# vMFER: Von Mises-Fisher Experience Resampling Based on Uncertainty of Gradient Directions for Policy Improvement

**Yiwen Zhu**[1,2,3], **Jinyi Liu**[4], **Wenya Wei**[1], **Qianyi Fu**[1], **Yujing Hu**[2*], **Zhou Fang**[1*], **Bo An**[3,5], **Jianye Hao**[4], **Tangjie Lv**[2] and **Changjie Fan**[2]

[1]Zhejiang University
[2]NetEase Fuxi AI Lab
[3]Nanyang Technological University
[4]Tianjin University
[5]Skywork AI

{evanzhu, wwy_vivian, qyfu, zfang}@zju.edu.cn, {jyliu, jianye.hao}@tju.edu.cn,
{huyujing, hzlvtangjie, fanchangjie}@corp.netease.com, boan@ntu.edu.sg

## Abstract

Reinforcement Learning (RL) is a widely employed technique in decision-making problems, encompassing two fundamental operations – policy evaluation and policy improvement. Enhancing learning efficiency remains a key challenge in RL, with many efforts focused on using ensemble critics to boost policy evaluation efficiency. However, when using multiple critics, the actor in the policy improvement process can obtain different gradients. Previous studies have combined these gradients without considering their disagreements. Therefore, optimizing the policy improvement process is crucial to enhance learning efficiency. This study focuses on investigating the impact of gradient disagreements caused by ensemble critics on policy improvement. We introduce the concept of uncertainty of gradient directions as a means to measure the disagreement among gradients utilized in the policy improvement process. Through measuring the disagreement among gradients, we find that transitions with lower uncertainty of gradient directions are more reliable in the policy improvement process. Building on this analysis, we propose a method called von Mises-Fisher Experience Resampling (vMFER), which optimizes the policy improvement process by resampling transitions and assigning higher confidence to transitions with lower uncertainty of gradient directions. Our experiments demonstrate that vMFER significantly outperforms the benchmark and is particularly well-suited for ensemble structures in RL.

## 1 Introduction

Over the past few years, there has been rapid progress in the field of reinforcement learning (RL), leading to impressive achievements in tackling complex tasks [Wu *et al.*, 2023; Radosavovic *et al.*, 2023; Abeyruwan *et al.*, 2023]. Despite these advancements, the challenge of enhancing learning efficiency persists.

In general, reinforcement learning involves two fundamental operations: policy evaluation and policy improvement [Sutton and Barto, 2018]. To enhance learning efficiency and optimality, numerous methods optimize the policy evaluation process by using ensemble critics, such as Double Q-learning [Hasselt, 2010], SAC [Haarnoja *et al.*, 2018a; Haarnoja *et al.*, 2018b], TD3 [Fujimoto *et al.*, 2018] and REDQ [Chen *et al.*, 2021]. Nevertheless, the utilization of ensemble critics often introduces disagreements in the direction of policy optimization during the policy improvement process. Existing methods like SAC, TD3, or REDQ simply aggregate the multiple gradients generated during policy improvement into a single gradient, without considering the disagreements among these gradients caused by ensemble critics. One alternative approach is to enhance the reliability of gradients in policy improvement, such as utilizing the delayed policy update method employed by REDQ and TD3. This method uses more reliable ensemble critics to ensure a more concentrated gradient direction. However, this approach does not account for the discrepancies among transitions, resulting in delayed updates for all sampled transitions.

We propose that by selectively avoiding delayed updates for transitions that can provide a reliable gradient under the current ensemble critics, the policy improvement process can be further optimized. We introduce additional indicators to measure the reliability of Q-ensembles under current ensemble critics. This allows us to identify which transition data is more appropriate for utilization in the policy improvement process. We posit that as the accuracy of the ensemble critics increases, the directions of policy gradients provided by the same transition under the ensemble structure will demonstrate a high concentration in the policy improvement process. Hence, we introduce the concept of uncertainty of gradient directions to identify the reliability of transitions under the current ensemble structure during the policy improvement process. From a directional statistics perspective [Mar-

*The corresponding authors.

dia *et al.*, 2000], these directions of the policy gradients can be modeled as a distribution. Considering the computational cost, we use the von Mises-Fisher distribution [Fisher, 1953] to quantify such uncertainty associated with each transition. Furthermore, we propose the von Mises-Fisher Experience Resampling (vMFER) algorithm which leverages the uncertainty of gradient directions to resample transitions for policy improvement. To improve the efficiency of the policy improvement process, we enhance the sampling probability of transitions with lower uncertainty while reducing the likelihood of sampling transitions with higher uncertainty during the policy improvement process.

Our primary contributions are threefold:

1. We introduce a metric to measure the uncertainty of gradient directions, aimed at evaluating the reliability of transitions used in the policy improvement process. This metric is calculated by analyzing the discrepancy in gradient directions, which are induced by ensemble critics for each transition.
2. We propose the vMFER algorithm to optimize policy improvement by resampling transitions based on the uncertainty of gradient directions. Moreover, it is compatible with most actor-critic algorithms utilizing the ensemble structure.
3. Our approach performs effectively in Mujoco control tasks [Brockman *et al.*, 2016]. This indicates the potential of vMFER for a wide range of applications.

## 2 Preliminary

### 2.1 Actor-critic Framework

The Actor-Critic framework is widely used in RL, consisting of two distinct modules: the actor network that learns the policy, and the critic network that learns the value function.

