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

Yiwen Zhu\textsuperscript{1,2,3}, Jinyi Liu\textsuperscript{4}, Wenyia Wei\textsuperscript{1}, Qianyi Fu\textsuperscript{1}, Yujing Hu\textsuperscript{2*}, Zhou Fang\textsuperscript{1*}, Bo An\textsuperscript{3,5}, Jianye Hao\textsuperscript{4}, Tangjie Lv\textsuperscript{2} and Changjie Fan\textsuperscript{2}

\textsuperscript{1}Zhejiang University
\textsuperscript{2}NetEase Fuxi AI Lab
\textsuperscript{3}Nanyang Technological University
\textsuperscript{4}Tianjin University
\textsuperscript{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-
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 minibatch of transitions in actor training is commonly expressed as

\[ E_{(s,a)} \sim D \left[ \log \pi(a|s) - \min_i Q_i(s,a) \right] \]

[Haarnoja et al., 2018a],

\[ E_{(s,a)} \sim D \left[ -Q_1(s,a) \right] \]

[Lillicrap et al., 2015; Fujimoto et al., 2018], or

\[ E_{(s,a)} \sim D \left[ \frac{1}{N} \sum_{i=1}^{N} \log \pi - Q_i(s,a) \right] \]

[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 \( x \sim vMF(k, \mu) \) is expressed in Eq. (1), where \( f_p \) represents the density function for vMF\((k, \mu)\).

\[ x \sim vMF(k, \mu), \]

\[ f_p(x; \mu, k) = C_p(k) \exp(k \mu^T x). \] (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].

\[ \mu = \bar{x} / R, \text{ where } \bar{x} = \frac{1}{N} \sum_{i} x_i, \ R = ||\bar{x}||_2. \] (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:

\[ \hat{k} = \frac{R(p - R^2)}{(1 - R^2)}. \] (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
In this study, we employ the uncertainty of the gradient direction in gradients during the policy improvement process. In a toy experiment, the presence of disagreements among gradients in the action-value function’s sensitivity to changes in action space. Figure 1 illustrates the direction and magnitude of the action-value function’s 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 \( s_t \) and action \( a \):

\[
Q_1(s_t, a) = -(a + \mathbf{1}_{2 \times 1} + \epsilon)^T (a + \mathbf{1}_{2 \times 1} + \epsilon),
\]

\[
Q_2(s_t, a) = -(a - \mathbf{1}_{2 \times 1} + \epsilon)^T (a - \mathbf{1}_{2 \times 1} + \epsilon).
\]

where \( a \in \mathbb{R}^{2 \times 1} \) 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: \( \max_a Q(s_t, a) \). This allows us to compute the gradient \( \frac{\partial Q(s_t, a)}{\partial a} \) 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
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 $k$ required is expressed simply. The efficacy and accuracy of this approach 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).

$$\mathbf{L}(s_t, a) = [l_1(s_t, a) \cdots l_n(s_t, a)]^T,$$

$$\partial \mathbf{L}(s_t, a) / \partial \theta = [\partial l_1(s_t, a) / \partial \theta \cdots \partial l_n(s_t, a) / \partial \theta]^T, \ a \sim \pi(\cdot | s_t). \quad (5)$$

where $n$ is the ensemble size of the critic networks. Estimating the parameter of the Von Mises-Fisher distribution for $\partial l_i/\partial \theta$ 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 $\partial / \partial a$ is constant for the same transition. As a result, the uncertainty in the gradient directions primarily arises from $\partial / \partial \theta$, $i \in [1, n]$. Therefore, calculating the uncertainty of $\partial l_i / \partial a$ rather than $\partial l_i / \partial \theta$ is a more cost-effective and scalable approach.

$$\partial l_i(s_t, a) / \partial \theta = \partial l_i(s_t, a) / \partial a \cdot \partial a / \partial \theta, \ \ a \sim \pi(\cdot | s_t). \quad (6)$$

Let $x_i(s_t) = ||\partial l_i(s_t, a)||_2^{-1} \cdot \partial l_i(s_t, a) / \partial \theta \sim \pi(\cdot | s_t)$ and $x(s_t) = \sum_i x_i(s_t)/n$, where $x_i(s_t)$ denote the direction of the gra-
Algorithm 1 Von Mises-Fisher Experience Resampling (Based on TD3)

