Variational OOD State Correction for Offline Reinforcement Learning

Ke Jiang[*1], **Wen Jiang**[*1], **Xiaoyang Tan**[1†]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
{ke_jiang, darren.jum, x.tan}@nuaa.edu.cn

Abstract

The performance of Offline reinforcement learning is significantly impacted by the issue of *state distributional shift*, and out-of-distribution (OOD) state correction is a popular approach to address this problem. However, previous methods correct the agent's transition distributions in a supervised way, which significantly degrades the flexibility and robustness. In this paper, we propose a novel method named Density-Aware Safety Perception (DASP) for OOD state correction. Specifically, our method encourages the agent to prioritize actions that lead to outcomes with higher data density, thereby promoting its operation within or the return to in-distribution (safe) regions. To achieve this, we optimize the objective within a variational framework that concurrently considers both the potential outcomes of decision-making and their density, thus providing crucial contextual information for safe decision-making. Finally, we validate the effectiveness and feasibility of our proposed method through extensive experimental evaluations on the offline MuJoCo and AntMaze suites.

Introduction

Deep reinforcement learning (RL) has achieved significant success in various domains, including robotics tasks (Mnih et al. 2015; Peng et al. 2017), game playing (Silver et al. 2017), and large language models (Achiam et al. 2023). However, its broader application is constrained by the challenges of interacting with real-world environments, which can be costly or risky (Garcıa and Fernández 2015). Offline reinforcement learning addresses these challenges by enabling agents to learn from previously corrected datasets (Zhang and Tan 2024), thereby avoiding high-risk interactions.

Despite this, deploying an online RL framework in an offline setting can significantly hinder the performance of the learned policy. This issue arises from the well-known *distributional shift* problem (Fujimoto, Meger, and Precup 2019; Kumar et al. 2020), where the TD target may be overestimated for actions with low data density, also known as out-of-distribution (OOD) actions, during training, resulting in extrapolation errors (Jin, Yang, and Wang 2021) that degrade the agent's performance. Previous works, such as Conservative Q-Learning (CQL)(Kumar et al. 2020), Bootstrapping Error Accumulation Reduction (BEAR)(Wu, Tucker, and Nachum 2019), and Supported Policy Optimization (SPOT)(Wu et al. 2022), have addressed this problem by suppressing OOD actions through specific regularization techniques. However, these methods primarily focus on avoiding OOD actions while neglecting the issue of *state distributional shift* (Jiang, Yao, and Tan 2023; Zhang et al. 2022), which occurs when encountering OOD or low-density states during test, leading to cumulative errors and task failure, i.e., the phenomenon of State deviation.

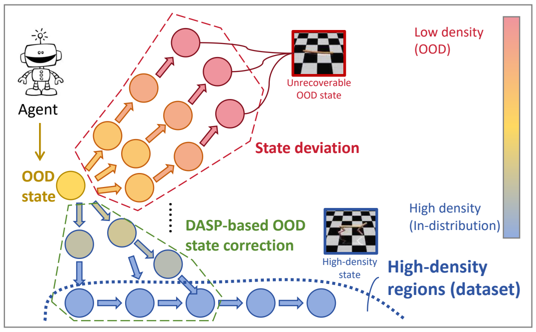
Agent
OOD state
State deviation
Unrecoverable OOD state
Low density (OOD)
High density (In-distribution)
DASP-based OOD state correction
High-density state
High-density regions (dataset)

Figure 1: The basic idea behind the proposed DASP-based OOD state correction - guiding the agent from OOD (low-density) to the high density regions according to the dataset.

By OOD states, we mean these states that experience low visitation frequency by the behavior policy. In other words, OOD states exhibit lower density compared to in-distribution states based on the offline dataset. From this perspective, as is shown in Figure 1, OOD state correction can be viewed as a process that guides the agent to transition from low-density states to high-density states, ensuring that decision-making is supported by sufficient data and thereby maintaining safety. Such guidance via density, comparing to previous supervision-style constraints, is more flexible and robust in penalizing the OOD state visitation, because the agent need not recover to specific states but to higher density regions (Kang et al. 2022), which to the best of our knowledge, has yet to be applied to OOD state correction in offline RL.

In this paper, we introduce a novel method called Density-Aware Safety Perception (DASP) to realize OOD state correction, hence dealing with the problem of *state distributional*

[*]These authors contributed equally.

[†]Corresponding author.

shift. The basic idea is to guide OOD state correction with an additional reward mechanism based on density optimization. For this purpose, inspired by the likelihood improvement mechanism commonly used in the deep generative model (e.g., diffusion model (Janner et al. 2022)), we propose a novel offline RL objective that encourages the new policy to prefer to choose those actions that lead to higher data density, besides obtaining higher return. Specifically, we optimize the objective within a variational framework, where DASP predicts the density based on the joint features of the inputted state-action pairs and their potential outcomes. This allows DASP to directly predict one-step forward features and estimate their density to assess the contextual safety of current decision-making, thereby guiding OOD state correction during policy optimization. In practical implementation, our method utilizes a modular algorithmic design, requiring only minor modifications to standard off-policy algorithms to be effective. Our experiments show that the proposed method outperforms several closely related state-of-the-art (SOTA) methods in offline MuJoCo control and AntMaze suites across various settings.

