OVD-Explorer: Optimism Should Not Be the Sole Pursuit of Exploration in Noisy Environments

Jinyi Liu[1*], **Zhi Wang**[2*], **Yan Zheng**[1†], **Jianye Hao**[1], **Chenjia Bai**[3], **Junjie Ye**[2], **Zhen Wang**[4], **Haiyin Piao**[4], **Yang Sun**[5]

[1]College of Intelligence and Computing, Tianjin University
[2]Independent Researcher
[3]Shanghai AI Laboratory
[4]Northwestern Polytechnical University
[5]SADRI Institute
{jyliu, yanzheng, jianye.hao}@tju.edu.cn, zhiwoong@163.com, baichenjia@pjlab.org.cn, kourenmu@gmail.com,
{w-zhen, haiyinpiao}@nwpu.edu.cn, yang.sun2010@gmail.com

Abstract

In reinforcement learning, the optimism in the face of uncertainty (OFU) is a mainstream principle for directing exploration towards less explored areas, characterized by higher uncertainty. However, in the presence of environmental stochasticity (noise), purely optimistic exploration may lead to excessive probing of high-noise areas, consequently impeding exploration efficiency. Hence, in exploring noisy environments, while optimism-driven exploration serves as a foundation, prudent attention to alleviating unnecessary over-exploration in high-noise areas becomes beneficial. In this work, we propose Optimistic Value Distribution Explorer (OVD-Explorer) to achieve a noise-aware optimistic exploration for continuous control. OVD-Explorer proposes a new measurement of the policy's exploration ability considering noise in optimistic perspectives, and leverages gradient ascent to drive exploration. Practically, OVD-Explorer can be easily integrated with continuous control RL algorithms. Extensive evaluations on the MuJoCo and GridChaos tasks demonstrate the superiority of OVD-Explorer in achieving noise-aware optimistic exploration.

Introduction

Efficient exploration is crucial for improving the reinforcement learning (RL) efficiency and ultimate policy performance (Sutton and Barto 2018), and many exploration strategies have been proposed in the literatures (Lillicrap et al. 2016; Osband et al. 2016; Chen et al. 2017; Ciosek et al. 2019). Most of them follows the *Optimism in the Face of Uncertainty* (OFU) principle (Auer, Cesa-Bianchi, and Fischer 2002) to guide exploration optimistically towards the area with high uncertainty (Chen et al. 2017; Ciosek et al. 2019). Conceptually, OFU-based methods regard the uncertainty as the ambiguity caused by insufficient exploration, and is high at those state-action pairs seldom visited, referred to as *epistemic uncertainty* (Osband et al. 2016).

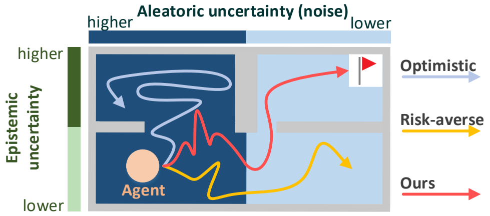


Figure 1: An intuitive example. The agent learns to move to the flag in a room filled with noise, and the noise in the left side is higher. The optimistic exploration strategy may overly explore the noisy area, and the risk-averse policy indiscriminately avoids noisy areas but explores insufficiently.

Another kind of uncertainty existed in RL is known as *aleatoric uncertainty*, caused by the randomness in the environment or policy, and referred to as *noise* (Kirschner and Krause 2018) or risk (Dabney et al. 2018a). The noise is ubiquitous in real world. For example, unpredictable wind shifts the trajectory after an robot's action, and rough ground changes the force point of objects, etc. However, overly visiting such noisy areas may cause severely unstable state transitions (the *Optimistic* arrow in the intuitive example in Fig. 1), thus is detrimental to the learning efficiency (Clements et al. 2019). For this, risk-averse policy is proposed to avoid visiting the areas with high aleatoric uncertainty estimation (Dabney et al. 2018b,a). Typical approaches use Conditional Variance at Risk (CVaR) to calculate a conservative value estimation and guide policy learning for easing the negative effect of the noise (Dabney et al. 2018a). However, indiscriminately avoiding noise may also yield no performance guarantee due to excessively conservation (the *risk-averse* arrow in Fig. 1).

Therefore, a more reasonable approach is integrating the risk-averse policy and optimistic exploration, guiding the agent to optimistically exploring the whole areas, while avoiding overly exploring the areas with high noise, like the

[*]These authors contributed equally.
[†]Corresponding author: Yan Zheng (yanzheng@tju.edu.cn)

Ours arrow in Fig. 1. Note that overly exploring noisy areas may damage performance, but moderate exploration of such area is necessary, which should be ensured by the ability of optimistic exploration. The similar concern has been demonstrated to be effective under discrete control tasks (Nikolov et al. 2019; Clements et al. 2019). However, for continuous control tasks, such concern has not yet been investigated well. A natural way to apply discrete control algorithms to solve continuous control problems is to discretize the continuous action space, but it suffers from the scalability issue due to the exponentially increasing discretized actions (Antos, Munos, and Szepesvári 2007; Li et al. 2021), and it may throw away crucial information in action space causing its performance to be compromised (Lillicrap et al. 2016; Tang and Agrawal 2020). Thus, designing such an optimistic exploration strategy that can avoid overly exploring noisy area for continuous action space is required.

In this work, we propose OVD-Explorer, a noise-aware optimistic exploration strategy that applies to continuous control tasks for the first time. Specifically, we propose a new policy's exploration ability measurement, quantifying both the ability of avoiding noise and pursuing optimisticity during exploration. To capture the noise, the value distribution is modelled. Further, to guide optimistic exploration, the upper bound distribution of return is approximated using Optimistic Value Distribution (OVD), representing the best returns that the policy can reach. Then we quantitatively measure such ability using OVD, and generates the behavior policy by maximizing the exploration ability measurement, thus names our approach as Optimistic Value Distribution Explorer (OVD-Explorer).