1: Initialize replay buffer $D$ and ensemble number $N$ 
2: Initialize critic networks $\{Q_{\phi_i}, i \in [1, N]\}$, and actor network $\pi_{\theta}$ with random parameters $\{\phi_i, i \in [1, N]\}$, $\theta' \leftarrow \theta$
3: Initialize target network $\{\hat{\phi}_i \leftarrow \phi_i, i \in [1, N]\}$
4: Initialize sampling factors $p_j = 1$ for each transition of $D$
5: for $t=1$ to $T$ do
6: sample action with noise $a \sim \pi_{\theta}(s) + \epsilon, \epsilon \sim N(0, \sigma)$
7: store transition $(s, a, r, s')$ in $D$
8: sample mini-batch of $b$ transitions $(s, a, r, s')$ from $D$
9: $a' \leftarrow \pi_{\theta}(s') + \epsilon \sim clip(N(0, \sigma), -c, c)$
10: $y \leftarrow r + \gamma \min_{n=1,2} Q_{\phi_i}(s', a')$
11: Update critics $\hat{\phi}_i \leftarrow \arg\min_{\phi_i} b^{-1} \sum(y - Q_{\phi_i}(s, a))^2$
12: if $t$ mod 2 then
13: for $j = 1$ to $b$ do
14: Resample transition $(s_j, a_j, r_j, s'_j) \sim P(j) = \frac{p_j}{\sum_m p_m}$ \text{▷ Eq. (8)}
15: Sample action $a_{j} \sim \pi_{\theta}(s_j) + \epsilon, \epsilon \sim clip(N(0, \sigma), -c, c)$
16: Calculate the actor losses $l_i(s_j, a_{j}) \leftarrow -Q_{\phi_i}(s_j, a_{j}), i \in [1, N]$
17: Calculate gradients of losses $g_i = \frac{\partial l_i(s_j, a_{j})}{\partial \theta}(a_{j} = a_{j_{true}}), i \in [1, N]$
18: Update sampling Factor $p_j \leftarrow \text{Update Sampling Factor } p_j$ (Algorithm 2)
23: Update $\theta$ by the deterministic policy gradient:
24: \( \nabla_{\theta} J(\theta) = -b^{-1} \sum_j \nabla_{\theta} l_i(s_j) \)
25: Update target networks:
26: $\phi_i' \leftarrow \tau \phi_i + (1 - \tau) \phi_i', \theta' \leftarrow \tau \theta + (1 - \tau) \theta'$

Algorithm 2 Update Sampling Factors $p_j$

Input: $p_j, s_j, \{g_{1j}, g_{2j}, \cdots, g_{Nj}\}, \{Q_{\phi_1}, Q_{\phi_2}, \cdots, Q_{\phi_N}\}, \pi_{\theta}$
1: Calculate normalized unit vector $x_i = \frac{x_i}{|x_i|}, i \in [1, N]$
2: Calculate other parameters:
3: $R(s_j) = \frac{2}{\sum_m p_m} \cdots$ \text{▷ Eq. (7)}
4: Choose index of critics used in the policy improvement:
5: $e = \arg\min_{i} Q_{\phi_i}(s_j, \pi_{\theta}(s_j))$ \text{SAC (TD3)}
6: $p_j \leftarrow \text{exp(} \frac{\|\text{rank} \cdot \text{exp(} (R(s_j)\mu(s_j)x(s_j)) \)}{\text{exp(} \text{rank} \cdot \text{exp(} (R(s_j)\mu(s_j)x(s_j)) \))^{-1}}$ \text{Eq. (9, 10)}

Output: $p_j$

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 $P(j|D) = \frac{1}{M}$, where $D$ denotes the transitions in the replay buffer and $M$ the total number of data points in it. Subsequently, we represent the conditional probability distribution as $P(x(j)|S, \{Q_{\phi}, \pi_{\theta}\})$.

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 $P(j|D) = \frac{1}{M}$, where $D$ denotes the transitions in the replay buffer and $M$ the total number of data points in it. Subsequently, we represent the conditional probability distribution.
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.

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

<table>
<thead>
<tr>
<th></th>
<th>SAC</th>
<th>TD3</th>
<th>SAC+PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>vMFER (rank)</td>
<td>106.84%</td>
<td>111.62%</td>
<td>102.09%</td>
</tr>
<tr>
<td>vMFER (uncertainty)</td>
<td>113.78%</td>
<td>117.75%</td>
<td>107.17%</td>
</tr>
</tbody>
</table>

Table 1: Average performance improvement of vMFER over baseline, calculated by aggregating performance gains across all tasks.
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 \( p_j = \exp(\mathbf{R}_\mu^T \mathbf{x}) = \exp(\mathbf{R} \cos \xi) \), 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 \( R \).

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


