SERA: Soft Ensemble Reliability Aggregation for Robust Multi-Agent Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Bootstrapped temporal-difference learning inherently introduces variance into value estimates, which often destabilizes learning due to value function oscillation between over- and under-estimation. Overestimation is commonly mitigated through pessimistic critic updates, but such bias-based approaches can introduce underestimation and do not address the estimation variance, which is often amplified in multi-agent reinforcement learning (MARL) due to its inherent learning complexities. To address this, we propose SERA, a soft ensemble reliability aggregation framework designed to reduce value estimation variance through reliability-aware critic aggregation. SERA constructs targets through soft reliability-weighted aggregation of critic estimates and introduces a novel decorrelation mechanism that adaptively tunes each critic's learning rate based on temporal-difference error uncertainty and the variance of target estimation error. This leads to more stable and reliable target estimation during training. Experiments on a wide range of multi-agent continuous-control benchmarks from MuJoCo and PettingZoo show that SERA consistently outperforms strong twin-critic and ensemble baselines, achieving performance improvements of up to 41.1%. We further demonstrate that the same framework generalizes well to single-agent continuous-control tasks, providing gains of up to 31.25% over established methods.

1 Introduction

Deep reinforcement learning (DRL) builds upon classical value-based methods and the actor–critic paradigm, both grounded in the principle of Temporal-Difference (TD) learning (Sutton et al., 1998). Even in the tabular setting, where convergence is guaranteed, TD-based Q-learning is known to produce overestimated value estimates during training despite asymptotic convergence (Tsitsiklis, 1994). This overestimation arises from uncertainty in value estimates and the effect of Jensen's inequality in the maximization step (Thrun & Schwartz, 2014), and is further exacerbated under function approximation (Kim et al., 2019; Duan et al., 2020).

This issue continues to persist under function approximation with deep neural networks such as Deep Q-Networks (DQN) (Mnih et al., 2015), deep deterministic policy gradient (DDPG) (Lillicrap et al., 2015), etc. In discrete-action settings,several methods have been proposed to mitigate overestimation such as Double DQN (Van Hasselt et al., 2016), Bootstrapped DQN (Osband et al., 2016), and ensemble-based approaches such as MeanQ (Liang et al., 2022), and Maxmin Q-learning (Lan et al., 2020). In continuous control, actor–critic methods such as DDPG (Lillicrap et al., 2015) suffer from similar overestimation issues under function approximation. To address this, algorithms such as Twin Delayed Deep Deterministic Policy Gradient (TD3) (Fujimoto et al., 2018) and Soft Actor–Critic (SAC) (Haarnoja et al., 2018) employ twin critics. Although these conservative critic updates can suppress overestimation, they may also push value estimates downward and still leave fluctuations in the target estimates insufficiently controlled.

In addition to bias-related issues, variance in bootstrapped value targets remain a major cause of instability in temporal-difference learning. Although TD learning is generally preferred over Monte Carlo methods for its lower variance, reliance on bootstrapped targets introduces an additional source of noise, as updates are computed from randomly sampled transitions rather than exact expectations (Xu et al., 2020). Moreover,

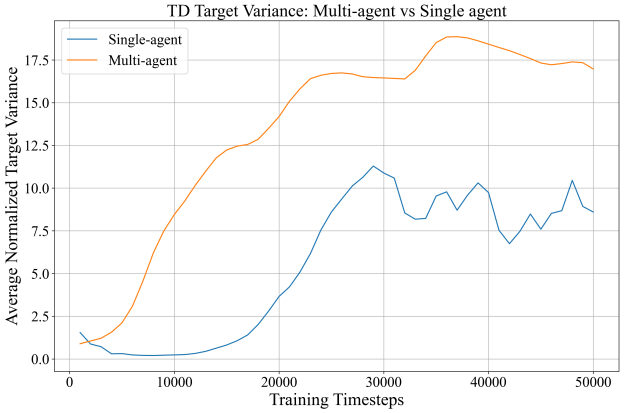


Figure 1: TD target variance during training for single-agent and multi-agent RL settings. The multi-agent setting exhibits consistently higher normalized target variance over a batch during training, indicating increased instability in value estimation.

in the presence of stochastic rewards and non-zero learning rates, this injected variance can propagate through successive updates, resulting in unstable value estimates (Pan & Schölkopf, 2025). When using function approximators such as neural networks, the bootstrapped value estimates are further affected by the bias–variance trade-off of the approximator, causing oscillations between under- and overestimation (Bengio et al., 2020). Thus, deep reinforcement learning is affected by variance-related issues from two sources: one arising from TD learning itself and the other from the inherent behavior of neural networks. As a result, reducing target estimation variance remains one of the challenges to address for efficient learning.

Multi-agent reinforcement learning (MARL) introduces additional sources of instability beyond those encountered in single-agent settings, including non-stationarity (Li et al., 2021), credit assignment difficulties (Sunehag et al., 2017), and partial observability (Omidshafiei et al., 2017). These factors increase the variability of value estimates and make critic learning significantly more difficult. As illustrated in Fig. 1, TD-target variance in multi-agent environments remains consistently higher than in comparable single-agent settings throughout training. Following (Lyu et al., 2021), this additional variability can be attributed to two major components: Multi-Action Variance (MAV), caused by the stochastic behavior of other agents, and Multi-Observation Variance (MOV), arising from differences in their local observations and histories. Both effects contribute to noisier TD targets and less stable policy updates, and their impact becomes increasingly severe as the number of interacting agents grows. Consequently, centralized critics in MARL are considerably more prone to unstable value estimation. Despite the growing use of ensemble critics in reinforcement learning, most existing approaches largely inherit design choices from single-agent methods and do not explicitly address estimation variance as a primary source of training instability. This raises an important question: can directly controlling estimation variance improve the stability and reliability of multi-agent learning?

To address this limitation we propose a soft ensemble reliability aggregation (SERA) framework for multi-agent reinforcement learning. Unlike prior ensemble methods that (i) rely on uniform averaging (e.g., MeanQ (Liang et al., 2022) in single-agent discrete settings), (ii) primarily encourage exploration (e.g., SUNRISE (Lee et al., 2021)), or (iii) reduce bias via pessimistic minimization of Q-values (e.g., MATD3 and MASAC in MARL), SERA treats estimation variance as a first-class learning signal that jointly governs target construction and critic updates. In particular, critic outputs are adaptively weighted based on their estimated reliability through a median-guided soft reliability aggregation, producing low-variance and robust training targets that attenuate noisy or inconsistent predictions while emphasizing more stable critic estimates. To further reduce correlation among ensemble members, we introduce a novel variance-aware decorrelation framework that adaptively adjusts each model critic's learning rate using temporal-difference error uncertainty and the variance of target estimation error. Together, these components improve target estimation stability, reduce critic correlation, and enable more sample-efficient learning. We evaluate SERA in

both multi-agent and single agent reinforcement learning settings to examine its effectiveness and generality across different learning scenarios. The key contributions are summarized as follows:

- *Reliability-Aware Target Aggregation:* We develop an ensemble critic framework in which the target values are constructed by assigning sample-wise soft reliability weights based on each critic's deviation from the median critic estimate, producing a unified and more stable target estimate.
- *Decorrelation Mechanism:* We propose a novel adaptive *learning rate* based critic decorrelation mechanism that adapts based on both temporal-difference error uncertainty and target estimation error variance.
- *Benchmarking in Multi-Agent Systems:* We adapt SERA to the centralized training–decentralized execution (CTDE) setting and evaluate it on nine multi-agent continuous control benchmarks from MuJoCo and PettingZoo (Terry et al., 2021). Under a fixed interaction budget of 300K steps, SERA achieves performance improvements of up to 41.1% and an average improvement of 28.7% compared to strong baselines.
- *Generalization to Single-Agent Continuous-Control Tasks:* We further extend the framework to single-agent continuous control, where SERA consistently surpasses recent ensemble-based methods and achieves gains of up to 31.25%.

2 RELATED WORKS

In this section, we briefly discuss the works that are closely related to our study.

2.1 Centralized Training with Decentralized Execution (CTDE)

Centralized Training with Decentralized Execution (CTDE) is the standard setup for MARL, enabling agents to leverage global information during training while maintaining decentralized policies at test time. Within CTDE, off-policy actor–critic baselines include MADDPG (Lowe et al., 2017), MAAC (Iqbal & Sha, 2019), and MATD3 (Ackermann et al., 2019), while entropy-regularized variants such as MASAC improve robustness. On-policy methods such as MAPPO (Yu et al., 2022) offer improved stability but at the cost of sample efficiency. Despite these advances, most CTDE methods continue to rely on standard target construction and critic update rules, leaving the variance of centralized value estimates largely uncontrolled in challenging multi-agent environments.

2.2 Ensemble-based Approaches

A number of ensemble-based methods have been proposed for multi-agent continuous control under the CTDE setting. EMAX (Lukas Schäfer & Mguni), and Implicit Ensemble Training (IET) (Shen & How, 2023) leverage ensembles primarily to encourage exploration, representation diversity, or robustness, rather than to explicitly control instability arising from noisy value estimates during critic learning. As a result, estimation variance remains largely untreated as a first-class factor in stabilizing critic updates.

In single-agent discrete-action learning, MeanQ (Liang et al., 2022) reduces overestimation by lowering target variance via ensemble averaging. However, it is formulated for discrete action-space and does not address variance propagation or instability in bootstrapped actor–critic learning, leaving estimation variance largely unregulated in continuous-control MARL. Methods such as ACE (Zhang & Yao, 2019) improve continuous-control learning by using an ensemble of actors to search for actions with higher critic values. This helps the policy update explore a richer set of candidate actions. However, ACE mainly introduces diversity on the actor side and does not explicitly consider how reliable each critic estimate is when forming target values. In contrast, SERA operates on the critic ensemble directly, assigning sample-wise reliability weights to target critics according to their deviation from a robust median anchor. Similarly, SUNRISE (Lee et al., 2021) leverages ensemble diversity to enhance exploration and robustness, rather than directly addressing variability in value estimation during critic updates.

Existing ensemble-based reinforcement learning methods typically improve learning through three main strategies: averaging critic estimates, using minimum operators for conservative target selection, or leveraging ensemble diversity to encourage exploration. In most existing methods, these ideas are handled separately. Averaging-based approaches primarily aim to reduce fluctuations in value estimates, while minimum-based targets are mainly used to control overestimation. Methods that exploit ensemble disagreement typically use it for exploration purposes instead of directly stabilizing the critic learning process.

