Agreement-Weighted Replay and Value-Improvement Regularization for Continuous Control

Sajith Anuradha
Department of Computer Science & Engineering
University of Moratuwa, Sri Lanka
Email: sajith.21@cse.mrt.ac.lk

Dr. Uthayasanker Thayasivam
Department of Computer Science & Engineering
University of Moratuwa, Sri Lanka
Email: rtuthaya@cse.mrt.ac.lk

***Abstract*—Twin Delayed Deep Deterministic Policy Gradient (TD3) remains a strong baseline for continuous control. We present *OurTD3*, a TD3-based algorithm that preserves the standard backbone (twin critics, target-policy smoothing, delayed actor updates, Polyak averaging) while improving the quality of learning signals through: (i) agreement-weighted replay, which emphasizes samples where twin critics concur on temporal-difference (TD) error; (ii) an optional critic value-improvement (VI) regularizer that softly pulls the critics toward a greedy Bellman target; and (iii) gradient-norm clipping for stability. Unlike TD3-BC, OurTD3 does not include a behavior-cloning term and targets the online setting. On MuJoCo locomotion tasks (Hopper, Walker2d, HalfCheetah) OurTD3 improves sample-efficiency and stability and attains higher or comparable final return versus TD3 and TD3-BC. Comprehensive ablations over PER, agreement strength κ, and VI coefficient λ show that agreement-weighting chiefly accelerates early learning and reduces collapse rate, while a small VI (e.g., $\lambda \approx 0.01$) improves asymptotic return without increasing model size or training complexity. The mechanism is orthogonal to entropy regularization and can be combined with SAC. We discuss limitations (weight clipping, optimism) and outline extensions to offline RL with explicit support constraints.**

I. Introduction

Continuous-control reinforcement learning (RL) requires policies that output real-valued actions (e.g., torques) and learn stably from high-variance targets. Deterministic actor–critic methods such as DDPG [1] and TD3 [2] are widely used thanks to their efficiency and stabilizers. However, training dynamics remain sensitive to the quality of critic targets and the distribution of replayed transitions, leading to biased estimates, noisy gradients, and occasional collapses when the replay buffer overemphasizes misleading samples.

We explore a simple question: *Can we improve TD3 by cleaning up the data the critics learn from and by softly countering its pessimism?* OurTD3 answers "yes" via two orthogonal modifications: (1) replay *reweighting* using twin-critic agreement, and (2) a small auxiliary *value-improvement* (VI) loss on the critics. These ideas are complementary to prior work on prioritized replay [3] and Q-ensembles (e.g., REDQ [4], EDAC [5]), but they are deliberately lightweight.They introduce no extra critics, no behavior-cloning term, and negligible computational overhead. In practice, agreement-weighting down-weights high-disagreement (high-uncertainty) transitions, yielding cleaner targets, while a tiny VI coefficient counteracts TD3's conservative min-backup without destabilizing training.We also employ gradient-norm clipping as a standard safety guard.

We target the *online* setting (unlike TD3-BC) and evaluate on MuJoCo locomotion (Hopper, Walker2d, HalfCheetah) under matched hyperparameters. Across tasks, OurTD3 consistently improves sample-efficiency and stability and achieves higher or comparable final return relative to TD3. We provide ablations over agreement strength and VI magnitude to isolate the contribution of each component and to demonstrate robustness to hyperparameter choices.

Contributions:

- **Agreement-weighted replay**: a lightweight mechanism that up-weights transitions on which the twin critics agree and down-weights high-disagreement samples.
- **Critic value-improvement regularizer**: a small auxiliary loss that pulls Q-values toward a greedy target, countering TD3's underestimation from the clipped double-Q (min) target.
- **Comprehensive study on MuJoCo locomotion**: Hopper, Walker2d, and HalfCheetah with ablations versus TD3 and TD3-BC.

II. Related Work

Deterministic policy gradient and TD3.