Several RL algorithms have been proposed based on the actor-critic framework. Algorithms like PPO [Schulman *et al.*, 2017] and DDPG [Lillicrap *et al.*, 2015] use a single critic structure, while others like TD3 [Fujimoto *et al.*, 2018] and SAC [Haarnoja *et al.*, 2018a] use an ensemble structure with multiple critics to overcome the problem of overestimation, arising due to the maximization of a noisy value estimate during the critic learning process [Thrun and Schwartz, 1993]. The ensemble structure results in multiple Q-values for each transition, allowing for the calculation of multiple actor network losses and generating multiple gradients for actor network parameter updates. The loss function for a mini-batch of transitions in actor training is commonly expressed as $\mathbb{E}_{(s,a)\sim D}\left[\log\pi(a|s) - \min_i Q_i(s,a)\right]$ [Haarnoja *et al.*, 2018a], $\mathbb{E}_{(s,a)\sim D}\left[-Q_1(s,a)\right]$ [Lillicrap *et al.*, 2015; Fujimoto *et al.*, 2018], or $\mathbb{E}_{(s,a)\sim D}\left[\frac{1}{N}\sum_i^N\left[\log\pi - Q_i(s,a)\right]\right]$ [Chen *et al.*, 2021]. Here, $D$ represents the replay buffer, and the subscript of critic $Q$ denotes the index number of ensemble critics, $\pi$ refers to the policy.

The gradients of mini-batch transitions provided for policy updates are usually integrated by averaging these gradients to update actor network parameters.

### 2.2 Von Mises-Fisher Distribution

The von Mises-Fisher (vMF) distribution [Fisher, 1953] is one of the most basic probability distributions in high-dimensional directional statistics [Mardia *et al.*, 2000]. It characterizes a probability distribution on the $(p-1)$-sphere in $\mathbb{R}^p$, defined on the unit hypersphere. To be more specific, the probability density function of the von Mises-Fisher distribution for a random $p$-dimensional unit vector $\mathbf{x} \sim \text{vMF}(k,\mu)$ is expressed in Eq. (1), where $f_p$ represents the density function for $\text{vMF}(k,\mu)$.

$$\begin{gathered}\mathbf{x} \sim \text{vMF}(k,\mu),\\ f_p(\mathbf{x};\mu,k) = C_p(k)\exp(k\mu^{\mathrm{T}}\mathbf{x}).\end{gathered} \tag{1}$$

In this function, $C_p(k)$ represents the normalization constant, and $k \geq 0$ represents the concentration parameter. Furthermore, the mean direction $\mu$ can be calculated as demonstrated in Eq. (2) [Mardia *et al.*, 2000].

$$\mu = \bar{\mathbf{x}}/\bar{\mathbf{R}}, \text{where } \bar{\mathbf{x}} = \frac{1}{N}\sum_i^N x_i,\ \bar{\mathbf{R}} = ||\bar{\mathbf{x}}||_2. \tag{2}$$

The concentration parameter $k$ is commonly used to indicate the degree of clustering and scattering of the vector direction distribution. However, estimating the concentration parameter $k$ using the Maximum-likelihood estimate is often challenging because of the difficulty in calculating the modified Bessel function [Sra, 2012]. Banerjee et al. [Banerjee *et al.*, 2005] proposed a simpler approximation given by:

$$\hat{k} = \frac{\bar{\mathbf{R}}(p-\bar{\mathbf{R}}^2)}{(1-\bar{\mathbf{R}}^2)}. \tag{3}$$

which avoids calculating Bessel functions [Sra, 2012].

## 3 Von Mises-Fisher Experience Resampling

In the actor-critic framework with the ensemble structure, each ensemble critic can theoretically contribute a gradient for updating the actor during the policy improvement process. Despite this, existing Actor-Critic frameworks [Fujimoto *et al.*, 2018; Haarnoja *et al.*, 2018a; Lillicrap *et al.*, 2015; Chen *et al.*, 2021] often merge these multiple gradients into a single gradient for policy improvement, disregarding the information that can be derived from the discrepancies among these gradients. Specifically, these frameworks often overlook the uncertain information in different gradients generated by the same transition. This information can be utilized to evaluate the reliability of the gradient provided by that particular transition for policy improvement.

To address these limitations, this paper proposes the use of the von Mises-Fisher distribution to describe the uncertainty of gradient directions. Subsequently, we propose the von Mises-Fisher Experience Resampling (vMFER) algorithm, which involves resampling the transitions using probabilities calculated based on the uncertainty of gradient directions for policy improvement. By decreasing the probability of sampling transitions with high uncertainty, the policy improvement can be made more reliable. Furthermore, we provide a straightforward example and a toy experiment to demonstrate

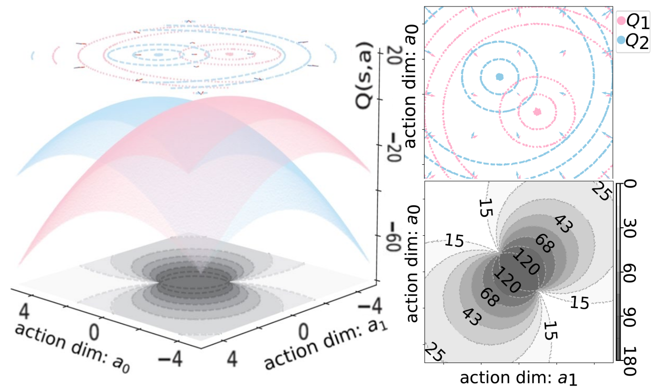



Figure 1: Multiple Q-values and their corresponding gradients $\frac{\partial Q_i(s_t,a)}{\partial a}$ are generated by the ensemble critic function for a given state input $s_t$, with each Q-value and gradient pair corresponding to a different action $a$. **Left:** The ensemble critic function $Q_1(s_t,\cdot)$ and $Q_2(s_t,\cdot)$, which are depicted as multi-dimensional surfaces. **Right Upper:**The gradients of the ensemble critics, represented as arrows with varied colors on the contour plots of the Q-values, illustrate the direction and magnitude of the action-value function's sensitivity to changes in action space. **Right Lower:** The angles between the gradients, $\frac{\partial Q_1(s_t,a)}{\partial a}$ and $\frac{\partial Q_2(s_t,a)}{\partial a}$, are used to quantify the uncertainty of gradients for different actions $a$.

the uncertainty of gradient directions in the ensemble structure, before explaining in-depth how this indicator contributes to enhancing algorithm performance.

