Proceedings of Machine Learning Research vol 242:1–14, 2024

Policy-regularized Offline Safe Reinforcement Learning with Preference Aligned Sampling

Cheng Tang * C-TANG21@MAILS.TSINGHUA.EDU.CN
Tsinghua University

Abstract

Offline safe reinforcement learning (RL) aims to learn a safe and relatively rewarding policy with a precollected dataset. One prevalent method to deal with this problem is offline policy-regularized method, which typically incorporates a behavior cloning mechanism into the policy learning to regularize the learned policy stay close enough to the behavior policy, hence mitigates the distribution shift challenge. However, this framework may suffer from suboptimality of behavior policy due to the imbalanced dataset. In this work, we propose DIAM (distribution aligned sampling), a preference aligned sampling method customized for policy-regularized offline safe algorithms. Comprehensive evaluation in various tasks illustrates the ability of DIAM in optimizing the behavior policy, hence benefits policy-regularized offline safe algorithms. DIAM shows superiority compared to other model-centric method and data-centric method, making it more applicable and universal, even with simple structure.

Keywords: Safe Reinforcement Learning, Policy-regularized Offline RL, Data-centric method

1. Introduction

Offline Reinforcement Learning (RL) focuses on learning high-reward policies from pre-existing datasets, a topic that has gained significant attention and shown great promise across various applications (Chen et al., 2021; Levine et al., 2020). This approach aims to leverage as much information as possible from collected trajectories while avoiding distribution shift (Fu et al., 2020; Kostrikov et al., 2021). Despite its success, real-world tasks often require more than simply maximizing a scalar reward function due to numerous constraints that limit feasible solutions (Gulcehre et al., 2020). Ensuring safety and satisfying constraints is particularly crucial for deploying RL algorithms in practical scenarios (Kim et al., 2022; Lu et al., 2023; Chen et al., 2023), such as autonomous driving (Sun et al., 2020; Lu et al., 2023) and robotics (Zhao et al., 2023; Ding et al., 2024).

Offline Safe RL, which aims to learn a relative rewarding policy within a constrained manifold (García and Fernández, 2015; Brunke et al., 2022), has shown its strength in achieving safe and robust objectives in applications (Gu et al., 2022). Several frameworks and techniques have been proposed to deal with offline safe RL problems, including stationary DIstribution CorrEction (DICE) family methods (Lee et al., 2022) which trains the policy by sampling state-action pairs with importance, and sequential based method which adapts the trained policy to different constraint thresholds (Liu et al., 2023b; Lin et al., 2023).

Policy-regularized offline safe RL is another prevalent framework to deal with offline safe RL, which typically integrates behavior cloning into policy learning to ensure the learned policy remains close to the behavior policy with a policy regularizer to manage shift (Fujimoto et al., 2019; Kumar et al., 2019a; Xu et al., 2022). However, Recent Policy-regularized offline safe RL may fail catastrophically if the behavior policy of the dataset is suboptimal: either too conservative or unfeasible due to the

* This work was done when doing summer research in safeai lab in Carnegie Mellon University

violation of the safety constraint. Recent works (Hong et al., 2023a,b) show the ability of dataset reweighting in offline RL, but as far as we know, the ability of such method under safe RL remains unknown.

In this paper, we focus on data-centric method in improving the performance of offline policy-regularizd safe RL algorithms. We propose DIAM (distribution aligned sampling), which computes the sampling weights of the dataset by solving an optimization oracle, specially designed for offline safe RL. We summarize the contribution of our contribution as follows:

- We study the suboptimality of behavior policy derived from the dataset. Our results show that suboptimality of behavior policy causes failure of Policy-regularizer in offline safe algorithms, hence affects algorithm performance.
- We propose DIAM, which is a data-centric method compatible with extensive policy regularized safe offline algorithms and can be computed without additional computation cost.
- We conduct comprehensive experiment to illustrate the superiority of DIAM. We compare DIAM with both model-centric baselines and data-centric baselines under varying tasks and datasets.

2. Related Work

Safe RL. Safe RL is often introduced as a constrained optimization problem, which focuses on maximizing the reward within the cost threshold (Garcıa and Fernández, 2015; Achiam et al., 2017; Zhang et al., 2020; Kim et al., 2024). One prevalent problem setting is online training, where the agent is able to interact with th nominal environment directly (Chow et al., 2017; Tessler et al., 2018; Wu et al., 2024). In order to learn a safe and relatively rewarding policy, Lagrangian-based methods apply a multiplier to penalize violations on constraint (Stooke et al., 2020; Chow et al., 2017), while variational inference based methods estimate optimal penalty multiplier directly (Liu et al., 2022a; Huang et al., 2022). Another prevalent problem setting is offline training, which focuses on training on a fixed dataset without interaction with the environment (Ernst et al., 2005). Besides policy-regularized algorithms, stationary DIstribution CorrEction (DICE)-style methods (Lee et al., 2022) which trains the policy by importance sampling and sequential modeling methods (Liu et al., 2023b; Guo et al., 2024) are two widely used approaches.

Policy-regularized Offline Safe RL. A significant challenge in offline RL is managing the distribution shift between state-action pairs in the dataset and those from the learned policy (Levine et al., 2020). Policy regularization (Fujimoto et al., 2019; Kumar et al., 2019a; Xu et al., 2022) emerges as a straightforward and effective solution. This technique integrates behavior cloning within policy learning to ensure the learned policy remains close to the behavior policy. BCQ (Fujimoto et al., 2019) uses a conditional variational auto-encoder to model behavior policies and develops a perturbation model for bounded action adjustments. BEAR (Kumar et al., 2019b) employs maximum mean discrepancy to regularize the policy, estimated through multiple samples from both the learned and behavior policies. CPQ (Xu et al., 2022) estimates Q-function in a conservative perspective with policy regularizer managing distribution shift.