To make OVD-Explorer tractable for continuous control, we generate the behavior policy using a gradient-based approach, and propose a scheme to incorporate it with policy-based RL algorithms. Practically, we demonstrate the exploration benefits based on SAC (Haarnoja et al. 2018), a well-performed continuous RL algorithm. Evaluations on various continuous RL tasks, including the GridChaos, MuJoCo tasks and their stochastic version, are conducted. The results demonstrate the effectiveness of OVD-Explorer in achieving an optimistic exploration that avoids overly exploring noisy areas, leading to a better performance.

Related Work

In this work, we consider the exploration strategy under the OFU principle (Auer, Cesa-Bianchi, and Fischer 2002), and aim to a noise-aware optimistic exploration strategy.

Overview of exploration approaches. Basic exploration strategies always lead to undirected exploration through random perturbations (Lillicrap et al. 2016; Sutton and Barto 2018; Haarnoja et al. 2018). With the increasing emphasis on exploration efficiency in RL, various exploration methods have been developed (Hao et al. 2023). One kind of methods uses intrinsic motivation to stimulate agent to explore (Martin et al. 2017; Bellemare et al. 2016; Savinov et al. 2019; Houthooft et al. 2016; Badia et al. 2020; Yuan et al. 2023). Some other methods, originating from tracking uncertainty, guide exploration under the OFU principle (Thompson 1933; Osband et al. 2016; Nikolov et al. 2019; Ciosek et al. 2019; Pathak, Gandhi, and Gupta 2019; Bai et al. 2021b,a). The key of OFU-based exploration methods is modeling the epistemic uncertainty (Osband et al. 2016; Gal and Ghahramani 2016; Qiu et al. 2022). Specifically, we use ensemble (Osband et al. 2016) for estimating epistemic uncertainty.

The issue of overly exploring noisy areas. There is another uncertainty in RL system, i.e., aleatoric uncertainty (a.k.a. noise (Kirschner and Krause 2018) or risk (Dabney et al. 2018a)), captured by return distribution (Bellemare, Dabney, and Munos 2017; Dabney et al. 2018a,b). Overly exploring the areas with high noise could make learning unstable and inefficient, thus many works seek a conservative and noise-averse (or risk-averse) policy to make the policy stable (Dabney et al. 2018a; Ma et al. 2020; Bai et al. 2022). Nevertheless, conservative alone without advanced exploration could induce low exploration efficiency, and exploration without avoiding noise could make interaction risky. Thus some recent works produce optimistic exploration strategies considering risk (Mavrin et al. 2019; Nikolov et al. 2019). However, such methods are complicated when deriving a behavior policy and only limited to discrete control.

Indeed, addressing noise in exploration poses a challenge for well-performing continuous RL algorithms (Haarnoja et al. 2018; Ma et al. 2020). While exploration strategies like OAC (Ciosek et al. 2019) are designed following OFU principle, guided by the upper bound of Q estimation, they overlook the potential impact of noise. This oversight can lead to misguided exploration, hampering the learning process. To address that, we propose OVD-Explorer to guide agent to explore optimistically, while avoiding overly exploring the noisy areas, improving the robustness of exploration especially facing heteroscedastic noise.

Preliminaries

Distributional Value Estimation

To capture the environment noise, we use quantile regression (Dabney et al. 2018b) to formulate Q-value distribution. Q-value distribution, represented by the quantile random variable Z, maps the state-action pair to a uniform probability distribution supported on the return values at all corresponding quantile fractions. Given state-action pair (s,a), we denote the i-th quantile fraction as τ_i, and the value at τ_i as $Z_{\tau_i}(s,a)$, where $\tau_i \in [0,1]$.

Based on the Bellman operator (Watkins and Dayan 1992), the distributional Bellman operator (Bellemare, Dabney, and Munos 2017) $\mathcal{T}_D^\pi$ under policy π is given as:

$$\mathcal{T}_D^\pi Z(s,a) \overset{D}{=} R(s,a) + \gamma Z(s',a'), a' \sim \pi(\cdot|s'). \tag{1}$$

Notice that $\mathcal{T}_D^\pi$ operates on random variables, $\overset{D}{=}$ denotes that distributions on both sides have equal probability laws. Based on operator $\mathcal{T}_D^\pi$, QR-DQN (Dabney et al. 2018b) trains quantile estimations via the quantile regression loss (Koenker and Hallock 2001), which is denoted as:

$$\mathcal{L}_{QR}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}[\rho_{\hat{\tau}_i}(\delta_{i,j})], \tag{2}$$

where θ and $\bar{\theta}$ is the parameters of the value distribution estimator and its target network, respectively, TD error $\delta_{i,j} = R(s,a) + \gamma Z_{\hat{\tau}_i}(s',a';\bar{\theta}) - Z_{\hat{\tau}_j}(s,a;\theta)$, the quantile Huber loss $\rho_\tau(u) = u * |\tau - \mathbb{1}_{u<0}|$, and $\hat{\tau}_i$ means the quantile midpoints, which is defined as $\hat{\tau}_i = \frac{\tau_{i-1}+\tau_i}{2}$.

Distributional Soft Actor-Critic

Distributional Soft Actor-Critic (DSAC) (Ma et al. 2020) seamlessly integrates distributional RL with Soft Actor-Critic (SAC) (Haarnoja et al. 2018). Basically, based on the Eq. 1, the distributional soft Bellman operator $\mathcal{T}^\pi_{DS}$ is defined considering the maximum entropy RL as follows:

$$\mathcal{T}^\pi_{DS} Z(s,a) \overset{D}{=} R(s,a) + \gamma[Z(s',a') - \alpha \log \pi(a'|s')], \tag{3}$$

where $a' \sim \pi(\cdot|s'), s' \sim \mathcal{P}(\cdot|s,a)$. Then, to overcome overestimation of Q-value estimation, DSAC extends clipped double Q-Learning (Fujimoto, van Hoof, and Meger 2018), maintaining two critic estimators $\theta_k, k = 1,2$. Thus, the quantile regression loss differs from Eq. 2 on TD loss of θ_l:

$$\begin{aligned}\delta^l_{i,j} =& R(s,a) + \gamma[\min_{k=1,2} Z_{\hat{\tau}_i}(s',a';\bar{\theta}_k) - \alpha \log \pi(a'|s';\bar{\phi})] \\ &- Z_{\hat{\tau}_j}(s,a;\theta_l),\end{aligned} \tag{4}$$

