
ENSEMBLE-MIX: ENHANCING SAMPLE EFFICIENCY IN MULTI-AGENT RL USING ENSEMBLE METHODS

Tom Danino

The Andrew and Erna Viterbi Faculty of
Electrical & Computer Engineering
Technion- Israel Institute of Technology
tdanino12@gmail.com

Nahum Shimkin

The Andrew and Erna Viterbi Faculty of
Electrical & Computer Engineering
Technion- Israel Institute of Technology
shimkin@ee.technion.ac.il

ABSTRACT

Value decomposition algorithms have demonstrated state-of-the-art performance across various cooperative multi-agent benchmark tasks. However, they struggle to perform effective exploration, as the learning process typically involves searching over a large joint action space that increases with the number of agents. To address this, we introduce a novel algorithm for efficient exploration that combines a centralized decomposed critic with decentralized ensemble learning. Our approach leverages ensemble kurtosis as an uncertainty measure to guide exploration toward high-uncertainty states and actions. We also introduce an architecture for uncertainty-weighted value decomposition, where each component of the global Q -function is weighted using the corresponding agent uncertainty, resulting in overall variance reduction during training. We employ a hybrid approach for training the actors, combining on-policy and off-policy loss functions, and provide theoretical results that bound the bias in the actor gradient updates. Empirical evaluation shows that our approach is highly effective, outperforming state-of-the-art baselines on the most challenging maps in the Starcraft II benchmark.

Keywords Multi-agent, Reinforcement learning, Ensemble learning, Selective exploration

1 Introduction

In recent years, *centralized training with decentralized execution* (CTDE) [1] algorithms have achieved state-of-the-art performance on benchmark tasks [2, 3], emerging as a highly challenging, yet promising area of research. In CTDE, the sharing of learning parameters and information between agents is enabled during training, while in execution time agents are deployed in a fully decentralized manner. Value decomposition algorithms play a crucial role in the success of CTDE. This family of algorithms suggests decomposing the global value function into individual per-agent components via a centralized mixing network, enabling agents to learn and coordinate their actions within a team more efficiently. Yet, value decomposition algorithms suffer from several drawbacks. For instance, the monotonic constraint imposed on the mixing network is known to result in poor exploration properties and may lead to agents learning suboptimal policies [4]. Another problem arises by the training variance exacerbated due to the presence of multiple agents [5]. This is known to be especially challenging in multi-agent policy gradient (MAPG) methods in which a single agent can induce variance through the centralized critic, potentially disrupting the training process of all other agents. Both issues are further discussed in Sec. 4.

The main objective of this paper is to enable sample-efficient exploration by all agents while avoiding the negative effect of agents inducing variance through the centralized critic thus disrupting the learning process. To achieve this, each agent identifies high-uncertainty states and actions using information learned centrally by the critics. Uncertainty quantification serves two key purposes: adjusting the exploration level to avoid redundant exploration and reducing variance by down-weighting noisy samples. To estimate uncertainty we utilize an ensemble of critics (Fig. 1). We use the ensemble approach for several reasons. First, variability in the ensemble is an effective indicator of uncertainty. Second, previous works from the single agent domain show ensemble methods to be highly sample-efficient [6]. This is essential in multi-agent settings in which training time can be notoriously long. Instead of directly using the ensemble

variance, we propose utilizing ensemble kurtosis as an alternative measure of uncertainty, which is further discussed in Section 5. While kurtosis [7, 8] has been previously suggested to capture uncertainty, to the best of our knowledge this is the first work to utilize it in the context of RL.

Our work offers several contributions in the areas of multi-agent and ensemble learning:

- We introduce an MAPG architecture for uncertainty-weighted value decomposition, in which each component of Q_{tot} is weighted based on an individual agent’s uncertainty. To further reduce variance and improve sample efficiency, we suggest using interpolated actors [9], which are trained with a mix of on-policy and off-policy gradient updates. We also provide a theoretical analysis demonstrating that the bias in the gradient updates of our approach is bounded.
- We suggest utilizing ensemble kurtosis to perform efficient exploration. We show that kurtosis can be used to identify high-uncertainty states and prioritize actions for exploration.
- We introduce a novel approach to facilitating diversity in an ensemble of critics via Bhattacharyya distance regularization.

The rest of the paper is organized as follows: Section 2 provides the necessary background. Section 3 reviews related work. Section 4 describes key challenges in multi-agent and ensemble learning. Section 5 presents our approach for combining ensemble learning with value decomposition MARL. The main theoretical results are presented in Section 6. Section 7 presents the evaluation process and results. Sections 8-9 provide discussion and conclude the paper. Appendix G outlines the theoretical guarantees and provides detailed proofs.

2 Background

Decentralised partially observable Markov decision processes (Dec-POMDPs). A Dec-POMDP [10] models the interaction between cooperative agents and a partially observable environment. It is represented by the tuple $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \Omega, \mathcal{O}, r, \gamma \rangle$, where \mathcal{I} denotes a finite set of k agents, and $s \in \mathcal{S}$ denotes the true state of the environment. Each agent selects an action $a_i \in \mathcal{A}$, forming the joint action space $\mathbf{a} = [a_i]_{i=1}^k \in \mathcal{A}$. $r(s, \mathbf{a})$ denotes the shared reward for executing the set of joint actions in the environment. Executing the set of actions results in a transition to the next state s' , according to the transition function $\mathcal{P}(\cdot|s, \mathbf{a})$. The Dec-POMDP considers partially observable settings where each agent i receives a partial observation $o_i \in \Omega$ according to the observation probability function $\mathcal{O}(o_i|s, a_i)$. The discount factor is denoted by $\gamma \in [0, 1)$ and the joint action-observation history of the set of agents is denoted by $\tau \in \mathcal{T}$.

For a joint set of policies $\{\pi_i\}_{i=1}^k$ and corresponding learning parameters $\{\theta_i\}_{i=1}^k$, the parameterized policies are denoted by $\{\pi_{\theta_i}\}_{i=1}^k$. We use the notation $\bar{z}^i = f_{\theta_i}(\tau_i)$ to denote the logits of the i th actor network before applying the softmax function.

Value decomposition and monotonic constraints in multi-agent reinforcement learning. Value decomposition in MARL involves decomposing the overall state value function into individual per-agent value functions, enabling an efficient credit assignment by ensuring that agents receive rewards based on their specific contributions to the total cumulative reward. To ensure that maximizing the global Q function also maximizes the per-agent Q functions the Individual-Global-Maximum (IGM) condition must be enforced:

$$\operatorname{argmax}_{\mathbf{a}} Q_{\text{tot}}(\tau, \mathbf{a}) = \left\{ \begin{array}{c} \operatorname{argmax}_{a_1} Q_1(\tau_1, a_1) \\ \vdots \\ \operatorname{argmax}_{a_k} Q_k(\tau_k, a_k) \end{array} \right\}.$$

Value-decomposition networks (VDN) express the global value function as the sum of individual Q -values. Maintaining additivity ensures IGM consistency and enables CTDE. QMIX enhances VDN by employing a centralized mixing network to represent Q_{tot} as a non-linear and monotonic combination of individual value functions.

MAPG architecture and loss functions. For the underlying architecture, we partially adopt the actor-critic scheme proposed by DOP [5]. DOP is a MAPG method that employs a centralized decomposed critic with decentralized actors (Figure 1). The centralized critic is trained by minimizing the following loss:

$$\mathcal{L}_{\text{critics}}(\phi) = \mathbf{c} \mathcal{L}^{\text{on-TD}(\lambda)}(\phi) + (1 - \mathbf{c}) \mathcal{L}^{\text{off-TB}}(\phi), \quad (1)$$

where \mathbf{c} is the tuning factor, used to balance between off-policy and on-policy updates. To construct the loss function two buffers are maintained, samples from the off-policy buffer yield $L^{\text{off-TB}}$, while $\mathcal{L}^{\text{on-TD}(\lambda)}$ is calculated with samples

