Reducing Overestimation by Measuring Critic Disagreement in Multi-Critics Architectures

Nitsan Soffair
Ben-Gurion University of the Negev
Be'er-Sheva, Israel
Nitsan.Soffair@gmail.com

Gilad Katz
Ben-Gurion University of the Negev
Be'er-Sheva, Israel
katz.gilad@gmail.com

ABSTRACT

We introduce Ensemble Std (ES), a lightweight regularizer for multi-critic deep reinforcement learning that mitigates overestimation by calibrating target values according to critic disagreement. ES treats the dispersion of target-value estimates across an ensemble of Q-functions as an uncertainty signal, applying an adaptive subtractive penalty when disagreement is high. This yields uncertainty-aware, conservative targets that preserve the learning signal where critics agree and temper optimism where they do not–complementing minimum-based targets without imposing uniformly pessimistic updates. ES operates directly on critic targets, requiring no architectural changes, and integrates seamlessly into a wide range of actor–critic algorithms. In practice, ES was easily plugged into TD3, SAC, and TD3+BC with negligible overhead, consistently improving stability and returns while reducing variance. Overall, ES offers a simple, conceptually transparent mechanism that turns ensemble disagreement into principled value regularization, making multi-critic learners more robust in noisy, uncertain, and data-limited regimes. Our code is available at https://github.com/anonymouszxcv16/ES.

KEYWORDS

Overestimation bias, Ensemble

ACM Reference Format:
Nitsan Soffair and Gilad Katz. 2026. Reducing Overestimation by Measuring Critic Disagreement in Multi-Critics Architectures. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/FYIR8341

1 INTRODUCTION

Deep reinforcement learning (DRL) combines deep neural networks with reinforcement learning to enable agents to make sequential decisions in complex, high-dimensional environments with large state–action spaces [31, 42, 43, 45]. This capability has driven notable progress in domains such as robotics [24], video games [34, 48], and autonomous driving [20, 35]. In continuous control in particular, popular off-policy actor–critic methods such as DDPG [26], TD3 [13], and SAC [15] learn directly from experience to produce smooth actions and have established strong baselines across MuJoCo benchmarks [5, 47].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/FYIR8341

Despite these advances, DRL remains vulnerable to overestimation bias – systematic overvaluation of state–action pairs – which can destabilize learning, degrade policy quality, and propagate errors through bootstrapped targets in off-policy settings [13, 15, 22, 41, 46]. This bias often arises because target construction selects maxima under noisy or misspecified value estimates, amplifying optimistic errors through the recursive Bellman update [4, 14, 21]. The challenge is exacerbated in limited-interaction or offline setups, where distributional shift and out-of-distribution (OOD) actions make target values particularly unreliable [10, 21, 25].

To counter overestimation, two complementary strategies have emerged. First, ensemble-based approaches use critic disagreement as a proxy for uncertainty, aggregating or subsampling multiple Q-functions to form more conservative targets [6, 13, 22, 48]. Representative methods include TD3, which takes the elementwise minimum of twin critics [13], REDQ, which scales ensemble size and update ratio to improve target fidelity [6], and TD7, which augments ensembles with learned state–action representations and prioritization [11]. Second, entropy-regularized and conservative value estimation methods temper optimistic targets or penalize high Q-values, as in SAC's maximum-entropy objective [15] and offline methods such as CQL, MCQ, TD3+BC, and SAC+BC that reduce overconfidence on out-of-distribution (OOD) data [12, 15, 21, 29]. While effective, many introduce extra computation, complex objectives, or brittle hyperparameter sensitivities that hinder scalability and practical deployment [15, 21, 29].

This work introduces Ensemble Std (ES), a simple, general regularizer for multi-critic architectures that reduces overestimation by directly moderating target values according to measured critic disagreement. ES assumes an ensemble of independently trained Q-functions and treats their dispersion – standard deviation over target estimates – as an uncertainty signal. High disagreement triggers a subtractive penalty on the target, yielding a more cautious, uncertainty-aware update [6, 13]. Conceptually, ES complements minimum-based target selection (as in TD3) by adapting the degree of pessimism to the strength of consensus among critics rather than imposing a fixed pessimistic bias everywhere. Unlike approaches that require action-space sweeps or auxiliary models, ES operates directly on the critic target, making it lightweight, easy to integrate, and broadly applicable to both online and offline regimes with negligible overhead [12, 15, 21].

We evaluate ES by instantiating it on top of the TD3 and SAC algorithms for online MuJoCo control tasks and on TD3+BC for offline D4RL benchmarks spanning random, medium, and expert datasets [5, 10, 11, 47]. Across these settings, ES consistently improves base-algorithm performance, particularly in challenging

offline conditions where uncertainty and OOD actions are prevalent, while also reducing return variance and maintaining near-zero computational overhead relative to the underlying learner [6, 13, 21]. Furthermore, ES integrates smoothly with stronger ensemble variants, providing complementary gains without additional architectural complexity or increased update-to-data ratios, as is the case in previous work [6, 8, 11].

Our contributions are threefold.

- We propose ES, a statistical target-regularization technique that leverages critic disagreement to mitigate overestimation bias in multi-critic DRL, unifying conservative behavior with uncertainty awareness in a single, lightweight mechanism [6, 13, 21].
- We demonstrate that ES yields consistent improvements over strong online and offline baselines – TD3, SAC, REDQ, TD3+BC, SAC+BC, MCQ, and CQL – across MuJoCo and D4RL tasks and hyperparameters, often with reduced variance and negligible runtime overhead.
- We show that ES complements existing ensemble mechanisms and credit-assignment dynamics by moderating value propagation where critic consensus is weak, thereby improving stability and robustness in data-limited regimes [6, 32, 36, 45].

2 RELATED WORK

Deep reinforcement learning has achieved remarkable success in complex sequential decision-making tasks, from game mastery [18, 42, 43] to physical system control [31, 45]. However, a persistent challenge is *overestimation bias*, where agents systematically overvalue state-action pairs, degrading stability and performance [13, 15, 22, 23, 33]. This bias is particularly problematic in off-policy algorithms, where bootstrapping and replayed data amplify early inaccuracies through recursive Bellman updates [4, 13, 14, 21].