To sum up, the core contributions of DASP are:

- **A Novel Paradigm for OOD State Correction:** We introduce DASP, a novel framework that reframes the problem of OOD state correction via density-scored guidance.
- **A Unified and Efficient Variational Framework:** We design a compact variational model capable of simultaneously performing forward dynamics prediction and estimating the density of the resulting state.
- **Superior Performance Validated by Extensive Experiments:** The effectiveness of DASP is demonstrated through experiments on different offline RL benchmarks.

Related Works

Policy Constraint Methods. Early offline RL methods like CQL (Kumar et al. 2020), BEAR (Wu, Tucker, and Nachum 2019), Supported Policy Optimization (SPOT) (Wu et al. 2022) and Explicit Behavior density (CPED) (Zhang et al. 2023) focus on OOD action suppression. They regularize the policy to prevent it from selecting actions not well-supported by the dataset. While effective at preventing initial deviation, these methods provide little recourse once the agent has already entered an OOD state, a common occurrence in stochastic environments.

OOD state correction. More recent work directly addresses OOD state correction, the central problem of this paper. The goal is to guide the agent from low-density (OOD) states back to high-density (in-distribution) regions. Existing methods like SDC (Zhang et al. 2022) achieve this by enforcing a strict alignment between the next-state distributions of the learned and behavior policies. Others, such as OSR (Jiang, Yao, and Tan 2023) and SCAS (Mao et al. 2024), rely on separately trained dynamics models to predict and select safe transitions. These approaches have key limitations: **Overly Restrictive Alignment:** Strict distributional matching (as in SDC) can unnecessarily constrain the policy, hindering its ability to discover better-yet-safe trajectories. **Dependence on Dynamics Models:** The performance of methods like OSR and SCAS is highly sensitive to the accuracy of the dynamics model, which is difficult to learn reliably from offline data and prone to compounding errors.

Preliminaries

Reinforcement learning is commonly framed as a Markov Decision Process (MDP), denoted by the tuple $(S, A, P, R, \gamma, \rho_0)$. In this representation, S signifies the state space, A indicates the action space, P is the transition probability matrix, R represents the reward function, γ is the discount factor, and ρ_0 is the initial state distribution. A policy $\pi : S \to A$ is established to make decisions during interactions with the environment.

Typically, the Q-value function is expressed as $Q^\pi(s,a) = (1-\gamma)\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(a_t|s_t))|s,a]$, which conveys the anticipated cumulative rewards. For ease of reference, the γ-discounted future state distribution (or stationary state distribution) is expressed as $d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr(s_t = s; \pi, \rho_0)$, with ρ_0 representing the initial state distribution and $(1-\gamma)$ acting as the normalization factor.

In an offline context, Q-Learning (Watkins and Dayan 1992) derives a Q-value function $\hat{Q}(s,a)$ and a policy π from a dataset $\mathcal{D}$ that is gathered via a behavior policy π_β. This dataset comprises quadruples $(s,a,r,s') \sim d^{\pi_\beta}(s)\pi_\beta(a|s)P(r|s,a)P(s'|s,a)$. The goal is to minimize the Bellman error across the offline dataset (Watkins and Dayan 1992), employing exact or approximate maximization techniques, such as CEM (Kalashnikov et al. 2018), to retrieve the greedy policy as follows:

$$\min_Q \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[r + \gamma\mathbb{E}_{\pi(a'|s')}Q(s',a') - Q(s,a)]^2 \quad (1)$$

$$\max_\pi \mathbb{E}_{s\sim\mathcal{D}}\mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s,a)]. \quad (2)$$

OOD State Correction. OOD state correction, also known as State recovery, based offline RL methods, such as SDC (Zhang et al. 2022), OSR (Jiang, Yao, and Tan 2023) and SCAS (Mao et al. 2024), have demonstrate their advantage in developing reliable and robust agents. The basic idea of such methods is to train a policy choosing actions whose state visitation frequency is as closer to that of the behavior policy as possible. It could be represented as follows,

$$\min_\pi \mathbb{E}_{s\sim\mathcal{D}} Dis\big(P(\cdot|s,\pi_\beta(\cdot|s)), P(\cdot|s,\pi(\cdot|s))\big) \quad (3)$$

where P is the dynamics model, and Dis is some kind of distance measure, which is Maximum Mean Discrepancy (MMD) in (Zhang et al. 2022) while Kullback-Leibler (KL) Divergence in (Jiang, Yao, and Tan 2023; Mao et al. 2024).

The Method

In this section, we provide a detailed description of the proposed density-aware safety perception framework, termed DASP, to address the issue of *state distributional shift* in offline reinforcement learning.

The Motivation

Out-of-distribution (OOD) states are defined as those with low density in the dataset, so the aim of OOD state correction is to guide the agent back to high-density regions, thereby ensuring that decision-making is supported by sufficient data. Intuitively, when aiming to identify the high - density regions of offline data, the approach is to leverage the s_t distribution information inherent in the dataset. Specifically, techniques such as the diffusion model (Janner et al. 2022) or score matching (Hyvärinen and Dayan 2005) can be employed to determine the direction in which the s_t likelihood experiences an increase, e.g., using a neural network to predict the vector of the score function for a given query state. Subsequently, during deployment, preference is given to the directions that exhibit a high degree of consistency with the likelihood - increasing direction estimated by the score function network. In essence, the action chosen by the agent is a weighted synthesis of two key elements: 1) actions associated with a relatively large reward; 2) actions whose resulting effects align with the direction of the score function.