### 3.1 Exploring Disagreements in Gradient Directions: A Simple Example

To interpret the uncertainty of gradient directions, we present an example of policy evaluation on a two-critic ensemble. To highlight the disagreement among ensemble critics during the learning process, we have formulated two critics given current state $\mathbf{s_t}$ and action $\mathbf{a}$:

$$\begin{aligned} Q_1(\mathbf{s_t},\mathbf{a}) &= -(\mathbf{a}+\mathbb{1}_{2\times1}+\epsilon)^T(\mathbf{a}+\mathbb{1}_{2\times1}+\epsilon), \\ Q_2(\mathbf{s_t},\mathbf{a}) &= -(\mathbf{a}-\mathbb{1}_{2\times1}+\epsilon)^T(\mathbf{a}-\mathbb{1}_{2\times1}+\epsilon). \end{aligned} \tag{4}$$

where $\mathbf{a}\in\mathbb{R}^{2\times1}$ and $\epsilon\sim N(0,0.01)$. The left of Figure 1 shows the variations of the output $Q$ for different action inputs of the two critic networks. The three axes represent two dimensions of the action and the value of $Q$, respectively.

Then we establish the optimization objective for the actor network, similar to the conventional continuous RL: $\max_{\mathbf{a}} Q(\mathbf{s_t},\mathbf{a})$. This allows us to compute the gradient $\frac{\partial Q(\mathbf{s_t},\mathbf{a})}{\partial\mathbf{a}}$ and determine the convergence direction for the desired action based on the current critic network and $s_t$. Due to the ensemble critics, multiple gradients can be computed for each transition in the policy improvement process, as shown in the right upper of Figure 1. From the perspective of policy improvement, the presence of disagreements among gradients becomes apparent, underscoring the importance of their judicious utilization.

Obviously, a metric is required to quantify the disagreement in gradients during the policy improvement process. In this study, we employ the uncertainty of the gradient directions as a measure of the extent of these disagreements. As demonstrated in the right lower of Figure 1, the larger angle represents more significant disagreements among gradient directions, indicating substantial conflicts in gradients for specific action inputs. Using this metric, we can assign a confidence level to the corresponding transition, where the magnitude of confidence should be inversely proportional to the level of uncertainty.

### 3.2 The Necessity of Measuring Gradient Uncertainty: A Toy Experiment

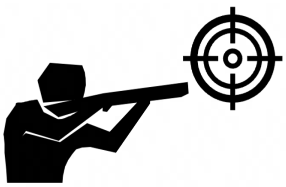
To demonstrate the influence of resampling on policy improvement, we introduce an artificial environment called the "Shooting" environment (depicted in Figure 2(a)). This environment is a one-step Markov Decision Process (MDP) with a continuous action space, and the optimal action is indicated by a green star in Figure 2(d). The closer the policy's action output is to the optimal action, the higher the reward obtained.

In this experimental setup, we employ three distinct approaches to investigate the effects of resampling transitions during the policy improvement phase in such a one-step MDP environment. The '**Uniform**' method involves uniformly sampling transitions. In contrast, the '**Uncertainty**' method selectively uses transitions with lower uncertainty of gradient directions. Finally, the '**Oracle**' method chooses transitions that can guide the updated action toward the optimal outcome.

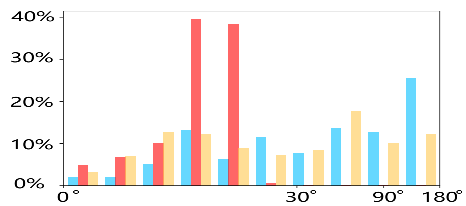
As we focus on a one-step MDP, the actions generated by the actor exhibit variation when subjected to different policy improvement techniques during training, yet the policy evaluation remains constant. This dynamic is carefully traced and depicted in Figure 2(d). Consistent with the setting outlined in Section 3.1, we utilize red and blue contour maps to represent Q-values. The uncertainty of gradient directions is quantified by measuring the angles between these gradients. Furthermore, this uncertainty is visually captured through a grey contour map, where darker shades indicate a higher degree of discrepancy in gradient directions for specific action inputs, as depicted in Figure 2(d). Figure 2(c) presents a comparison of episode rewards achieved by agents trained using different approaches. Additionally, Figure 2(b) captures the distribution of angles between gradients associated with the actions output by policies, which have been updated through various methods across the entire training process. Notably, we find that the '**Oracle**' and '**Uniform**' methods yield a relatively even distribution of uncertainty levels in the transitions used during the policy improvement phase. Conversely, the '**Uncertainty**' method tends to select transitions with lower uncertainty for policy improvement, namely smaller angles.

In contrast, the '**Oracle**' method represents an ideal approach. However, its practical application is limited due to the prerequisite of knowing the optimal policy beforehand, which is often not feasible in real-world scenarios. Intriguingly, while the '**Uncertainty**' method may exhibit lower efficiency in policy improvement compared to the '**Oracle**' method, it excels in terms of the policy update trajectory and offers greater practical applicability. Moreover, both the '**Oracle**' and '**Uncertainty**' outperform the '**Uniform**' approach.

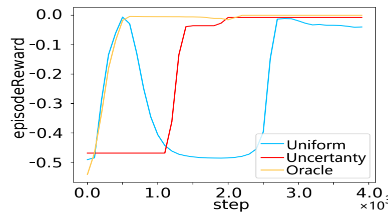
Our research indicates that the careful selection of transitions for policy improvement is crucial for enhancing learning efficiency. Specifically, the resampling method that leverages



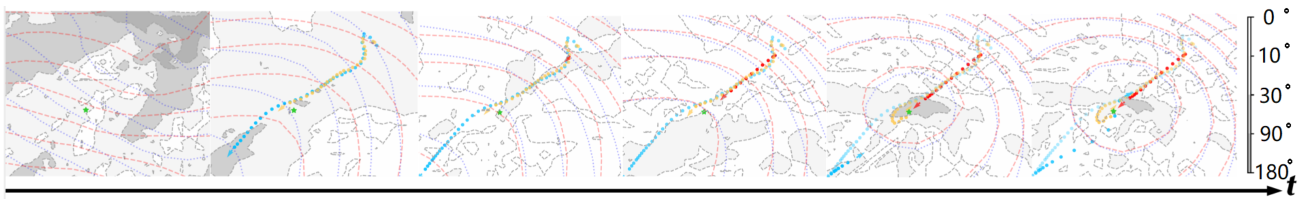
(a) Shooting environment: an one step MDP




