Improving Generalization in Offline Reinforcement Learning via Latent Distribution Representation Learning

Da Wang[1], Lin Li[1], Wei Wei[1*], Qixian Yu[1], Jianye Hao[2], Jiye Liang[1]

[1]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
{sxu_wd, yy17835438310}@163.com, {lilynn1116, weiwei, ljy}@sxu.edu.cn, jianye.hao@tju.edu.cn

Abstract

Dealing with the distribution shift is a significant challenge when building offline reinforcement learning (RL) models that can generalize from a static dataset to out-of-distribution (OOD) scenarios. Previous approaches have employed pessimism or conservatism strategies. More recently, data-driven work has taken a distributional perspective, treating offline data as a domain adaptation problem. However, these methods use heuristic techniques to simulate distribution shifts, resulting in a limited diversity of artificially created distribution gaps. In this paper, we propose a novel perspective: offline datasets inherently contain multiple latent distributions, with behavior data from diverse policies potentially following different distributions and data from the same policy across various time phases also exhibiting distribution variance. We introduce the Latent Distribution Representation Learning (LAD) framework, which aims to characterize the multiple latent distributions within offline data and reduce the distribution gaps between any pair of them. LAD consists of a min-max adversarial process: it first identifies the "worst-case" distributions to enlarge the diversity of distribution gaps and then reduces these gaps to learn invariant representations for generalization. We derive a generalization error bound to support LAD theoretically and verify its effectiveness through extensive experiments.

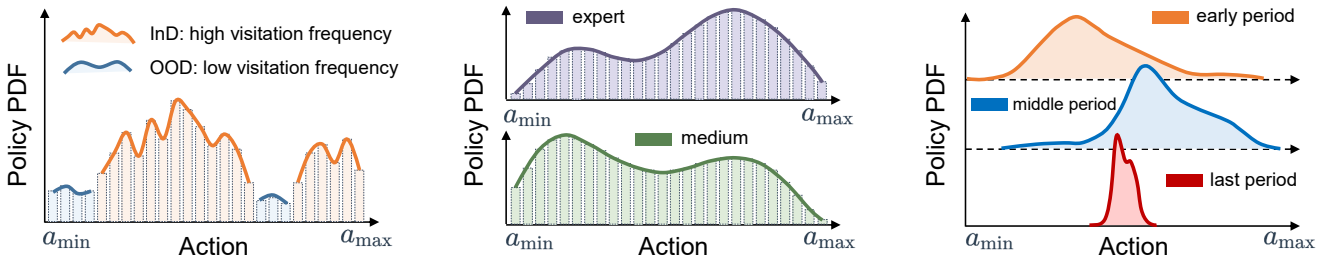
1 Introduction

Offline reinforcement learning (RL) (Levine et al. 2020) enables the execution of safe and efficient learning utilizing pre-collected static datasets without the need for further interaction with the environment. Compared with its online counterpart (Sutton and Barto 2018), offline RL offers an attractive potential for saving resources and mitigating risks. The distribution shift represents one of the core challenges in offline RL, with out-of-distribution (OOD) (Fujimoto, Meger, and Precup 2019) data being not only grossly overestimated but also exacerbated through bootstrapping (Kumar et al. 2019), which can precipitate a catastrophic decline in performance. Rich experience has been accumulated in addressing the distribution shift issue, with strategies containing the implementation of policy constraints (Fakoor et al. 2021; Fujimoto and Gu 2021; Wu et al. 2022; Ran et al. 2023; Li et al. 2023), the adoption of conservative Q values (Kumar et al. 2020; Kostrikov et al. 2021; Ma, Jayaraman, and Bastani 2021; Wang, Hunt, and Zhou 2022; Lyu et al. 2022), and the utilization of uncertainty estimation techniques (An et al. 2021; Bai et al. 2022; Wu et al. 2021).

Unlike the aforementioned methods that employ pessimism or conservatism, the alternative research (Qi et al. 2022; Wang et al. 2024) introduces a new way of thinking, conceptualizing offline data-driven decision-making as domain adaptation (Ben-David et al. 2010; Zhao et al. 2019). The primary goal is to ensure accurate predictions for the value of optimized decisions – the "target domain" – when training only on the provided dataset – the "source domain". IOM (Qi et al. 2022) addresses distribution shift by enforcing invariance between the learned representations of the training dataset and optimized decisions. Coincidentally, the recent ADS (Wang et al. 2024) framework employs dataset splitting to create simulations of distribution shifts, an endeavor grounded in the concept of domain adaptation. This framework is modeled as a min-max optimization challenge, which takes inspiration from meta-learning, and it enforces the model to achieve generalization over the distribution shifts simulated from the train/validation subsets splitting of the dataset. This novel approach, added to the line of offline RL algorithms, brings a fresh perspective to the field. However, the existing techniques still rely on heuristic designs for simulating distribution shifts. The diversity inherent in the distribution gaps they produce and the potential to extract richer information from the original offline datasets are issues deserving of deeper investigation.

With the goal of building models that can generalize across unknown distributions, this paper steers a different course from prior studies by introducing an innovative perspective: offline datasets inherently contain multiple latent distributions, with behavior data from diverse policies potentially following different distributions and data from the same policy across various time phases also exhibiting distribution variance (Figure 1). The former scenario emerges in particular cases, such as situations where the offline dataset arises from sampling various policies. On the other hand, the latter corresponds to the dynamically changing scenarios within time series, where such dynamic distributions exhibit diversity (Lu et al. 2023). We start by examining the connection between latent distributions and OOD

*Corresponding Author.