where $\bar{\theta}$ and $\bar{\phi}$ represents their target networks respectively. The objective of actor is the same as SAC,

$$\mathcal{J}_\pi(\phi) = \mathop{\mathbb{E}}_{\substack{s\sim\mathcal{D}\\ \epsilon\sim\mathcal{N}}}[\, \log \pi(f_\phi(s,\epsilon)|s) - Q(s, f_\phi(s,\epsilon);\theta)\,], \tag{5}$$

where $\mathcal{D}$ is the replay buffer, $f_\phi(s,\epsilon)$ means sampling action with re-parameterized policy and ϵ is a noise vector sampled from any fixed distribution, like standard spherical Gaussian. Here, Q value is the minimum value of the expectation on totally N quantile fractions, as

$$Q(s,a;\theta) = \min_{k=1,2} \mathbb{E}_{i\sim\mathcal{U}(1,N)} Z_{\hat{\tau}_i}(s,a;\theta_k). \tag{6}$$

Optimistic Value Distribution Explorer

In noisy environments, a more efficient exploration strategy entails being noise-aware optimistic, especially to avoid excessive exploration in noisy areas. Over exploration towards the areas with high noise may damage the exploration performance, but indiscriminately avoiding visiting such areas could also compromise performance due to excessively conservation and insufficient exploration. In this work, we propose OVD-Explorer to achieve a noise-aware optimistic exploration in continuous RL. Accordingly, the key insight, theoretical derivation and formulation, and analysis of OVD-Explorer are outlined below.

Noise-aware Optimistic Exploration

Several previous optimistic exploration strategies for continuous control typically estimate the upper bound of Q-value, and guide exploration by maximizing this upper bound (Ciosek et al. 2019; Lee et al. 2021). While such upper bounds provide valuable guidance for optimistic exploration, they fail to capture the noise in the environment. To address that, we propose to incorporate the value distribution into the definition of the upper bound to capture noise, and define the upper bound distribution of Q-value. Additionally, we introduce a novel exploration ability measurement for policy distribution $\pi(\cdot)$ using such upper bound distribution, to characterize a policy's ability for noise-aware optimistic exploration. We then derive the behavior (exploration) policy by maximizing this ability measurement.

Firstly, we define the upper bound distribution of the Q-value at each state-action pair as $\bar{Z}^\pi(s,a)$.

Definition 1 (The upper bound distribution of Q-value) *Given state-action pair (s,a), the upper bound distribution of its Q-value, denoted as $\bar{Z}^\pi(s,a)$, is a value distribution satisfying that at each quantile fraction $\tau_i \in [0,1]$, its value $\bar{Z}^\pi_{\tau_i}(s,a)$ is the upper bound of possible estimations:*

$$\bar{Z}^\pi_{\tau_i}(s,a) := \sup_\theta Z^\pi_{\tau_i}(s,a;\theta), \tag{7}$$

where θ represents different estimators of value distribution, $Z^\pi_{\tau_i}(s,a;\theta)$ represent the value at quantile fraction τ_i of the value distribution estimation $Z^\pi(s,a;\theta)$.

We expect an effective exploration policy to approach the upper bound of Q-value. With the distribution-based definition of such an upper bound, we then employ mutual information to evaluate the correlation between the policy distribution and the upper bound distribution, which forms the basis for our definition of exploration capability. Overall, given current state s, we quantitatively measure the policy's exploration ability, denoted as $\mathbf{F}^\pi(s)$, by the integral of mutual-information between policy $\pi(\cdot|s)$ and the upper bound distributions of Q-value over the action space:

$$\mathbf{F}^\pi(s) = \int_{a'} \mathbf{MI}(\bar{Z}^\pi(s,a');\pi(\cdot|s)|s)\,\mathrm{d}a' \tag{8}$$

where $a' \in \mathbf{A}$ denotes any legal action. Now we state how to approximate the exploration ability in Proposition 2.

Proposition 2 *The mutual information in Eq. 8 at state s can be approximated as:*

$$\mathbf{F}^\pi(s) \approx \frac{1}{C} \mathop{\mathbb{E}}_{\substack{a\sim\pi(\cdot|s)\\ \bar{z}(s,a)\sim\bar{Z}^\pi(s,a)}} \left[\Phi_{Z^\pi}(\bar{z}(s,a)) \log \frac{\Phi_{Z^\pi}(\bar{z}(s,a))}{C}\right]. \tag{9}$$

$\Phi_x(\cdot)$ is the cumulative distribution function (CDF) of random variable x, $\bar{z}(s,a)$ is the sampled upper bound of return from its distribution $\bar{Z}^\pi(s,a)$ following policy π, Z^π describes the current return distribution of the policy π, and C is a constant (see proof in the Appendix).

Note that, to optimize the above objective, we need to formulate two critical components at any state-action pair (s,a) under policy π: ❶ the return distribution $Z^\pi(s,a)$ and ❷ the upper bound distribution of return $\bar{Z}^\pi(s,a)$. We detail the formulations in Sec. 4.2.

Proposition 2 reveals that $\mathbf{F}^\pi(s)$ is only proportional to the CDF value $\Phi_{Z^\pi}(\bar{z}(s,a))$, which is also proportional to $\bar{z}(s,a)$, an upper bound of Q-value, thus a higher $\mathbf{F}^\pi(s)$ represents the higher ability of optimistic exploration, following traditional OFU principle. Meanwhile, $\Phi_{Z^\pi}(\bar{z}(s,a))$

increases as the variance of current return distribution Z^π becomes lower, thus the higher $\mathbf{F}^\pi(s)$ means the higher ability of exploring towards the areas with low variance of return distribution, i.e., low noise. A more detailed analysis is given in Sec. 4.3.

Given current state s, OVD-Explorer aims to find the behavior policy π_E which has the best exploration ability $\mathbf{F}^\pi(s)$ in the policy space Π, as follows:

$$\pi_E = \arg\max_{\pi\in\Pi} \mathbf{F}^\pi(s). \tag{10}$$

For continuous action space, generating the analytical solution π_E in Eq. 10 is intractable. Hence, we propose to perform the gradient ascent based on the policy π, so as to iteratively deriving a behavior policy with high ability of noise-aware optimistic exploration. In short, given the policy π_ϕ parameterized by ϕ, we calculate the derivative $\nabla_\phi \Phi_{Z^{\pi_\phi}}(\bar{z}(s,a))$ and guide ϕ along the gradient direction to improve the exploration ability (more details in Sec. 5).

