ReDS: Offline Reinforcement Learning With Heteroskedastic Datasets via Support Constraints

Anikait Singh[1,*], **Aviral Kumar**[1,*], **Quan Vuong**[2], **Yevgen Chebotar**[2], **Sergey Levine**[1]
[1]UC Berkeley, [2]Google DeepMind (*Equal contribution)
asap7772@berkeley.edu

Abstract

Offline reinforcement learning (RL) learns policies entirely from static datasets. Practical applications of offline RL will inevitably require learning from datasets where the variability of demonstrated behaviors changes non-uniformly across the state space. For example, at a red light, nearly all human drivers behave similarly by stopping, but when merging onto a highway, some drivers merge quickly, efficiently, and safely, while many hesitate or merge dangerously. Both theoretically and empirically, we show that typical offline RL methods, which are based on distribution constraints fail to learn from data with such non-uniform variability, due to the requirement to stay close to the behavior policy **to the same extent** across the state space. Ideally, the learned policy should be free to choose **per state** how closely to follow the behavior policy to maximize long-term return, as long as the learned policy stays within the support of the behavior policy. To instantiate this principle, we reweight the data distribution in conservative Q-learning (CQL) to obtain an approximate support constraint formulation. The reweighted distribution is a mixture of the current policy and an additional policy trained to mine poor actions that are likely under the behavior policy. Our method, CQL (ReDS), is theoretically motivated, and improves performance across a wide range of offline RL problems in games, navigation, and pixel-based manipulation.

1 Introduction

Recent advances in offline RL [39, 36] hint at exciting possibilities in learning high-performing policies, entirely from offline datasets, without requiring dangerous [19] or expensive [25] active interaction. Analogously to the importance of data diversity in supervised learning [9], the practical benefits of offline RL depend heavily on the *coverage* of behavior in the offline datasets [35]. Intuitively, the dataset must illustrate the consequences of a diverse range of behaviors, so that an offline RL method can determine what behaviors lead to high returns, ideally returns that are significantly higher than the best single behavior in the dataset.

One easy option to attain this kind of coverage is to combine many realistic sources of data, but doing so can lead to the variety of demonstrated behaviors varying in highly non-uniform ways across the state space, i.e. the dataset is *heteroskedastic*. For example, a driving dataset might show very high variability in driving habits, with some drivers being timid and some more aggressive, but remain remarkably consistent in "critical" states (e.g., human drivers are extremely unlikely to swerve in an empty road or drive off a bridge). A good offline RL algorithm should combine the *best* parts of each behavior in the dataset – e.g., in the above example, the algorithm should produce a policy that is *as good as the best human in each situation*, which would be better than *any* human driver overall. At the same time, the learned policy should not attempt to extrapolate to novel actions in subset of the state space where the distribution of demonstrated behaviors is narrow (e.g., the algorithm should not attempt to drive off a bridge). How effectively can current offline RL methods selectively choose on a *per-state* basis how closely to stick to the behavior policy?

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Most existing methods [32, 33, 28, 27, 53, 17, 23] constrain the learned policy to stay close to the behavior policy with so-called "distribution constraints". Using a combination of empirical and theoretical evidence, we first show that distribution constraints are insufficient when the heteroskedasticity of the demonstrated behaviors varies non-uniformly across states, because the strength of the constraint is state-agnostic, and may be overly conservative at some states even when it is not conservative enough at other states. We also devise a measure of heteroskedasticity that enables us to determine if certain offline datasets would be challenging for distribution constraints.

Our second contribution is a simple observation: distribution constraints against a *reweighted* version of the behavior policy give rise to support constraints. That is, the return-maximization optimization process can freely choose per-state how much the learned policy should stay close to the behavior policy, so long as the learned policy remains within the data support. We show that it is convenient to instantiate this insight on top of conservative Q-learning (CQL) [33], a recent offline RL method. The new method, CQL (ReDS), changes minimally the form of regularization, design decisions employed by CQL and inherits existing hyper-parameter values. CQL (ReDS) attains better performance than recent distribution constraints methods on a variety of tasks with more heteroskedastic distributions.

2 Preliminaries

The goal in offline RL is find the optimal policy in a Markov decision process (MDP) specified by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r, \mu_0, \gamma)$. $\mathcal{S}, \mathcal{A}$ denote the state and action spaces. $T(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ and $r(\mathbf{s}, \mathbf{a})$ represent the dynamics and reward function. $\mu_0(s)$ denotes the initial state distribution. $\gamma \in (0, 1)$ denotes the discount factor. We wish to learn a policy that maximizes return, denoted by $J(\pi) := \frac{1}{1-\gamma}\mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t)\sim\pi}[\sum_t \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)]$. We must find this policy while only having access to an offline dataset of transitions collected using a behavior policy π_β, $\mathcal{D} = \{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')\}$.

Offline RL via distributional constraints. Most offline RL algorithms regularize the learned policy π from querying the target Q-function on unseen actions [17, 30], either implicitly or explicitly. For our theoretical analysis, we will abstract the behavior of distributional constraint offline RL algorithms into a generic formulation following Kumar et al. [33]. As shown in Equation 1, we consider the problem where we must maximize the return of the learned policy π (in the empirical MDP) $\widehat{J}(\pi)$, while also penalizing the divergence from π_β:

$$\max_{\pi} \; \mathbb{E}_{\mathbf{s}\sim\widehat{d}^{\pi}} \left[\widehat{J}(\pi) - \alpha D(\pi, \pi_\beta)(\mathbf{s}) \right], \tag{1}$$

where D denotes a divergence between the learned policy π and the behavior policy π_β at state $\mathbf{s}$.

Conservative Q-learning. [33] enforces the distributional constraint on the policy *implicitly*. To see why this is the case, consider the CQL objective, which consists of two terms:

$$\min_{\theta} \; \alpha \underbrace{\left(\mathbb{E}_{\mathbf{s}\sim\mathcal{D}, \mathbf{a}\sim\pi}\left[Q_\theta(\mathbf{s}, \mathbf{a})\right] - \mathbb{E}_{\mathbf{s}, \mathbf{a}\sim\mathcal{D}}\left[Q_\theta(\mathbf{s}, \mathbf{a})\right]\right)}_{\mathcal{R}(\theta)} + \frac{1}{2}\mathbb{E}_{\mathbf{s}, \mathbf{a}, \mathbf{s}'\sim\mathcal{D}}\left[\left(Q_\theta(\mathbf{s}, \mathbf{a}) - \mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a})\right)^2\right], \tag{2}$$

where $\mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a})$ is the Bellman backup operator applied to a delayed target Q-network, $\bar{Q}$: $\mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a}) := r(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{a}'\sim\pi(\mathbf{a}'|\mathbf{s}')}[\bar{Q}(\mathbf{s}', \mathbf{a}')]$. The second term (in blue) is the standard TD error [40, 18, 22]. The first term $\mathcal{R}(\theta)$ (in red) attempts to prevent overestimation in the Q-values for out-of-distribution (OOD) actions by minimizing the Q-values under a distribution $\mu(\mathbf{a}|\mathbf{s})$, which is automatically chosen to pick actions with high Q-values $Q_\theta(\mathbf{s}, \mathbf{a})$, and counterbalances by maximizing the Q-values of the actions in the dataset. Kumar et al. [33] show that Equation 2 gives rise to a pessimistic Q-function that modifies the optimal Q function by the ratios of densities, $\pi(\mathbf{a}|\mathbf{s})/\pi_\beta(\mathbf{a}|\mathbf{s})$ at a given state-action pair $(\mathbf{s}, \mathbf{a})$. Formally, the Q-function obtained after one iteration is given by:

$$Q_\theta(\mathbf{s}, \mathbf{a}) := \mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a}) - \alpha \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi_\beta(\mathbf{a}|\mathbf{s})} - 1\right]. \tag{3}$$

The Q function is unchanged only if the density of the learned policy π matches that of the behavior policy π_β. Otherwise, for state-action pairs where $\pi(\mathbf{a}|\mathbf{s}) < \pi_\beta(\mathbf{a}|\mathbf{s})$, Eq. 3 increases their Q values and encourages the policy π to assign more mass to the action. Vice versa, if $\pi(\mathbf{a}|\mathbf{s}) > \pi_\beta(\mathbf{a}|\mathbf{s})$, Eq. 3 encourages the policy π to assign smaller density to the action $\mathbf{a}$. In Eq. 3, α is a constant for every state, and hence the value function learned by CQL is altered by the ratio of action probabilities to the same extent at all possible state-action pairs. As we will discuss in the next section, this can be sub-optimal when the learnt policy should stay close to the behavior policy in some states, but not others. We elaborate on this intuition in the next section.

3 Why Distribution Constraints Fail with Heteroskedastic Data

In statistics, heteroskedasticity is typically used to refer to the condition when the standard deviation in a given random variable varies non-uniformly over time (see for example, Cao et al. [6]). We call a offline dataset heteroskedastic when the variability of the behavior differs in different regions of the state space: for instance, if for certain regions of the state space, the observed behaviors in the dataset assign the most probability mass to a few actions, but in other regions, the observed behaviors are more diverse. Realistic offline datasets are often heteroskedastic as they are typically generated by multiple policies, each with its own characteristics, under different conditions. E.g., driving datasets come from multiple humans [11], and many robotic datasets are collected by multiple teleoperators [10], resulting in systematic variability in different regions of the state space.

3.1 A Didactic Example

To understand why distribution constraints are insufficient with heteroskedastic data, we present a didactic example. Motivated by the driving scenario, we consider a maze navigation task shown in Fig. 1. The task is to navigate from the position labeled as "Start" to the position labeled as "Goal" using five actions at every possible state (L: $\leftarrow$, R: $\rightarrow$, U: $\uparrow$, D: $\downarrow$, No: No Op), while making sure that the executed actions do not hit the walls of the grid.

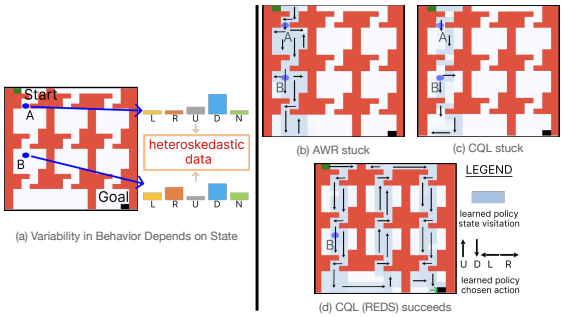

Figure 1: **Failure mode of distribution constraints.** In this navigation task, an offline RL algorithm must find a path from the start state to the goal state as indicated in (a). The offline dataset provided exhibits non-uniform coverage at different states, e.g., in the state marked as "B" located in a wide room has more uniform action distribution, whereas the states in the narrow hallways exhibit a more narrow action distribution. This is akin to how the behavior of human drivers varies in certain locations ("B"), but is very similar in other situations ("A"). To perform well, an algorithm must stay close to the data in the hallways ("A"), but deviate significantly from the data in the rooms ("B"), where the data supports many different behaviors (most are not good). AWR and CQL get stuck because they stay too close to the bad behavior policy in the rooms, e.g. the left and right arrows near State B in Fig (b) and (c). Our method, CQL (ReDS), learns to ignore the bad behavior action in state B and prioritizes the good action, indicated by the downward arrow near B in (d).

Dataset construction. To collect a heteroskedastic dataset, we consider a mixture of several behavior policies that attain a uniform occupancy over different states in the maze. However, the dataset action distributions differ significantly in different states. The induced action distribution is heavily biased to move towards the goal in the narrow hallways (e.g., the behavior policy moves upwards at state A)). In contrast, the action distribution is quite diverse in the wider rooms. In these rooms, the behavior policy often selects actions that do not immediately move the agent towards the goal (e.g., the behavior policy at state B), because doing so does not generally hit the walls as the rooms are wider, and hence the agent is not penalized. Whereas, the agent must take utmost precaution to not hit the walls in the narrow hallways. More details are in Appendix B.

Representative distribution constraint algorithms such as AWR [44, 43] and CQL [33] fail to perform the task, as shown in Figure 1. To ensure fair comparison, we tune each method to its best evaluation performance using online rollouts. The visualization in Figure 1 demonstrates that these two algorithms fail to learn reasonable policies because the learned policies match the random behavior of the dataset actions too closely in the wider rooms, and therefore are unable to make progress towards the Goal position. This is a direct consequence of enforcing too strong of a constraint on the learned policy to stay close to the behaviors in the dataset. Therefore, we also evaluated the performance of CQL and AWR in this example, with lower amounts of conservatism (Appendix B) and found that utilizing a lower amount of conservatism suffers from the opposite failure mode: it is unable to prevent the policies from hitting the walls in the narrow hallways. This means that conservatism prevents the algorithm from making progress in the regions where the behavior in the dataset is more diverse, whereas not being conservative enough hurts performance in regions where the behaviors in the dataset agree with each other. The method we propose in this paper to tackle this challenge, indicated as "CQL (ReDS)", effectively traverses the maze, 80% of the time.