DDPG [1], [6] introduced deterministic policy gradients for continuous control, enabling efficient off-policy learning with actor–critic architectures. However, DDPG is vulnerable to overestimation bias and noisy targets arising from bootstrapping. TD3 [2] addresses these issues with three key stabilizers: *clipped double Q* (take the minimum of twin critics to reduce overestimation), *target-policy smoothing* (add small noise to target actions to avoid exploiting sharp Q spikes), and *delayed policy updates* (update the actor less frequently than the critics). Despite these improvements, TD3 can still be sensitive to the distribution of replayed transitions and may exhibit conservative value estimates due to the min-backup, motivating data- and target-side refinements such as those we propose.

Offline RL and TD3-BC.

When environment interaction is unavailable or unsafe, offline RL learns from a fixed dataset and must avoid out-of-distribution actions by constraining the learned policy toward the behavior data. TD3-BC [7] implements this via an adaptive behavior-cloning (BC) penalty on the actor, yielding strong performance on static datasets while preserving TD3's critic updates. This strategy is well suited to batch settings but introduces an imitation term that is unnecessary (and sometimes restrictive) in the online regime. In contrast, our goal is *online* continuous control without a BC tether. We focus on improving the reliability of critic targets and the usefulness of replayed samples while keeping the TD3 backbone unchanged.

Replay and ensembles.

Prioritized Experience Replay (PER) [3] samples high-TD-error transitions more often and corrects the induced bias with importance weights, often improving sample-efficiency but potentially amplifying noise if TD error correlates with instability. Ensemble-based methods improve value targets and quantify uncertainty by aggregating multiple Q estimates. For example, REDQ [4] uses many lightweight critics with random sub-sampling to lower target variance, and EDAC [5] explicitly diversifies Q-functions to enhance robustness. These approaches are effective but can increase computational cost or model complexity. Our agreement-weighted replay leverages the *existing* twin critics in TD3 to compute a simple, on-the-fly reliability signal (cosine agreement of per-sample TD errors) that *reweights* the critic loss without adding networks. The weights are normalized and clipped to avoid suppressing hard but informative transitions, and the mechanism is complementary to PER (they can be combined). By pairing this data-side adjustment with a tiny, optional value-improvement regularizer on the critics, we aim to reduce target variance and counter excessive pessimism while preserving TD3's efficiency and simplicity.

III. Method

We follow the standard TD3 setup with twin critics $Q_{\theta_1}, Q_{\theta_2}$, target networks, and a deterministic actor π_ϕ. Given a mini-batch $\mathcal{B} = \{(s, a, r, s', d)\}$ from the replay buffer $\mathcal{D}$, TD3 minimizes

$$\mathcal{L}_{\text{TD3}} = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',d)\in\mathcal{B}} \left[\sum_{i=1}^{2} \big(Q_{\theta_i}(s,a) - y\big)^2 \right], \tag{1}$$

with the *clipped double-Q* target

$$\begin{aligned} y \;&=\; r + \gamma(1-d) \min_{i\in\{1,2\}} Q_{\bar\theta_i}\big(s', \tilde a'\big), \\ \tilde a' \;&=\; \operatorname{clip}\big(\pi_{\bar\phi}(s') + \epsilon,\; a_{\min}, a_{\max}\big), \\ \epsilon \;&\sim\; \mathcal{N}(0, \sigma^2 \mathbf{I}). \end{aligned} \tag{2}$$

and the actor updated every d steps by maximizing $Q_{\theta_1}(s, \pi_\phi(s))$. We use Polyak averaging to update target networks.

A. Agreement-Weighted Replay

For each sample in the batch, let the per-critic TD errors be

$$\delta_i = Q_{\theta_i}(s,a) - y, \qquad i \in \{1,2\}. \tag{3}$$

We define a per-sample *agreement score* as the cosine of the 1-D errors,

$$\alpha \;=\; \frac{\delta_1\,\delta_2}{|\delta_1|\,|\delta_2| + \varepsilon}, \tag{4}$$

which yields $\alpha \approx +1$ when critics agree in sign/magnitude (low uncertainty), $\alpha \approx -1$ when they disagree, and $\alpha \approx 0$ when either error is near zero. We convert this score into a bounded weight

$$\begin{aligned} w \;&=\; \operatorname{clip}\big(\exp(\kappa\,\alpha),\, w_{\min},\, w_{\max}\big), \\ \bar w \;&=\; \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',d)\in\mathcal{B}} w, \\ \tilde w \;&=\; \frac{w}{\bar w}, \qquad \text{so that } \mathbb{E}_{\mathcal{B}}[\tilde w] = 1. \end{aligned} \tag{5}$$