SERA instead treats estimation variance as an important quantity to manage during training. The framework combines reliability-based target aggregation with adaptive critic updates so that unreliable critic estimates contribute less to learning, while diversity within the ensemble is still maintained. As a result, variance is controlled both when forming the bootstrapped TD targets and when updating the critic networks. This is different from most existing ensemble actor–critic methods, which usually emphasize either target aggregation or exploration independently.

3 Preliminaries

3.1 Markov Games

We model the multi-agent environment as a *Markov game* (Littman, 1994), which extends the standard Markov Decision Process (MDP) framework to settings involving multiple interacting agents. The game is represented by the tuple $(\mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{I}}, \mathcal{P}, \{R_i\}_{i\in\mathcal{I}}, \{\Omega_i\}_{i\in\mathcal{I}}, \gamma)$, where $\mathcal{I}$ denotes the collection of agents and $\mathcal{S}$ corresponds to the global state space. At each time step, the agents jointly execute an action vector $\mathbf{a} = (a_1, \ldots, a_N) \in \mathcal{A}$, resulting in a state transition governed by the probability mapping $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathrm{Dist}(\mathcal{S})$. Each agent i observes a local observation $o_i \in \Omega_i$ and receives a reward signal determined by the reward function $R_i(s, \mathbf{a})$. Decision making is performed through stochastic policies of the form $\pi_i(a_i \mid o_i)$. The objective of every agent is to maximize the expected discounted cumulative reward given by $J_i = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_i^{(t)}\right]$, where $\gamma \in (0, 1]$ represents the discount factor. In this work, we specifically consider cooperative multi-agent settings in which all agents aim to achieve a common or mutually aligned objective.

3.2 Centralized Training and Decentralized Execution (CTDE)

CTDE paradigm decouples training from execution. During training, critics are allowed to access global information, while during execution each agent acts using only its local observations. Let the global state space be denoted as $\mathcal{S}$, the action space of agent i is denoted as A_i, and O_i its local observation space. Each agent follows a stochastic policy $\pi_i : O_i \to \mathrm{Dist}(A_i), \quad a_i \sim \pi_i(\cdot \mid o_i)$, which depends only on its local observation $o_i \in O_i$. The critic is trained with centralized information. For agent i, the action-value function is defined as

$$Q_i^{\pi}(s, a_1, \ldots, a_N) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i^t \,\middle|\, s_0 = s,\ a_j \sim \pi_j(o_j)\ \forall j\right],$$

where $s \in S$ is the global state, $(a_1, \ldots, a_N)$ denotes the joint action, and $\gamma \in (0, 1)$ is the discount factor.

Each agent seeks to maximize its expected return, given by $J_i(\pi_i) = \mathbb{E}_{s\sim\rho^{\pi},\, a\sim\pi}\left[Q_i^{\pi}(s, a_1, \ldots, a_N)\right]$. While the critics benefit from the global state to mitigate non-stationarity, the policies π_i only require local observations o_i, making the approach suitable for decentralized deployment under partial observability. For more details, the readers are referred to (Lowe et al., 2017).

3.3 Temporal Difference Learning and Value Bootstrapping

Critics are commonly trained using temporal-difference (TD) learning. For a critic Q_i with parameters θ_i, the update is based on a bootstrapped target of the form

$$y_i = r_i + \gamma Q_i(s', \mathbf{a}'; \theta_i^-), \tag{1}$$

where $\bar{\theta}_i$ denotes a target network that is updated more slowly than the online parameters. This delayed update helps maintain stability by limiting rapid changes in the regression target.

Temporal-difference learning estimates future returns using its own current value predictions. Because of this recursive update structure, estimation errors can accumulate over time and influence later updates. Under function approximation, the bootstrapped targets may therefore contain both bias and variance from inaccurate value estimates, which can make training unstable (Pan & Schölkopf, 2025; Bengio et al., 2020).

These effects are further amplified in multi-agent environments. The joint action $\mathbf{a}' = (a'_1, \ldots, a'_N)$ is generated by multiple agents whose policies evolve concurrently, making the target distribution non-stationary. As a result, TD targets can become increasingly noisy during training. To address this issue, we introduce SERA, which stabilizes target estimation through reliability-weighted aggregation of critic outputs.

3.4 Ensemble Critics

In reinforcement learning, an ensemble refers to the use of multiple value functions trained for the same task. The collective behavior of an ensemble can be used to quantify uncertainty or to produce a lower-variance estimate, compared to a single estimator. This is used by prior works like (Lee et al., 2021; Liang et al., 2022; Fujimoto et al., 2018).

Let $\{Q_{\theta_k}\}_{k=1}^{K}$ denote a collection of K critics, each with its own parameters. For a given state-action pair $(s, \mathbf{a})$, the ensemble yields a set of predictions $\{Q_{\theta_k}(s, \mathbf{a})\}_{k=1}^{K}$. Maintaining multiple critics can help stabilize learning by mitigating noisy estimates and providing a measure of uncertainty through their disagreement. In this work, we make use of this ensemble to form more reliable targets during training.

4 SERA: Soft Ensemble Reliable Aggregation

In this section, we present the main components of the proposed Soft Ensemble Reliable Aggregation (SERA) framework. While SERA builds on the general idea of using multiple critics to stabilize temporal-difference learning, it takes a different direction by explicitly estimating how reliable each critic is and using this information to reduce noise in target estimation. This notion of reliability is used consistently in both how targets are formed and how critics are trained.

At the core of SERA is a soft aggregation scheme (reliability-aware target aggregation), where critics that behave more consistently have a stronger influence on the target, while those that deviate significantly are naturally down-weighted. Alongside this, we introduce a novel decorrelation mechanism designed to keep the critics sufficiently diverse, avoiding redundancy and improving overall stability. We begin by describing the reliability-aware target aggregation, and then present the proposed decorrelation strategy.

4.1 Reliability-aware Target Aggregation

To obtain a more stable target, we use a sample-wise aggregation scheme that combines the robustness of the median with an adaptive notion of reliability based on per-sample agreement among critics.

Let $\{Q_{\theta_i}(s, a)\}_{i=1}^{N}$ denote the ensemble of critics, and let $\{Q_{\bar{\theta}_i}(s, a)\}_{i=1}^{N}$ represent the corresponding target critics. For each sample b in a mini-batch, we first compute the median anchor using the target critics:

$$Q_{\text{med}}^{(b)}(s, a) = \text{median}\left(\{Q_{\bar{\theta}_i}^{(b)}(s, a)\}_{i=1}^{N}\right), \tag{2}$$

and denote its index by m_b. This median serves as a stable reference point and reduces the effect of extreme critic estimates.

We then measure how much each critic deviates from this reference:

$$d_i^{(b)} = \left|Q_{\bar{\theta}_i}^{(b)}(s, a) - Q_{\text{med}}^{(b)}(s, a)\right|. \qquad \forall i \in \{1, \ldots, N\} \tag{3}$$

Based on these deviations, we assign weights to the non-median critics using a soft exponential scheme:

$$w_i^{(b)} = \frac{\exp\left(-d_i^{(b)}/\tau\right)}{\sum_{k \neq m_b} \exp\left(-d_k^{(b)}/\tau\right)}, \qquad \tau > 0, \tag{4}$$

where τ controls how sharply the weights concentrate. Since the median critic has zero deviation by definition, it is excluded from the weighting procedure to prevent it from dominating the aggregation. This design assigns larger weights to critics whose predictions stay nearer to the median estimate, which is the central consensus. Lemma 4.1 formally establishes that critics with smaller deviations from the consensus always obtain higher weights.

Lemma 4.1 (Sample-wise reliability ordering of SERA)**.** *For any two non-median critics i, j such that $i, j \neq m_b$, $d_i^{(b)} < d_j^{(b)} \quad \Rightarrow \quad w_i^{(b)} > w_j^{(b)}$. Also, if for all $b \in \mathcal{B}$, $d_i^{(b)} \leq d_j^{(b)}$, with strict inequality for at least one sample, then*

$$\frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} w_i^{(b)} > \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} w_j^{(b)}. \qquad \forall i, j \neq m_b$$

Therefore, critics that remain closer to the consensus throughout the mini-batch are assigned larger average reliability weights under SERA. This weighting behavior shapes the final aggregation by reducing the influence of outlier critics.

The resulting aggregated estimate from the non-median critics is:

$$Q^{(b)}_{\text{sera}\setminus\text{med}}(s,a) = \sum_{i\neq m_b} w_i^{(b)}\, Q^{(b)}_{\bar{\theta}_i}(s,a). \tag{5}$$

We then combine this estimate with the median anchor to form the final SERA target:

$$Q^{(b)}_{\text{SERA}}(s,a) = \alpha\, Q^{(b)}_{\text{med}}(s,a) + (1-\alpha)\, Q^{(b)}_{\text{sera}\setminus\text{med}}(s,a), \tag{6}$$

where $\alpha \in [0,1]$ controls the trade-off between robustness and adaptivity.

Remark 1 (Median-anchored reliability estimation)**.** *Since the true target $y^\star$ is not directly accessible, the median is used as a robust anchor for comparing critic estimates. Its resistance to outliers makes it a stable reference for measuring deviations (Hampel, 1971; Rousseeuw & Leroy, 2003).*

In summary, the aggregation is performed at the individual sample level, assigning greater importance to critics that align more closely with the central tendency and suppressing the influence of outlier estimates. We now analyze some fundamental properties of the method, particularly the critic ranking induced by the weights and the characteristics of the final target.

Proposition 4.2 (Convexity and Boundedness)**.** *For each sample b, $Q^{(b)}_{\text{SERA}}$ is a convex combination of the target critic estimates. Consequently, the SERA estimate is bounded by the ensemble's extremes:*

$$\min_i Q^{(b)}_{\bar{\theta}_i} \leq Q^{(b)}_{\text{SERA}} \leq \max_i Q^{(b)}_{\bar{\theta}_i}.$$

Thus, unlike min-based aggregation such as MATD3 and MASAC, $Q^{(b)}_{\text{SERA}}$ does not reduce to the minimum critic estimate, thus avoiding excessive pessimistic underestimation. At the same time, it does not exceed the largest estimate, ensuring that the aggregation remains within the range of the ensemble predictions. This establishes bounded avoidance of extreme critic aggregation.

Proposition 4.3 (Variance reduction relative to a single critic)**.** *Suppose the target critic estimates satisfy $Q^{(b)}_{\bar{\theta}_i} = y^\star + \varepsilon_i$, where $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$, and the critic errors $\{\varepsilon_i\}_{i=1}^N$ are mutually independent. Then, the SERA target satisfies*

$$\text{Var}\Big(Q^{(b)}_{\text{SERA}}\Big) < \sigma^2 = \text{Var}\Big(Q^{(b)}_{\bar{\theta}_i}\Big), \forall i \in \{1,\ldots,N\}$$

for any individual critic j.

