A Case for Validation Buffer in Pessimistic Actor-Critic

Michał Nauman[1] , **Mateusz Ostaszewski**[2] , **Marek Cygan**[1,3]
[1]University of Warsaw
[2]Warsaw University of Technology
[3]Nomagic
nauman.mic@gmail.com

Abstract

In this paper, we investigate the issue of error accumulation in critic networks updated via pessimistic temporal difference objectives. We show that the critic approximation error can be approximated via a recursive fixed-point model similar to that of the Bellman value. We use such a recursive definition to retrieve the conditions under which the pessimistic critic is unbiased. Building on these insights, we propose Validation Pessimism Learning (VPL) algorithm. VPL uses a small validation buffer to adjust the levels of pessimism throughout the agent's training, with the pessimism set such that the approximation error of the critic targets is minimized. We investigate the proposed approach on a variety of locomotion and manipulation tasks and report improvements in sample efficiency and performance.

1 Introduction

Approximation errors, although ubiquitous in machine learning, are particularly exaggerated in the context of value-based Reinforcement Learning (RL). Such exaggeration stems from Temporal Difference (TD) in which the critic is supervised via value estimate calculated at a different state [Silver *et al.*, 2014], [Mnih *et al.*, 2015].

Inaccuracies in this estimate lead to propagated errors in state-action updates, and the use of maximization in value estimation inherently promotes overestimation. Addressing such overestimation has proven to be an effective strategy in discrete and continuous action environments [Hasselt, 2010], [Van Hasselt *et al.*, 2016], [Hessel *et al.*, 2018], [Haarnoja *et al.*, 2018]. Clipped Double Q-Learning (CDQL), a common solution to overestimation in continuous action actor-critic algorithms aims to mitigate overestimation by balancing errors against a pessimistic lower bound value approximation [Fujimoto *et al.*, 2018]. However, challenges arise if the lower bound is insufficiently pessimistic, leading to continued overestimation, or overly pessimistic, causing underestimation [Cetin and Celiktutan, 2023]. The latter, though less recognized, can significantly reduce sample efficiency and degrade actor-critic agents' performance in both low and high replay ratio settings which we show in Figure 1.

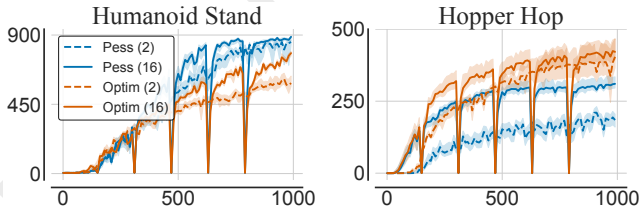

Figure 1: Pessimism adjustment can yield performance benefits exceeding those of increased replay ratio and full-parameter resets. The pessimistic algorithm dominates Humanoid, whereas the optimistic algorithm dominates Hopper. 10 seeds and 95% CI.

In this paper, we investigate the relationship between pessimism in Q-value approximation and error accumulation in critic networks. We start by characterization of existing strategies for pessimism adjustment. Furthermore, we analyze the pessimistic critic approximation error and show that such error can be represented recursively forming a fixed-point model, akin to values. This recursive representation helps us highlight the bias inherent in pessimistic actor-critic algorithms, examine their convergence dynamics, and identify the conditions under which pessimistic critics can achieve zero error. Building on these insights, we propose the Validation Pessimism Learning (VPL) algorithm. VPL employs a small validation replay buffer to adjust the pessimism levels online, aiming to minimize the approximation error of critic targets while preventing overfitting to accumulated experience. We evaluate VPL against existing pessimism adjustment methods on DeepMind control [Tassa *et al.*, 2018] and MetaWorld [Yu *et al.*, 2020]. Our findings demonstrate that VPL not only achieves performance improvements but also exhibits less sensitivity to hyperparameters compared to the baseline algorithms. We list our contributions below:

- We show that critic error can be defined recursively through a fixed-point model. We demonstrate that pessimistic TD learning, a method often used in RL, converges to the true value under strict conditions.
- We present an empirical analysis showing that the performance loss associated with not including every transition in the replay buffer diminishes as training progresses. This observation challenges the traditional belief that every transition must be used in value learning for sample-efficient RL and builds a case for employing

a validation buffer in an online RL setting.

- We propose VPL, an algorithm that uses a small validation buffer for online adjustment of pessimism associated with lower bound Q-value approximation. We test the effectiveness of VPL and other pessimism adjustment strategies in low and high replay regimes. We show that VPL offers performance improvements across a variety of locomotion and manipulation tasks.

2 Background

2.1 Maximum Entropy Reinforcement Learning

We consider an infinite-horizon Markov Decision Process (MDP) [Puterman, 2014], represented by the tuple (S, A, r, p_0, γ), where S and A are continuous state and action spaces, $r_{s,a}$ is the reward, $p_0(s)$ is the initial state distribution, and $\gamma \in (0, 1]$ is the discount factor. The policy $\pi(a|s)$ is a distribution of actions given states. The goal of Maximum Entropy Reinforcement Learning (MaxEnt RL) [Haarnoja *et al.*, 2017] is to maximize the expected cumulative discounted return augmented with an entropy term:

$$\pi^* = \arg\max_{\pi} \mathop{\mathrm{E}}_{p_0,\pi} \sum_{t=0}^{\infty} \gamma^t \big(r_{s_t,a_t} + \alpha \mathcal{H}(\pi(\cdot|s_t))\big), \quad (1)$$

where α controls the balance between reward and entropy [Haarnoja *et al.*, 2018]. The soft Q-value is defined as:

$$Q^\pi(s,a) = r_{s,a} + \gamma \mathop{\mathrm{E}}_{s' \sim p} \big[V^\pi(s')\big], \quad (2)$$

and the soft value is given by:

$$V^\pi(s) = \mathop{\mathrm{E}}_{a \sim \pi} \big[Q^\pi(s,a) - \alpha \log \pi(a|s)\big]. \quad (3)$$

