The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Learning Pessimism for Reinforcement Learning

Edoardo Cetin[1], **Oya Celiktutan**[1]

[1] King's College London
edoardo.cetin@kcl.ac.uk, oya.celiktutan@kcl.ac.uk

Abstract

Off-policy deep reinforcement learning algorithms commonly compensate for overestimation bias during temporal-difference learning by utilizing pessimistic estimates of the expected target returns. In this work, we propose *Generalized Pessimism Learning (GPL)*, a strategy employing a novel *learnable* penalty to enact such pessimism. In particular, we propose to learn this penalty alongside the critic with *dual TD-learning*, a new procedure to estimate and minimize the magnitude of the target returns bias with trivial computational cost. *GPL* enables us to accurately counteract overestimation bias *throughout* training without incurring the downsides of overly pessimistic targets. By integrating *GPL* with popular off-policy algorithms, we achieve *state-of-the-art* results in both competitive proprioceptive and pixel-based benchmarks.

1 Introduction

Sample efficiency and generality are two directions in which reinforcement learning (RL) algorithms are still lacking, yet, they are crucial for tackling complex real-world problems (Mahmood et al. 2018). Consequently, many RL milestones have been achieved through simulating large amounts of experience and task-specific parameter tuning (Mnih et al. 2013; Silver et al. 2017). Recent off-policy model-free (Chen et al. 2021) and model-based algorithms (Janner et al. 2019) advanced RL's sample-efficiency on several benchmark tasks. We attribute such improvements to two main linked advances: more expressive models to capture uncertainty and better strategies to counteract detrimental biases from the learning process. These advances yielded the stabilization to adopt more aggressive optimization procedures, with particular benefits in lower data regimes.

Modern policy gradient algorithms learn behavior by optimizing the expected performance as predicted by the *critic*, a trained parametric model of the agent's performance in the environment. Within this process, overestimation bias naturally arises from the maximization performed over the critic's performance predictions, and consequently, also over the critic's possible errors. In the context of off-policy RL, the critic is trained to predict the agent's future returns via temporal difference (TD-) learning. A common strategy to counteract overestimation is to parameterize this model with multiple, independently-initialized networks and optimize the agent's behavior over the minimum value of the relative outputs (Fujimoto, Van Hoof, and Meger 2018). Empirically, this strategy consistently yields pessimistic target performance measures, avoiding overestimation bias propagating through the TD-learning target bootstraps. However, this approach directly links the critic's parameterization to bias counteraction and appears to suffer from suboptimal exploration due to *underestimation* bias (Ciosek et al. 2019).

Based on these observations, we propose ***Generalized Pessimism Learning (GPL)***, a new strategy that *learns* to counteract biases by optimizing a new dual objective. GPL makes use of an explicit penalty to correct the critic's target predictions. We design this penalty as a weighted function of epistemic uncertainty, computed as the expected Wasserstein distance between the return distributions predicted by the critic. We learn the penalty's weight with *dual TD-learning*, a new procedure to estimate and counteract any arising bias in the critic's predictions with dual gradient descent. GPL is the first method to freely learn unbiased performance targets throughout training.

We extend GPL by introducing *pessimism annealing*, a new procedure motivated by the principle of *optimism in the face of uncertainty* (Brafman and Tennenholtz 2002). This procedure leads the agent to adopt risk-seeking behavior, by utilizing a purposely biased estimate of the performance in the initial training stages. This allows it to trade expected immediate performance for improved directed exploration, incentivizing the visitation of states with high uncertainty from which it would gain more information.

We incorporate GPL with the *Soft Actor-Critic (SAC)* (Haarnoja et al. 2018a,b) and *Data-regularized Q (DrQ)* algorithms (Yarats, Kostrikov, and Fergus 2021; Yarats et al. 2021), yielding *GPL-SAC* and *GPL-DrQ*. On challenging Mujoco tasks from OpenAI Gym (Todorov, Erez, and Tassa 2012; Brockman et al. 2016), GPL-SAC outperforms both model-based (Janner et al. 2019) and model-free (Chen et al. 2021) state-of-the-art algorithms, while being more computationally efficient. For instance, in the Humanoid environment GPL-SAC recovers a score of 5000 in less than 100K steps, more than nine times faster than regular SAC. Additionally, on pixel-based environments from the DeepMind Control Suite (Tassa et al. 2018), GPL-DrQ provides significant performance improvements from the recent state-of-

the-art *DrQv2* algorithm. We validate the statistical significance of our improvements using *Rliable* (Agarwal et al. 2021), further highlighting the effectiveness and applicability of GPL. We share our code to facilitate future extensions.

In summary, we make three main contributions toward improving off-policy reinforcement learning:

- We propose Generalized Pessimism Learning, using the first dual optimization method to estimate and precisely counteract overestimation bias throughout training.
- To improve exploration, we extend GPL with pessimism annealing, a strategy that exploits epistemic uncertainty to actively seek for more informative states.
- We integrate our method with SAC and DrQ, yielding new state-of-the-art results with trivial computational overheads on both proprioceptive and pixel tasks.