(a) The distribution shift of state-action visitation frequency

(b) Diverse policies potentially following different distributions

(c) The same policy across various time phases exhibiting distribution variance

Figure 1: Abstract representation of our perspective: offline datasets inherently contain multiple latent distribution.

generalization. Following this examination, we introduce a framework, *Latent Distribution Representation Learning* (LAD), which subtly utilizes these latent distributions. The goal of LAD is to characterize naturally existing latent distributions in offline datasets and then learn generalizable models by reducing the distribution gaps between any pair of them. To be specific, the LAD framework works through an iterative min-max adversarial game: initially, it identifies and characterize the multiple latent distributions that have the maximum gap; subsequently, it bridges the gap between these distributions, thereby learning invariant representations that enhance the model's ability to generalize. To validate the effectiveness of our proposed LAD framework, we offer theoretical insights accompanied by an extensive array of experiments. LAD can enhance the backbone agent's performance across various continuous control tasks on the D4RL dataset, outperforming several state-of-the-art offline RL algorithms. Additionally, we provide parameter analysis and ablation studies, along with visual evidence of LAD's successful representation of latent distributions, which together robustly attest to the efficacy of LAD.

In summary, our contribution is four-fold:

- We introduce a novel data-driven perspective to address the distribution shift issue in offline RL: simulating OOD generalization scenarios with a higher diversity of distribution gaps, achieved by characterizing the multiple latent distributions inherently present in offline data.
- We propose the Latent Distribution Representation Learning (LAD) framework, which consists of an iterative min-max adversarial game. LAD first identifies latent distribution representations in scenarios with the maximum distribution gap. It then works to reduce these gaps, thus learning invariant representations crucial for generalization.
- We derive a generalization error bound for multiple distribution scenarios and provide theoretical insights to analyze the effectiveness of LAD.
- By applying LAD to commonly used algorithms, we significantly enhance their performance and competitiveness. Importantly, LAD's successful characterization of latent distributions validates our proposition that promoting the model to generalize across distribution gaps with maximum diversity is essential for offline learning.

2 Preliminaries

An infinite-horizon discounted Markov Decision Process (MDP) is denoted as (S, A, P, r, γ), where S and A are finite state and action spaces, $P(s'|s,a) : S \times A \times S \mapsto [0,1]$ is the state transition probability function, $r(s,a) : S \times A \mapsto [0, R_{\max}]$ is the reward function, and $\gamma \in (0,1)$ is the discount factor. RL aims to find an optimal policy $\pi(\cdot|s)$ that maximizes the expected cumulative discounted reward $J(\pi) := \mathbb{E}_\pi[\Sigma_{t=0}^{\infty}\gamma^t r(s_t, a_t)]$. A major approach, Q-learning, which learns a Q-function $Q^\pi(s,a) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r\left(s_t, a_t\right) \mid s_0 = s, a_0 = a\right]$ to obtain the optimal policy. For a policy π, the Bellman operator for iteratively updating the Q-function as $\mathcal{T}Q(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}[r(s,a) + \gamma \mathbb{E}_{a'\sim\pi(\cdot|s')}Q(s',a')]$.

In offline RL, where real-world online interaction is not practicable, the agent must learn a policy based on a static dataset $\mathcal{D} = \{(s,a,r,s')\}$ that was previously generated from an unknown behavior policy $\mu(\cdot|s)$. We assume that (s,a) is generated $i.i.d.$ from the data distribution μ in the training process. The learned policy, however, whose action $a' \sim \pi(\cdot|s')$ utilized during the Bellman backup might lie outside the support of μ. Such a distribution shift between π and μ leads to extrapolation errors due to the arbitrarily wrong estimation upon a'. Offline RL aims to obtain an optimal policy by leveraging solely the static dataset under the influence of distribution shift.

3 Our Method

In this section, we first examine the connection between latent distributions and OOD generalization. Then, we propose the Latent Distribution Representation Learning (LAD) framework, which subtly utilizes latent distributions and thus learns generalizable models. Finally, we theoretically derive a generalization error bound for multiple distribution scenarios and provide theoretical insight to understand our method.

3.1 Motivation

The distribution shift in offline RL refers to the disparity in the state-action visitation frequency between the learned policy and the dataset. An abstract illustration of the visitation frequency for In-Distribution (InD) data and Out-of-Distribution (OOD) data is depicted in Figure 1(a). In the

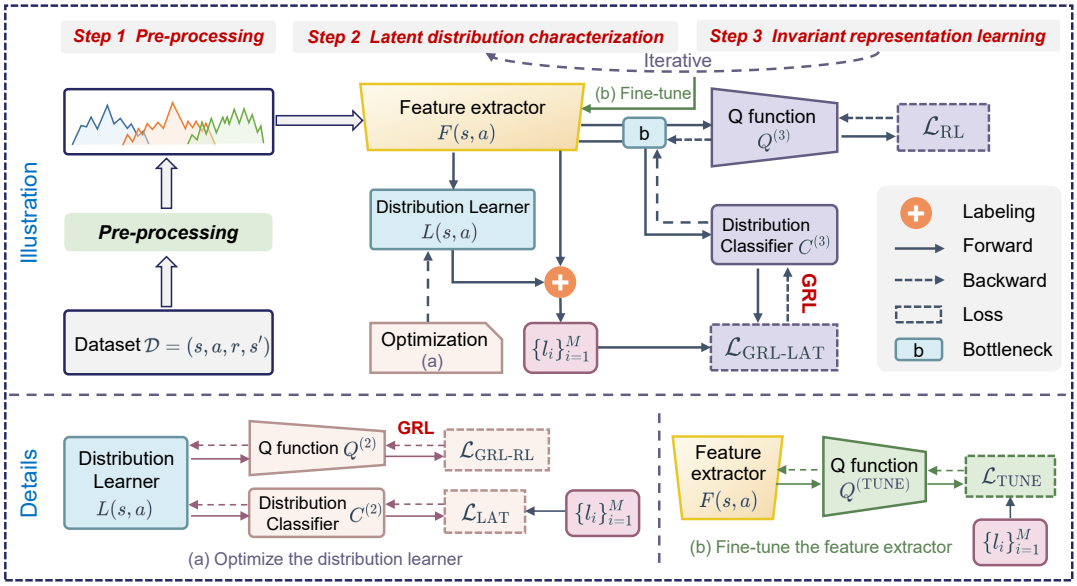


Figure 2: Illustration of the LAD framework. Details (a) corresponds to the `Optimization` in step 2 and (b) corresponds to the `Fine-tune` in the iterative process.