MaxEnt RL algorithms, such as Soft Actor-Critic (SAC), parameterize the policy (actor) π_θ and the Q-value (critic) Q_ϕ, which are optimized iteratively using objectives derived from policy iteration [Haarnoja *et al.*, 2018]. The critic ensemble is often used to address Q-value overestimation via Clipped Double Q-Learning (CDQL) [Fujimoto *et al.*, 2018]. In CDQL, the value lower bound is approximated as:

$$V^{lb}_\phi(s) \approx Q^{lb}_\phi(s,a) - \alpha \log \pi_\theta(a|s), \quad a \sim \pi_\theta, \quad (4)$$

where $Q^{lb}_\phi(s,a) = \min(Q^1_\phi(s,a), Q^2_\phi(s,a))$.

2.2 Pessimism Adjustment

Building upon CDQL, recent works propose parameterizing the lower bound of Q-values as:

$$Q^{lb}_\phi(s,a) = Q^\mu_\phi(s,a) - \beta Q^\sigma_\phi(s,a), \quad (5)$$

where Q^μ_ϕ and Q^σ_ϕ represent the mean and standard deviation of the critic ensemble, respectively [Ciosek *et al.*, 2019], [Nauman and Cygan, 2023]. This formulation allows the adjustment of the pessimism level β, controlling the influence of critic disagreement. Algorithms such as Generalized Pessimism Learning (GPL) [Cetin and Celiktutan, 2023] and On-Policy Pessimism Learning (OPL) [Kuznetsov *et al.*, 2021] optimize β online by aligning pessimism with approximation error. For example, GPL treats this adjustment as a dual optimization problem:

$$\beta = \arg\min_{\beta} \mathop{\mathrm{E}}_{p_0,\pi} \beta\big(Q^\pi(s,a) - r_{s,a} - \mathtt{sg}(\gamma V^{lb}_\phi(s'))\big), \quad (6)$$

where $V^{lb}_\phi(s') \approx Q^\mu_\phi(s,a) - \beta Q^\sigma_\phi(s,a) - \alpha \log \pi_\theta(a|s)$ and $\mathtt{sg}$ denotes the stop-gradient operator. This approach risks overfitting, as the adjustment heavily relies on critic outputs. Alternative methods, such as Tactical Optimism and Pessimism (TOP) [Moskovitz *et al.*, 2021], use external bandit controllers for β, but they can be less effective with continuous pessimism adjustments. We include additional discussion of critic disagreement in the appendix, as well as key comparisons of pessimism adjustment methods in Table 2.

3 Approximation Error and Pessimism

In this section, we focus on the analysis of critic approximation errors within the framework of pessimistic updates. For simplicity, we consider a fixed policy π_θ and use $V(s)$ and $Q(s,a)$ to represent the value and Q-value under this policy. We define the mean and lower bound approximation errors denoted as U^μ_ϕ and U^{lb}_ϕ respectively:

$$\begin{aligned} U^\mu_\phi(s,a) &Q(s,a) - Q^\mu_\phi(s,a), \\ U^{lb}_\phi(s,a) &Q(s,a) - Q^{lb}_\phi(s,a). \end{aligned} \quad (7)$$

Here, $Q(s,a)$ denotes the true Q-value, the term $Q^\mu_\phi(s,a)$ represents the mean Q-value estimated by an ensemble of k critics, calculated as $Q^\mu_\phi(s,a) = \frac{1}{k}\sum^k Q^i_\phi(s,a)$, and $Q^{lb}_\phi(s,a)$ is the lower bound Q-value as defined in Equation 5. Additionally, we introduce the mean and lower bound temporal critic errors, denoted as u^μ_ϕ and u^{lb}_ϕ, respectively:

$$\begin{aligned} u^\mu_\phi(s,a,s') r_{s,a} + \gamma V^\mu_\phi(s') - Q^\mu_\phi(s,a), \\ u^{lb}_\phi(s,a,s') r_{s,a} + \gamma V^{lb}_\phi(s') - Q^\mu_\phi(s,a). \end{aligned} \quad (8)$$

These temporal critic errors quantify the deviation between the Q-values $Q^\mu_\phi(s,a)$ and the mean or lower bound Temporal Difference (TD) targets. The value $V^{lb}_\phi(s)$ is equal to the expected value of $Q^{lb}_\phi(s,a)$ over all state-action pairs under policy π, such that $V^{lb}_\phi(s) = \mathrm{E}_\pi Q^{lb}_\phi(s,a) - \log \pi_\theta(a|s)$.

Lemma 3.1 (Approximation error operator)**.** *Given policy π, k on-policy q-value approximations $Q^1_\phi, Q^2_\phi, ..., Q^k_\phi$, sample mean Q^μ_ϕ and standard deviation Q^σ_ϕ, the mean and lower bound approximation errors follow a recursive formula:*

$$\begin{aligned} U^\mu_\phi(s,a) &= u^\mu_\phi(s,a,s') + \gamma \mathop{\mathrm{E}}_{a' \sim \pi} U^\mu_\phi(s',a'), \\ U^{lb}_\phi(s,a) &= u^{lb}_\phi(s,a,s') + \beta Q^\sigma_\phi(s,a) + \gamma \mathop{\mathrm{E}}_{a' \sim \pi} U^{lb}_\phi(s',a'), \\ U^{lb}_\phi(s,a) &= U^\mu_\phi(s,a) + \beta Q^\sigma_\phi(s,a). \end{aligned}$$

We expand on Lemma 3.1 in Appendix A. The lemma reveals that approximation errors exhibit a recurrent pattern analogous to Q-values. Specifically, the temporal errors function as an immediate signal, akin to rewards, while the future approximation errors serve as the bootstrap signal. Furthermore, this observation formalizes the intuitive concept that

minimizing the lower-bound error necessitates a precise calibration of the pessimistic correction against the temporal error and the approximation errors of subsequent states.