Thus, under the same independence and equal-variance assumptions commonly used to justify variance reduction in ensemble averaging, SERA also yields a lower-variance target than any single critic. The reduction factor is determined by $\sum_i (\tilde{w}_i^{(b)})^2$, which acts as the effective concentration of the reliability weights.

Comparison with mean aggregation: Proposition 4.3 establishes that SERA reduces variance relative to a single critic under standard assumptions. Compared with simple averaging, the arithmetic mean achieves the minimum variance when all critics have identical and independent noise, yielding variance σ^2/N. However, in deep reinforcement learning, critic noise is often uneven across the ensemble. In such cases, uniform averaging becomes sensitive to unreliable critics since all estimates contribute equally. SERA instead adjusts critic contributions according to their agreement with the median estimate, assigning smaller weights to critics with larger deviations. This reduces the influence of unstable critics while still preserving the benefits of ensemble aggregation, making SERA more robust in high-noise or heterogeneous settings.

Proposition 4.2 and 4.3 together show that the SERA target remains within the range of the ensemble predictions while reducing the variance of the aggregated estimate. As a result, critics that produce unstable or highly inconsistent estimates have a smaller influence on the final target, without forcing the update toward overly pessimistic values as in hard minimum selection (MATD3 or MASAC). This makes the TD targets less affected by noisy or outlier predictions, leading to more stable critic learning dynamics during training.

While Proposition 4.3 assumes independent critic errors for the variance analysis, critics in deep reinforcement learning are usually correlated due to shared replay data and similar training dynamics. Such correlation can weaken the advantages of ensemble aggregation by reducing critic diversity. SERA addresses this directly by incorporating mechanisms that encourage decorrelation during training. These include separate parameter initialization, heterogeneous network structures, and the variance-adaptive learning rate strategy described in Section 4.2. Collectively, these components help preserve diversity across critics and strengthen the robustness of the aggregation process.

4.2 Decorrelation Strategy

Effective variance reduction through ensemble aggregation relies on maintaining diversity among critics, as averaging identical estimates provides no benefit. Accordingly, SERA encourages decorrelation so that critics retain sufficiently diverse estimation errors, enabling meaningful variance reduction. To promote diversity among the ensemble critics, three complementary strategies are employed. First, all networks are initialized with different random parameters so that each critic begins training from a unique starting point. Second, the critics are designed with heterogeneous hidden-layer architectures, encouraging them to capture different feature representations during learning. In addition, a learning-rate-driven decorrelation strategy is introduced to further reduce similarity among the critic updates.

4.2.1 Variance-Adaptive Learning Rate

In addition to the reliability-aware target construction, we incorporate a novel mechanism that adjusts each critic's update strength according to the spread of the ensemble.

Let $Q^\star(s,a)$ denote the true action–value for a given state–action pair (s,a). $\{Q_{\theta_i}(s,a)\}_{i=1}^N$ denote the ensemble of critics, and $\{Q_{\bar{\theta}_i}(s,a)\}_{i=1}^N$ the corresponding target critics. For a transition (s,a,r,s'), the temporal-difference (TD) target is defined as: $y = r + \gamma Q_{\bar{\theta}}(s,a)$.

To characterize uncertainty in learning, we introduce two variance terms. The uncertainty associated with the critic's current estimate is defined as

$$P_t = \mathrm{Var}\big(Q_t - Q^\star(s,a)\big) \tag{7}$$

and is termed as estimation variance, while the uncertainty of the TD target is defined as

$$R_t = \mathrm{Var}\big(y - Q^\star(s,a)\big), \tag{8}$$

and is termed as measurement variance. The estimation variance reflects the uncertainty in the current critic estimate, while the measurement variance captures the noise present in the TD target. The following lemma, termed Variance-aware adaptive critic step size, can be established.

Lemma 4.4 (Variance-aware adaptive critic step size)**.** *Assume that the critic estimation error $e_t = Q_t - Q^\star(s,a)$ and the TD target error $\varepsilon_t = y_t - Q^\star(s,a)$ are unbiased and uncorrelated. Then, the step size β_t*

that minimizes the expected mean-square error $\mathbb{E}(Q_{t+1} - Q^\star(s,a))^2$ *is given by*

$$\beta_t^\star = \frac{P_t}{P_t + R_t}.$$

Remark 2 (Practical validity of assumptions)**.** *Lemma 4.4 assumes unbiased and uncorrelated estimation errors for analytical clarity. SERA does not require these conditions to hold exactly in practice; the resulting step-size is used as an uncertainty-gated heuristic, akin to adaptive optimization and filtering methods in non-linear systems. Empirically, this variance-adaptive update remains effective in MARL.*

The adaptive learning rate reflects a balance between the trustworthiness of the current critic's prediction estimate P and the reliability of incoming TD targets R. This balance ensures stable learning while still allowing rapid adaptation when reliable information is available.The sensitivity analysis of $\beta_t^\star$ is provided in the appendix. This form is inspired by the Kalman gain, which balances the uncertainty in the estimate and the target (Li et al., 2015).

4.2.2 Practical design of the adaptive learning rate

In practice, the true variances P_t and R_t are not directly observable, as they depend on the unknown quantity $Q^\star(s,a)$. We therefore approximate them using empirical batch statistics. For each critic i, we estimate the measurement variance as

$$R_{\text{gain},i} = \text{Var}_{j\in\mathcal{B}}\left(Q_{\bar{\theta}_i}(s'_j, \mathbf{a}'_j) - Q_{\text{SERA}}(s'_j, \mathbf{a}'_j)\right), \tag{9}$$

where $\mathcal{B}$ denotes the mini-batch. Here, $Q_{\text{SERA},j}$ serves as the aggregated target, and the deviation from it reflects the critic's estimation error.

Within the ensemble framework, every critic is updated using its own adaptive learning rate, allowing the update magnitude to vary according to both the critic-specific uncertainty and the disagreement from the ensemble consensus. For the i-th critic, the prediction variance at training step t, denoted by $v_{Q,i}^{(t)}$, is computed from the variance of the temporal-difference (TD) residuals over the sampled minibatch. This quantity is tracked using an exponential moving average to obtain a stable estimate during training:

$$v_{Q,i}^{(t)} \leftarrow \alpha_Q v_{Q,i}^{(t-1)} + (1-\alpha_Q)\,\text{Var}_{j\in\mathcal{B}}\left(\delta_j^{(i)}\right), \tag{10}$$

where $\delta_j^{(i)} = y_j - Q_{\theta_i}(s_j, \mathbf{a}_j)$ is the TD error of critic i and α_Q controls the smoothness.

The estimation variance P is then updated by accumulating the prediction-variance:

$$P_t^{(i)} \leftarrow \alpha_P P_{t-1}^{(i)} + (1-\alpha_P) v_{Q,i}^{(t)}. \tag{11}$$

Here, $P_{t-1}^{(i)}$ denotes the posterior estimation variance of critic i after incorporating all information available up to time $t-1$. This term captures the remaining uncertainty in the critic's value estimate.

The adaptive scaling factor for critic i, from Lemma 5.4, is defined as

$$\kappa_t^{(i)} = \frac{P_t^{(i)}}{P_t^{(i)} + R_{\text{gain},i} + \eta}, \qquad \kappa_t^{(i)} \in (0,1). \tag{12}$$

where $\eta > 0$ ensures numerical stability. The critic parameters are then updated with a variance-adaptive gradient step:

$$\theta^{(i)} \;\leftarrow\; \theta^{(i)} + \frac{\alpha_Q}{S}\sum_{j=1}^{S} \kappa_t^{(i)}\,\delta_j^{(i)}\,\nabla_{\theta^{(i)}} Q_{\theta^{(i)}}(s_j, \mathbf{a}_j), \tag{13}$$

This design is inspired by the Kalman filter under simplifying assumptions; further details are provided in the appendix. To understand the relationship between R and R_{gain} is shown in appendix.

4.3 SERA Algorithm and its Complexity

SERA combines the novel learning rate based decorrelating mechanism along with the reliability-aware target formulation. The whole training procedure is shown in Algorithm 2. We chose MASAC updating rules for our base. The whole training mechanism is shown in Algorithm 2 (please refer the supplementary file).

Computational Complexity: The per-iteration computational cost of the proposed SERA framework can be expressed as $\mathcal{O}\big(NS\,C_Q + MS\,C_\pi\big)$, where the individual terms represent different stages of the training procedure. In this formulation, N corresponds to the number of critic networks in the ensemble, S denotes the minibatch size, and M indicates the number of decentralized policy networks. The terms C_Q and C_π represent the computational costs associated with the forward and backward passes of a single critic and actor network, respectively. The component $NS\,C_Q$ arises from computing and updating all ensemble critics, while $MS\,C_\pi$ captures the cost of updating the actor policies. In addition, the aggregation operation grows linearly with respect to both the ensemble size and minibatch size, introducing only a minor overhead compared to the overall neural network optimization process.

5 Experimental Evaluations

We evaluate the proposed SERA framework on a diverse collection of cooperative multi-agent continuous control tasks from the MuJoCo and PettingZoo benchmarks. To examine the scalability of the method under varying levels of coordination complexity, the selected environments include systems ranging from two to six agents. SERA is compared against four representative CTDE-based baselines, namely IET, MATD3, MASAC, and MAPPO, using identical network architectures and training protocols for fair comparison. Additional implementation details and hyperparameter configurations are included in the appendix. Performance is measured using the average episodic return throughout training, and results are reported with standard deviation across 10 random seeds to evaluate both stability and consistency.