2.1 Ensemble and Conservative Approaches to Overestimation Bias

One prominent family of solutions leverages *Q-function ensembles* [1, 6, 13, 22, 34] to aggregate or subsample multiple critics, better capturing uncertainty. In these approaches, disagreement among critics signals uncertainty, enabling conservative behavior in ambiguous state-action regions. TD3 [13] exemplifies this principle by employing twin critics with minimum value selection. REDQ [6] further advances this direction by increasing both ensemble size and update frequency to improve target quality under limited data conditions.

Complementing ensemble methods, conservative value estimation approaches like Conservative Q-Learning (CQL) [21] and Mildly Conservative Q-Learning (MCQ) [29] directly address overestimation by explicitly penalizing high Q-values. These methods constrain Q-functions to remain pessimistic on out-of-distribution data, proving particularly crucial in offline RL settings where distributional shift poses significant challenges.

Additional strategies have emerged to address related aspects of the overestimation problem. Entropy regularization, exemplified by Soft Actor-Critic (SAC) [15], incentivizes stochastic policies to promote broader exploration and avoid premature convergence to spurious high-value actions. Meanwhile, approaches that increase the update-to-data ratio (UTD $\gg$ 1), explored in REDQ [6] and SR-SAC [8], yield more stable value targets through frequent critic updates, improving sample efficiency and reducing bias accumulation.

2.2 Architectural Improvements and Optimization

Beyond algorithmic innovations targeting overestimation bias, numerous architectural and optimization techniques have been developed to enhance learning stability more broadly. Layer normalization [2, 49] modulates per-sample activations to reduce internal covariate shift and improve gradient flow in deep critics and policies, while orthogonal initialization [38] preserves signal norms in the linear regime to mitigate exploding/vanishing gradients during early training. Meta regularization [3] and continual-learning-inspired penalties complement these effects by discouraging disruptive parameter drift, thus curbing catastrophic forgetting in nonstationary RL settings.

At the optimization level, trust region updates (TRPO) [39] bound policy changes to maintain local improvement guaranties, while clipped policy gradients (PPO) [9] provide a practical surrogate that stabilizes updates under high-variance estimates. Gradient clipping and adaptive optimizers like AdamW [28] further combat gradient instability by capping extreme updates and decoupling weight decay from adaptive moments, respectively, which collectively help mitigate catastrophic forgetting [30].

Beyond these widely adopted components, several complementary studies refine stabilization practice without overlapping the specific mechanisms emphasized above. The authors of [50] offer a batch-size-agnostic alternative to BatchNorm by normalizing within channel groups, improving stability when batch statistics are unreliable and pairing well with attention-based architectures where per-batch variance is problematic. Huang et al. [16] introduced train-time orthogonalization that maintains near-orthogonal weights beyond initialization, sustaining well-conditioned signal propagation deeper into training than fixed orthogonal init. The authors of [17] theoretically and empirically links orthogonal initialization to faster training and better generalization in the large learning-rate regime, supporting its use in deep critics.

To address gradient variance and step-size fragility in policy gradients from a different angle than PPO/TRPO, [52] studies PGPE and derives optimal baselines that reduce estimator variance, while [7] analytically integrates over actions to stabilize gradients and improve sample efficiency relative to DPG-style exploration. Finally, [19] characterizes failure modes from overly aggressive steps that cause premature entropy collapse, motivating conservative step-size control that complements clipping and trust-region concepts.

2.3 Temporal Credit Assignment

The challenge of temporal credit assignment – determining which trajectory actions are responsible for future rewards [44] – becomes particularly complex when using ensemble methods. Classic solutions like eligibility traces [36] (e.g., TD(λ) [45]) enable multi-step

reward propagation, while modern approaches such as Retrace [32] and V-trace extend these concepts with off-policy corrections. However, a high disagreement between the ensemble critics can produce noisy gradients that harm learning efficiency.

Our approach contributes to this landscape by proposing regularized ensemble targets that adjust value propagation based on critic consensus. This mechanism not only reduces overestimation bias but also improves credit assignment robustness by attenuating the influence of unreliable or out-of-distribution actions–particularly valuable in noisy learning regimes. ES builds upon ensemble-based uncertainty estimation, conservative Q-learning, value-target regularization, and robust credit assignment, offering a simple, general, low-overhead solution that integrates seamlessly into existing multi-critic frameworks while improving stability and performance across diverse DRL tasks.

3 THE PROPOSED APPROACH

3.1 Overview

We introduce a new method, **Ensemble Std** (ES), that improves stability and reliability in reinforcement learning by tempering overconfident value estimates arising from limited experience. ES encourages conservative updates precisely when internal value models disagree, reducing the risk of propagating spurious optimism through bootstrapped targets.

ES operates by maintaining an ensemble of independently trained value estimators (Q-functions) and using their dispersion as an uncertainty signal. Disagreement is quantified via the standard deviation across critics' target estimates; when this statistic is large, ES subtracts a proportional penalty from the target, yielding cautious updates where uncertainty is high while preserving learning signal where critics agree.

Unlike other popular methods like CQL [21] and SAC [15], which compute uncertainty based on either action selection or full action-space evaluations (which can be computationally heavy), our approach directly regularizes the value estimation process itself. This results in a simple, efficient, and effective method that improves training stability and reduces bias without requiring major changes to the algorithm's structure.

3.2 Ensemble Std

Let us define the learning setup formally. We consider a standard reinforcement learning setting modeled as a Markov Decision Process (MDP) [4, 37], defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Here, $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $P(s' \mid s, a)$ denotes the transition probability of reaching state $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ after taking action $a \in \mathcal{A}$, $r(s, a)$ is the immediate reward received for taking action a in state s, and $\gamma \in [0, 1)$ is the discount factor that determines the importance of future rewards.

3.2.1 Using Multiple Value Estimators (Critics). Instead of using just one value estimator, we use an ensemble of K different Q-functions, denoted by $\{Q_{\theta_i}(s, a)\}_{i=1}^{K}$. To estimate the value of an action, we take the smallest value predicted by any of these Q-functions (based on TD3's target estimation [13]):

$$Q_{\min}(s, a) = \min_i Q_{\phi_i}(s, a) \tag{1}$$

This conservative estimate helps the agent avoid becoming overly optimistic about uncertain actions.