(b) The distributions of angles between gradients during training process by various approaches




(c) Episode rewards achieved by various approaches




(d) Trajectory of policy changes during update

Figure 2: A toy experiment illustrating the advantage of considering the uncertainty of gradient directions on the learning efficiency of policy improvement. The experiment compares three approaches: '**Uniform**' involves uniformly sampling from the transitions, '**Uncertainty**' utilizes only transitions with low uncertainty of gradient directions, and '**Oracle**' employs only transitions that update the action in the direction of the optimal action.

the uncertainty of gradient directions is effective in optimizing the learning process. Essentially, less concentrated gradient directions signal higher uncertainty, with more uncertainty indicating greater divergence in these directions.

**Remark 1.** *Transitions under the current ensemble critics with higher uncertainty of gradient directions should be less likely to be employed for policy improvement.*

### 3.3 How To Measure Gradient Uncertainty: Via Von Mises-Fisher Distribution

Our aim is to identify a metric that can determine which transitions contribute reliable gradients for actor updates. Figure 1 illustrates that this metric should describe the concentration of gradients contributed by the same transition under different indices of the ensemble critics. It's essential to emphasize that, in this context, the direction of the gradient is more crucial than its length. Comparatively, an incorrect gradient descent direction is less acceptable than an incorrect magnitude. Because, unlike the latter, the former implies ineffective optimization. Moreover, the metric is not too complex to compute, since in theory, we need to compute the metric for each transition before updating the actor.

In the field of directional statistics, few distributions align with our criteria. The Bingham [Bingham, 1974] and Kent [Kent, 1982] distributions, while noteworthy, require the computation of the Bessel function [Bowman, 2012], thereby not fulfilling our need for low computational overhead. In contrast, the von Mises-Fisher distribution [Fisher, 1953; Watson, 1982; Mardia *et al.*, 2000], particularly when employing Banerjee's [Banerjee *et al.*, 2005] method for parameter estimation, circumvents the need for Bessel function calculations. Notably, the concentration parameter $k$ we required is expressed simply. The efficacy and accuracy of this approximation method are well-demonstrated by [Sra, 2012]. Additionally, employing vMF to model gradient directions offers advantages of scalability and a threshold-free setup, compared to using the angle between gradients as mentioned in Section 3.1 and Section 3.2.

Hence, we assume that the directions of the gradients are sampled from the von Mises-Fisher distribution. To measure the uncertainty of the gradient directions, the parameters of the distribution need to be estimated. Using the ensemble structure enables us to compute numerous actor losses and their corresponding gradients with respect to the actor network parameters $\theta$ using the same transition $(s_t, a_t, r_t, s'_t)$, as illustrated in Eq. (5).

$$\begin{aligned} \mathbf{L}(s_t,a) &= \begin{bmatrix} l_1(s_t,a) & \cdots & l_n(s_t,a) \end{bmatrix}^{\mathrm{T}}, \\ \frac{\partial \mathbf{L}(s_t,a)}{\partial \theta} &= \begin{bmatrix} \frac{\partial l_1(s_t,a)}{\partial \theta} & \cdots & \frac{\partial l_n(s_t,a)}{\partial \theta} \end{bmatrix}^{\mathrm{T}}, \ a \sim \pi(\cdot|s_t). \end{aligned} \tag{5}$$

where $n$ is the ensemble size of the critic networks. Estimating the parameter of the Von Mises-Fisher distribution for $\frac{\partial l_i}{\partial \theta}$ can be challenging due to the high dimensionality of $\theta$. This can result in issues such as increased computational cost and a large scale of the estimated concentration parameter $k$. To mitigate these issues, it is advisable to reduce the dimensionality of the gradients. By applying the chain rule, it is evident that $\frac{\partial a}{\partial \theta}$ is constant for the same transition. As a result, the uncertainty in the gradient directions primarily arises from $\frac{\partial l_i}{\partial a}, i \in [1, n]$. Therefore, calculating the uncertainty of $\frac{\partial l_i}{\partial a}$ rather than $\frac{\partial l_i}{\partial \theta}$ is a more cost-effective and scalable approach.

$$\frac{\partial l_i(s_t,a)}{\partial \theta} = \frac{\partial l_i(s_t,a)}{\partial a}\frac{\partial a}{\partial \theta}, \qquad a \sim \pi(\cdot|s_t). \tag{6}$$

Let $x_i(s_t) = ||\frac{\partial l_i(s_t,a)}{\partial a}||_2^{-1} \cdot \frac{\partial l_i(s_t,a)}{\partial a}|_{a\sim\pi(\cdot|s_t)}$ and $\mathbf{x}(s_t) = \sum_i x_i(s_t)/n$, where $x_i(s_t)$ denote the direction of the gra-

