OVD-EXPLORER: A GENERAL INFORMATION-THEORETIC EXPLORATION APPROACH FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many exploration strategies are built upon the optimism in the face of the uncertainty (OFU) principle for reinforcement learning. However, without considering the aleatoric uncertainty, existing methods may over-explore the state-action pairs with large randomness and hence are non-robust. In this paper, we explicitly capture the aleatoric uncertainty from a distributional perspective and propose an information-theoretic exploration method named Optimistic Value Distribution Explorer (OVD-Explorer). OVD-Explorer follows the OFU principle, but more importantly, it avoids exploring the areas with high aleatoric uncertainty through maximizing the mutual information between policy and the upper bounds of policy's returns. Furthermore, to make OVD-Explorer tractable for continuous RL, we derive a closed form solution, and integrate it with SAC, which, to our knowledge, for the first time alleviates the negative impact on exploration caused by aleatoric uncertainty for continuous RL. Empirical evaluations on the commonly used Mujoco benchmark and a novel GridChaos task demonstrate that OVD-Explorer can alleviate over-exploration and outperform state-of-the-art methods.

1 INTRODUCTION

The exploration and exploitation trade-off is critical in reinforcement learning (Thompson, 1933; Sutton & Barto, 2018). Many exploration strategies have been proposed in the literature (Osband et al., 2016; Lillicrap et al., 2016; Martin et al., 2017; Pathak et al., 2017; Ciosek et al., 2019). Among these strategies, those following the *Optimism in the Face of Uncertainty* (OFU) principle (Auer et al., 2002) provide efficient guidance for exploration (Chen et al., 2017; Ciosek et al., 2019). Generally, OFU-based methods regard the uncertainty as the result of insufficient exploration for the state-action pair, and refer to such uncertainty as *epistemic uncertainty*. Besides, there is another uncertainty called *aleatoric uncertainty*, which captures environmental stochasticity:

- *Epistemic uncertainty* (a.k.a. parametric uncertainty) represents the ambiguity of the model arisen from insufficient knowledge, and is high at those state-action pairs seldom visited (Dearden et al., 1998; Osband et al., 2016; Moerland et al., 2017).
- *Aleatoric uncertainty* (a.k.a. intrinsic uncertainty) is the variation arisen from environment randomness, caused by the stochasticity of policy, reward and/or transition probability, and is characterized by return distribution. (Bellemare et al., 2017; Moerland et al., 2017).

In the heteroscedastic stochastic tasks where the environment randomness of different state-action pairs differs, those OFU-based methods may be inefficient without properly tackling aleatoric uncertainty (Nikolov et al., 2019). As shown in Figure 1, if the epistemic uncertainty estimation is disturbed due to the volatility caused by aleatoric uncertainty, the exploration strategy (i.e., policy A) optimistic about such estimated uncertainty may lead to explore areas with high aleatoric uncertainty. Visiting such areas results in unstable and risky transitions. This issue, to explore overly the state-action pairs visited frequently but with high aleatoric uncertainty, is referred to as the *over-exploration issue*. Intuitively, such issue could be solved by avoiding exploring optimistically about aleatoric uncertainty (i.e., policy B in Figure 1).

The similar concern has been raised in the discrete RL, where Nikolov et al. (2019) proposes to use the Information-Directed Sampling (IDS) to avoid the over-exploration issue in the environments with heteroscedastic noise. However, IDS needs to calculate information-regret ratio for each action, thus applying IDS for continuous RL with explosive or even infinite action space could be non-trivial and ineffective. Meanwhile, many advanced continuous RL approaches suffer from the over-exploration issue. Soft Actor-Critic (SAC) is a well performed continuous RL algorithm, but does not account for efficient exploration beyond maximizing policy entropy (Haarnoja et al., 2018). OAC improves the exploration following OFU principle, but it ignores to avoid exploration towards higher aleatoric uncertainty (Ciosek et al., 2019). Therefore, an effective exploration method to address the over-exploration issue for continuous RL is of urgent need.

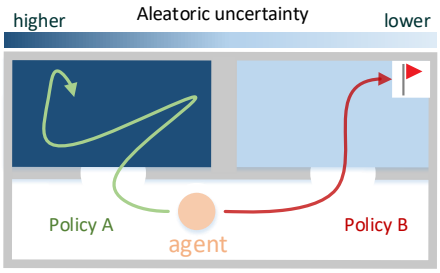


Figure 1: An intuitive example. Red/green lines denote good/bad exploration policies.