Key Insight

Lemma 3.1 shows that approximation errors propagate through recursive updates, just as Q-values propagate through Bellman equations. The temporal errors, u^{μ}_{ϕ} and u^{lb}_{ϕ}, act as immediate signals, while future approximation errors (weighted by γ) bootstrap over time. This highlights the need to balance immediate errors and long-term disagreement for effective learning.

It can be shown that similarly to the Bellman operator, both mean and lower bound error approximation operators are monotonic contractions:

Theorem 3.2 (Approximation error contraction)**.** *Let $\mathcal{F}$ be the space of functions on domain $S \times A$. We define the mean error and lower bound error operators $\mathcal{U}^{\mu}, \mathcal{U}^{lb} : \mathcal{F} \to \mathcal{F}$ as:*

$$\mathcal{U}^{\mu}\big(f(s,a)\big) u^{\mu}_{\phi}(s,a,s') + \gamma \mathop{\mathrm{E}}_{a' \sim \pi} f(s',a'),$$
$$\mathcal{U}^{lb}\big(f(s,a)\big) u^{lb}_{\phi}(s,a,s') + \beta Q^{\sigma}_{\phi}(s,a) + \gamma \mathop{\mathrm{E}}_{a' \sim \pi} f(s',a').$$

Above, $f(s,a) : S \times A \to R$ represents an estimate of the approximation error, and we assume that $Q^{\sigma}_{\phi}(s,a) \to 0$ as training progresses. Then it follows that both $\mathcal{U}^{\mu}$ and $\mathcal{U}^{lb}$ are monotonic contractions for any f_1 and f_2:

$$||\mathcal{U}(f_1) - \mathcal{U}(f_2)||_{\infty} \leq \gamma ||f_1 - f_2||_{\infty}.$$

We provide the relevant derivations in Appendix A.

Key Insight

Theorem 3.2 formalizes that both mean and lower-bound approximation error operators are contractions. This ensures that repeated applications of these operators converge to a fixed point. Intuitively, this means that as training progresses, the approximation errors stabilize, allowing the model to converge to a consistent value function.

As follows from Theorem 3.2, repeated application of the approximation error operator yields a Cauchy sequence, and therefore leads to a fixed point:

Corollary 3.3 (Approximation error fixed point)**.** *We denote repeated k applications of either approximation error operator to function f as $\mathcal{U}_k(f)$. Then, due to Banach fixed point theorem:*

$$\mathcal{U}^{\infty}(f) = f^* \ \wedge \ \mathcal{U}(f^*) = f^*.$$

The corollary shows that the approximation error of values can be effectively modeled using a fixed-point approach, analogous to treating values themselves. The potential ramifications and applications of this concept are further explored in Appendix A. Principally, the convergence of a pessimistic

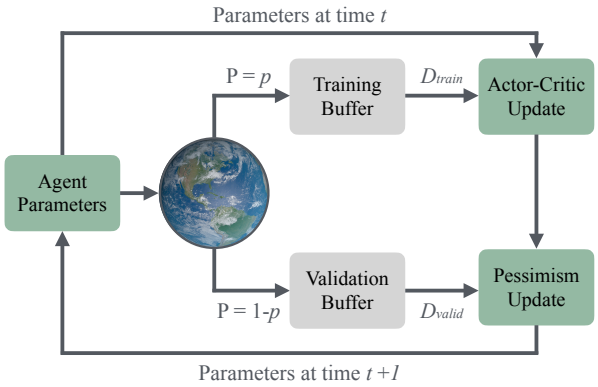

Figure 2: High-level overview of the proposed approach. After each environment step, the transition is stored in either the training buffer (used for updating actor-critic modules) or the validation buffer (used for updating the pessimism module). The pessimism is updated via a "reverse" TD loss, optimisation of which on the training buffer would be prone to overfitting.

value model signifies that the approximation errors converge to zero, implying $U^{\mu}_{\phi} = U^{lb}_{\phi} = 0$. The convergence proof of CDQL indicates that the value model should align with the true on-policy values under the conventional Q-learning convergence assumptions [Watkins and Dayan, 1992], [Fujimoto *et al.*, 2018]. Lemma 3.1 explicitly shows that for all s, a and s', both approximation errors equate to zero iff the following conditions are satisfied:

$$Q^{\mu}_{\phi}(s,a) = r + \gamma V^{\mu}_{\phi}(s') \ \wedge \ \beta Q^{\sigma}_{\phi}(s,a) = 0. \tag{9}$$

The convergence of a pessimistic model necessitates either the absence of critic ensemble disagreement (i.e., $Q^{\sigma}_{\phi}(s,a) = 0$ for all state-action pairs) or an algorithmic ability to diminish the level of pessimism over time, culminating in $\beta = 0$ asymptotically. We emphasize that this pertains specifically to convergence toward the zero-error fixed point defined in Lemma 3.1; an algorithm may still converge to a biased solution even if these conditions are not met, but exact fixed-point convergence with zero approximation error is unattainable otherwise. Figure 10 shows that the critic disagreement does not completely diminish on popular DeepMind Control and MetaWorld benchmarks. Given the improbability of achieving zero critic disagreement in overparameterized deep RL contexts, the adjustment of β emerges as a compelling strategy. Additionally, it can be demonstrated that under the scenario of critic underestimation, the lower-bound error exceeds the mean approximation error:

$$U^{\mu}_{\phi}(s,a) > 0 \implies |U^{\mu}_{\phi}(s,a)| \leq |U^{lb}_{\phi}(s,a)|. \tag{10}$$

The inequality follows from the third formula from Lemma 3.1, where $\beta \geq 0$ and $Q^{\sigma}(s,a) \geq 0$ as the standard deviation of the critic ensemble. We refer the reader to Appendix B for detailed proof.