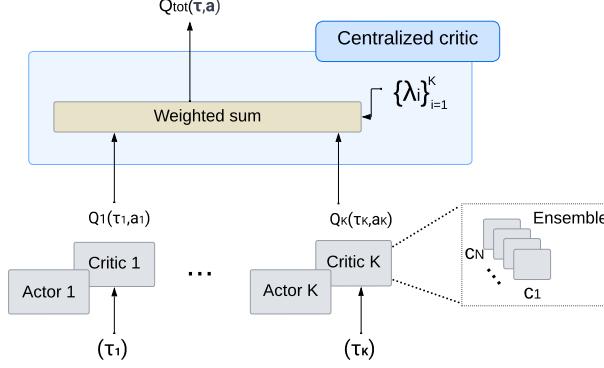


Figure 1: Ensemble MIX architecture

from the smaller on-policy buffer that stores only recent experiences. $\mathcal{L}_{\text{off}}(\phi)$ is defined as $E_{\pi}[(y^{\text{off}} - Q_{\text{tot}}^{\phi}(\tau, \mathbf{a}))^2]$, and y^{off} is the tree backup update target, as suggested by [5, 11]:

$$y^{\text{off}} = Q_{\text{tot}}^{\phi'}(\tau, \mathbf{a}) + \sum_{t=0}^{m-1} \prod_{l=0}^t \lambda \pi(\mathbf{a}_l | \tau_l) [r_t + \gamma E_{\mathbf{a}_{t+1} \sim \pi}[Q_{\text{tot}}^{\phi'}(\tau_{t+1}, \mathbf{a}_{t+1})] - Q_{\text{tot}}^{\phi'}(\tau_t, \mathbf{a}_t)], \quad (2)$$

where $\tau_0 = \tau, \mathbf{a}_0 = \mathbf{a}$. The target mixing network is denoted by $Q_{\text{tot}}^{\phi'}$ and the joint policy π is defined by the product of individual policies: $\pi(\mathbf{a}_l | \tau_l) = \prod_{i=1}^k \pi_i(a_i^l | \tau_i^l)$. The second loss is defined as $\mathcal{L}^{\text{on-TD}(\lambda)}(\phi) = E_{\pi}[(y^{\text{on}} - Q_{\text{tot}}^{\phi}(\tau, \mathbf{a}))^2]$, with the following TD(λ) target:

$$y^{\text{on}} = Q_{\text{tot}}^{\phi'}(\tau, \mathbf{a}) + \sum_{t=0}^{\infty} (\gamma \lambda)^t [r_t + \gamma Q_{\text{tot}}^{\phi'}(\tau_{t+1}, \mathbf{a}_{t+1}) - Q_{\text{tot}}^{\phi'}(\tau_t, \mathbf{a}_t)], \quad (3)$$

where Q_{tot} is a linear combination of individual Q-function, formed by the (monotonically constrained) mixing network. Note that individual critics $\{Q_i^{\phi_i}\}_{i=1}^k$ are learned implicitly by minimizing the global objective $\mathcal{L}_{\text{critics}}(\phi)$. Decomposing the central critic to individual critics yields per-agent Q-functions $\{Q_i\}_{i=1}^k$ that used by each corresponding agent. The i^{th} actor, parameterized by θ_i , is trained by maximizing a performance objective with the following gradient:

$$\nabla_{\theta_i} J(\theta_i) = E_{\pi_i}[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i | \tau_i) U_i^{\phi_i}(\tau_i, a_i)], \quad (4)$$

where U_i is the advantage function:

$$U_i^{\phi_i}(\tau_i, a_i) = \lambda_i(\tau)(Q_i^{\phi_i}(\tau_i, a_i) - \sum_{x \in \mathcal{A}} \pi_{\theta_i}(x | \tau_i) Q_i^{\phi_i}(\tau_i, x)). \quad (5)$$

Kurtosis. The kurtosis of a random variable x is defined as the fourth standardized moment [12]:

$$\kappa[x] = E[(\frac{x - \mu}{\sigma})^4] = \frac{E[(x - \mu)]^4}{(E[(x - \mu)^2])^2}, \quad (6)$$

where σ denotes the standard deviation and μ the mean of the distribution. When measuring the kurtosis of N samples it is custom to use the excess kurtosis:

$$g_2(\{x_i\}_{i=1}^N) = \frac{\frac{1}{n} \sum_{i=1}^N (x_i - \bar{x})^4}{[\frac{1}{N} \sum_{i=1}^n (x_i - \bar{x})^2]^2} - 3, \quad (7)$$

where x_i is the i^{th} sample, \bar{x} denotes the samples mean, and 3 is the kurtosis of the normal distribution. In our work, each sample corresponds to a distinct ensemble member, and $\kappa_{Q_i}^{\phi_i}(\tau_i, a_i) = g_2(\{Q_{i,j}^{\phi_i}(\tau_i, a_j)\}_{j=1}^N)$ denotes the i^{th} critic ensemble's predictions for action a_i . Note that the kurtosis indicates the distribution's tails and the presence of

outliers within it. Subtracting the constant 3 sets the normal distribution as a reference point. Previous works from other domains made a similar use of the normal distribution to estimate the excess kurtosis. For instance, [13] estimated value at risk (VaR) in financial time series using excess kurtosis, measuring it relatively to the normal distribution. [14] applied kurtosis for outlier detection in financial time series and used excess kurtosis as a criterion for portfolio selection.

3 Related Work

Several methods integrated MAPG architectures with value decomposition. DOP [5] proposed a decomposed shared critic that improves sample efficiency by combining on-policy and off-policy loss functions to train the critic. PAC [15] suggested an off-policy actor-critic with assistive information to enhance learning. [16] addressed efficiency in MARL by grouping agents before the decomposition process to enable a more efficient collaboration. Exploration in MARL has also been addressed in several works. [17] suggested collaborative and efficient exploration by guiding the agents to explore only meaningful states. [15] added an entropy term to the central critic with a decaying temperature to avoid over-exploration. [18] suggested efficient exploration via dynamically adjusted entropy temperature that performs more exploration when higher return is expected. Diversity in ensembles has been mainly addressed in single-agent RL. These include maximizing inequality measures (e.g., the Theil index) [19], using random initialization of learning parameters, and training ensemble members on different sample sets [20].

4 Challenges in Ensemble and Multi-Agent Reinforcement Learning

Variance in multi-agent RL. Variance in RL typically refers to the variability in rewards received by an agent as it interacts with the environment. It is a measure of how much the actual rewards obtained by the agent deviate from the expected or average reward. High variance can lead to inconsistent and unstable learning, making it difficult for the agent to converge to an optimal policy or value function. This issue is further exacerbated in multi-agent settings, as MARL methods suffer from higher variance during training compared to their single-agent counterpart [21]. Unlike single-agent settings, where the variance of an individual agent usually stems only from randomness in the reward or the environment, in multi-agent settings, both suboptimally and exploration of other agents induce variance. This is known to be especially challenging in (multi-agent) policy gradient methods in which a single agent can induce variance on the entire system through the centralized critic, disrupting the training process of the other agents. The focus of this work is on centralized training methods, where variance propagates between agents through the centralized critic/mixing network. As a result, $\text{Var}[Q_{\text{tot}}]$ can become very large. For instance, assuming a linearly decomposed critic, denoted by $Q_{\text{tot}}^{\phi}(\tau, \mathbf{a}) = \sum_i \lambda_i(\tau) Q_i^{\phi_i}(\tau_i, a_i)$, its variance $\text{Var}_{\mathbf{a} \sim \pi}[Q_{\text{tot}}^{\phi}(\tau, \mathbf{a})]$, can be expressed as follows:

$$\text{Var}_{a_1 \sim \pi_1, \dots, a_k \sim \pi_k} [\lambda_1 Q_1^{\phi_1}(\tau_1, a_1) + \dots + \lambda_k Q_k^{\phi_k}(\tau_k, a_k)]. \quad (8)$$

It can be seen the variance of Q_{tot} depends on the actions taken by all the agents and their individual Q -functions. Since Q_{tot} is backpropagated back to all critics, each agent is exposed to variance induced by the other agents.

Ensemble diversity. Recent studies show that the effectiveness of ensemble-based techniques heavily relies on promoting diversity among its members [22]. We define diversity as the variability or differences in the representation learned by individual members within the ensemble. Lack of diversity, or homogeneity, is a phenomenon commonly observed in ensemble-based algorithms, in which different ensemble members converge to a similar solution [19]. Homogeneity can significantly impact performance, as it tends to persist even when ensemble members are trained on different samples, often leading them to converge toward similar representations despite slight differences in the training data. Various methods have been proposed to enhance diversity in ensembles. These include maximizing inequality measures (e.g., the Theil index) [20], using random initialization of learning parameters, and training ensemble members on different sample sets.

Exploration in multi-agent RL. In multi-agent systems, the joint action space expands exponentially as the number of agents increases, making effective exploration more challenging. Moreover, inefficient or random exploration in the joint action space can exacerbate learning instability and increase overall training time [17]. Applying maximum entropy methods, such as the soft actor-critic scheme [23], commonly used in single-agent reinforcement learning, to multi-agent environments can lead to excessive exploration and inefficient search of the action space. This is because maximizing entropy encourages agents to explore as randomly as possible [24]. In the next section, we propose a way to cope with those challenges with a cost as minimal as possible, in terms of sample efficiency.