latest work, ADS (Wang et al. 2024), the offline datasets are segmented to create simulated distribution shifts by establishing training/validation subsets that embody distribution gaps. Carrying forward the analysis of data distributions from ADS, we concentrate on a more innovative and realistic perspective: the offline datasets inherently contain multiple latent distributions, denoted as $\mathcal{P}(x, y) = \sum \mathcal{P}_i(x, y)$ ($\mathcal{P}_i$ is the i-th latent distribution, x is the input, and $y = r + \gamma \max_{a'} Q(s', a')$ is the output of target Q-network). For instance, some datasets, such as the medium-expert in the D4RL (Fu et al. 2020), have data collected from various policies, as depicted in Figure 1(b), where each distinct policy corresponds to behavior data from a different distribution. Furthermore, the same policy can exhibit different behaviors across various time phases, thus resulting in distinct data distributions (such as the medium-replay in the D4RL), as shown in Figure 1(c). Well-studied offline datasets can generally satisfy one or both of the aforementioned scenarios and are equally applicable to unknown data distributions. This factor also highlights the challenge of building models capable of generalizing knowledge to OOD data.

Leveraging the processing of data distributions to simulate distribution shifts, ADS (Wang et al. 2024) conducts an adversarial search to identify the splitting that exhibits the most significant distribution gap, subsequently employing meta-learning techniques to develop a generalizable model. In contrast, our approach involves characterizing the multiple latent distributions present within the dataset and seeking to minimize the "worst-case" scenario: an adversarial search targeting instances where the distribution gaps between any $\mathcal{P}_i$ and $\mathcal{P}_j$ are maximized. The advantage is that it fosters a model capable of handling a more diverse array of distribution gaps, thereby yielding a more comprehensive and robust model with enhanced generalization capabilities.

3.2 The LAD Framework

In this section, we introduce the LAD framework, whose core is to characterize multiple latent distributions followed by reducing the gaps between these distributions. Figure 2 describes the principal procedures of LAD, containing a data pre-processing phase (Step 1) and an iterative process: it initially identifies the "worst-case" distribution scenarios from the provided dataset (Step 2) and subsequently works to reduce the distribution gaps between each pair of them (Step 3).

Step 1 Pre-processing Directly characterizing latent distributions from an extensive array of raw offline datasets is inefficient. A practical approach is to handle a large batch of samples in each epoch (such as M samples) (Wang et al. 2024), and then, we focus on characterizing the latent distributions for just these M samples. Motivated by earlier studies (Sun, Zhou, and Li 2020; Wang et al. 2024), in our approach to create varied data distributions, we employ a Gaussian Mixture Model (GMM) (Bishop 2006) to cluster the state-action pairs, which yields K initial latent distributions[1].

Step 2 Latent Distribution Characterization This step aims to characterize the latent distributions and learn representations that maximize the distribution gaps, thereby enlarging diversity. For this process, we define a feature extractor $F(s, a)$, a specific distribution learner $L(s, a)$, a classifier $C^{(2)}(s, a)$ for the latent distribution, and a Q-network $Q^{(2)}(s, a)$. In Step 1, we employ GMM to initiate the K latent distributions. This method is inherently a self-supervised approach to pseudo-labeling, effectively assigning an initial latent distribution l ($l \in [1, 2, ..., K]$) to

[1] K is a hyperparameter which we test in Section 5.2

each sample. With the representation serving as the output of $F(s,a)$, we first determine the centroid of the current latent distribution:

$$\tilde{\sigma}_k = \frac{\sum_{(s,a)\in\mathcal{D}} \eta_k(C^{(2)}(L(F(s,a))))L(F(s,a))}{\sum_{(s,a)\in\mathcal{D}} \eta_k(C^{(2)}(L(F(s,a))))}, \quad (1)$$

where $\tilde{\sigma}_k$ is the initial centroid of the k^{th} latent distribution and η_k is the k^{th} element of the logit soft-max output. Based on the current centroid $\tilde{\sigma}_k$, we use a distance function d to label the latent distribution through the nearest centroid classifier:

$$\tilde{l}(s,a) = \arg\min_k d(L(F(s,a), \tilde{\sigma}_k). \quad (2)$$

We then compute the accurate centroids:

$$\sigma_k = \frac{\sum_{(s,a)\in\mathcal{D}} \mathbb{I}(\tilde{l}(s,a) = k)L(F(s,a))}{\sum_{(s,a)\in\mathcal{D}} \mathbb{I}(\tilde{l}(s,a) = k)}, \quad (3)$$

where $\mathbb{I}$ is the indicator function that returns 1 if and only if its argument is valid. The difference between Equation (3) and Equation (1) lies in the fact that the centroid $\tilde{\sigma}_k$, as computed by the latter, is subject to the influence of η_k, whereas Equation (3) provides a more accurate determination of the centroid (reflecting the divergence between the soft-max function and the output [0,1]). As a result, we can attain the updated latent distribution labels:

$$l(s,a) = \arg\min_k d(L(F(s,a), \sigma_k). \quad (4)$$

Once we attach the updated latent distributions $\{l_i\}_{i=1}^M$ for the M-selected samples, our subsequent task is to learn the representation that exhibits the maximal gaps within these distributions. Drawing inspiration from Ganin et al. (2016), we adopt an adversarial training approach, employing the following loss function:

$$\begin{aligned}\mathcal{L}_{\text{LAT}} + \mathcal{L}_{\text{GRL-RL}} = \mathbb{E}_{(s,a)\in\mathcal{D}}\ell\left(C^{(2)}(L(F(s,a))), l\right) \\ + \ell\left(Q^{(2)}(\texttt{GRL}(L(F(s,a)))), y\right).\end{aligned} \quad (5)$$