Dataset Reweighting. Dataset reweighting is a possible approach to formulate different dataset distribution which assigns each data point with customized weight when sampling. Hong et al. (2023a) first proposed to sample through applying more weights on more rewarding trajectory-wise data point, with an entropy regularization term reducing variance. Hong et al. (2023b) further assigned

weights by solving optimization oracle on each (s, a)-pairs to obtain a more sophisticated sampling method design. Yao et al. (2024a) first applied reweighting method in offline safe RL setting.

3. Problem Formulation

In this section, we introduce the framework of Offline safe reinforcement learning with regularization.

3.1. Constrained Markov Decision Process

The Constrained Markov Decision Process (CMDP) $\mathcal{M}$ is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathbf{c}, \mu_0)$ (Altman, 1998), where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, and $\mu_0 : \mathcal{S} \to [0, 1]$ is the initial state distribution. Compared to traditional MDPs, CMDPs consider MDP with an additional element $\mathbf{c} = \{c_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}, i = 1, 2, 3...N\}$ to capture the violation cost through the constrains, where N is the cost dimension. A safe RL problem with multi-constrains is specified by a CMDP and constraint threshold $\kappa_i \in \mathbb{R}_{\geq 0}$. Let $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denote the policy and $\tau = \{s_1, a_1, ...\}$ denote the trajectory. The trajectory-wise reward returns and cost returns are defined as $R(\tau) = \sum_\tau r$, and the cost returns $C_i(\tau) = \sum_\tau c_i, i = 1, 2, ...N$. which is the expectation of discounted return under the policy π and the initial state distribution μ_0. The goal of the safe RL problem with multi-constrains is to find the policy that maximizes the reward return while constraining the cost return under the pre-defined threshold κ_i:

$$\pi^* = \arg\max_\pi \mathbb{E}_{\tau\sim\pi}[R(\tau)], \quad \text{s.t.} \quad \mathbb{E}_{\tau\sim\pi}[C_i(\tau)] \leq \kappa_i, i = 1, 2, 3...N \tag{1}$$

3.2. Offline Safe RL with Regularization

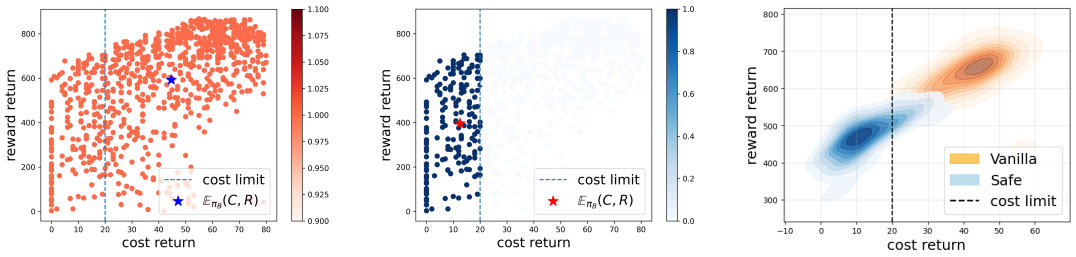
In this report, we focus on the offline safe RL with multi-constrain, where the agent can only access to a precollected dataset $\mathcal{D} = \{\tau_1, ...\tau_N\}$, where $\mathcal{D}$ is collected by a behavior policy π_B. Due to the **Distribution shift** (Kostrikov et al., 2021) challenge in offline RL, which refers to the poor generalization ability of the agent when facing the OOD situations (Lin et al., 2024), we often introduce a regularization penalty term to alleviate the OOD issue (Yao et al., 2024a), that is, the constrained safe RL problem is converted to:

$$\pi_r^* = \arg\max_\pi \mathbb{E}_{\tau\sim\pi}[R(\tau)] - wL(\pi, \pi_B), \quad \text{s.t.} \quad \mathbb{E}_{\tau\sim\pi}[C_i(\tau)] \leq \kappa_i, i = 1, 2, 3...N, \tag{2}$$

where $w > 0$ is a constant weight, $L(\pi, \pi_B)$ is a regularization distance. In practice, the regularization is often chosen as MSE or evidence lower bound regularization (Yao et al., 2024a). In order to solve the problem in Eq.(2), the most common way is to convert the multi-constrain optimization problem into dual form by the lagrange method (Liu et al., 2023b):

$$(\pi_r^*, \boldsymbol{\lambda}^*) = \arg\max_{\boldsymbol{\lambda}} \min_\pi \sum_{i=1}^{N} \lambda_i \left(\mathbb{E}_{\tau\sim\pi} C_i(\tau) - \kappa_i\right) - \mathbb{E}_{\tau\sim\pi} R(\tau) + wL(\pi, \pi_B) \tag{3}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ..., \lambda_N]^T$ is the Lagrangian multiplier corresponding to the primary problem (3). In practice, we can update $(\pi, \boldsymbol{\lambda})$ iteratively (Stooke et al., 2020).



(*a*) Unfeasible (*b*) Conservative (*c*) Behavior policy π_B

Figure 1: (a) example of unfeasible dataset. (b) example of conservative dataset. (c) behavior policy obtained from unfeasible dataset and conservative dataset.