2 Related Work

Modern model-free off-policy algorithms utilize different strategies to counteract overestimation bias arising in the critic's TD-targets (Thrun and Schwartz 1993; Pendrith, Ryan et al. 1997; Mannor et al. 2007). Many approaches combine the predictions of multiple function approximators to estimate the expected returns, for instance, by independently selecting the bootstrap action (Hasselt 2010). In discrete control, such a technique appears to mitigate the bias of the seminal *DQN* algorithm (Mnih et al. 2013), consistently improving performance (Van Hasselt, Guez, and Silver 2016; Hessel et al. 2018). In continuous control, similar strategies successfully stabilize algorithms based on the policy gradient theorem (Silver et al. 2014). Most notably, Fujimoto, Van Hoof, and Meger (2018) proposed to compute the critic's TD-targets by taking the minimum over the outputs of two different action-value models. This minimization strategy has become ubiquitous, being employed in many popular subsequent works (Haarnoja et al. 2018b). For a better trade-off between optimism and pessimism, Zhang, Pan, and Kochenderfer (2017) proposed using a weighted combination of the original and minimized targets. Instead, Kuznetsov et al. (2020) proposed to parameterize a distributional critic and drop a fixed fraction of the predicted quantiles to compute the targets. Alternative recently proposed strategies for bias-counteraction also entail combining the different action-value predictions with a Softmax function (Pan, Cai, and Huang 2020) and computing action-value targets with convex combinations of predictions obtained from multiple actors (Lyu et al. 2022). Similarly to our approach, several works considered explicit penalties based on heuristic measures of epistemic uncertainty (Lee, Defourny, and Powell 2013; Ciosek et al. 2019). Recently, Kumar, Gupta, and Levine (2020) proposed to complement these strategies by further reducing bias propagation through actively weighing the TD-loss of different experience samples. All these works try to *hand-engineer* a fixed penalization to counteract the critic's bias. In contrast, we show that any fixed penalty would be inherently suboptimal (Section 5.2) and propose a novel strategy to precisely adapt the level of penalization throughout training.

Within model-based RL, recent works have achieved remarkable sample efficiency by learning large ensembles of dynamic models for better predictions (Chua et al. 2018; Wang and Ba 2019; Janner et al. 2019). In the model-free framework, prior works used large critic ensembles for more diverse scopes. Anschel, Baram, and Shimkin (2017) proposed to build an ensemble using several past versions of the value network to reduce the magnitude of the TD-target's bias. Moreover, Lan et al. (2020) showed that different sampling procedures for the critic's ensemble predictions can regulate underestimation bias. Their work was extended to the continuous setting by Chen et al. (2021), which showed that large ensembles combined with a high update-to-data ratio can outperform the sample efficiency of contemporary model-based methods. Ensembling has also been used to achieve better exploration following the optimism in the face of uncertainty principle in both discrete (Chen et al. 2017) and continuous settings (Ciosek et al. 2019). In addition to these advantages, we show that GPL can further exploit large ensembles to better estimate and learn to counteract bias.

In the same spirit as this work, multiple prior methods attempted to learn the components and parameters of underlying RL algorithms. Several works have approached this problem by utilizing expensive meta-learning strategies to obtain new learning objectives based on the multi-task performance from low-computation environments (Oh et al. 2020; Xu et al. 2020; Bechtle et al. 2021). More related to our method, *Tactical Optimism and Pessimism* (Moskovitz et al. 2021) introduced the concept of adapting a bias penalty online. Together with similar later work (Kuznetsov et al. 2021), they proposed step-wise updates to the bias correction parameters based on the performance of recent trajectories. Instead, GPL proposes a new method to precisely estimate bias and reduce its magnitude via dual gradient descent. We provide a direct empirical comparison in the extended version of this work (Cetin and Celiktutan 2021).

3 Preliminaries

In RL, we aim to autonomously recover optimal agent behavior for performing a particular task. Formally, we describe this problem setting as a Markov Decision Process (MDP), defined as the tuple $(S, A, P, p_0, r, \gamma)$. At each timestep of interaction, the agent observes some state in the state space, $s \in S$, and performs some action in the action space, $a \in A$. The transition dynamics function $P : S \times A \times S \to \mathbb{R}$ and the initial state distribution $p_0 : S \to \mathbb{R}$ describe the evolution of the environment as a consequence of the agent's behavior. The reward function $r : S \times A \to \mathbb{R}$ quantifies the effectiveness of each performed action, while the discount factor $\gamma \in [0, 1]$ represents the agent's preference for earlier rewards. A policy $\pi : S \times A \to \mathbb{R}$ maps each state to a probability distribution over actions and represents the agent's behavior. An episode of interactions between the agent and the environment yields a trajectory τ containing the transitions experienced, $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, ...)$. The RL objective is then to find an optimal policy π^* that maximizes

the expected sum of discounted future rewards:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{p_\pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right], \tag{1}$$

where $p_\pi(\tau)$ represents the distribution of trajectories stemming from the agent's interaction with the environment. Off-policy RL algorithms commonly utilize a critic model to evaluate the effectiveness of the agent's behavior. A straightforward choice for the critic is to represent the policy's action-value function $Q^\pi : S \times A \rightarrow \mathbb{R}$. This function quantifies the expected sum of discounted future rewards after executing some particular action from a given state:

$$Q^\pi(s, a) = \mathbb{E}_{p_\pi(\tau|s_0=s, a_0=a)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \tag{2}$$

Most RL algorithms consider learning parameterized models for both the policy, π_θ, and the corresponding action-value function, Q^π_ϕ. In particular, after storing experience transitions (s, a, s', r) in a replay data buffer D, we learn Q^π_ϕ by iteratively minimizing a squared TD-loss of the form:

$$\begin{aligned} J_Q(\phi) &= \mathbb{E}_{s,a,s',r\sim D} \left[(Q^\pi_\phi(s, a) - y)^2 \right], \\ y &= r + \gamma \mathbb{E}_{a\sim\pi(s')} \left[\hat{Q}^\pi_{\phi'}(s', a) \right]. \end{aligned} \tag{3}$$

Here, the TD-targets y are obtained by computing a 1-step bootstrap with a *target action-value estimator* $\hat{Q}^\pi_{\phi'}$. Usually, $\hat{Q}^\pi_{\phi'}$ is a regularized function of action-value predictions from a target critic model using delayed parameters ϕ'. Following the policy gradient theorem (Sutton et al. 2000; Silver et al. 2014), we can then improve our policy by maximizing the expected returns as predicted by the critic. This corresponds to minimizing the negation of the critic's expected target action-value estimates:

$$J_\pi(\theta) = -\mathbb{E}_{s\sim D, a\sim\pi_\theta(s)} \left[\hat{Q}^\pi_\phi(s, a) \right]. \tag{4}$$