3.2.2 Adding a Penalty. The core idea is to adjust the learning target based on *the degree of disagreement within the ensemble of Q-functions*. We measure this disagreement using the standard deviation of their value estimates, and subtract a proportional penalty from the target when disagreement is high. This produces more cautious updates in uncertain situations while leaving confident estimates largely unchanged. The updated Q-target is:

$$y = r + \gamma \cdot [Q_{\min} - \alpha \cdot \sigma_Q] \tag{2}$$

where:

$$\sigma_Q = \mathrm{std}_i[Q_{\phi_i}(s'_j, a'_j)] \tag{3}$$

and α controls how strongly we penalize actions with high uncertainty.

By calibrating target conservatism to critic disagreement, this mechanism prevents overconfident value estimation while preserving learning signal where critics reach consensus—particularly valuable in offline regimes where out-of-distribution actions lack reliable value estimates (as seen in Table 2)

3.2.3 Training the Critics. Each Q-function (critic) is trained to match the regularized target value using the standard mean squared error:

$$\frac{1}{B} \sum_{j=1}^{B} [Q_{\phi_i}(s_j, a_j) - y]^2 \tag{4}$$

3.2.4 Updating the Policy. The agent's policy is updated using feedback from all critics (e.g., average Q-values):

$$\nabla_\theta \frac{1}{K} \sum_{i=1}^{K} Q_{\phi_i}(s, a) \tag{5}$$

This encourages the policy to choose actions that multiple Q-functions agree are good.

3.2.5 Algorithm Summary. The full process combines standard actor-critic learning with our regularization. The step-by-step process of each training iteration is presented in Algorithm 1:

The training loop begins by collecting experience from the current policy with added exploration (line 2), storing it in the replay buffer (line 3). A minibatch of transitions is sampled (line 4), and for each sample, the next action is drawn from the current policy (line 6). All Q-values in the ensemble are evaluated (line 7), and a pessimistic value is estimated using the minimum across critics (line 8), penalized by the standard deviation across ensemble outputs (line 9). The critic target is then computed (line 10), incorporating both pessimism and uncertainty regularization. The critic networks are updated using the shared target (line 12), and the actor is updated with feedback from all critics (line 13). This procedure can be embedded in online (interacting with environment) or offline (from dataset) settings.

Algorithm 1 Training with Ensemble Regularization (ES)

Require: Initial policy π_θ, critics $\{Q_{\phi_i}\}_{i=1}^{K}$, replay buffer $\mathcal{D}$, temperature α

while training **do**
  Collect a transition (s, a, r, s') using π_θ with exploration noise
  Store (s, a, r, s') in replay buffer $\mathcal{D}$
  Sample minibatch $\{(s_j, a_j, r_j, s'_j)\}_{j=1}^{B}$ from $\mathcal{D}$
  for all $j \in \{1, \ldots, B\}$ **do**
    Compute next action $a'_j \sim \pi_\theta(s'_j)$
    Evaluate ensemble: $\{Q_{\phi_i}(s'_j, a'_j)\}_{i=1}^{K}$
    Compute pessimistic value: $Q_{\min} = \min_i Q_{\phi_i}(s'_j, a'_j)$
    Compute uncertainty penalty: $\sigma_Q = \mathrm{std}_i[Q_{\phi_i}(s'_j, a'_j)]$
    Compute target: $y = r + \gamma \cdot [Q_{\min} - \alpha \cdot \sigma_Q]$
  end for
  Update each critic Q_{ϕ_i} to minimize: $\mathcal{L}_{\text{critic}} = \frac{1}{B}\sum_{j=1}^{B}[Q_{\phi_i}(s_j, a_j) - y]^2$
  Update policy π_θ using feedback from all critics (e.g., average Q-values)
end while

4 EXPERIMENTAL RESULTS

We evaluate our proposed approach on standard continuous control benchmarks to assess its effectiveness in improving learning stability and sample efficiency. Our experimental evaluation encompasses both *online* (real-time environment interaction) and *offline* (learning from static datasets) learning paradigms. Specifically, we conduct experiments on six popular locomotion environments from the MuJoCo physics simulator accessed via OpenAI Gym [5, 47], and four benchmark tasks from the D4RL offline dataset collection [10]. These environments span diverse control complexities and exhibit varying data quality characteristics, providing a comprehensive testbed for evaluating ES's robustness.

We focus evaluation on continuous action domains because our proposed regularizer targets off-policy actor-critic methods with differentiable action selection – the dominant family developed for continuous control – and comparability with field-standard benchmarks (e.g., MuJoCo/D4RL) requires aligning with where these algorithms are best defined and stress-tested. Our decision mirrors common practice in recent atudies that primarily conduct their evaluation on continuous control, including [27, 40, 51], all of which concentrate their empirical evidence on continuous-control tasks where uncertainty calibration and overestimation mitigation are most directly exercised.

4.1 Experimental Setup

Benchmark Environments and Baselines.

We evaluate ES across two learning paradigms: **online** reinforcement learning with real-time environment interaction, and **offline** learning from pre-collected datasets. For online evaluation, we used six standard MuJoCo locomotion tasks: *Hopper-v2*, *HalfCheetah-v2*, *Walker2d-v2*, *Ant-v2*, *Humanoid-v2*, and *Standup-v2* [5, 47]. We compare ES-augmented algorithms against three widely adopted baselines: TD3 [13], SAC [15], and REDQ [6]. We do not integrate our approach into REDQ, but include its results to provide a point of reference against a modern actor-critic method that employs a distinct update-to-data ratio (UTD $\gg 1$) and a novel target computation scheme. Unlike TD3 and SAC, REDQ departs from the classic actor-critic design, and is therefore less suitable for direct comparison with our lightweight ES modification.

For our offline evaluation, we employ four D4RL benchmark environments (*hopper*, *cheetah*, *walker*, *ant*) [10], each evaluated across three dataset quality regimes: *Random* (data from untrained agents), *Medium* (data from partially trained agents), and *Expert* (data from near-optimal agents). We compare against established offline RL methods including TD3+BC [12], SAC+BC [35], MCQ [29], and CQL [21].

Integration and Training Protocol. To assess ES's effectiveness, we integrate our method into TD3 and SAC for online learning, and into TD3+BC for offline scenarios. All experiments follow a standardized training protocol: online agents interact with environments for 1 million steps, while both online and offline settings perform 1 million gradient updates (except REDQ with UTD $\gg 1$). Each configuration is evaluated across 5 random seeds, with reported metrics representing the average of best-performing scores and their standard deviations. To quantify the statistical significance of our results, we used the paired-T statistical significance test.