where ℓ is the cross-entropy loss and GRL is the gradient reverse layer (Ganin et al. 2016) facilitating adversarial training through reversing gradients. We use the loss $\mathcal{L}_{\text{LAT}} + \mathcal{L}_{\text{GRL-RL}}$ to optimize the network $Q^{(2)}$, $C^{(2)}$, and L, except for the feature extractor F. Training the specific distribution learner L results in a representation that, while it achieves separability of the latent distribution, fails to correct prediction − this constitutes the worst-case scenario. We offer visualization results (Section 5.4) of features with distribution separation extracted by the distribution learner. The parameters of the feature extractor F, which remain constant, are sustained for copy in forthcoming processes, while the distribution learner L enters a fresh iteration.

Step 3 Invariant Representation Learning In the final phase, we use invariant representation learning to develop generalizable models, ensuring effective learning even in the worst-case scenarios. We define a bottleneck $B^{(3)}(s,a)$, a new latent distribution classifier $C^{(3)}(s,a)$, and a Q-network $Q^{(3)}(s,a)$. We still use adversarial training:

$$\begin{aligned}\mathcal{L}_{\text{RL}} + \mathcal{L}_{\text{GRL-LAT}} = \mathbb{E}_{(s,a)\in\mathcal{D}}\ell\left(Q^{(3)}B((F(s,a))), y\right) \\ + \ell\left(C^{(3)}(\texttt{GRL}(B(F(s,a)))), l\right).\end{aligned} \quad (6)$$

In contradistinction to Equation (5), we utilize the gradient reverse layer to update the latent distribution classifier loss $\mathcal{L}_{\text{GRL-LAT}}$. This approach equips the learned model with the capability to address distributional gaps, thereby enhancing its generalizability. We set Q-network $Q^{(3)}$ as the principal critic within our RL algorithm, serving the purpose of actor training.

Fine-tune the feature extractor. We emphasize that the feature extractor F undergoes updates exclusively during the fine-tuning phase. The purpose of this step is to extract the finer-grained knowledge information embedded in the latent distribution and perform feature updating to obtain a fine-grained representation. This implies that the samples from the current latent distribution obtain a new representation from the updated feature extractor. We introduce a Q-network $Q^{(\text{TUNE})}$, a replica of $Q^{(3)}$, to participate in the feature fine-tuning process. Concurrently, we refine the target Q-function following the current latent distribution: $y' = y + \delta \cdot l$. Here, $l \in [1, 2, ..., K]$, signifies the pseudo-label assigned to the latent distribution, a result derived from Step 2. The corresponding fine-tuning loss is:

$$\mathcal{L}_{\text{TUNE}} = \mathbb{E}_{(s,a)\in\mathcal{D}}\ell\left(Q^{(\text{TUNE})}(F(s,a)), y'\right). \quad (7)$$

To ensure the stability of the training process, it is also essential that the parameters of $Q^{(2)}$ in Step 2 are copied from the updated $Q^{(3)}$ in each subsequent iteration. More details on the iteration and training of the LAD framework can be found in Appendix A.

3.3 Theoretical Analysis

To facilitate a more profound grasp of our method, we derive a generalization error bound applicable to scenarios involving multiple distributions. Drawing on the principles of domain adaptation, we conceptualize the generalization challenge in offline RL as the crucial task of minimizing the model's generalization error when trained on a source domain $\mathcal{P}$ and then evaluated on an unseen target domain $\mathcal{Q}$. A specific premise is that the source domain $\mathcal{P}$ is a collection of multiple distributions $\{\mathcal{P}_i\}_{i=1}^K$. Our theory aims to elucidate the relationship between the generalization error $\epsilon_{\mathcal{Q}}$ and the mitigation of the distribution gaps between any two $\mathcal{P}_i$ and $\mathcal{P}_j$, facilitating an assessment of the efficacy of our LAD framework.

Theorem 3.1. *Let $\mathcal{H}$ be a family of functions mapping from X to $[0,1]$, $\mathcal{Q}$ and the collection $\{\mathcal{P}_i\}_{i=1}^K$ be distributions over X. Given a set of distributions $\mathcal{O}$, for $\forall\mathcal{M} \in \mathcal{O}$ and $\forall h \in \mathcal{H}$, we have,*

$$\begin{aligned}\epsilon_{\mathcal{Q}}(h) \le \sum_i \epsilon_{\mathcal{P}_i}(h) + \frac{1}{2}\min_{\mathcal{M}\in\mathcal{O}} d_{\mathcal{H}}(\mathcal{M}, \mathcal{Q}) \\ + \frac{1}{2}\max_{i,j} d_{\mathcal{H}}(\mathcal{P}_i, \mathcal{P}_j) + \lambda^\star,\end{aligned} \quad (8)$$