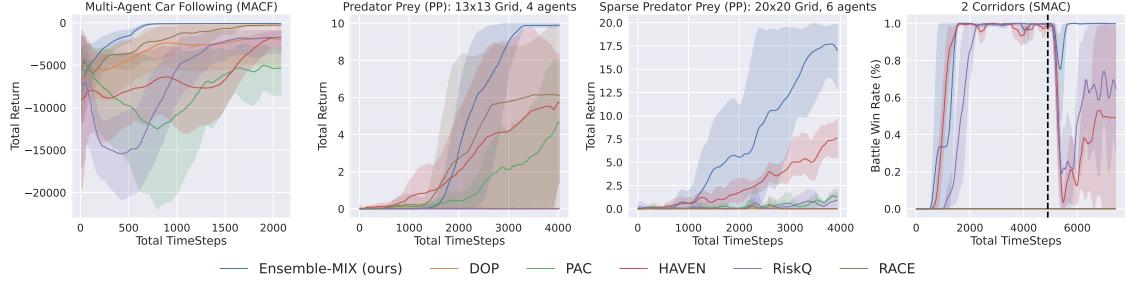


Figure 2: Results on predator-prey, MACF, and 2 corridors map (SMAC)

5 Ensemble Mix

This section presents our method for efficient exploration. Our approach comprises 3 main components: (1) uncertainty-weighted learning of Q_{tot} , combined with actors trained with mixed off-and-on loss functions, (2) action selection through kurtosis-based prioritization, and (3) diversity regularization within the ensemble based on Bhattacharyya distance.

5.1 Uncertainty weighted decomposed critic with decentralized actors

Each individual critic i is composed of an ensemble of N sub-critics:

$$Q_i^{\phi_i}(\boldsymbol{\tau}, a_i) = \frac{1}{N} \sum_{j=1}^N Q_{i,j}^{\phi_i}(\tau_i, a_i). \quad (9)$$

The mean Q -value of the i_{th} agent is simply denoted by $Q_i^{\phi_i}$, while each member in the ensemble is denoted by $Q_{i,j}^{\phi_i}$. The global Q -function is decomposed into a linear combination of individual functions:

$$Q_{tot}^{\phi}(\boldsymbol{\tau}, \mathbf{a}) = \sum_{i=1}^K k_i(\tau_i, a_i) \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) + b(\boldsymbol{\tau}), \quad (10)$$

with K denotes the number of agents. To maintain the monotonic constraint, as suggested by QMIX, the mixing network coefficients λ_i are restricted to be positive. Q_{tot} comprises several additional components: a learnable bias, denoted by b , and uncertainty weights K_i , where the novelty of our approach lies in the addition of k_i to the value decomposition process. Our proposed architecture is depicted in Figure 1. In the rest of the subsection, we present the specification of the weighting function k_i and the intuition underlying this approach.

Uncertainty weighting. To cope with the noisy and high variance updates previously discussed in section 4, we propose uncertainty weighting value decomposition in which Q_{tot} is weighted based on ensemble uncertainty. Down-weighting Q_{tot} alleviates the negative impact of high variance samples and reduces the noise such updates induce on the centralized critic. Note that we handle each component of Q_{tot} separately, meaning each Q_i is weighted based on the corresponding uncertainty of the i_{th} critic. A single weighting, implemented as the sum or mean of uncertainties can result in an overly pessimistic shared critic, as one agent's Q -function is down-weighted due to the high variance of other agents. Previous studies from the single agent domain show such pessimism to hinder exploration [25] and hurt performance. This issue can persist in the multi-agent domain as a pessimistic shared critic is decomposed into over-pessimistic individual critics. To detect noisy samples we employ the kurtosis of the ensemble, which is considered more effective than variance in terms of outlier detection [12], as it allows for more variability around the mean while punishing positive excess-tailedness (outliers). We adopt the weight function proposed by [20] and define k_i as follows :

$$k_i(\tau_i, a_i) = 0.5 + \mathcal{S}(-C_1 \kappa_{Q_i}^{\phi_i}(\tau_i, a_i)), \quad (11)$$

here, $\mathcal{S} : \mathbb{R} \rightarrow (0, 1)$ denotes the sigmoid function. $\kappa_{Q_i}^{\phi_i}(\tau_i, a_i)$ is the kurtosis of the i_{th} critic ensemble's predictions for action a_i . C_1 is a scaling factor. Note that the kurtosis remains positive, as it is not computed relative to the normal distribution, and no subtraction of 3 takes place, ensuring that $\frac{1}{2} \leq k_i(\tau_i, a_i) \leq 1$.

Underlying architecture and loss functions. For the underlying architecture, we partially adopt the actor-critic scheme proposed by DOP [5]. DOP suggests that the critics be trained with a mix of off-policy and on-policy samples, improving both sample efficiency and overall performance. However, the authors of DOP limit themselves to using

off-policy samples only for training the critics while the actors are trained with on-policy experiences. To improve their scheme, we propose to train the actors by combining gradients from on-policy and off-policy loss functions:

$$\begin{aligned} D_\nu^\beta(\pi_i, \boldsymbol{\pi}_{-i}) \\ = (1 - \nu) \mathbb{E}_{\rho^\boldsymbol{\pi}, \pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) U_{\pi_i}^{\phi_i}(\tau_i, a_i)] \\ + \nu \mathbb{E}_{\rho^\beta} [\nabla_{\theta_i} \bar{Q}_{\text{tot}}^\phi(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i}))], \end{aligned} \quad (12)$$

where $\bar{Q}_{\text{tot}}^\phi(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i})) = \mathbb{E}_{\pi_i} [Q_{\text{tot}}^\phi(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i}))]$, and ν is the relative scaling factors. The joint off-policy is denoted by $\beta(\mathbf{a} | \boldsymbol{\tau}) = \prod_{i=1}^k \beta_i(a_i, \tau_i)$, and ρ^β is the joint off-policy state distribution. Note that using off-policy data to train the actors is more sample-efficient. Still, it may come at the cost of stability during training [26] which is especially important in multi-agent settings. Combining both approaches is known to produce good results [9] with higher stability. Finally, the critics are trained with a mix of off-policy and on-policy data, similar to DOP and according to Eq. 1. Note that the weighting is applied only when calculating the Q-function for the next state $Q_{\text{tot}}^\phi(\boldsymbol{\tau}_{t+1}, \cdot)$.

5.2 Exploration based on ensemble kurtosis

We propose leveraging ensemble kurtosis to adjust the level of exploration and prioritize actions associated with higher kurtosis values. Our algorithm applies kurtosis in a twofold manner. First, each agent detects high-uncertainty states based on the mean value of the kurtoses taken over all actions in the action space. Second, once a relevant state is detected, prioritization is applied by weighting actions according to their kurtosis. This two-step approach allows us to meet the efficiency requirement discussed in Sec. 4 and avoid over-exploration.

In each iteration of action selection, every i_{th} agent calculates the kurtosis over all M_i actions:

$$\bar{g}_i(\tau_i, \{Q^{\phi_{i,j}}\}_{j=1}^N) = \frac{\sum_{z=1}^{M_i} g_2(\{Q^{\phi_{i,j}}(\tau_i, a_z)\}_{j=1}^N)}{M_i} \quad (13)$$

where g_2 is the excess kurtosis described in Eq. 7. If $\bar{g}_i > 0$ then an exploration step is performed by adding the kurtosis of each action to the corresponding logits, otherwise, a standard action selection is executed. Conditioning on the positivity of \bar{g}_i guarantees exploration is performed only on areas of positive excess kurtosis. Formally, given the M_i logits of actor i on a given state, denoted by $\bar{z}^i = \{z_1^i, \dots, z_{M_i}^i\}$, and the corresponded ensemble kurtosis, denoted by $\{\kappa_{Q_i}^{\phi_i}(\tau_i, a_1), \dots, \kappa_{Q_i}^{\phi_i}(\tau_i, a_{M_i})\}$, then the j_{th} weighted logit of actor i is given by:

$$\tilde{z}_j^i = \begin{cases} z_j^i + \beta \kappa_{Q_i}^{\phi_i}(\tau_i, a_i), & \text{if } g_i > 0 \\ z_j^i, & \text{otherwise} \end{cases} \quad (14)$$

where β is a hyperparameter, chosen to be sufficiently small so as not to dominate the value of \tilde{z}_j^i . Similarly to standard actor-critic algorithms, the softmax operator is applied to the weighted logits to create a distribution for action selection. The complete algorithm is provided in Appendix A. In terms of efficiency, our approach is preferable to existing exploration methods for two reasons. First, unlike entropy maximization and ϵ -greedy methods, our exploration procedure is performed only in high uncertainty states that correspond to positive excess kurtosis. Second, actions are visited in an informed manner, i.e., once visited, the priority given to an action is decreased with respect to the ensemble kurtosis. Note that prior to our work, [20, 27] suggested leveraging ensemble variance as an exploration bonus. In section 8 we discuss the difference between the two approaches, emphasizing efficiency improvement, which is especially crucial for successful exploration in multi-agent settings.