**Algorithm 1** Von Mises-Fisher Experience Resampling (Based on TD3)

Initialize replay buffer $\mathcal{D}$ and ensemble number $N$
Initialize critic networks $\{Q_{\phi_i} \mid i \in [1, N]\}$, and actor network $\pi_\theta$ with random parameters $\{\phi_i \mid i \in [1, N]\}, \theta$
Initialize target network $\{\phi_i{}' \leftarrow \phi_i \mid i \in [1, N]\}, \theta' \leftarrow \theta$
Initialize sampling factors $p_j = 1$ for each transition of $\mathcal{D}$
**for** $t$=1 to $T$ **do**
  sample action with noise $a \leftarrow \pi_\theta(s) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma)$
  store transition $(s, a, r, s')$ in $\mathcal{D}$
  sample mini-batch of $b$ transitions $(s, a, r, s')$ from $\mathcal{D}$
  $a' \leftarrow \pi_\theta(s') + \epsilon,\ \epsilon \sim clip(\mathcal{N}(0, \sigma), -c, c)$
  $y \leftarrow r + \gamma \min_{i=1,2} Q_{\phi_i}(s', a')$
  Update critics $\phi_i \leftarrow argmin_{\phi_i} b^{-1} \sum (y - Q_{\phi_i}(s, a))^2$
  **if** $t$ mod 2 **then**
    **for** $j = 1$ to $b$ **do**
      \\Resample transition
      $(s_j, a_j, r_j, s'_j) \sim P(j) = \frac{p_j}{\sum_m p_m}$ ▷ Eq. (8)
      \\Sample action
      $\hat{a}_j \leftarrow \pi_\theta(s_j) + \epsilon,\ \epsilon \sim clip(\mathcal{N}(0, \sigma), -c, c)$
      \\Calculate the actor losses
      $l_i(s_j, \hat{a}_j) = -Q_{\phi_i}(s_j, \hat{a}_j), \quad i \in [1, N]$
      \\Calculate gradients of losses
      $g_i = \frac{\partial l_i(s_j, a)}{\partial a}|_{a=\hat{a}_j}, \quad i \in [1, N]$
      $p_j \leftarrow$ Update Sampling Factor $p_j$ (Algorithm 2)
    Update $\theta$ by the deterministic policy gradient:
    $\nabla_\theta J(\theta) = -b^{-1} \sum_j \nabla_\theta l_1(s_j)$
    Update target networks:
    $\phi'_i \leftarrow \tau \phi_i + (1 - \tau)\phi'_i, \theta' \leftarrow \tau\theta + (1 - \tau)\theta'$

dient contributed by the current actor loss, assuming that $x_i \sim$ vMF$(k, \mu)$, then according to Banerjee's method we can estimate the concentration parameter $k$ and mean direction $\mu$ as demonstrated in Eq. (7).

$$\mathbf{R}(s_t) = ||\mathbf{x}(s_t)||_2, \quad \mu(s_t) = \frac{\mathbf{x}(s_t)}{\mathbf{R}(s_t)},$$
$$k(s_t) = \frac{\mathbf{R}(\mathbf{s_t})(p - \mathbf{R}^2(s_t))}{(1 - \mathbf{R}^2(s_t))} \quad \propto \quad \mathbf{R}(s_t). \tag{7}$$

Here, $p$ denotes the dimension of the action. The concentration parameter $k$ is used to articulate the uncertainty present in the current gradient directions. Clearly, we can prove that $\mathbf{R}$ is proportional to $k$. Hence, $\mathbf{R}$ possesses a similar capability to represent uncertainty.

### 3.4 How To Use vMF Distribution: Von Mises-Fisher Experience Resampling

Our objective is to independently fit von Mises-Fisher distributions to the gradients from each transition and evaluate their uncertainty levels to ascertain the probability of each transition's utilization. A higher level of uncertainty implies a reduced likelihood of sampling the data. The gradient directions of each data corresponds to a distinct von Mises-Fisher distribution. We define the likelihood of the prior distribution for each transition being sampled as $P(j|\mathcal{D}) = \frac{1}{M}$, where $\mathcal{D}$ denotes the transitions in the replay buffer and $M$ the total number of data points in it. Subsequently, we represent the conditional probability distribution

**Algorithm 2** Update Sampling Factors $p_j$

**Input:** $p_j, s_j, \{g_1, g_2, \cdots, g_N\}, \{Q_{\phi_1}, Q_{\phi_2}, \cdots, Q_{\phi_N}\}, \pi_\theta$
Calculate normalized unit vector $x_i = \frac{g_i}{||g_i||_2}, i \in [1, N]$
Calculate other parameters:
  $\mathbf{R}(s_j) = ||\frac{\sum_i x_i}{N}||_2$ and $\mu(s_j) = \frac{\sum_i x_i}{N\mathbf{R}(s_j)}$ ▷ Eq. (7)
Choose index of critics used in the policy improvement:
  $e = \begin{cases} \underset{i}{argmin}\, Q_{\phi_i}(s_j, \pi_\theta(s_j)) & \textbf{(SAC)} \\ 1 & \textbf{(TD3)} \end{cases}$
$p_j \leftarrow \begin{cases} \exp(\mathbf{R}(s_j)\mu^{\mathrm{T}}(s_j)x_e(s_j)) & \textbf{(uncertainty)} \\ rank(\exp(\mathbf{R}(s_j)\mu^{\mathrm{T}}(s_j)x_e(s_j)))^{-1} & \textbf{(rank)} \end{cases}$ ▷ Eq. (9,10)
**Output:** $p_j$

as $P(x(s_j)|j, \mathcal{D}) = C_p(k(s_j)) \exp(k(s_j)\mu^{\mathrm{T}}(s_j)x(s_j))$. The posterior distribution, which we aim to achieve, is denoted by $P(j|x(s_j), \mathcal{D}) = \frac{p_j}{\sum_m p_m}$. Here, $p_j$ indicates the sampling factor for the specific transition $(s_j, a_j, r_j, s'_j)$. This framework enables us to derive the posterior distribution of the resampling probability for the current transition after sampling the gradient direction $x(s_j) \sim$ vMF$(k(s_j), \mu(s_j))$, as elaborated in Eq. (8).

$$\prod_{j=1}^{M} \frac{1}{M} C_p(k(s_j)) \exp(k(s_j)\mu(s_j)^{\mathrm{T}} x(s_j))$$
$$\propto \prod_{j=1}^{M} P(j|x(s_j), \mathcal{D}) = \prod_{j=1}^{M} \frac{p_j}{\sum_m p_m}. \tag{8}$$

As $k \propto R$, and the dimensionality of action heavily affects the value range of $k$, we simplify the calculation and set the probability of sampling each transition using Eq. (9).