Distributions of Return's Upper Bound and Return

Now, we introduce the formulation of the return distribution $Z^\pi(s,a)$ and its upper bound distribution $\bar{Z}^\pi(s,a)$.

In specific, we use two value distribution estimators $\hat{Z}(s,a;\theta_1)$ and $\hat{Z}(s,a;\theta_2)$ parameterized by θ_1 and θ_2, as ensembles to formulate $\bar{Z}^\pi$ and Z^π differently. Unless stated otherwise, (s,a) is omitted hereafter to ease notation. As mentioned earlier, two types of uncertainties are involved, epistemic uncertainty and aleatoric uncertainty (noise), denoted as $\sigma^2_{\text{epistemic}}(s,a)$ and $\sigma^2_{\text{aleatoric}}(s,a)$, respectively. Due to space limitations, the computational details regarding uncertainty value are presented in detail in the Appendix.

Formulation of $\bar{Z}^\pi$. The $\bar{Z}^\pi$ denotes the upper bound distribution of return that policy π can reach. We propose Gaussian distribution with optimistic mean value $\mu_{\bar{Z}}(s,a)$ for formulation to formulate $\bar{Z}^\pi(s,a)$ as follows, and accordingly refer to it as *Optimistic Value Distribution* (OVD):

$$\bar{Z}^\pi(s,a) \sim \mathcal{N}(\mu_{\bar{Z}}(s,a), \sigma^2_{\text{aleatoric}}(s,a)), \tag{11}$$

where $\sigma^2_{\text{aleatoric}}(s,a)$ is its variance. Notable, Chen et al. (2017) discovers the optimisticity is beneficial for better estimating the upper bound, which motivates us to optimistically estimate $\mu_{\bar{Z}}(s,a)$ as averaged upper bound value of return by considering epistemic uncertainty as follows:

$$\begin{aligned} &\mu_{\bar{Z}}(s,a) = \mu(s,a) + \beta\sigma_{\text{epistemic}}(s,a), \\ &\text{s.t. } \mu(s,a) = \mathbb{E}_{i\sim\mathcal{U}(1,N)}\mathbb{E}_{k=1,2}\hat{Z}_{\tau_i}(s,a;\theta_k) \end{aligned} \tag{12}$$

where $\mu(s,a)$ represents the expected Q-value estimation, and uncertainty value is weighted by β, $\mathcal{U}$ is uniform distribution, N is the number of quantiles, and $\hat{Z}_{\tau_i}(s,a;\theta_k)$ is the value of the i-th quantile drawn from $\hat{Z}(s,a;\theta_k)$.

Leveraging optimistic value estimations together with explicitly modeling the noise, the upper bound distribution $\bar{Z}^\pi$ can be comprehensively formulated, known as OVD. Such an optimistic distribution can guides effectively optimistic exploration for OVD-Explorer.

Formulation of Z^π. Z^π estimates the return distribution obtained following policy π. Following Fujimoto, van Hoof, and Meger (2018), to alleviate overestimation, we formulate Z^π in a pessimistic way. In practice, Z^π can be measured in two ways. First, similar to formulating $\bar{Z}^\pi$ in Eq. 11, Z^π can also be formulated as Gaussian distribution as follows:

$$\begin{aligned} &Z^\pi(s,a) \sim \mathcal{N}(\mu_{Z^\pi}(s,a), \sigma^2_{\text{aleatoric}}(s,a)), \\ &s.t. \quad \mu_{Z^\pi}(s,a) = \mu(s,a) - \beta\sigma_{\text{epistemic}}(s,a), \end{aligned} \tag{13}$$

where $\mu(s,a)$, $\sigma_{\text{aleatoric}}(s,a)$ and $\sigma_{\text{epistemic}}(s,a)$ are the same defined in Eq. 12. Differently, $\sigma_{\text{epistemic}}(s,a)$ is subtracted from $\mu(s,a)$ to reveal the pessimistic estimation.

Another way is to formulate Z^π pessimistically as multivariate uniform distribution as:

$$\begin{aligned} &Z^\pi(s,a) \sim \mathcal{U}\{z_i^\pi(s,a;\theta)\}_{i=1,\ldots,N}, \\ &s.t.\ z_i^\pi(s,a;\theta) = \min_{k=1,2} \hat{Z}_{\tau_i}(s,a;\theta_k), \end{aligned} \tag{14}$$

where each quantile value $z_i^\pi(s,a;\theta)$ is the minimum estimated value among ensemble estimators (i.e., $\hat{Z}_{\tau_i}(s,a;\theta_k)$.

OVD-Explorer formulates the value distribution Z^π in two ways using Eq. 13 and Eq. 14, abbreviated in the following as OVDE_G and OVDE_Q, respectively. Intuitively, Gaussian distribution is expected to help more when the environment randomness follows a unimodal distribution, and multivariate uniform distribution is more flexible and suitable for scenarios with multi-modal distributions.

Analysis of OVD-Explorer

To analyzes how OVD-Explorer *optimistically* explores the whole areas and performs *noise-aware* exploration at the same time, an intuitive example involving two actions is adopted. According to Proposition 2, the behavior policy of OVD-Explorer maximizes $\mathbf{F}^\pi(s)$, which is proportional to the CDF value $\Phi_{Z^\pi}(\bar{z}(s,a))$. Supposing an agent need to select an actions between a_1 and a_2 to explore, Fig. 2(a) and (b) illustrate the CDF value (shaded area) for each action.

In these cases, the value distribution $Z^\pi(s,a)$ is specified as Gaussian (Eq. 13), and the sampled optimistic value $\bar{z}(s,a)$ is specified as the mean of OVD $\mu_{\bar{Z}}(s,a)$ (Eq. 12). At state s, we assume that the means of Z^π at actions a_1 and a_2 are the same for ease of clarification.