4 Addressing Overestimation Bias

4.1 Bias in Q-Learning

In off-policy RL, several works have identified an accumulation of overestimation bias in the action-value estimates as a consequence of TD-learning (Thrun and Schwartz 1993; Mannor et al. 2007). Formally, we quantify the target action-value bias $B(s, a, s')$ as the difference between the actual and estimated TD-targets of a transition (Chen et al. 2021):

$$B(s, a, s') = \gamma \mathbb{E}_{a'\sim\pi(s')} \left[\hat{Q}^\pi_{\phi'}(s', a') - Q^\pi(s', a') \right]. \tag{5}$$

Positive bias arises when the target action-values are obtained directly from the outputs of a parameterized action-value function, i.e., $\hat{Q}^\pi_{\phi'} = Q^\pi_{\phi'}$ (Fujimoto, Van Hoof, and Meger 2018). The reason for this phenomenon is that the policy is trained to locally maximize the action-value estimates from Eqn. 4. Hence, its actions will exploit potential model errors to obtain higher scores, implying that $\mathbb{E}_{s,a\sim\pi(s)}[Q^\pi_{\phi'}(s, a)] > \mathbb{E}_{s,a\sim\pi(s)}[Q^\pi(s, a)]$. Instabilities then arise as the errors can quickly propagate through the bootstrap operation, inherently causing the phenomenon of *positive bias accumulation*. To counteract this phenomenon, Fujimoto, Van Hoof, and Meger (2018) proposed *clipped double Q-learning*. This technique consists in learning two separate action-value functions and computing the target action-values by taking the minimum over their outputs:

$$\hat{Q}^\pi_{\phi'_{min}}(s, a) = \min \left(Q^\pi_{\phi'_1}(s, a), Q^\pi_{\phi'_2}(s, a) \right). \tag{6}$$

The role of the minimization is to consistently produce overly pessimistic estimates of the target action-values, preventing positive bias accumulation. This approach is an empirically effective strategy for different benchmark tasks and has become standard practice.

4.2 The Uncertainty Regularizer

In this work, we take a more general approach for computing the target action-values. Particularly, we use a parameterized function, the *uncertainty regularizer* $p_\beta(s, a, \phi, \theta)$, for trying to approximate the bias in the critic's action-value predictions for on-policy actions. Thus, we specify an action-value estimator that penalizes the action-value estimates via the uncertainty regularizer:

$$\begin{aligned} &\hat{Q}^\pi_{\phi'}(s, a|\beta) = Q^\pi_{\phi'}(s, a) - p_\beta(s, a, \phi', \theta), \\ &\textit{where} \quad p_\beta(s, a, \phi', \theta) \approx Q^\pi_{\phi'}(s, a) - Q^\pi(s, a). \end{aligned} \tag{7}$$

A consequence of this formulation is that as long as p_β is unbiased *for on-policy actions*, so will the action-value estimator $\hat{Q}^\pi_{\phi'}$. Therefore, this would ensure that the expected target action-value bias is zero, preventing the positive bias accumulation phenomenon without requiring overly pessimistic action-value estimates. Based on these observations, we now specify a new method that learns an unbiased p_β and continuously adapts it to reflect changes in the critic and policy.

5 Generalized Pessimism Learning

Generalized Pessimism Learning (GPL) entails learning a particular uncertainty regularizer p_β, which we precisely specify in Eqn. 8. Our approach makes p_β adapt to changes in both actor and critic models throughout the RL process, to keep the target action-values unbiased. Hence, GPL allows for preventing positive bias accumulation without overly pessimistic targets. With any fixed penalty, we argue that it would be infeasible to maintain the expected target action-value bias close to zero due to the number of affecting parameters and stochastic factors in different RL experiments.

5.1 Uncertainty Regularizer Parameterization

We strive for a parameterization of the uncertainty regularizer that ensures low bias and variance estimation of the target action-values. Similar to prior works (Ciosek et al. 2019; Moskovitz et al. 2021), GPL uses a linear model of some measure of the epistemic uncertainty in the critic. Epistemic uncertainty represents the uncertainty from the model's learned parameters towards its possible predictions. Hence, when using expressive deep models, the areas of the state and action spaces where the critic's epistemic uncertainty is elevated are the areas in which the agent did not

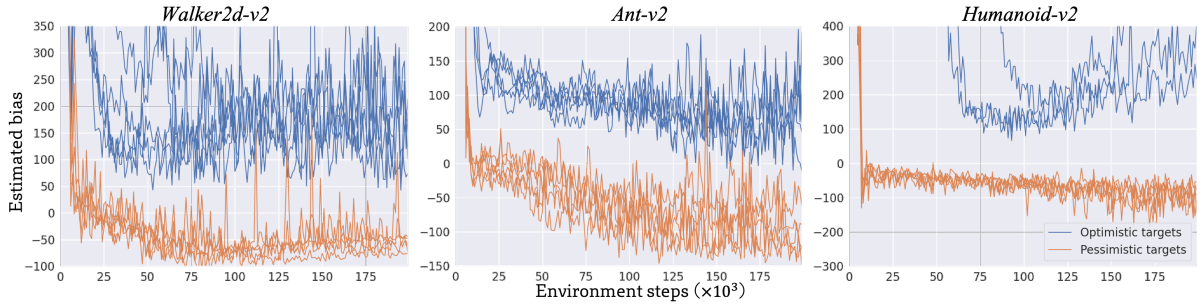

Figure 1: Recorded estimated bias for ten runs of two simple extensions of the *SAC* algorithm.

yet observe enough data to reliably predict its returns and, for this reason, the magnitude of the critic's error is expectedly higher. Consequently, if a policy yields behavior with high epistemic uncertainty in the critic, it is likely exploiting positive errors and overestimating its expected returns. As we use the policy to compute the TD-targets, the higher the uncertainty, the higher the expected positive bias.