Computational Infrastructure. All experiments were conducted on a GPU cluster equipped with NVIDIA RTX 6,000 Ada Generation GPUs, ensuring consistent computational resources across all evaluations.

4.2 Online Control Benchmarks

Setup and Metrics. Table 1 reports online performance when integrating ES into TD3 and SAC across six MuJoCo control tasks, evaluated at four temperatures ($\alpha \in \{0.1, 0.2, 0.5, 1\}$), and compared to TD3, SAC, and REDQ [6] baselines, including average return, standard deviation, and relative improvement over the corresponding base algorithm (TD3 or SAC) [6]. For REDQ, we include results without applying ES to that method, following its standard configuration with increased ensemble size and update ratio [6].

Aggregate gains over TD3 and SAC.. ES yields consistent, positive average improvements over both TD3 and SAC across all temperatures, with the strongest average gains at $\alpha = 0.2$ (TD3: +18.5% ; SAC: +18.8%), despite requiring no architectural changes and incurring negligible overhead [6]. Even the most conservative ($\alpha = 0.1$) and most aggressive ($\alpha = 1$) settings produce non-trivial mean improvements, indicating robustness of ES to the temperature hyperparameter [6].

Statistical significance. To assess whether observed gains reflect systematic effects rather than stochastic variability, we conducted paired t-tests across random seeds for each base algorithm and temperature. For SAC-based experiments, all ES configurations attain p-values of $p < 0.05$ or lower, with the strongest strongest significance obained for $\alpha = 0.1$ with $p < 0.01$ for all evaluated tasks. For our TD3-based experiments, our approach achieved $p < 0.05$ for the configuration of $\alpha = 0.1$.

Table 1: Online evaluation results for our method (ES) integrated into the TD3 and SAC algorithms. Each row reports performance on a specific MuJoCo task, comparing various ES temperature configurations to the corresponding TD3 or SAC baseline. The best result in each row is highlighted in bold. The final row (Improvement) shows the average percentage gain of each ES variant over the baseline. REDQ-SAC is included for additional comparison against standard TD3 and SAC. Asterisks (*) denote configurations where the average p-value across tasks falls below the 5% threshold, indicating statistically significant improvement.

Env.	TD3	REDQ	ES ($\alpha = 0.1$)	ES ($\alpha = 0.2$)	ES ($\alpha = 0.5$)	ES ($\alpha = 1$)
Hopper	3.5K ± 70	2.9K ± 740	3.5K ± 41	3.4K ± 149	3.4K ± 310	2.6K ± 1K
Cheetah	9.6K ± 5.7K	9.1K ± 5.4K	**12.5K ± 361**	12.5K ± 1.7K	10K ± 5.9K	12K ± 1K
Walker	5.1K ± 901	4.9K ± 557	4.1K ± 2.3K	5.1K ± 952	**5.3K ± 771**	4.6K ± 1.5K
Ant	3.8K ± 4.1K	4.1K ± 3.1K	6.1K ± 652	6.4K ± 520	5.8K ± 1.7K	**6.4K ± 855**
Humanoid	5.2K ± 2.7K	4.3K ± 2.4K	5.4K ± 2.9K	6.2K ± 142	**6.5K ± 295**	5.1K ± 2.7K
Standup	155K ± 10.1K	145.9K ± 15.3K	**183K ± 47K**	152K ± 16K	156K ± 19K	151K ± 16K
Improvement	–	-6.6%	15.3%*	**18.5%**	13.1%	9.0%
Env.	SAC	REDQ	ES ($\alpha = 0.1$)	ES ($\alpha = 0.2$)	ES ($\alpha = 0.5$)	ES ($\alpha = 1$)
Hopper	2.4K ± 1.3K	2.9K ± 740	**3.5K ± 110**	3.3K ± 387	3K ± 898	3.1K ± 910
Cheetah	9.4K ± 5.7K	9.1K ± 5.4K	10.8K ± 2.5K	**11.9K ± 649**	11.5K ± 790	9.2K ± 5.5K
Walker	5.4K ± 410	4.9K ± 557	5.4K ± 555	5.4K ± 783	5.4K ± 868	4.9K ± 640
Ant	**6.3K ± 460**	4.1K ± 3.1K	5.9K ± 510	4.5K ± 2.9K	5.8K ± 941	3.6K ± 3.8K
Humnaoid	6.5K ± 417	4.3K ± 2.4K	3.7K ± 3.2K	6.6K ± 374	5.1K ± 2.8K	5.1K ± 2.7K
Standup	119K ± 5.1K	145.9K ± 15.3K	152K ± 6.2K	**153K ± 3.1K**	122K ± 35K	135K ± 18K
Improvement	–	0.5%	15.1%*	**18.8%***	14.2%*	14.0%*

Comparison to REDQ. Although REDQ leverages larger ensembles and higher update-to-data ratios to improve targets, it does not consistently surpass TD3 or SAC under this evaluation protocol, showing an average decrease in perofrmance of -6.6% versus TD3 and a minor improvement of +0.5% versus SAC. In contrast, ES achieves substantially higher average returns while avoiding REDQ's additional computational cost from frequent target updates and multiple critics.

Variance and seed robustness. ES reduces return variance relative to TD3 and SAC across multiple tasks, improving stability under stochastic dynamics [6]. For instance, in Ant with TD3, the standard deviation decreases from 4.1K to 520 at with $\alpha = 0.2$, and similar reductions are observed in Humanoid, reflecting improved reproducibility across seeds and more stable training dynamics.

summary. Overall, ES provides a lightweight, uncertainty-aware regularization that improves actor–critic performance and stability across online MuJoCo tasks, with $\alpha = 0.2$ performing best on average and consistently outperforming the TD3 and SAC baselines as well as REDQ in this benchmark. These results support ES as a practical, plug-in enhancement for online control with strong mean gains and statistically significant improvements under standard paired testing across seeds.

4.3 Offline Control Benchmarks

We evaluate ES on four D4RL tasks (Hopper, Walker2d, HalfCheetah, and Ant) under three dataset qualities—**Random** (low-quality, off-policy), **Medium** (moderate-quality), and **Expert** (high-quality, near-optimal)—by applying ES on top of TD3+BC and comparing against SAC+BC, TD3+BC+CQL, and MCQ across temperature values $\alpha \in \{0.1, 0.2, 0.5, 1.0\}$ that scale the ES regularization strength.