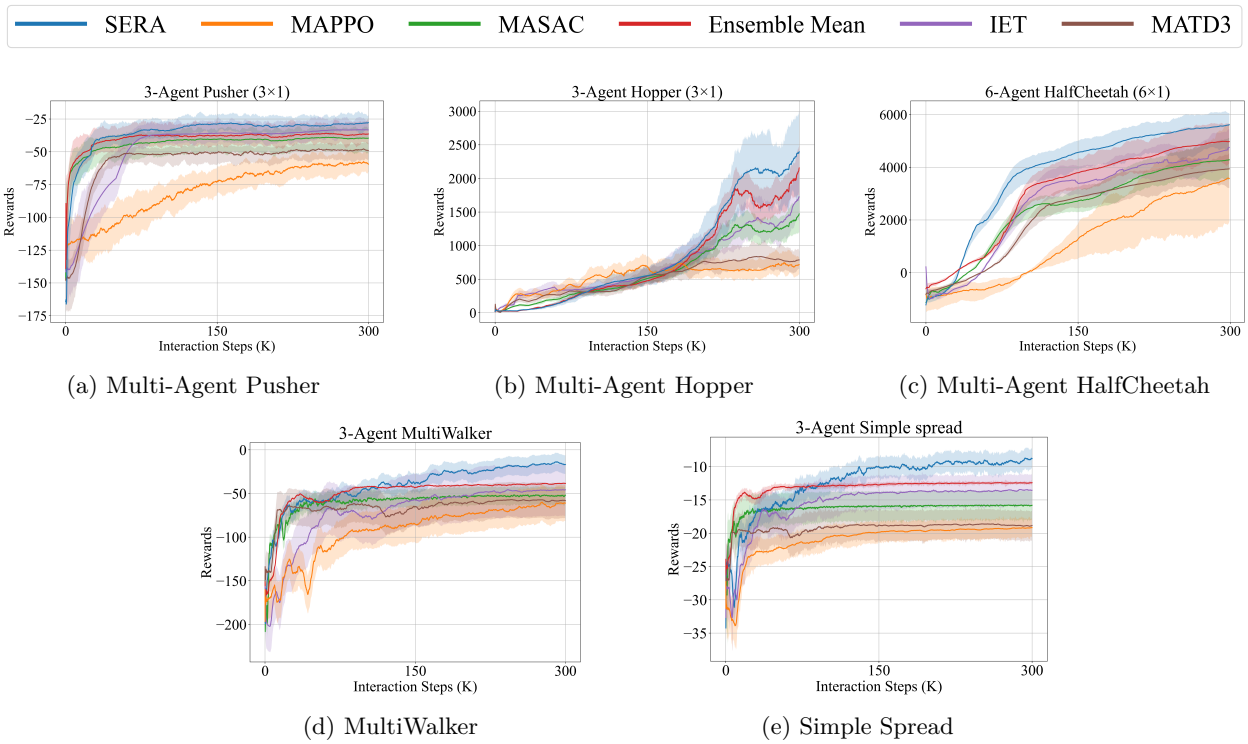
Alongside SERA, we also consider **Ensemble Mean** as a comparison baseline to examine the effect of variance-aware target aggregation. Experiments are conducted on five multi-agent continuous control benchmarks, including three MaMuJoCo environments and two cooperative PettingZoo tasks. These benchmarks cover a range of coordination challenges, interaction patterns, and control difficulties. We focus on continuous-action environments since value estimation variance can become more pronounced in continuous domains because of the larger action space and the coupled learning dynamics among agents. In addition, MASAC and similar actor–critic approaches are mainly developed for continuous control, making these benchmarks appropriate for evaluating the proposed framework.

5.1 Results

Fig. 2 presents the average reward curves and standard deviation profiles for both the MuJoCo and PettingZoo environments. The rewards are averaged over ten independent runs to reflect not only overall performance but also the consistency of training behavior across different random seeds. The selected benchmarks include a variety of coordination patterns and control dynamics, allowing evaluation under diverse multi-agent settings.

Table 1: Performance comparison of SERA with baseline algorithms across multi-agent environments.

Environment	SERA	Ensemble Mean	IET	MASAC	MATD3	MAPPO
HalfCheetah (6 agents)	**5520** $\pm$ 711	4831 $\pm$ 721	4608 $\pm$ 783	4500 $\pm$ 802	3815 $\pm$ 912	3623 $\pm$ 812
Pusher (3 agents)	**-27** $\pm$ 8	-30 ± 7	-29 ± 9	-31 ± 8	-50 ± 11	-57 ± 13
Hopper (3 agents)	**2380** $\pm$ 519	2190 $\pm$ 611	1701 $\pm$ 600	1530 $\pm$ 200	840 $\pm$ 400	790 $\pm$ 100
Multiwalker (3 agents)	**-27.4** $\pm$ 5	-39 ± 5	-46.5 ± 5	-51 ± 10	-56 ± 10	-60.3 ± 10
Simple Spread (3 agents)	**-9** $\pm$ 2	-13 ± 2.3	-14 ± 4	-16 ± 5.1	-17.5 ± 6	-18 ± 2



(a) Multi-Agent Pusher (b) Multi-Agent Hopper (c) Multi-Agent HalfCheetah

(d) MultiWalker (e) Simple Spread

Figure 2: Performance comparison across different multi-agent cooperative environments.

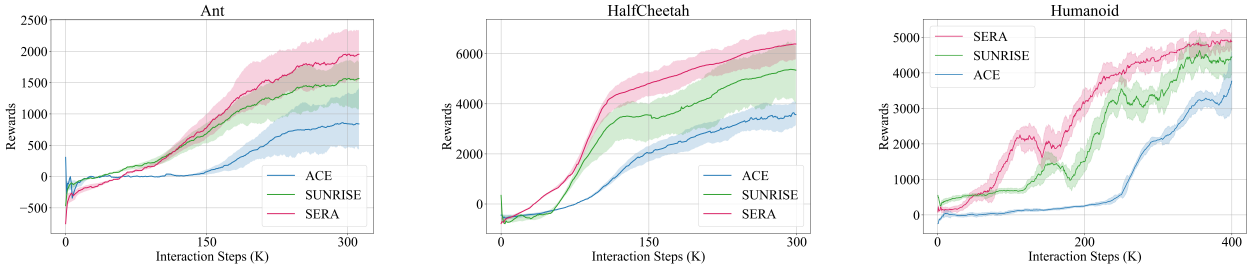
The results in Fig. 2 show that SERA consistently achieves stronger performance than all compared baselines across the tested environments. It also performs better than its ensemble counterpart, **Ensemble Mean**. Although simple ensemble averaging can sometimes improve over standard baselines, its gains are not consistently maintained against IET. SERA, on the other hand, shows stable improvements across tasks, highlighting the benefit of combining sample-aware aggregation with critic decorrelation for more reliable multi-agent learning.

As reported in Table 1, the proposed SERA framework produces the largest performance improvement in the Multiwalker environment, achieving a gain of 41.1%. Across all evaluated tasks, the method yields an average improvement of 28.7% over the competing approaches. SERA consistently reaches higher cumulative returns, suggesting better utilization of collected samples and improved learning efficiency. Among the baseline algorithms, MASAC generally provides the strongest competitive performance due to the stability offered by entropy regularization and its twin-critic structure, with MATD3 showing comparable behavior in several environments. MAPPO, however, records the lowest overall performance among the considered methods.

5.2 Generalization to single agent continuous control

To further examine the generalization capability of SERA, experiments are also conducted on single-agent continuous control tasks, namely Ant, HalfCheetah, and Humanoid. As the multi-agent study already includes widely used CTDE-based methods such as MATD3, MASAC, and MAPPO, their corresponding single-agent variants are not considered again in order to avoid repetitive comparisons. Instead, the evaluation focuses on recent ensemble-oriented reinforcement learning approaches, including SUNRISE (Lee et al., 2021) and ACE (Zhang & Yao, 2019). This allows the comparison to remain centered on ensemble-based learning strategies.

From Fig. 3, it is evident that the competing methods obtain relatively similar performance in the *HalfCheetah* and *Humanoid* tasks, where SERA achieves improvements of 17.12% and 15.23%, respectively. A larger



(a) Ant (b) HalfCheetah (c) Humanoid

Figure 3: Training reward trajectories comparing the proposed SERA framework with recent ensemble-based baselines across three single-agent MuJoCo environments.

Table 2: Performance comparison of SERA with ensemble-based baselines on continuous single-agent control tasks.

Environment	SERA	SUNRISE	ACE
Ant-v5	**1983.7** $\pm$ 602	1530 $\pm$ 605	890 $\pm$ 654
HalfCheetah-v5	**6056** $\pm$ 439	5613 $\pm$ 570	3923 $\pm$ 480
Humanoid-v5	**4982.6** $\pm$ 510	4515.3 $\pm$ 600	3908.6 $\pm$ 582

performance gain is observed in the *Ant* environment, where the proposed method improves the results by nearly 31.25%. These observations indicate that SERA remains effective even in single-agent continuous control problems.

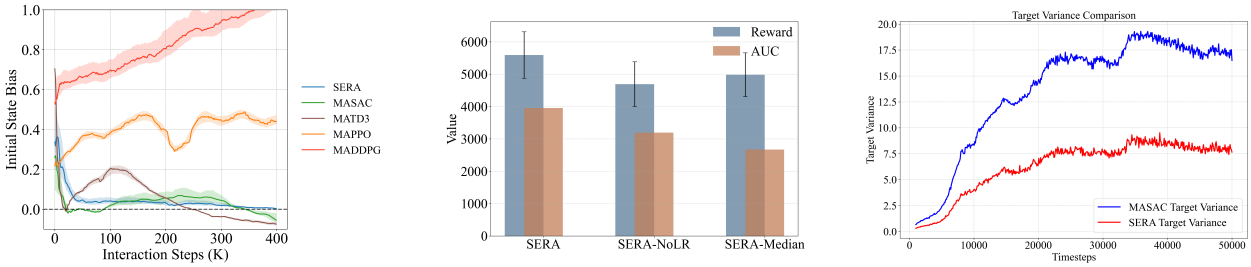
5.3 Discussion

To verify that the observed improvements are not due to random variation, we performed Welch's t-tests using the final evaluation returns across different random seeds. SERA showed statistically significant improvements over MASAC ($p = 0.021$) and IET ($p = 0.029$), indicating consistent gains across runs. The comparison with Ensemble Mean resulted in a larger p-value ($p = 0.072$), suggesting that part of the improvement arises from the variance reduction effect of the ensemble itself. It should also be noted that Ensemble Mean uses the same adaptive learning rate mechanism proposed in SERA.

We also analyze the computational overhead associated with ensemble critics by comparing wall-clock training time against MASAC. Using MASAC as the reference point (1.0×), IET requires about 1.48× more training time, SERA requires approximately 1.6× the training time of MASAC. Even with this additional cost, the overhead remains moderate relative to the gains in training stability and learning efficiency. Overall, the results indicate that the extra computation introduced by SERA is justified by its improved and more reliable training behavior.