As follows, pessimistic learning is advantageous only in overestimation, whereas it becomes detrimental in cases of underestimation. To this end, the pessimism levels should be adjusted in tandem with changes in the approximation errors. In practical terms, achieving a zero approximation error for either mean or lower bound is unrealistic. Given that

$U_\phi(s,a) \in R$, one might be interested in optimization of norm of $U^\mu_\phi(s,a)$ or $U^{lb}_\phi(s,a)$. This leads to the possibility of defining an "optimal" level of pessimism, where optimality is considered in relation to minimizing the respective approximation error norm. We note that our analysis yields a different approach to updating pessimism as compared to the method derived from dual optimization [Cetin and Celiktutan, 2023], which we discuss in Section 2.2.

Key Insight

Equation 10 highlights that pessimistic updates are beneficial in situations of overestimation, but they may harm learning when underestimation occurs. Therefore, adjusting the level of pessimism dynamically is critical to balance the benefits of avoiding overestimation with the risks of underestimation.

4 Validation Pessimism Learning Algorithm

Building on the analysis conducted in the previous Section, we propose the Validation Pessimism Learning module (VPL). The goal of the VPL module is to adjust the pessimism parameter such that the critic targets (lower bound Q-value approximation) has the least approximation error. As such, VPL can be used as an alternative to CDQL or GPL in conjuction with any off-policy actor-critic algorithm. For our analysis, we utilize the Soft Actor-Critic (SAC) [Haarnoja *et al.*, 2018] as the backbone algorithm. VPL is based on a simple premise of adjusting pessimism via a TD loss. Given that the critic concurrently optimizes this loss function, such setup is especially prone to overfitting. To mitigate this, the optimization of the pessimism parameter is conducted on a distinct set of *validation* data, which remains unseen by the actor-critic modules. From a theoretical standpoint, VPL can be interpreted as a strategy for pessimism model selection, with the selection process aimed at minimizing the lower bound approximation error delineated in the previous section. A critical aspect of VPL involves conducting the pessimism model selection on validation data. The model selection is achieved through gradient-based optimization of the proposed pessimism loss. The utilization of validation data in this process reduces the probability of overfitting to bootstrapped supervision signals used by TD learning. We summarize VPL approach in Figure 2 and share pseudo-code in Section B.1, where we colour changes wrt. regular SAC.

4.1 Validation Buffer

The employment of validation data is a well-established practice in supervised learning frameworks [Bishop and Nasrabadi, 2006]. It serves a dual purpose: providing an unbiased assessment of model performance trained on the training dataset, and facilitating regularization techniques such as early stopping [Prechelt, 2002] or hyperparameter tuning [Bergstra and Bengio, 2012]. However, the integration of validation data entails a trade-off, notably the reduction of the training set size. In supervised learning, the regret associated with decreasing the training set can be quantitatively evaluated through the lens of neural scaling laws [Rosenfeld *et al.*, 2019]. Such regret is, to the best of our knowledge, a relatively understudied area in the context of online RL. In online RL, the notion of a validation buffer is not popular, primarily due to the requisite sacrifice of actor-critic learning on the validation transitions. Given inherent sample inefficiency of RL, this cost is often deemed as overly burdening. Contrary to supervised learning setup, RL is characterized by a high

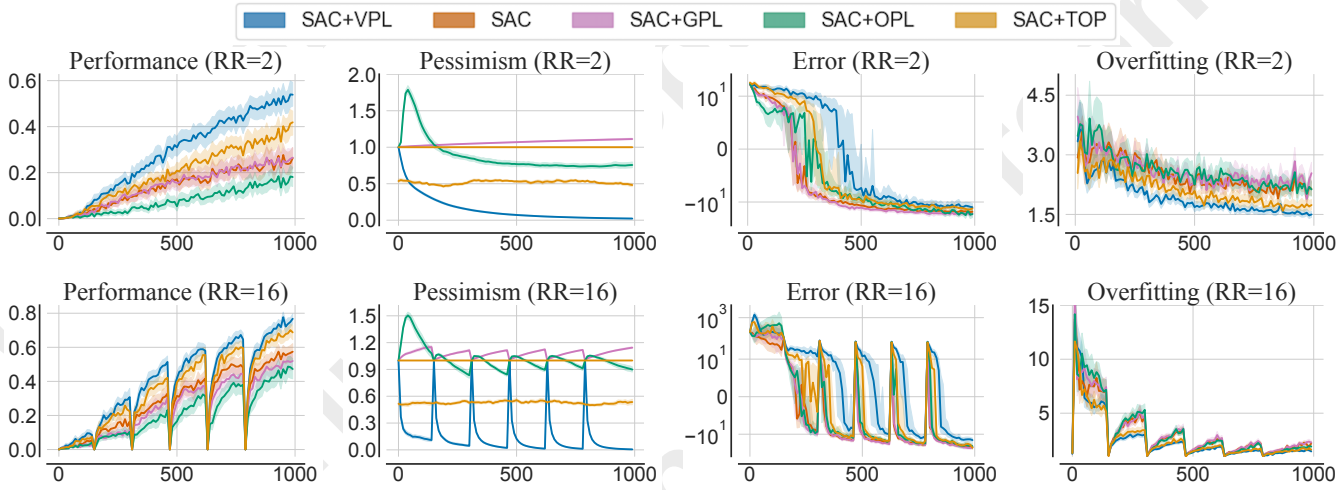

Figure 3: We integrate the Soft Actor-Critic (SAC) and the Scaled-By-Resetting SAC (SR-SAC) with various pessimism adjustment algorithms. Performance is evaluated in both low replay (top row) and high replay (bottom row) regimes. All tested algorithms use the same network architectures and hyperparameter settings, so performance differences arise solely from the pessimism adjustment strategies. Despite similar motivations, each algorithm exhibits different levels of pessimism. Our proposed Validation Pessimism Learning (VPL) module demonstrates the lowest approximation error and mitigates value overfitting more effectively than alternative approaches, leading to improvements in performance and sample efficiency. The experimental setting is detailed in Sections 5 and E. Results are based on 20 tasks with 10 seeds per task, presented as interquartile mean (IQM) and 95% confidence intervals (CI).