3.2 Challenges with Distribution Constraints

Having seen that distribution constraints can fail in certain scenarios, we now formally characterize when offline RL datasets is heteroskedastic, and why distribution constraints may be ineffective in

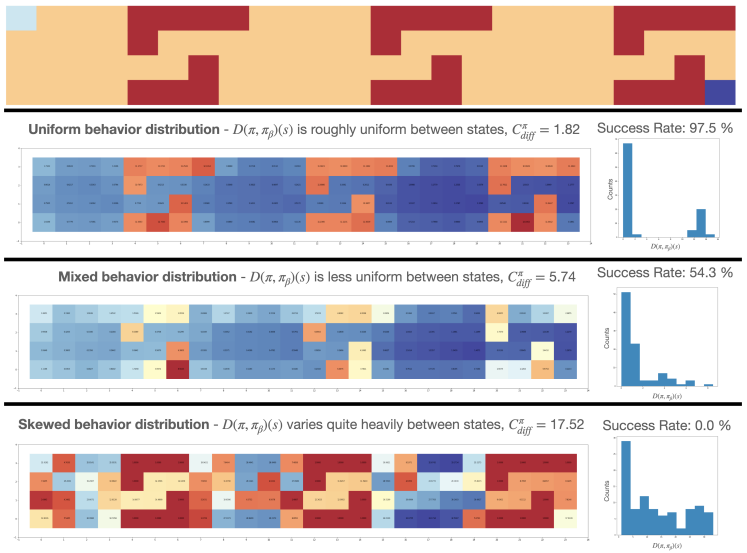

Figure 2: **Empirically computing C^π_{diff}** with three datasets: uniform (top), mixed (middle) and skewed (bottom) on a gridworld. We also visualize $D(\pi, \pi_\beta)(\mathbf{s})$ across states in the maze as the colors on different cells, a histogram of $D(\pi, \pi_\beta)(\mathbf{s})$ to visualize variation in this quantity and the performance of running standard CQL. **Top:** The uniform distribution leads to low C^π_{diff}, uniform $D(\pi, \pi_\beta)(s)$, and highest success. **Middle:** The mixed distribution leads to medium C^π_{diff}, less uniformly distributed $D(\pi, \pi_\beta)(s)$, and a drop in task success. **Bottom**: The skewed distribution leads to a high C^π_{diff}, non-uniform $D(\pi, \pi_\beta)(s)$, and poor performance.

such scenarios. Similar to how standard analyses utilize concentrability coefficient [46], which upper bounds the ratio of state-action visitation under a policy $d^\pi(\mathbf{s}, \mathbf{a})$ and the dataset distribution μ, i.e., $\max_{\mathbf{s},\mathbf{a}} d^\pi(\mathbf{s}, \mathbf{a})/\mu(\mathbf{s}, \mathbf{a}) \leq C^\pi$, we introduce a new metric called *differential concentrability*, which measures dataset heteroskedasticity (i.e., the variability in the dataset behavior across different states).

Definition 3.1 (Differential concentrability.)**.** Given a divergence D over the action space, the differential concentrability of a given policy π with respect to the behavioral policy π_β is given by:

$$C^\pi_{\text{diff}} = \mathop{\mathbb{E}}_{\mathbf{s}_1, \mathbf{s}_2 \sim d^\pi} \left[\left(\sqrt{\frac{D(\pi, \pi_\beta)(\mathbf{s}_1)}{\mu(\mathbf{s}_1)}} - \sqrt{\frac{D(\pi, \pi_\beta)(\mathbf{s}_2)}{\mu(\mathbf{s}_2)}} \right)^2 \right]. \tag{4}$$

Eq. 4 measures the variation in the divergence between a given policy $\pi(\mathbf{a}|\mathbf{s})$ and the behavior policy $\pi_\beta(\mathbf{a}|\mathbf{s})$ weighted inversely by the density of these states in the offline dataset (i.e., $\mu(\mathbf{s})$ in the denominator). For simplicity, let us revisit the navigation example from Section 3.1 and first consider a scenario where $\mu(\mathbf{s}) = \text{Unif}(\mathcal{S})$. For any given policy π, if there are states where π chooses actions that lie on the fringe of the data distribution (e.g., in the wider rooms), as well as states where the policy π chooses actions at the mode of the data distribution (e.g., as in the narrow passages), then C^π_{diff} would be large any policy π that we learn. Crucially, C^π_{diff} would be small even if the learned policy π deviates significantly from the behavior policy π_β, such that $D(\pi, \pi_\beta)(\mathbf{s})$ is large, but $|D(\pi, \pi_\beta)(\mathbf{s}_1) - D(\pi, \pi_\beta)(\mathbf{s}_2)|$ is small, indicating the dataset is not heteroskedastic.

Connection between variability in the action distribution and high C^π_{diff}. Consider a simpler formula where we remove the counts $n(\mathbf{s})$ from the expression of differential concentrability and set π in C^π_{diff} to be the uniform distribution over actions. Then, we can show that the value of C^π_{diff} is *exactly* equal to twice the variance of $D(\pi, \pi_\beta)(\mathbf{s})$ across states. Therefore, we will demonstrate in Section 5 that arbitrary policy checkpoints π learned by offline RL algorithms generally attain a low value of the variance in $D(\pi, \pi_\beta)(\mathbf{s})$ on offline datasets from non-heteroskedastic sources, such as those covered in the D4RL [13] benchmark. Of course, we cannot always exclude the counts of states $n(\mathbf{s})$, however, we note that in high-dimensional state spaces, such as those in our experiments, each state in the offline data is likely to be unique, thus validating the condition that $n(\mathbf{s}) = 1$. That said, we do compute the exact value of C^π_{diff} (with $n(\mathbf{s})$) in a didactic gridworld maze shown in Figure 2. In this case, we find that our definition of C^π_{diff} is actually able to reflect the intuitive notion of heteroskedasticity.

We now use the definition of differential concentrability to bound both the improvement and deprovement of π w.r.t. π_β for distribution constraint algorithms using the framework of safe policy

improvement [37, 33]. We show that when C^{π}_{diff} is large, then constraints (Eq. 1) may not improve significantly over π_β, even for the best value for the weight α (proof in Appendix C):

Theorem 3.2 (Informal; Limited policy improvement via distributional constraints.)**.** *W.h.p.* $\geq 1-\delta$*, for any prescribed level of safety* ζ*, the maximum possible policy improvement over choices of* α*,* $\max_\alpha \; [J(\pi_\alpha) - J(\pi_\beta)] \leq \zeta^+$*, where* ζ^+ *is given by:*

$$\zeta^+ := \max_{\alpha} \quad \frac{h^*\left(\alpha\right)}{(1-\gamma)^2} \text{s.t.} \; \frac{c_1\sqrt{\log\frac{|\mathcal{S}||\mathcal{A}|}{\delta}}}{(1-\gamma)^2}\sqrt{\frac{C^{\pi_\alpha}_{diff}}{|\mathcal{D}|}} - \frac{\alpha\mathbb{E}_{\mathbf{s}\sim\widehat{d}^{\pi_\alpha}}\left[D(\pi_\alpha,\pi_\beta)(\mathbf{s})\right]}{1-\gamma} \leq \zeta, \tag{5}$$

where h^* *is a monotonically decreasing function of* α*, and* $h(0) = \mathcal{O}(1)$*.*

Theorem 3.2 quantifies the fundamental tradeoff with distribution constraints: to satisfy a given ζ-safety constraint in problems with larger C^{π}_{diff}, we would need a larger α. Since the maximum policy improvement ζ^+ is upper bounded by $h^*(\alpha)$, the policy may not necessarily improve over the behavior policy if α is large. On the flip side, if we choose to fix the value of α to be small in hopes to attain more improvement in problems where C^{π}_{diff} is high for all policies, we would end up compromising on the safety guarantee as ζ needs to be large for a small α and large C^{π}_{diff}. Thus, in this case, the policy may not improve over the behavior policy reliably.

Note that a larger value of C^{π}_{diff} need not imply large $\mathbb{E}_{\mathbf{s}\sim\widehat{d}^{\pi}}[D(\pi,\pi_\beta)(\mathbf{s})]$ because the latter does not involve $\mu(\mathbf{s})$. C^{π}_{diff} also measures the dispersion of $D(\pi,\pi_\beta)(\mathbf{s})$, while the latter performs a mean over states. In addition, Theorem 3.2 characterizes the *maximum possible* improvement with an *oracle* selection of α, though is not feasible in practice. Thus, when C^{π}_{diff} is large, distribution constraint algorithms could either not safely improve over π_β or would attain only a limited improvement with *any possible* value of α. Finally, we remark that complementing [32, 39] that discuss failure modes of distribution constraints with high-entropy behavior policies, Theorem 3.2 quantifies when this would be the case: this happens when C^{π}_{diff} is large.

4 Support Constraints As Reweighted Distribution Constraints

Thus far, we have seen that distribution constraints can be ineffective with heteroskedastic datasets. If we can impose the distribution constraint such that the constraint strength can be modulated per state, then in principle, we can alleviate the issue raised in Theorem 3.2 and Section 3.1.

Our key insight is that by reweighting the action distribution in the data before utilizing a distribution constraint, we can obtain a method that enforces a per-state distribution constraint, which corresponds to an approximate *support* constraint. This will push down the values of actions that are outside the behavior policy support, but otherwise not impose a severe penalty for in-support actions, thus enabling the policy to deviate from the behavior policy by different amounts at different states. Rather than having a distribution constraint between π and π_β (Eq. 1), if we can impose a constraint between π and a *reweighted* version of π_β, where the reweighting is state-dependent, then we can obtain an approximate support constraint. Let the reweighted distribution be π^{re}. Intuitively, if $\pi(\cdot|\mathbf{s})$ is within the support of the $\pi_\beta(\cdot|\mathbf{s})$, then one can find a reweighting $\pi^{re}(\cdot|\mathbf{s})$ such that $D(\pi,\pi^{re})(\mathbf{s}) = 0$, whereas if $\pi(\cdot|\mathbf{s})$ is not within the support of $\pi^{re}(\cdot|\mathbf{s})$, then $D(\pi,\pi^{re})(\mathbf{s})$ still penalizes π when π chooses out-of-support actions, since no reweighting π^{re} can put non-zero probability on out-of-support actions. This allows us to handle the failure mode from Section 3: at states with wide behavior policy, even with a large α, π is not anymore constrained to the behavior distribution, whereas at other "critical" states, where π_β is narrow, a large enough α will constrain $\pi(\cdot|\mathbf{s})$ to stay close to $\pi_\beta(\cdot|\mathbf{s})$. We call this **Re**weighting **D**istribution constraints to **S**upport (ReDS).

4.1 Instantiating the Principle Behind ReDS

One option is to reweight π_β to π^{re}, and enforce a distribution constraint $D(\pi,\pi^{re})$ between π and π^{re}. However, this is problematic because the π^{re} would typically be estimated by using importance weighting or by fitting a parametric model, and prior work has shown that errors in estimating the behavior policy [43, 20] using only one action sample often get propagated and lead to poor downstream performance. For CQL, this issue might be especially severe if we push up the Q-values under π^{re}, because then these errors might lead to severe Q-value over-estimation.