4. Distribution aligned sampling for Offline Safe Reinforcement Learning

In this section, we first state the suboptimality of the behavior policy π_B, which shows either unfeasible or conservative property. Then we focus on how the behavior policy π_B affects policy regularized safe offline RL algorithms, from which we identify that the policy regularizer will contribute most to the conservative or unfeasible of the learned policy. At last we propose our sampling method, Distribution Aligned Sampling (DIAM) to ease the suboptimality of behavior policy π_B thus be beneficial to the policy-regularized safe offline RL.

4.1. Problem Statement: suboptimality of π_B

In this part, we first formulate the problem of the suboptimality of the behavior policy π_B, then we present our distribution aligned sampling method to deal with this problem. In problem formulation, in order to avoid OOD issue, we choose to solve the regularized form of constrained safe RL problem (2). This regularization term depends on the behavior policy π_B, hence the performance gap between the regularized optimal policy π_r^* and the optimal policy π^* is bounded as:

$$\mathbb{E}_{\tau\sim\pi^*}[R(\tau)] - \mathbb{E}_{\tau\sim\pi_r^*}[R(\tau)] \leq wL(\pi^*, \pi_B). \tag{4}$$

The proof is given in Appendix. Hence the suboptimality property of recent method depends highly on the suboptimality of the behavior policy π_B. Many collections of dataset $\mathcal{D}$ depend on the online RL algorithms (Liu et al., 2023a), which encourage exploration and learn from mistakes, hence the dataset collected by the algorithms may show suboptimality of π_B. In practice, conditioned on a specific threshold κ, the behavior policy π_B shows either **unfeasible** or **conservative** property. Intuitively, unfeasible policy is more rewarding but violates cost constrains (Liu et al., 2022b), while conservative policy obeys the safety constrains well but is less rewarding. Figure 1 illustrates typical examples of unfeasible dataset and conservative dataset. Due to the exploration practice in the online RL, if sampled evenly, the behavior policy of the collected dataset by online RL may be unfeasible. On the other hand, the behavior policy of the safe dataset may be conservative due to the insufficient exploration in the safe trajectories. Hence, given a dataset $\mathcal{D}$ and cost threshold κ, the most common case is that the induced behavior policy π_B of the dataset $\mathcal{D}$ is sub-optimal and will cause the learned policy π to be either conservative or unfeasible.

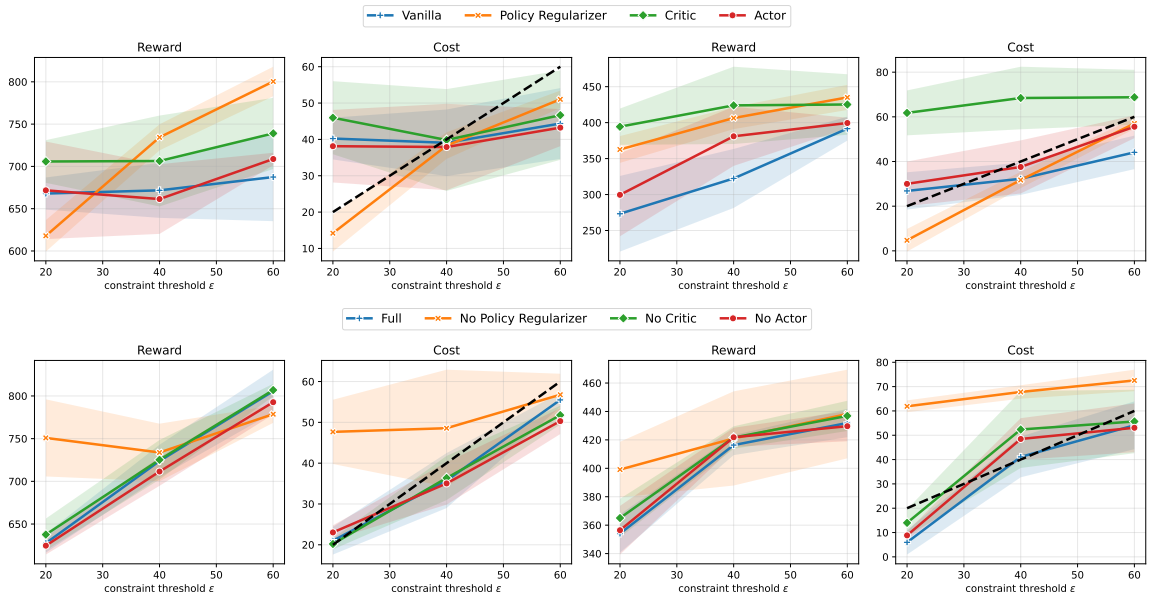


Figure 2: experiment on the function of suboptimal behavior policy π_B. The tasks are chosen as Ball Circle (left) and Car Circle (right). Vanilla: raw dataset. Policy Regularizer: only policy regularizer access to expert dataset. No Policy Regularizer: only policy regularizer access to raw suboptimal dataset.

4.2. How does suboptimality of behavior policy affect policy-regularized safe offline RL algorithms?

Policy-regularized safe offline RL algorithms mainly obtain three parts: **Policy regularizer**, **Critic** and **Actor**. **Policy regularizer** is mainly applied to encourage the learned policy close enough to behavior policy π_B. **Critic** acts as estimated Q-function in practice, which encourages the update of learned policy. **Actor** is a framework that generate action guided by the **Policy regularizer** and **Critic**. To illustrate how the unfeasible dataset and conservative dataset affect the performance of policy regularized safe offline RL algorithms, we conduct experiment shown in Figure 2, which indicates that sub-optimal dataset mainly influences the **Policy regularizer** hence affect the performance of policy-regularized safe offline RL algorithms. This is because that the regularizer pushes the learned policy to be close to the behavior policy. The **Critic** and **Actor** show little variation under different dataset with same support. Hence, in order to improve the performance of policy-regularized safe offline RL algorithms, one potential direction is to optimize the suboptimality of the policy regularizer caused by the unfeasible or conservative behavior policy π_B.