5.3 Diversity in ensemble via Bhattacharyya regularization

To promote diversity in the ensemble we propose utilizing the Bhattacharyya distance as a regularization term. Bhattacharyya distance is widely used in classification tasks due to its ability to measure distributional overlap. It can effectively measure separability between features [28] from different classes, and aid in clustering tasks. To our knowledge, there have been no prior efforts to apply the Bhattacharyya distance in the context of RL, despite studies indicating it to be more stable than other similarities measures such as KL divergence [29]. The Bhattacharyya term is calculated as the sum of distances between the mean Q-function of the ensemble $Q_i^{\phi_i}$ and all the distinct Q-functions:

$$\delta_{B_{\text{total}}}^{Q_i}(\tau_i) = \sum_{j=1}^N \delta_B \left(\sigma(Q_i^{\phi_i}(\tau_i, \cdot)), \sigma(Q^{\phi_{i,j}}(\tau_i, \cdot)) \right). \quad (15)$$

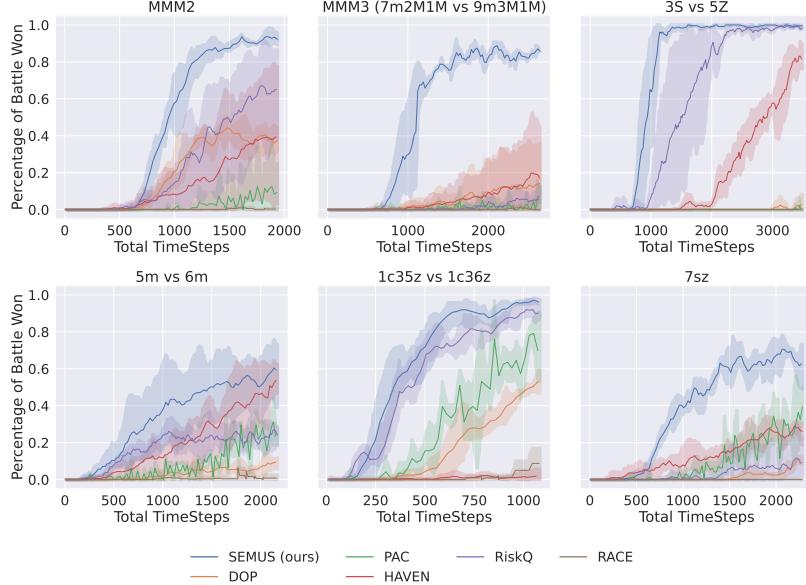


Figure 3: Results on the SMAC benchmark

The sum extends over all members within the ensemble and σ denotes the soft-max function. The distance δ_B is given by [30]:

$$-\ln \left(\sum_{a \in \mathcal{A}} \sqrt{\sigma(Q_i^{\phi_i}(\tau_i, a))\sigma(Q_{i,j}^{\phi_i}(\tau_i, a))} \right).$$

The regularization term is added to the critic's loss with a negative sign:

$$\mathcal{L}_{\text{total}}(\phi) = \mathcal{L}_{\text{critics}}(\phi) - C_2 \sum_{i=1}^K \delta_{B_{\text{total}}}^{Q_i}(\tau_i). \quad (16)$$

To prevent $\delta_{B_{\text{total}}}^{Q_i}$ from dominating the loss function, we set C_2 to be sufficiently small (see Appendix B).

6 Bounds on the Bias of Mixed Actors

In this section, we bound the bias of the gradient update for each individual agent. Since each agent independently updates its policy, analyzing the bias on a per-agent basis is more intuitive. To do so, we first define the true Q-function of the i_{th} agent [5]:

$$Q_i^\pi(\tau, a_i) = \sum_{\mathbf{a}_{-i}} \pi_{-i}(\mathbf{a}_{-i} | \tau_{-i}) Q_{\text{tot}}^\pi(\tau, (a_i, \mathbf{a}_{-i})), \quad (17)$$

and the corresponding gradient update:

$$\nabla_{\theta_i} J(\pi_i, \pi_{-i}) = \mathbb{E}_{\tau \sim \rho^\pi, a_i \sim \pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) Q_i^\pi(\tau, a_i)], \quad (18)$$

where $\pi_{-i}(\mathbf{a}_{-i} | \tau_{-i}) = \prod_{j \neq i} \pi_j(a_j | \tau_j)$. We denote the discrepancy between the true Q-function and the approximated function of the i_{th} agent by:

$$Q_{\phi_i}^\pi(\tau, a_i) = Q_i^\pi(\tau, a_i) - \lambda_i(\tau) Q_i^{\phi_i}(\tau_i, a_i). \quad (19)$$

The bias can then be bound by estimating the difference between the gradient of the true objective $\nabla_{\theta_i} J(\pi_i, \pi_{-i})$ and the mixed gradient update $D_\nu^\beta(\pi_i, \pi_{-i})$ we use in the paper.

Proposition 1. Let the maximal KL divergence between two policies π, β be denoted by $D_{\max}^{\text{KL}}(\pi, \beta) = \max_{\tau} D_{\text{KL}}(\pi(\cdot | \tau), \beta(\cdot | \tau))$, and let $\omega_1 = \max_{\tau, a} \left\| \nabla_{\theta_i} \log (\pi_i(a_i | \tau_i; \theta_i)) \lambda_i(\tau) Q_{\phi_i}^\pi(\tau, a_i) \right\|_1$, $\omega_2 =$

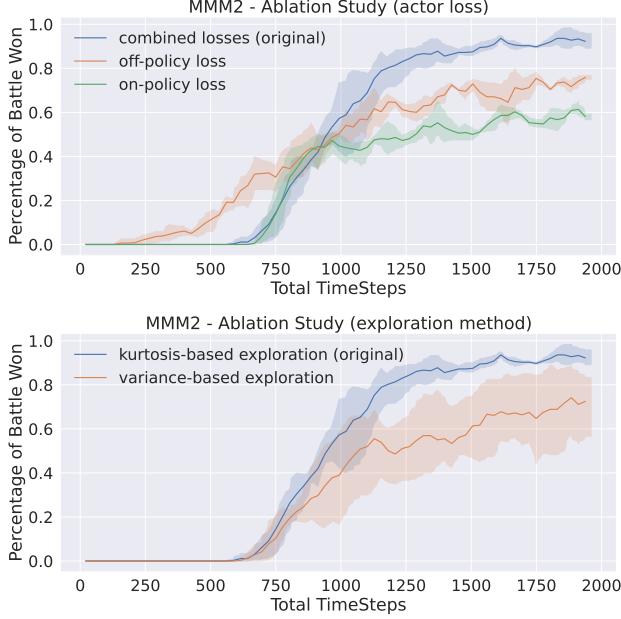


Figure 4: Ablation study on MMM2 map. The upper graph tests different variations of gradient updates, and the bottom graph tests different exploration methods

$\max_{\boldsymbol{\tau}} \left\| \nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right\|_1$. Then, the learning bias of each agent can be bounded as follows:

$$\left\| \nabla_{\theta_i} J(\pi_i, \boldsymbol{\pi}_{-i}) - D_{\nu}^{\beta}(\pi_i, \boldsymbol{\pi}_{-i}) \right\|_1 \leq \frac{\omega_1}{1-\gamma} + \frac{2\omega_2\nu\gamma}{(1-\gamma)^2} \sqrt{D_{\max}^{\text{KL}}(\boldsymbol{\pi}, \boldsymbol{\beta})}. \quad (20)$$

The bound tells us that the bias of each agent’s gradient update depends on two main factors: errors in the approximation of $Q_i^{\boldsymbol{\pi}}$, as measured by $Q_i^{\phi_i}$, and second, how much the joint on-policy diverge from the (joint) policy used to collect the off-policy data (similarly to the bound in the single agent case [9]). A detailed proof is given in Appendix G.