Abstract idea of CQL (ReDS). Instead, we devise an alternative formulation for ReDS that modifies the learned policy π to π^{re}, such that applying a distribution constraint on this modified policy imposes a support constraint. Thus, with CQL, now we instead *push down* the Q-values under π^{re}. We define π^{re} as a mixture distribution of the learned policy π and a reweighted version of the behavior policy as follows:

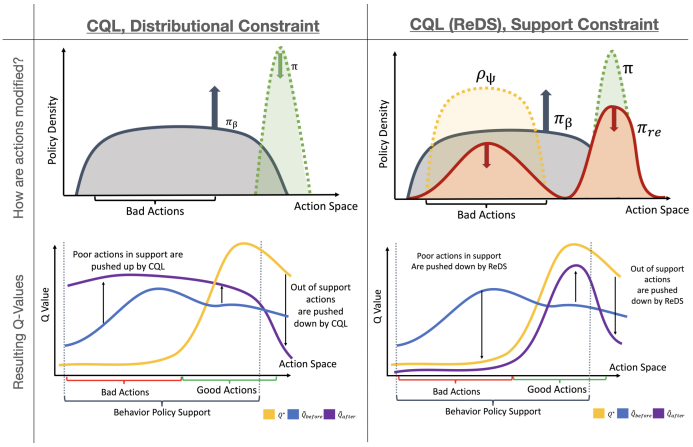

Figure 3: **Comparison between support and distributional constraints: Left:** CQL pushes down the Q-function under the policy π, while pushing up the function under the behavior policy π_β. This means that the Q-values for bad actions can go up. **Right:** In contrast, ReDS re-weights the data distribution to push down the values of bad actions, alleviating this shortcoming.

$$\pi^{re}(\cdot|\mathbf{s}) := \frac{1}{2}\pi(\cdot|\mathbf{s}) + \frac{1}{2}\left[\pi_\beta(\cdot|\mathbf{s}) \cdot g\left(\pi(\cdot|\mathbf{s})\right)\right], \tag{6}$$

where $g(\cdot)$ is a monotonically decreasing function. We will demonstrate how pushing down the Q-values under π^{re} modifies CQL to enable a support constraint while reusing existing components of CQL that impose a distribution constraint. As shown in Figure 3, the second term in Equation 6 increases the probability of actions that are likely under the behavior policy, but are less likely under the learned policy (due to g being a decreasing function). We will show in Lemma 4.1 that utilizing π^{re} in CQL enforces a support constraint on π. Thus, the learned policy π can be further away from π_β, allowing π to assign more probability to good actions that are within the behavior policy support, even if they have lower probabilities under π_β. Section 4.2 illustrates theoretically why pushing down the Q-values under Eq. 6 approximates a support constraint in terms of how it modifies the resulting Q-values. For an illustration, please see Figure 3.

How should we pick g in practice? Since we wish to use π^{re} as a replacement for π in the minimization term in the CQL regularizer (Equation 2), we aim to understand how to design the re-weighting g in practice. Since specifically CQL enforces a distribution constraint by maximizing the Q-value on *all* actions sampled from the behavior policy π_β, our choice of g should aim to counter this effect by instead minimizing the Q-value on "bad" actions within the support of the behavior policy. Equation 6 quantifies the notion of these "bad" actions using a monontonically decreasing function $g(\pi(\mathbf{a}|\mathbf{s}))$ of the policy probability. In practice, we find it convenient to define g to be a function of the advantage estimate: $A_\theta(\mathbf{s}, \mathbf{a}) := Q_\theta(\mathbf{s}, \mathbf{a}) - E_{\mathbf{a}\sim\pi}[Q_\theta(\mathbf{s}, \mathbf{a})]$, that the policy π is seeking to maximize. In fact, if entropy regularization is utilized for training the policy (akin to most offline RL algorithms), the density of an action under a policy is directly proportional to exponentiated advantages, i.e., $\pi(\mathbf{a}|\mathbf{s}) \propto \exp(A_\theta(\mathbf{s}, \mathbf{a}))$. Hence, we choose $g(x) = 1/x$, such that $g(\exp(A(\mathbf{s}, \mathbf{a}))) = \exp(-A(\mathbf{s}, \mathbf{a}))$ (a decreasing function).

For the rest, we approximate the product distribution $\pi(\mathbf{a}|\mathbf{s}) \cdot g(\pi(\mathbf{a}|\mathbf{s}))$ by fitting a parametric function approximator $\rho_\psi(\mathbf{a}|\mathbf{s})$. Since $\rho_\psi(\mathbf{a}|\mathbf{s})$ is being trained to approximate a re-weighted version of the behavior policy, we fit ρ_ψ by minimizing using a weighted maximum log-likelihood objective, as shown in prior work [44, 43]. The concrete form for our objective for training ρ_ψ is shown below (τ is a temperature hyperparameter typically introduced in prior work [44, 43]):

$$\rho_\psi(\cdot|\mathbf{s}) = \arg\max_{\rho_\psi} \; \mathbb{E}_{\mathbf{s}\sim\mathcal{D}, \mathbf{a}\sim\pi_\beta(\cdot|\mathbf{s})}[\log \rho_\psi(\mathbf{a}|\mathbf{s}) \cdot \exp\left(-A_\theta(\mathbf{s}, \mathbf{a})/\tau\right)]. \tag{7}$$

The crucial difference between this objective and standard advantage-weighted updates is the difference of the sign. While algorithms such as AWR [43] aim to find an action that attains a high advantage while being close to the behavior policy, and hence, uses a positive advantage, we utilize the *negative* advantage to mine for poor actions that are still quite likely under the behavior policy.

The final objective for the Q-function combines the regularizer in Eq. 8 with a standard TD objective:

$$\mathcal{R}(\theta; \rho) = \frac{1}{2}\left(\mathop{\mathbb{E}}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\pi}\left[Q_\theta(\mathbf{s},\mathbf{a})\right] + \mathop{\mathbb{E}}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\rho}\left[Q_\theta(\mathbf{s},\mathbf{a})\right]\right) - \mathop{\mathbb{E}}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}\left[Q_\theta(\mathbf{s},\mathbf{a})\right] \tag{8}$$

$$\min_\theta \; J_Q(\theta) = \mathcal{R}(\theta;\rho) + \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}[(Q_\theta(\mathbf{s},\mathbf{a}) - \mathcal{B}^\pi\bar{Q}(\mathbf{s},\mathbf{a}))^2] \tag{9}$$

4.2 Theoretical Analysis of CQL (ReDS)

Next, we analyze CQL (ReDS), showing how learning using the regularizer in Eq. 8 modifies the Q-values and justifies our choice of the distribution ρ in the previous section.

Lemma 4.1 (Per-state change of Q-values.)**.** *Let $g(\mathbf{a}|\mathbf{s})$ be a shorthand for $g(\mathbf{a}|\mathbf{s}) = g\left(\tau\cdot\pi(\mathbf{a}|\mathbf{s})\right)$. In the tabular setting, the Q-function obtained after one iteration of objective in Eq. 9 is given by:*

$$Q_\theta(\mathbf{s},\mathbf{a}) := \mathcal{B}^\pi\bar{Q}(\mathbf{s},\mathbf{a}) - \alpha\frac{\pi(\mathbf{a}|\mathbf{s}) + \pi_\beta(\mathbf{a}|\mathbf{s})g(\mathbf{a}|\mathbf{s}) - 2\pi_\beta(\mathbf{a}|\mathbf{s})}{2\pi_\beta(\mathbf{a}|\mathbf{s})} \tag{10}$$

where $\mathcal{B}^\pi\bar{Q}(\mathbf{s},\mathbf{a})$ is the Bellman backup operator applied to a delayed target Q-network.

Eq. 10 illustrates why the modified regularizer in Eq. 8 leads to a "soft" support constraint whose strength is modulated per-state. Since g is a monotonically decreasing function of π, for state-action pairs where $\pi(\mathbf{a}|\mathbf{s})$ has high values, $g(\mathbf{a}|\mathbf{s})$ is low and therefore the Q-value $Q(\mathbf{s},\mathbf{a})$ for such state-action pairs are underestimated less. Vice versa, for state-action pairs where $\pi(\mathbf{a}|\mathbf{s})$ attains low values, $g(\mathbf{a}|\mathbf{s})$ is high to counter-acts the low $\pi(\mathbf{a}|\mathbf{s})$ values. Also, since $\pi_\beta(\mathbf{a}|\mathbf{s})$ appears in the denominator, for out-of-support actions, where $\pi_\beta(\mathbf{a}|\mathbf{s}) = 0$, $\pi(\mathbf{a}|\mathbf{s})$ must also assign 0 probability to the actions for the Q values to be well defined. An illustration of this idea is shown in Figure 9. We can use this insight to further derive the closed-form objective optimized by ReDS.

Lemma 4.2 (CQL (ReDS) objective.)**.** *Assume that for all policies $\pi\in\Pi, \forall(\mathbf{s},\mathbf{a}), \pi(\mathbf{a}|\mathbf{s}) > 0$. Then, CQL (ReDS) solves the following optimization problem:*

$$\max_{\pi\in\Pi}\; \widehat{J}(\pi) - \frac{\alpha}{2(1-\gamma)}\mathbb{E}_{\mathbf{s}\sim\widehat{d}^\pi}\left[D(\pi,\pi_\beta)(\mathbf{s}) + \mathop{\mathbb{E}}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})}\left[g\left(\tau\cdot\pi(\mathbf{a}|\mathbf{s})\right)\mathbb{I}\left\{\pi_\beta(\mathbf{a}|\mathbf{s}) > 0\right\}\right]\right]. \tag{11}$$

$\widehat{J}(\pi)$ corresponds to the empirical return of the learned policy, i.e., the return of the policy under the learned Q-function. The objective in Lemma 4.3 can be intuitively interpreted as follows: The first term, $D(\pi,\pi_\beta)(\mathbf{s})$, is a standard distribution constraint, also present in naïve CQL, and it aims to penalize the learned policy π if it deviates too far away from π_β. ReDS adds an additional second term that effectively encourages π to be "sharp" within the support of the behavior policy (as g is monotonically decreasing), enabling π to potentially put mass on actions that lead to a high $\widehat{J}(\pi)$.

Specifically, this second term allows us control the strength of the distribution constraint per state: at states where the support of the policy is narrow, i.e., the volume of actions such that $\pi_\beta(\mathbf{a}|\mathbf{s}) > 0$ is small (say, only a single action), the penalty in Equation 51 reverts to a standard distributional constraint by penalizing divergence from the behavioral policy via $D(\pi,\pi_\beta)(\mathbf{s})$ as the second term cannot be minimized. At states where the policy π_β is broad, the second term counteracts the effect of the distributional constraint within the support of the behavior policy, by enabling π to concentrate its density on only good actions within the support of π_β with the same multiplier α. Thus even when we need to set α to be large to stay close to $\pi_\beta(\cdot|\mathbf{s})$ at certain states (e.g., in narrow hallways in the example in Sec. 3.1), $D(\pi,\pi_\beta)(\mathbf{s})$ is not heavily constrained at other states.

In fact, we formalize this intuition below to show that for the best possible value of the hyperparameters appearing in the training objective for CQL (ReDS) (Equation 51), CQL (ReDS) is guaranteed to outperform the best-tuned version of CQL for any offline RL problem. A proof is in Appendix C.

Lemma 4.3 (CQL (ReDS) formal guarantee)**.** *We will add the following guarantee to show that the policy learned by ReDS for the best possible value of τ (Equation 51) and α in CQL (Equation 3) outperforms the best CQL policy. That is, formally we show:*

$$\max_{\alpha,\tau} J(\pi_{\text{ReDS};\alpha,\tau}) \geq \max_{\alpha} J(\pi_{\text{CQL};\alpha}). \tag{12}$$

5 Experimental Evaluation

The goal of our experiments is to understand how CQL (ReDS) compares to distributional constraint methods when learning from heteroskedastic offline datasets. In order to perform our experiments,

we construct new heteroskedastic datasets that pose challenges representative of what we would expect to see in real-world problems. We first introduce tasks and heteroskedastic datasets that we evaluate on, and then present our results compared to prior state-of-the-art methods. We also evaluate ReDS on some of the standard D4RL [13] datasets which are not heteroskedastic in and find that the addition of ReDS, as expected, does not help, or hurt on those tasks.

5.1 Comparison on the D4RL Benchmark

Dataset	BC	10%BC	DT	AWAC	Onestep RL	TD3+BC	COMBO	CQL	IQL	**Ours**
halfcheetah-medium-replay	36.6	40.6	36.6	40.5	38.1	44.6	**55.1**	45.5	44.2	**52.3**
hopper-medium-replay	18.1	75.9	82.7	37.2	**97.5**	60.9	89.5	95.0	94.7	**101.5**
walker2d-medium-replay	26.0	62.5	66.6	27.0	49.5	**81.8**	56.0	77.2	73.9	**85.0**
halfcheetah-medium-expert	55.2	92.9	86.8	42.8	**93.4**	90.7	90.0	**91.6**	86.7	89.5
hopper-medium-expert	52.5	110.9	107.6	55.8	103.3	98.0	**111.1**	105.4	91.5	**110.0**
walker2d-medium-expert	107.5	109.0	108.1	74.5	**113.0**	110.1	103.3	108.8	109.6	**112.0**
locomotion total	295.9	491.8	488.4	277.8	494.8	486.1	505	**523.5**	500.6	**550.3**

Table 1: Performance comparison on the D4RL benchmark. (Top 2 **bolded**)

Heteroskedastic data is likely to exist in real-world problems such as driving and manipulation, where datasets are collected by multiple policies that agree and disagree at different states. While standard benchmarks (D4RL [13] and RLUnplugged [21]) include offline datasets generated by mixture policies (e.g. the "medium-expert" generated by two policies with different performance), these policies are trained via RL methods (SAC) that constrain the entropy of the action distribution at each state to be uniform. To measure heteroskedasticity, we utilize an approximation to C^{π}_{diff}: the standard deviation in the value of $D(\pi, \pi_\beta)(\mathbf{s})$ across states in the dataset, using a fixed policy π obtained by running CQL. We didn't use C^{π}_{diff} directly, as it is challenging to compute in continuous spaces. In Table 3, the standard deviation is lower for the D4RL antmaze datasets, corroborating our intuition that these datasets are significantly less heteroskedastic.