4.3. Distribution aligned sampling: optimize behavior policy

In this subsection, we propose to use distribution aligned sampling to optimize the behavior policy, which is to provide a sampling weight for each trajectory τ. Without changing the dataset support, the aligned sampling is equivalent to change the behavior policy.

Unbias estimation of $\mathbb{E}_{\tau\sim\pi_w}$. To start up, we first show how the resampling of trajectories emulates sampling transitions generated by an implicit behavior policy different from the behavior that

collected the dataset. According to the work in (Hong et al., 2023a), the newly induced weighted state-action distribution has the form of

$$d_W(s,a) = \sum_{i=1}^{N} w_i d_{\pi_B}(s)\pi_B(a|s)$$

where the weight w_i is the weight of τ_i, $W = \{w_0, w_1, ..., w_{N-1}\}$, and d_{π_B} is defined as state occupancy measure induced by the behavior policy π_B hence to train a better policy regularizer benefit regularized safe algorithms. The behavior policy of the weighted dataset $\mathcal{D}_W$ can be therefore formed as $\pi_W = d_W(s,a)/\sum_{i=1}^{N} w_i d_{\pi_B}(s)$. Followed by the (Hong et al., 2023a), with the assumption that the discount factor of $\mathcal{D}_W$ is less than 1, the trajectory-wise expected return and cost of the weighted behavior policy can be bounded by the Hoeffding's inequality (Serfling, 1974) as:

$$\mathbb{P}\left[\left|\mathbb{E}_{\tau\sim\pi_W}[R(\tau)] - \sum_{k=1}^{N} w_i R(\tau_k)\right| \geq \epsilon\right] \leq 2\exp\left(\frac{2\epsilon^2}{\delta_R^2 \sum_{k=1}^{N} w_k^2}\right),$$

where $\delta_R = \max R(\tau_k) - \min R(\tau_k)$ is the reward interval amplitude. Similarly, the cost of the weighted behavior policy can also be center-bound: for any $i \in \{1, 2, 3...N\}, \epsilon > 0$

$$\mathbb{P}\left[\left|\mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)] - \sum_{i=1}^{N} w_k C_i(\tau_k)\right| \geq \epsilon\right] \leq 2\exp\left(\frac{2\epsilon^2}{\delta_{C_i}^2 \sum_{k=1}^{N} w_k^2}\right),$$

where $\delta_{C_i} = \max C_i(\tau_k) - \min C_i(\tau_k)$. With the concentration inequality above, we have the unbiased estimation of $\mathbb{E}_{\tau\sim\pi_W}[R(\tau)]$ and $\mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)]$ as $\sum_{k=1}^{N} w_i R(\tau_k), \sum_{i=1}^{N} w_k C_i(\tau_k)$ respectively.

Optimize Weighted Behavior Policy π_W. We start with a conservative policy π induced dataset $\mathcal{D}$, i.e, $\mathbb{E}_{\tau\sim\pi}[C_i(\tau)] \leq \kappa_i, \forall i$. In practice, this dataset can be achieved by filtering unsafe trajectory to obtain. We formalize our goal as to maximize the following term:

$$\max_W \mathbb{E}_{\tau\sim\pi_W}[R(\tau)], \quad \text{s.t.} \quad \mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)] \leq \kappa_i, i = 1, 2, 3...N.$$

By the unbias estimation of $\mathbb{E}_{\tau\sim\pi_w}$, the maximization problem $\max_W \mathbb{E}_{\tau\sim\pi_W}[R(\tau)]$ can be transformed into

$$\max_W \sum_{k=1}^{N} w_i R(\tau_k) + \beta H(W),$$

where $H(\cdot)$ is information entropy that acts as a regularization, i.e. $H(W) = -\sum w_i \log w_i$, and β is a temperature parameter to control the regularization. Now it remains to formulate the safety constraint. As safety is the top priority in offline safe reinforcement learning, we hope the weighted sampled dataset can achieve at least as conservative as the conservative policy π, that is, we formalize the safety constrain as:

$$\frac{\kappa_i - \mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)]}{\sqrt{\text{Var}_{\tau\sim\pi_W}[C_i(\tau)]}} \geq \frac{\kappa_i - \mathbb{E}_{\tau\sim\pi}[C_i(\tau)]}{\sqrt{\text{Var}_{\tau\sim\pi}[C_i(\tau)]}}, \forall i \in [N].$$

The threshold set above garrantees the optimization problem solvable in practice as the $\mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)]$ can be estimated through $\sum_{i=1}^{N} w_k C_i(\tau_k)$. This sampling weight can be directly computed by optimization tool box without loss of precision.

5. Experiment

In the experiment section, we aim to answer the following questions: (1) how does DIAM perform compared to other offline safe RL baselines? (2) how does DIAM compare to baseline methods? (3) How robust is DIAM in terms of the hyperparameters? To validate our method, we set the following experiments for evaluation.

Tasks and Datasets. We adopt robot locomotion control tasks and datasets in the public benchmark `Bullet-Safety-Gym` (Gronauer, 2022), `Safety-gymnasium` (Ji et al., 2023), and public offline RL dataset `DSRL` (Liu et al., 2023a) for evaluation. We consider two tasks (`Circle` and `Run`) with three types of agent, `Ball`, `Car` and `Drone`. These experiments are commonly used in previous works (Liu et al., 2023b; Zhang et al., 2020).

Baselines. We compare our method with two categories of regularized algorithms baselines: model-centric approaches and data-centric approaches.