Nevertheless, a notable limitation of the aforementioned straightforward solution lies in its ignorance of the knowledge context of offline reinforcement learning, failing to account for the impact of factors such as the behavior policy and the environment model during the modeling procedure. In light of this, this paper puts forward a more integrated objective function (Eq.(4)), as presented in the next section.

Density-Aware Safety Perception

Given a state s, we first formulate the objective for OOD state correction as follows:

$$\max_{\pi} \mathbb{E}_{a\sim\pi(\cdot|s),s'\sim P(\cdot|s,a)} \log d^{\pi_\beta}(s') \tag{4}$$

where $P(\cdot|s,a)$ represents the dynamics of the environment, and d^{π_β} is the stationary state distribution of the behavior policy π_β. The objective in Eq. (4) is referred to as Density-Aware Safety Perception (DASP), which evaluates the safety of the input state-action pairs based on the data density of their consequences. We then utilize DASP as a regularization term in policy optimization to prioritize actions that lead the agent toward regions of higher density, thus satisfying safety requirements.

In OOD state correction objective in Eq.(4), the $P(\cdot|s,a)$ and d^{π_β} are two complicated distributions that are hard to estimate explicitly. Therefore, we implicitly estimate them or their lower bound with the framework of variational inference. First, we approximate the d^{π_β} via maximum likelihood estimation, i.e.,

$$d^{\pi_\beta} \approx \arg\max_{d} \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \log d(s') \tag{5}$$

$$= \arg\max_{d} \mathbb{E}_{(s,a)\sim\mathcal{D},s'\sim P(s'|s,a)} \log d(s') \tag{6}$$

Then we remark that the estimation of one-step forward density, i.e., $\mathbb{E}_{s'\sim P(s'|s,a)} \log d(s')$, is the core to realize the OOD state correction. Then Theorem 1 gives the solution by estimating the lower bound of the term $\mathbb{E}_{s'\sim P(s'|s,a)} \log d(s')$ by introducing two variational distributions.

Theorem 1. *The term $\mathbb{E}_{s'\sim P(s'|s,a)} \log d(s')$ could be lower bounded by solving the following optimization problem in the offline setting,*

$$\max_{q_1,q_2} \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \Bigg[\int dz \cdot q_1(z|s') \log P(s'|z) - KL(q_2(z|s,a)\|P(z)) - KL(q_1(z|s')\|q_2(z|s,a)) \Bigg] \tag{7}$$

where $q_1(z|s')$ and $q_2(z|s,a)$ are two variational distributions. $KL(\cdot\|\cdot)$ is the KL-divergence between two distributions. $P(s'|z)$ is the poster distribution.

The proof is found in Appendix. In Eq.(7) the first term represents the reconstruction loss of the consequence s'; the second term measures the divergence between the encoding distribution $q_2(z|s,a)$ and the prior distribution $P(z)$, which should be minimized; the third term enables the encoder q_2 to directly predict the consequential feature distribution $q_1(z|s')$. This embeds the contextual information into the feature, thereby enabling the decoder to reconstructing the outcome states from either themselves or their previous state-action pairs. The most advantage of this solution is that we can reuse the models to approximate both the dynamics model $P(s'|s,a)$ and the density model $d^{\pi_\beta}(s)$: by the combination of the encoder $q_2(z|s,a)$ and the poster distribution (decoder) $P(s'|z)$, we can predict the consequence of the inputted (s,a); on the other hand, after we have the estimated consequence, we can calculate its density by the variational result in Eq.(7). The detailed utilization would be discussed in the next section.

Finally, with the objective in Eq.(7), we can learn the two variational distribution estimators q_1 and q_2, through which the one-step forward density $\mathbb{E}_{s'\sim P(\cdot|s,a)} d^{\pi_\beta}(s')$ could be variationally estimated. Then, in the next section, we introduce how to utilize this module, also named as DASP, to conduct OOD state correction in an offline manner.

DASP-based OOD State Correction

First of all, in order to generate OOD states for training, like previous works (Jiang, Yao, and Tan 2023; Zhang et al. 2022; Mao et al. 2024), we attach Gaussian noise $\mathcal{N}(0,\sigma^2)$ onto the states s from the dataset $\mathcal{D}$, denoted as $\hat{s}$. For OOD state correction in this paper, once the agent entering those OOD states $\hat{s}$, we aim to correct it to restore to safe states with high data density according to the offline dataset. Note that this objective can be reformulated as follows,

$$\max_{\pi} \mathbb{E}_{s\sim\mathcal{D},\hat{s}\sim\mathbb{B}_\sigma(s)} \mathbb{E}_{a\sim\pi(\cdot|\hat{s}),\hat{s}'\sim P(\cdot|\hat{s},a)} \log d^{\pi_\beta}(\hat{s}') \tag{8}$$

where the $\mathbb{B}_\sigma(s)$ is a Gaussian perturbation ball with center s and radius σ. The objective in Eq.(8) utilizes a one-step forward density module to attach the preference of the actions that could lead to consequences with high data density onto the new policy, hence satisfying the safety requirements for offline RL. Then the practical implementation based on the variational results are as follows,

Parametrization and construction of dynamics model. Before we handle the policy optimization regularization in Eq.(8), we need to parameterize the three distribution in Eq.(7) : the poster distribution $P(s'|z)$ is parameterized with

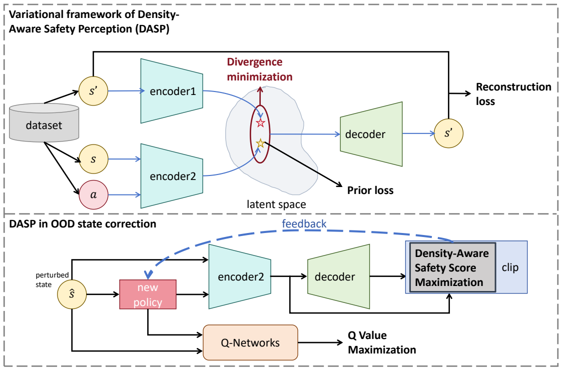

Figure 2: The framework of the proposed DASP and its utilization for OOD state correction. In the top figure: the reconstruction loss, prior loss and divergence minimization are the 3 terms in Eq.(10) respectively. The procedure in the buttom figure represents the policy optimization in Eq.(12) .