To address the over-exploration issue, we propose that the characterization of aleatoric uncertainty and a new exploration principle are necessary. The new principle need to properly trade-off between the epistemic and aleatoric uncertainties, so as to explore seldom visited state-action pairs while avoiding trapped in areas with high aleatoric uncertainty. In this paper, we propose a general information-theoretic exploration approach OVD-Explorer, which enriches the OFU exploration principle with a complement: "avoid areas with high aleatoric uncertainty". Furthermore, to guide the exploration following the new principle, OVD-Explorer maximizes the mutual information between policy and corresponding upper bounds. Notable, the upper bounds represent the best return that policy can reach, formulated using *Optimistic Value Distributions* (OVD), and the value distribution characterises the aleatoric uncertainty. From the theoretical derivation, we show that maximizing such mutual information urges the OVD-Explorer to guide exploration towards the areas with higher epistemic uncertainty and alleviate over-exploration issue by avoiding the areas with higher aleatoric uncertainty.

To make OVD-Explorer tractable for continuous action space, we derive its closed form, and propose a scheme to incorporate OVD-Explorer with any policy-based RL algorithm. Practically, we demonstrate the exploration benefits based on SAC. Evaluations on many challenging continuous RL tasks, including a toy task GridChaos, tasks in Mujoco as well as their stochastic variants are conducted, and the results verify our analysis. To the best of our knowledge, OVD-Explorer firstly addresses the negative impact of aleatoric uncertainty for exploration in continuous RL.

2 Preliminaries

2.1 Distributional RL

To model the randomness in the observed long-term return, which characterises aleatoric uncertainty, distributional RL methods (Bellemare et al., 2017; Dabney et al., 2018b;a) are used to estimate Q-value distribution rather than expected Q-value (Mnih et al., 2015). In our paper, we focus on quantile regression used in QR-DQN (Dabney et al., 2018b), where the distribution of Q-value is represented by the quantile random variable Z. Z maps the state-action pair to a uniform probability distribution supported on the values at all corresponding quantile fractions. Based on value distribution estimator parameterized as θ, we denote the value at quantile fraction τ_i as $\hat{Z}_i(s, a, \theta)$ given state-action pair (s, a).

Similar to the Bellman operator in the traditional Q-Learning (Watkins & Dayan, 1992), the distributional Bellman operator (Bellemare et al., 2017) $\mathcal{T}_D^\pi$ under policy π is given as:

$$\mathcal{T}_D^\pi Z(s_t, a_t) \stackrel{D}{=} R(s_t, a_t) + \gamma Z(s_{t+1}, a_{t+1}), a_{t+1} \sim \pi(\cdot|s_{t+1}). \tag{1}$$

Notice that $\mathcal{T}_D^\pi$ operates on random variables, $\stackrel{D}{=}$ denotes that distributions on both sides have equal probability laws. Based on operator $\mathcal{T}_D^\pi$, Dabney et al. (2018b) propose QR-DQN to train quantile

estimations via the quantile regression loss (Koenker & 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 TD error $\delta_{i,j} = R(s_t, a_t) + \gamma\hat{Z}_i(s_{t+1}, a_{t+1}; \bar{\theta}) - \hat{Z}_j(s_t, a_t; \theta)$, 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 $\hat{\tau}_i = \frac{\tau_{i+1}+\tau_i}{2}$.

2.2 Distributional Soft Actor-Critic method

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 Equation 1, the distributional soft Bellman operator $\mathcal{T}^{\pi}_{DS}$ is defined considering the maximum entropy RL as follows:

$$\mathcal{T}^{\pi}_{DS}Z(s_t, a_t) \overset{D}{=} R(s_t, a_t) + \gamma[Z(s_{t+1}, a_{t+1}) - \alpha\log\pi(a_{t+1}|s_{t+1})], \tag{3}$$

where $a_{t+1} \sim \pi(\cdot|s_{t+1}), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. Then, the quantile regression loss differs from Equation 2 on $\delta_{i,j}$, by extending clipped double Q-Learning (Fujimoto et al., 2018) based on the maximum entropy RL framework, to overcome overestimation of Q value estimation:

$$\delta^k_{i,j} = R(s_t, a_t) + \gamma[\min_{k=1,2}\hat{Z}_i(s_{t+1}, a_{t+1}; \bar{\theta}_k) - \alpha\log\pi(a_{t+1}|s_{t+1}; \bar{\phi})] - \hat{Z}_j(s_t, a_t; \theta_k), \tag{4}$$

where $\bar{\theta}$ and $\bar{\phi}$ represents their target networks respectively. The actor is optimized by minimizing the following actor loss, the same as SAC,

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

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

$$Q(s_t, a_t; \theta) = \min_{k=1,2} Q(s_t, a_t; \theta_k) = \frac{1}{N}\min_{k=1,2}\sum_{i=0}^{N-1}\hat{Z}_i(s_t, a_t; \theta_k). \tag{6}$$

3 Methodology

We introduce a general exploration method, named OVD-Explorer, to achieve efficient and robust exploration for continuous RL. Overall, OVD-Explorer estimates the upper bounds of the policy's return, and uses the bounds as criteria to find a behavior policy for online exploration from the information-theoretic perspective. Innovatively, to avoid over-exploration, OVD-Explorer particularly takes the aleatoric uncertainty into account, leveraging the optimistic value distribution (OVD) to formulate the upper bounds of policy's return. Then, OVD-Explorer derives the exploration policy via maximizing the mutual information between policy and corresponding upper bounds for an optimistic exploration. Meanwhile, the derived behavior policy avoids being optimistic to aleatoric uncertainty, thus avoiding over-exploration issue.