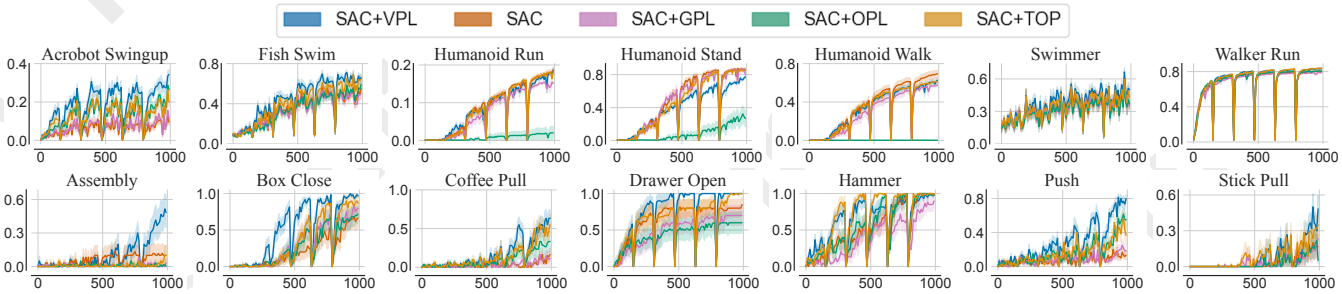

Figure 4: Task-specific performance of high-replay configurations in 14 out of 20 considered tasks. VPL achieves performance improvements, especially in the manipulation tasks. In the case of DMC tasks the y-axis denotes evaluation returns, whereas for MetaWorld tasks it denotes the evaluation success ratio. The x-axis shows environment steps (in thousands). We detail the setting in Section 5.1. 10 seeds per task.

correlation between successive samples, thereby diminishing the marginal utility of processing additional samples from the same trajectory. Consequently, we posit that in online RL, the cost associated with the use of validation data can be counterbalanced, provided the validation data is leveraged to enhance the learning process. In the case of the VPL, we allocate the validation transitions exclusively for the adjustment of the pessimism parameter. This approach presents a novel utilization of validation data in online RL in a manner that is RL-specific, diverging from supervised methodologies.

4.2 Pessimism Update Rule

The persistence of critic disagreement throughout training implies that the standard convergence guarantees of the pessimistic temporal difference update towards on-policy values are not upheld when $\beta \neq 0$. Moreover, in cases where minimizing the mean approximation error is not achievable, particularly in scenarios characterized by strong overestimation, the presence of non-zero critic disagreement can be leveraged to decrease the lower bound approximation error by increasing β. This observation forms the basis for our proposed method of adjusting β. The aim is to minimize the expected lower bound approximation error $U^{lb}_{\phi}(s,a)$, formulated as follows:

$$\beta^* = \arg\min_{\beta} \mathop{\mathrm{E}}_{p_0,\pi} \sum_{t=0}^{\infty} \gamma^t U^{lb}_{\phi}(s,a). \tag{11}$$

Unfortunately, obtaining $U^{lb}_{\phi}(s,a)$ is challenging as it necessitates an estimate of the true on-policy Q-value. Typically, such estimates are derived through methods like Monte-Carlo (MC) rollouts, TD(n), or TD(λ), with MC being the only unbiased method. However, in the context of off-policy learning or non-terminating environments, employing MC rollouts is impractical. Consequently, we leverage the simple approach proposed by [Cetin and Celiktutan, 2023] in which it is assumed that the critic output for prerecorded off-policy actions is unbiased. Therefore, we assume that $Q^{\pi}(s,a) = Q^{\mu}_{\phi}(s,a)$ for actions that do not maximize the output of the policy. Additionally, akin to the approach in off-policy actor-critic algorithms, the policy-induced distribution is approximated using an off-policy replay buffer. This approach leads to the formulation of the following:

$$\beta^* \approx \arg\min_{\beta} \mathop{\mathrm{E}}_{\mathcal{D}_v} \left(Q^{\mu}_{\phi}(s,a) - r_{s,a} - \gamma V^{lb}_{\phi}(s')\right)^2. \tag{12}$$

In this formulation, $\mathcal{D}_V$ represents the validation replay buffer, with s, a, s' denoting transitions sampled from this buffer. In line with other stochastic policy algorithms, we further approximate $V^{lb}_{\phi}(s')$ with the critic output for a single action $a' \sim \pi_{\theta}(a'|s')$. As follows, VPL adjusts the pessimism under the assumption that $Q^{\mu}_{\phi}(s,a)$ is a good representation of $Q^{\pi}(s,a)$. Since the actions at which $Q^{\mu}_{\phi}(s,a)$ is evaluated are sampled from the validation buffer and are off-policy, these actions are likely to produce less overestimation than the adversarial actions sampled from a value-maximizing policy.

The inclusion of a square in the loss function ensures that the optimization remains non-negative and convex with respect to β, focusing the updates on reducing significant errors, which are particularly impactful in reinforcement learning scenarios. Additionally, removing the stop-gradient operator allows β to be updated dynamically based on both the approximation error and the disagreement among critics (Q^{σ}), tightly coupling the level of pessimism to the degree of uncertainty in the critic estimates. This ensures that the level of pessimism is adjusted adaptively during training, aligning with the intuition that higher disagreement indicates greater uncertainty, necessitating increased pessimism.