DATASET	TD3BC	TD3BC +ADS	**TD3BC +LAD**	MCQ	MCQ +ADS	**MCQ +LAD**
HALFCHEETAH-M	48.2 ± 0.5	49.0 ± 2.7	**50.2 ± 1.3**	62.5 ± 3.1	63.2 ± 2.7	**66.4 ± 3.4**
HOPPER-M	60.8 ± 3.4	**73.7 ± 13.0**	69.8 ± 3.7	78.4 ± 4.3	103.0 ± 1.5	**103.6 ± 2.5**
WALKER2D-M	84.4 ± 2.1	85.0 ± 1.1	**87.1 ± 1.8**	91.0 ± 1.1	94.5 ± 2.9	**94.9 ± 2.1**
HALFCHEETAH-MR	45.0 ± 0.5	46.1 ± 2.9	**49.0 ± 1.5**	56.2 ± 2.7	59.4 ± 3.1	**61.0 ± 1.7**
HOPPER-MR	67.3 ± 13.2	**100.3 ± 2.2**	89.2 ± 11.3	101.6 ± 1.0	105.0 ± 0.9	**105.8 ± 1.8**
WALKER2D-MR	83.4 ± 7.0	91.3 ± 0.9	**93.6 ± 1.8**	91.3 ± 1.8	96.1 ± 0.6	**98.1 ± 2.3**
HALFCHEETAH-ME	90.7 ± 2.7	96.6 ± 3.1	**97.7 ± 2.4**	80.1 ± 3.8	78.9 ± 0.5	**101.5 ± 1.2**
HOPPER-ME	91.4 ± 11.3	**114.0 ± 1.9**	112.0 ± 4.3	87.8 ± 2.0	105.8 ± 0.2	**112.7 ± 2.0**
WALKER2D-ME	110.2 ± 0.3	114.0 ± 1.1	**114.5 ± 1.5**	114.2 ± 0.9	108.3 ± 1.0	**116.6 ± 1.3**

Table 1: Results of different algorithms and the ones equipped with ADS, LAD. We **bold** the highest score.

where $\epsilon_{\mathcal{Q}}(h) = \mathbb{E}_{x\sim\mathcal{Q}}|h(x) - h^*(x)|$ and h^* is an ideal function, $d_{\mathcal{H}}(\mathcal{P}, \mathcal{Q})$ is the $\mathcal{H}$-divergence[2], which measures differences in distribution, $\lambda^\star$[3] is the error of an ideal joint hypothesis for $\mathcal{P}$, $\mathcal{Q}$.

We provide the proof of Theorem 3.1 in Appendix B and proceed to elucidate how our LAD framework implicitly minimizes the rest of the terms in Equation (8), with the exception of the constant $\lambda^\star$. The first term $\sum_i \epsilon_{\mathcal{P}_i}(h)$ corresponds to training samples across all distributions within the source domain $\mathcal{P}$, the typical RL learning process that can be minimized with the loss $\mathcal{L}_{\text{RL}}$ in Equation (6). For the second item, $\frac{1}{2}\min_{\mathcal{M}\in\mathcal{O}} d_{\mathcal{H}}(\mathcal{M}, \mathcal{Q})$, the inaccessibility of the target domain $\mathcal{Q}$ prompts us to incorporate an intermediary variant $\mathcal{M} \in \mathcal{O}$. To minimize $d_{\mathcal{H}}(\mathcal{M}, \mathcal{Q})$, therefore, directly associated with enlarging the range of $\mathcal{O}$. Given the existence of a relationship such that $d_{\mathcal{H}}(\sum_i \mathcal{P}_i, \mathcal{M}) \leq \max_{i,j} d_{\mathcal{H}}(\mathcal{P}_i, \mathcal{P}_j)$, it is reasonable that we prioritize the maximization of $d_{\mathcal{H}}(\mathcal{P}_i, \mathcal{P}_j)$ under this circumstance – in essence, seeking the "worst-case" scenario among the multiple latent distributions during Step 2. The third item $\frac{1}{2}\max_{i,j} d_{\mathcal{H}}(\mathcal{P}_i, \mathcal{P}_j)$ measures the maximum gaps among source domains, which corresponds to Step 3 in LAD. The final term $\lambda^\star$ is frequently overlooked, as it is typically negligible in practice. To summarize our discussion, the LAD framework is related to minimizing the upper bound in Equation (8).

4 Related Work

With an increasing focus on the distribution shift challenges within offline RL in the community, diverse solutions have come to the forefront. In certain traditional conservative studies, learned policies are constrained by the dataset's support to limit the production of OOD actions, with techniques including explicit policy constraints (Fakoor et al. 2021; Fujimoto and Gu 2021; Wu et al. 2022; Li et al. 2023), the penalization of value functions (Kumar et al. 2020; Ma, Jayaraman, and Bastani 2021; Wang, Hunt, and Zhou 2022), the application of uncertainty quantification (An et al. 2021; Bai et al. 2022; Wu et al. 2021), and the use of imitation learning as part of the approach (Chen et al. 2020; Kostrikov et al. 2021).

Expanding on the foundation applied by prior investigations, subsequent works have focused on refining the approach to mitigate overly conservatism, thus promoting improved generalization capabilities. MCQ (Lyu et al. 2022) actively training OOD actions by constructing them pseudo target values. STR (Mao et al. 2023) performs trust region policy optimization within the support of the behavior policy, relaxing the implicit density constraint of Exponentiated Advantage-Weighted Behavior Cloning (EAWBC) methods to a support constraint. Some data-driven work such as PRDC (Ran et al. 2023) find that regularizing the policy towards the nearest state-action pair can be more effective and allows the learned policy to choose actions that do not appear in the dataset along with the given state. Other work (Qi et al. 2022; Wang et al. 2024) innovatively models the OOD generalization problem for offline RL as domain adaptation from a distribution perspective. More recently, the ADS (Wang et al. 2024) framework simulates distribution shift by splitting the dataset into train/validation subsets, adversarially searching for the divisions with the largest distribution gaps, and then using meta-learning approach to learn a generalizable model. Inspired by ADS, we start from a novel perspective: characterizing multiple latent distributions that naturally exist in offline datasets and then training generalizable model by bridging the distribution gap of the "worst-case" distribution scenario. In contrast, our considerations are more macroscopic and comprehensive, and our method is able to characterize a greater diversity of distribution shifts, leading to better generalization.