In the following, we first give the intuition together with the theoretical derivation of OVD-Explorer. Then we practically describe how we formulate OVD based on uncertainty estimation. Lastly, analysis are given to further illustrate why OVD-Explorer can explore effectively and robustly.

3.1 Optimistic value distribution guided exploration

Most OFU-based algorithms follow the exploration principle: "select the action that leads to areas with high uncertainty". However, this principle is incomplete as the high uncertainty may also be caused by the aleatoric uncertainty, which misleads the exploration direction. Hence, we complement the principle by introducing a constrain: ***"Not only select the action that leads to areas with high uncertainty, but also avoid the ones that only lead to the area with high aleatoric uncertainty"***. Notable, this new principle intuitively describes a good exploration ability, and we will

illustrate how to quantitatively measure a policy's exploration ability in the later section. By following this principle, OVD-Explorer derives the behavior policy for exploration.

Intuitively, each policy π in the policy space Π has different exploration ability (at state s), denoted by $\mathbf{F}^\pi(\mathbf{s})$. OVD-Explorer aims at finding the exploration policy π_E with the best exploration ability by solving the following optimization problem:

$$\pi_E = \arg\max_{\pi \in \Pi} \mathbf{F}^\pi(s). \tag{7}$$

We propose to quantitatively measure the policy's exploration ability from an information-theoretic perspective by measuring the mutual-information of multi-variables as follows:

$$\mathbf{F}^\pi(s) = \mathbf{MI}(\bar{Z}^\pi(s, a_0), \cdots, \bar{Z}^\pi(s, a_k), \pi(s)|s), \tag{8}$$

where $\pi(s)$ is the action random variable, and $\bar{Z}^\pi(s, a_i)$ the random variable describing the upper bounds of return that action a_i can reach under policy π at state s. Here, $a_i \in \mathbf{A}$ denotes any legal action, thus k could be infinite in continuous action space. The higher $\mathbf{F}^\pi(\mathbf{s})$ is, the better ability that policy π has in exploring towards higher epistemic uncertainty while simultaneously avoiding higher aleatoric uncertainty, which satisfies the principle we raised. (see analysis in Section 3.3.)

Now we state the Theorem 1 to measure the mutual information above in continuous action space.

Theorem 1. *The mutual information of $\bar{Z}^\pi(s, a_0), \cdots, \bar{Z}^\pi(s, a_k)$ and $\pi(s)$ at state s is:*

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

$\Phi_x(\cdot)$ 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 Appendix B.1).

Theorem 1 reveals that $\mathbf{F}^\pi(s)$ is only proportional to $\Phi_{Z^\pi}(\bar{z}(s, a))$, by maximizing which the policy's exploration ability can be improved. Concretely, given any policy π_θ (parameterized by θ), the derivative $\nabla_\theta \Phi_{Z^\pi_\theta}(\bar{z}(s, a))$ can be measured, thus gradient ascent can be iteratively performed to derive a better exploration policy. (The complete closed form solution is given in later Section 4)

Note that, to perform above optimization procedure, we need to formulate two critical components in $\Phi_{Z^\pi}(\bar{z}(s, a))$: ❶ the return of the policy $Z^\pi(s, a)$ and ❷ upper bound of the return of the policy $\bar{Z}^\pi(s, a)$. Therefore, we first details the formulations in Section 3.2, and then analyze why maximizing $\Phi_{Z^\pi}(\bar{z}(s, a))$ can derive a better exploration in Section 3.3.

3.2 Formulating the Return of Policy and Corresponding Upper Bound

Now, we introduce the formulation of the return of policy $Z^\pi(s, a)$ and corresponding upper bound $\bar{Z}^\pi(s, a)$. As mentioned before, most OFU-based methods neglect the aleatoric uncertainty when formulating the return and upper bound of return (Mavrin et al., 2019; Chen et al., 2017), resulting in the over-exploration issue. Therefore, OVD-Explorer particularly takes into account the aleatoric uncertainty, leveraging distributional RL paradigm to characterize aleatoric uncertainty (Bellemare et al., 2017; Dabney et al., 2018b;a). OVD-Explorer uses 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 ensemble estimators to formulate $\bar{Z}^\pi$ and Z^π in different ways. Unless stated otherwise, the (s, a) is omitted hereafter to ease notation.

