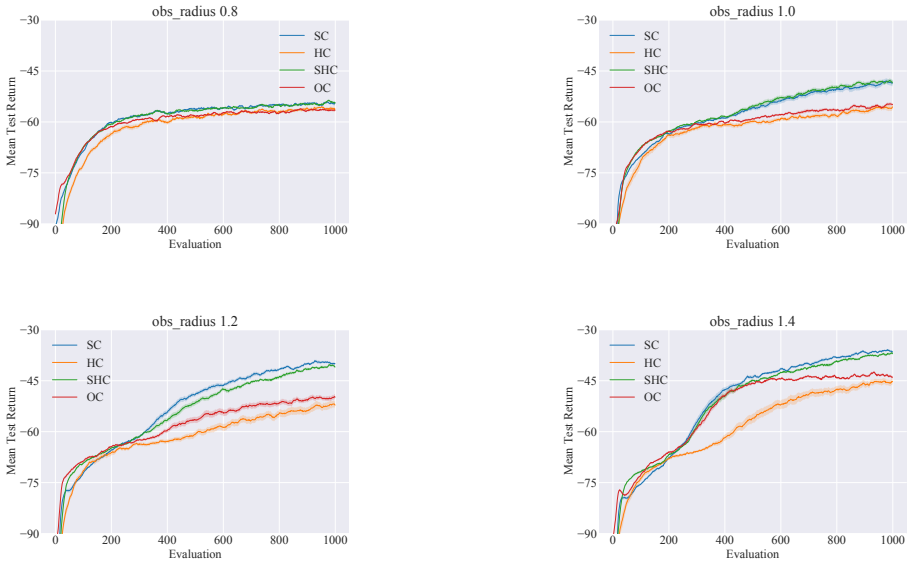

Figure 6: Performance comparison in Cooperative Navigation.

F.4 Meeting-in-a-Grid Small

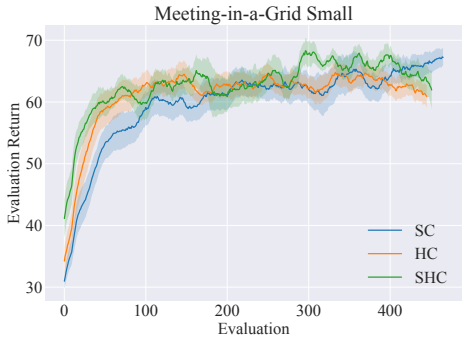

Figure 7: Meeting-in-a-Grid-Small environment (Bernstein, Hansen, and Zilberstein 2005)

F.5 COMA

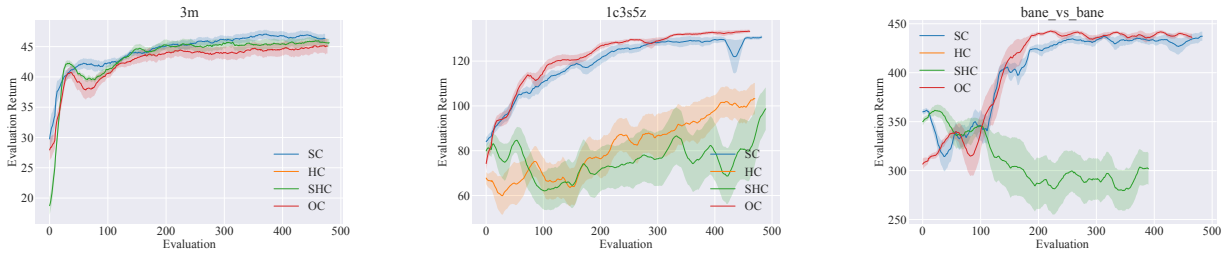

Figure 8: Performance comparison in scenarios of SMAC using various critics, sum of reward is plotted.