Initial-State Bias: The initial-state bias can be defined as the difference between the Q-value prediction at the starting state and the empirical discounted return obtained from evaluation rollouts. For an ensemble with K critics, the aggregated value estimate is computed as $V(s_0) = \frac{1}{K}\sum_{k=1}^{K} Q_{\theta_k}(s_0, a_0)$, while the corresponding bias is given by $b = V(s_0) - \sum_t \gamma^t r_t$. Here smaller bias reflects more reliable temporal-difference estimation.

Fig. 4a presents the evolution of the initial-state bias during training. SERA consistently maintains the smallest bias, suggesting more stable and accurate value estimation. MADDPG, on the other hand, shows clear overestimation behavior, whereas twin-critic methods such as MATD3 and MASAC display a tendency toward underestimation (Ren et al., 2021; Lyu et al., 2022). These observations suggest that controlling estimation variance helps SERA avoid both overly optimistic and overly pessimistic value estimates during MARL training.



(a) Initial state bias. (b) Ablation study on Humanoid. (c) Target variance curve.

Figure 4: Experimental analysis of SERA showing initial state bias, ablation results and target variance comparison on Halfcheetah environment.

Ablation Study Fig. 4b reports a focused ablation study of SERA's two key design choices on the Halfcheetah environment. Disabling the variance-adaptive learning rate (SERA-NoLR) or including the median in the soft aggregation (SERA-Median) consistently degrades performance, indicating that both uncertainty-aware updates and robust anchoring play a critical role in SERA's stability and effectiveness. SERA-Median becomes highle conservative as it downweights the other critic values as discussed in section 4.1. In the results section, we compared SERA with Ensemble Mean, which also plays a part of ablation study. The ensemble mean underperforms SERA, providing empirical evidence that soft-aggregation is more effective than simple averaging, in line with the theoretical insights of Lemma 4.1 and proposition 4.2. Here, in Fig. 4b, the AUC plots are divided by the total number of training episodes to have a better plot.

Variance Reduction Figure 4c presents the evolution of TD target variance during training. Compared to MASAC, the proposed SERA approach maintains noticeably lower variance throughout the learning process. The difference becomes larger in the later stages of training, where critic estimates generally become more unstable due to accumulated approximation errors and growing disagreement across critics. The smoother variance profile observed with SERA indicates that the proposed aggregation strategy produces more stable target estimates and reduces the influence of unreliable critic predictions.

Limitations. Despite improving the stability of value estimation, SERA requires maintaining multiple critic networks, which increases both training time and computational cost compared to conventional twin-critic approaches. The proposed aggregation mechanism also relies on the ensemble retaining sufficiently different critic estimates during training. If the critics become overly similar, the overall benefit obtained from ensemble-based variance reduction can decrease. In addition, using larger ensembles may further improve robustness, but this comes at the expense of higher memory usage and additional optimization overhead.

6 Conclusion

In this work, we investigated the role of critic disagreement and estimation variance in RL, and introduced SERA, a reliability-aware ensemble critic framework for stable target estimation. Instead of relying on hard minimum selection or uniform averaging, SERA constructs targets through a soft reliability-based aggregation centered around the median critic estimate. This allows the framework to suppress unstable or inconsistent critics while avoiding the excessive pessimism often introduced by minimum-based target operators. The proposed method is developed within the centralized training and decentralized execution paradigm and can be integrated into standard off-policy MARL pipelines with minimal modification. In addition, the adaptive entropy mechanism further aligns exploration with the stability of critic predictions, improving the robustness of policy updates during training. Experiments on multiple cooperative continuous-control benchmarks demonstrate that SERA consistently outperforms the strong baseline methods. The results show that reliability-aware target aggregation can serve as an effective alternative to conventional ensemble reduction strategies.

References

Johannes Ackermann, Volker Gabler, Takayuki Osa, and Masashi Sugiyama. Reducing overestimation bias in multi-agent domains using double centralized critics. *arXiv preprint arXiv:1910.01465*, 2019.

Emmanuel Bengio, Joelle Pineau, and Doina Precup. Interference and generalization in temporal difference learning. In *International Conference on Machine Learning*, pp. 767–777. PMLR, 2020.

Jingliang Duan, Yang Guan, Yangang Ren, Shengbo Eben Li, and Bo Cheng. Addressing value estimation errors in reinforcement learning with a state-action return distribution function. *arXiv preprint arXiv:2001.02811*, 2020.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

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*, pp. 1861–1870. Pmlr, 2018.

Frank R Hampel. A general qualitative definition of robustness. *The annals of mathematical statistics*, 42 (6):1887–1896, 1971.

Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International conference on machine learning*, pp. 2961–2970. PMLR, 2019.

Seungchan Kim, Kavosh Asadi, Michael Littman, and George Konidaris. Deepmellow: removing the need for a target network in deep q-learning. In *Proceedings of the twenty eighth international joint conference on artificial intelligence*, 2019.

Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. Maxmin q-learning: Controlling the estimation bias of q-learning. *arXiv preprint arXiv:2002.06487*, 2020.

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*, pp. 6131–6141. PMLR, 2021.

Qiang Li, Ranyang Li, Kaifan Ji, and Wei Dai. Kalman filter and its application. In *2015 8th international conference on intelligent networks and intelligent systems (ICINIS)*, pp. 74–77. IEEE, 2015.

Wenhao Li, Xiangfeng Wang, Bo Jin, Junjie Sheng, and Hongyuan Zha. Dealing with non-stationarity in marl via trust-region decomposition. *arXiv preprint arXiv:2102.10616*, 2021.

Litian Liang, Yaosheng Xu, Stephen McAleer, Dailin Hu, Alexander Ihler, Pieter Abbeel, and Roy Fox. Reducing variance in temporal-difference value estimation via ensemble of deep networks. In *International Conference on Machine Learning*, pp. 13285–13301. PMLR, 2022.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Stephen McAleer Yali Du Stefano V. Albrecht Lukas Schäfer, Oliver Slumbers and David Mguni. Ensemble value functions for efficient exploration in multi-agent reinforcement learning.

Jiafei Lyu, Xiaoteng Ma, Jiangpeng Yan, and Xiu Li. Efficient continuous control with double actors and regularized critics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7655–7663, 2022.

Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. *arXiv preprint arXiv:2102.04402*, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International conference on machine learning*, pp. 2681–2690. PMLR, 2017.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Hsiao-Ru Pan and Bernhard Schölkopf. On the variance of temporal difference learning and its reduction using control variates. In *Eighteenth European Workshop on Reinforcement Learning*, 2025.

Zhizhou Ren, Guangxiang Zhu, Hao Hu, Beining Han, Jianglun Chen, and Chongjie Zhang. On the estimation bias in double q-learning. *Advances in Neural Information Processing Systems*, 34:10246–10259, 2021.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2003.

Macheng Shen and Jonathan P How. Implicit ensemble training for efficient and robust multiagent reinforcement learning. *Transactions on Machine Learning Research*, 2023.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Jordan Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In *Proceedings of the 1993 connectionist models summer school*, pp. 255–263. Psychology Press, 2014.

John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Tengyu Xu, Zhe Wang, Yi Zhou, and Yingbin Liang. Reanalysis of variance reduced temporal difference learning. *arXiv preprint arXiv:2001.01898*, 2020.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Shangtong Zhang and Hengshuai Yao. Ace: An actor ensemble algorithm for continuous control with tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5789–5796, 2019.

A Appendix

A.1 Proofs

Proof of Lemma 4.1

Proof. For a fixed sample $b \in \mathcal{B}$, the denominator of the SERA weight is common to all non-median critics:

$$Z^{(b)} = \sum_{k \neq m_b} \exp\left(-d_k^{(b)}/\tau\right).$$

Since $\tau > 0$, the function

$$f(d) = \exp(-d/\tau)$$

is strictly decreasing in d. Therefore, if

$$d_i^{(b)} < d_j^{(b)},$$

then

$$\exp\left(-d_i^{(b)}/\tau\right) > \exp\left(-d_j^{(b)}/\tau\right).$$

Dividing both sides by the same positive denominator $Z^{(b)}$ gives $w_i^{(b)} > w_j^{(b)}$.

This proves the sample-wise ordering.

Now suppose that $d_i^{(b)} \leq d_j^{(b)} \qquad \forall b \in \mathcal{B}$, with strict inequality for at least one sample. By the sample-wise result,

$$w_i^{(b)} \geq w_j^{(b)} \qquad \forall b \in \mathcal{B},$$

and

$$w_i^{(b)} > w_j^{(b)}$$

for at least one sample. Summing over all samples gives

$$\sum_{b \in \mathcal{B}} w_i^{(b)} > \sum_{b \in \mathcal{B}} w_j^{(b)}.$$

Dividing by $|\mathcal{B}|$ yields

$$\frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} w_i^{(b)} > \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} w_j^{(b)}.$$

Hence, a critic that stays closer to the median anchor across the mini-batch receives a larger average SERA reliability weight. □

Proof of Proposition 4.2

Proof. Define

$$\lambda_{\text{med}} = \alpha, \qquad \lambda_i = (1-\alpha) w_i^{(b)}, \quad i \neq m_b.$$

Since $0 \leq \alpha \leq 1$ and $w_i^{(b)} \geq 0$, it follows that

$$\lambda_{\text{med}} \geq 0, \qquad \lambda_i \geq 0.$$

In addition,

$$\lambda_{\text{med}} + \sum_{i \neq m_b} \lambda_i = \alpha + (1-\alpha) \sum_{i \neq m_b} w_i^{(b)} = \alpha + (1-\alpha) = 1.$$

Hence, $Q_{\text{SERA}}^{(b)}$ can be written as a convex combination of the target critic estimates.

Because the median $Q_{\text{med}}^{(b)}$ is one of the target critic values, the SERA estimate is a convex combination of the set $\{Q_{\bar{\theta}_i}^{(b)}\}_{i=1}^{N}$. Therefore, it lies within their convex hull.

Since the critic outputs are scalar, this convex hull reduces to the interval

$$\left[\min_i Q_{\bar{\theta}_i}^{(b)},\ \max_i Q_{\bar{\theta}_i}^{(b)}\right].$$

Thus,

$$\min_i Q_{\bar{\theta}_i}^{(b)} \leq Q_{\text{SERA}}^{(b)} \leq \max_i Q_{\bar{\theta}_i}^{(b)}.$$

This shows that the aggregation is not forced to select the smallest critic value, as in min-based methods, while at the same time remaining within the range of the ensemble predictions. □

This result shows that SERA does not only reduce variance at the critic-output level, but also induces a lower-variance bootstrapped TD target. Since critic learning is driven by regression toward this target, the resulting update signal is less noisy than that obtained from a single target critic.