• **Model-centric approach:** (1) Imitation Learning: `BC` (Behavior cloning), `BC-safe` (Behavior cloning with safe dataset) (Xu et al., 2022); (2) Q-Learning based methods: `BCQ-Lag` (Fujimoto et al., 2019), `BEAR-Lag` (Kumar et al., 2019a), and `CPQ` (Xu et al., 2022); (3) Distribution Correction Estimation: `COptiDICE` (Lee et al., 2021). (4) Sequential modeling: CDT (Liu et al., 2023b).

• **Data-centric approach:** (1) `Vanilla-BCQ-Lag`, (2) `Safe-BCQ-Lag`, (3) `Safe-10-BCQ-Lag`, and (4) `Top-10-BCQ-Lag`. `Safe-BCQ-Lag` represents sampling with solely safe data, i.e. $w_k = 0$ if there exists $i \in [N], C_i(\tau_k) \leq \kappa_i$. `Top-10-BCQ-Lag` represents sampling with data that obtains the top 10% reward. `Safe-10-BCQ-Lag` represents sampling with safe data that obtains the top 10% reward.

Metrics. We adopt the normalized reward return and normalized cost as evaluation metrics, which is based on previous offline safe RL works (Yao et al., 2024a). The normalized reward and normalized cost are defined as:

$$R_{\text{normalized}} = R_\pi / r_{\max}(\mathcal{D}), C_{\text{normalized}} = C_\pi / \kappa,$$

where R_π, C_π are the reward return and cost return respectively, $r_{\max}(\mathcal{D})$ is the maximum theoretical return of certain task given dataset $\mathcal{D}$. Each evaluation is done over 3 seeds and taken averaged results and standard deviations in order to avoid avoid randomness.

5.1. How does DIAM perform compared to other offline safe RL baselines

We compare DIAM on two types of baselines, Model-centric approach and Data-centric approach.

Model-centric approach. The comparison results on model-centric are presented in Table.1 with cost threshold $\kappa = 20$. The results of BC and BC-Safe illustrate the suboptimality of the behavior policy π_B, which is either tempting or conservative with whole dataset and safe dataset respectively. The results of BCQ-Lag and BEAR-Lag further prove the importance of behavior policy π_B in regularized Q-learning based algorithms. The CPQ shows notable reward degradation across all tested tasks due to the over-conservative behaviors learned from pessimistic estimation methods, while CDT learns a rather tempting policy in many tasks, violating the top priority of considering the safety-constrain. Compared to those algorithms, DIAM is able to balance well the demand for safety and reward, learning a safe and relatively rewarding behavior.

Data-centric approach. We then conduct experiment to compare DIAM with Data-centric approach. The Data-centric approach is based on the regularized Q-based algorithm BCQ-Lag. we also set up two additional constrains, **High Velocity Constrain** and **Low Velocity Constrain** (Yao et al., 2024b),

Algorithm	Stats	BallCircle	CarCircle	DroneCircle	BallRun	CarRun	DroneRun
BC	reward ↑	0.76 ± 0.03	0.50 ± 0.04	0.84 ± 0.05	0.73 ± 0.04	**0.98 ± 0.00**	0.42 ± 0.18
	cost ↓	2.24 ± 0.17	2.66 ± 0.88	3.19 ± 0.56	3.02 ± 0.16	**0.72 ± 0.78**	1.33 ± 1.23
BC-Safe	reward ↑	0.55 ± 0.03	0.35 ± 0.11	**0.62 ± 0.01**	**0.20 ± 0.01**	**0.98 ± 0.00**	**0.57 ± 0.02**
	cost ↓	1.08 ± 0.27	1.01 ± 0.28	**0.42 ± 0.17**	**0.89 ± 0.17**	**0.01 ± 0.00**	**0.27 ± 0.38**
BCQ-Lag	reward ↑	0.71 ± 0.04	0.59 ± 0.03	0.57 ± 0.87	0.04 ± 0.09	**0.91 ± 0.04**	0.67 ± 0.05
	cost ↓	1.96 ± 0.26	1.78 ± 0.14	3.57 ± 0.49	1.97 ± 1.50	**0.00 ± 0.00**	5.28 ± 0.31
BEAR-Lag	reward ↑	0.80 ± 0.04	0.85 ± 0.05	0.87 ± 0.03	0.56 ± 0.42	0.62 ± 0.29	0.16 ± 0.14
	cost ↓	2.49 ± 0.26	3.08 ± 0.88	3.61 ± 0.26	3.08 ± 2.18	3.62 ± 2.85	4.11 ± 2.42
CPQ	reward ↑	**0.62 ± 0.03**	**0.65 ± 0.03**	**0.01 ± 0.02**	0.37 ± 0.03	0.98 ± 0.01	0.34 ± 0.01
	cost ↓	**0.78 ± 0.14**	**0.45 ± 0.63**	**0.64 ± 0.36**	3.08 ± 0.66	1.69 ± 1.57	1.08 ± 0.79
COptiDICE	reward ↑	0.72 ± 0.01	0.44 ± 0.05	**0.42 ± 0.01**	0.63 ± 0.03	**0.95 ± 0.01**	0.67 ± 0.03
	cost ↓	2.30 ± 0.11	3.45 ± 0.56	**0.85 ± 0.34**	3.12 ± 0.15	**0.00 ± 0.00**	3.75 ± 0.16
CDT	reward ↑	0.70 ± 0.00	**0.72 ± 0.02**	0.64 ± 0.00	**0.30 ± 0.00**	1.00 ± 0.00	**0.60 ± 0.00**
	cost ↓	1.06 ± 0.04	**0.82 ± 0.01**	1.07 ± 0.04	**0.28 ± 0.40**	1.01 ± 0.20	**0.33 ± 0.13**
DIAM	reward ↑	**0.71 ± 0.01**	**0.66 ± 0.02**	**0.64 ± 0.01**	**0.30 ± 0.11**	**0.96 ± 0.02**	**0.34 ± 0.04**
	cost ↓	**0.98 ± 0.12**	**0.30 ± 0.25**	**0.45 ± 0.17**	**0.57 ± 0.46**	**0.45 ± 0.63**	**0.69 ± 0.55**