7 Experiments and Results

The objective of the empirical evaluation is to assess the performance of our approach. We measure and compare the percentage of battle wins of the team of agents on the challenging Starcraft *II* benchmark on different maps and scenarios [31]. The hyperparameters and network architecture are fixed for all of the experiments, full specification is given in Appendix B. We use 3 different seeds for each map. A comparison is conducted between our approach and DOP, MAIC, PAC, CIA, and NA2Q. All baseline methods are described in the related work section. We set the size of the ensemble to be $N = 10$. Each critic in the ensemble is implemented with a fully connected 2 layers network. To ensure a fair comparison, all network architectures are identical to those of DOP. We use 2 different Starcraft *II* maps: MMM2, MMM3 (7m2M1M vs 9m3M1M [32]), 3s vs 5z, 27m vs 30m, 1c35z vs 1c36z, 7sz, and 2 corridors. Our main focus is on scenarios that require the agents to learn diverse skills or exhibit some level of exploration to solve the task. Note that we use the term diversity to describe inter-agent diversification, meaning different agents acquiring different sets of skills. Further explanation of the SMAC maps used in our experiments is provided in Appendix C.

Performance evaluation. Figure 3 describes the total score throughout the training process, shaded areas around the learning curves represent the standard deviation. All measurements are taken as follows: every T steps the training process is paused and a fixed amount of U testing episodes is executed. The mean score over the test episodes is recorded and used in the graphs. To maintain a decentralized execution, in test time the actors in our method execute a standard softmax policy, without the ensemble kurtosis added to the logits. It can be seen that Ensemble mix outperforms all other methods on the tested maps. We specifically excel on the MMM2 and MMM3 which are considered extremely challenging and categorized as super-hard. Note that the enhanced exploration properties of our approach lead to good performance in maps that pose challenges such as diversity among agents (e.g., MMM2) or intensive exploration scenario such as 27m vs 30m, which requires extensive exploration due to the large action space that increases with the number of agents. When evaluating baseline performance, no single method consistently outperformed on all maps. PAC achieved strong performance on maps like 27m vs 30m, primarily due to the implementation of entropy

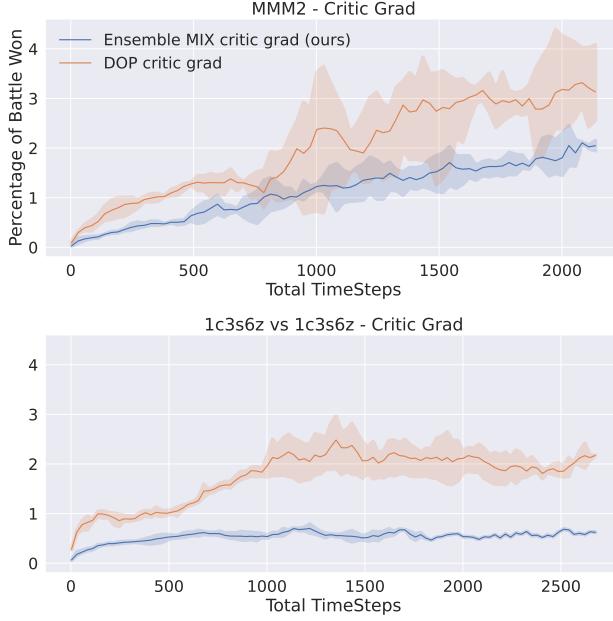


Figure 5: Critics grad during training (DOP versus Ensemble-mix)

regularization. In contrast, MAIC excelled on maps that require agents’ diversity (e.g. MMM2, MMM3), attributed to the use of mutual information regularization between agents. NA2Q and CIA performed relatively similar on all maps. DOP, an older method, generally underperformed compared to other baselines, except on the MMM maps. This result can be partially explained by DOP utilization of softmax policies, which enable greater stochasticity and may promote more diverse behavior.

2 corridors. To show that our method performs effective exploration, we conduct an additional experiment with the 2 corridors map [4], in which 2 allies fight an enemy unit with the shorter corridor closed mid-training. This requires the agents to explore the terrain and learn how to use the second corridor when attacking the enemy unit. The results show that our method adapts to the change faster than the baseline methods while achieving the best overall results.

8 Discussion

Computation requirements. All experiments are conducted with an ensemble of size $N = 10$. Achieving good performance with a relatively small value of N is made possible due to the diversity regularization we apply via the Bhattacharyya distance. This finding is consistent with previous studies that show diversity among members of the ensemble can reduce the size of N from a few hundred to less than a few dozen [22, 19].

Comprison to previous works. Prior to our work, [20, 27] suggested leveraging ensemble variance as an exploration bonus for Q -learning algorithms. Applying it directly to actor-critic and multi-agent settings yields suboptimal performance, as demonstrated in our ablation studies. Our approach differs from [27] by incorporating excess kurtosis and selectively applying it as exploration bonuses. This ensures that agents do not always engage in exploration and maintain higher efficiency, which becomes critical as the joint action space grows with the number of agents.

Variance reduction. One of the key claims we made throughout the paper is that uncertainty weighting can reduce variance and stabilize the learning process in multi-agent settings. We verify variance reduction by measuring the total norm of the neural network (NN) parameters’ gradients. Measurements are conducted in each training step for the critics’ networks. Figure 5 presents the variance of our approach versus DOP on two maps, 1c3s6z vs 1c3s6z and MMM2. It can be seen that our scheme maintains a lower gradient compared to DOP which exhibits higher volatility with peaks that can potentially destabilize learning and lead to slower convergence.

8.1 Ablation study

Variance versus kurtosis for exploration. In a separate set of experiments, we attempted to use variance rather than kurtosis to adjust exploration. The process is similar to the algorithm suggested in Section 5.2, but instead of applying it only to high-uncertainty states that correspond to positive excess kurtosis, the variance is added to the actor logits in all

time steps, regardless of the level of uncertainty. Results are presented in the bottom part of Fig. 4 for the MMM2 map. It can be seen that kurtosis-based exploration achieves superior results compared to the variance-based approach.

Actors loss and off-policy samples. We conducted an additional set of experiments to evaluate the impact of training actors using a mix of off-policy and on-policy samples. Results for the MMM2 map are shown in upper part of Fig. 4. We compare three approaches: training the actors solely with on-policy samples using the original DOP loss, training only with off-policy samples, and training with a mix of off-policy and on-policy loss functions. While using off-policy achieves better results compared to using only on-policy data, the combined approach outperforms both methods,

9 Conclusion

This paper presents a novel algorithm for efficient exploration in MARL, leveraging ensemble kurtosis to guide the visitation of high-uncertainty states and actions. To address the high variance in MARL, we propose an uncertainty-weighted value decomposition architecture, where components of the global Q -function are weighted using individual agent uncertainties. By training the actors with a combination of on-policy and off-policy loss functions our method achieves both sample efficiency and higher stability in comparison to relying solely on off-policy learning. We also highlight the importance of ensemble diversity, introducing a simple approach to enhance diversity using Bhattacharyya regularization. Our method generalizes to environments with homogeneous or heterogeneous agents with different action spaces. Empirical results demonstrate superior performance over state-of-the-art baselines on challenging StarCraft II maps.

References

- [1] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10199–10210, 2020.
- [2] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869*, 2020.
- [3] Yongsheng Mei, Hanhan Zhou, and Tian Lan. Projection-optimal monotonic value function factorization in multi-agent reinforcement learning. In *Proceedings of the 2024 International Conference on Autonomous Agents and Multiagent Systems*, 2024.
- [4] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. Maven: Multi-agent variational exploration. *Advances in neural information processing systems*, 32, 2019.
- [5] Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. Dop: Off-policy multi-agent decomposed policy gradients. In *International conference on learning representations*, 2020.
- [6] Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021.
- [7] Jon Anda, Alexander Golub, and Elena Strukova. Economics of climate change under uncertainty: Benefits of flexibility. *Energy Policy*, 37(4):1345–1355, 2009.
- [8] IP Zakharov and OA Botsyura. Calculation of expanded uncertainty in measurements using the kurtosis method when implementing a bayesian approach. *Measurement techniques*, 62:327–331, 2019.
- [9] Shixiang Shane Gu, Timothy Lillicrap, Richard E Turner, Zoubin Ghahramani, Bernhard Schölkopf, and Sergey Levine. Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.
- [10] Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.
- [11] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.
- [12] Peter H Westfall. Kurtosis as peakedness, 1905–2014. rip. *The American Statistician*, 68(3):191–195, 2014.
- [13] Alexandros Gabrielsen, Axel Kirchner, Zhuoshi Liu, and Paolo Zagaglia. Forecasting value-at-risk with time-varying variance, skewness and kurtosis in an exponential weighted moving average framework. *Annals of Financial Economics*, 10(01):1550005, 2015.
- [14] Nicola Loperfido. Kurtosis-based projection pursuit for outlier detection in financial time series. *The European Journal of Finance*, 26(2-3):142–164, 2020.