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

$$\bar{Z}^\pi(s, a) \sim \mathcal{N}(\mu_{\bar{Z}}(s, a), \sigma^2_{\text{aleatoric}}(s, a)), \tag{10}$$

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

$$\mu_{\bar{Z}}(s, a) = \mu(s, a) + \beta \sigma_{\text{epistemic}}(s, a), \text{ s.t. } \begin{cases} \mu(s, a) = \mathbb{E}_{i \sim \mathcal{U}(1,N)} \mathbb{E}_{k=1,2} \hat{Z}_i(s, a; \theta_k) \\ \sigma^2_{\text{epistemic}}(s, a) = \mathbb{E}_{i \sim \mathcal{U}(1,N)} \text{var}_{k=1,2} \hat{Z}_i(s, a; \theta_k) \end{cases}, \tag{11}$$

Table 1: The comparison about two scenarios. Note that for the ease of clarity, without causing ambiguity, we omit part of the notation in the table.

Fig. 2	Aleatoric uncertainty σ_a	Epistemic uncertainty σ_e	Optimistic value estimation $\mu_{\bar{Z}}(s,a)$	CDF value $\Phi(a)$	Action a
(a)	$\sigma_a(a_1) = \sigma_a(a_2)$	$\sigma_e(a_1) > \sigma_e(a_2)$	$\mu_{\bar{Z}}(s,a_1) > \mu_{\bar{Z}}(s,a_2)$	$\Phi(a_1) > \Phi(a_2)$	a_1
(b)	$\sigma_a(a_1) < \sigma_a(a_2)$	$\sigma_e(a_1) = \sigma_e(a_2)$	$\mu_{\bar{Z}}(s,a_1) = \mu_{\bar{Z}}(s,a_2)$	$\Phi(a_1) > \Phi(a_2)$	a_1

where $\mu(s,a)$ represents the Q-value estimation, $\sigma_{\text{epistemic}}(s,a)$ is the epistemic uncertainty, β is a hyper-parameter controlling the magnitude of epistemic uncertainty, $\mathcal{U}$ is uniform distribution, N is the number of quantiles, and $\hat{Z}_i(s,a;\theta_k)$ is the value of the i-th quantile drawn from $\hat{Z}(s,a;\theta_k)$.

Moreover, the aleatoric uncertainty should be considered as it will affect the policy's performance, thus we propose to model it as the variance of $\bar{Z}^\pi$. In practice, the aleatoric uncertainty $\sigma_{\text{aleatoric}}(s,a)$ can be captured by value distribution estimators (Clements et al., 2019) as follows:

$$\sigma^2_{\text{aleatoric}}(s,a) = \text{var}_{i\sim\mathcal{U}(1,N)}\left[\mathbb{E}_{k=1,2}\hat{Z}_i(s,a;\theta_k)\right]. \tag{12}$$

Leveraging optimistic value estimations (via epistemic uncertainty) together with explicitly modeling aleatoric uncertainty, the upper bound of the value $\bar{Z}^\pi$ can be effectively formulated as an optimistic Gaussian distribution. Such an upper bound value estimation provides a useful guidance for OVD-Explorer to derive the behavior policy for online exploration.

Formulation of Z^π. Z^π estimates the value obtained following policy π. Inspired by TD3 (Fujimoto et al., 2018), to alleviate overestimation, we propose to 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 Equation 10, Z^π can also be formulated as Gaussian distribution as follows:

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

where $\mu(s,a)$, $\sigma_{\text{aleatoric}}(s,a)$ and $\sigma_{\text{epistemic}}(s,a)$ are the same defined in Equation 11. Differently, $\sigma_{\text{epistemic}}(s,a)$ is subtracted from $\mu(s,a)$, which reveals the pessimistic value estimation of Z^π.

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

$$Z^\pi(s,a) \sim \mathcal{U}\{z_i^\pi(s,a;\theta)\}_{i=1,\dots,n}, \quad \textit{s.t.}\ z_i^\pi(s,a;\theta) = \min_{k=1,2}\hat{Z}_i(s,a;\theta_k), \tag{14}$$

where each quantile value $z_i^\pi(s,a;\theta)$ equals to the minimum estimated value among ensemble estimators (i.e., $z_i(s,a;\theta_k)$. As such, the estimated value of Z^π is relative pessimistic.

Adopting pessimistic value estimation avoids overestimation issue, thereby the value distribution Z^π can be more accurately formulated. Additional, Gaussian distribution helps more when the environment randomness follows a unimodal distribution, and multivariate uniform distribution can be more flexible and suitable for scenarios with multi-modal distributions. (see more in Section 5.3.)

3.3 Analysis of OVD-Explorer

This section analyzes how OVD-Explorer optimistically explores the informative areas and avoids the over-exploration issue. According to Theorem 1, the exploration policy OVD-Explorer derives can maximize $\mathbf{F}^\pi(s)$, which is proportional to CDF value $\Phi_{Z^\pi}(\bar{z}(s,a))$. In the following, Figure 2(a) and 2(b) illustrate such CDF values for different actions (shaded area), and Table 1 shows how uncertainties affect the exploration.

In these cases, the value distribution $Z^\pi(s,a)$ is specified as a Gaussian distribution (in Equation 13), and the sampled optimistic value $\bar{z}(s,a)$ is specified as the mean of OVD $\mu_{\bar{Z}}(s,a)$ (in Equation 11). Specifically, 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.