5.2 Comparisons on Heteroskedastic datasets

Heteroskedastic datasets. To stress-test our method and prior distribution constraint approaches, we collected new datasets for the medium and large mazes used in the antmaze navigation tasks from D4RL: `noisy` datasets, where the behavior policy action variance differs in different regions of the maze, representative of user variability in navigation, and `biased` datasets, where the behavior policy admits a systematic bias towards certain behaviors in different regions of the maze, representative of bias towards certain routes in navigation problems. Table 3 shows that these datasets are significantly more heteroskedastic to the D4RL datasets.

Dataset	std	max
noisy (Ours)	**18**	**253**
biased (Ours)	**9**	**31**
diverse (D4RL)	2	11
play (D4RL)	2	13

Table 3: The new antmaze datasets (Ours) are significantly more heteroskedastic than the standard D4RL datasets. We measure heteroskedasticity using the std and max of $D(\pi, \pi_\beta)(\mathbf{s})$ across states in the offline dataset.

Using these more heteroskedastic datasets, we compare CQL (ReDS) with CQL and IQL [27], recent popular methods, and two prior methods, BEAR [30] and EDAC [3], that also enforce support constraints. For each algorithm, including ours, we utilize hyperparameters directly from the counterpart tasks in D4RL. Due to the lack of an effective method for offline policy selection (see Fu et al. [14]), we utilize oracle checkpoint selection for every method. We compute the mean and standard deviation across 3 seeds. Table 2 shows that the largest gap between CQL (ReDS) and prior methods is on `noisy` datasets, which are particularly more heteroskedastic (Table 3).

We also compare CQL (ReDS) with recent offline RL algorithms on D4RL, including DT [8], AWAC [42], onestep RL [5], TD3+BC [16] and COMBO [57]. Table 1 shows that CQL (ReDS) obtains similar performance as existing distributional constraint methods and outperforms BC-based baselines. This is expected given that the D4RL datasets exhibit significantly smaller heteroscedasticity, as previously explained. Also, a large fraction of the datasets is trajectories with high returns. BC using the top 10% trajectories with the highest episode returns already has strong performance. The previous results compares CQL (ReDS) to baselines in tasks where the MDP states are low-dimensional vectors. Next, we study vision-based robotic manipulation tasks.

Visual robotic manipulation. We consider two types of manipulation tasks. In the "Pick & Place" task, the algorithm controls a WidowX robot to grasp an object and place it into a tray located at a test

Task & Dataset	EDAC	BEAR	CQL	IQL	INAC	RW & AW	EQL	SQL	XQL-C	**Ours**
medium-noisy	0	0	55	44	0	5	0.0	0.7	4.3	**73**
medium-biased	0	0	73	48	0	0	6.5	8.0	11.7	**74**
large-noisy	0	0	42	39	0	10	7.1	2.9	11.3	**53**
large-biased	0	0	50	41	0	8	8.5	0.5	7.3	45

Table 2: CQL (ReDS) outperforms prior offline RL methods including methods (IQL, XQL-C), and prior support constraint methods (BEAR, EDAC, SQL, EQL, RW & AW) on three out of four scenarios when learning from heteroskedastic data in the antmaze task. The improvement over prior methods is larger when learning from the `noisy` datasets, which are more heteroskedactic, as in Table 3, compared to `biased` datasets.

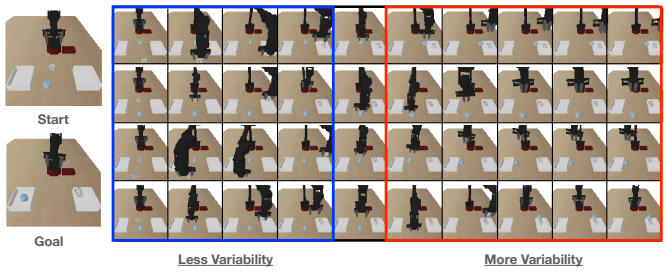

Figure 4: **Examples rollouts in the heteroskedastic bin-sort data.** In this task, an offline RL method must sort objects in front of it into two bins with a dataset that has non-uniform coverage at different states, using visual input. In the first half of the trajectory, the states exhibit a more narrow action distribution but the second half admits a more uniform action distribution.

location, directly from raw 128×128×3 images and sparse 0/1 reward signal. The dataset consists of behavior from suboptimal grasping and placing policies, and the positions of the tray in the offline dataset very rarely match the target test location. The placing policies exhibit significant variability, implying these datasets are heteroskedastic under our definition. We also consider "Bin Sort" task (see Figure 4), where a WidowX robot is controlled to sort two objects into two separate bins. Here, heteroskedacity is introduced when sorting objects into the desirable bins. Similar to the Pick & Place task, the placing policy exhibits significant variability, showing an object placed in the incorrect bin (e.g., recyclable trash thrown into the non-recyclable bin). However, the grasping policy is more expert-like grasping the object with low variability. More details in Appendix E.

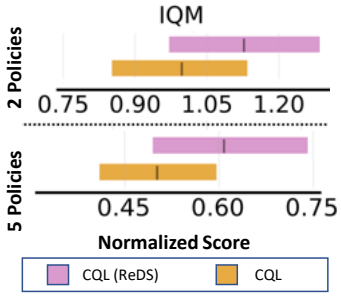

Figure 5: **CQL vs ReDS**: IQM normalized score for 10 Atari games. We consider two dataset compositions.

Table 4 presents the results on these tasks. We utilize oracle policy selection analogous to the antmaze experiments from Table 2. Table 4 shows that CQL (ReDS) outperforms CQL attaining a success rate of about 15.1% for the visual pick and place task, whereas CQL only attains 6.5% success. While performance might appear low in an absolute sense, note that both CQL and ReDS do improve over the behavior policy, which only attains a success rate of 4%. Thus offline RL does work on this task, and utilizing ReDS in conjunction with the standard distributional constraint in CQL does result in a boost in performance with this heteroskedastic dataset. For the "Bin Sorting", our method outperforms CQL by **3.5x** when learning from more heteroskedastic datasets. This indicates the effectiveness of our method in settings with higher heteroskedasticity.

Task	CQL	CQL (ReDS)	std $D(\pi, \pi_\beta)(\mathbf{s})$	max $D(\pi, \pi_\beta)(\mathbf{s})$
Pick & Place	6.5 ± 0.4	$\mathbf{15.1 \pm 0.4}$	48.7	307.4
Bin Sort (Easy)	$\mathbf{31.2 \pm 0.3}$	$\mathbf{31.4 \pm 0.3}$	7.9	81.6
Bin Sort (Hard)	6.1 ± 0.2	$\mathbf{23.1 \pm 0.7}$	59.6	988.3

Table 4: **CQL (ReDS) vs CQL** on robotic manipulation tasks. CQL (ReDS) outperforms CQL significantly when learning from more heteroskedastic datasets, as measured by C^{π}_{diff}: the standard deviation and the maximum of $D(\pi, \pi_\beta)(\mathbf{s})$ across states.

Atari games. We collect data on 10 Atari games from multiple policies that behave differently at certain states while having similar actions otherwise. We consider a case of **two** such policies, and a harder scenario of **five**. We evaluate the performance of CQL (ReDS) on the Atari games using the evaluation metrics from prior works [2, 34]. Figure 5 shows that in both testing scenarios: with the mixture of two policies (top figure) and the mixture of five policies (bottom figure), CQL (ReDS) outperforms CQL in aggregate.

To summarize, our results indicate that incorporating CQL (ReDS) outperforms distribution constraints with heteroskedastic datasets in a variety of domains.

6 Related Work

Offline Q-learning methods utilize mechanisms to prevent backing up unseen actions [39], by applying an explicit behavior constraint that forces the learned policy to be "close" to the behavior policy [23, 53, 44, 49, 53, 30, 28, 27, 52, 15], or by learning a conservative value function [33, 54, 41, 57, 56, 47, 24, 53]. Most of these offline RL methods utilize a distribution constraint, explicit (e.g., TD3+BC [15]) or implicit (e.g., CQL [33]), and our empirical analysis of representative algorithms from either family indicates that these methods struggle with heteroskedastic data, especially those methods that use an explicit constraint. Model-based methods [26, 56, 4, 51, 45, 38, 57] train value functions using dynamics models, which is orthogonal to our method.

Some prior works have also made a case for utilizing support constraints instead of distribution constraints, often via didactic examples [30, 29, 39], and devised algorithms that impose support constraints in theory, by utilizing the maximum mean discrepancy metric [30] or an asymmetric f-divergences [53] for the policy constraint [53]. Empirical results on D4RL [13] and the analysis by Wu et al. [53] suggest that support constraints are not needed, as strong distribution constraint algorithms often have strong performance. As we discussed in Sections 3.2 (Theorem 3.2 indicates that this distribution constraints may not fail when C^{π}_{diff} is small, *provided these algorithms are well-tuned*.) and 4, these benchmark datasets are not heteroskedastic, as they are collected from policies that are equally wide at all states and centered on good actions (e.g., Antmaze domains in [13], control suite tasks in Gulcehre et al. [21]) and hence, do not need to modulate the distribution constraint strength. To benchmark with heteroskedastic data, we developed some novel tasks which may be of independent interest beyond this work, and find that our method ReDS can work well here.

7 Discussion, Future Directions, and Limitations

We studied the behavior of distribution constraint offline RL algorithms when learning from heteroskedastic datasets, a property we are likely encounter in the real world. Naïve distribution constraint algorithms can be highly ineffective in such settings both in theory and practice, as they fail to modulate the constraint strength per-state. We propose ReDS, a method to convert distributional constraints into support-based constraints via reweighting, and validate it in CQL. A limitation of ReDS is that it requires estimating the distribution ρ_ψ to enforce a support constraint, which brings about its some additional compute overhead. Additionally, the instantiation of ReDS we develop in Section 4.1 is specific to methods that utilize a conservative regularizer such as CQL (or related approaches like COMBO). We clarify that our main contribution in this work is an analysis of when distributional constraints fail (which we study for AWR and CQL), and developing a principle for reformulating distributional constraints to approximate support constraints via reweighting. Devising approaches for enforcing support constraints that do not require extra machinery is a direction for future work. Understanding if support constraints are less sensitive to hyperparameters or are more amenable to model election is also a direction for future work.

References

[1] Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 22–31. JMLR. org, 2017.

[2] Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020.

[3] An, G., Moon, S., Kim, J.-H., and Song, H. O. Uncertainty-Based Offline Reinforcement Learning with Diversified Q-Ensemble. *arXiv e-prints*, art. arXiv:2110.01548, October 2021.

[4] Argenson, A. and Dulac-Arnold, G. Model-based offline planning. *arXiv preprint arXiv:2008.05556*, 2020.

[5] Brandfonbrener, D., Whitney, W. F., Ranganath, R., and Bruna, J. Offline RL without off-policy evaluation. *CoRR*, abs/2106.08909, 2021. URL https://arxiv.org/abs/2106.08909.

[6] Cao, K., Chen, Y., Lu, J., Arechiga, N., Gaidon, A., and Ma, T. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020.

[7] Castro, P. S., Moitra, S., Gelada, C., Kumar, S., and Bellemare, M. G. Dopamine: A Research Framework for Deep Reinforcement Learning. 2018. URL http://arxiv.org/abs/1812.06110.

[8] Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*, 2021.

[9] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[10] Ebert, F., Yang, Y., Schmeckpeper, K., Bucher, B., Georgakis, G., Daniilidis, K., Finn, C., and Levine, S. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[11] Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9710–9719, 2021.

[12] Fu, J., Kumar, A., Soh, M., and Levine, S. Diagnosing bottlenecks in deep Q-learning algorithms. *arXiv preprint arXiv:1902.10250*, 2019.