We propose measuring epistemic uncertainty with the expected Wasserstein distance between the critic's predicted return distributions Z^π (Bellemare, Dabney, and Munos 2017). In our main experiments we consider the usual non-distributional case where we parameterize the critic with multiple action-value functions, in which case we view each action-value estimate as a Dirac delta function approximation of the return distribution, $Z^\pi_\phi(s,a) = \delta_{Q^\pi_\phi(s,a)}$. Our uncertainty regularizer then consists of linearly scaling the expected Wasserstein distance via a learnable parameter β:

$$p_\beta(s,a,\phi,\theta) = \beta \times \mathbb{E}_{a,\phi_1,\phi_2}\left[W(Z^\pi_{\phi_1}(s,a), Z^\pi_{\phi_2}(s,a))\right]. \quad (8)$$

We estimate the expectation in Eqn. 8 by learning $N \geq 2$ independent critic models with parameters $\{\phi_i\}_{i=1}^N$, and averaging the distances between the corresponding predicted return distributions. Notably, the Wasserstein distance has easy-to-compute closed forms for many popular distributions. For Dirac delta functions, it is equivalent to the distance between the corresponding locations, hence, $W(\delta_{Q^\pi_{\phi_1}(s,a)}, \delta_{Q^\pi_{\phi_2}(s,a)}) = |Q^\pi_{\phi_1}(s,a) - Q^\pi_{\phi_2}(s,a)|$.

Our quantification of epistemic uncertainty is an interpretable measure for any distributional critic. Moreover, for some fixed β, increasing the number of critics decreases the estimation variance but leaves the expected magnitude of the uncertainty regularizer unchanged. This is because the sample mean of the Wasserstein distances is always an unbiased estimate of Eqn. 8 for $N \geq 2$. Assuming we can approximately model the distribution of different action-value predictions with a Gaussian, we can show our penalty is proportional to the standard deviation of the distribution of action-value predictions. We can also restate clipped double Q-learning using our uncertainty regularizer with $N = 2$ and $\beta = 0.5$, allowing us to replicate its penalization effects for $N > 2$ by simply fixing β. In contrast, Ciosek et al. (2019) proposed the sample standard deviation of the action-value predictions to quantify epistemic uncertainty. However, the sample standard deviation does not have a clear generalization to arbitrary distributional critics and its expected magnitude is dependent on the number of models. All formal derivations are provided in (Cetin and Celiktutan 2021).

5.2 Dual TD-Learning

We hypothesize that the expected bias present in the action-value targets is highly dependent on several unaccountable factors from the stochasticity in the environment and the learning process. This *extends* recent results showing that the effectiveness of any bias penalty highly varies across tasks (Moskovitz et al. (2021), Fig. 4). We empirically validate our hypothesis by running multiple experiments with simple extensions of the SAC algorithm in different Gym environments and periodically recording estimates of the action-value bias by comparing the actual and estimated discounted returns. As shown in Figure 1, the bias in the predicted action-values notably varies across environments, agents, training stages, and even across different random seeds. These results validate our thesis that there is no *fixed* penalty able to account for the many sources of stochasticity in the RL process, even for a single task. Hence, this shows the necessity of learning p_β alongside the policy and critic to accurately counteract bias.

When using the uncertainty regularizer, we will denote the target bias for a transition as $B(s,a,s'|\beta)$, to highlight its dependency on the current value of β. Furthermore, note that $B(s,a,s'|\beta)$ would take on a positive or negative value in the cases where β yields either insufficient or excessive regularization, respectively. Therefore, to recover unbiased targets, we propose to optimize β as a dual variable by enforcing the expected target action-value bias to be zero:

$$\arg\min_\beta -\beta \times \mathbb{E}_{s,a,s'\sim D}\left[B(s,a,s'|\beta)\right]. \quad (9)$$

To estimate $B(s,a,s'|\beta)$, we use the property that for *arbitrary off-policy* actions the action-value estimates are not directly affected by the positive bias from the policy gradient optimization (Fujimoto, Van Hoof, and Meger 2018). Consequently, we make the assumption that Q^π_ϕ itself provides *initially* unbiased estimates of the expected returns, i.e., $\mathbb{E}_{s,a\sim D}[Q^\pi_\phi(s,a)] \approx \mathbb{E}_{s,a\sim D}[Q^\pi(s,a)]$. This assumption directly implies that any expected error in the Bellman relationship between $r + \gamma \mathbb{E}_{a'\sim\pi(s')}\left[\hat{Q}^\pi_{\phi'}(s',a'|\beta)\right]$ and $Q^\pi_\phi(s,a)$ is due to bias present in our action-value esti-

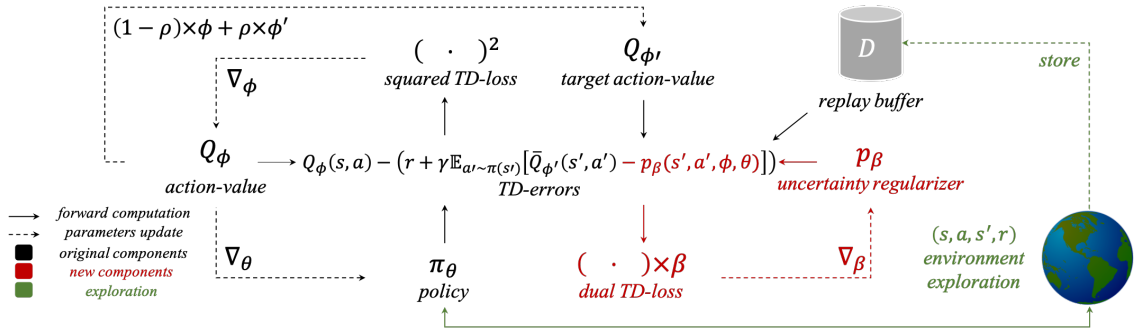

Figure 2: Schematic overview of the training and exploration processes involved in the GPL framework. The TD-errors play a *central role* both for updating the critic and for estimating the current bias to update the uncertainty regularizer.