Basically, Figure 2(a) shows OVD-Explorer can achieve more optimistic exploration. We assume the aleatoric uncertainty at a_1 and a_2 are the same, i.e., PDF curves overlap, but epistemic uncertainty is higher at a_1, causing $\mu_{\bar{Z}}(s,a_1) > \mu_{\bar{Z}}(s,a_2)$. Thus the CDF value is larger at a_1, which means the action with higher epistemic uncertainty is preferred.

More crucially, Figure 2(b) shows that OVD-Explorer could guide to avoid the area with higher aleatoric uncertainty given the equal optimism. We assume that the epistemic uncertainty at a_1 and a_2 are the same, causing their optimistic value estimation to be equal to $\mu_{\bar{Z}}(s,a)$, but aleatoric uncertainty is lower at a_1, i.e., PDF curve of $Z^\pi(s,a_1)$ is "thinner and taller". Thus the CDF value is larger at a_1, meaning that the action with lower aleatoric uncertainty is preferred.

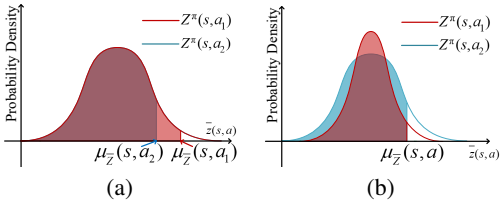


Figure 2: How OVD-Explorer explores (a) optimistically about epistemic uncertainty and (b) pessimistically about aleatoric uncertainty.

In the limit of $t \to \infty$, the epistemic uncertainty tend to be 0, and the aleatoric uncertainty estimated in Equation 12 converges to represent the true environment randomness. This is comparable to the second case (Figure 2(b)), and OVD-Explorer tends to choose the one with lower aleatoric uncertainty, which is a great advantage of our method over other OFU methods.

To summarize, the exploration guided by OVD-Explorer trades off between two criteria: exploring the areas with higher epistemic uncertainty, to ensure exploring optimistically following the OFU principle, and avoiding the areas with higher aleatoric uncertainty, to avoid over-exploration caused by high environment randomness.

4 OVD-Explorer for modern RL algorithms

For continuous action space, the argmax operator in Equation 7 is intractable. To address that, in this section, we derive the behavior policy π_E in closed form and state the scheme to incorporate it with existing policy-based algorithms.

First, we define the target policy π_T as the policy learned by any policy-based algorithm. To avoid the gap between training data and the policy, we need to constrain the difference between behavior policy π_E and target policy π_T, thus we derive π_E in the vicinity of π_T.

Second, to derive exploration π_E for modern RL algorithms with stochastic policy based on OVD-Explorer, where both the exploration $\pi_E = \mathcal{N}(\mu_E, \Sigma_E)$ and target policy $\pi_T = \mathcal{N}(\mu_T, \Sigma_T)$ are Gaussian distributions, we introduce the following proposition:

Proposition 1. *The OVD-Explorer behavior policy $\pi_E = \mathcal{N}(\mu_E, \Sigma_E)$ is as follows:*

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

and

$$\Sigma_E = \Sigma_T. \tag{16}$$

In specific, α is the step size controlling the exploration level and $m = \log \frac{\Phi_{Z^\pi(s,\mu_T)}(\bar{z}(s,\mu_T))}{C} + 1$ (see proof in Appendix B.2).

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

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

Algorithm 1 summarizes the overall procedure of OVD-Explorer, including the formulation of $\bar{Z}^\pi$ and Z^π (Line 2 and line 3) and behavior policy generation (Line 4). The generated behavior policy can be integrated with any modern policy-based RL algorithms to render the stable and well-performed algorithm. Please see Appendix C for the entire pseudo-code and details about how the OVD-Explorer-based behavior policy is incorporated with DSAC.

5 Experiments

To reveal the consistency between our theoretical analysis and the performance of OVD-Explorer algorithm, we conduct experiments to address the following questions:

Algorithm 1 The behavior policy (i.e., exploration policy) derived from OVD-Explorer.

Input: Current state s_t, current value distribution estimators θ_1, θ_2, current policy network ϕ.
Output: Behavior policy π_E.

1: Obtain target policy from policy-based RL algorithm $\pi_T(\cdot|s_t;\phi) \sim \mathcal{N}(\mu_T(s_t;\phi), \sigma_T(s_t;\phi))$
2: Derive OVD $\bar{Z}^\pi(s_t, \mu_T(s_t;\phi))$ using Eq. 10.
3: Construct value distribution of behavior policy $Z^\pi(s_t, \mu_T(s_t;\phi))$ using Eq. 13 or 14.
4: Calculate the mean of behavior policy distribution μ_E using Eq. 17.
5: **return** $\pi_E \sim \mathcal{N}(\mu_E, \sigma_T(s_t;\phi))$

RQ1 (Exploration): Can OVD-Explorer explore optimistically while avoiding over-exploration simultaneously?
RQ2 (Performance): Can OVD-Explorer handle complex and even stochastic tasks?
RQ3 (Sensitivity to α): Is OVD-Explorer sensitive to α that controls exploration magnitude?