Optimistic exploration: Fig. 2(a) illustrates how OVD-Explorer achieves an optimistic exploration. Assuming the noise at a_1 and a_2 is equal, but epistemic uncertainty is higher at a_1, then $\mu_{\bar{Z}}(s,a_1) > \mu_{\bar{Z}}(s,a_2)$ and the CDF value is larger at a_1. Therefore, OVD-Explorer prefers a_1 with high epistemic uncertainty for an optimistic exploration.

Noise-aware exploration: Fig. 2(b) demonstrates how OVD-Explorer behave noise-aware to avoid the area with higher noise (aleatoric uncertainty). When both actions have equal epistemic uncertainty, $\mu_{\bar{Z}}(s,a_1) = \mu_{\bar{Z}}(s,a_2)$, and noise is lower at a_1 (PDF curve of $Z^\pi(s,a_1)$ is "thinner and taller"), the CDF value will be larger at a_1. In such a case, OVD-Explorer prefers action a_1 with lower aleatoric uncertainty (i.e., lower noise) for a noise-aware exploration.

Adaptivity. In the early training, the noise estimations of all actions are nearly identical (Fig. 2(a)), and the exploration is primarily guided by epistemic uncertainty. After sufficient training, the epistemic uncertainty decreases,

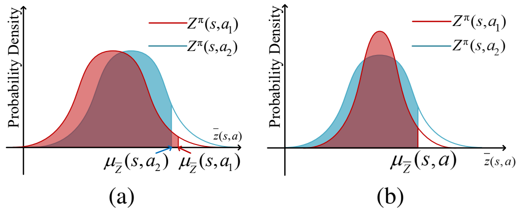


Figure 2: How OVD-Explorer explores (a) optimistically about epistemic uncertainty, (b) pessimistically about noise.

while the noise estimation converges to true environment randomness (Fig. 2(b)). At this point, exploration strategy tends to be noise-avoiding. OVD-Explorer seeks an adaptive balance of noise-aware optimistic exploration throughout the exploration process, which is a significant advantage compared to other OFU-based methods.

OVD-Explorer for RL Algorithms

For continuous RL, solving the argmax operator in Eq. 10 is intractable. In this section, aiming at maximizing $\mathbf{F}^\pi(s)$, we use a gradient-based approach to generate the behavior policy, and incorporate it with policy-based algorithms.

We denote the policy learned by any policy-based algorithm as π_ϕ, parameterized by ϕ. To avoid the gap between π_ϕ and the training data collected by behavior policy π_E, we derive π_E in the vicinity of π_ϕ. Then, aiming at maximizing $\mathbf{F}^{\pi_\phi}(s)$, we derive its gradient regarding the policy $\nabla_\phi \mathbf{F}^{\pi_\phi}(s)$ using automatic differentiation and generate behavior policy π_E by performing gradient ascent based on π_ϕ. Thus π_E can guide exploration towards maximizing the exploration ability continuously, performing noise-aware optimistic exploration. Concretely, Proposition 3 shows how to calculate π_E.

Proposition 3 *Based on any policy $\pi_\phi = \mathcal{N}(\mu_\phi, \sigma_\phi)$, the OVD-Explorer behavior policy $\pi_E = \mathcal{N}(\mu_E, \Sigma_E)$ at given state s is as follows:*

$$\mu_E = \mu_\phi + \alpha \mathbb{E}_{\bar{Z}^\pi}\left[m \times \frac{\partial \bar{z}(s,a)}{\partial a}|_{a=\mu_\phi}\right], \tag{15}$$

and

$$\Sigma_E = \sigma_\phi. \tag{16}$$

In specific, $m = \log \frac{\Phi_{Z^\pi(s,\mu_\phi)}(\bar{z}(s,\mu_\phi))}{C} + 1$, $\bar{z}(s,a)$ is a sample from OVD $\bar{Z}^\pi$, and α controls the step size of the update along the gradient direction, representing the exploration degree (see proof in the Appendix).

The expectation $\mathbb{E}_{\bar{Z}^\pi}$ can be estimated by K samples, then Eq. 15 is simplifies as:

$$\mu_E = \mu_\phi + \frac{\alpha m}{K} \sum_{i=1}^{K} \frac{\partial \bar{z}_i(s,a)}{\partial a}|_{a=\mu_\phi}. \tag{17}$$

Algorithm 1 summarizes the procedure to generate a behavior policy at step t of OVD-Explorer. Following Algorithm 1, OVD-Explorer can be integrated with any existing policy-based RL algorithms from a distributional perspective, to render a stable and well-performed algorithm.

Algorithm 1: Behavior policy generation at step t.

```
Input: Current state $s_t$, current value distribution estimators $\theta_1, \theta_2$, current policy network $\phi$.
Output: Behavior policy $\pi_E$.
1: Obtain policy $\pi_\phi(\cdot|s_t) \sim \mathcal{N}(\mu_\phi(s_t), \sigma_\phi(s_t))$
2: // Construct the distributions of return and upper bound
3: Construct OVD $\bar{Z}^\pi(s_t, \mu_\phi(s_t))$ using Eq. 11
4: Construct $Z^\pi(s_t, \mu_\phi(s_t))$ using Eq. 13 or 14
5: // Calculate the behavior policy
6: Calculate the behavior policy's mean $\mu_E$ using Eq. 17
7: return $\pi_E \sim \mathcal{N}(\mu_E, \sigma_\phi(s_t))$
```

Specifically, given state s_t, based on the current policy (Line 1), by constructing the optimistic value distribution $\bar{Z}^\pi$ as well as the value distribution Z^π of the policy (Line 3-4), the behavior policy derived from OVD-Explorer can be calculated directly using Proposition 3 (Line 6-7).

Experiments

To reveal the consistency between our theoretical analysis and the performance of OVD-Explorer, and demonstrate the significant advantage over other advanced methods, we conduct experiments mainly for the following questions:
RQ1 (Exploration ability): Can OVD-Explorer explore as a noise-aware optimistic manner as expected?
RQ2 (Performance): Can OVD-Explorer perform notable advantages on common continuous control benchmarks?
Due to space constraints, more experimental details and evaluation results can be found in the Appendix.