[13] Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[14] Fu, J., Norouzi, M., Nachum, O., Tucker, G., ziyu wang, Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., Paduraru, C., Levine, S., and Paine, T. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kWSeGEeHvF8.

[15] Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *arXiv preprint arXiv:2106.06860*, 2021.

[16] Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *CoRR*, abs/2106.06860, 2021. URL https://arxiv.org/abs/2106.06860.

[17] Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. *arXiv preprint arXiv:1812.02900*, 2018.

[18] Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning (ICML)*, pp. 1587–1596, 2018.

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

[20] Ghasemipour, S. K. S., Schuurmans, D., and Gu, S. S. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *International Conference on Machine Learning*, pp. 3682–3691. PMLR, 2021.

[21] Gulcehre, C., Wang, Z., Novikov, A., Paine, T. L., Colmenarejo, S. G., Zolna, K., Agarwal, R., Merel, J., Mankowitz, D., Paduraru, C., et al. Rl unplugged: Benchmarks for offline reinforcement learning. 2020.

[22] Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., and Levine, S. Soft actor-critic algorithms and applications. Technical report, 2018.

[23] Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

[24] Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020.

[25] Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pp. 651–673, 2018.

[26] Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

[27] Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[28] Kostrikov, I., Tompson, J., Fergus, R., and Nachum, O. Offline reinforcement learning with fisher divergence critic regularization. *arXiv preprint arXiv:2103.08050*, 2021.

[29] Kumar, A. Data-driven deep reinforcement learning. https://bair.berkeley.edu/blog/2019/12/05/bear/, 2019. BAIR Blog.

[30] Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11761–11771, 2019.

[31] Kumar, A., Fu, J., Tucker, G., and Levine, S. Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. *arXiv e-prints*, art. arXiv:1906.00949, June 2019.

[32] Kumar, A., Fu, J., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. 2019. URL http://arxiv.org/abs/1906.00949.

[33] Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

[34] Kumar, A., Agarwal, R., Ghosh, D., and Levine, S. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=O9bnihsFfXU.

[35] Kumar, A., Hong, J., Singh, A., and Levine, S. Should i run offline reinforcement learning or behavioral cloning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AP1MKT37rJ.

[36] Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

[37] Laroche, R., Trichelair, P., and Combes, R. T. d. Safe policy improvement with baseline bootstrapping. *arXiv preprint arXiv:1712.06924*, 2017.

[38] Lee, B.-J., Lee, J., and Kim, K.-E. Representation balancing offline model-based reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QpNz8r_Ri2Y.

[39] Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[40] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[41] Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[42] Nair, A., Dalal, M., Gupta, A., and Levine, S. Accelerating online reinforcement learning with offline datasets. *CoRR*, abs/2006.09359, 2020. URL https://arxiv.org/abs/2006.09359.

[43] Nair, A., Dalal, M., Gupta, A., and Levine, S. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

[44] Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

[45] Rafailov, R., Yu, T., Rajeswaran, A., and Finn, C. Offline reinforcement learning from images with latent space models. *Learning for Decision Making and Control (L4DC)*, 2021.

[46] Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

[47] Rezaeifar, S., Dadashi, R., Vieillard, N., Hussenot, L., Bachem, O., Pietquin, O., and Geist, M. Offline reinforcement learning as anti-exploration. *arXiv preprint arXiv:2106.06431*, 2021.

[48] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

[49] Siegel, N. Y., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., and Riedmiller, M. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

[50] Singh, A., Yu, A., Yang, J., Zhang, J., Kumar, A., and Levine, S. Cog: Connecting new skills to past experience with offline reinforcement learning. *arXiv preprint arXiv:2010.14500*, 2020.

[51] Swazinna, P., Udluft, S., and Runkler, T. Overcoming model bias for robust offline deep reinforcement learning. *arXiv preprint arXiv:2008.05533*, 2020.

[52] Wang, Z., Novikov, A., Żołna, K., Springenberg, J. T., Reed, S., Shahriari, B., Siegel, N., Merel, J., Gulcehre, C., Heess, N., et al. Critic regularized regression. *arXiv preprint arXiv:2006.15134*, 2020.

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

[54] Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

[55] Yarats, D., Brandfonbrener, D., Liu, H., Laskin, M., Abbeel, P., Lazaric, A., and Pinto, L. Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning. *arXiv preprint arXiv:2201.13425*, 2022.

[56] Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[57] Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. Combo: Conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*, 2021.

Appendices

A Practical ReDS Algorithm

Algorithm 1 CQL (ReDS) pseudo-code

1: Initialize Q-function, Q_θ, a policy, π_ϕ, and distribution ρ_ψ
2: **for** step t in $\{1, \ldots, \mathrm{N}\}$ **do**
3: Update ρ_ψ with the objective in Eq. 7:
4: $\psi_t := \psi_{t-1} + \eta_\rho \mathop{\mathbb{E}}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}[\log \rho_\psi(\mathbf{a}|\mathbf{s}) \cdot \exp(-A_\theta(\mathbf{s},\mathbf{a})/\tau)]$
5: Train the Q-function using $J_Q(\theta)$ in Eq. 9: $\theta_t := \theta_{t-1} - \eta_Q \nabla_\theta J_Q(\theta)$
6: Train π_ϕ as: $\phi_t := \phi_{t-1} + \eta_\pi \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\pi_\phi(\cdot|\mathbf{s})}[Q_\theta(\mathbf{s},\mathbf{a}) - \log \pi_\phi(\mathbf{a}|\mathbf{s})]$
7: **end for**

Our algorithm, CQL (ReDS) is illustrated in Algorithm 1. A complete implementation of the functions in python-style is provided in Appendix D.2. CQL (ReDS) first performs one gradient descent step on ρ_ψ using Eq. 7 (Lines 3), then performs one gradient step on the Q-function using the objective in Eq. 9 (Line 4), then performs one gradient descent step on the policy π_ϕ (Line 5), and repeats.

B Details of the Didactic Navigation Example from Section 3.1

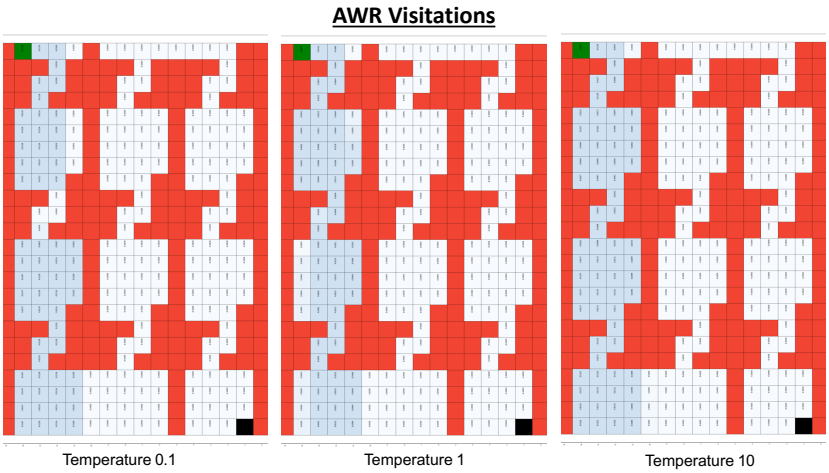

Figure 6: **Results of running AWR on the gridworld maze for a variety of temperature values** annotated with state-occupancy values. Observe that AWR is unable to reach the goal across a wide range of temperatures for the AWR hyperparameter. Even though for some of these hyperparameters, such as $\tau = 0.1$, it is able to traverse quickly into the third region, it does spend a larger fraction of its state visitation in the narrow hallways in this case (e.g., the final narrow hallway it is able to reach to) indicating that it gets stuck. Increasing the temperature to $\tau = 1.0$ does not solve it either, since now it does not even reach this hallway with as high of an occupancy.

In this section, we present some details regarding the navigation example we considered in Section 3.1. To create this example, we modified the gridworld code from Fu et al. [12] (code taken from: `https://github.com/justinjfu/diagnosing_qlearning`) to create the corresponding gridworld maze. We utilize a 24×16 gridworld as shown in the figures below, larger than the 8×8 or 16×16 gridworlds studied in this repository in the past. We utilize a smooth representation of the observation space. This is constructed by first sampling random Gaussian feature vectors from $\mathbb{R}^{50}$ for each grid cell (each grid cell is a state). Smoothing of these features vectors is then done locally, following the protocol for "grid-*-smoothobs" in Fu et al. [12]. This observation type presents a challenge for Q-learning algorithms that may often generalize incorrectly. This is because aliasing occurs with the predictions of nearby states that exhibit very different dynamics but not so different

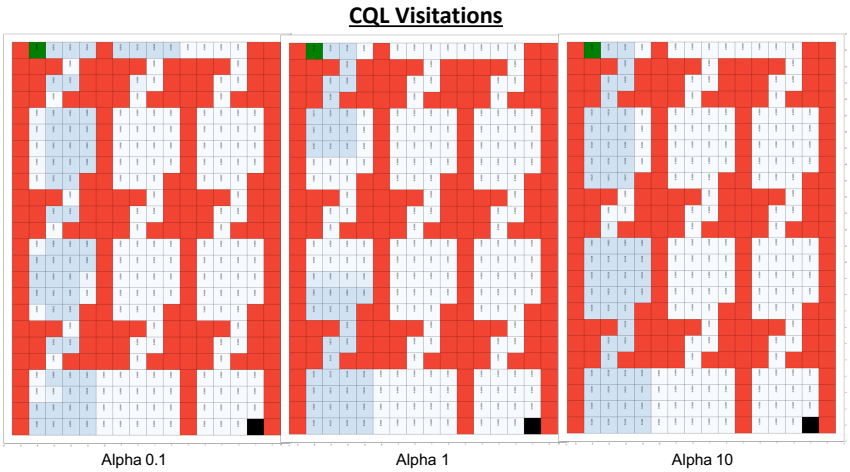

Figure 7: **Results of running CQL on the gridworld maze for a variety of temperature values.** Observe that CQL is unable to reach the goal and gets stuck for all the α values we studied.

observation vectors. The agents gets a sparse binary 0/1 reward: +1 is attained only when the agent reaches the goal (marked in black) from the start location (marked in green).

The behavior policy in each part of the maze is based on a mixture of different policies. In the wider rooms of the maze, one of the policies is a uniform policy, that uniformly chooses every action at a state. The second policy is a biased policy, that drives the agent away from the goal. In the narrow hallways, the behavior policy deterministically drives the agent towards the goal. For wide rooms, in contrast, a bias exists for actions taken in the direction away from the goal. This bias was set to be 0.8. This means that for rooms where you need to exit the wide passage by going down, the action of going up was selected 80% of the time by the behavior policy and each of the other action was randomly sampled with 20% probability. The opposite decision was taken for rooms where the goal was towards the top of the wide passage where 80% of the time the down action was selected.

Result visualizations. We now present some visualizations of the policies learned by various methods: AWR, CQL and CQL (ReDS). Since the reward values are binary and sparse, return curves for AWR and CQL are not as informative, since they attain a 0 return in any episode, even if they make some progress. Therefore, we present the results in the form of state-visitation density plots under rollouts from the learned policy. Observe in Figures 6 and 7 that neither AWR or CQL are able to actually successfully traverse the maze, and get stuck in it. Note that τ and α control the strength of the distributional constraint in the methods AWR and CQL respsectively. Varying the temperature hyperparameter τ for AWR and α for CQL does modify the density in the narrow hallways, but utilizing too small of a parameter leads the agent to more frequent crashes in the narrow hallways. This makes the agent spend a higher fraction of its visitation in one such narrow region, whereas a higher τ or α does not even reach the third narrow hallway with a high-enough visitation.

On the other hand, the method we propose in this paper, ReDS, when combined with CQL is able to successfully traverse this maze, as shown in Figure 8.

Analysis. These experiments are consistent with our expectations. When distributional constraints are faced with highly heteroskedastic action distributions at subsequent states, they failed to perform good actions at these consecutive states. A strong distributional constraint leads to the agent being too close to the behavior policy, whereas a weaker constraint fails to identify good actions at states where the behavior policy is narrow.

B.1 Why and When Do Distributional Constraints Fail in scenarios with Heteroskedastic Data?