5.1 Experiment Setup

Baselines include SAC (Haarnoja et al., 2018), DSAC (Ma et al., 2020), and DOAC, which is the distributional variant of OAC (Ciosek et al., 2019). We implement Gaussian and quantile formulations of Z^π as in Equation 13 and 14, which are denoted in the following as OVDE_G and OVDE_Q, respectively. We test OVD-Explorer on a novel task GridChaos and several tasks in Mujoco (Todorov et al., 2012) including the stochastic variants.

The appendix gives more details, including environment settings, implementation details, hyperparameters settings, and computing infrastructure used, as well as more experiment evaluations, including tasks with different noise scales, tasks with different episode horizon, the sensitivity to β and ablation study on the pessimistic formulation of $Z^\pi(s, a)$.

5.2 Exploration in GridChaos (RQ1)

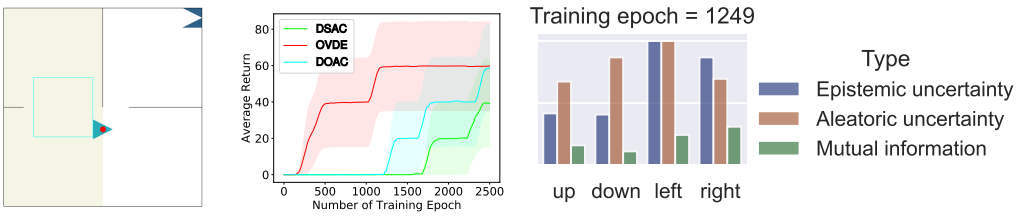
To illustrate the exploration pattern of OVD-Explorer and show the advantage of OVD-Explorer over DSAC and DOAC, we evaluate OVD-Explorer on a novel continuous and stochastic control task called GridChaos.

Figure 3(a) shows the map of GridChaos, in which we control the cyan triangle (agent) aiming to reach the fixed dark blue goal. The state is the current coordinate, and the action is a two-dimensional vector including the movement angle and distance. One episode terminates when the agent reaches the goal or maximum steps (i.e., MAX_STEP, typically 100). Also, it receives a +100 reward when reaching the goal otherwise 0. The reason why it is *chaos* is that the randomness of the transition is heterogeneous in different parts of this environment. Table 2 shows the environment settings. Moreover, Figure 3(b) shows that OVD-Explorer can reach the goal faster, with more efficient exploration.

Table 2: Settings of GridChaos.

	Value	Description
Observation[0]	[-1, 1]	X-coordinate
Observation[1]	[-1, 1]	Y-coordinate
Action[0]	[-1, 1]	Degree, mapped to $[-\pi, \pi]$
Action[1]	[-1, 1]	Distance, mapped to [0, MAX_STEP]
Noise_0	0.5	Variance of Gaussian noise in the left half of the map (default)
Noise_1	0.1	Variance of Gaussian noise in the right half of the map (default)

Figure 3(c) shows the values of uncertainty estimation and exploration objective (mutual information) taken four different actions at the state shown in Figure 3(a) and at training epoch 1249. Basically, Figure 3(c) illustrates that the estimated aleatoric uncertainty of left is higher than that of right, indicating that OVD-Explorer models aleatoric uncertainty properly. Further, OVD-explorer encourages to explore the right side at that time, since the value of exploration objective (in green) is highest. It implies that OVD-explorer tends to explore areas with higher epistemic uncertainty and avoiding higher aleatoric uncertainty, which is in accordance with our exploration principle. On the other hand, if high epistemic uncertainty is considered only like in OAC(Ciosek et al., 2019), the



(a) (b) (c)

Figure 3: GridChaos. (a) The cyan triangle can move to reach the goal at the top right. (b) The performance on it. (c) The values about uncertainty and exploration objective (mutual information).

agent would be guided to the left. Then the agent may be trapped in the left side due to the high aleatoric uncertainty, and it is the possible reason about why OAC fails tackling such heteroscedastic stochastic task. In appendix, we show more analysis about the exploration pattern, please refer to Appendix E.2 and Appendix E.10.

Besides, we conduct the evaluation when the randomness is high around the goal, as well as many other noise settings, the results enhance the ability of OVD-Explorer. More experiments that evaluate the performance on several other noise settings, as well as more detailed value histogram on different epoch in training process can be found in Appendix.

5.3 Performance on Mujoco tasks (RQ2)

To demonstrate the performance of OVD-Explorer more generally, we evaluate it on several Mujoco tasks. For 5 standard Mujoco tasks [1], the transition is deterministic and the randomness is only from the stochastic policy. For 5 noisy tasks, the heteroscedastic Gaussian noise is added in Mujoco tasks. Specifically, in each state transition on noisy tasks, Gaussian noise of different scales is randomly injected following a certain probability. Overall, we have the following findings.