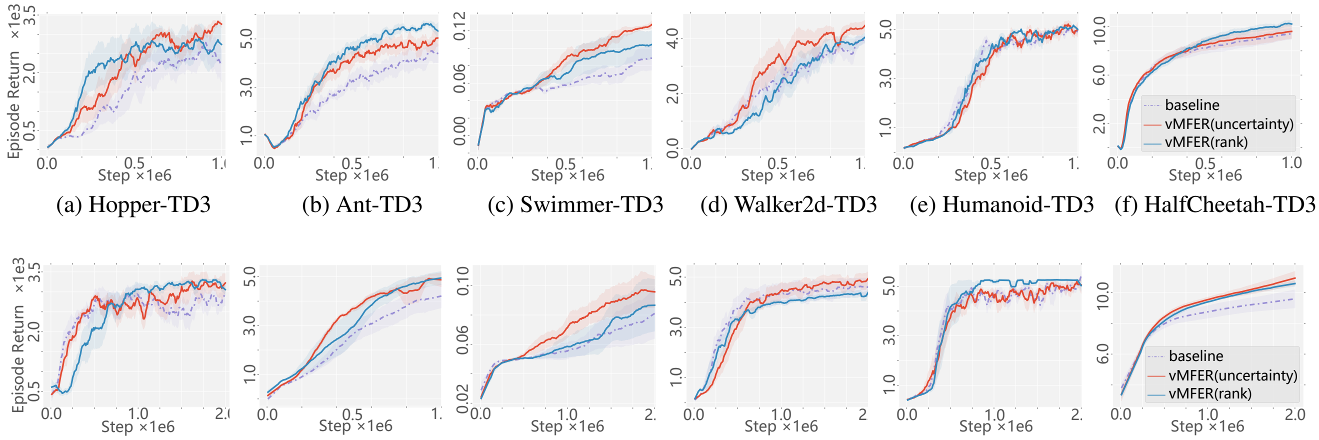
$$P(j|x(s_j), \mathcal{D}) = \frac{\exp(\mathbf{R}(s_j)\mu^{\mathrm{T}}(s_j)x(s_j))}{\sum_i^M \exp(\mathbf{R}(s_i)\mu^{\mathrm{T}}(s_i)x(s_i))}. \tag{9}$$

The second approach involves an indirect rank-based method. Here, the probability of the sampling transition is calculated as shown in Eq. (10). The rank of transition $(s_j, a_j, r_j, s'_j)$ is determined by sorting the replay memory based on $\exp(\mathbf{R}(s_j)\mu^{\mathrm{T}}(s_j)x(s_j))$ from high to low.

$$P(j|x(s_j), \mathcal{D}) = \frac{rank(\exp(\mathbf{R}(s_j)\mu^{\mathrm{T}}(s_j)x(s_j)))^{-1}}{\sum_i^M rank(\exp(\mathbf{R}(s_i)\mu^{\mathrm{T}}(s_i)x(s_i)))^{-1}}. \tag{10}$$

Our method seamlessly integrates with any Actor-Critic framework algorithm, given that the critic network employs an ensemble structure. Taking TD3 as an example, we combine it with our approach, as detailed in Algorithm 1. The key modifications in our enhanced version, marked in brown in Algorithm 1, primarily optimize the policy improvement process. The probability of resampling each transition is guided by the uncertainty of gradient directions, and these resampled transitions are subsequently used for policy improvement, aiding in determining the actor's update direction.

It is evident that our method can act as a versatile plugin, which, through Algorithm 2, updates and maintains the sampling factor $p_j$ for each transition. This allows for effective resampling during the policy improvement process.




(a) Hopper-TD3 (b) Ant-TD3 (c) Swimmer-TD3 (d) Walker2d-TD3 (e) Humanoid-TD3 (f) HalfCheetah-TD3

(g) Hopper-SAC (h) Ant-SAC (i) Swimmer-SAC (j) Walker2d-SAC (k) Humanoid-SAC (l) HalfCheetah-SAC

Figure 3: Examining the performance of vMFER on the Mujoco environment. The baseline curves represents pure TD3 or SAC, while vMFER (uncertainty) and vMFER (rank) represent two distinct forms of vMFER utilized in policy improvement combined with baseline.

## 4 Experimental Results

We conducted a series of experiments to evaluate the effectiveness of our vMFER algorithm when combined with off-policy algorithms like TD3 and SAC. We aim to compare the performance between different RL methods that incorporate uncertainty-based and rank-based probability updating strategies in a variety of environments. Furthermore, we have integrated vMFER with Prioritized Experience Replay (PER) [Schaul *et al.*, 2015], based on SAC, to showcase the flexibility and compatibility of our method. These experiments are mainly centered around the Mujoco robotic control environment [Brockman *et al.*, 2016].

Additionally, to enhance learning efficiency in sparse reward scenarios, we have merged vMFER with Hindsight Experience Replay [Andrychowicz *et al.*, 2017], achieving notable results in robotic arm control tasks with sparse rewards [Plappert *et al.*, 2018]. We also conducted ablation studies on the impact of the update-to-data (UTD) ratio [Chen *et al.*, 2021] on our algorithm, and found that utilizing vMFER could further improve the performance of the algorithm with different UTD ratio values, demonstrating the compatibility between vMFER and UTD ratio.

**Implementation Details.** Apart from hyperparameters associated with baseline algorithms, like ensemble numbers, vMFER requires no fine-tuning of hyperparameters. This facilitates its seamless and efficient integration with any Actor-Critic algorithm to enhance performance. We follow the hyperparameter configurations specified in the respective papers of the baseline algorithms. Besides, the reported results are based on 5 trials, with curves representing means and shaded areas denoting variances.

**Performance Improvement.** Results presented in Figure 3 demonstrate significant performance improvement for both TD3 and SAC algorithms compared to their respective baseline algorithms. Furthermore, in Figure 4, we investigate the influence of vMFER on the combination of PER, based on SAC. Table 1 presents the average performance improvement achieved by integrating vMFER with various algorithms (SAC, TD3, PER) in Mujoco tasks, compared to their baseline counterparts.