While we acknowledge that providing a strict mathematical justification for all aspects of the proposed rule is challenging, our empirical results (Figure 3) demonstrate its practical effectiveness in reducing overfitting and improving performance across diverse environments. VPL mitigates the risk of overfitting by computing the pessimism loss exclusively on validation samples that are not used by the actor-critic modules. Unlike GPL, VPL fully incorporates critics' disagreement into the optimization process and dynamically adjusts β, enabling more precise control of the level of pessimism based on the observed uncertainty.

5 Experiments

Our experiments are based on the JaxRL codebase [Kostrikov, 2021]. Since all considered algorithms use SR-SAC [D'Oro *et al.*, 2022] as their backbone, we align the common hyperparameters with those recommended for Scaled-By-Resetting SAC (SR-SAC) as per [D'Oro *et al.*, 2022]. This includes using the same network architectures and a two-critic ensemble in accordance with established practices [Fujimoto *et al.*, 2018], [Haarnoja *et al.*, 2018],

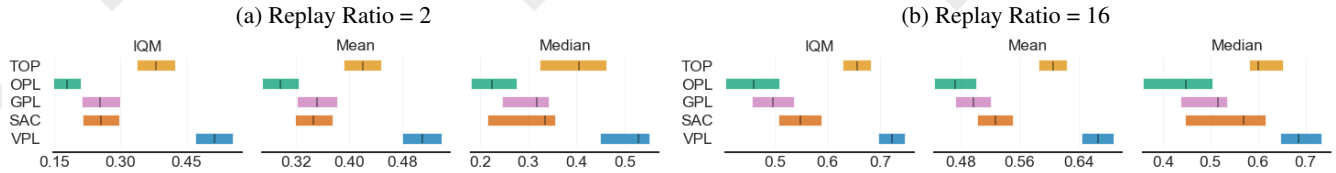

Figure 5: RLiable final performance metrics for the main experiment detailed in Section 5.1. VPL outperforms baseline algorithms in both replay regimes. The performance metrics are calculated on 20 tasks listed in Table 3 with 10 random seeds per task.

[Ciosek *et al.*, 2019], [Moskovitz *et al.*, 2021], [Cetin and Celiktutan, 2023]. We conduct our experiments in two environments: the DeepMind Control (DMC) suite [Tassa *et al.*, 2018] and the single-task MetaWorld [Yu *et al.*, 2020]. Our study encompasses two replay regimes: a compute-efficient setup with 2 gradient steps per environment step without resets, and a sample-efficient setup with 16 gradient steps per environment step, including full-parameter resets every 160k steps, as suggested by [D'Oro *et al.*, 2022]. We provide robust analysis using the RLiable package [Agarwal *et al.*, 2021] and detail the experimental setting in Appendix E.

5.1 Performance and Sample Efficiency

Firstly, we test the performance and sample efficiency of the proposed approach. To this end, we compare SR-SAC [D'Oro *et al.*, 2022] (DMC state of the art) to four algorithms that extend SR-SAC with online pessimism adjustment: GPL [Cetin and Celiktutan, 2023]; OPL [Kuznetsov *et al.*, 2021]; TOP [Moskovitz *et al.*, 2021]; and VPL (the proposed approach). We run the tested algorithms in both replay regimes for 1mln environment steps on 20 medium to hard tasks (10 from DMC and 10 from MetaWorld). We discuss the chosen baselines in Sections 2.2 & C. We discuss hyperparameter selection in Appendix G and the tested tasks in F. We report the results of this experiment in Figures 3, 4 & 5. We find that the proposed approach surpasses baseline algorithms, demonstrating 48% and 27% higher performance than the baseline SR-SAC in low and high replay regimes, respectively. As depicted in Figure 4, VPL exhibits particular effectiveness in MetaWorld manipulation tasks, developing robust policies in environments where other approaches fail, such as the assembly task. To provide further practical context, we evaluate the computational overhead introduced by each of the tested pessimism adjustment methods. Table 1 shows that VPL maintains a low wall-clock overhead compared to SR-SAC, with only a 3.5% increase in the low-replay regime and 3.8% in the high-replay regime. This minimal computational overhead highlights VPL's suitability for large-scale and real-time reinforcement learning applications.

METHOD	GPL	OPL	TOP	VPL
RR= 2	0.3%	6.3%	0.3%	3.5%
RR= 16	0.5%	1.1%	0.1%	3.8%

Table 1: We measure runtimes for 2000 runs of each algorithm and find that the pessimism adjustment methods have trivial wall-clock overhead as compared to SAC/SR-SAC.

5.2 Validation Buffer Regret

To understand the impact of a validation buffer on online RL training, we analyze three distinct agent setups: *baseline* SR-SAC, which operates without a validation buffer, thus updating actor-critic modules with all experienced transitions; *regret* SR-SAC, which maintains a validation buffer but does not employ validation transitions for pessimism adjustment; and SR-SAC-VPL, which not only maintains a validation buffer but also utilizes validation transitions for pessimism adjustment (we present these results in Figure 6).

This comparative analysis aims to isolate the performance loss attributable to the presence of a validation buffer and the efficiency gains derived from employing VPL for updating pessimism. We evaluate these agents in high-replay regime on 4 tasks (listed in Table 4) over 1mln environment steps, using varying ratios of validation to training samples, specifically at proportions of $\frac{1}{128}$, $\frac{1}{32}$, $\frac{1}{8}$, and $\frac{1}{2}$. The results for this experiment are presented in Figure 6. We observe that the regret associated with maintaining a validation buffer, and thus not utilizing it for actor-critic updates, diminishes over the course of training. Specifically, the *regret* SR-SAC reaches parity with the SR-SAC in performance for all validation proportions except at $\frac{1}{2}$. We note that the rate of regret reduction correlates with the size of the validation proportion, with smaller proportions converging to baseline performance more rapidly. When examining the effectiveness of pessimism adjustment, we observe its most pronounced impact during the early stages of training. This trend aligns with the expectation of reducing critic disagreement over time. Additionally, the extent of performance gain appears to be influenced by the size of the validation buffer, where larger proportions yield greater improvements. This effect is likely due to the increased diversity of environment transitions available for pessimism adjustment in larger buffers. When considering the combined effects on performance, our findings indicate that, except for the $\frac{1}{2}$ proportion, all validation proportions successfully compensate for the performance loss due to validation buffer maintenance. This result is in line with the broader experimental results presented in Figures 3 & 5.