Basically, OVD-Explorer can perform stably in the standard tasks with slight or little randomness, guiding efficient exploration. From the results[2] in standard Mujoco tasks in table 3, it is shown that in the relatively easy tasks, i.e., Hopper-v2, Reacher-v2, and HalCheetah-v2, OVD-Explorer does not obtain much gain beyond the baselines, which seems that these tasks are not profoundly demanding for exploration. In the high-dimension tasks, i.e., Ant-v2 as shown in Figure 4(a), OVD-Explorer significantly outperforms DSAC and DOAC. It means that the aleatoric uncertainty caused by policy indeed degrades the performance of DOAC, where the aleatoric and epistemic uncertainties are not distinguished.

Besides, OVD-Explorer can avoid the impact of heteroscedastic aleatoric uncertainty on exploration and thus improve robustness. As the results of noisy Mujoco tasks shown in Table 3, DOAC is worse than DSAC in most cases, which means that heteroscedastic aleatoric uncertainty causes significant degrades of DOAC. Simultaneously, OVD-Explorer significantly outperforms DOAC and DSAC, especially in Noisy Ant-v2 as shown in Figure 4(b).

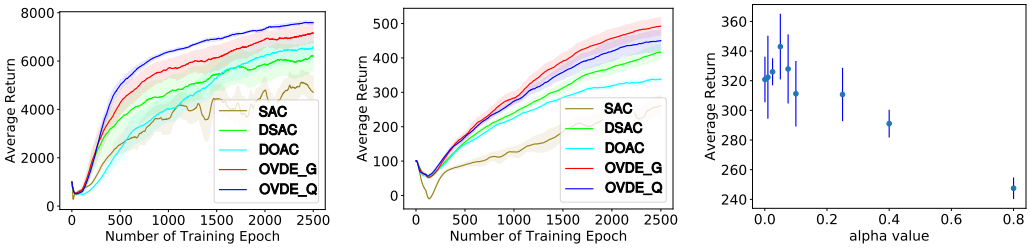
Furthermore, OVDE_G takes the Gaussian prior and concretely obtains better estimation of critic in the noisy tasks, while OVDE_Q performs better in the standard tasks due to the flexibility of quantile distribution. Generally, the difference between OVDE_G and OVDE_Q is that they use different formulations of $Z^\pi(s, a)$. Theoretically, the value function distribution in OVDE_Q is more flexible and should perform better than OVDE_G, as the reported results on five standard Mujoco tasks show. On the other hand, OVDE_G performs better in the noisy tasks, and it is because that the transition probability is Gaussian in the stochastic Mujoco tasks.

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

[2] Here, the reported results may be slightly different from previously reported results, partly due to the statistic approach, and partly due to the implementation. Nevertheless, the patterns of these baselines (e.g., DSAC outperforms SAC in the vast majority of cases) are consistent with previous results. The details about the implementation of baselines are described in the Appendix $D.2$.

Table 3: Comparisons of algorithms on five standard and five noisy tasks in Mujoco. We report the averaged performance and standard deviation of 5 runs. Each trail uses the mean undiscounted episodic return over the last 8% epoch (or at most the last 100 epoch) to avoid bias, and the total epoch number is shown in column *epoch*. The maximum value of each row is shown in bold.

TASK	EPOCH	SAC	DSAC	DOAC	OVD-EXPLORER_G	OVD-EXPLORER_Q
ANT-V2	2500	4706.2±1338.9	6206.9±1202.5	6586.7±1023.3	7160.6±763.2	**7590.3**±154.9
HALFCHEETAH-V2	2500	12373.8±860.9	13890.0±3424.4	12977.0±140.4	14084.5±1579.8	**14792.4**±997.4
HOPPER-V2	1250	**2751.8**±775.7	2199.7±602.7	2215.1±557.1	2239.5±428.2	2619.3±457.0
REACHER-V2	250	-21.6±2.5	-11.9±0.5	-19.8±1.7	-11.6±2.4	**-10.8**±1.4
INVDBPENDULUM-V2	300	9344.0±28.4	**9359.6**±0.1	5109.4±3638.7	9128.0±460.6	9351.3±16.3
N-ANT-V2	2500	261.5±57.6	416.38±42.16	337.39±11.96	**492.54**±50.44	450.34±58.42
N-HALFCHEETAH-V2	1250	351.91±6.68	431.39±35.68	417.47±39.62	**445.28**±37.52	429.63±34.45
N-HOPPER-V2	1250	207.06±19.49	244.53±4.71	242.74±7.87	**252.09**±7.82	237.68±13.11
N-PUSHER-V2	1250	-46.92±12.12	-25.31±2.46	-39.13±9.06	**-23.41**±0.69	-28.51±4.53
N-INVDBPENDULUM-V2	300	934.36±1.91	932.78±4.02	496.61±205.8	933.67±1.54	**934.64**±0.95