$P_\phi(s'|z)$, which could also be seen as the decoder module; the two variational distribution $q_1(z|s')$ and $q_2(z|s,a)$ are parameterized with $q_\psi(z|s')$ and $q_\theta(z|s,a)$ (corresponding to two encoders respectively in Figure 2(top)) . In this way, we reformulate the optimization problem by,

$$\theta^*, \psi^*, \phi^* = \arg\max_{\theta,\psi,\phi} \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \Big[\int q_\psi(z|s') \log P_\phi(s'|z) dz \\ - KL(q_\theta(z|s,a) \| P(z)) - KL(q_\psi(z|s') \| q_\theta(z|s,a)) \Big] \quad (9)$$

Then the above optimization could be further solved by methods like in (Doersch 2016; Burda, Grosse, and Salakhutdinov 2015). Specially, all the parameterized distributions are assumes as Gaussian - $q_\theta(z|s,a) = \mathcal{N}(\mu_\theta, \sigma_\theta; s, a)$, $q_\psi(z|s') = \mathcal{N}(\mu_\psi, \sigma_\psi; s')$ and $P_\phi(s'|z) = \mathcal{N}(\mu_\phi, \sigma_\phi; s')$. Suppose the prior distribution $P(z) = \mathcal{N}(0, I)$, then the above formulation in Eq.(9) could be transferred into the loss function as,

$$\mathcal{L}_{dasp}(s, a, s'; \theta, \psi, \phi) = \mathbb{E}_{z\sim q_\theta(z|s,a)} \|\mu_\phi(z) - s'\|_2^2 \\ -\frac{1}{2}\Big[\sum_i^K (1 + \log(\sigma_{\theta,i}^2) - \mu_{\theta,i}^2 - \sigma_{\theta,i}^2)\Big] \\ -\frac{1}{2}\Big[\sum_{i=1}^K (1 + \log\frac{\sigma_{\psi,i}^2}{\sigma_{\theta,i}^2} - \frac{\sigma_{\psi,i}^2 + (\mu_{\psi,i} - \mu_{\theta,i})^2}{\sigma_{\theta,i}^2})\Big] \quad (10)$$

where i represents the value of i^{th} dimension of the K-dimensional variable.

The forward dynamics model $P(s'|s,a)$ could be estimated by the combination of $P_{\phi^*}(s'|z)$ and $q_{\theta^*}(z|s,a)$. Here the $q_{\theta^*}(z|s,a)$ could be seen as an approximation of $q_\psi(z|s')$ due to the minimization of the divergence between the representations generated by these two encoders, hence the combined module could predict the $s' \sim P(s'|s,a)$ from $z \sim q_{\theta^*}(z|s,a)$ with low bias. Then the approximated dynamical model is denoted as $\hat{P}(s'|s,a)$.

DASP-based actor regularization. We construct the estimation term[1] for the objective in Eq.(8) based on the variational results and the parameterization. To be specific, the OOD state correction term could be approximated by,

$$\mathcal{R}(\hat{s}, a) = \mathbb{E}_{\hat{s}'\sim\hat{P}(\cdot|\hat{s},a)} f_\tau(\mathcal{L}_{dasp}(\hat{s}, a, \hat{s}'; \theta^*, \psi^*, \phi^*)) \quad (11)$$

where $(\theta^*, \psi^*, \phi^*)$ is the solution by minimizing the DASP loss $\mathcal{L}_{dasp}$ over the dataset $\mathcal{D}$ and f_τ is a clip function with threshold τ. The use of the clipping function f_τ is motivated by our objective, which is not to maximize likelihood but to regularize the agent's visitation to ensure sufficient density, specifically above a specified threshold τ. Please note that, instead of pretraining the dynamics model separately, the $\hat{P}(s'|s,a)$ is constructed by the modules in $\mathcal{L}_{dasp}$, hence formulating the indicator $\mathcal{R}(\hat{s}, a)$ a more compact implementation compared with other methods (Zhang et al. 2022; Jiang, Yao, and Tan 2023; Mao et al. 2024). The actor loss is,

$$\max_\pi \mathbb{E}_{s\sim\mathcal{D}} \Big[\mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s,a)] + \alpha \cdot \mathbb{E}_{\hat{s}\sim\mathbb{B}_\epsilon(s), a\sim\pi(\cdot|\hat{s})} \mathcal{R}(\hat{s}, a) \Big] \quad (12)$$

where α is the balance coefficient of the DASP term. Besides, we also utilize a momentum-based optimizer, e.g. Adam, in implementation to avoid the problem of local optimum.