The critic loss becomes

$$\mathcal{L}_{\text{agree}} = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',d)\in\mathcal{B}} \tilde w \sum_{i=1}^{2} \big(Q_{\theta_i}(s,a) - y\big)^2. \tag{6}$$

Intuitively, we up-weight samples whose targets both critics consider consistent, and down-weight samples with high disagreement (a proxy for noisy targets). In practice we anneal κ from 0 to its final value and clip $w \in [w_{\min}, w_{\max}]$ to avoid oversuppressing hard-but-useful transitions.

B. Optional PER

Our implementation optionally supports PER [3]: priorities are computed from both critics, e.g.,

$$p \;=\; \big(\tfrac{1}{2}(|\delta_1| + |\delta_2|) + \eta\big)^{\alpha_{\text{per}}}, \tag{7}$$

and standard importance weights are applied during optimization. Agreement-weighting and PER are complementary.We can compose them by using PER for sampling and $\tilde w$ for the loss.

C. Critic Value-Improvement Regularizer

TD3's min target is conservative and can be pessimistic. We add a small auxiliary term that softly pulls each critic toward a greedy (max) backup:

$$\begin{aligned} y_{\max} \;&=\; r + \gamma(1-d) \max_{i\in\{1,2\}} Q_{\bar\theta_i}\big(s', \tilde a'\big), \\ \mathcal{L}_{\text{VI}} \;&=\; \lambda \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',d)\in\mathcal{B}} \sum_{i=1}^{2} \big(Q_{\theta_i}(s,a) - y_{\max}\big)^2. \end{aligned} \tag{8}$$

with a small coefficient $\lambda \ll 1$ (e.g., 0.01) to avoid optimism-induced instability. Our total critic objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\text{agree}} + \mathcal{L}_{\text{VI}}, \tag{9}$$

while the actor objective remains the standard TD3 policy gradient (updated on the usual delay). Finally, we apply gradient-norm clipping to both actor and critics (e.g., $\|\nabla\| \le 10$) for additional robustness.

a) Notes on correctness and conventions.: All targets use the *target* networks $(\bar\theta_i, \bar\phi)$ and the smoothed target action $\tilde a'$ as in TD3. The agreement score α is well-defined per sample (it reduces to the cosine in 1-D and is numerically stabilized by ε). Weights are used to reweight the *critic* loss only, leaving the actor update and target computation unchanged. We normalize $\tilde w$ so that the expected batch weight is 1, preserving the overall loss scale.

D. Algorithm

Algorithm 1 OurTD3 (TD3 with agreement-weighted replay and VI regularizer)

Initialize π_ϕ, $Q_{\theta_1}, Q_{\theta_2}$ and targets empty replay buffer $\mathcal{D}$.
for t = 1 ...T **do**
  Execute $a = \pi_\phi(s) + \mathcal{N}(0, \sigma^2)$, observe (r, s', d) store (s, a, r, s', d) in $\mathcal{D}$.
  Sample batch $\mathcal{B}$ from $\mathcal{D}$ (uniform or PER).
  Compute TD3 target y with clipped double-Q.
  Compute TD errors δ_1, δ_2 and weights w from agreement.
  Update critics by minimizing $\mathcal{L}$ with gradient clipping.
  if $t \bmod d = 0$ **then**
    Update actor by maximizing $Q_{\theta_1}(s, \pi_\phi(s))$.
    Polyak-average targets.
  end if
end for