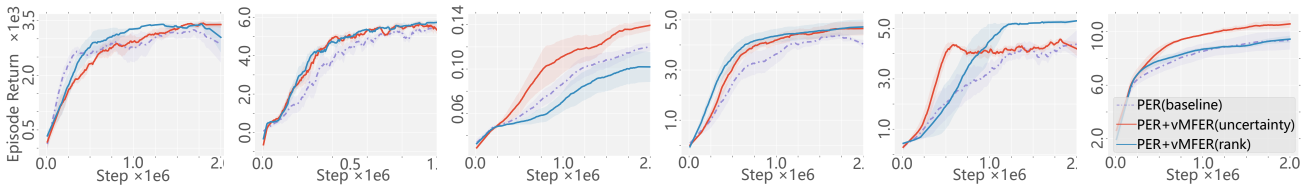
| | SAC | TD3 | SAC+PER |
|---|---|---|---|
| baseline | 100% | 100% | 100% |
| vMFER (rank) | 106.84% | 111.62% | 102.09% |
| vMFER (uncertainty ) | **113.78%** | **117.75%** | **107.17%** |

Table 1: Average performance improvement of vMFER over baseline, calculated by aggregating performance gains across all tasks.

Irrespective of whether the resampling probability of transition was updated directly through uncertainty or rank, the overall performance is notably superior to the baseline. In various environments, vMFER exhibits distinct performance enhancements with rank and uncertainty. In summary, an average performance improvement of over 10% compared to the baseline was accomplished. These findings highlight the importance of avoiding the blind use of transitions during the policy improvement process, which may reduce efficiency. Our method of reassigning the confidence of transitions by the uncertainty of gradient directions during the policy improvement process is more efficient.

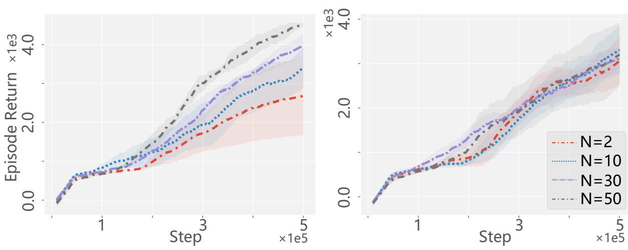
**Extended Analyses.** Our integration of vMFER with PER, known for its transition sampling probability redistribution based on TD error, still yields significant performance improvements. This suggests that both PER and vMFER methods independently exert influence on the SAC algorithm. The superior performance of the combined PER and vMFER algorithm compared to PER alone implies that their effects on the algorithm are somewhat orthogonal. Indeed, while PER primarily optimizes the policy evaluation process, vMFER enhances the policy improvement process.

Moreover, we also find that policy improvement using vMFER, which relies on the uncertainty in transition sampling




(a) Hopper (b) Ant (c) Swimmer (d) Walker2d (e) Humanoid (f) HalfCheetah

Figure 4: An experiment conducted in the Mujoco environment to explore the effect of various forms of VMFER on policy improvement. The experiment builds upon the PER combined with SAC.




(a) Rank (b) Uncertainty

Figure 5: The impact of ensemble number on vMFER.

probability, is more stable and effective than using rank-based methods, challenging the robustness associated with rank in PER. This divergence can be attributed to the finite range of uncertainty in our approach. The vMFER algorithm calculates the sampling factor $p_j$ as $p_j = \exp(\mathbf{R}\mu^{\mathrm{T}}x) = \exp(\mathbf{R}\cos\xi)$, where $\xi$ is the angle between the selected gradient of the RL algorithm and the mean direction $\mu$ of the distribution. Here, the uncertainty is quantified on a smaller scale by substituting $k$ with $\mathbf{R}$.

**Ablation on Ensemble Number.** In addition, we also investigate how the use of different numbers of ensemble critics for calculating gradient uncertainty in vMFER influences its performance, as depicted in Figure 5. It is important to emphasize that these additional critics are solely employed to enhance the calculation of gradient uncertainty and do not affect policy evaluation. Our findings reveal that modifications in ensemble size significantly affect the performance of vMFER when the resampling probability is determined by rank, as illustrated in Figure 5(a). Higher ensemble size results in improved performance. However, varying the ensemble size has minimal effect on vMFER when the resampling probability is determined by uncertainty, observed in Figure 5 (b).

## 5 Related Work

**Ensemble Structure in RL.** Ensemble structures enhance RL algorithm performance [Buckman *et al.*, 2018; Lee *et al.*, 2021; Shen and How, 2021; Song *et al.*, 2023; Lee *et al.*, 2022], addressing overestimation in stochastic MDPs, as seen in Q-learning [Watkins and Dayan, 1992]. Double Q-learning [Hasselt, 2010] initially used dual critics to counteract overestimation. Averaged Q-estimates by Anschel et al. [Anschel *et al.*, 2017] reduced Q-learning variance, while Lan et al. [Lan *et al.*, 2020] and Ciosek et al. [Ciosek *et al.*, 2019] utilized ensembles for exploration and conservative updates. In offline RL, ensemble critics are used for training more stable, or even conservative critics [Agarwal *et al.*, 2020; An *et al.*, 2021; Zhao *et al.*, 2023].

**Uncertainty Measure in RL.** Uncertainty estimation is widely used in RL [Chen *et al.*, 2017; Lockwood and Si, 2022; Kalweit and Boedecker, 2017; Zhang *et al.*, 2020; Clements *et al.*, 2019] for exploration [Audibert *et al.*, 2009; Yang *et al.*, 2021; Liu *et al.*, 2024], Q-learning [Dearden *et al.*, 1998; Wang and Zou, 2021], and planning [Wu *et al.*, 2022]. Bootstrapped DQN [Osband *et al.*, 2016] uses an ensemble of Q-functions for uncertainty quantification in Q-values, enhancing exploration. Osband et al. [Osband *et al.*, 2018] propose a Q-ensemble with Bayesian prior functions. Abbas et al. [Abbas *et al.*, 2020] introduce uncertainty-incorporated planning with imperfect models. In offline RL, MOPO [Yu *et al.*, 2020] and MOReL [Kidambi *et al.*, 2020] employ model prediction uncertainty measures to address uncertainty-penalized policy optimization.