Proof of Proposition 4.3

Proof. For a fixed sample b, let the target critic estimates be $Q_{\bar{\theta}_i}^{(b)} = y^\star + \varepsilon_i$, where $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$, and the critic errors $\{\varepsilon_i\}_{i=1}^{N}$ are mutually independent. Since the median anchor corresponds to one of the target critic estimates, there exists an index m_b such that $Q_{\text{med}}^{(b)} = Q_{\bar{\theta}_{m_b}}^{(b)}$. The SERA target can therefore be written as

$$Q_{\text{SERA}}^{(b)} = \alpha Q_{\bar{\theta}_{m_b}}^{(b)} + (1-\alpha)\sum_{i \neq m_b} w_i^{(b)} Q_{\bar{\theta}_i}^{(b)}.$$

Define the effective aggregation weights

$$\tilde{w}_i^{(b)} = \begin{cases} \alpha, & i = m_b, \\ (1-\alpha) w_i^{(b)}, & i \neq m_b. \end{cases}$$

Then,

$$Q_{\text{SERA}}^{(b)} = \sum_{i=1}^{N} \tilde{w}_i^{(b)} Q_{\bar{\theta}_i}^{(b)}.$$

Because

$$\sum_{i \neq m_b} w_i^{(b)} = 1,$$

the effective weights satisfy

$$\sum_{i=1}^{N} \tilde{w}_i^{(b)} = \alpha + (1-\alpha) = 1.$$

Hence, $Q_{\text{SERA}}^{(b)}$ is a convex combination of the critic estimates. Substituting $Q_{\bar{\theta}_i}^{(b)} = y^\star + \varepsilon_i$, we obtain

$$Q_{\text{SERA}}^{(b)} = y^\star + \sum_{i=1}^{N} \tilde{w}_i^{(b)} \varepsilon_i.$$

Conditioned on the aggregation weights, the variance becomes

$$\text{Var}\left(Q_{\text{SERA}}^{(b)}\right) = \text{Var}\left(\sum_{i=1}^{N} \tilde{w}_i^{(b)} \varepsilon_i\right).$$

Using the independence of the critic errors,

$$\mathrm{Var}\Big(Q_{\mathrm{SERA}}^{(b)}\Big) = \sum_{i=1}^{N} \Big(\tilde{w}_i^{(b)}\Big)^2 \mathrm{Var}(\varepsilon_i).$$

Since each critic has variance σ^2,

$$\mathrm{Var}\Big(Q_{\mathrm{SERA}}^{(b)}\Big) = \sigma^2 \sum_{i=1}^{N} \Big(\tilde{w}_i^{(b)}\Big)^2.$$

The effective weights are nonnegative and sum to one. Therefore,

$$\sum_{i=1}^{N} \Big(\tilde{w}_i^{(b)}\Big)^2 \leq 1.$$

Moreover, because $0 < \alpha < 1$, at least two effective weights are nonzero, implying

$$\sum_{i=1}^{N} \Big(\tilde{w}_i^{(b)}\Big)^2 < 1.$$

Thus,

$$\mathrm{Var}\Big(Q_{\mathrm{SERA}}^{(b)}\Big) < \sigma^2.$$

Since each individual critic satisfies

$$\mathrm{Var}\Big(Q_{\bar{\theta}_i}^{(b)}\Big) = \sigma^2,$$

we conclude that

$$\mathrm{Var}\Big(Q_{\mathrm{SERA}}^{(b)}\Big) < \mathrm{Var}\Big(Q_{\bar{\theta}_i}^{(b)}\Big)$$

for any critic i. □

Reduced variance of the SERA TD target : For a fixed transition, define the TD target based on a single critic j as

$$y_j = r + \gamma Q_{\bar{\theta}_j}(s', \mathbf{a}'),$$

and the SERA TD target as

$$y_{\mathrm{SERA}} = r + \gamma Q_{\mathrm{SERA}}(s', \mathbf{a}').$$

If the reward term is fixed for the sampled transition, then $\mathrm{Var}(y_{\mathrm{SERA}}) = \gamma^2 \mathrm{Var}(Q_{\mathrm{SERA}})$ and $\mathrm{Var}(y_j) = \gamma^2 \mathrm{Var}(Q_{\bar{\theta}_j})$. Therefore, by Proposition 4.3, $\mathrm{Var}(y_{\mathrm{SERA}}) < \mathrm{Var}(y_j)$.

Proof of Lemma 4.4

Proof. Let the estimation error at step t be

$$e_t = Q_t - Q^\star(s, a), \qquad \varepsilon_t = y_t - Q^\star(s, a).$$

Both quantities are assumed to be zero-mean, so that $\mathbb{E}[e_t] = 0$ and $\mathbb{E}[\varepsilon_t] = 0$.

Consider the critic update written as a convex combination of the current estimate and the target:

$$Q_{t+1} = (1 - \beta_t)Q_t + \beta_t y_t.$$

Subtracting $Q^\star(s, a)$ from both sides gives

$$\begin{aligned} e_{t+1} &= Q_{t+1} - Q^\star(s, a) \\ &= (1 - \beta_t)Q_t + \beta_t y_t - Q^\star(s, a) \\ &= (1 - \beta_t)\big(Q_t - Q^\star(s, a)\big) + \beta_t\big(y_t - Q^\star(s, a)\big) \\ &= (1 - \beta_t)e_t + \beta_t \varepsilon_t. \end{aligned}$$

Squaring both sides yields

$$e_{t+1}^2 = (1-\beta_t)^2 e_t^2 + \beta_t^2 \varepsilon_t^2 + 2(1-\beta_t)\beta_t e_t \varepsilon_t.$$

Taking expectations, we obtain

$$\mathbb{E}[e_{t+1}^2] = (1-\beta_t)^2 \mathbb{E}[e_t^2] + \beta_t^2 \mathbb{E}[\varepsilon_t^2] + 2(1-\beta_t)\beta_t \mathbb{E}[e_t \varepsilon_t].$$

Since e_t and ε_t are uncorrelated, the cross term vanishes:

$$\mathbb{E}[e_t \varepsilon_t] = 0.$$

Therefore,

$$\mathbb{E}[e_{t+1}^2] = (1-\beta_t)^2 \mathbb{E}[e_t^2] + \beta_t^2 \mathbb{E}[\varepsilon_t^2].$$

Using the definitions $P_t = \mathbb{E}[e_t^2]$ and $R_t = \mathbb{E}[\varepsilon_t^2]$, we obtain

$$\mathbb{E}\big[(Q_{t+1} - Q^\star(s,a))^2\big] = (1-\beta_t)^2 P_t + \beta_t^2 R_t.$$

To obtain the variance-minimizing step size, differentiate with respect to β_t:

$$\frac{\partial}{\partial \beta_t}\Big[(1-\beta_t)^2 P_t + \beta_t^2 R_t\Big] = -2(1-\beta_t)P_t + 2\beta_t R_t.$$

Setting the derivative to zero gives

$$-2(1-\beta_t)P_t + 2\beta_t R_t = 0,$$

which simplifies to

$$\beta_t^\star = \frac{P_t}{P_t + R_t}.$$

□

A.2 Sensitivity Analysis of the Adaptive Learning Rate

In this subsection, we analyze how the adaptive learning rate $\beta_t^\star$ responds to variations in two key sources of uncertainty: uncertainty in the critic's current estimate and uncertainty in the temporal-difference (TD) target. We further discuss how these factors jointly influence the resulting learning behavior.

- **High critic uncertainty, low target uncertainty.** When P_t is large and R_t is small, $\beta_t^\star$ moves closer to one, producing a larger update step. This occurs when the critic estimate is uncertain but the TD target remains reliable, making stronger updates desirable.
- **High critic uncertainty, high target uncertainty.** If both P_t and R_t are large, the update is automatically moderated because the denominator also increases. Although the critic is uncertain, the target itself is noisy, so overly aggressive updates are avoided.
- **Low critic uncertainty, low target uncertainty.** When both uncertainty measures are small, the learning rate stays moderate. In this setting, the critic estimate is already stable and the TD target is reliable, so learning mainly performs gradual refinement.
- **Low critic uncertainty, high target uncertainty.** When P_t is small but R_t is large, the learning rate decreases. Here, the critic estimate is comparatively reliable, whereas the TD target is noisy, and smaller updates help maintain stability.

Overall, the adaptive learning rate does not depend on critic uncertainty alone; rather, it reflects a balance between the trustworthiness of the current value estimate and the reliability of incoming TD targets. This balance enables stable learning while still allowing rapid adaptation when informative and reliable targets are available.

A.3 Decorrelation strategy

Relationship between R_t and R_{gain}: We now clarify the connection between R_t and R_{gain}. For simplicity, we first present the derivation using a generic target critic $Q_{\bar{\theta}}$, and later extend the formulation to each ensemble member $Q_{\bar{\theta}_i}$.

For a transition (s, a, r, s'), $y = r + \gamma Q_{\bar{\theta}}(s', a')$, denotes TD target for training, where $Q_{\bar{\theta}}(s', a')$ is the target estimate (used by DDPG, SAC and TD3).

Bellman optimality equation is,

$$Q^\star(s, a) = \mathbb{E}[r + \gamma Q^\star(s', a') \mid s, a],$$

where $a' = \pi^\star(s') = \arg\max_{\bar{a}} Q^\star(s', \bar{a})$.

Now, subtracting $Q^\star(s, a)$ from y, and by addition and subtraction of $\gamma Q^\star(s', a')$ in the RHS, we get:

$$y - Q^\star(s, a) = \Big(r + \gamma Q^\star(s', a') - Q^\star(s, a)\Big) + \gamma\Big(Q_{\bar{\theta}}(s', a') - Q^\star(s', a')\Big).$$

Let us define the first term as Bellman sampling noise and take variance on both sides,

$$\text{Var}\left(y_t - Q^\star(s, a)\right) = \text{Var}(\epsilon_B) + \gamma^2 \,\text{Var}\left(Q_{\bar{\theta}}(s', a')) - Q^\star(s', a')\right) + 2\gamma \,\text{Cov}\Big(\epsilon_B, Q_{\bar{\theta}}(s', a') - Q^\star(s', a')\Big).$$

In practice, the Bellman sampling noise and the target estimation error are expected to have only limited correlation. Therefore, the covariance term is treated as sufficiently small and is ignored to simplify the analysis. The first term reflects transition noise and is independent of the critic. On ignoring the scaling factor γ^2, we get

$$R_t := \text{Var}\left(y_t - Q^\star(s, a)\right) \approx \text{Var}\left(Q_{\bar{\theta}}(s', a') - Q^\star(s', a')\right). \tag{2}$$

Since $Q^\star(s', a')$ is not known, we approximate it using the ensemble-based estimate $Q_{\text{SERA}}(s', a')$ introduced in Section 5.1 as a stable target, consistent with the DRL ensemble papers like MeanQ, EMAX and others. Thus,

$$R_t \approx \text{Var}\left(Q_{\bar{\theta}}(s', a') - Q_{\text{SERA}}(s', a')\right). \tag{3}$$

This equation (3) matches the formulation of R_{gain}. Thus R_{gain} serves as the practical proxy of R_t, that is, for each critic i, we estimate the measurement variance as

$$R_{\text{gain},i} = \text{Var}_{j \in \mathcal{B}}\left(Q_{\bar{\theta}_i}(s'_j, \mathbf{a}'_j) - Q_{\text{SERA}}(s'_j, \mathbf{a}'_j)\right), \tag{14}$$

where $\mathcal{B}$ denotes the mini-batch. Here, $Q_{\text{SERA},j}$ serves as the aggregated target, and the deviation from it reflects the critic's estimation error.