- [15] Hanhan Zhou, Tian Lan, and Vaneet Aggarwal. Pac: Assisted value factorization with counterfactual predictions in multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35:15757–15769, 2022.
- [16] Yifan Zang, Jinmin He, Kai Li, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Automatic grouping for efficient cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [17] Yonghyeon Jo, Sunwoo Lee, Junghyuk Yeom, and Seungyul Han. Fox: Formation-aware exploration in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12985–12994, 2024.
- [18] Woojun Kim and Youngchul Sung. An adaptive entropy-regularization framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 16829–16852. PMLR, 2023.
- [19] Hassam Sheikh, Mariano Phielipp, and Ladislau Boloni. Maximizing ensemble diversity in deep reinforcement learning. In *International Conference on Learning Representations*, 2021.
- [20] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*, pages 6131–6141. PMLR, 2021.
- [21] Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Haifeng Zhang, David Mguni, Jun Wang, Yaodong Yang, et al. Settling the variance of multi-agent policy gradients. *Advances in Neural Information Processing Systems*, 34:13458–13470, 2021.
- [22] Chao Li, Chen Gong, Qiang He, and Xinwen Hou. Keep various trajectories: Promoting exploration of ensemble policies in continuous control. *Advances in Neural Information Processing Systems*, 36, 2024.
- [23] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.
- [24] Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations*, 2022.
- [25] Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. *Advances in Neural Information Processing Systems*, 32, 2019.
- [26] Nessrine Hammami and Kim Khoa Nguyen. On-policy vs. off-policy deep reinforcement learning for resource allocation in open radio access network. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1461–1466, 2022.
- [27] Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.
- [28] Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9172–9182, 2022.
- [29] Yifei Lou, Andrei Irimia, Patricio A Vela, Micah C Chambers, John D Van Horn, Paul M Vespa, and Allen R Tannenbaum. Multimodal deformable registration of traumatic brain injury mr volumes via the bhattacharyya distance. *IEEE Transactions on Biomedical Engineering*, 60(9):2511–2520, 2013.
- [30] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, 1967.
- [31] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.
- [32] Wei-Fang Sun, Cheng-Kuang Lee, Simon See, and Chun-Yi Lee. A unified framework for factorizing distributional value functions for multi-agent reinforcement learning. *Journal of Machine Learning Research*, 24(220):1–32, 2023.
- [33] Owen Lockwood and Mei Si. A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 155–162, 2022.
- [34] Eduard Hofer, Martina Kloos, Bernard Krzykacz-Hausmann, Jörg Peschke, and Martin Wolterbeck. An approximate epistemic uncertainty analysis approach in the presence of epistemic and aleatory uncertainties. *Reliability Engineering & System Safety*, 77(3):229–238, 2002.

- [35] Gregory Kahn, Tianhao Zhang, Sergey Levine, and Pieter Abbeel. Plato: Policy learning using adaptive trajectory optimization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3342–3349. IEEE, 2017.

Algorithm 1: Action Selection

Input: actor policies $\{\pi_i(a_i|\tau_i; \theta_i)\}_{i=1}^k$, individual Q-functions $\{Q_i^{\phi_i}\}_{i=1}^k$, joint trajectory τ , num_agents, action space sizes $\{M_i\}_{i=1}^k$

Output: Selected actions set \mathbf{a}

for agent $i = 1$ to num_agents **do**

 Get action logits $\bar{z}^i = f_{\theta_i}(\tau_i)$

 Calculate kurtosis over all actions for each agent:

$$\bar{g}_i(\tau_i, \{Q_j^{\phi_{i,j}}\}_{j=1}^N) = \frac{\sum_{z=1}^{M_i} g_2(\{Q_j^{\phi_{i,j}}(\tau_i, a_z)\}_{j=1}^N)}{M_i}. \quad (21)$$

 Update action logits according to Eq. 14 and the value of \bar{g}_i

 Select action $a_i \sim \sigma(\tilde{z}_j)$ with the updated logits using epsilon greedy policy.

end for

Algorithm 2: Ensemble MIX Algorithm

Input: L (num of episodes), T_{max} (episode length), K (num of agents)

Initialize centralized mixing network Q_{tot}^ϕ , for each agent i : ensemble critics $\{Q_j^{\phi_{i,j}}\}_{j=1}^N$, actor networks $\{\pi_{\theta_i}\}_{i=1}^k$

Initialize target centralized critic $Q_{tot}^{\phi'} = Q_{tot}^\phi$, target ensemble critics $Q_j^{\phi'_{i,j}} = Q_j^{\phi_{i,j}}$

Initialize on-policy replay buffer \mathcal{D}_{on} , off-policy replay buffer \mathcal{D}_{off}

for episode = 1 to L **do**

 Initialize environment and agents

for $t = 1$ to T **do**

 Observe state s_t

 obtain joint action set \mathbf{a}_t by calling Algorithm 1

 Execute joint actions \mathbf{a}_t

 Receive reward r_t and next state s_{t+1}

 Store transition $(s_t, \mathbf{a}_t, r_t, s_{t+1})$ in replay buffers \mathcal{D}_{on} and \mathcal{D}_{off}

 Sample minibatch N_1 from replay buffer \mathcal{D}_{on}

 Sample minibatch N_2 from replay buffer \mathcal{D}_{off}

 Calculate the Bhattacharyya loss $\delta_{B_{total}}^{Q_i}$ for sampled batches $\{N_2, N_1\}$

 Update the critics with the loss in Eq. 16 using a mix of on-policy and off-policy losses.

 Update decentralized actors using a mix of on-policy and off-policy loss functions (Eq. ??)

if $t \bmod X = 0$ **then**

 Update target networks: $Q_j^{\phi'_{i,j}} = Q_j^{\phi_{i,j}}$, $Q_{tot}^{\phi'} = Q_{tot}^\phi$

end if

end for

end for

A Algorithm

The full specifications of our approach are described in Algorithms 1& 2. The main training loop begins in algorithm 2. The learning parameters are initialized at the beginning of the algorithm and the training loop is executed iteratively for X episodes. Action selection is performed in algorithm 1. Note that our proposed exploration is executed when the excess kurtosis is positive, otherwise, a standard epsilon greedy selection is performed. The critics' loss function can be divided into two parts, the Bhattacharyya term which ensures diversity in the ensemble, and the critic loss, which utilizes the weighted Q_{tot} to reduce variance in the centralized critic. Note that both the critics and the actors are trained with a mix of offline and online data. The online data is sampled from a smaller buffer \mathcal{D}_{on} and $|N_2| > |N_1|$.

B Training Parameters and computation resources

Configurations and hyper-parameters values are provided in table 1.

Table 1: Hyperparameter Table - SMAC

Algorithm Parameters	
HyperParameters	Value
T (train interval)	20K
U (num test episodes)	24
C_1	0.01
C_2	0.001
β	0.001
v	0.5
N_1	32
N_2	5000
ensemble size (N)	10

All experiments are conducted on stations with 64 Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.10GHz, 251G RAM, and NVIDIA GeForce GTX 1080 Ti 12GB.

C Starcraft II scenarios

Our main focus is on scenarios that require the agents to learn diverse skills or exhibit some level of exploration to solve the task. We use the term diversity to describe inter-agent diversification, meaning different agents acquiring different sets of skills. Note that then the state space in all SMAC maps is augmented with the unit type, which allows the agents to acquire different roles based on the unit type.

MMM2. Two identical teams battle each other, aiming to eliminate the units of the opposing team. Each team controls 3 types of units. The first type is a flying, healing unit called Medivac, capable of restoring health to nearby units. The second are the Marines units, which are capable of attacking both ground and air enemy units, and third, the Marauders, bulkier units only capable of attacking enemy ground units. The MMM2 scenario requires diversity among agents, where each type of agent needs to learn a different set of strategies.

MMM3 (7m2M1M vs 9m3M1M). This is a ultra hard version of the MMM2 map. The allies control a smaller team of 10 agents compared to the larger enemy team of 13 agents.

1c3s5z vs 1c3s6z. A scenario where both teams control 3 types of agents, Colossus, Stalkers, and Zealots. The enemy team has a higher number of Zealots.

3s vs 5z. In this scenario, 3 Stalkers (ranged units) are controlled by agents facing 5 Zealots (melee units). The challenge lies in overcoming the numerical disadvantage by leveraging Stalkers' kiting abilities, coordinating attacks, and maintaining effective positioning to avoid Zealots' close-range damage.

D Implementation & Reproducibility

All source code will be made available on publication under an open-source license. We refer the reader to the included README file, which contains instructions to recreate the experiments discussed in this paper.