## 6 Conclusion

We have advanced the policy improvement process by incorporating the consideration of gradient direction disagreements under an ensemble structure. Distinct from prior methodologies, our approach utilizes von Mises-Fisher distributions to model gradient directions and quantify the uncertainty of these directions under current critics for each transition during policy improvement. Building on this, we introduce the vMFER algorithm, which assigns confidence levels to all transitions in the replay buffer and resamples them based on their probability, determined by the uncertainty of gradient directions. In this way, the transition with high confidence can be used to update actors more frequently, thereby enhancing the efficiency of the policy improvement process.

The impact of gradient uncertainty on the policy improvement process, considered in this paper, is an aspect that has been scarcely addressed in existing research. This insight prompts future researchers to be aware of the potential effects of ensemble gradients. Future studies could delve deeper into the uncertainty of gradients, extending from solely directional uncertainty to the joint uncertainty of both direction and magnitude. Furthermore, exploring the quantification of gradient uncertainty, its impact in offline RL, and its advantages in practical implementations holds substantial value.

## References


[Abbas *et al.*, 2020] Zaheer Abbas, Samuel Sokota, Erin Talvitie, and Martha White. Selective dyna-style planning under limited model capacity. In *International Conference on Machine Learning*, pages 1–10, 2020.

[Abeyruwan *et al.*, 2023] Saminda Wishwajith Abeyruwan, Laura Graesser, David B D'Ambrosio, and et al. i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops. In *Conference on Robot Learning*, pages 212–224. PMLR, 2023.

[Agarwal *et al.*, 2020] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114, 2020.

[An *et al.*, 2021] Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in Neural Information Processing Systems*, 34:7436–7447, 2021.

[Andrychowicz *et al.*, 2017] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, and et al. Hindsight experience replay. *Advances in Neural Information Processing Systems*, 30, 2017.

[Anschel *et al.*, 2017] Oron Anschel, Nir Baram, and Nahum Shimkin. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International Conference on Machine Learning*, pages 176–185, 2017.

[Audibert *et al.*, 2009] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[Banerjee *et al.*, 2005] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

[Bingham, 1974] Christopher Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974.

[Bowman, 2012] Frank Bowman. *Introduction to Bessel functions*. Courier Corporation, 2012.

[Brockman *et al.*, 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[Buckman *et al.*, 2018] Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[Chen *et al.*, 2021] Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021.

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

[Clements *et al.*, 2019] William R Clements, Bastien Van Delft, Benoît-Marie Robaglia, and et al. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*, 2019.

[Dearden *et al.*, 1998] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian q-learning. *AAAI*, 1998:761–768, 1998.

[Fisher, 1953] Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

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

[Haarnoja *et al.*, 2018a] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

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

[Hasselt, 2010] Hado Hasselt. Double q-learning. *Advances in Neural Information Processing Systems*, 23, 2010.

[Kalweit and Boedecker, 2017] Gabriel Kalweit and Joschka Boedecker. Uncertainty-driven imagination for continuous deep reinforcement learning. In *Conference on Robot Learning*, pages 195–206, 2017.

[Kent, 1982] John T Kent. The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982.

[Kidambi *et al.*, 2020] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:21810–21823, 2020.

[Lan *et al.*, 2020] Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. Maxmin q-learning: Controlling the estimation bias of q-learning. *arXiv preprint arXiv:2002.06487*, 2020.

[Lee *et al.*, 2021] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*, pages 6131–6141. PMLR, 2021.

[Lee *et al.*, 2022] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, pages 1702–1712. PMLR, 2022.

[Lillicrap *et al.*, 2015] 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. *arXiv preprint arXiv:1509.02971*, 2015.

[Liu *et al.*, 2024] Jinyi Liu, Zhi Wang, Yan Zheng, Jianye Hao, Chenjia Bai, Junjie Ye, Zhen Wang, Haiyin Piao, and Yang Sun. Ovd-explorer: Optimism should not be the sole pursuit of exploration in noisy environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13954–13962, Mar. 2024.

[Lockwood and Si, 2022] Owen Lockwood and Mei Si. A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 155–162, 2022.

[Mardia *et al.*, 2000] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional Statistics*, volume 2. Wiley Online Library, 2000.

[Osband *et al.*, 2016] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in Neural Information Processing Systems*, 29, 2016.

[Osband *et al.*, 2018] Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[Plappert *et al.*, 2018] Matthias Plappert, Marcin Andrychowicz, Alex Ray, and et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

[Radosavovic *et al.*, 2023] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pretraining. In *Conference on Robot Learning*, pages 416–426, 2023.

[Schaul *et al.*, 2015] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Shen and How, 2021] Macheng Shen and Jonathan P How. Robust opponent modeling via adversarial ensemble reinforcement learning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 31, pages 578–587, 2021.

[Song *et al.*, 2023] Yanjie Song, PN Suganthan, Witold Pedrycz, Junwei Ou, Yongming He, and Yingwu Chen. Ensemble reinforcement learning: A survey. *arXiv preprint arXiv:2303.02618*, 2023.

[Sra, 2012] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $I_s$ (x). *Computational Statistics*, 27:177–190, 2012.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[Thrun and Schwartz, 1993] Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*, volume 255, page 263. Hillsdale, NJ, 1993.

[Wang and Zou, 2021] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

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

[Watson, 1982] Geoffrey S Watson. Distributions on the circle and sphere. *Journal of Applied Probability*, 19(A):265–280, 1982.

[Wu *et al.*, 2022] Zifan Wu, Chao Yu, Chen Chen, Jianye Hao, and Hankz Hankui Zhuo. Plan to predict: Learning an uncertainty-foreseeing model for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:15849–15861, 2022.

[Wu *et al.*, 2023] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, pages 2226–2240, 2023.

[Yang *et al.*, 2021] Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: a comprehensive survey. *arXiv preprint arXiv:2109.06668*, 2021.

[Yu *et al.*, 2020] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

[Zhang *et al.*, 2020] Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 33:10571–10583, 2020.

[Zhao *et al.*, 2023] Kai Zhao, Yi Ma, Jinyi Liu, Yan Zheng, and Zhaopeng Meng. Ensemble-based offline-to-online reinforcement learning: From pessimistic learning to optimistic exploration. *CoRR*, abs/2306.06871, 2023.