mator. Hence, we propose to approximate $B(s,a,s'|\beta)$, with the expected difference between the current *on-policy* TD-targets and action-value predictions for *off-policy* actions:

$$B(s,a,s'|\beta) \approx r + \gamma \mathbb{E}_{a'\sim\pi(s')}\left[\hat{Q}^{\pi}_{\phi'}(s',a'|\beta)\right] - Q^{\pi}_{\phi}(s,a). \quad (10)$$

In practice, GPL alternates the optimizations of β for the current bias, and both actor and critic parameters, with the corresponding updated RL objectives. This is similar to the automatic exploration temperature optimization proposed by Haarnoja et al. (2018b), approximating dual gradient descent (Boyd and Vandenberghe 2004). We can estimate the current bias according to Eqn. 10 by simply negating the already-computed errors from the TD-loss, with trivial cost. Thus, we name this procedure *dual TD-learning*. We provide a schematic overview of GPL with dual TD-learning in Fig. 2.

Limitations. When using deep networks and approximate stochastic optimization, we recognize that the unbiasedness assumption of Eqn. 10 might not necessarily hold. Therefore, given initially biased action-value targets, some of the bias might propagate to the critic model, influencing the approximation in Eqn. 10. This property of our method makes it hard to provide a formal analysis beyond the tabular setting. However, in practice, we still find that GPL's performance and the optimization dynamics are robust to different levels of initial target bias, and that dual TD-learning appears to always improve over fixed penalties. See (Cetin and Celiktutan 2021) for an analysis of the empirical behavior of GPL with different initial values of β and, thus, different bias in the initial targets. An intuition for our results is that the relative difference between the off-policy and on-policy action-value predictions should always push β to counteract new bias stemming from model errors in the policy gradient action maximization, and thus improve over non-adaptive methods which are also affected by initial bias. We further validate dual TD-learning in (Cetin and Celiktutan 2021), comparing with optimizing β by minimizing the squared norm of the bias and by using the bandit-based optimization from TOP (Moskovitz et al. 2021). We also note that integrating GPL adds non-trivial complexity by introducing an entirely new optimization step which could be unnecessary for low-dimensional and easy-exploration problems. This inevitable limitation could further exacerbate the reproducibility of off-policy RL (Islam et al. 2017).

5.3 Pessimism Annealing for Directed Exploration

As described in Section 3, the policy learns to maximize the unbiased action-values predicted by action-value estimator $\hat{Q}^{\pi}_{\phi}$. Motivated by the principle of optimism in the face of uncertainty (Brafman and Tennenholtz 2002), we propose to make use of a new *optimistic* policy gradient objective:

$$\begin{aligned} J^{opt}_{\pi}(\theta) &= -\mathbb{E}_{s\sim D, a\sim\pi(s)}[\hat{Q}^{\pi_{opt}}_{\phi}(s,a|\beta)], \\ \textit{where} \quad \hat{Q}^{\pi_{opt}}_{\phi}(s,a|\beta) &= Q^{\pi}_{\phi}(s,a) - p_{\beta_{opt}}(s,a,\phi,\theta). \end{aligned} \quad (11)$$

This objective utilizes an *optimistic shifted uncertainty regularizer*, $p_{\beta_{opt}}$, calculated with parameter $\beta_{opt} = \beta - \lambda_{opt}$, for a decaying *optimistic shift value*, $\lambda_{opt} \geq 0$. This new objective trades off the traditional exploitative behavior of the policy with directed exploration. As λ_{opt} is large, π will be incentivized to perform actions for which the outcome has high epistemic uncertainty. Therefore, the agent will experience transitions that are increasingly informative for the critic but expectedly sub-optimal. Hence, we name the process of decaying λ_{opt} *pessimism annealing*, enabling to achieve improved exploration early on in training without biasing the policy's final objective.

6 Experiments

To evaluate the effectiveness of GPL, we integrate it with two popular off-policy RL algorithms. GPL itself introduces trivial computational and memory costs as it optimizes a single additional weight, re-utilizing the errors in the TD-loss to estimate the bias. Moreover, we implement the critic's ensemble as a single neural network, using linear non-fully-connected layers evenly splitting the nodes and dropping the weight connections between the splits. Practically, when evaluated under the same hardware, this results in our algorithm running more than 2.4 times faster than the implementation from Chen et al. (2021) while having a similar algorithmic complexity (see (Cetin and Celiktutan 2021)).

We show that GPL significantly improves the performance and robustness of off-policy RL, concretely surpassing prior algorithms and setting new state-of-the-art results. In our evaluation, we repeat each experiment with five random seeds and record both mean and standard deviation over

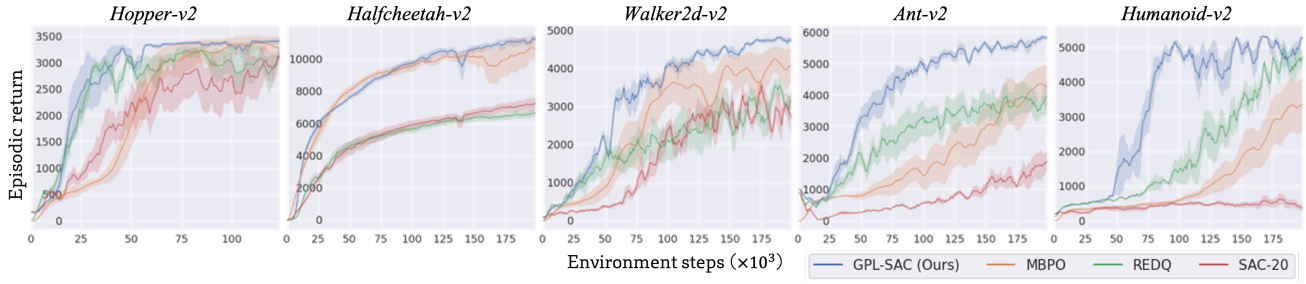

Figure 3: Performance curves for the considered complex Mujoco environments from OpenAI Gym, over five random seeds.

the episodic returns. Moreover, we validate statistical significance using tools from *Rliable* (Agarwal et al. 2021). In the extended version (Cetin and Celiktutan 2021), we report all details of our experimental settings and utilized hyper-parameters. We also provide comprehensive extended results analyzing the impact of all relevant design choices, testing several alternative implementations, and reporting all training times.