Table 1: Evaluation of results through single constraint. The cost threshold is 1. Gray: unsafe agent. **Bold**: Safe agent with all cost threshold less than 1. **Safe**: Safe agent with all cost threshold less than 1.

to make our problem more challenging and practical. The high velocity constrain induces cost when the agent exceed the upper velocity limit, while the low velocity constrain induces cost when the agent falls below the lower velocity limit. The costs are all binary. The tasks setting is `Circle`, with various robots (`Ball, Car, Drone`). The results are shown in Table.2. When the cost constrain is more complex and challenging, behavior policy is more difficult to stay within safe boundaries because the feasible set for policy is shrinkage. Our results can be found in Table.2. The Vanilla and Top-10 data show that tempting dataset will induce unsafe policy in regularized Q-learning based algorithm. The violation of cost constraint in Safe-Top-10 group illustrates that filtering sub-optimal data intuitively will reduce the dataset support, hence induce insufficient coverage of $(s, a)-$ pairs in the environment producing greater variance. The Safe dataset, on the contrary, will induce a relative conservative policy with low reward, and due to the fact that some constrains are positively related to the reward, the learned policy can be rather tempting in terms of some safety constrains. DIAM mitigates those drawbacks by sampling with certain weights and achieves high reward without violating the safety constraint, which shows strength in those cases.

5.2. How does DIAM benefit policy-regularized safe reinforcement learning method?

In this part, we conduct extensive experiment to illustrate the benefit of DIAM. Apart from ablation study in Figure.2, further experiment is conducted to evaluate the function of DIAM. We focus on the comparison with Data-centric approach. Figure.2 illustrates that the suboptimality of behavior policy π_B can induce the policy regularizer to learn a either unfeasible or conservative policy. To step further, as shown in Figure.3, training with whole dataset or top-10% dataset will learn a feasible policy as expected, while safe dataset will induce a relative conservative policy. Safe-top-10 dataset cannot learn a ideal policy in some cases, but due to the limited dataset support, it shows high variance and tempting property, which indicates that optimizing behavior policy π_B is non-trivial and

Tasks	Stats	Safe-Top-10	Safe	Vanilla	Top-10	DIAM
BallCircle	reward ↑	0.85 ± 0.02	0.68 ± 0.04	0.75± 0.03	0.89 ± 0.01	**0.77 ± 0.02**
	cost ↓	1.06 ± 0.03	0.48 ± 0.06	1.08 ± 0.14	1.35 ± 0.04	**0.96 ± 0.04**
	high vel cost ↓	0.84 ± 0.11	1.18 ± 0.27	1.05 ± 0.12	0.77 ± 0.30	**0.88 ± 0.07**
	low vel cost ↓	1.17 ± 0.02	0.92 ± 0.05	0.95 ± 0.04	1.06 ± 0.01	**0.84 ± 0.04**
CarCircle	reward↑	0.73 ± 0.02	**0.56 ± 0.04**	0.92± 0.03	0.79 ± 0.02	**0.67 ± 0.01**
	cost ↓	1.82 ± 0.59	**0.39 ± 0.09**	2.35 ± 0.04	1.29 ± 0.34	**0.91 ± 0.16**
	high vel cost ↓	1.04 ± 0.63	**0.61 ± 0.23**	0.25 ± 1.12	0.76 ± 0.48	**0.89 ± 0.13**
	low vel cost ↓	0.85 ± 0.01	**0.87 ± 0.03**	0.91 ± 0.02	0.96 ± 0.06	**0.86 ± 0.03**
DroneCircle	reward ↑	0.57 ± 0.03	**0.63 ± 0.01**	0.71 ± 0.02	0.72 ± 0.09	**0.63 ± 0.01**
	cost ↓	1.38 ± 0.40	**0.62 ± 0.34**	1.38 ± 0.38	3.84 ± 0.22	**0.50 ± 0.38**
	high vel cost ↓	0.94 ± 0.43	**0.45 ± 0.33**	0.86 ± 0.50	1.07 ± 1.49	**0.46 ± 0.33**
	low vel cost ↓	0.53 ± 0.04	**0.49 ± 0.01**	0.48 ± 0.03	0.54 ± 0.05	**0.48 ± 0.04**

Table 2: Evaluation of results through multi-constraint. The cost threshold is 1. Gray: unsafe agent. **Bold**: Safe agent with all cost threshold less than 1. **Safe**: Safe agent with all cost threshold less than 1.

requires sophisticated design. In contrast, DIAM shows its strength to obtain a feasible and relatively rewarding behavior policy that is able to guide policy-regularized safe RL algorithms.