In this section, we shall discuss why distributional constraints are especially worse than support constraints in scenarios with heteroskedastic data. Let's first consider a simple single-state bandit problem. In this simple one-state problem, we must find a single optimal action by using offline data. At first, it might appear that when faced with a wide behavior policy, distributional constraints

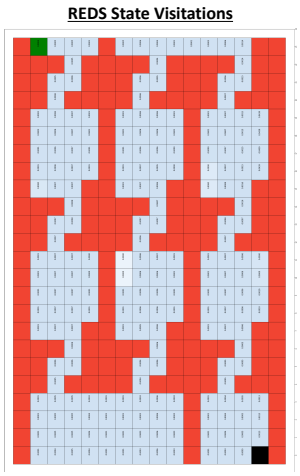

Figure 8: **Results of running CQL (ReDS) on the gridworld.** Note that CQL (ReDS) does actually succeed at solving the task.

would clearly fail since they would not deviate far away from the behavior policy. However, note that by carefully choosing the strength of the distributional constraint, we can, in principle, control the strength of the distributional constraint quite effectively. To see this concretely, consider applying CQL (Equation 2) to a single-state bandit problem. The Q-values on actions not observed in the training dataset will be clearly pushed down to $-\infty$. In addition, choosing a non-zero α will allow us to precisely control how close the action taken by the learned policy is to the best in-support action vs how close the learned policy is to the behavior policy. Therefore, for an optimally chosen value of α, distributional constraint methods should already work well in the single-state setting.

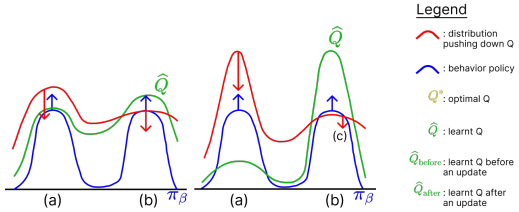

Figure 9: **An illustration of how ReDS modulates the strength of the distributional constraint per state. (i) Left:** let (a) and (b) denote the two modes of the behavior policy $\pi_\beta(\cdot|\mathbf{s})$ at a state $\mathbf{s}$. Consider in this setting that the learned Q-values at these modes are also roughly similar. In this case, the reweighted distribution π^{re} (in red) puts a roughly identical weight on pushing down Q-values on actions appearing in each of these modes. The resulting outcome is that the ReDS constraint behaves similarly to a distributional constraint. **(ii) Right:** In this case, let's consider a setting where actions appearing in one mode of the behavior policy attain very low Q-values (mode (a)). The re-weighted distribution π^{re} pushes down the actions appearing in the mode with lower Q-values with a larger weight than the mode with higher Q-values. This counteracts the effect of the push-up term in CQL that attempts to push up the Q-values for all actions under the behavior policy more strongly in region (a). In effect, this ensures that the Q-value for any action in the high-density regions of the behavior policy is not pushed up blindly, but the Q-values for only good, in-support actions (b) is pushed up. The ReDS constraint, therefore allows the learned policy to selectively pick out the better mode in this case.

However, the precise challenge with distributional constraints arises in the multi-state setting, where finding a single value of α that can work well at all states might be can be exceedingly challenging in practice. This is the setting with heteroskedastic data that our paper considers, including in our didactic example above. In such a setting, a strong distributional constraint is able to take a desirable action when the behavior policy is narrow (for example in the narrow room). However, the action distribution has wider support in the wider room and the strong distributional constraint would not take a desirable action in this setting due to the high entropy of the behavioral distribution in this setting. In contrast, with a weaker distributional constraint, a bad out-of-distribution action may be taken where the behavior policy is narrow which can lead to instability in the policy. Note that these challenges are not, however, present in the single-state scenario. Our method CQL (ReDS) can tackle

the problem in this setting by modulating the strength of the constraint per state. This allows the agent to stay close to the behavior distribution in the narrow room as well as learning to output the most desirable in-support action in the wider room.

B.2 Heteroskedastic Data vs Exploratory Data

There is a distinction between exploratory data [55] and heteroskedasticity. In the context of offline RL, exploratory data typically refers to being able to cover **all** possible state-action pairs, as uniformly as possible. On the other hand, heteroskedasticity is used to refer to **disproportionate** coverage: when the action distribution in the offline data is more uniform at certain states, but very narrow at others. Thus, a heteroskedastic dataset does not have high coverage in the sense of observing all possible state-action pairs.

Mathematically, while the notion of coverage is typically quantified by the concentrability coefficient which denotes the **worst case** density ratio between the distribution of the training dataset and the distribution of the learned policy, as shown below:

$$C^{\pi} = \max_{\mathbf{s},\mathbf{a}} \frac{d^{\pi}(\mathbf{s},\mathbf{a})}{\mu(\mathbf{s},\mathbf{a})},$$

where $\mu(\mathbf{s},\mathbf{a})$ denotes the state-action distribution induced by the dataset, while $d^{\pi}(\mathbf{s},\mathbf{a})$ denotes the state-action visitation distribution of a policy π. $\max_{\pi} C^{\pi}$ is generally expected to be smaller when $\mu(\mathbf{s},\mathbf{a})$ is more uniform. In contrast, we quantify the notion of heteroskedasticity by the variation in density ratios (in this case, action density ratios, $\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi_{\beta}(\mathbf{a}|\mathbf{s})}$) across the state space in Equation 4 (reproduced below for convenience).

$$C^{\pi}_{\text{diff}} = \mathop{\mathbb{E}}_{\mathbf{s}_1,\mathbf{s}_2\sim d^{\pi}} \left[\left(\sqrt{\frac{D(\pi,\pi_{\beta})(\mathbf{s}_1)}{\mu(\mathbf{s}_1)}} - \sqrt{\frac{D(\pi,\pi_{\beta})(\mathbf{s}_2)}{\mu(\mathbf{s}_2)}} \right)^2 \right].$$

Even if the dataset has high coverage (i.e., it has a low concentrability coefficient, C^{π}), it can attain a higher C^{π}_{diff}, because of higher variation in the density ratios across state spaces.

Out of all methods considered in Yarats et al. [55], we note that the methods classified as "Data" (APT, ProtoRL) attempt to maximize state coverage, the methods classified as "Knowledge" (ICM, Disagreement, RND) attempt to find surprising state-action pairs and visit them, and the methods classified as "Competence" (SMM, DIAYN, APS) that maximize the mutual information between the state distribution attained by the learned policy and the learned skills also implicitly optimize for state coverage. In effect, all of these methods aim to visit all state-action pairs (or all states), which is akin to the standard definition of concentrability, whereas heteroskedastic data corresponds distinctly to disproportionate action coverage at different states.

B.3 Differential concentrability in theory vs practice

The intuitive explanation for what makes a dataset heteroskedastic – namely, that the variability in the policy is different in different states – is provided primarily to build intuition. Unfortunately, this notion by itself does not make a great formal definition, because it does not capture the count of states $n(s)$ that shows up in a safe-policy improvement bound (the constraint in Equation 5). Therefore, in our analysis, we use the somewhat more complex notion of differential concentrability, which captures a similar intuition (we make the mathematical connection between the intuition of heteroskedasticity and our definition precise below) but provides us with the foundation on which to build our theoretical analysis.

To better understand the relationship between the variance of the action distribution and our Definition 3.1, consider a simpler scenario where we remove the counts $n(s)$ from the expression of differential concentrability and set π in $C^{\pi}_{diff}$to be the uniform distribution over actions. Then, we can show that the value of C^{π}_{diff} is exactly equal to twice the variance of entropy of the dataset action distribution across states, which matches with the metric (variance = square of the standard deviation) we measure in our experiments.

Of course, we cannot exclude the counts of states $n(s)$ for technical accuracy, however, it should be noted that in high-dimensional state space tasks, such as those examined in our experiments, each

state in the offline dataset is likely to be unique, thus validating the assumption that $n(s) = 1$. This means that measuring the variance in the entropy of the action distribution is an accurate estimate of differential concentrability in our experiments.

C Proofs

In this appendix, we will provide proofs for the various theoretical results in the main paper: Theorem 3.2. We will first discuss some preliminaries and notation, then present the proof for Theorem 3.2, and finally the remaining results.

C.1 Notation and Preliminaries

Let $\pi_\beta(\mathbf{a}|\mathbf{s})$ denote the behavior policy. Note that the dataset, $\mathcal{D}_i$ is generated from the marginal state-action distribution of π_β, i.e., $\mathcal{D} \sim d^{\pi_\beta}(\mathbf{s})\pi_\beta(\mathbf{a}|\mathbf{s})$. Define $\widehat{d^\pi}$ as the state marginal distribution introduced by π under the empirical MDP defined by the transitions in the dataset. Let $D_{\text{CQL}}(p, q)$ denote the following distance between two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ with equal support $\mathcal{X}$:

$$D_{\text{CQL}}(p, q) := \sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1 \right).$$

We drop the subsrcipt "CQL" from D_{CQL} for clarity. [33] showed that when optimizing the generic distributional constraint algorithm shown in Equation 1, the resulting policy π^* attains a high probability safe-policy improvement guarantee, i.e., $J(\pi^*) \geq J(\pi_\beta) - \zeta$, where ζ is:

$$\zeta = \mathcal{O}\left(\frac{1}{(1-\gamma)^2} \right) \mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi^*}} \left[\sqrt{\frac{D(\pi^*, \pi_\beta)(\mathbf{s}) + 1}{|\mathcal{D}(\mathbf{s})|}} \right] + \frac{\alpha}{1-\gamma} \mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi}} [D(\pi^*, \pi_\beta)(\mathbf{s})]. \tag{13}$$

We can further express $|\mathcal{D}(\mathbf{s})| = |\mathcal{D}||\mu(\mathbf{s})|$. The first term in Equation 13 corresponds to the decrease in performance due to sampling error and this term is high when the learned policy π^* visits low density states under the dataset distribution (i.e., $\mu(\mathbf{s})$ is small) and when the divergence from the behavior policy π_β is higher under these states. We will use this safe policy improvement guarantee in our proofs.

C.2 Formal Restatement of Theorem 3.2

Theorem C.1 (Formal version of Theorem 3.2)**.** *For any prescribed level of safety ζ, i.e., the learned policy satisfies $J(\pi) - J(\pi_\beta) \geq -\zeta$, the maximum possible policy improvement over choices of α:*

$$\max_\alpha \left[J(\pi_\alpha) - J(\pi_\beta) \right] \leq \zeta^+,$$

where ζ^+ is given by:

$$\zeta^+ := \max_\alpha \quad h^*(\alpha) \cdot \frac{1}{(1-\gamma)^2} \qquad \textit{s.t.} \quad \frac{c_1 \sqrt{\log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}}}{(1-\gamma)^2} \sqrt{\frac{C^{\pi_\alpha}_{\textit{diff}}}{|\mathcal{D}|}} - \zeta \leq \underbrace{\alpha \cdot \left(\frac{\mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi_\alpha}} [D(\pi_\alpha, \pi_\beta)(\mathbf{s})]}{1-\gamma} \right)}_{:= g(\alpha)} \tag{14}$$

where h^* is a monotonically decreasing function of α, and $h(0) = const.$

Intuition behind the theorem: We will show in the proof of this theorem that $g(\alpha)$ is a monotonically increasing function of α and $g(0) = 0$. This means that if the value of C^π_{diff} is large for all policies π, the value of $\alpha = g^{-1}\left(\sqrt{C^\pi_{\text{diff}}} c' - \zeta\right)$ (where c' is the constant and subsumes the term containing δ) is also large. Further, note that $h^*(\alpha)$ is decreasing in α. This means that the larger the value of α, the smaller the value of $h^*(\alpha)$, and hence, smaller the value of ζ^+. This means that as C^π_{diff} increases, α needs to take larger values to satisfy the safety constraint, and this reduces the value of maximal improvement, ζ^+.

Conversely, to attain a larger improvement ζ^+, if we choose a smaller α for a problem where C^π_{diff} is large for all policies, then, we must give up on the safety guarantee, and ζ would be large. This means that the learned policy π_α might substantially degrade beyond the behavior policy.

C.3 Proof of Theorem 3.2

In order to prove this result, we we will utilize a basic algebraic inequality mentioned below in Lemma C.2.

Lemma C.2 (Algebraic variation)**.** *Given any N positive real numbers, $x_1, x_2, \cdots, x_N$:*

$$\left(\sum_{i=1}^{N} \sqrt{x_i}\right)^2 \geq \sum_{i=1}^{N} x_i \geq \frac{1}{(N-1)} \sum_{i<j} (\sqrt{x_i} - \sqrt{x_j})^2. \tag{15}$$

Proof. For every x_i, define $y_i = \sqrt{x_i}$. Then, the difference between the two sides in the equation above is given by:

$$\sum_i y_i^2 - \frac{1}{(N-1)} \sum_{i<j} (y_i^2 + y_j^2 - 2y_i y_j) = \sum_i y_i^2 - \frac{N-1}{N-1} y_i^2 + \frac{1}{N-1} \sum_{i<j} 2y_i y_j \tag{16}$$

$$= \frac{1}{N-1} \sum_{i<j} 2y_i y_j \geq 0, \tag{17}$$

where the first step follows by rearranging y_i and y_j, and the final step follows by noting that $y_i \geq 0$ for all i. The other inequality follows trivially by noting that $\sqrt{x_i}$ are positive, and applying the standard formula for sum of squares. □

We will also require a Lemma that allows us to upper bound the performance difference $J(\pi) - J(\pi_\beta)$ in terms of the metric $D_{\text{CQL}}(\pi, \pi_\beta)$ that appears in the safe policy improvement guarantee in Equation 13.