6.1 Continuous Control from Proprioception

GPL-SAC. First, we integrate GPL as a plug-in addition to *Soft Actor-Critic (SAC)* (Haarnoja et al. 2018a,b), a popular model-free off-policy algorithms that uses a weighted entropy term in its objective to incentivize exploration. Specifically, we only substitute SAC's clipped double Q-learning with our uncertainty regularizer, initialized with $\beta = 0.5$. Inline with the other considered state-of-the-art baselines (Chen et al. 2021; Janner et al. 2019), we use an increased ensemble size and update-to-data (UTD) ratio for the critic. We found both these choices necessary for sample-efficient learning in the evaluated experience regimes. We denote the resulting algorithm GPL-SAC. We would like to note that all other practices, unrelated to counteracting overestimation bias (such as learning the entropy bonus) were already present in SAC and are utilized by all baselines.

Baselines. We compare ***GPL-SAC*** with prior state-of-the-art model-free and model-based algorithms with similar or greater computational complexity, employing high UTD ratios: ***REDQ*** (Chen et al. 2021), state-of-the-art model-free algorithm on OpenAI Gym. This algorithm learns multiple action-value functions and utilizes clipped double Q-learning over a sampled pair of outputs to compute the critic's targets. ***MBPO*** (Janner et al. 2019), state-of-the-art, model-based algorithm on OpenAI Gym. This algorithm learns a large ensemble of world models with *Dyna*-style (Sutton 1991) optimization to train the policy. ***SAC-20***, simple SAC extension where with an increased UTD ratio of 20.

Results. We evaluate GPL-SAC compared to the described baselines on five of the more challenging Mujoco environments from OpenAI Gym (Brockman et al. 2016), involving complex locomotion problems from proprioception. We collect the returns over five evaluation episodes every 1000 environment steps. In Figure 3, we show the different performance curves. GPL-SAC is consistently the best performing algorithm on all environments, setting new state-of-the-art results for this benchmark at the time of writing. Moreover, the performance gap is greater for tasks with larger state and action spaces. We motivate this by noting that increased task-complexity appears to correlate with an increased stochasticity affecting target bias (Fig. 1), increasing the necessity for an adaptive counteraction strategy. Furthermore, as all baselines use fixed strategies to deal with overestimation bias, they also require overly pessimistic estimates of the returns to avoid instabilities. Hence, the resulting policies are likely overly conservative, hindering exploration and efficiency, with larger effects on higher-dimensional tasks. For instance, on *Humanoid*, GPL-SAC remarkably surpasses a score of 5000 after only 100K steps, more than $9\times$ faster than SAC and $2.5\times$ faster than REDQ.

Metric\Alg.	GPL-DrQ+An.	GPL-DrQ	DrQv2	CURL	SAC
Milestone	1.5M frames				
Average score	**640.20**	620.24	544.67	302.74	50.34
# Top scores	**11/12**	8/12	3/12	0/12	0/12
Milestone	3.0M frames				
Average score	**744.09**	720.29	670.95	318.38	59.38
# Top scores	**10/12**	7/12	4/12	0/12	0/12

Table 1: Results summary for the DeepMind Control Suite experiments. Per-task results in (Cetin and Celiktutan 2021).

6.2 Continuous Control from Pixels

GPL-DrQ. We also incorporate GPL to a recent version of *Data-regularized Q* (DrQv2) (Yarats et al. 2021), an off-policy, model-free algorithm achieving state-of-the-art performance for pixel-based control problems. DrQv2 combines image augmentation from DrQ (Yarats, Kostrikov, and Fergus 2021) with several advances such as n-step returns (Sutton and Barto 2018) and scheduled exploration noise (Amos et al. 2021). Again, we only substitute DrQv2's clipped double Q-learning with our uncertainty regularizer. To bolster exploration, we also integrate pessimism annealing from Section 5.3, with λ_{opt} linearly decayed from 0.5 to 0.0 together with the exploration noise in DrQv2. We leave the rest of the hyper-parameters and models unaltered to evaluate the generality of applying GPL. We name the re-