5.3 Other Experiments

We investigate the sensitivity of VPL to varying pessimism learning rates as compared other pessimism adjustment algorithms. Given the dependency of such learning rate on reward scales and environmental dynamics, determining an optimal rate a priori is challenging, which is a significant restriction for practical applications. To address this, we test the performance of VPL, GPL, and OPL across four environments

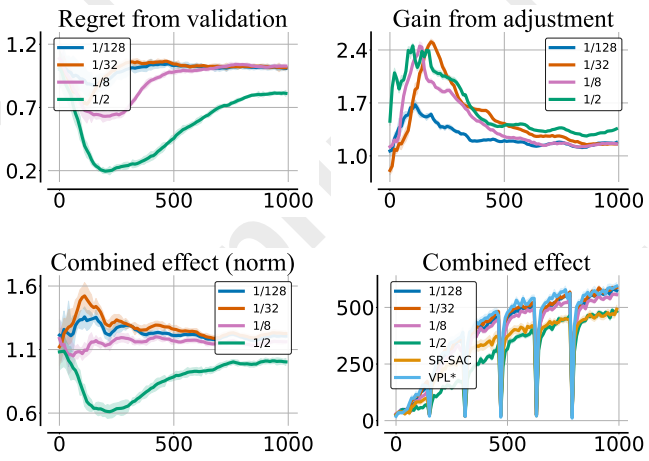

Figure 6: We examine the impact of maintaining a validation buffer on performance distinct from pessimism adjustment across varying proportions of validation samples. Upper-left figure demonstrates whether validation agents can match the performance of their validation-free counterparts without utilizing validation samples for pessimism updates, enabling quantification of the regret associated with allocating samples to a validation buffer. Upper-right figure quantifies the performance gains attributable to pessimism by contrasting agents that do not update pessimism against those that do. Figures in the bottom illustrate the cumulative effect of validation pessimism adjustment for different validation ratios, benchmarking against the baseline performance of SR-SAC and VPL with "free" validation (denoted as VPL*). X-axis denotes environments steps (in thousands) and y-axis denotes performance.

detailed in Table 4 in the high-replay regime. We evaluate agents after 500k environments steps for learning rates of $[5e-5, 5e-4, 5e-3, 5e-2]$. The results, presented in Figure 7, indicate that VPL exhibits less sensitivity to changes in the pessimism learning rate than the other considered algorithms. Furthermore, we investigate the importance of the two proposed design elements: the use of a validation buffer and the VPL pessimism loss as formulated in Equation 12. To this end, we compare the performance of six agents, each employing different combinations of pessimism loss – either the dual optimization pessimism loss or the VPL pessimism loss – along with varying sources for pessimism updates. These sources include samples from the replay buffer, the validation buffer, and the most recent transitions. The results of this analysis are presented in Figure 9. In our final analysis, we focus on validating the premise of VPL: its effectiveness in reducing approximation error and mitigating overfitting compared to baseline algorithms. Our methodology for quantifying approximation error and overfitting are described in Appendix E. We conducted these measurements across both low and high replay regimes, using a selection of 20 tasks from the DMC and MetaWorld as listed in Table 3. The findings, depicted in Figure 3 and Appendix H, confirm that VPL achieves the lowest levels of critic overfitting and approximation error in both replay scenarios.

6 Limitations

The primary challenge of VPL lies in estimating the lower-bound approximation error necessary for the pessimism adjustment mechanism. This estimation currently relies on a simplified assumption, inherited from GPL and discussed in Section 4.2, that Q^{μ} provides a reasonable approximation of Q^{π}. While this assumption works well in many standard benchmarks, it may limit the applicability of VPL in environments with high policy dynamics or entropy, where off-policy actions are less representative of on-policy behavior. Future research could explore alternative estimation methods for the lower-bound approximation error that do not rely on this assumption, potentially leading to more robust algorithms. Surprisingly, our experiments (see Figure 6) reveal that using a validation buffer does not detrimentally impact agent performance in high-replay scenarios, except in extremely sample-scarce environments (e.g., fewer than 250k environment steps). Addressing the limitations of the current error estimation framework may also mitigate this sample-scarcity sensitivity.

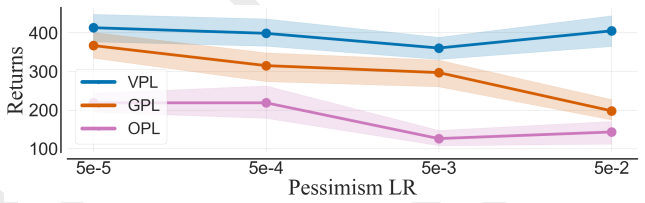

Figure 7: VPL exhibits substantially less sensitivity to the learning rate of the pessimism module. 4 tasks, 10 seeds per task.

7 Conclusions

This paper examined the approximation error in critic networks optimized via temporal difference variants. We introduced a fixed-point model for estimating mean and lower bound errors and used this model to analyze the convergence of pessimistic actor-critic algorithms. We proposed the VPL algorithm, which dynamically adjusts pessimism levels to minimize approximation errors of critic supervision in validation samples. We tested VPL against baseline algorithms in various locomotion and manipulation tasks, showing performance and sample efficiency improvements. We explored the impact of VPL components and their sensitivity to hyperparameter selection. Our results confirm VPLs effectiveness in complex continuous action tasks.