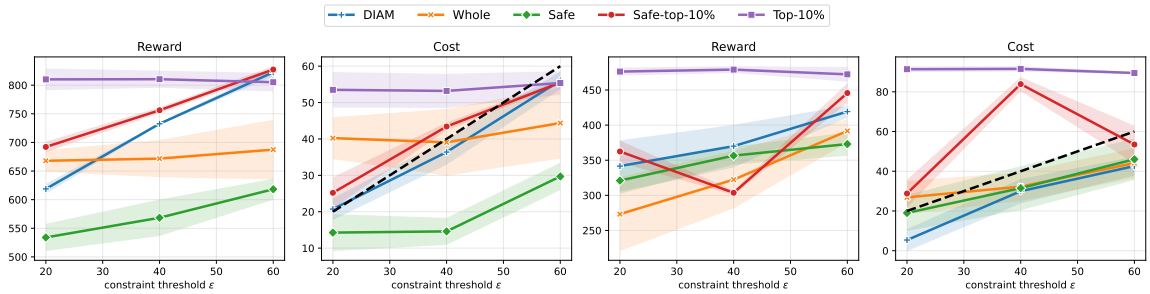


Figure 3: Comparison between Data-centric approach.

5.3. How does DIAM applied to different policy-regularized offline safe RL algorithms?

In this subsection, we aim to show the compatibility of DIAM with policy-regularized offline safe RL algorithms. The result of our experiment is shown in Figure.4. The original BC shows highly tempting property due to the suboptimality of the behavior policy, while DIAM shows ability to form a safe and rewarding behavior with raw dataset. The BCQ-Lag and BEAR-Lag are both compatible with DIAM as policy-regularized offline safe RL algorithms

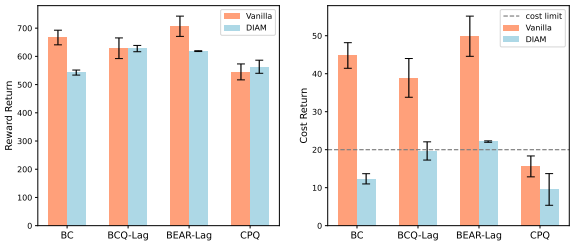


Figure 4: Results of comparison between under different offline policy-regularized RL algorithms.

are easily affected by behavior policy. Notably,
due to the pessimistic estimation methods inducing conservative policy, CPQ will be more conservative after DIAM hence show little improvement of the performance.

6. Conclusion

In this paper, we focus on improving the performance of policy-regularized offline safe algorithms in comprehensive offline safe-RL tasks. We first identify the suboptimality of behavior policy π_B, influencing the policy-regularized offline safe algorithms by forming unfeasible or conservative policy in policy regularizer, which is critical in policy-regularized method. To deal with this issue, we propose DIAM, a preference aligned sampling method customized for policy-regularized offline safe algorithms. We conduct extensive experiment to illustrate the superiority of DIAM compared to both model-centric baselines and data-centric baselines. We also show the ability of DIAM in mitigating the suboptimality of behavior policy and the compatibility of DIAM with different policy-regularized offline safe RL algorithms.
Admittedly, our work has at least two limitations. First, our method lacks theoretical guarantees due to the absence of a unified theoretical analysis framework on the offline RL algorithms with data-centric method. Second, sampling method cannot change the support of dataset, which means that our method probably cannot improve the performance of certain offline RL algorithms without sufficient coverage of state-action pairs.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

Eitan Altman. Constrained markov decision processes with total cost criteria: Lagrangian approach and dual linear program. *Mathematical methods of operations research*, 48(3):387–417, 1998.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

Baiming Chen, Zuxin Liu, Jiacheng Zhu, Mengdi Xu, Wenhao Ding, Liang Li, and Ding Zhao. Context-aware safe reinforcement learning for non-stationary environments. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10689–10695. IEEE, 2021.

Zhaorun Chen, Binhao Chen, Tairan He, Liang Gong, and Chengliang Liu. Progressive adaptive chance-constrained safeguards for reinforcement learning. *arXiv preprint arXiv:2310.03379*, 2023.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems*, 36, 2024.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

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

Sven Gronauer. Bullet-safety-gym: Aframework for constrained reinforcement learning. 2022.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Thomas Paine, Sergio Gómez, Konrad Zolna, Rishabh Agarwal, Josh S Merel, Daniel J Mankowitz, Cosmin Paduraru, et al. Rl unplugged: A suite of benchmarks for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:7248–7259, 2020.

Zijian Guo, Weichao Zhou, and Wenchao Li. Temporal logic specification-conditioned decision transformer for offline safe reinforcement learning. *arXiv preprint arXiv:2402.17217*, 2024.

Zhang-Wei Hong, Pulkit Agrawal, Rémi Tachet des Combes, and Romain Laroche. Harnessing mixed offline reinforcement learning datasets via trajectory weighting. *arXiv preprint arXiv:2306.13085*, 2023a.

Zhang-Wei Hong, Aviral Kumar, Sathwik Karnik, Abhishek Bhandwaldar, Akash Srivastava, Joni Pajarinen, Romain Laroche, Abhishek Gupta, and Pulkit Agrawal. Beyond uniform sampling: Offline reinforcement learning with imbalanced datasets. *Advances in Neural Information Processing Systems*, 36:4985–5009, 2023b.

Sandy Huang, Abbas Abdolmaleki, Giulia Vezzani, Philemon Brakel, Daniel J Mankowitz, Michael Neunert, Steven Bohez, Yuval Tassa, Nicolas Heess, Martin Riedmiller, et al. A constrained multi-objective reinforcement learning framework. In *Conference on Robot Learning*, pages 883–893. PMLR, 2022.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36, 2023.

Dohyeong Kim, Yunho Kim, Kyungjae Lee, and Songhwai Oh. Safety guided policy optimization. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2462–2467. IEEE, 2022.

Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Scale-invariant gradient aggregation for constrained multi-objective reinforcement learning. *arXiv preprint arXiv:2403.00282*, 2024.

Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pages 5774–5783. PMLR, 2021.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems*, 32, 2019a.