Baseline Algorithms and Implementation Details

Our baseline algorithms include SAC (Haarnoja et al. 2018), DSAC (Ma et al. 2020), and DOAC, an extension of the scalar Q-value within the OAC (Ciosek et al. 2019) to distributional Q-value. Our implementation of OVD-Explorer is based on the OAC repository, also refers to the code of DSAC [1] and softlearning [2]. We implement OVD-Explorer_G and OVD-Explorer_Q (or abbreviated as OVDE_G and OVDE_Q), representing approaches to formulate the value distribution Z^π using Eq. 13 (`torch.distributions.Normal`) or Eq. 14, respectively. The key hyper-parameters associated with the exploration, i.e., the exploration ratio α and the uncertainty ratio β, are determined by grid search, with detailed information presented in the Appendix. Moreover, the hyperparameters related to the training procedure remain consistent across all algorithms.

All experiments are performed on NVIDIA GeForce RTX 2080 Ti 11GB graphics card. To counteract the randomness from a statistical perspective, we conduct multiple trials using different seeds. The final results of each trial are collected based on the mean undiscounted episodic return over the last 8% epoch (or up to the last 100 epochs) to ensure impartiality and minimize bias.

[1] https://github.com/xtma/dsac

[2] https://github.com/rail-berkeley/softlearning

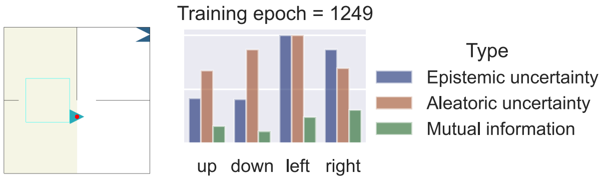


Figure 3: GridChaos. Left: In this map, the areas with darker background color have higher noise injected, and the agent aims at reaching the goal at the top right. Right: The values of uncertainty and exploration ability.

Exploration in GridChaos (RQ1)

To illustrate that OVD-Explorer guides noise-aware optimistic exploration, we first evaluate OVD-Explorer on GridChaos. The GridChaos task is characterized by heterogeneous noise and sparse reward, making it particularly challenging and necessitating a robust capacity for noise-aware optimistic exploration to successfully accomplish the task.

GridChaos Task GridChaos is built on OpenAI's Gym toolkit, as shown in Fig. 3. In GridChaos, the cyan triangle is under the agent's control, with the objective of reaching the fixed dark blue goal located at the top right corner. The state is its current coordinate, and the action is a two-dimensional vector including the movement angle and distance. An episode terminates when the agent reaches the goal or maximum steps (typically 100). Also, it receives a +100 reward when reaching the goal, and otherwise 0. To simulate noise, heterogeneous Gaussian noise is injected into the state transitions.

Exploration Patterns Analysis We first analyse the exploration pattern facilitated by OVD-Explorer. We compare values of uncertainty and the exploration ability measurement (in Proposition 2) corresponding to distinct actions at the state in Fig. 3. It shows the values obtained at the 1249th training epoch. Basically, it shows that estimated aleatoric uncertainty (noise) of left is the highest, aligning with the environment's inherent attribute. This indicates that OVD-Explorer models the noise properly. Further, OVD-explorer encourages to explore towards right, where the exploration ability (in green) is higher. It implies that OVD-explorer balances the optimistic and noise in exploration, aligning with our intended objective. Nevertheless, if the noise is not considered in exploration, the agent would be guided towards left, where the epistemic uncertainty is higher, then the agent may be trapped due to the high randomness in this area, potentially explaining why DOAC fails to effectively address such a stochastic task.

Evaluation on Various Noise Scales To further empirically prove our strength, we test OVD-Explorer in GridChaos with various noise scales settings, as outlined in row *A-D* in Tab. 1. Note that the variance is not reported, as the mean values of 5 seeds offer a comprehensive representation of results. For instance, when the average result approaches 20, it indicates that only one seed successfully achieved the goal (obtaining a reward of +100) in the end. The row *S* is the standard GridChaos as shown in Fig. 3. Remarkably, the results indicate that OVD-Explorer consistently achieves better performance across all the tested settings. This underscores OVD-Explorer's exploration capability in such noisy tasks, leading to more efficient learning and faster convergence towards the goal (see column *FRG*).

	Average return			FRG epoch		
	DSAC	OVDE	DOAC	DSAC	OVDE	DOAC
s	0.00	**59.30**	3.02	1250+	**229**	1222
a	18.94	**58.99**	38.42	1161	**180**	662
b	39.78	**79.52**	18.71	694	**144**	846
c	0.05	**39.64**	20.59	1250+	**180**	309
d	20.00	**40.46**	39.99	**284**	**276**	**321**
e	0.00	**20.20**	14.60	1250+	**185**	1118

Table 1: The averaged return of 5 runs for GridChaos (the first part). FRG epoch means the minimum training epochs to Firstly Reach the Goal before totally 1250 epochs.

	Average return				FRG epoch		
	DSAC	OVDE	OVDE(m)	DOAC	DSAC	OVDE	DOAC
f	0.00	19.84	**39.96**	20.14	1250+	**188**	**233**
g	0.00	**20.69**	**20.07**	0.00	1250+	**247**	1250+
h	0.00	40.00	**60.00**	39.99	1250+	**200**	301
i	20.00	**39.98**	20.00	20.00	**236**	**312**	**296**

Table 2: The averaged return of 5 runs (the second part).

Also, we make observations in the case without noise (row *E*). Here, DSAC fails in any run across 5 seeds, while DOAC reaches the goal in one run but at a slower pace. OVD-Explorer achieves the goal swiftly in one run. This highlight the capability of OVD-Explorer and the highly challenging nature of the task, emphasizing the significance of employing a robust exploration strategy.