Table 2 reports results across all benchmarks, from which three themes emerge: robustness to dataset quality, consistent trends with the temperature α, and competitiveness against strong offline baselines

Robustness to Dataset Quality. ES exhibits larger gains as dataset quality declines (i.e., more OOD actions), with the strongest improvements in the **Random** regime where $\alpha = 1$ yields an average **+11.8%** over TD3+BC and ranks first in 3 of 4 tasks; all ES variants outperform SAC+BC and MCQ and match or exceed CQL, indicating strong resilience to noisy, off-policy data. In **Medium**, lighter regularization is preferable: $\alpha = 0.1$ provides a **+4.3%** average improvement, reflecting that cleaner data benefits from weaker penalization that preserves confident value updates. In **Expert**, ES remains competitive with lower regularization, with $\alpha = 0.5$ achieving the best average gain of **+6.5%** over TD3+BC and outperforming SAC+BC and CQL on average, showing that ES does not degrade performance even with near-optimal data.

Temperature trends (α). The effect of ES follows a clear α-dependent pattern tied to uncertainty penalization: for **Random** datasets, larger temperature values (e.g., $\alpha = 1$) consistently dominates smaller values, aligning with the need for stronger caution under high uncertainty and OOD exposure. For **Expert** datasets, smaller temperature values (e.g., $\alpha = 0.5$) performs best, as aggressive regularization can unnecessarily dampen learning when targets are reliable.

Consistency across regimes. Across **Random**, **Medium**, and **Expert**, **every ES configuration improves the TD3+BC average**, demonstrating robustness to the choice of α and supporting ES as a plug-in regularizer for offline learning. Paired t-tests indicate these

Table 2: Offline evaluation results for our method (ES) integrated into the TD3+BC algorithm across three difficulty levels of D4RL benchmark tasks. Each row reports performance on a specific task, comparing different ES configurations against the original TD3+BC baseline. The best score in each row is highlighted in bold. The last row (Improvement) shows the average percentage gain of each method over TD3+BC. Results are also shown for other baselines: CQL combined with TD3+BC, SAC+BC, and MCQ. Asterisks (*) indicate settings where the average p-value across tasks is below the standard 5% threshold, denoting statistically significant improvement.

Env.	TD3+BC	MCQ	SAC+BC	CQL (TD3+BC)	ES ($\alpha = 0.1$)	ES ($\alpha = 0.2$)	ES ($\alpha = 0.5$)	ES ($\alpha = 1$)
Random								
Hopper	27.1 ± 10.5	24.9 ± 11.0	**31.5 ± 0.5**	21.9 ± 13.2	26.9 ± 10.9	22.7 ± 12.4	24.5 ± 11.1	28.2 ± 8.1
Cheetah	17.1 ± 1.9	11.7 ± 0.8	16.6 ± 1.0	16.4 ± 0.6	17.3 ± 0.9	17.7 ± 0.5	17.1 ± 0.8	**18.4 ± 1.6**
Walker	9.6 ± 7.6	6.8 ± 0.7	11.9 ± 8.0	13.1 ± 7.9	12.4 ± 10.5	**14.5 ± 9.6**	12.1 ± 10.0	11.7 ± 9.8
Ant	58.3 ± 2.8	34.0 ± 5.2	52.9 ± 9.9	57.8 ± 5.5	57.7 ± 3.7	62.8 ± 3.7	**67.0 ± 3.3**	66.5 ± 6.0
Improvement	–	-27.8%*	4.6%	3.1%	7.1%	11.5%*	7.8%*	**11.8%***
Medium								
Hopper	85.8 ± 14.9	28.8 ± 20.5	78.1 ± 25.3	81.7 ± 13.2	**93.2 ± 12.1**	87.2 ± 7.0	89.2 ± 15.4	88.3 ± 12.6
Cheetah	57.1 ± 0.3	12.5 ± 22.5	57.0 ± 0.8	37.0 ± 4.6	57.3 ± 1.0	57.2 ± 0.7	57.2 ± 0.5	**58.0 ± 0.1**
Walker	82.2 ± 9.8	7.0 ± 14.6	77.6 ± 11.1	83.8 ± 3.6	**87.0 ± 2.2**	86.0 ± 7.3	86.3 ± 1.0	85.4 ± 7.1
Ant	132.4 ± 2.1	29.8 ± 0.6	128.2 ± 6.1	63.7 ± 18.6	**135.4 ± 2.3**	130.4 ± 7.2	131.5 ± 3.9	132.1 ± 5.8
Improvement	–	-78.4%*	-4.5%*	-22.5%*	**4.3%***	1.2%	2.1%	2.0%
Expert								
Hopper	92.3 ± 29.5	76.5 ± 36.6	77.9 ± 26.5	**99.7 ± 22.1**	94.6 ± 26.9	98.3 ± 18.5	98.9 ± 10.5	96.1 ± 12.3
Cheetah	78.2 ± 9.8	28.3 ± 26.7	79.7 ± 6.6	**82.4 ± 4.4**	77.7 ± 13.9	80.7 ± 4.6	80.2 ± 6.2	80.3 ± 5.1
Walker	112.2 ± 0.3	41.7 ± 57.6	112.2 ± 0.2	111.9 ± 0.2	112.1 ± 0.1	112.1 ± 0.4	112.1 ± 0.2	**112.2 ± 0.4**
Ant	76.3 ± 9.3	29.8 ± 0.6	**89.9 ± 24.2**	60.8 ± 12.4	82.4 ± 24.3	83.2 ± 40.8	89.0 ± 21.0	85.3 ± 26.0
Improvement	–	-51.2%*	1.0%*	-1.8%*	2.4%	4.6%	**6.5%***	4.6%

improvements are statistically significant in key settings: in **Random**, high-temperature variants ($\alpha \in \{0.2, 0.5, 1\}$) achieve p-values below $p < 0.05$. In **Medium**, $\alpha = 0.5$ achieves the highest average improvement (4.3%) and has the highest statistical significance of $p < 0.05$. In **Expert**, $\alpha = 0.5$ both delivers the highest average gain (6.5%) and has the highest statistical significance with $p < 0.05$.