F.6 Reactive Policies

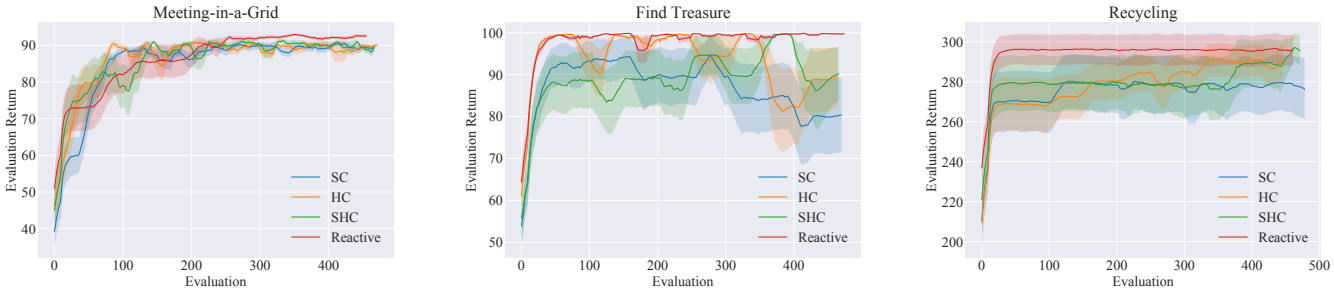

Figure 9: Performance evaluation including reactive policies

F.7 MAAC

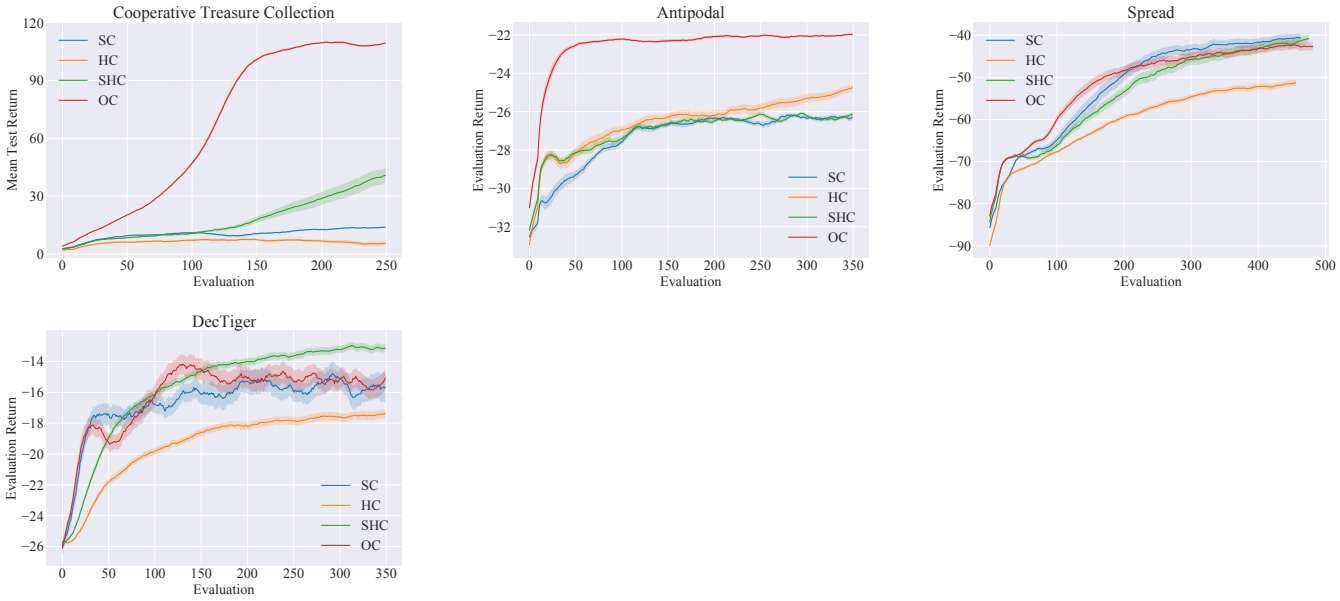

Figure 10: Performance evaluation using MAAC with different types of critics