E Epistemic and aleatoric uncertainty in multi-agent RL

In MARL environments, aleatoric uncertainty, or inherent randomness [33], arises from various sources, including the stochastic dynamics of the environment, inherent noise, agents' stochastic policies (e.g., epsilon-greedy), and policies shift that introduce non-stationarity. Epistemic uncertainty (lack of knowledge [34]), on the other hand, stems from limited familiarity with the environment's model (e.g., rewards and transitions) or an incomplete understanding of agents' policies in relation to one another. The presence of multiple agents within the same environment introduces both aleatoric and epistemic uncertainty, as agents' exploration and stochastic policies induce aleatoric uncertainty [33] and unfamiliarity of agents with other agents' models and the environment is epistemic.

In our uncertainty-weighted critic, we apply the weighting by utilizing ensemble kurtosis, i.e., disagreement between ensemble members, which is usually associated with epistemic uncertainty. Yet, the noise in our settings stems both sources: randomness in agent policies (i.e., aleatory) and epistemic uncertainty originated from the unfamiliarity of agents with the model of one another. Similar concerns have been addressed by recent works [20], which showed ensemble variability to be effective in capturing aleatoric uncertainty. [20] injected noise into the environment and

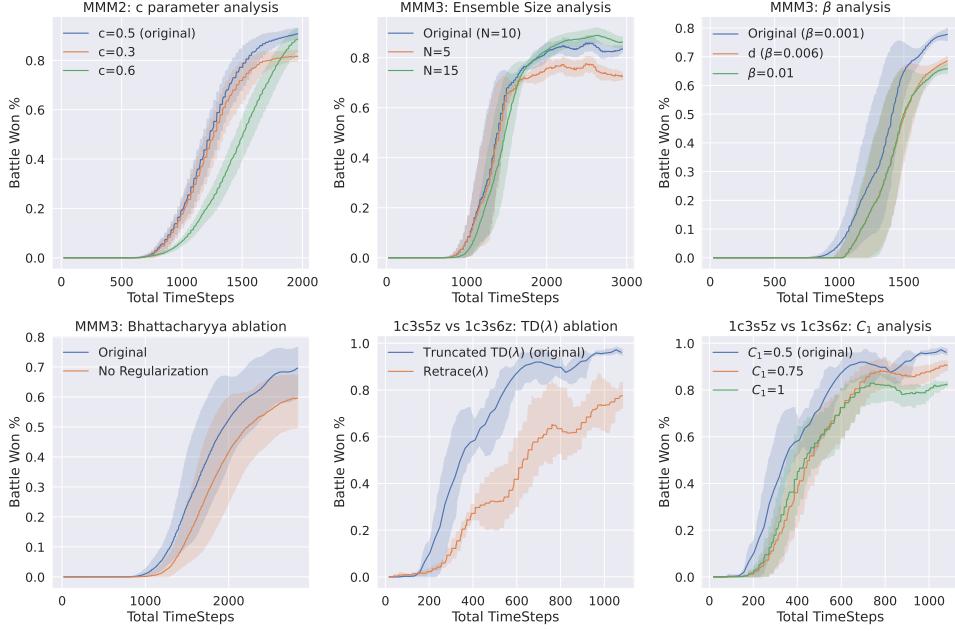


Figure 6: Parameter analysis on MMM2 and MMM3 maps for c , β , C_1 , and ensemble sizes. Ablations on the Bhattacharyya distance and our version of $\text{TD}(\lambda)$ versus $\text{retrace}(\lambda)$ are presented in the bottom graphs.

detected noisy states via ensemble variability. Motivated by those results, we apply the ensemble approach in our settings, which mixes both epistemic and aleatoric sources.

F Ablation study and parameter analysis

Additional details and results of the ablation study, which were omitted from the main paper, are provided in this section.

Variance versus kurtosis for exploration. In a separate set of experiments, we attempt to use the variance rather than the kurtosis to tune exploration. The process is similar to the algorithm suggested in Section 5.2, but instead of applying it only to high-uncertainty states that correspond to positive excess kurtosis, the variance is added to the actor logits in all time steps, regardless of the level of uncertainty. Note that, unlike kurtosis, where it is custom to use the normal distribution as a reference point, there isn’t a similar threshold for the variance. Results are presented in Figure 4 the MMM2 map. It can be seen that kurtosis-based exploration achieves superior results compared to the variance in both scenarios. It is interesting to note that previous studies from the single agent domain which suggested utilizing variance for exploration had more success [20, 27]. We hypothesize the main reason for the difference in performance lies in the transition from the single-agent to the multi-agent domain. Multi-agent systems require the learning algorithm to account for efficiency issues to achieve good results, i.e., when all agents are performing exploration all of the time, it becomes harder to form collaboration due to the large joint action space.

Parameter analysis and additional ablation. Fig. 6 shows analysis with different sizes of β , C_1 , c and varying ensemble size. Ensemble-MIX shows relatively low sensitivity to ensemble size, yet different sizes of β yield greater effect on performances, which is consistent with previous studies that show excessive exploration in multi-agent RL hurts performances. The bottom graphs present ablations on the diversity regularization mechanism and a comparison to the $\text{retrace}(\lambda)$ algorithm.

G Theoretical Analysis

The theoretical analysis objective is to bound the bias in the gradient update for each individual agent. Since each agent independently updates its policy, analyzing the bias on a per-agent basis is more intuitive. Our analysis shows that an agent’s updates are influenced not only by errors in the approximation of its Q-function but also by other agents’ policies.

To derive a bound on the bias we first outline key definitions used in the main paper, beginning with the mixed gradient update:

$$D_\nu^\beta(\pi_i, \boldsymbol{\pi}_{-i}) = (1 - \nu)\mathbb{E}_{\rho^\boldsymbol{\pi}, \pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) U_{\pi_i}^{\phi_i}(\tau_i, a_i)] + \nu \mathbb{E}_{\rho^\beta} [\nabla_{\theta_i} \bar{Q}_{\text{tot}}^\phi(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i}))], \quad (22)$$

where $\bar{Q}_{\text{tot}}^\phi(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i})) = \mathbb{E}_{\pi_i}[Q_{\text{tot}}^\phi(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i}))]$, and ν is the relative scaling factor. The on-policy term in Eq. 22 can be simplified as follows:

$$\mathbb{E}_{\pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | \tau_i) U_{\pi_i}^{\phi_i}(\tau_i, a_i)] \quad (23)$$

$$= \mathbb{E}_{\pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i) \lambda_i(\boldsymbol{\tau}) \left(Q_i^{\phi_i}(\tau_i, a_i) - \sum_x \pi_i(x | \tau_i) Q_i^{\phi_i}(\tau_i, x) \right) \right] \quad (24)$$

$$= \mathbb{E}_{\pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | \tau_i) \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i)]. \quad (25)$$

Note that the notation $D_\nu^\beta(\pi_i, \boldsymbol{\pi}_{-i})$ has been chosen to distinguish the mixed update from the gradient update of the true objective, which we denote by $\nabla_{\theta_i} J$. To express $\nabla_{\theta_i} J$ we first define the true Q-function of the i^{th} agent [5]:

$$Q_i^\pi(\boldsymbol{\tau}, a_i) = \sum_{\mathbf{a}_{-i}} \boldsymbol{\pi}_{-i}(\mathbf{a}_{-i} | \boldsymbol{\tau}_{-i}) Q_{\text{tot}}^\pi(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i})), \quad (26)$$

then:

$$\nabla_{\theta_i} J(\pi_i, \boldsymbol{\pi}_{-i}) = \mathbb{E}_{\boldsymbol{\tau} \sim \rho^\pi, a_i \sim \pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) Q_i^\pi(\boldsymbol{\tau}, a_i)], \quad (27)$$

where $\boldsymbol{\pi}_{-i}(\mathbf{a}_{-i} | \boldsymbol{\tau}_{-i}) = \prod_{j \neq i} \pi_j(a_j | \tau_j)$. The discrepancy between the true Q-function and the weighted approximated Q-function of the i^{th} agent is given by:

$$Q_{\phi_i}^\pi(\boldsymbol{\tau}, a_i) = Q_i^\pi(\boldsymbol{\tau}, a_i) - \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i). \quad (28)$$

We are now ready to prove proposition 1 which gives us a bound on the bias of each agent's updates. To achieve this we measure the difference between the gradient of the true objective $\nabla_{\theta_i} J(\pi_i, \boldsymbol{\pi}_{-i})$ and the mixed gradient update $D_\nu^\beta J(\pi_i, \boldsymbol{\pi}_{-i})$ we use in the paper.