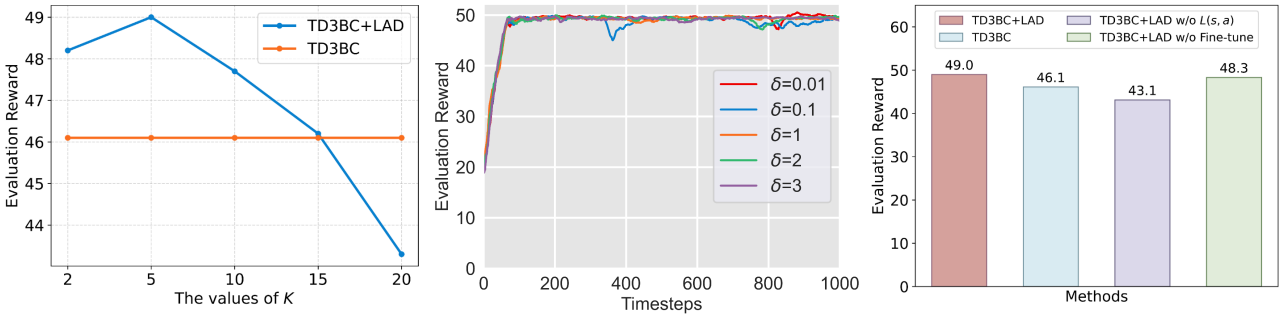
For representation learning in offline RL, prior work (Yang and Nachum 2021; Xiao et al. 2022; Geng et al. 2022; Ma et al. 2023) mainly focuses on the implications of learned representations in TD-based methods. Yang and Nachum (2021) finds that pre-training and fixing the state representations can dramatically improve downstream learning. Geng et al. (2022) demonstrates that regularizing representation takes a more critical role in improving offline RL methods than merely imposing pessimism. Ma et al. (2023) provides an examination of the explicit representation distinction between in-sample and OOD state-action pairs for offline RL. In contrast, our approach is orthogonal to the above work. We introduce additional neural networks for feature extraction, enabling the learning of spaces where the latent distribution exhibits divisibility.

[2] $d_{\mathcal{H}}(\mathcal{P}, \mathcal{Q}) = 2\sup_{h\in\mathcal{H}} |\mathbb{E}_{\mathcal{P}}[h = 1] - \mathbb{E}_{\mathcal{Q}}[h = 1]|$

[3] $\lambda^\star \geq \inf_{h'\in\mathcal{H}}\{\epsilon_{\mathcal{P}}(h') + \epsilon_{\mathcal{Q}}(h')\}$

DATASET	BC	IQL	CQL	SPOT	POR	PRDC	STR	DIFFUSION -QL	CQL +ADS	**MCQ +LAD**
HALFCHEETAH-M	42.9	47.4	49.4	58.4	48.8	63.5	51.8	51.1	**73.9**	66.4
HOPPER-M	56.1	65.7	59.1	86.0	78.6	100.3	101.3	90.5	101.0	**103.6**
WALKER2D-M	76.6	81.1	83.6	86.4	81.1	85.2	85.9	87.0	91.3	**94.9**
HALFCHEETAH-MR	36.6	44.2	47.0	52.2	43.5	55.0	47.5	47.8	49.6	**61.0**
HOPPER-MR	19.3	94.8	98.6	100.2	98.9	100.1	100.0	101.3	102.4	**105.8**
WALKER2D-MR	24.8	77.3	71.3	91.6	76.6	92.0	85.7	95.5	93.7	**98.1**
HALFCHEETAH-ME	53.1	88.0	93.0	86.9	94.7	94.5	94.9	96.8	93.5	**101.5**
HOPPER-ME	52.7	106.2	90.0	111.4	99.3	109.2	111.9	111.1	**113.3**	112.7
WALKER2D-ME	102.5	108.3	109.8	112.0	109.1	111.2	110.2	110.1	112.1	**116.6**
AVERAGE ABOVE	51.6	79.2	80.4	85.9	80.1	90.1	87.7	87.9	92.3	**95.6**

Table 2: Average normalized scores of different methods on the benchmark. We **bold** the highest mean.



(a) Parameter K (b) Parameter δ (c) Ablation study

Figure 3: Results of parameter variations and ablation study.

5 Experiments

In this section, we conduct experiments to validate the effectiveness of LAD in terms of performance, hyperparameter robustness, ablation study, and visualization of latent distribution characterization using the D4RL benchmark (Fu et al. 2020). For simply, we abbreviate the names of the datasets from {`MEDIUM`, `MEDIUM-REPLAY`, `MEDIUM-EXPERT`} to {`M`, `MR`, `ME`} in all the tables. Regarding the algorithms relevant to our comparison, baseline results at the 1M gradient step are procured either by re-running the official code or directly extracting from the original papers. Our algorithm, in turn, is tested across five random seeds, with the reported findings reflecting the mean normalized results garnered from the last ten evaluations.

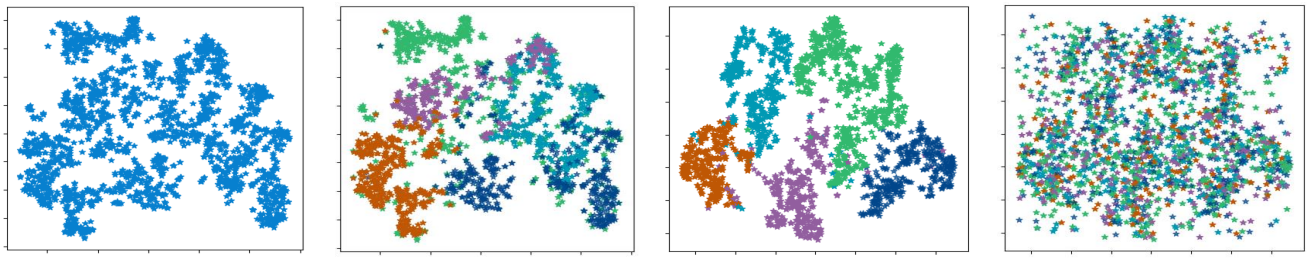
5.1 Performance

Application to strong baseline algorithms. Our LAD, a highly effective data-driven framework, can effectively integrate with strong baseline algorithms. Inspired by ADS (Wang et al. 2024), we combine LAD with TD3BC (Fujimoto and Gu 2021) and MCQ (Lyu et al. 2022), then compare with the baselines and ADS. Table 1 shows that LAD boosts the performance of the baselines and surpasses ADS in most cases. Furthermore, the results reveal a noteworthy phenomenon: our method exhibits superior performance on the `-MR` and `-ME` datasets. This enhanced performance can be attributed to the greater diversity of the latent distribution, which arises when the dataset contains a richer mix of samples. In this situation, our LAD framework can make full use of the favorable diversity to train models with better generalization, thereby yielding superior performance. This result aligns with our expectations and confirms the effectiveness of our approach. More results equipped with CQL (Kumar et al. 2020) are shown in Appendix C.