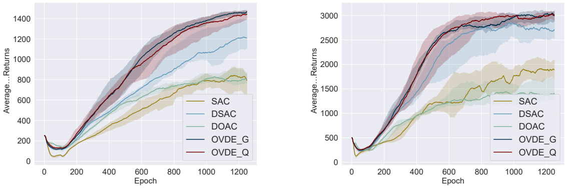
Evaluation on Tasks in Which the Noise is High around the Goal To verify whether the noise avoidance ability of OVD-Explorer dominates the exploration process when the noise around the target is higher, we conduct the experiment where the noise in the right half (where the goal is located), is set larger. The results are shown in row *F-I* in Tab. 2. Note that we use OVDE to denote the usual implementation that pessimistically estimates the value distribution (i.e., using Eq. 13). Besides, OVDE(m) denotes the implementation that does not pessimistically estimate the value distribution (i.e., we modify the mean of Gaussian distribution Z^{π} in Eq. 13 from the lower bound to expected value of the Q estimation $\mu(s, a)$ as in Eq. 12).

Overall, in most cases, OVD-Explorer guides better exploration and perform better than baselines. This highlights OVD-Explorer's ability to handle various scenarios effectively, even in tasks with higher noise levels around the goal.

Moreover, an intriguing observation is that OVD-Explorer may exhibit better performance when the pessimistic estimation is turned off in the presence of higher noise around the goal (see column OVDE(m)). This finding suggests that

Task	**SAC**	**DSAC**	**DOAC**	**OVD-Explorer_G**	OVD-Explorer_Q
Ant-v2	4867.8±1658.7	6385.9±1287.2	6625.4±746.8	7175.3±789.0	**7382.3**±466.6
HalfCheetah-v2	11619.8±1642.01	13348.4±1957.1	12987.6±148.1	14796.2±1473.2	**16484.3**±1373.75
Hopper-v2	**2593.5**±574.7	2506.0±390.56	2353.0±754.1	2394.6±496.6	2559.3±384.5
Reacher-v2	-22.7±2.0	-12.2±1.6	-18.7±1.7	-11.6±1.0	**-11.3**±1.2
InvDbPendulum-v2	9306.2±89.5	8916.9±1041.7	5798.5±3439.0	9263.8±189.1	**9355.0**±12.1
N-Ant-v2	222.96±41.93	465.34±53.94	344.71±20.39	**524.16**±10.54	513.77±17.87
N-HalfCheetah-v2	368.57±28.01	431.81±39.41	402.26±37.27	447.30±38.57	**453.56**±55.97
N-Hopper-v2	213.71±21.97	238.62±19.89	**252.53**±13.07	234.88±15.24	239.43±9.90
N-Pusher-v2	-50.57±20.65	-27.33±3.79	-29.82±4.29	**-25.69**±3.57	-26.13±3.63
N-InvDbPendulum-v2	931.63±7.14	**932.81**±1.87	381.87±139.36	**932.70**±2.26	**933.54**±2.69
Average (standard tasks)	5672.92	6229.00	5549.16	6723.66	**7153.92**
Average (noisy tasks)	337.26	408.25	270.31	**422.67**	**422.83**

Table 3: Comparisons of algorithms on five standard and five noisy tasks. The averaged performance and standard deviation of 10 runs are reported. The training epoch count is shown in column epoch, and the best values of each row are shown in bold.



(a) 250 steps

(b) 500 steps

Figure 4: Training curves on Noisy Ant-v2 tasks with different maximum episodic length setup. The sub-title of each figure represents the episodic horizon. We report the median of returns and the interquartile range of 10 runs. Curves are smoothed uniformly for visual clarity.

excessive pessimism may not be necessary when there is a crucial need to explore noisy areas. In such cases, a more balanced approach may lead to improved results.

Performance on Mujoco Tasks (RQ2)

To showcase the broader efficacy of OVD-Explorer, we conduct experiments encompassing 5 standard and 5 stochastic tasks based on wildly-used continuous control benchmark, Gym Mujoco. Tab. 3 shows the averaged performance and standard deviation of 10 seeds. It is important to note that for the standard tasks[3], the dynamics are deterministic, and any observed noise is ascribed to the stochastic policy employed. Conversely, in the case of the five noisy tasks (indicated by the prefix N-), Gaussian noise of varying scales is randomly injected into each state transition. The investigation yields compelling insights into the capabilities of OVD-Explorer.

Primarily, OVD-Explorer can perform stably in standard task. The results reveal that DSAC consistently outperforms SAC, underscoring the advantage that value distribution brings to policy evaluation. Notably, OVD-Explorer surpasses baseline algorithms significantly, particularly in high-dimensional tasks, such as Ant-v2 and HalfCheetah-v2. These evaluations on standard tasks convincingly demonstrate OVD-Explorer's remarkable proficiency in promoting optimistic exploration, highlighting its universal efficacy in exploration capabilities.

Second, these results underscore the efficacy of OVD-Explorer in exploring noisy environments while avoiding the adverse impact of noise. This is exemplified by experiments on 5 noisy tasks (Tab. 3). Notably, DOAC's performance is even inferior to DSAC in most tasks, indicating that the presence of heteroscedastic noise significantly interferes with the exploration process guided by DOAC. In contrast, OVD-Explorer exhibits substantial advantages over SAC and DOAC, and outperforms DSAC in most cases.

Furthermore, regarding the two implementations (namely OVDE_G and OVDE_Q) of OVD-Explorer, we observe that OVDE_Q consistently demonstrates greater stability. The key distinction between these implementations lies in the formulations of $Z^{\pi}(s, a)$. OVDE_Q's employment of quantile distribution offers higher flexibility, allowing for a more accurate characterization of the value distribution. Conversely, OVDE_ G, reliant on the Gaussian prior, exhibits limited capacity in this regard, leading to a relatively diminished performance in some cases.

Conclusion

In this paper, we have presented OVD-Explorer, a novel noise-aware optimistic exploration method for continuous RL. By introducing a unique measurement of exploration ability and maximizing it, OVD-Explorer effectively generates a behavior policy that adheres to the OFU principle. Also, it intelligently avoids excessive exploration in areas with high noise, thereby mitigating the adverse effects of noise. Consistently across tasks with no noise as well as various forms of noise, the experiment underscores our performance advantages. Moving forward, we recognize the potential for extending OVD-Explorer to discrete tasks and even Multi-agent tasks. This will enhance the versatility of OVD-Explorer, affording it the capability to effectively confront the challenges of exploration in noisy environments that are widespread across diverse real-world scenarios.