Lemma C.3 (Tight upper bound on policy improvement.)**.** *Assume that the reward function $r(\mathbf{s}, \mathbf{a})$ of the MDP is bounded such that $\forall \mathbf{s}, \mathbf{a}, r(\mathbf{s}, \mathbf{a}) \in [-R_{\max}, R_{\max}]$. For any two policies π and π_β, we have the following:*

$$J(\pi) - J(\pi_\beta) \lesssim \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right) \cdot \mathbb{E}_{\mathbf{s} \sim d^\pi} \left[D(\pi(\cdot|\mathbf{s}), \pi_\beta(\cdot|\mathbf{s}))\right] \cdot R_{\max}. \tag{18}$$

Proof. The core of the proof of this lemma relies on the fact that for any given function $\nu(\mathbf{x})$ over some space $\mathbf{x}$, we can upper bound, $\Delta_\nu(p, q) := \mathbb{E}_{\mathbf{x} \sim p}[\nu(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim q}[\nu(\mathbf{x})]$ in terms of $D(p, q)$. To show this, we expand this expression:

$$\Delta_\nu(p, q) := \sum_{\mathbf{x}} (p(\mathbf{x}) - q(\mathbf{x})) \cdot \nu(\mathbf{x}) \tag{19}$$

$$= \sum_{\mathbf{x}} q(\mathbf{x}) \cdot \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1\right) \cdot \nu(\mathbf{x}) \tag{20}$$

$$= \sum_{\mathbf{x}: \frac{p(\mathbf{x})}{q(\mathbf{x})} \geq 1, \nu(\mathbf{x}) \geq 0} q(\mathbf{x}) \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1\right) \nu(\mathbf{x}) + \sum_{\mathbf{x}: \frac{p(\mathbf{x})}{q(\mathbf{x})} < 1, \nu(\mathbf{x}) \geq 0} q(\mathbf{x}) \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1\right) \nu(\mathbf{x})$$

$$+ \sum_{\mathbf{x}: \frac{p(\mathbf{x})}{q(\mathbf{x})} \geq 1, \nu(\mathbf{x}) \leq 0} q(\mathbf{x}) \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1\right) \nu(\mathbf{x}) + \sum_{\mathbf{x}: \frac{p(\mathbf{x})}{q(\mathbf{x})} < 1, \nu(\mathbf{x}) \leq 0} q(\mathbf{x}) \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1\right) \nu(\mathbf{x}).$$

Each of the four terms above can be bounded independently as follows: for the first two terms, we multiply by $\frac{p(\mathbf{x})}{q(\mathbf{x})}$, the third term is clearly negative, and the final term is upper bounded by multiplying

by $1 + \frac{p(\mathbf{x})}{q(\mathbf{x})}$:

$$\sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}\geq 1,\nu(\mathbf{x})\geq 0} q(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\nu(\mathbf{x}) \leq \sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}\geq 1,\nu(\mathbf{x})\geq 0} q(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\nu(\mathbf{x}) \tag{21}$$

$$\sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}< 1,\nu(\mathbf{x})\geq 0} q(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\nu(\mathbf{x}) \leq \sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}\leq 1,\nu(\mathbf{x})\geq 0} q(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\nu(\mathbf{x}) \tag{22}$$

$$\sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}\geq 1,\nu(\mathbf{x})\leq 0} q(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\nu(\mathbf{x}) \leq 0 \tag{23}$$

$$\sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}< 1,\nu(\mathbf{x})\leq 0} q(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\nu(\mathbf{x}) \leq \sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}< 1,\nu(\mathbf{x})\leq 0} q(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\left(1+\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)\nu(\mathbf{x}). \tag{24}$$

Finally, for each of these terms, we can now upper bound $\nu(\mathbf{x})$ by its maximum absolute value, $|\nu(\mathbf{x})| \leq \nu_0$, and combine the terms to get the following bound on $\Delta_\nu(p,q)$:

$$\Delta_\nu(p,q) \leq \nu_0 \sum_{\mathbf{x}:\nu(\mathbf{x})>0} p(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right) + 0 + \nu_0 \sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}<1,\nu(\mathbf{x})<0}\left(\frac{p^2(\mathbf{x})}{q(\mathbf{x})}-q(\mathbf{x})\right) \tag{25}$$

$$\leq \nu_0\left[\sum_{\mathbf{x}}\frac{p^2(\mathbf{x})}{q(\mathbf{x})}-1\right], \tag{26}$$

where Equation 26 follows from using the fact that for the case when $p(\mathbf{x})/q(\mathbf{x}) > 1$ but $\nu(\mathbf{x}) \leq 0$, $p(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right) > 0$, and hence it upper bounds the RHS of Equation 23. For the last case, where $\mathbf{x} : \frac{p(\mathbf{x})}{q(\mathbf{x})} < 1, \nu(\mathbf{x}) < 0$, we note that $\sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}<1,\nu(\mathbf{x})<0} q(\mathbf{x}) \geq \sum_{\mathbf{x}:\frac{p(\mathbf{x})}{q(\mathbf{x})}<1,\nu(\mathbf{x})<0} p(\mathbf{x})$, and hence the upper bound on this term in Equation 26 follows. To complete the argument note that D_{CQL} exactly takes the form obtained in the final equation, and hence:

$$\Delta_\nu(p,q) \leq \nu_0 \cdot D(p,q).$$

We can now use this result to bound the return differences, by using standard results for bounding the performance difference between policies [1, 48] in terms of $\frac{1}{(1-\gamma)^2} \times D(\pi,\pi_\beta)$. At the core of these results is a bound on $\mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})}[f(\mathbf{s},\mathbf{a})] - \mathbb{E}_{\mathbf{a}\sim\pi_\beta(\cdot|\mathbf{s})}[f(\mathbf{s},\mathbf{a})]$, and hence the result proven above directly applies. This proves the required result. □

C.4 Additional Technical Lemmas

In this section, we will provide proofs of two technical lemmas, that allow us to conclude the proof of Theorem 3.2. For these lemmas, we will consider a generic optimization problem,

$$\max_{\mathbf{x}} \; f(\mathbf{x}) + \alpha\, g(\mathbf{x}), \tag{27}$$

where $g(\mathbf{x}) > 0$ for any $\mathbf{x}$, and α can only take non-negative values.

Lemma C.4 (Value of $g(\mathbf{x})$ as a function of α)**.** *Let $\mathbf{x}^*_\alpha$ be the value of $\mathbf{x}$ that maximizes Equation 27 for a given fixed value of α. Then the following statements hold:*

1. *For any $\alpha \geq \beta \geq 0$, $g(\mathbf{x}^*_\alpha) \geq g(\mathbf{x}^*_\beta)$.*
2. *For any $\alpha \geq \beta \geq 0$, $\alpha g(\mathbf{x}^*_\alpha) \geq \beta g(\mathbf{x}^*_\beta)$.*

Proof. For any given α, $\mathbf{x}^*_\alpha$ satisfies the following inequality:

$$\forall \mathbf{x}', \quad f(\mathbf{x}^*_\alpha) + \alpha g(\mathbf{x}^*_\alpha) \geq f(\mathbf{x}') + \alpha g(\mathbf{x}'). \tag{28}$$

Using the above relation, we can write down two inequalities relating $\mathbf{x}^*_\alpha$ and $\mathbf{x}^*_\beta$:

$$f(\mathbf{x}^*_\alpha) + \alpha g(\mathbf{x}^*_\alpha) \geq f(\mathbf{x}^*_\beta) + \alpha g(\mathbf{x}^*_\beta) \tag{29}$$

$$f(\mathbf{x}^*_\beta) + \beta g(\mathbf{x}^*_\beta) \geq f(\mathbf{x}^*_\alpha) + \beta g(\mathbf{x}^*_\alpha) \tag{30}$$

Now, adding the two inequalities above, and cancelling the terms $f(\mathbf{x}^*_\alpha) + f(\mathbf{x}^*_\beta)$ from both sides, we obtain:

$$(\alpha - \beta)\, g(\mathbf{x}^*_\alpha) \geq (\alpha - \beta)\, g(\mathbf{x}^*_\beta). \tag{31}$$

Since $\alpha - \beta$ is non-negative, it is either equal to 0, in which case $g(\mathbf{x}^*_\alpha) = g(\mathbf{x}^*_\beta)$ or it is positive, in which case, $g(\mathbf{x}^*_\alpha) \geq g(\mathbf{x}^*_\beta)$. Combining these two cases, we get the desired result in (1). For proving the second part (2), note that we can write the difference of the two sides as:

$$\alpha g(\mathbf{x}^*_\alpha) - \beta g(\mathbf{x}^*_\beta) = \alpha g(\mathbf{x}^*_\alpha) - \alpha g(\mathbf{x}^*_\beta) + \alpha g(\mathbf{x}^*_\beta) - \beta g(\mathbf{x}^*_\beta) \tag{32}$$

$$= \alpha \left[g(\mathbf{x}^*_\alpha) - g(\mathbf{x}^*_\beta) \right] + (\alpha - \beta) g(\mathbf{x}^*_\beta) \tag{33}$$

$$\geq 0 + 0 = 0, \tag{34}$$

where the last inequality follows from the fact that $g(\mathbf{x}) \geq 0$ and $\alpha > \beta$. □

We will now provide a proof for Theorem 3.2.

Theorem C.5 (Theorem 3.2 restated)**.** *W.h.p. $\geq 1 - \delta$, for any prescribed level of safety ζ, the maximum possible policy improvement over choices of α, $J(\pi_\alpha) - J(\pi_\beta) \leq \zeta^+$, where ζ^+ is given by:*

$$\zeta^+ := \max_{\alpha} \quad h^*(\alpha) \cdot \frac{1}{(1-\gamma)^2} \qquad \textit{s.t.} \quad \frac{c_1}{(1-\gamma)^2} \sqrt{\frac{C^{\pi_\alpha}_{\textit{diff}}}{|\mathcal{D}|}} - \frac{\alpha}{1-\gamma} \mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi_\alpha}} [D(\pi_\alpha, \pi_\beta)(\mathbf{s})] \leq \zeta, \tag{35}$$

where h^* is a monotonically decreasing function of α, and $h(0) = \mathcal{O}(1)$.

Proof of Theorem C.5. To prove this theorem, we will apply Lemma C.2 on $x_i = \frac{D(\pi_\alpha(\cdot|\mathbf{s}_i)||\pi_\beta(\cdot|\mathbf{s}_i))}{\mu(\mathbf{s}_i)}$ and combine it with a the safe policy improvement guarantee for behavior regularization methods that admit updates of the form shown in Equation 1.

First, we note by applying Lemma C.2 in its expectation form that:

$$\left(\mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi_\alpha}} \left[\sqrt{\frac{D(\pi_\alpha(\cdot|\mathbf{s})||\pi_\beta(\cdot|\mathbf{s}))}{\mu(\mathbf{s})}} \right] \right)^2 \geq \mathbb{E}_{\mathbf{s}_1 \sim \widehat{d}^{\pi_\alpha}, \mathbf{s}_2 \sim \widehat{d}^{\pi_\alpha}} \left[\sqrt{\frac{D(\pi_\alpha(\cdot|\mathbf{s}_1)||\pi_\beta(\cdot|\mathbf{s}_1))}{\mu(\mathbf{s}_1)}} - \sqrt{\frac{D(\pi_\alpha(\cdot|\mathbf{s}_2)||\pi_\beta(\cdot|\mathbf{s}_2))}{\mu(\mathbf{s}_2)}} \right]^2,$$

where the term on the RHS of the above equation corresponds to C^{π}_{diff}.

Now we can plug this into the safe-policy improvement guarantee to obtain the resulting result as follows. Note that the first term in the bound in Equation 13 can be lower bounded using the differential concentrability as discussed above, and therefore, we get the following lower bound on ζ:

$$\zeta \geq \mathcal{O}\left(\frac{1}{(1-\gamma)^2} \right) \sqrt{\frac{C^{\pi_\alpha}_{\text{diff}}}{|\mathcal{D}|}} + \alpha \mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi_\alpha}} \left[D(\pi_\alpha, \pi_\beta)(\mathbf{s}) \right], \tag{36}$$

which is exactly the same as the expression for the constraint in Theorem 3.2.