Comparison to other offline methods. MCQ shows large negative averages across all regimes (Random: -27.8%; Medium: -78.4%; Expert: -51.2%), indicating sensitivity to value mis-estimation and weak generalization. SAC+BC yields modest positives in Random (4.6%) and Expert (1.0%), but underperforms in Medium (-4.5%), whereas ES outperforms SAC+BC in all regimes. CQL is competitive in Expert and Random but collapses in Medium (-2.5%), while ES achieves higher overall average gains and substantially lower variance across dataset qualities.

Takeaways. ES provides a **lightweight, robust** improvement to TD3+BC that scales across dataset regimes, adheres to predictable α selection heuristics (higher α for noisier data; lower α for cleaner data), and consistently outperforms strong baselines without architectural overhead. The combination of average improvements, favorable regime-wise rankings, and statistically significant t-test results clearly shows ES to be a practical choice for real-world offline RL pipelines.

5 ANALYSIS

To elucidate the mechanisms underlying ES's performance improvements, we conduct a comprehensive analysis examining two critical aspects of the learning dynamics: (1) ensemble disagreement patterns across different data quality regimes, and (2) computational efficiency relative to baseline methods. These analyses provide theoretical insight into when and why ES excels, while demonstrating its practical utility.

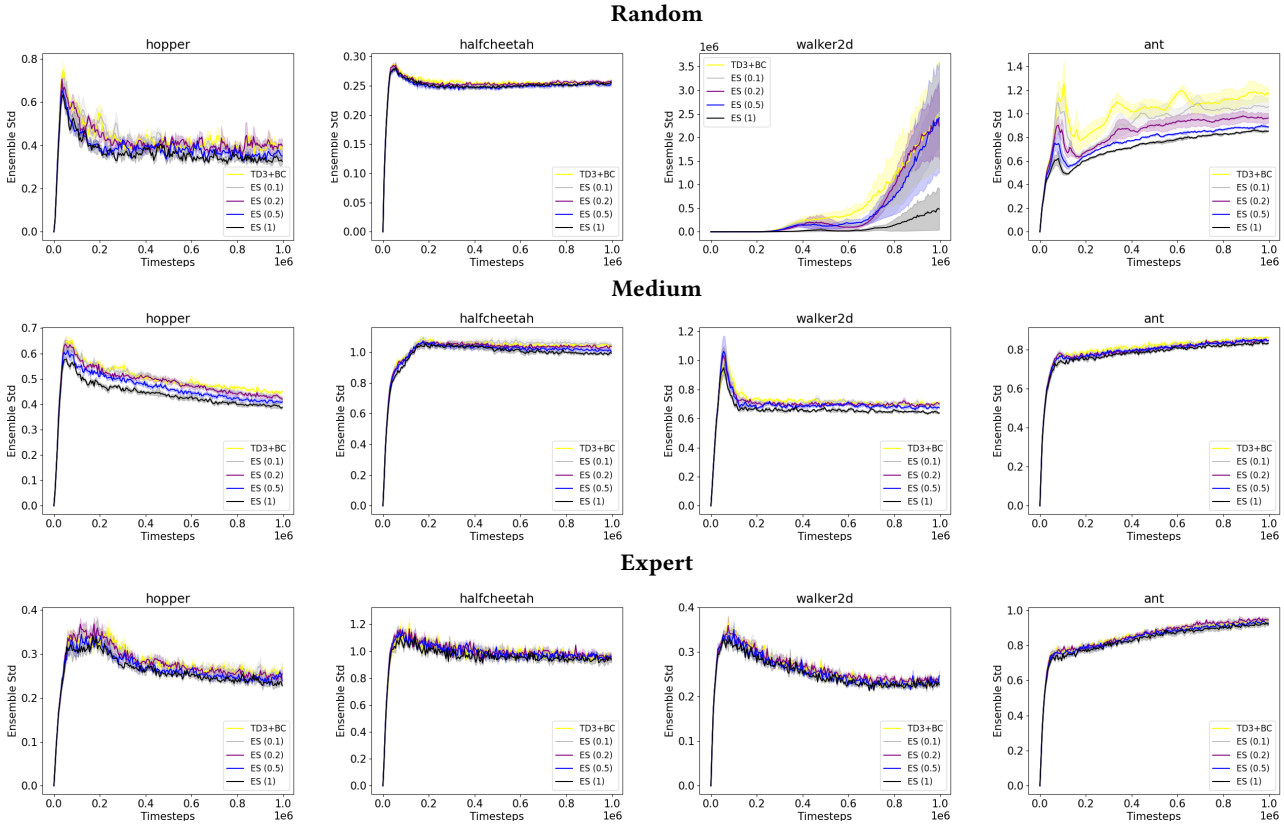
5.1 Critic Disagreement Dynamics

Understanding how ES modulates ensemble behavior requires examining the evolution of critic disagreement throughout training. We quantify this through the standard deviation of Q-value estimates across ensemble critics, which serves as a direct indicator of uncertainty and model confidence in value predictions.

Figure 1 presents critic disagreement trajectories for TD3+BC baseline and ES variants across four D4RL locomotion tasks under three data quality conditions. This comprehensive visualization reveals how ES's uncertainty-aware regularization adapts to both algorithmic context and dataset characteristics, providing mechanistic insight into the observed performance improvements.

5.1.1 Systematic Disagreement Reduction Across Data Regimes. **Consistent Uncertainty Mitigation.** Our analysis reveals that ES systematically reduces ensemble standard deviation compared to TD3+BC across all experimental conditions—**random**, **medium**, and **expert** data regimes. These reductions are most pronounced

Figure 1: Ensemble Std over 1M gradient steps comparing TD3+BC and ES across MuJoCo tasks. The figure is organized in three blocks corresponding to dataset quality: the first block shows results for the *Random* setting, the second for *Medium*, and the third for *Expert*. Each task (*Hopper*, *Cheetah*, *Walker*, *Ant*) appears once per block. Reported values represent the mean ± standard error over 5 seeds.



in challenging scenarios where baseline methods exhibit high uncertainty.

Quality-Dependent Regularization Effects. The magnitude of disagreement reduction scales with data quality challenges. In the **random** regime, characterized by high-variance trajectories and prevalent out-of-distribution actions, ES achieves the most substantial uncertainty reduction. This strong effect directly correlates with the largest performance improvements observed in our experimental results, validating ES's design principle of applying stronger regularization when critics encounter unreliable value signals.