Remark 1. *A lower $\mathcal{L}_{dasp}$ value indicates a better density model, which in turn provides a more accurate $\mathcal{R}(\hat{s}, a)$ score. Therefore, maximizing $\mathcal{R}(\hat{s}, a)$ aligns with our intuition: it encourages the policy π to select actions a with higher next step data density (obtained implicitly via the trained encoder and decoder in Eq.(10)).*

Overall Algorithm. Figure 2 gives the network architecture of the proposed DASP approach, while the whole training algorithm is shown in Algorithm 1.

Algorithm 1: DASP-based offline RL framework

Input: offline dataset $\mathcal{D}$, maximal update iterations T,
Parameter: policy network π, Q-networks Q_1, Q_2, DASP module $\mathcal{R}$,
Output: learnt policy network π

```
Initialize the policy network, Q-networks and the DASP module.
Pretrain the DASP module $\mathcal{R}$ according to Eq.(10).
Let $t = 0$.
while $t < T$ do
    Sample mini-batch of N samples $(s, a, r, s')$ from $\mathcal{D}$.
    Perturb $s$ with Gaussian Noise and get $\hat{s}$.
    Feed $\hat{s}$ into the policy network, get the action $a$ and calculate the DASP score $\mathcal{R}(\hat{s}, a)$.
    Update the Q-networks according to Eq.(1),
    Update the policy network $\pi$ according to Eq.(12).
end while
return learnt policy network $\pi$.
```

Experiments

In experiments we answer the following three key questions:

1) Does DASP achieve the state-of-the-art performance on standard MuJoCo benchmarks compared to the latest closely related methods?

[1] The validation study for this term is shown in Sec..

Table 1: Results of **DASP(ours)**, CQL, PBRL, SPOT, SVR, EDAC, RORL, SDC , OSR-10 and SCAS on D4RL averaged over 4 seeds. We bold the highest scores in each task.

		CQL	PBRL	SPOT	SVR	EDAC	RORL	SDC	OSR-10	SCAS	DASP(Ours)
halfcheetah	r	17.5	11.0	35.3	27.2	28.4	28.5	**36.2**	26.7	12.2	32.4±0.9
	m	47.0	57.9	58.4	60.5	65.9	66.8	47.1	67.1	46.6	**70.4**±2.9
	m-e	75.6	92.3	86.9	94.2	106.3	107.8	101.3	108.7	91.7	**112.1**±2.0
	m-r	45.5	45.1	52.2	52.5	61.3	61.9	47.3	64.7	44.0	**67.1**±3.9
	e	96.3	92.4	97.6	96.1	106.8	105.2	106.6	106.3	106.6	**107.4**±1.8
hopper	r	7.9	26.8	33.0	31.0	25.3	31.4	10.6	30.4	31.4	**33.1**±0.3
	m	53.0	75.3	86.0	103.5	101.6	104.8	91.3	105.5	102.5	**108.6**±0.9
	m-e	105.6	110.8	99.3	111.2	110.7	112.7	112.9	113.2	109.7	**116.0**±6.3
	m-r	88.7	100.6	100.2	103.7	101.0	102.8	48.2	103.1	101.6	**104.1**±1.1
	e	96.5	110.5	112.3	111.1	110.1	112.8	112.6	**113.6**	112.8	113.5±1.0
walker2d	r	5.1	8.1	21.6	2.2	16.6	21.4	14.3	19.7	1.4	**23.9**±0.8
	m	73.3	89.6	86.4	92.4	92.5	102.4	81.1	102.0	82.3	**108.6**±2.7
	m-e	107.9	110.8	112.0	109.3	114.7	121.2	105.3	**123.4**	108.4	123.0±2.6
	m-r	81.8	77.7	91.6	95.6	87.1	90.4	30.3	93.8	78.1	**99.5**±1.7
	e	108.5	108.3	109.7	110.0	115.1	**115.4**	108.3	115.3	115.0	115.3±1.6
average		67.4	74.4	78.8	80.0	82.9	85.7	70.2	86.2	76.3	**89.0**
antmaze	umaze	82.6	-	93.5	-	-	**96.7**	81.4	89.9	90.4	94.6±3.2
	umaze-div	10.2	-	40.7	-	-	**90.7**	49.6	74.0	63.8	65.5±6.1
	med-play	59.0	-	74.7	-	-	76.3	55.0	66.0	76.6	**79.0**±4.6
	med-div	46.6	-	79.1	-	-	69.3	56.6	80.0	**80.4**	79.6±4.9
	large-play	16.4	-	35.3	-	-	16.3	20.8	37.9	49.0	**49.3**±8.5
	large-div	3.2	-	36.3	-	-	41.0	25.8	37.9	**50.6**	43.4±9.3
average		36.3	-	59.9	-	-	65.1	48.2	64.3	68.5	**68.6**

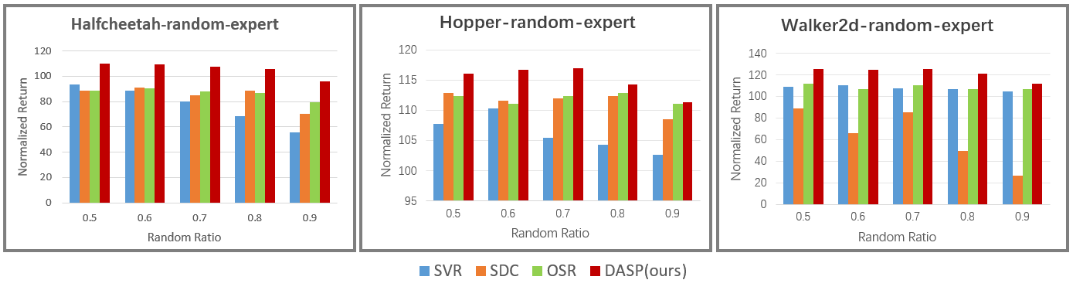

Figure 3: The results on the MuJoCo benchmarks with different levels of sub-optimal data.