(a) Ant-v2 (b) Noisy Ant-v2 (c) Noisy Ant-v2

Figure 4: Training curves on (a) Ant-v2 and (b) Noisy Ant-v2. The x-axis indicates the number of training epochs, while the y-axis is the evaluation result represented by the average episode return. The shaded region denotes the half standard deviation of average evaluation over 5 seeds. Curves are smoothed uniformly for visual clarity. (c) Sensitivity to α. The x-axis indicates different α settings, while the y-axis is the evaluation result represented by average episode return in the last 100 epoch before total 1250 epoch. Error bars indicate half standard deviation of average evaluation over 5 seeds. The 9 different α values are 0.0005, 0.01. 0.025, 0.05, 0.075, 0.1, 0.25, 0.4, 0.8, respectively.

5.4 SENSITIVITY TO α (RQ3)

The step size α in Equation 15 controls how much the behavior policy derived from OVD-Explorer, i.e. π_E, is far away from the target policy π_T, which can be essential for exploration benefit. To investigate the appropriate range of it, we test several α value on Noisy Ant-v2 task using OVDE_G, and the result is shown in Figure 4(c). If α is quite small, OVD-Explorer degenerates to DSAC and implies little exploration. In contrast, if α is larger, the performance becomes worse because of the huge gap between behavior policy and target policy. The result demonstrates that α should be taken in a suitable range to facilitate more adequate exploration. In our experiments, we uniformly use α to be equal to 0.05 to show the performance of the OVD-Explorer. In addition, we find that a smaller α leads to higher gains when the task is much more difficult, and the details can be found in Appendix E.6. We also verify the sensitivity of OVD-Explorer to β in Appendix E.7 and found that there is a broad range of settings for β, which can lead to well performance.

6 CONCLUSION

This paper proposes an information-theoretic exploration method OVD-Explorer, which introduces a novel measurement of exploration ability, i.e., the mutual information between the policy and the upper bounds of return. By maximizing the mutual information, OVD-Explorer is able to derive the behavior policy, that follows the OFU principle, and further avoids exploring the areas with high aleatoric uncertainty. Integrated with SAC, OVD-Explorer addresses the negative impact of aleatoric uncertainty for exploration in continuous RL for the first time.

REFERENCES

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.

Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Max-value entropy search for multi-objective bayesian optimization with constraints. *CoRR*, abs/2009.01721, 2020.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 530–539. PMLR, 2018.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017.

Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1779–1788. PMLR, 2020. URL `http://proceedings.mlr.press/v119/cheng20b.html`.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 1785–1796, 2019.

William R. Clements, Benoît-Marie Robaglia, Bastien Van Delft, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *CoRR*, abs/1905.09638, 2019.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1104–1113. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2892–2901. AAAI Press, 2018b.

Richard Dearden, Nir Friedman, and Stuart J. Russell. Bayesian q-learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA*, pp. 761–768, 1998.

Lior Fox, Leshem Choshen, and Yonatan Loewenstein. Dora the explorer: Directed outreaching reinforcement action-selection. In *International Conference on Learning Representations*, 2018.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1582–1591. PMLR, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1856–1865. PMLR, 2018.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory, COLT 1993, Santa Cruz, CA, USA, July 26-28, 1993*, pp. 5–13. ACM, 1993.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pp. 1109–1117, 2016.

Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. EMI: exploration with mutual information. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3360–3369, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Johannes Kirschner and Andreas Krause. 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*, pp. 358–384. PMLR, 2018.

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

Shibo Li, Wei Xing, Robert M. Kirby, and Shandian Zhe. Multi-fidelity bayesian optimization via deep neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Xiaoteng Ma, Qiyuan Zhang, Li Xia, Zhengyuan Zhou, Jun Yang, and Qianchuan Zhao. Distributional soft actor critic for risk sensitive learning. *CoRR*, abs/2004.14547, 2020.

Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2471–2478, 2017.

Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4424–4434. PMLR, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937, 2016.

Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Efficient exploration with double uncertain value networks. *CoRR*, abs/1711.10789, 2017.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 271–279, 2016.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4026–4034, 2016.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2721–2730, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2778–2787, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5062–5071. PMLR, 2019.

Valerio Perrone, Iaroslav Shcherbatyi, Rodolphe Jenatton, Cédric Archambeau, and Matthias W. Seeger. Constrained bayesian optimization with max-value entropy search. *CoRR*, abs/1910.07003, 2019.

Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphaël Marinier, Marc Pollefeys, Timothy P. Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

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

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2753–2762, 2017.

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

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pp. 5026–5033. IEEE, 2012.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. volume 70 of *Proceedings of Machine Learning Research*, pp. 3627–3635, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Christopher J. C. H. Watkins and Peter Dayan. Technical note q-learning. *Mach. Learn.*, 8:279–292, 1992.

Fan Zhou, Jianing Wang, and Xingdong Feng. Non-crossing quantile regression for distributional reinforcement learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.