The *Cheetah* environment presents a notable exception, where both ES and TD3+BC maintain uniformly low disagreement levels. This behavior likely reflects the task's inherent simplicity and the tendency for all methods to converge toward similar suboptimal policies, highlighting a boundary condition where additional regularization provides minimal marginal benefit.

Progressive Stabilization with Improved Data Quality. In *medium* and *expert* regimes, baseline ensemble disagreement naturally decreases due to improved data consistency and reduced distributional shift. ES maintains its uncertainty reduction advantage, though with diminished absolute differences. This adaptive behavior demonstrates ES's key strength: providing strong regularization when critics disagree substantially while preserving learning signal when consensus emerges.

Mechanistic Validation. The observed disagreement patterns provide direct empirical support for ES's theoretical foundation. By systematically penalizing value estimates in proportion to critic disagreement, ES addresses overestimation bias precisely in regions of highest uncertainty–where individual critics are most likely to produce unreliable estimates. The consistent relationship between disagreement reduction and performance improvement across diverse conditions validates this uncertainty-aware approach to conservative value learning.

Key Insights. ES demonstrates robust capacity for uncertainty mitigation across all data quality regimes, with effectiveness scaling directly with the level of noise and out-of-distribution samples in the dataset. This analysis confirms ES's utility as an adaptive uncertainty-aware regularizer and supports the hypothesis that disagreement-based penalties provide a principled mechanism for addressing overestimation bias in multi-critic architectures.

Table 3: Offline results showing the computation time (in minutes) when using our method (ES) with the TD3+BC algorithm on three D4RL tasks. Each row compares the time taken by ES to the original TD3+BC, with the best time highlighted in bold. The last row shows the average extra computation required by ES. We also include results for CQL (TD3+BC) and SAC+BC for comparison.

Env.	TD3+BC	SAC+BC	CQL (TD3+BC)	ES ($\alpha = 0.1$)	ES ($\alpha = 0.2$)	ES ($\alpha = 0.5$)	ES ($\alpha = 1$)
Random							
Hopper	162.3 ± 20.8	154.7 ± 5.3	183.2 ± 9.3	206.0 ± 14.5	146.0 ± 5.5	147.1 ± 5.2	**143.1 ± 3.0**
Cheetah	174.7 ± 5.5	187.6 ± 5.3	210.8 ± 6.5	262.3 ± 10.9	175.9 ± 3.7	174.7 ± 7.7	179.0 ± 6.0
Walker	145.8 ± 6.6	152.5 ± 5.3	184.3 ± 3.2	**142.4 ± 10.6**	147.4 ± 2.3	146.2 ± 5.5	149.0 ± 3.7
Ant	187.0 ± 5.3	195.3 ± 4.9	222.8 ± 6.1	186.7 ± 4.9	186.1 ± 4.8	184.0 ± 3.5	**176.7 ± 6.1**
Improvement	–	2.9%	19.8%	18.6%	-2.2%	-2.7%	**-3.2%**
Medium							
Hopper	165.4 ± 7.8	165.1 ± 6.2	199.4 ± 4.0	167.1 ± 5.8	160.4 ± 4.9	162.2 ± 4.0	**159.9 ± 3.4**
Cheetah	180.4 ± 12.0	186.5 ± 3.5	212.7 ± 8.3	180.7 ± 17.1	176.4 ± 7.2	174.0 ± 4.5	**176.7 ± 5.0**
Walker	164.8 ± 4.6	172.2 ± 3.7	202.2 ± 5.5	169.5 ± 2.6	164.3 ± 6.0	165.4 ± 5.2	**163.0 ± 7.7**
Ant	**182.8 ± 5.2**	195.4 ± 5.5	207.7 ± 10.4	183.8 ± 6.5	186.0 ± 8.8	183.4 ± 4.7	190.1 ± 6.5
Improvement	–	3.6%	18.7%	1.1%	-1.0%	**-1.2%**	-0.6%
Expert							
Hopper	160.9 ± 7.4	165.9 ± 3.7	196.9 ± 10.5	164.9 ± 8.2	172.9 ± 40.1	167.4 ± 21.4	**152.0 ± 3.6**
Cheetah	**174.7 ± 8.9**	196.5 ± 23.1	210.2 ± 3.6	**181.5 ± 6.6**	177.1 ± 3.8	176.0 ± 4.3	177.2 ± 4.6
Walker	176.8 ± 6.2	191.2 ± 5.3	214.7 ± 7.0	174.0 ± 5.8	182.0 ± 6.2	**173.8 ± 3.4**	175.0 ± 2.8
Ant	169.4 ± 5.4	183.5 ± 3.6	214.5 ± 17.4	168.9 ± 5.4	167.7 ± 2.7	168.7 ± 3.6	**164.4 ± 4.3**
Improvement	–	8%	22.7%	1.1%	2.7%	0.7%	**-2.0%**

5.2 Computational Efficiency Analysis

Table 3 presents offline training times (in minutes) for TD3+BC with and without ES across three D4RL data regimes, alongside SAC+BC and CQL baselines for comparison.

Minimal Runtime Overhead. ES demonstrates excellent computational efficiency across all experimental conditions. With only one minor exception (ES $\alpha = 0.1$ on random data), all other ES configurations ($\alpha = 0.2, 0.5, 1.0$) exhibit *near-zero or negative overhead* relative to TD3+BC. Notably, ES with $\alpha = 1.0$ reduces training time by 3.2% in the random setting and 2.0% in the expert setting, demonstrating that the regularization mechanism can actually accelerate convergence.

This efficiency holds consistently across data quality regimes: in the challenging random regime (high-noise, out-of-distribution data), ES maintains negative average overhead for most α values; in medium and expert regimes, ES continues to match or improve baseline efficiency, with $\alpha = 1.0$ reducing wall-clock time in 3/4 environments across both settings.

Comparison with Alternative Methods. In stark contrast, CQL introduces substantial computational burden with an average +20.4% increase in training time across all settings, confirming that its stronger regularization comes at significant cost. SAC+BC also incurs additional overhead, particularly pronounced in the expert regime where behavioral complexity is highest.

Practical Implications. ES achieves its performance improvements while maintaining high computational efficiency across hyperparameters, environments, and data quality conditions. This combination of effectiveness and efficiency makes ES particularly suitable for resource-constrained applications and large-scale deployments where computational overhead directly impacts feasibility.