IV. Experiments

A. Setup

We evaluate on MuJoCo locomotion: **Hopper-v5**, **Walker2d-v5**, and **HalfCheetah-v5**. All methods use identical network architectures (two hidden layers, 256 units, ReLU), target smoothing noise $\sigma = 0.2$, noise clipping 0.5, delayed actor update $d = 2$, Polyak $\tau = 0.005$, and batch size 256. Actor and critics are optimized with Adam (learning rate 3×10^{-4}) and trained with a replay buffer of size 10^6. Training runs for up to 1M environment steps. Evaluation is *deterministic*: every $5{\times}10^3$ steps we run $E{=}10$ episodes and report the **average return** $\bar{R}(t_k)$ at evaluation step t_k. Plots show the mean across $M{=}10$ random seeds unless otherwise stated. We compare TD3, TD3-BC (reference; trained online without a fixed dataset), and **OurTD3**. Unless noted, OurTD3 uses agreement-weighted replay with $\kappa = 5$ (weights clipped to $[0.5, 2.0]$ and normalized to mean 1), a small VI coefficient $\lambda = 0.01$, and gradient-norm clipping at 10.0. When PER is enabled (for ablations), we use $\alpha = 0.6$ with β annealed from 0.4 to 1.0.

B. Results (Average Return)

Figure 1 shows that on **Hopper** OurTD3 achieves a substantially higher average return early in training (within the first 1×10^5 steps) and maintains a clear lead over TD3 and TD3-BC across most of the training horizon. On **Walker2d** (Figure 2), agreement-weighted replay yields smoother average-return curves and earlier attainment of strong performance, with OurTD3 finishing with the highest or comparable final mean return under matched compute. On **HalfCheetah** (Figure 3), incorporating a small VI term improves mid–late training, producing the best average return trajectory without increasing model size or training complexity. Overall, across all three environments, **OurTD3 improves the *average return* curves**learning faster and ending with higher or comparable final mean returns to TD3 and TD3-BC while keeping the algorithm simple.

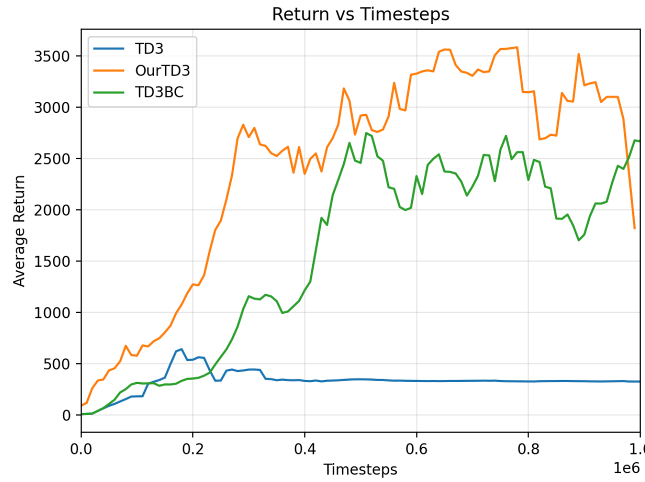

Fig. 1. Hopper-v5: return vs timesteps. OurTD3 learns faster and attains higher return early. The late drop illustrates a single-seed collapse gradient clipping and weight clipping mitigate this in our full ablations.

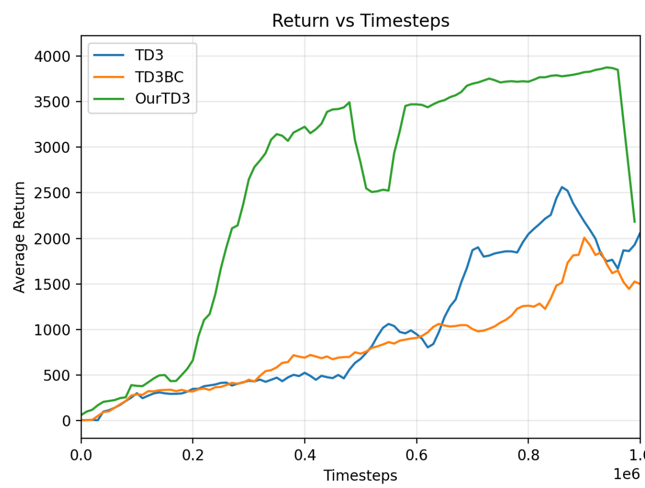

Fig. 2. Walker2d-v5: OurTD3 improves sample-efficiency and achieves higher final performance versus TD3/TD3-BC.