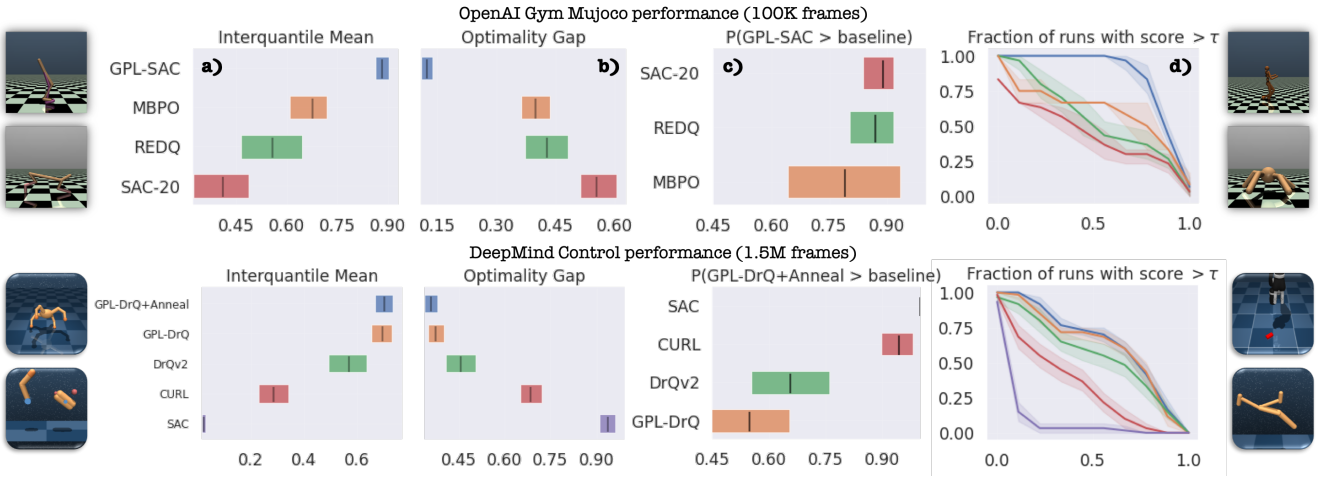

Figure 4: Aggregate performance metrics using the evaluation protocol proposed by *Rliable*.We report a) *interquantile mean* b) *optimality gap* c) *probability of improvement*, and d) full *performance profiles*. Ranges/shaded regions correspond to 95% CIs.

sulting algorithms GPL-DrQ and *GPL-DrQ+Anneal*.

Baselines. We compare our ***GPL-DrQ*** and ***GPL-DrQ+Anneal*** with model-free baselines for continuous control from pixels: The aforementioned state-of-the-art ***DrQv2***. ***CURL*** (Srinivas, Laskin, and Abbeel 2020), recent algorithm combining off-policy learning with a contrastive objective for representation learning. ***SAC***, simple integration of SAC with a convolutional encoder.

Results. We evaluate GPL-DrQ and GPL-DrQ+Anneal on the environments from the DeepMind Control Suite (Tassa et al. 2018) modified to yield pixel observations. We use the medium benchmark evaluation as described by Yarats et al. (2021), consisting of 12 complex tasks involving control problems with hard exploration and sparse rewards. In Table **??**, we report the mean returns obtained after experiencing 3M and 1.5M environment frames together with the number of tasks where each algorithm achieves *top scores* within half a standard deviation from the highest recorded return. For each run, we average the returns from 100 evaluation episodes collected in the 100K steps preceding each of these milestones. We provide the full per-environment results in (Cetin and Celiktutan 2021). Both GPL-DrQ and GPL-DrQ+Anneal significantly improve the performance of DrQv2 and all other baseline algorithms in the great majority of the tasks (7-11 out of 12). DrQv2 yields inconsistent returns on some tasks, likely due to a lack of exploration from its overly pessimistic critic. GPL generally appears to resolve this issue, while pessimism annealing further aids precisely in the tasks where under-exploration is more frequent. Overall, these results show both the generality and effectiveness of GPL for improving the current state-of-the-art through simple integrations, providing a novel framework to better capture and exploit bias.

6.3 Statistical Significance

To validate the statistical significance of the performance gains of GPL over the considered state-of-the-art algorithms, we follow the evaluation protocol proposed by Agarwal et al. (2021). For both Mujoco and DeepMind Control (DMC) benchmarks, we calculate various informative aggregate statistical measures of each algorithm halfway through training, normalizing each task's score within $[0, 1]$. Error bars/shaded regions correspond to the 95% stratified bootstrap confidence intervals (CIs) for each algorithm's performance (Efron 1992). As reported in Figure 4, in both benchmarks GPL achieves considerably *higher interquantile mean* (a) and *lower optimality gap* (b) than any of the baselines, with non-overlapping CIs. Furthermore, we analyze *probability of improvement* (c) from the Mann-Whitney U statistic (Mann and Whitney 1947), which reveals that GPL's improvements are *statistically meaningful* as per the Neyman-Pearson statistical testing criterion (Bouthillier et al. 2021). Lastly, we calculate the different *performance profiles* (d) (Dolan and Moré 2002), which show that GPL-based algorithms *stochastically dominate* all considered state-of-the-art baselines (Dror, Shlomov, and Reichart 2019). We believe these results convincingly validate the effectiveness and future potential of GPL.

7 Discussion and Future Work

We proposed Generalized Pessimism Learning, a strategy that adaptively *learns* a penalty to recover an unbiased performance objective for off-policy RL. Unlike traditional methods, GPL achieves training stability without necessitating overly pessimistic estimates of the target returns, thus, improving convergence and exploration. We show that integrating GPL with modern algorithms yields state-of-the-art results for both proprioceptive and pixel-based control tasks. Moreover, GPL's penalty has a natural generalization to different distributional critics and variational representations of the weights posterior. Hence, our method has the potential to facilitate research in off-policy reinforcement learning, going beyond action-value functions and model ensembles. Future extensions could also have implications for offline RL, a problem setting particularly sensitive to overestimation.

Acknowledgments

Edoardo Cetin would like to acknowledge the support from the Engineering and Physical Sciences Research Council [EP/R513064/1]. Oya Celiktutan would also like to acknowledge the support from the LISI Project, funded by the Engineering and Physical Sciences Research Council [EP/V010875/1]. Furthermore, we thank Toyota Motor Europe and Toyota Motor Corporation for providing support towards funding the utilized computational resources.

References

Agarwal, R.; Schwarzer, M.; Castro, P. S.; Courville, A. C.; and Bellemare, M. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34.

Amos, B.; Stanton, S.; Yarats, D.; and Wilson, A. G. 2021. On the model-based stochastic value gradient for continuous reinforcement learning. In *Learning for Dynamics and Control*, 6–20. PMLR.

Anschel, O.; Baram, N.; and Shimkin, N. 2017. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International conference on machine learning*, 176–185. PMLR.