6 CONCLUSION AND FUTURE WORK

We introduced Ensemble Std (ES), a lightweight regularization method that addresses overestimation bias in multi-critic deep reinforcement learning by penalizing value estimates proportional to critic disagreement. This uncertainty-aware approach consistently reduces ensemble disagreement across diverse experimental conditions, validating its effectiveness in mitigating overestimation.

Our comprehensive evaluation spanning online MuJoCo environments and offline D4RL datasets demonstrates that ES consistently outperforms or matches multiple strong baselines across varying task complexities and data quality regimes. Critically, ES achieves these performance gains with *near-zero computational overhead*. Furthermore, ES's practical advantages include seamless integration into existing multi-critic architectures through a simple regularization term requiring no architectural modifications.

For future work, we consider the development of adaptive scheduling to be promising extension of out approach. We plan to achieve this goal through meta-learning, automatically tuning regularization strength based on environment dynamics patterns to enhance generalization across diverse tasks.

REFERENCES

[1] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. 2020. An optimistic perspective on offline reinforcement learning. In *International conference on machine learning*. PMLR, 104–114.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. 2020. Learning to continually learn. In *ECAI 2020*. IOS Press, 992–1001.

[4] Richard Bellman. 1957. A Markovian Decision Process. *Journal of Mathematics and Mechanics* 6, 5 (1957), 679–684.

[5] G Brockman. 2016. OpenAI Gym. *arXiv preprint arXiv:1606.01540* (2016).

[6] Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. 2021. Randomized Ensembled Double Q-Learning: Learning Fast Without a Model. In *International Conference on Learning Representations*. https://openreview.net/forum?id=AY8zfZm0tDd

[7] Kamil Ciosek and Shimon Whiteson. 2020. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research* 21, 52 (2020), 1–51.

[8] Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. 2022. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *Deep Reinforcement Learning Workshop NeurIPS 2022*.

[9] Mónika Farsang and Luca Szegletes. 2021. Decaying Clipping Range in Proximal Policy Optimization. *arXiv preprint arXiv:2102.10456* (2021).

[10] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219* (2020).

[11] Scott Fujimoto, Wei-Di Chang, Edward J Smith, Shixiang Shane Gu, Doina Precup, and David Meger. 2023. For SALE: State-Action Representation Learning for Deep Reinforcement Learning. *arXiv preprint arXiv:2306.02451* (2023).

[12] Scott Fujimoto and Shixiang Shane Gu. 2021. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems* 34 (2021), 20132–20145.

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

[14] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*. PMLR, 2052–2062.

[15] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, 1861–1870.

[16] Lei Huang, Li Liu, Fan Zhu, Diwen Wan, Zehuan Yuan, Bo Li, and Ling Shao. 2020. Controllable orthogonalization in training dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6429–6438.

[17] Wei Huang, Weitao Du, and Richard Yi Da Xu. 2020. On the neural tangent kernel of deep networks with orthogonal initialization. *arXiv preprint arXiv:2004.05867* (2020).

[18] Steven James, George Konidaris, and Benjamin Rosman. 2017. An analysis of monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[19] Scott M Jordan, Samuel Neumann, James E Kostas, Adam White, and Philip S Thomas. 2024. The Cliff of Overcommitment with Policy Gradient Step Sizes. In *Reinforcement Learning Conference*.

[20] Alex Kendall et al. 2019. Learning to drive in a day. In *International Conference on Robotics and Automation (ICRA)*.

[21] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 1179–1191.

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

[23] Hojoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian, Peter R Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. 2024. Simba: Simplicity bias for scaling up parameters in deep reinforcement learning. *arXiv preprint arXiv:2410.09754* (2024).

[24] Sergey Levine et al. 2016. End-to-end training of deep visuomotor policies. In *JMLR*, Vol. 17. 1334–1373.

[25] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv preprint arXiv:2005.01643* (2020).

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

[27] Zhixuan Lin, Pierluca D'Oro, Evgenii Nikishin, and Aaron Courville. 2024. The Curse of Diversity in Ensemble-Based Exploration. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR 2024.

[28] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[29] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. 2022. Mildly conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 1711–1724.

[30] Michael C. McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *In Psychology of Learning and Motivation* 24 (1989), 109–165.

[31] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[32] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. 2016. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems* 29 (2016).

[33] Michal Nauman, Mateusz Ostaszewski, Krzysztof Jankowski, Piotr Miłoś, and Marek Cygan. 2024. Bigger, regularized, optimistic: scaling for compute and sample efficient continuous control. *Advances in neural information processing systems* 37 (2024), 113038–113071.

[34] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep exploration via bootstrapped DQN. *Advances in Neural Information Processing Systems (NeurIPS)* (2016), 4026–4034.

[35] Dean A Pomerleau. 1989. ALVINN: An autonomous land vehicle in a neural network. In *Proceedings of the 1st International Conference on Neural Information Processing Systems (NIPS)*. MIT Press, 305–313.

[36] Doina Precup, Richard S Sutton, and Satinder Singh. 2000. Eligibility traces for off-policy policy evaluation. (2000).

[37] Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (1st ed.). John Wiley & Sons, Inc., USA.

[38] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120* (2013).

[39] John Schulman, Sergey Levine, Pieter Abbeel, Michael I Jordan, and Philipp Moritz. 2015. Trust Region Policy Optimization. *International Conference on Machine Learning* (2015), 1889–1897.

[40] Younggyo Seo and Pieter Abbeel. 2024. Reinforcement learning with action sequence for data-efficient robot learning. (2024).

[41] Tali Sharot. 2011. The optimism bias. *Current biology* 21, 23 (2011), R941–R945.

[42] D. Silver et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.

[43] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[44] Richard Stuart Sutton. 1984. *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst.

[45] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[46] Sebastian Thrun and Anton Schwartz. 1993. Issues in using function approximation for reinforcement learning. *Advances in neural information processing systems* 5 (1993).

[47] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 5026–5033.

[48] Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double Q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (2016).

[49] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[50] Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[51] Junghyuk Yeom, Yonghyeon Jo, Jeongmo Kim, Sanghyeon Lee, and Seungyul Han. 2024. Exclusively penalized q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 113405–113435.

[52] Tingting Zhao, Hirotaka Hachiya, Gang Niu, and Masashi Sugiyama. 2011. Analysis and improvement of policy gradient estimation. *Advances in Neural Information Processing Systems* 24 (2011).