Proposition 1. Let the maximal KL divergence between two policies $\boldsymbol{\pi}, \boldsymbol{\beta}$ be denoted by $D_{\max}^{\text{KL}}(\boldsymbol{\pi}, \boldsymbol{\beta}) = \max_{\boldsymbol{\tau}} D_{\text{KL}}(\boldsymbol{\pi}(\cdot | \boldsymbol{\tau}), \boldsymbol{\beta}(\cdot | \boldsymbol{\tau}))$, and let

$\omega_1 = \max_{\boldsymbol{\tau}, \mathbf{a}} \|\nabla_{\theta_i} \log (\pi_i(a_i | \tau_i; \theta_i)) \lambda_i(\boldsymbol{\tau}) Q_{\phi_i}^\pi(\boldsymbol{\tau}, a_i)\|_1$, $\omega_2 = \max_{\boldsymbol{\tau}} \|\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i}[Q_i^{\phi_i}(\tau_i, a_i)]\|_1$. Then, the learning bias of each agent can be bounded as follows:

$$\|\nabla_{\theta_i} J(\pi_i, \boldsymbol{\pi}_{-i}) - D_\nu^\beta(\pi_i, \boldsymbol{\pi}_{-i})\|_1 \leq \frac{\omega_1}{1 - \gamma} + \frac{2\omega_2\nu\gamma}{(1 - \gamma)^2} \sqrt{D_{\max}^{\text{KL}}(\boldsymbol{\pi}, \boldsymbol{\beta})}. \quad (29)$$

The bound tells us that the bias of each agent's gradient update depends on two main factors: errors in the approximation of Q_i^π , as measured by $Q_{\phi_i}^\pi$, and second, how much the joint on-policy diverge from the (joint) policy used to collect the off-policy data.

Proof. Let the discounted joint state distribution be denoted by:

$$\rho^\pi(\boldsymbol{\tau}) = \sum_{t=0}^{\infty} \gamma^t \rho_t^\pi(\boldsymbol{\tau}), \quad (30)$$

where $\rho_t^\pi(\boldsymbol{\tau}) = p^\pi(\boldsymbol{\tau}_t = \boldsymbol{\tau})$ denotes the state distribution at time t , following a joint policy $\boldsymbol{\pi}$. Then, we can use the following lemma from [35]:

Lemma 1. Let ρ_t^π and ρ_t^β be two state distributions at time t . Then the following bound holds:

$$\|\rho_t^\pi - \rho_t^\beta\|_1 \leq 2t \sqrt{D_{\max}^{\text{KL}}(\boldsymbol{\pi}, \boldsymbol{\beta})}. \quad (31)$$

Given Lemma 1 we are ready to prove the bound in Proposition 1:

$$\begin{aligned}
& \left\| \nabla_{\theta_i} J(\pi_i, \boldsymbol{\pi}_{-i}) - D_{\nu}^{\beta}(\pi_i, \boldsymbol{\pi}_{-i}) \right\|_1 \\
&= \left\| \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}, a_i)] - (1 - \nu) \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right] \right. \\
&\quad \left. - \nu \mathbb{E}_{\rho^{\boldsymbol{\beta}}} \left[\nabla_{\theta_i} \bar{Q}_{\text{tot}}^{\phi}(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i})) \right] \right\|_1 \\
&\stackrel{(*)}{=} \left\| \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \left(Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}, a_i) - \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right) \right] + \nu \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right] \right. \\
&\quad \left. - \nu \mathbb{E}_{\rho^{\boldsymbol{\beta}}} \left[\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right] \right\|_1 \\
&\stackrel{(**)}{\leq} \left\| \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \left(Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}, a_i) - \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right) \right] \right\|_1 \\
&\quad + \nu \left\| \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right] - \mathbb{E}_{\rho^{\boldsymbol{\beta}}} \left[\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right] \right\|_1 \\
&\stackrel{(***)}{\leq} \omega_1 \sum_{t=0}^{\infty} \gamma^t + \nu \omega_2 \sum_{t=0}^{\infty} \gamma^t \left\| \rho_t^{\boldsymbol{\pi}} - \rho_t^{\boldsymbol{\beta}} \right\|_1 \\
&\leq \frac{\omega_1}{1 - \gamma} + 2\nu \omega_2 \left(\sum_{t=0}^{\infty} \gamma^t t \right) \sqrt{D_{\max}^{\text{KL}}(\boldsymbol{\pi}, \boldsymbol{\beta})} \\
&= \frac{\omega_1}{1 - \gamma} + \frac{2\omega_2 \nu \gamma}{(1 - \gamma)^2} \sqrt{D_{\max}^{\text{KL}}(\boldsymbol{\pi}, \boldsymbol{\beta})}. \tag{32}
\end{aligned}$$

Transition $(*)$ leverages the additive structure of Q_{tot} by replacing $Q_{\text{tot}}^{\phi}(\boldsymbol{\tau}, (a_i, \mathbf{a}_{-i}))$ with $Q_i^{\phi_i}(\boldsymbol{\tau}, a_i)$, this is allowed since applying the gradient with respect to θ_i cancels out all the components of Q_{tot} that are independent of Q_i . Transition $(**)$ rearranges the terms and utilizes the triangle inequality to separate the original term into 2 different expressions. In transition $(***)$ and onward we bound the bias by taking the maximum value of each expression and extracting ω_1 and ω_2 from the respective sums. The first term is bounded with ω_1 as follows:

$$\begin{aligned}
& \left\| \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \left(Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}, a_i) - \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right) \right] \right\|_1 \\
&= \left\| \sum_{t=0}^{\infty} \gamma^t \sum_{\boldsymbol{\tau}} \rho_t^{\boldsymbol{\pi}}(\boldsymbol{\tau}) \sum_{a_i \in \mathcal{A}} \pi_i(a_i | \tau_i; \theta_i) \nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \left(Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}, a_i) - \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right) \right\|_1 \\
&\leq \max_{\boldsymbol{\tau}, \mathbf{a}} \left\| \nabla_{\theta_i} \log (\pi_i(a_i | \tau_i; \theta_i)) \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\boldsymbol{\tau}, a_i) \right\|_1 \sum_{t=0}^{\infty} \gamma^t \sum_{\boldsymbol{\tau}} \rho_t^{\boldsymbol{\pi}}(\boldsymbol{\tau}), \tag{33}
\end{aligned}$$

with $\sum_{\boldsymbol{\tau}} \rho_t^{\boldsymbol{\pi}}(\boldsymbol{\tau}) = 1$ and ω_1 equal the maximum over $\boldsymbol{\tau}$ and \mathbf{a} . The second term, which we bound using ω_2 , represents the difference between two distinct mean expressions. To derive a bound, we need to ensure that the same expression appears inside both means. Using the following transition:

$$\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) = \frac{\nabla_{\theta_i} \pi_i(a_i | \tau_i; \theta_i)}{\pi_i(a_i | \tau_i; \theta_i)}, \tag{34}$$

we can write:

$$\begin{aligned}
& \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\nabla_{\theta_i} \log \pi_i(a_i | \tau_i; \theta_i) \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right] \\
&= \mathbb{E}_{\rho^{\boldsymbol{\pi}}, \pi_i} \left[\frac{\nabla_{\theta_i} \pi_i(a_i | \tau_i; \theta_i)}{\pi_i(a_i | \tau_i; \theta_i)} \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right] \\
&= \mathbb{E}_{\rho^{\boldsymbol{\pi}}} \left[\sum_{a_i \in \mathcal{A}} \pi_i(a_i | \tau_i; \theta_i) \frac{\nabla_{\theta_i} \pi_i(a_i | \tau_i; \theta_i)}{\pi_i(a_i | \tau_i; \theta_i)} \lambda_i(\boldsymbol{\tau}) Q_i^{\phi_i}(\tau_i, a_i) \right] \\
&= \mathbb{E}_{\rho^{\boldsymbol{\pi}}} \left[\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right]. \tag{35}
\end{aligned}$$

This gives us the same expression in both means which differ only by state distributions:

$$\begin{aligned}
& \left\| \mathbb{E}_{\rho^\pi} \left[\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right] - \mathbb{E}_{\rho^\beta} \left[\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right] \right\|_1 \\
& \leq \sum_{t=0}^{\infty} \gamma^t \left\| \mathbb{E}_{\rho_t^\pi} \left[\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right] - \mathbb{E}_{\rho_t^\beta} \left[\nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right] \right\|_1 \\
& \leq \max_{\boldsymbol{\tau}} \left\| \nabla_{\theta_i} \lambda_i(\boldsymbol{\tau}) \mathbb{E}_{\pi_i} [Q_i^{\phi_i}(\tau_i, a_i)] \right\|_1 \sum_{t=0}^{\infty} \gamma^t \left\| \rho_t^\pi - \rho_t^\beta \right\|_1,
\end{aligned} \tag{36}$$

where the technique for bounding the difference between two means is similar to the proof presented in [9]. \square