[3] https://github.com/openai/gym/tree/master/gym/envs/mujoco

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No. 2022ZD0116402), the National Natural Science Foundation of China (Grant Nos. 92370132, 62106172).

References

Antos, A.; Munos, R.; and Szepesvári, C. 2007. Fitted Q-iteration in continuous action-space MDPs. In *Advances in NIPS, Vancouver, British Columbia, Canada, December 3-6, 2007*, 9–16. Curran Associates, Inc.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.*, 47(2-3): 235–256.

Badia, A. P.; Sprechmann, P.; Vitvitskyi, A.; Guo, D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A.; and Blundell, C. 2020. Never Give Up: Learning Directed Exploration Strategies. *arXiv preprint arXiv:2002.06038*.

Bai, C.; Wang, L.; Han, L.; Garg, A.; Hao, J.; Liu, P.; and Wang, Z. 2021a. Dynamic bottleneck for robust self-supervised exploration. In *Advances in Neural Information Processing Systems*, volume 34, 17007–17020.

Bai, C.; Wang, L.; Han, L.; Hao, J.; Garg, A.; Liu, P.; and Wang, Z. 2021b. Principled Exploration via Optimistic Bootstrapping and Backward Induction. In *International Conference on Machine Learning*, volume 139, 577–587.

Bai, C.; Xiao, T.; Zhu, Z.; Wang, L.; Zhou, F.; Garg, A.; He, B.; Liu, P.; and Wang, Z. 2022. Monotonic quantile network for worst-case offline reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th ICML, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 449–458. PMLR.

Bellemare, M. G.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in NIPS*, 1471–1479.

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

Ciosek, K.; Vuong, Q.; Loftin, R.; and Hofmann, K. 2019. Better Exploration with Optimistic Actor Critic. In *Advances in NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 1785–1796.

Clements, W. R.; Robaglia, B.; Delft, B. V.; Slaoui, R. B.; and Toth, S. 2019. Estimating Risk and Uncertainty in Deep Reinforcement Learning. *CoRR*, abs/1905.09638.

Dabney, W.; Ostrovski, G.; Silver, D.; and Munos, R. 2018a. Implicit Quantile Networks for Distributional Reinforcement Learning. In *Proceedings of the 35th ICML, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 1104–1113. PMLR.

Dabney, W.; Rowland, M.; Bellemare, M. G.; and Munos, R. 2018b. Distributional Reinforcement Learning With Quantile Regression. In *AAAI, New Orleans, Louisiana, USA, February 2-7, 2018*, 2892–2901. AAAI Press.

Fujimoto, S.; van Hoof, H.; and Meger, D. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th ICML, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 1582–1591. PMLR.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33nd ICML, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 1050–1059. JMLR.org.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 1856–1865. PMLR.

Hao, J.; Yang, T.; Tang, H.; Bai, C.; Liu, J.; Meng, Z.; Liu, P.; and Wang, Z. 2023. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*.

Houthooft, R.; Chen, X.; Duan, Y.; Schulman, J.; De Turck, F.; and Abbeel, P. 2016. Vime: Variational information maximizing exploration. In *Advances in NIPS*, 1109–1117.

Kirschner, J.; and Krause, A. 2018. Information Directed Sampling and Bandits with Heteroscedastic Noise. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, 358–384. PMLR.

Koenker, R.; and Hallock, K. F. 2001. Quantile regression. *Journal of economic perspectives*, 15(4): 143–156.

Lee, K.; Laskin, M.; Srinivas, A.; and Abbeel, P. 2021. SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. In *ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 6131–6141. PMLR.

Li, B.; Tang, H.; Zheng, Y.; Hao, J.; Li, P.; Wang, Z.; Meng, Z.; and Wang, L. 2021. HyAR: Addressing Discrete-Continuous Action Reinforcement Learning via Hybrid Action Representation. *CoRR*, abs/2109.05490.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2016. Continuous control with deep reinforcement learning. In *ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ma, X.; Zhang, Q.; Xia, L.; Zhou, Z.; Yang, J.; and Zhao, Q. 2020. Distributional Soft Actor Critic for Risk Sensitive Learning. *CoRR*.

Martin, J.; Sasikumar, S. N.; Everitt, T.; and Hutter, M. 2017. Count-Based Exploration in Feature Space for Reinforcement Learning. In *Proceedings of the Twenty-Sixth IJCAI*, 2471–2478.

Mavrin, B.; Yao, H.; Kong, L.; Wu, K.; and Yu, Y. 2019. Distributional Reinforcement Learning for Efficient Exploration. In *ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 4424–4434. PMLR.

Nikolov, N.; Kirschner, J.; Berkenkamp, F.; and Krause, A. 2019. Information-Directed Exploration for Deep Reinforcement Learning. In *ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Osband, I.; Blundell, C.; Pritzel, A.; and Roy, B. V. 2016. Deep Exploration via Bootstrapped DQN. In *Advances in NIPS 2016, December 5-10, 2016, Barcelona, Spain*, 4026–4034.

Pathak, D.; Gandhi, D.; and Gupta, A. 2019. Self-Supervised Exploration via Disagreement. In *Proceedings of the 36th ICML, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 5062–5071. PMLR.

Qiu, S.; Wang, L.; Bai, C.; Yang, Z.; and Wang, Z. 2022. Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, 18168–18210. PMLR.

Savinov, N.; Raichuk, A.; Vincent, D.; Marinier, R.; Pollefeys, M.; Lillicrap, T. P.; and Gelly, S. 2019. Episodic Curiosity through Reachability. In *ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Tang, Y.; and Agrawal, S. 2020. Discretizing Continuous Action Space for On-Policy Optimization. In *AAAI, New York, NY, USA, February 7-12, 2020*, 5981–5988. AAAI Press.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4): 285–294.

Watkins, C. J. C. H.; and Dayan, P. 1992. Technical Note Q-Learning. *Mach. Learn.*, 8: 279–292.

Yuan, Y.; Hao, J.; Ni, F.; Mu, Y.; Zheng, Y.; Hu, Y.; Liu, J.; Chen, Y.; and Fan, C. 2023. EUCLID: Towards Efficient Unsupervised Reinforcement Learning with Multi-choice Dynamics Model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.