2) Is DASP able to recover from out-of-distribution (OOD) states successfully?
3) Is DASP term robust enough to deal with unfavorable conditions, such as sub-optimal demonstrations or inefficient samples, in practical deployments?

Our experimental section is organized as follows: First, by fairly comparing the performance of learning policies using traditional methods on standard MuJoCo benchmarks, we verify that the proposed method DASP achieves superior performance among these methods, answering Question 1. Then, to answer Question 2, we verify the ability of DASP to recover from OOD states using the Out-of-sample MuJoCo (OOSMuJoCo) benchmarks, as described in (Jiang, Yao, and Tan 2023). Finally, to answer Question 3, we evaluate DASP on benchmarks under the settings of sub-optimal data and inefficient data (Zhang et al. 2022). Additionally, we conducted an ablation study and designed an experiment to analysis the validity of the DASP regular term. A brief introduction of our code is available in Appendix.

Comparisons on Standard Benchmarks

In this section, we compare the two proposed implementations of our method with several significant methods, including CQL (Kumar et al. 2020), PBRL (Bai et al. 2022), SPOT (Wu et al. 2022), SVR (Mao et al. 2023), EDAC (An et al. 2021), RORL (Yang et al. 2022), SDC (Zhang et al. 2022), OSR-10 (Jiang, Yao, and Tan 2023) and SCAS (Mao et al. 2024), based on the D4RL (Fu et al. 2020) dataset in the standard MuJoCo benchmarks and AntMaze tasks.

MuJoCo (D4RL). The MuJoCo domain have three types of high-dimensional control environments representing different robots in D4RL: Hopper, Halfcheetah and Walker2d, and five kinds of datasets: ’random’, ’medium’, ’medium-replay’, ’medium-expert’ and ’expert’. The **AntMaze** domain is a more challenging navigation domain with sparse rewards and multitask data, which contains three types of datasets, namely ‘umaze’, ‘medium’, and ‘large’.

The results is shown in Table 1, where part of the results for the comparative methods are obtained by (Yang et al.

Table 2: Results of RORL, SDC, OSR-10 and DASP in OOSMuJoCo setting on the normalized return and decrease metric averaged over 4 seeds. The noteworthy results are bolded.

	RORL		SDC		OSR-10		DASP	
Task name	score	dec.(%)	score	dec.(%)	score	dec.(%)	score	dec.(%)
Halfcheetah-OOS-slight	55.3	17.2	45.1	**4.3**	**59.4**	11.5	58.5±1.2	14.8
Halfcheetah-OOS-moderate	47.6	28.7	39.8	**15.5**	56.5	15.8	**56.9±2.2**	17.2
Halfcheetah-OOS-large	35.4	47.0	34.0	27.8	50.8	24.3	**54.6±4.2**	**20.5**
Hopper-OOS-slight	100.4	4.2	85.7	6.1	100.8	4.5	**101.9±0.2**	**4.1**
Hopper-OOS-moderate	94.4	9.9	82.9	9.2	98.3	**6.8**	**98.5±0.5**	**7.3**
Hopper-OOS-large	82.1	21.7	75.5	17.3	**94.7**	**10.2**	89.5±2.4	15.8
Walker2d-OOS-slight	92.9	**9.3**	71.0	12.5	92.4	9.4	**93.3±0.7**	10.4
Walker2d-OOS-moderate	86.5	15.5	69.5	14.3	90.3	**11.5**	**91.4±1.1**	**12.2**
Walker2d-OOS-large	71.8	29.9	65.3	19.5	88.6	**13.1**	**89.1±4.6**	**14.4**

2022; Jiang, Yao, and Tan 2023; Mao et al. 2024). On the MuJoCo tasks, we have observed that the performance of all methods experiences a significant decrease when learning from datasets such as 'random', 'medium', 'medium-replay', and 'medium-expert', which are collected by sub-optimal behavior policies. This highlights the inherent difficulty in getting rid of the influence on the sub-optimal behavior strategy in practical settings. However, our proposed methods, DASP, consistently outperform other approaches across most benchmarks, particularly surpassing methods that rely on behavior cloning such as CQL, PBRL, and EDAC. Furthermore, DASP achieve state-of-the-art performance in terms of the average score. Additionally, we would like to emphasize that DASP demonstrates significant improvements over the state-of-the-art conservative methods (e.g., SVR and OSR) on the 'medium' and 'medium-replay' datasets. This notable margin can be attributed to DASP's ability to avoid aligning the transition of the dataset through its flexibility in correcting the consequences. This further underscores the advantages of DASP in effectively handling sub-optimal offline data. In the following section, we will explore DASP's ability to recover from OOD states. On the AntMaze tasks, DASP outperforms all the methods in total score, and is very close to SOTA method in each item.

Evaluation on Out-of-sample MuJoCo Setting

To investigate the agent's behavior in unseen (OOD) states and assess whether the proposed DASP enables recovery from out-of-sample situations, we introduce the OOSMuJoCo benchmarks from (Jiang, Yao, and Tan 2023) and implement other related methods: RORL, SDC, and OSR-10 on 'medium' datasets. OOSMuJoCo simulates external forces to push the agent into out-of-sample states in Halfcheetah, Walker2d, and Hopper, with three levels of force: slight, moderate, and large.