Aviral Kumar, Xue Bin Peng, and Sergey Levine. Reward-conditioned policies. *arXiv preprint arXiv:1912.13465*, 2019b.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021.

Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. Coptidice: Offline constrained reinforcement learning via stationary distribution correction estimation. *arXiv preprint arXiv:2204.08957*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Qian Lin, Bo Tang, Zifan Wu, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Safe offline reinforcement learning with real-time budget constraints. In *International Conference on Machine Learning*, pages 21127–21152. PMLR, 2023.

Qian Lin, Chao Yu, Zongkai Liu, and Zifan Wu. Policy-regularized offline multi-objective reinforcement learning. *arXiv preprint arXiv:2401.02244*, 2024.

Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, pages 13644–13668. PMLR, 2022a.

Zuxin Liu, Zijian Guo, Zhepeng Cen, Huan Zhang, Jie Tan, Bo Li, and Ding Zhao. On the robustness of safe reinforcement learning under observational perturbations. *arXiv preprint arXiv:2205.14691*, 2022b.

Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303*, 2023a.

Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. *arXiv preprint arXiv:2302.07351*, 2023b.

Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp, Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560. IEEE, 2023.

Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

Zifan Wu, Bo Tang, Qian Lin, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Off-policy primal-dual safe reinforcement learning. *arXiv preprint arXiv:2401.14758*, 2024.

Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized q-learning for safe offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8753–8760, 2022.

Yihang Yao, Zhepeng Cen, Wenhao Ding, Haohong Lin, Shiqi Liu, Tingnan Zhang, Wenhao Yu, and Ding Zhao. Oasis: Conditional distribution shaping for offline safe reinforcement learning. *arXiv preprint arXiv:2407.14653*, 2024a.

Yihang Yao, Zuxin Liu, Zhepeng Cen, Peide Huang, Tingnan Zhang, Wenhao Yu, and Ding Zhao. Gradient shaping for multi-constraint safe reinforcement learning. In *6th Annual Learning for Dynamics & Control Conference*, pages 25–39. PMLR, 2024b.

Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 2020.

Weiye Zhao, Rui Chen, Yifan Sun, Ruixuan Liu, Tianhao Wei, and Changliu Liu. Guard: A safe reinforcement learning benchmark. *arXiv preprint arXiv:2305.13681*, 2023.

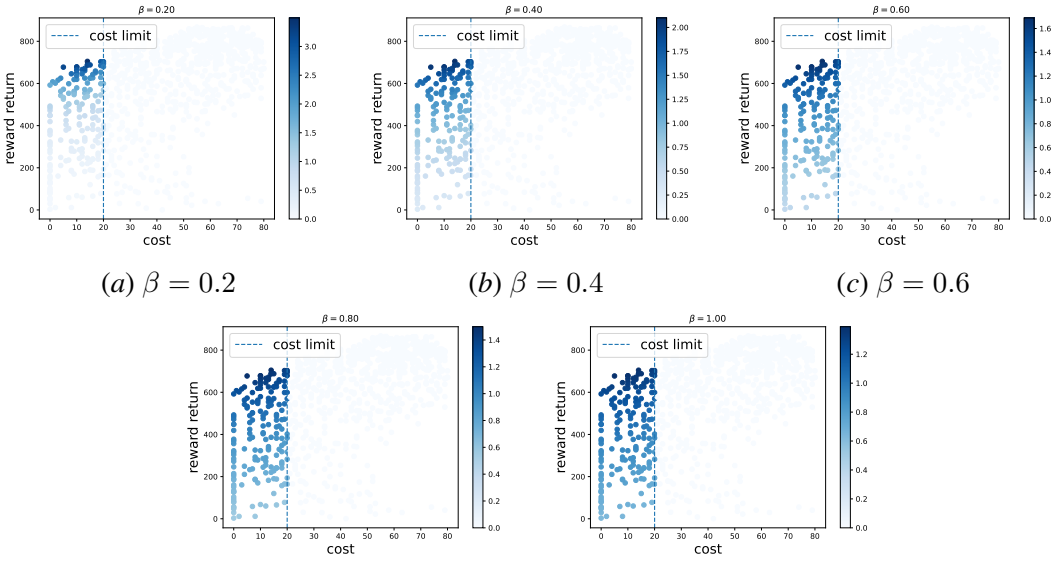
Appendix A. Proof of Eq. (4)

Proof Recall the definition of π^* and π_r^* in equation 1 and equation 2 respectively, we have

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\pi^*}[R(\tau)] - \mathbb{E}_{\tau\sim\pi_r^*}[R(\tau)] \\
&= (\mathbb{E}_{\tau\sim\pi^*}[R(\tau)] - wL(\pi^*, \pi_B)) - (\mathbb{E}_{\tau\sim\pi_r^*}[R(\tau)] - wL(\pi_r^*, \pi_B)) + wL(\pi^*, \pi_B) - wL(\pi_r^*, \pi_B) \\
&\leq w(L(\pi^*, \pi_B) - L(\pi_r^*, \pi_B)) \\
&\leq wL(\pi^*, \pi_B),
\end{aligned}$$

hence we conclude the proof. ■

Appendix B. Illustration of how β forms the dataset sampling



(*a*) $\beta = 0.2$ (*b*) $\beta = 0.4$ (*c*) $\beta = 0.6$

(*d*) $\beta = 0.8$ (*e*) $\beta = 1.0$

Figure 5: Tasks under Ball Circle. When the temperature parameter $\beta \to 0$, the sampled dataset is more tempting and concentrated. When the $\beta \to \infty$, the sampling method is closer to uniform sampling.