Comparison with related methods. We opt for the representative algorithm that demonstrates the most outstanding performance, MCQ+LAD, and compare it against other related classical algorithms as well as those designed to enhance generalizability, including BC, IQL (Kostrikov et al. 2021), CQL (Kumar et al. 2020), SPOT (Wu et al. 2022), POR (Xu et al. 2022), PRDC (Ran et al. 2023), DIFFUION-QL (Wang, Hunt, and Zhou 2022), and CQL+ADS (Wang et al. 2024). Table 2 exhibits the outcomes of different methods within representative gym environments. The results indicate that our method is not only competitive in the majority of cases but also has a distinct advantage in the average performance, providing reassurance of its effectiveness. Additional evidence on other tasks is provided in Appendix C.

5.2 Hyperparameter Robustness

The key parameters that define our LAD framework include the pre-processing sample size M, the initial latent distribution count K, and the fine-tuning parameter δ. Recognizing that M affects the computational resources for adjusting the latent distribution labels l, we assign a conservative value,



(a) The original dataset (b) The initial latent distributions obtained by Pre-processing (c) Separable latent distributions obtained by distribution learner L (d) Inseparable latent distributions obtained by bottleneck B in Step 3

Figure 4: The t-SNE visualization of latent distribution characterization by LAD.

such as $M = 5000$. An exhaustive list of parameters is listed in Appendix C. Next, we primarily analysis K and δ.

The effect of the initial latent distribution count K. The K is a noticeable and crucial parameter in our LAD framework. Typically, the larger the value of K, the fewer samples constituting each latent distribution, which can lead to an increase in the discrepancy of data points within each latent distribution, while the gaps between the latent distributions decrease as their boundaries become less distinct. We report the results of TD3BC+LAD on `HALFCHEETAH-MR` across various K in Figure 3(a). Observations indicate that the performance trend corresponding to the K aligns fundamentally with our preceding expectations. Consequently, we choose K = 5 as an appropriate selection.

The effect of the fine-tuning parameter δ. The parameter δ denotes the extent of fine-tuning the feature extractor. Figure 3(b) illustrates the performance of TD3BC+LAD on `HALFCHEETAH-MR` across various δ. The results show that LAD is more robust to δ. A natural consideration is whether omitting the use of the latent distribution information for fine-tuning the feature extractor would lead to a decline in performance. For this reason, we conduct an ablation study for the fine-tuning process in the following subsection.

5.3 Ablation Study

We perform an ablation study over the components in our method. First, we evaluate the significance of the latent distribution learner, which comes into play during Step 2. Figure 3(c) indicates that removing the distribution learner (w/o $L(s, a)$) results in a degradation in performance. That is because the role of the latent distribution learner is to act as an intermediate network that records the worst-case features learned, and learning directly through the feature extractor is problematic. We then verify the importance of using Equation (7) for fine-tuning during the iterations after Step 3 through ablation experiments. Using the original target y in place of y' in Equation (7) to update the feature network could skew the representation towards a Q-predictive bias, preventing the acquisition of the updated latent distribution features. This ultimately leads to suboptimal performance as depicted in Figure 3(c) (w/o Fine-tune). Hence, it is crucial to fine-tune the feature extractor with fine-grained latent distribution information. By doing so, we can align the representation to prioritize features associated with the latent distribution, thus facilitating a new round of iterations.

5.4 Visualization

To further verify the efficacy of LAD, we offer the t-SNE (der Maaten and Hinton 2008) visualization, representing the latent distribution's successful characterization by LAD, as shown in Figure 4. The original dataset of `HALFCHEETAH-MR` is displayed in Figure 4(a). Following that, Figure 4(b) illustrates the data distribution after the pre-processing phase, which presents an initial appearance of a multiple latent distribution by clustering with GMM. In Figure 4(c), we observe the data distribution characterized by LAD in Step 2, showcasing a maximization of distribution gaps. This visualization is derived from the feature extraction enabled by the distribution learner $L(s, a)$. Evidently, LAD successfully identifies latent distributions with a remarkable diversity, thereby bolstering generalization. In Figure 4(d), we present the results of using the bottleneck $B^{(3)}(s, a)$ in Step 3 to record the feature of inseparable latent distribution. Compared to the above visualization results, our LAD successfully characterizes the diverse latent distributions and effectively utilizes them, which supports the efficacy of LAD.

6 Conclusion

In this paper, we introduce a novel data-driven perspective for improving the generalization in offline RL. The principle is to characterize the multiple latent distributions that inherently exist in offline data and enable models to bridge distribution gaps, thereby enhancing generalization. We present the Latent Distribution Representation Learning (LAD) framework, which first identifies "worst-case" distributions to enlarge the diversity of distribution gaps and then reduces these gaps to learn invariant representations. We provide the theoretical insights behind LAD, demonstrating its ability to implicitly minimize the upper bound on the generalization error of offline RL in multi-distribution scenarios. Our extensive experiments convincingly support the efficacy of LAD. However, the iterative nature of our method currently limits its scalability to large data in the pre-processing phase. This limitation underscores the need for further research. In our future work, we plan to explore the use of the diffusion model for yielding diverse trajectories and fully leverage our LAD framework. We believe that our unique perspective on data distribution will stimulate further community discussion and inspire new approaches to RL.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2020AAA0106100), the National Natural Science Foundation of China (62276160), and the Natural Science Foundation of Shanxi Province, China (202203021211294, 202203021211291). We extend our heartfelt gratitude to the anonymous reviewers for their invaluable and constructive comments, and we express our deep appreciation to Ting Guo and Yi Ma for their significant support.