Bechtle, S.; Molchanov, A.; Chebotar, Y.; Grefenstette, E.; Righetti, L.; Sukhatme, G.; and Meier, F. 2021. Meta learning via learned loss. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 4161–4168. IEEE.

Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 449–458. PMLR.

Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; et al. 2021. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3: 747–769.

Boyd, S.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Brafman, R. I.; and Tennenholtz, M. 2002. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Cetin, E.; and Celiktutan, O. 2021. Learning pessimism for robust and efficient off-policy reinforcement learning. *arXiv preprint arXiv:2110.03375*.

Chen, R. Y.; Sidor, S.; Abbeel, P.; and Schulman, J. 2017. UCB exploration via Q-ensembles. *arXiv preprint arXiv:1706.01502*.

Chen, X.; Wang, C.; Zhou, Z.; and Ross, K. 2021. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*.

Chua, K.; Calandra, R.; McAllister, R.; and Levine, S. 2018. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *arXiv preprint arXiv:1805.12114*.

Ciosek, K.; Vuong, Q.; Loftin, R.; and Hofmann, K. 2019. Better exploration with optimistic actor-critic. *arXiv preprint arXiv:1910.12807*.

Dolan, E. D.; and Moré, J. J. 2002. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91(2): 201–213.

Dror, R.; Shlomov, S.; and Reichart, R. 2019. Deep dominance-how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2773–2785.

Efron, B. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, 569–593. Springer.

Fujimoto, S.; Van Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018a. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. 2018b. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Hasselt, H. 2010. Double Q-learning. *Advances in neural information processing systems*, 23: 2613–2621.

Hessel, M.; Modayil, J.; Van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M.; and Silver, D. 2018. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Islam, R.; Henderson, P.; Gomrokchi, M.; and Precup, D. 2017. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. *arXiv preprint arXiv:1708.04133*.

Janner, M.; Fu, J.; Zhang, M.; and Levine, S. 2019. When to trust your model: Model-based policy optimization. *arXiv preprint arXiv:1906.08253*.

Kumar, A.; Gupta, A.; and Levine, S. 2020. Discor: Corrective feedback in reinforcement learning via distribution correction. *arXiv preprint arXiv:2003.07305*.

Kuznetsov, A.; Grishin, A.; Tsypin, A.; Ashukha, A.; and Vetrov, D. 2021. Automating Control of Overestimation Bias for Continuous Reinforcement Learning. *arXiv preprint arXiv:2110.13523*.

Kuznetsov, A.; Shvechikov, P.; Grishin, A.; and Vetrov, D. 2020. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, 5556–5566. PMLR.

Lan, Q.; Pan, Y.; Fyshe, A.; and White, M. 2020. Maxmin q-learning: Controlling the estimation bias of q-learning. *arXiv preprint arXiv:2002.06487*.

Lee, D.; Defourny, B.; and Powell, W. B. 2013. Bias-corrected q-learning to control max-operator bias in q-learning. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 93–99. IEEE.

Lyu, J.; Ma, X.; Yan, J.; and Li, X. 2022. Efficient continuous control with double actors and regularized critics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7655–7663.

Mahmood, A. R.; Korenkevych, D.; Vasan, G.; Ma, W.; and Bergstra, J. 2018. Benchmarking reinforcement learning algorithms on real-world robots. *arXiv preprint arXiv:1809.07731*.

Mann, H. B.; and Whitney, D. R. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1): 50 – 60.

Mannor, S.; Simester, D.; Sun, P.; and Tsitsiklis, J. N. 2007. Bias and variance approximation in value function estimates. *Management Science*, 53(2): 308–322.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Moskovitz, T.; Parker-Holder, J.; Pacchiano, A.; Arbel, M.; and Jordan, M. I. 2021. Tactical Optimism and Pessimism for Deep Reinforcement Learning. arXiv:2102.03765.

Oh, J.; Hessel, M.; Czarnecki, W. M.; Xu, Z.; van Hasselt, H.; Singh, S.; and Silver, D. 2020. Discovering reinforcement learning algorithms. *arXiv preprint arXiv:2007.08794*.

Pan, L.; Cai, Q.; and Huang, L. 2020. Softmax deep double deterministic policy gradients. *Advances in Neural Information Processing Systems*, 33: 11767–11777.

Pendrith, M. D.; Ryan, M. R.; et al. 1997. *Estimator variance in reinforcement learning: Theoretical problems and practical solutions*. University of New South Wales, School of Computer Science and Engineering.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; and Riedmiller, M. 2014. Deterministic policy gradient algorithms. In *International conference on machine learning*, 387–395. Pmlr.

Srinivas, A.; Laskin, M.; and Abbeel, P. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*.

Sutton, R. S. 1991. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4): 160–163.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057–1063.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; Casas, D. d. L.; Budden, D.; Abdolmaleki, A.; Merel, J.; Lefrancq, A.; et al. 2018. Deepmind control suite. *arXiv preprint arXiv:1801.00690*.

Thrun, S.; and Schwartz, A. 1993. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*, 255–263. Hillsdale, NJ.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. IEEE.

Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*.

Wang, T.; and Ba, J. 2019. Exploring model-based planning with policy networks. *arXiv preprint arXiv:1906.08649*.

Xu, Z.; van Hasselt, H.; Hessel, M.; Oh, J.; Singh, S.; and Silver, D. 2020. Meta-gradient reinforcement learning with an objective discovered online. *arXiv preprint arXiv:2007.08433*.

Yarats, D.; Fergus, R.; Lazaric, A.; and Pinto, L. 2021. Mastering Visual Continuous Control: Improved Data-Augmented Reinforcement Learning. *arXiv preprint arXiv:2107.09645*.

Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *International Conference on Learning Representations*.

Zhang, Z.; Pan, Z.; and Kochenderfer, M. J. 2017. Weighted Double Q-learning. In *IJCAI*, 3455–3461.