MLE-style justification of the SERA proxy. Since the true value $Q^\star(s', a')$ is not directly available during training, we motivate the use of $Q_{\text{SERA}}(s', a')$ as a reliability-weighted proxy from a likelihood perspective. For a fixed next state–action pair (s', a'), assume that each target critic provides a noisy estimate of the true value:

$$Q_{\bar{\theta}_i}(s', a') = Q^\star(s', a') + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, \sigma_i^2),$$

where σ_i^2 represents the uncertainty associated with critic i.

Let q denote a candidate estimate of $Q^\star(s', a')$. Under the Gaussian noise assumption, the likelihood of observing the ensemble predictions is

$$p\big(\{Q_{\bar{\theta}_i}(s', a')\}_{i=1}^N \mid q\big) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\big(Q_{\bar{\theta}_i}(s', a') - q\big)^2}{2\sigma_i^2}\right).$$

Maximizing this likelihood is equivalent to minimizing the negative log-likelihood:

$$\mathcal{L}(q) = \sum_{i=1}^{N} \frac{\big(Q_{\bar{\theta}_i}(s', a') - q\big)^2}{2\sigma_i^2}.$$

Differentiating with respect to q gives

$$\frac{\partial \mathcal{L}(q)}{\partial q} = \sum_{i=1}^{N} \frac{q - Q_{\bar{\theta}_i}(s', a')}{\sigma_i^2}.$$

Setting the derivative to zero,

$$\sum_{i=1}^{N} \frac{q - Q_{\bar{\theta}_i}(s', a')}{\sigma_i^2} = 0,$$

which yields

$$q \sum_{i=1}^{N} \frac{1}{\sigma_i^2} = \sum_{i=1}^{N} \frac{Q_{\bar{\theta}_i}(s', a')}{\sigma_i^2}.$$

Therefore, the maximum-likelihood estimate is

$$\hat{q}_{\mathrm{MLE}} = \frac{\sum_{i=1}^{N} \sigma_i^{-2} Q_{\bar{\theta}_i}(s', a')}{\sum_{i=1}^{N} \sigma_i^{-2}}.$$

Thus, under Gaussian critic noise, the likelihood-optimal estimate of $Q^\star(s', a')$ becomes an inverse-uncertainty weighted combination of the target critic estimates. In practice, however, the exact variances σ_i^2 are not known. SERA instead uses a median-anchored reliability measure, assigning larger weights to critics that remain closer to the median while suppressing high-deviation critics. This leads to the reliability-weighted estimate $Q_{\mathrm{SERA}}(s', a')$, which serves as a practical approximation of the unknown value $Q^\star(s', a')$.

Using $Q_{\mathrm{SERA}}(s', a')$ as a practical approximation of $Q^\star(s', a')$, the critic-specific target uncertainty can be estimated through the observable residual variance

$$\mathrm{Var}\left(Q_{\bar{\theta}_i}(s', a') - Q_{\mathrm{SERA}}(s', a')\right).$$

Estimating this quantity across a mini-batch gives

$$R_{\mathrm{gain},i} = \mathrm{Var}_{j \in \mathcal{B}}\left(Q_{\bar{\theta}_i}(s'_j, a'_j) - Q_{\mathrm{SERA}}(s'_j, a'_j)\right),$$

which serves as a practical proxy for the critic-dependent component of the measurement variance R_t.

Connection between P_t and TD residual variance. The ideal estimation variance is defined as

$$P_t = \mathrm{Var}\big(Q_t - Q^\star(s, a)\big),$$

which measures the uncertainty in the critic's current value estimate. In temporal-difference learning, each critic is updated by comparing its current estimate against its corresponding bootstrapped target network. Therefore, the uncertainty of an individual critic is naturally reflected through the variability of its own TD residuals. Since the true value $Q^\star(s, a)$ is unknown, this quantity cannot be computed directly during training.

Consider the temporal-difference residual

$$\delta_t = y_t - Q_t,$$

where y_t denotes the bootstrapped TD target. When the target estimate is approximately unbiased, we have

$$y_t \approx Q^\star(s, a).$$

Substituting this into the TD residual gives

$$\delta_t = y_t - Q_t \approx Q^\star(s, a) - Q_t.$$

Therefore,

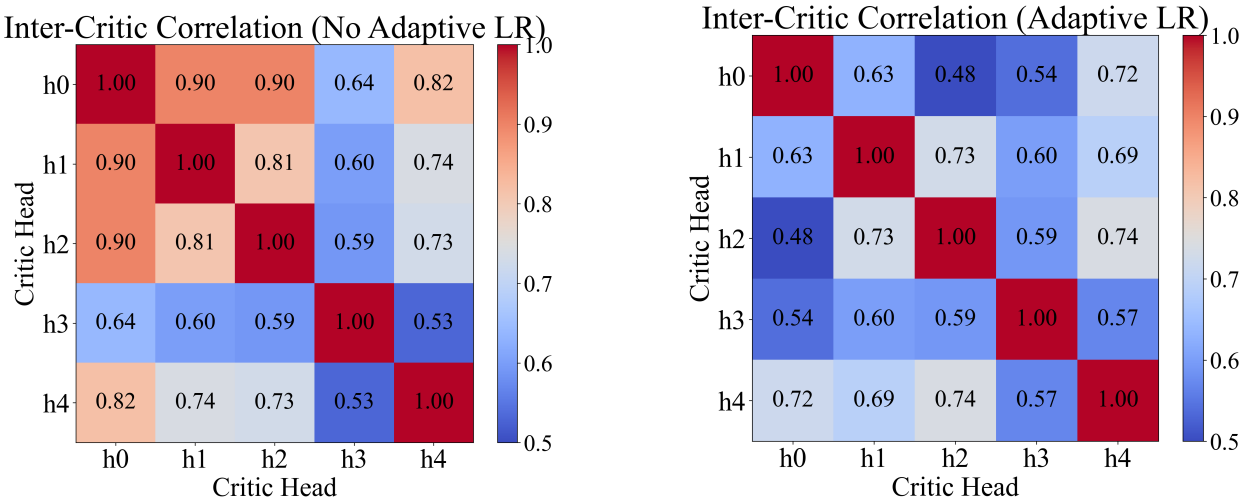
$$\mathrm{Var}(\delta_t) \approx \mathrm{Var}\big(Q_t - Q^\star(s,a)\big) = P_t.$$

This shows that the variance of TD residuals can be used as an empirical proxy for the estimation variance. Accordingly, for critic i, we estimate this quantity over a mini-batch using

$$v_{Q,i}^{(t)} \leftarrow \alpha_Q v_{Q,i}^{(t-1)} + (1-\alpha_Q)\,\mathrm{Var}_{j\in\mathcal{B}}\left(\delta_j^{(i)}\right).$$

A.4 Effect of Adaptive Learning Rates on Inter-Critic Correlation

Figure 5 shows the correlation patterns among critic heads for both fixed and adaptive learning rates, in the early phase of training. When a fixed learning rate is used, several critic pairs become highly correlated, in some cases with correlation values above 0.9, suggesting that the ensemble members tend to evolve in a strongly synchronized manner.ncy. With the adaptive learning rate, the critic correlation pattern becomes more balanced, reducing excessively strong correlations while maintaining overall agreement among critics. This suggests that the adaptive update mechanism helps prevent critics from becoming overly coupled, allowing them to follow the target values more independently instead of converging to nearly identical estimates. Although the correlations remain positive, they are less tightly clustered near one, indicating lower synchronization while still preserving consistent learning behavior.



(a) No Adaptive Learning Rate

(b) Adaptive Learning Rate

Figure 5: Inter-critic correlation matrices for a five-head ensemble critic after 50k timesteps. Critics exhibit strong correlation and near-duplicate behavior when the adaptive learning rate is not used.

B Algorithm requirements

Adaptive Temperature via Critic Disagreement. In addition to constructing stable TD targets, critic disagreement is also used to adapt the exploration behavior of the policy. For each state–action pair, disagreement among critics is measured using the mean absolute deviation from the median estimate, denoted by $U(s,a)$. This quantity reflects how consistent the critic predictions are within the ensemble.

The entropy target is then modified as

$$\mathcal{H}_{\text{target}}^{\text{eff}} = \mathcal{H}_{\text{target}} - \beta\, U(s,a), \tag{15}$$

where $\beta \geq 0$ determines the influence of the disagreement term. The temperature parameter is subsequently updated using the standard SAC objective with this adjusted entropy target.

When disagreement among critics becomes large, the value estimates are likely to be less reliable. In such cases, lowering the entropy target reduces excessively stochastic policy updates driven by noisy estimates. On the other hand, when critics produce similar predictions, the entropy target stays close to its default value, allowing exploration to continue normally. As a result, the exploration level adapts according to the stability of the value estimates, complementing the SERA aggregation strategy.