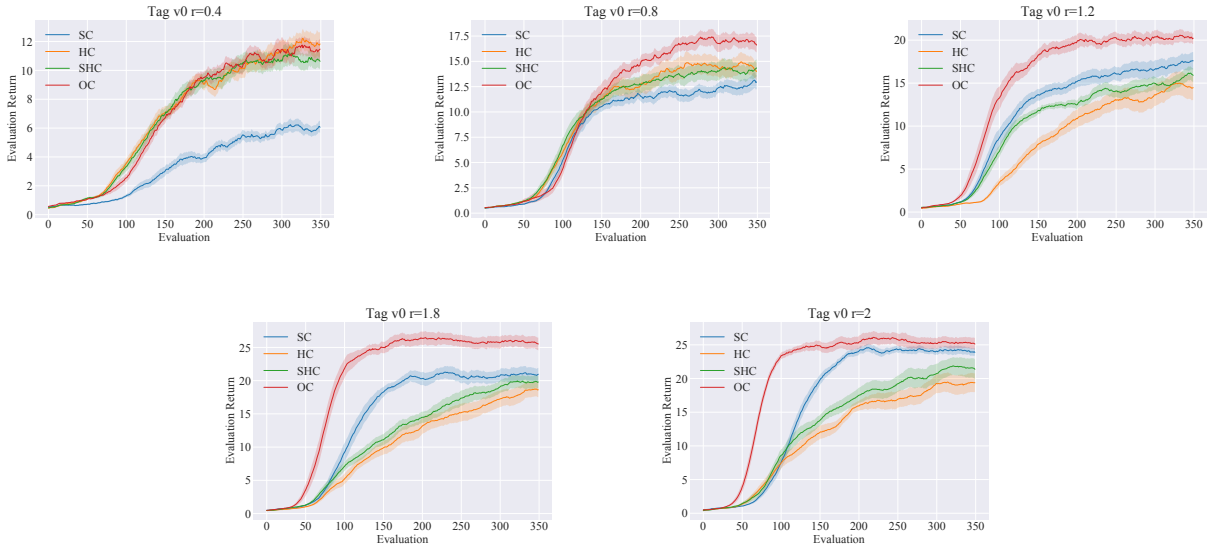

Figure 11: MAAC Performance comparison in Cooperative Tag V0.

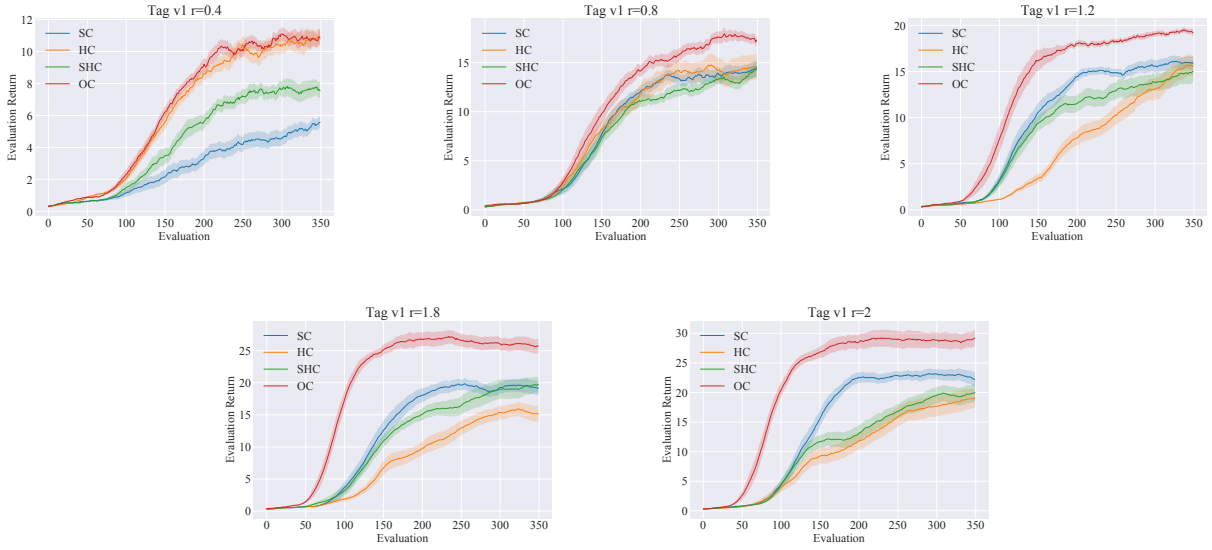

Figure 12: MAAC Performance comparison in Cooperative Tag V1.