References

An, G.; Moon, S.; Kim, J. H.; and Song, H. O. 2021. Uncertainty-based Offline Reinforcement Learning with Diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, volume 34, 7436–7447.

Bai, C.; Wang, L.; Yang, Z.; Deng, Z.; Garg, A.; Liu, P.; and Wang, Z. 2022. Pessimistic Bootstrapping for Uncertainty-Driven Offline Reinforcement Learning. In *Proceedings of the Tenth International Conference on Learning Representations*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A Theory of Learning from Different Domains. *Machine Learning*, 79: 151–175.

Bishop, C. M. 2006. Pattern Recognition and Machine Learning. *Springer google schola*, 2: 531–537.

Chen, X.; Zhou, Z.; Wang, Z.; Wang, C.; Wu, Y.; and Ross, K. 2020. Bail: Best-action Imitation Learning for Batch Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, 18353–18363.

der Maaten, L. V.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Fakoor, R.; Mueller, J. W.; Asadi, K.; Chaudhari, P.; and Smola, A. J. 2021. Continuous Doubly Constrained Batch Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34, 11260–11273.

Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; and Levine, S. 2020. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. arXiv:2004.07219.

Fujimoto, S.; and Gu, S. S. 2021. A Minimalist Approach to Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34, 20132–20145.

Fujimoto, S.; Meger, D.; and Precup, D. 2019. Off-policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning*, 2052–2062.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59): 1–35.

Geng, X.; Li, K.; Gupta, A.; Kumar, A.; and Levine, S. 2022. Effective Offline RL needs Going beyond Pessimism: Representations and Distributional Shift. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*.

Kostrikov, I.; Fergus, R.; Tompson, J.; and Nachum, O. 2021. Offline Reinforcement Learning with Fisher Divergence Critic Regularization. In *Proceedings of the 38th International Conference on Machine Learning*, 5774–5783.

Kumar, A.; Fu, J.; Soh, M.; Tucker, G.; and Levine, S. 2019. Stabilizing off-policy Q-learning via Bootstrapping Error Reduction. In *Advances in Neural Information Processing Systems*, volume 32.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative Q-learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, 1179–1191.

Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv:2005.01643.

Li, J.; Zhang, E.; Yin, M.; Bai, Q.; Wang, Y.; and Wang, W. Y. 2023. Offline Reinforcement Learning with Closed-form Policy Improvement Operators. In *Proceedings of the 40th International Conference on Machine Learning*, 20485–20528.

Lu, W.; Wang, J.; Sun, X.; Chen, Y.; and Xie, X. 2023. Out-of-distribution Representation Learning for Time Series Classification. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Lyu, J.; Ma, X.; Li, X.; and Lu, Z. 2022. Mildly Conservative Q-learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 35, 1711–1724.

Ma, Y.; Jayaraman, D.; and Bastani, O. 2021. Conservative Offline Distributional Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34, 19235–19247.

Ma, Y.; Tang, H.; Li, D.; and Meng, Z. 2023. Reining Generalization in Offline Reinforcement Learning via Representation Distinction. In *Advances in Neural Information Processing Systems*, volume 36, 40773–40785.

Mao, Y.; Zhang, H.; Chen, C.; Xu, Y.; and Ji, X. 2023. Supported Trust Region Optimization for Offline Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning*, 23829–23851.

Qi, H.; Su, Y.; Kumar, A.; and Levine, S. 2022. Data-driven Offline Decision-making via Invariant Representation Learning. In *Advances in Neural Information Processing Systems*, volume 35, 13226–13237.

Ran, Y.; Li, Y.; Zhang, F.; Zhang, Z.; and Yu, Y. 2023. Policy Regularization with Dataset Constraint for Offline Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning*, 28701–28717.

Sun, P.; Zhou, W.; and Li, H. 2020. Attentive Experience Replay. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, 5900–5907.

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

Wang, D.; Li, L.; Wei, W.; Yu, Q.; Jianye, H.; and Liang, J. 2024. Improving Generalization in Offline Reinforcement Learning via Adversarial Data Splitting. In *Proceedings of the 41st International Conference on Machine Learning*.

Wang, Z.; Hunt, J. J.; and Zhou, M. 2022. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning. In *Proceedings of the Tenth International Conference on Learning Representations*.

Wu, J.; Wu, H.; Qiu, Z.; Wang, J.; and Long, M. 2022. Supported Policy Optimization for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 35, 31278–31291.

Wu, Y.; Zhai, S.; Srivastava, N.; Susskind, J. M.; Zhang, J.; Salakhutdinov, R.; and Goh, H. 2021. Uncertainty Weighted Actor-Critic for Offline Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, 11319–11328.

Xiao, C.; Dai, B.; Mei, J.; Ramirez, O. A.; Gummadi, R.; Harris, C.; and Schuurmans, D. 2022. Understanding and Leveraging Overparameterization in Recursive Value Estimation. In *Proceedings of the Tenth International Conference on Learning Representations*.

Xu, H.; Jiang, L.; Li, J.; and Zhan, X. 2022. A Policy-guided Imitation Approach for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 35, 4085–4098.

Yang, M.; and Nachum, O. 2021. Representation Matters: Offline Pretraining for Sequential Decision Making. In *Proceedings of the 38th International Conference on Machine Learning*, 11784–11794.

Zhao, H.; Combes, R. T. D.; Zhang, K.; and Gordon, G. 2019. On Learning Invariant Representations for Domain Adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, 7523–7532.