Acknowledgements

Marek Cygan was partially supported by an NCBiR grant POIR.01.01.01-00-0433/20. Mateusz Ostaszewski was partially supported by the National Science Centre, Poland, under a grant 2023/51/D/ST6/01609. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018377

References

[Agarwal *et al.*, 2021] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Belle-

mare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

[Ball *et al.*, 2023] Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. 2023.

[Bellemare *et al.*, 2017] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

[Bergstra and Bengio, 2012] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[Bishop and Nasrabadi, 2006] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[Cetin and Celiktutan, 2023] Edoardo Cetin and Oya Celiktutan. Learning pessimism for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6971–6979, 2023.

[Chen *et al.*, 2020] Xinyue Chen, Che Wang, Zijian Zhou, and Keith W Ross. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2020.

[Ciosek *et al.*, 2019] Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. *Advances in Neural Information Processing Systems*, 32, 2019.

[D'Oro *et al.*, 2022] Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *The Eleventh International Conference on Learning Representations*, 2022.

[Farahmand *et al.*, 2010] Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.

[Foret *et al.*, 2020] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

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

[Ha and Schmidhuber, 2018] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[Haarnoja *et al.*, 2017] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[Hasselt, 2010] Hado Hasselt. Double q-learning. *Advances in neural information processing systems*, 23, 2010.

[Hessel *et al.*, 2018] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Hiraoka *et al.*, 2021] Takuya Hiraoka, Takahisa Imagawa, Taisei Hashimoto, Takashi Onishi, and Yoshimasa Tsuruoka. Dropout q-functions for doubly efficient reinforcement learning. In *International Conference on Learning Representations*, 2021.

[Januszewski *et al.*, 2021] Piotr Januszewski, Mateusz Olko, Michał Królikowski, Jakub Swiatkowski, Marcin Andrychowicz, Łukasz Kuciński, and Piotr Miłoś. Continuous control with ensemble deep deterministic policy gradients. In *Deep RL Workshop NeurIPS 2021*, 2021.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kostrikov, 2021] Ilya Kostrikov. JAXRL: Implementations of Reinforcement Learning algorithms in JAX, 10 2021.

[Kumar *et al.*, 2019] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

[Kumar *et al.*, 2020] Aviral Kumar, Abhishek Gupta, and Sergey Levine. Discor: Corrective feedback in reinforcement learning via distribution correction. *Advances in Neural Information Processing Systems*, 33:18560–18572, 2020.

[Kuznetsov *et al.*, 2020] Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, pages 5556–5566. PMLR, 2020.

[Kuznetsov *et al.*, 2021] Arsenii Kuznetsov, Alexander Grishin, Artem Tsypin, Arsenii Ashukha, Artur Kadurin, and Dmitry Vetrov. Automating control of overestimation bias for reinforcement learning. *arXiv preprint arXiv:2110.13523*, 2021.

[Lee *et al.*, 2023] Hojoon Lee, Hanseul Cho, Hyunseung Kim, Daehoon Gwak, Joonkee Kim, Jaegul Choo, Se-Young Yun, and Chulhee Yun. Plastic: Improving input and label plasticity for sample efficient reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Li *et al.*, 2022] Qiyang Li, Aviral Kumar, Ilya Kostrikov, and Sergey Levine. Efficient deep reinforcement learning requires regulating overfitting. In *The Eleventh International Conference on Learning Representations*, 2022.

[Lyle *et al.*, 2023] Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. *arXiv preprint arXiv:2303.01486*, 2023.

[Mnih *et al.*, 2015] 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.

[Moskovitz *et al.*, 2021] Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano, Michael Arbel, and Michael Jordan. Tactical optimism and pessimism for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12849–12863, 2021.

[Munos and Szepesvári, 2008] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[Munos, 2005] Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[Nauman and Cygan, 2023] Michal Nauman and Marek Cygan. On the theory of risk-aware agents: Bridging actor-critic and economics. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2023.

[Nauman *et al.*, 2024] Michal Nauman, Michał Bortkiewicz, Piotr Miłoś, Tomasz Trzcinski, Mateusz Ostaszewski, and Marek Cygan. Overestimation, overfitting, and plasticity in actor-critic: the bitter lesson of reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. PMLR 235:37342-37364.

[Nikishin *et al.*, 2022] Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International conference on machine learning*, pages 16828–16847. PMLR, 2022.

[Prechelt, 2002] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[Puterman, 2014] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Rosenfeld *et al.*, 2019] Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2019.

[Rowland *et al.*, 2023] Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023.

[Seyde *et al.*, 2022] Tim Seyde, Wilko Schwarting, Sertac Karaman, and Daniela Rus. Learning to plan optimistically: Uncertainty-guided deep exploration via latent model ensembles. In *Conference on Robot Learning*, pages 1156–1167. PMLR, 2022.

[Shang *et al.*, 2016] Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *international conference on machine learning*, pages 2217–2225. PMLR, 2016.

[Silver *et al.*, 2014] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.

[Tassa *et al.*, 2018] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

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

[Van Hasselt *et al.*, 2016] 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.

[Wang *et al.*, 2022] Zhihai Wang, Jie Wang, Qi Zhou, Bin Li, and Houqiang Li. Sample-efficient reinforcement learning via conservative model-based actor-critic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8612–8620, 2022.

[Watkins and Dayan, 1992] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

[Yao *et al.*, 2021] Yao Yao, Li Xiao, Zhicheng An, Wanpeng Zhang, and Dijun Luo. Sample efficient reinforcement learning via model-ensemble exploration and exploitation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4202–4208. IEEE, 2021.

[Yu *et al.*, 2020] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[Zhang *et al.*, 2017] Zongzhang Zhang, Zhiyuan Pan, and Mykel J Kochenderfer. Weighted double q-learning. In *IJCAI*, pages 3455–3461, 2017.