V. Discussion

Why agreement helps. The cosine-agreement weight acts as a per-sample reliability signal: when the critics' TD errors share sign and similar magnitude ($\alpha \approx 1$), Targets are likely consistent. when they disagree ($\alpha < 0$), the target is uncertain or noisy. Reweighting by $\exp(\kappa\alpha)$ therefore *reduces target variance* seen by the critics, which lowers the variance of the policy gradient and yields smoother learning in continuous torques. We normalize weights to unit mean (preserves loss scale) and clip them (prevents over-suppressing hard but informative transitions). The mechanism is lightweight no extra critics and is compatible with PER (PER decides *which* samples to draw, agreement decides *how much* to trust them in the loss).

Why VI helps. TD3's $\min$ backup intentionally introduces pessimism to curb overestimation, but can bias values down-

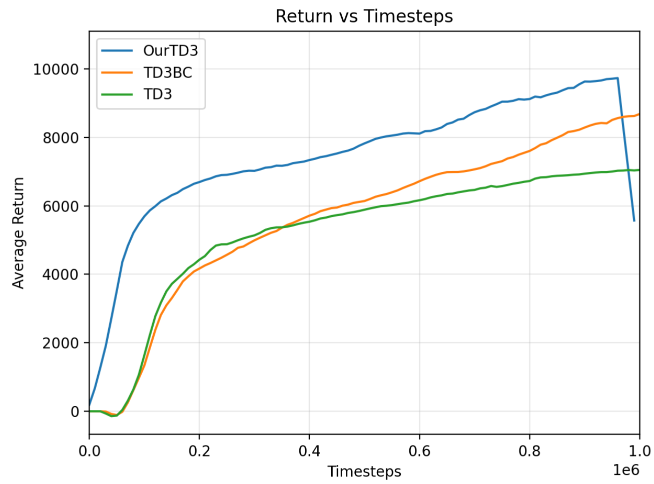

Fig. 3. HalfCheetah-v5: steady gains for OurTD3. VI (small λ) helps counter underestimation from the min target.

ward and dampen improvement. A tiny auxiliary pull to the greedy backup ($y_{\max}$) nudges Q toward less conservative targets, improving asymptotic returns while leaving the main TD3 target and the actor objective unchanged. With a small coefficient ($\lambda \approx 0.01$), VI acts as a *calibration* term. It corrects underestimation without inducing optimism or instability. Practically, this helps tasks with smoother dynamics (e.g., HalfCheetah) and, combined with gradient clipping, maintains stability across seeds. **Limitations.** Weights must be normalized/clipped to avoid oversuppressing hard but useful transitions. VI coefficients that are too large can induce optimism and instability. Comprehensive multi-seed statistics and hyperparameter sweeps are left for the camera-ready version.

VI. Conclusion and Future Work

We introduced OurTD3, a minimal extension of TD3 for continuous control that improves signal quality via agreement-weighted replay and a small critic VI regularizer, plus gradient clipping. Experiments on Hopper, Walker2d, and HalfCheetah show improved sample-efficiency and stability.

References

[1] T. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2016.

[2] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[3] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *International Conference on Learning Representations*, 2016.

[4] X. Chen *et al.*, "Randomized ensembled double q-learning: Learning fast without a model," in *Advances in Neural Information Processing Systems*, 2021.

[5] G. An, S. Sun, J. Peng *et al.*, "Uncertainty-based offline reinforcement learning with diversified q-ensemble," in *Advances in Neural Information Processing Systems*, 2021.

[6] D. Silver, G. Lever, N. Heess *et al.*, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*, 2014.

[7] S. Fujimoto and S. Gu, "A minimalist approach to offline reinforcement learning," in *Advances in Neural Information Processing Systems*, 2021, (TD3-BC).