Table 2 presents the scores and performance decreases of these policies across the 9 OOSMuJoCo benchmarks. The performance decrease is calculated as the percentage reduction in scores from OOSMuJoCo compared to the standard MuJoCo environments shown in Table 1. The results indicate that the proposed DASP outperforms other methods in scores, particularly in the 'Halfcheetah' and 'Walker2d' benchmarks with larger perturbations, likely due to these benchmarks' higher sensitivity to OOD situations. Additionally, we note that DASP and OSR-10 exhibit comparable performance decreases across the environments, suggesting that methods incorporating the DASP constraint are at least as robust as OSR-10 and RORL in handling OOD situations. Next, we will explore DASP's capabilities in sub-optimal demonstrations.

Evaluation on Sub-Optimal Datasets

In this section, we further investigate the feasibility of the proposed DASP on different levels of sub-optimal offline datasets, where 'expert' and 'random' datasets are mixed in various ratios. This setting is widely used, as seen in (Zhang et al. 2022; Mao et al. 2023; Jiang, Yao, and Tan 2023). In this paper, the proportions of 'random' data are 0.5, 0.6, 0.7, 0.8, and 0.9 for 'Halfcheetah', 'Hopper', and 'Walker2d'.

We compare the proposed DASP with SVR (Mao et al. 2023), OSR (Jiang, Yao, and Tan 2023), and SDC (Zhang et al. 2022). As shown in Figure 3, our method outperforms the other three methods across the three control environments in terms of normalized scores. We observed that our proposed method exhibits a significantly lower decrease rate over the 'Halfcheetah' benchmark compared to the other two methods as the random ratio increases, which can be attributed to the agent's heightened sensitivity to the quality of data collection in this environment. Furthermore, when testing on the 'Hopper' and 'Walker2d' benchmarks, we note that DASP demonstrates the least decrease in performance among all methods when the random ratio reaches 0.9. This highlights the advantage of the implicit implementation in addressing more complex tasks and learning from lower-quality data in practical scenarios. Therefore, we emphasize that our method is better equipped for learning with sub-optimal data and exhibits improved stability and performance across various benchmarks.

Evaluation on Data Inefficient Benchmarks

Sub-optimal data can be considered as a form of noisy-labeled data, where certain states 's' are associated with sub-optimal (incorrect) labels, denoted as action 'a'. Previous studies (Wang and Tan 2014; Bootkrajang and Kabán 2012) have shown that learning performance is significantly influenced by the size of the training data. This motivated

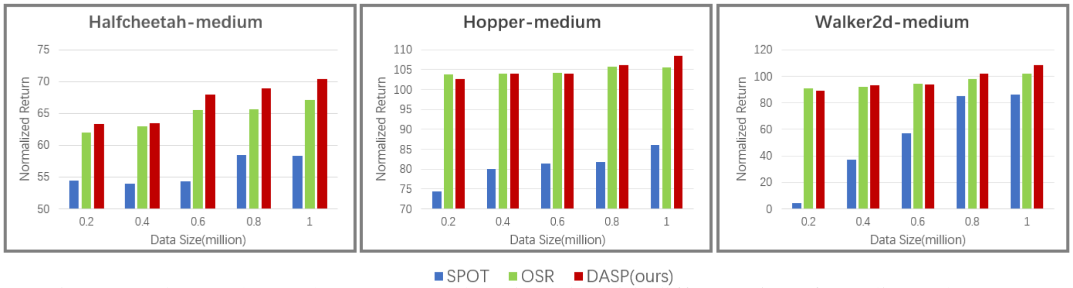

Figure 4: The results on three MuJoCo benchmarks with different size of 'medium' datasets.

us to investigate the performance of different methods under varying sizes of sub-optimal data.

In this section, as depicted in Figure 4, we compare our proposed DASP method with typical offline RL approaches, namely SPOT and OSR-10, using different sizes of training data (0.2, 0.4, 0.6, 0.8 million). We select the 'medium' datasets as the sub-optimal training data. Our observations reveal that the DASP method consistently outperforms the other two methods across all data sizes. Notably, both DASP and OSR-10 exhibit superior performance compared to SPOT by a significant margin. Furthermore, the advantage of DASP over OSR-10 becomes more pronounced as the data size increases. These findings demonstrate that the challenges of dealing with OOD states in offline RL would diminish with massive data sizes. However, when the data is insufficient, OOD state correction methods ,including our proposed DASP, exhibit better generalization capabilities.

Validity Analysis of DASP Regularization

In this section, we perform a experiments within the MuJoCo environment to Analysis the validity of key components in Eq. 12. We first generated two sets of actions for a given set of states from dataset: one set with safe outcomes, generated by a well trained policy in the medium-expert dataset; the other set with unsafe outcomes, composed of a series of random actions. We then utilized either the true dynamics model (TDM) or our DASP model to predict the next states of these actions and assess their safety as $score = \mathbb{E}_{s\sim D, a\sim\pi(\cdot|s)} \exp(\mathcal{R}(s,a))$.

Table 3: Validation study of DASP term.