Algorithm 1 Adaptive Temperature via Critic Disagreement

1: **Input:** Ensemble critics $\{Q_{\theta^{(i)}}\}_{i=1}^{N}$, policy π_ϕ, entropy temperature α_{ent}, nominal entropy target $\mathcal{H}_{\text{target}}$, adaptation strength β
2: Sample minibatch $\{(s_j, \mathbf{a}_j)\}_{j=1}^{S}$ from replay buffer $\mathcal{D}$
3: **for** each sample $j = 1, \ldots, S$ **do**
4: Compute critic estimates: $q_j^{(i)} = Q_{\theta^{(i)}}(s_j, \mathbf{a}_j), \qquad i = 1, \ldots, N.$
5: Compute median critic estimate: $Q_{\text{med},j} = \text{median}\Big(\{q_j^{(i)}\}_{i=1}^{N}\Big)$.
6: Compute critic disagreement:

$$U_j = \frac{1}{N} \sum_{i=1}^{N} \left| q_j^{(i)} - Q_{\text{med},j} \right|.$$

7: Compute effective entropy target:

$$\mathcal{H}_{\text{target},j}^{\text{eff}} = \mathcal{H}_{\text{target}} - \beta U_j.$$

8: **end for**
9: Update entropy temperature using the SAC temperature objective:

$$J(\alpha_{\text{ent}}) = -\frac{1}{S} \sum_{j=1}^{S} \alpha_{\text{ent}} \left(\log \pi_\phi(\mathbf{a}_j | s_j) + \mathcal{H}_{\text{target},j}^{\text{eff}} \right).$$

10: Update α_{ent} by minimizing $J(\alpha_{\text{ent}})$.

Actor Update: While Q_{SERA} is employed to construct a stable and variance-controlled target for critic learning, the policy network is updated using the median ensemble estimate Q_{med}. The requirements of critic learning and actor optimization are fundamentally different. Critic updates mainly benefit from reduced target variance, whereas actor updates require a reliable optimization signal that is less affected by temporary estimation errors within the ensemble. Directly optimizing the actor with the adaptive SERA aggregation can encourage the policy to exploit critics that momentarily receive larger weights during training. Using the median estimate alleviates this issue by providing a robust central estimate of the ensemble predictions. At the same time, it avoids the excessive conservatism often introduced by minimum-based operators while remaining resistant to outlier critic estimates.

Using different aggregation strategies for critic targets and actor updates is also common in ensemble reinforcement learning methods. For instance, TD3 Fujimoto et al. (2018) constructs critic targets using the clipped minimum of two critics, whereas the actor is optimized using a single critic estimate. Similarly, SUNRISE Lee et al. (2021) incorporates ensemble uncertainty into critic target estimation through weighted Bellman backups, while policy optimization is handled separately from the uncertainty-aware aggregation process. Motivated by these observations, SERA uses variance-aware aggregation to stabilize critic learning and employs the median ensemble estimate to provide a robust and stable signal for actor optimization.

B.1 Hyper-parameters and Network Architectures

In this section, we present the network architectures and training hyperparameters used in the simulation study, which form the basis for all experimental evaluations.

B.1.1 Network Architectures

Table 3 summarizes the network designs used for each algorithm. Here, "$H \times U$" denotes H hidden layers with U units per layer.

Table 3: Network designs used for each algorithm.

Algorithm	#Actors	#Critics	Ensemble	Actor MLP	Critic MLP	Activation
MATD3	N	$2N$ (twin critic)	1	2×256	2×256	ReLU
MASAC	N	$2N$ (twin critic)	1	2×256	2×256	ReLU
MAPPO	N (shared/sep)	N (shared/sep)	1	2×128	2×128	Tanh / ReLU
Ensemble Mean	N	N	5	2×256	–	ReLU
SERA (ours)	N	N	5	2×256	–	ReLU

Notes: N denotes the number of agents. For MAPPO, actor and critic networks may be shared or separate depending on the update strategy. Ensemble sizes are shown as K=5 and can be adjusted depending on the experimental setup. To implement IET, we followed the same hyper-parameters used by Shen & How (2023).

B.1.2 Ensemble Critic Architectures

For Ensemble Mean and SERA, we employ heterogeneous ensemble critics with K=5. All critics use ReLU activations between hidden layers and linear output heads. The architectural diversity is summarized in Table 4.

Table 4: Ensemble critic architectures (K=5).

Critic	Hidden Layers	Notes
$Q^{(1)}$	2×256	Baseline width
$Q^{(2)}$	128–256–128	Deeper, tapered
$Q^{(3)}$	256–128	Asymmetric, narrower tail
$Q^{(4)}$	512–256	Wider front layer
$Q^{(5)}$	3×128	Deeper, uniformly narrow

B.1.3 Training Hyper-parameters

Table 5 reports the key training hyper-parameters used across all algorithms.

Table 5: Key training hyper-parameters.

Algorithm	LR (A/C)	γ	τ	Batch	Buffer
MATD3	$3\times10^{-4}/3\times10^{-4}$	0.99	0.005	256	10^6
MASAC	$3\times10^{-4}/3\times10^{-4}$	0.99	0.005	256	10^6
MAPPO	$3\times10^{-4}/3\times10^{-4}$	0.99	–	2048 (traj)	–
Ensemble Mean	$3\times10^{-4}/3\times10^{-4}$	0.99	0.005	256	10^6
SERA (ours)	adaptive (variance aware)	0.99	0.005	256	10^6

We did not compare with **EMAX** as it is not tested in continuous task and also serves different purpose of learning. EMAX was proposed for better exploration.

B.1.4 Variance-Adaptive Learning Rate Rescaling

To prevent the variance-adaptive gain from collapsing critic updates under high uncertainty, we rescale the effective learning rate. The per-critic gain is defined as $\kappa_t^{(i)}$ given by (11), which would otherwise yield an effective learning rate $\eta_i = \alpha_Q \kappa_t^{(i)} \to 0$ as $\kappa_t^{(i)} \to 0$.

To avoid vanishing updates, we apply a design mapping:

$$\tilde{\kappa}_t^{(i)} = \kappa_{\min} + (1 - \kappa_{\min})\kappa_t^{(i)}, \tag{16}$$

with $\kappa_{\min} = 0.2$, and define the critic step size as $\eta_i = \alpha_{\max}\tilde{\kappa}_t^{(i)}$.

Setting $\alpha_{\max} = 5 \times 10^{-4}$ yields $\eta_i \in [10^{-4}, 5 \times 10^{-4}]$. When uncertainty is low ($\kappa_t^{(i)} \approx 1$), critics take a full update, while under high uncertainty ($\kappa_t^{(i)} \approx 0$), they still perform conservative but non-negligible updates. The actor learning rate is fixed at 3×10^{-4} across all experiments.

Note: The complete code will be released upon acceptance.

Algorithm 2 Soft Ensemble Reliability Aggregation (SERA)

Initialize: M actors π_{ϕ_m}, ensemble critics $Q_{\theta^{(i)}}$, target critics $\bar{Q}_{\theta^{(i)}}$, replay buffer $\mathcal{D}$, and per-head estimation variances $P_0^{(i)}$
for each training step **do**
  Sample minibatch $\{(s_j, \mathbf{a}_j, r_j, s'_j, \gamma_j)\}_{j=1}^{S}$ from $\mathcal{D}$
  Sample target actions $a'_{j,m} \sim \bar{\pi}_{\phi_m}(\cdot|o'_{m,j})$, form $\mathbf{a}'_j$, and compute $\ell'_j = \sum_m \log \bar{\pi}_{\phi_m}(a'_{j,m}|o'_{m,j})$
  SERA: Reliability-aware target aggregation
  for each sample $j = 1, \ldots, S$ **do**
    Compute target critic values: $q_j^{(i)} = \bar{Q}_{\theta^{(i)}}(s'_j, \mathbf{a}'_j), \qquad i = 1, \ldots, N.$
    Compute median anchor and its index:

$$Q_{\text{med},j} = \text{median}\Big(\{q_j^{(i)}\}_{i=1}^{N}\Big), \qquad m_j = \text{median\_index}\Big(\{q_j^{(i)}\}_{i=1}^{N}\Big).$$

    Compute median-based deviations: $d_j^{(i)} = \left|q_j^{(i)} - Q_{\text{med},j}\right|$.
    Assign soft reliability weights to non-median critics:

$$w_j^{(i)} = \frac{\exp\Big(-d_j^{(i)}/\tau_{\text{sera}}\Big)}{\sum_{k \neq m_j} \exp\Big(-d_j^{(k)}/\tau_{\text{sera}}\Big)}, \qquad i \neq m_j.$$

    Aggregate the non-median critics: $Q_{\text{sera}\setminus\text{med},j} = \sum_{i \neq m_j} w_j^{(i)} q_j^{(i)}$.
    Construct the SERA target estimate: $Q_{\text{SERA},j} = \alpha Q_{\text{med},j} + (1-\alpha) Q_{\text{sera}\setminus\text{med},j}$.
    Construct TD target: $y_j = r_j + \gamma_j \left(Q_{\text{SERA},j} - \alpha_{\text{ent}} \ell'_j\right)$.
  end for
  Variance-adaptive critic updates
  for each critic $i = 1, \ldots, N$ **do**
    Estimate target uncertainty: $R_{\text{gain},i} = \text{Var}_{j \in B}\left(\bar{Q}_{\theta^{(i)}}(s'_j, \mathbf{a}'_j) - Q_{\text{SERA},j}\right)$.
    Compute TD error: $\delta_j^{(i)} = y_j - Q_{\theta^{(i)}}(s_j, \mathbf{a}_j)$.
    Track prediction variance using TD residuals: $v_{Q,i}^{(t)} \leftarrow \alpha_Q v_{Q,i}^{(t-1)} + (1-\alpha_Q)\,\text{Var}_{j \in B}\left(\delta_j^{(i)}\right)$.
    Update estimation variance: $P_t^{(i)} \leftarrow \alpha_P P_{t-1}^{(i)} + (1-\alpha_P) v_{Q,i}^{(t)}$.
    Compute adaptive update scale:

$$\kappa_t^{(i)} = \frac{P_t^{(i)}}{P_t^{(i)} + R_{\text{gain},i} + \eta}.$$

    Update critic:

$$\theta^{(i)} \leftarrow \theta^{(i)} + \frac{\alpha_Q}{S} \sum_{j=1}^{S} \kappa_t^{(i)} \delta_j^{(i)} \nabla_{\theta^{(i)}} Q_{\theta^{(i)}}(s_j, \mathbf{a}_j).$$

  end for
  Compute centralized value estimate for actor update: $Q_{\text{cent}}(s, \mathbf{a}) = \text{median}_i\, Q_{\theta^{(i)}}(s, \mathbf{a})$.
  for each actor $m = 1, \ldots, M$ **do**
    Update actor:

$$\nabla_{\phi_m} J(\phi_m) = \frac{1}{S} \sum_{j=1}^{S} \nabla_{\phi_m} \left[\alpha_{\text{ent}} \log \pi_{\phi_m}(a_{m,j}|o_{m,j}) - Q_{\text{cent}}(s_j, \mathbf{a}_j)\right].$$

  end for
  Soft update target networks
end for