	Halfcheetah	Hopper	Walker2d
TDM w. safe action	0.61	0.44	0.42
TDM w. unsafe action	0.37	0.21	0.30
DASP w. safe action	0.64	0.47	0.44
DASP w. unsafe action	0.38	0.19	0.27

Table 3 shows the results. Comparing the results of the first and second rows, we observe that our safety score is sensitive to whether the consequences of actions are in-distribution (ID) or OOD, which supports the validity of this measurement. Analyzing the results from the third and fourth rows, we observe a notable score disparity in the density indicator between the two types of actions when utilizing the DASP model. This difference is similar to what we see in the first and second rows. It indicates that the DASP model performs well enough to differentiate between safe and unsafe actions.

Ablation study

The DASP weight α is the hyperparameter that control the magnitude of how the DASP term influence the training. Its influence to DASP is as shown in Table 4, where three agents are all trained on the 'meidum' datasets. From the results, we note that the best choice for α in this implementation is around 0.1 for the "halfcheetah" and "hopper" tasks, while for the "walker2d" task, the optimal α is 0.05. We utilized these parameters in our experiments to achieve the best performance across the different tasks.

Table 4: The ablation study results of α. We bold the highest scores in each task.

α	Ha.-m	Ho.-m	Wa.-m
0.01	66.0	104.8	101.6
0.05	67.8	105.1	**108.6**
0.1	**70.4**	**108.6**	100.0
0.5	66.8	105.1	104.7
3	65.0	104.3	102.5
10	64.8	103.2	97.6
100	51.6	100.8	85.6

The results suggest that while moderate values of α enhance performance by balancing conservatism and generalization, excessive values lead to instability and poorer decision-making.

Conclusion

In this paper, we propose a novel method called Density-Aware Safety Perception (DASP) to perform OOD state correction for a more robust and reliable offline reinforcement learning. To be specific, DASP is designed under a variational framework to achieve a more source-efficiency structure, which formulates the one-step forward dynamics model and the density model in a compact manner. Empirical results show that the proposed DASP outperforms most SOTA methods in offline RL, hence demonstrating the advantages of our method, which only uses an indicator instead of estimating specific distributions for OOD state correction.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (62476128), and Doctoral Visiting Program of Nanjing University of Aeronautics and Astronautics (241206DF16).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

An, G.; Moon, S.; Kim, J.; and Song, H. O. 2021. Uncertainty-Based Offline Reinforcement Learning with Diversified Q-Ensemble. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Bootkrajang, J.; and Kabán, A. 2012. Label-noise robust logistic regression and its applications. In *Joint European conference on machine learning and knowledge discovery in databases*, 143–158. Springer.

Burda, Y.; Grosse, R.; and Salakhutdinov, R. 2015. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Doersch, C. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

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

Fujimoto, S.; Meger, D.; and Precup, D. 2019. Off-Policy Deep Reinforcement Learning without Exploration. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 2052–2062. PMLR.

Garcıa, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1): 1437–1480.

Hyvärinen, A.; and Dayan, P. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

Janner, M.; Du, Y.; Tenenbaum, J. B.; and Levine, S. 2022. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*.

Jiang, K.; Yao, J.-Y.; and Tan, X. 2023. Recovering from Out-of-sample States via Inverse Dynamics in Offline Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jin, Y.; Yang, Z.; and Wang, Z. 2021. Is Pessimism Provably Efficient for Offline RL? In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 5084–5096. PMLR.

Kalashnikov, D.; Irpan, A.; Pastor, P.; Ibarz, J.; Herzog, A.; Jang, E.; Quillen, D.; Holly, E.; Kalakrishnan, M.; Vanhoucke, V.; and Levine, S. 2018. Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, 651–673. PMLR.

Kang, K.; Gradu, P.; Choi, J. J.; Janner, M.; Tomlin, C.; and Levine, S. 2022. Lyapunov density models: Constraining distribution shift in learning-based control. In *International Conference on Machine Learning*, 10708–10733. PMLR.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mao, Y.; Wang, Q.; Chen, C.; Qu, Y.; and Ji, X. 2024. Offline Reinforcement Learning with OOD State Correction and OOD Action Suppression. *CoRR*, abs/2410.19400.

Mao, Y.; Zhang, H.; Chen, C.; Xu, Y.; and Ji, X. 2023. Supported Value Regularization for Offline Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Peng, X. B.; Berseth, G.; Yin, K.; and Van De Panne, M. 2017. Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *Acm transactions on graphics (tog)*, 36(4): 1–13.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

Wang, D.; and Tan, X. 2014. Robust distance metric learning in the presence of label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

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

Wu, J.; Wu, H.; Qiu, Z.; Wang, J.; and Long, M. 2022. Supported Policy Optimization for Offline Reinforcement Learning. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wu, Y.; Tucker, G.; and Nachum, O. 2019. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Yang, R.; Bai, C.; Ma, X.; Wang, Z.; Zhang, C.; and Han, L. 2022. RORL: Robust Offline Reinforcement Learning via Conservative Smoothing. *CoRR*, abs/2206.02829.

Zhang, H.; Shao, J.; Jiang, Y.; He, S.; Zhang, G.; and Ji, X. 2022. State Deviation Correction for Offline Reinforcement Learning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, 9022–9030. AAAI Press.

Zhang, J.; Zhang, C.; Wang, W.; and Jing, B. 2023. Constrained Policy Optimization with Explicit Behavior Density For Offline Reinforcement Learning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhang, Z.; and Tan, X. 2024. An Implicit Trust Region Approach to Behavior Regularized Offline Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16944–16952.