Next we provide an upper bound on the maximal improvement that can be possible, in terms of $D(\pi_\alpha, \pi_\beta)$. For this, we will utilize Lemma C.3, and we can directly upper bound $J(\pi_\alpha) - J(\pi_\beta)$ as follows:

$$J(\pi_\alpha) - J(\pi_\beta) \lesssim \frac{1}{(1-\gamma)^2} \mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi_\alpha}} \left[D(\pi_\alpha, \pi_\beta)(\mathbf{s}) \right] \cdot R_{\max}. \tag{37}$$

Finally, we express this upper bound in terms of α.

Now, note that the RHS in Equation 37 depends on $\mathbb{E}_{\mathbf{s} \sim \widehat{d}^{\pi_\alpha}} \left[D(\pi_\alpha, \pi_\beta)(\mathbf{s}) \right]$, which is directly the term that a generic distributional constraint algorithm minimizes (Equation 1). We wish to understand how this term evolves a function of α.

Now we will invoke Lemma C.4 to understand the behavior of the term above when solving the optimization problem in Equation 1. To do so, consider any α, α' and note that $f(\mathbf{x}) = \widehat{J}(\pi)$ and $g(\mathbf{x}) = -\mathbb{E}_{\mathbf{s}\sim\widehat{d}^{\pi}}[D(\pi, \pi_\beta)(\mathbf{s})]$. Now applying Lemma C.4, Part (1), we note that:

$$\mathbb{E}_{\mathbf{s}\sim\widehat{d}^{\pi_\alpha}(\mathbf{s})}[D(\pi_\alpha, \pi_\beta)(\mathbf{s})] \leq \mathbb{E}_{\mathbf{s}\sim\widehat{d}^{\pi_{\alpha'}}(\mathbf{s})}[D(\pi_{\alpha'}, \pi_\beta)(\mathbf{s})], \tag{38}$$

for $\alpha' \leq \alpha$. This means that we can upper bound this quantity by a function $h^*(\alpha)$ that is monotonically decreasing in α.

Therefore, the maximal improvement is upper bounded by: $h^*(\alpha)\mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$, which completes the proof of this theorem. □

C.5 Proof of Lemma 4.1

Lemma C.6 ((Lemma 4.1 restated) Per-state modification.)**.** *Let g represents $g\left(\tau \cdot \pi(\cdot|\mathbf{s})\right)$. The Q-function obtained after one TD-learning iteration using the objective in Eq. 9 is:*

$$Q_\theta(\mathbf{s}, \mathbf{a}) := \mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a}) - \alpha \frac{\pi(\mathbf{a}|\mathbf{s}) + \pi_\beta(\mathbf{a}|\mathbf{s})g - 2\pi_\beta(\mathbf{a}|\mathbf{s})}{2\pi_\beta(\mathbf{a}|\mathbf{s})} \tag{39}$$

where $\mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a})$ is the Bellman backup operator applied to a delayed target Q-network.

Recall from Section 2 that the objective of CQL consists of two terms

$$\min_\theta \; \alpha \underbrace{\left(\mathbb{E}_{\mathbf{s}\sim\mathcal{D}, \mathbf{a}\sim\pi}\left[Q_\theta(\mathbf{s}, \mathbf{a})\right] - \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}\left[Q_\theta(\mathbf{s}, \mathbf{a})\right]\right)}_{\mathcal{R}(\theta)} + \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q_\theta(\mathbf{s}, \mathbf{a}) - \mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a})\right)^2\right], \tag{40}$$

where $\mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a})$ is the Bellman backup operator applied to a delayed target Q-network. In tabular setting, the Q-function obtained after one iteration of TD-learning using the objective function in Eq. 40 is given by:

$$Q_\theta(\mathbf{s}, \mathbf{a}) := \mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a}) - \alpha\left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi_\beta(\mathbf{a}|\mathbf{s})} - 1\right]. \tag{41}$$

In CQL, the result in Eq. 41 is obtained by setting the derivative of the objective in Eq. 40 with respect to the Q-values to 0, and solve for $Q_\theta(\mathbf{s}, \mathbf{a})$ [33]. The objective for ReDS is given by:

$$\min_\theta \quad J_Q(\theta) = \mathcal{R}(\theta; \rho) + \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q_\theta(\mathbf{s}, \mathbf{a}) - \mathcal{B}^\pi \bar{Q}(\mathbf{s}, \mathbf{a})\right)^2\right] \tag{42}$$

Notice that the main difference between the original CQL objective in Eq. 40 and the new objective in Eq. 42 is the distribution with which we push down Q values. The objective in Eq. 40 pushes Q values down under the learned policy π, whereas the objective in Eq. 42 pushes Q values down under a mixture of π and ρ, i.e. $\frac{1}{2}\pi + \frac{1}{2}\rho$. Since ρ is parameterized by a neural network whose input does not contain the Q values, its gradient with respect to the Q values is 0. Additionally, since the mixture $\frac{1}{2}\pi + \frac{1}{2}\rho$ plays the same role that π plays in the objective function in Eq. 40, we therefore can obtain the solution for the Q-values after updating the Q-function using the objective function in Eq. 42 simply by replacing π in Eq. 41 with the mixture. That is, in tabular setting, after updating the Q-function using the objective in Eq.42, the Q-values are:

$$
\begin{aligned}
Q_\theta(\mathbf{s},\mathbf{a}) &= \mathcal{B}^\pi \bar{Q}(\mathbf{s},\mathbf{a}) - \alpha \left[\frac{\frac{1}{2}\pi(\mathbf{a}|\mathbf{s}) + \frac{1}{2}\rho(\mathbf{a}|\mathbf{s})}{\pi_\beta(\mathbf{a}|\mathbf{s})} - 1 \right]. \\
&= \mathcal{B}^\pi \bar{Q}(\mathbf{s},\mathbf{a}) - \alpha \left[\frac{\pi(\mathbf{a}|\mathbf{s}) + \rho(\mathbf{a}|\mathbf{s})}{2\pi_\beta(\mathbf{a}|\mathbf{s})} - 1 \right]. \\
&= \mathcal{B}^\pi \bar{Q}(\mathbf{s},\mathbf{a}) - \alpha \left[\frac{\pi(\mathbf{a}|\mathbf{s}) + \rho(\mathbf{a}|\mathbf{s}) - 2\pi_\beta(\mathbf{a}|\mathbf{s})}{2\pi_\beta(\mathbf{a}|\mathbf{s})} \right]. \\
&= \mathcal{B}^\pi \bar{Q}(\mathbf{s},\mathbf{a}) - \alpha \left[\frac{\pi(\mathbf{a}|\mathbf{s}) + \pi_\beta(\mathbf{a}|\mathbf{s}) g(.) - 2\pi_\beta(\mathbf{a}|\mathbf{s})}{2\pi_\beta(\mathbf{a}|\mathbf{s})} \right].
\end{aligned}
$$

since ρ subsumes $\pi_\beta \cdot g$, giving us the desired result.

C.6 Proof of Lemma 4.3

Intuition: Comparison of CQL and CQL (ReDS) objectives We will first intuitively compare the objectives in CQL and CQL (ReDS) to understand where the difference arises from. We write down the CQL (ReDS) objective below:

$$
\max_{\pi \in \Pi} \widehat{J}(\pi) - \frac{\alpha}{2(1-\gamma)} \mathbb{E}_{\mathbf{s}\sim\widehat{d}^\pi} \left[D(\pi,\pi_\beta)(\mathbf{s}) + \cdot \mathop{\mathbb{E}}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})} \left[g\left(\tau \cdot \pi(\mathbf{a}|\mathbf{s})\right) \mathbb{I}\left\{\pi_\beta(\mathbf{a}|\mathbf{s}) > 0\right\} \right] \right]. \tag{43}
$$

and now the CQL objective:

$$
\max_{\pi \in \Pi} \widehat{J}(\pi) - \frac{\alpha}{(1-\gamma)} \mathbb{E}_{\mathbf{s}\sim\widehat{d}^\pi} \left[D(\pi,\pi_\beta)(\mathbf{s}) \right]. \tag{44}
$$

Observe that the only difference between CQL and ReDS stems from the fact that while the regularizer in CQL only optimizes the policy to stay close to the behavior policy, the regularizer in ReDS minimizes an additional term $\sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{s}) g(\tau \cdot \pi(\mathbf{a}|\mathbf{s}))$. This term attempts to make the behavior policy more "sharp", as g is a monotonically decreasing function of its argument, by preventing π to be less sharp in states where π_β is broad. This enables π to find the action in the dataset support that maximizes the learned return $\widehat{J}(\pi)$, even when π_β is broad.

To prove Lemma 4.3, we will consider the following abstract update form for the policy evaluation version of CQL (ReDS), that obtains the next Q-function iterate Q_{k+1}:

$$
\min_Q \alpha \left(\mathop{\mathbb{E}}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\pi^{re}} \left[Q(\mathbf{s},\mathbf{a})\right] - \mathop{\mathbb{E}}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\pi_\beta} \left[Q(\mathbf{s},\mathbf{a})\right] \right) + \frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}} \left[\left(Q(\mathbf{s},\mathbf{a}) - \mathcal{B}^\pi Q_k(\mathbf{s},\mathbf{a})\right)^2 \right], \tag{45}
$$

Lemma C.7 (CQL (ReDS) restated more completely.)**.** *CQL (ReDS) solves the following optimization problem, when α is large enough:*

$$
\max_\pi \ \widehat{J}(\pi) - \frac{\alpha}{2(1-\gamma)} \mathbb{E}_{\mathbf{s}\sim\widehat{d}^\pi} \left[D(\pi,\pi_\beta)(\mathbf{s}) + \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})} \left[g\left(\tau \cdot \pi(\mathbf{a}|\mathbf{s})\right) \mathbb{I}\left\{\pi_\beta(\mathbf{a}|\mathbf{s}) \geq 0\right\} \right] \right].
$$

Proof. For proving Lemma 4.3, we follow an argument similar to the proof of Theorem 3.1 from Kumar et al. [33]. By differentiating the above objective w.r.t. Q, we note that the ReDS + CQL objective above exhibits the following effective Bellman backup

$$
\forall \mathbf{s},\mathbf{a} \in \mathcal{D}, Q_{k+1}(\mathbf{s},\mathbf{a}) := \left(\mathcal{B}^\pi Q_k\right)(\mathbf{s},\mathbf{a}) - \alpha \left(\frac{\pi^{re}(\mathbf{a}|\mathbf{s})}{\pi_\beta(\mathbf{a}|\mathbf{s})} - 1 \right). \tag{46}
$$

This backup is equivalent to running pessimistic RL with a reward bonus equal to $-\alpha \left(\frac{\pi^{re}(\mathbf{a}|\mathbf{s})}{\pi_\beta(\mathbf{a}|\mathbf{s})} - 1 \right)$, and therefore, the policy obtained by maximizing the resulting Q-function can be expressed as:

$$
\max_\pi \ \widehat{J}(\pi) - \alpha \frac{1}{1-\gamma} \mathbb{E}_{\mathbf{s}\sim\widehat{d}^\pi} \left[\mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})} \left[\left(\frac{\pi^{re}(\mathbf{a}|\mathbf{s})}{\pi_\beta(\mathbf{a}|\mathbf{s})} - 1 \right) \right] \right] \tag{47}
$$

$$
\equiv \max_\pi \ \widehat{J}(\pi) - \alpha \frac{1}{2(1-\gamma)} \mathbb{E}_{\mathbf{s}\sim\widehat{d}^\pi} [D(\pi,\pi_\beta)(\mathbf{s})] - \frac{\alpha}{2(1-\gamma)} \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\widehat{d}^\pi} \left[\mathbb{I}\{\pi_\beta(\mathbf{a}|\mathbf{s}) > 0\} g\left(\tau \cdot \pi(\mathbf{a}|\mathbf{s})\right) \right]. \tag{48}
$$

We retain the notion of support $\mathbb{I}\{\pi_\beta(\mathbf{a}|\mathbf{s}) > 0\}$ because $\pi_\beta(\mathbf{a}|\mathbf{s})$ in the denominator only cancels out for actions within the support of the behavior policy, as for all other actions, this term the term would be ill-defined as $\pi_\beta(a|s) = 0$ appears in the denominator. Hence this term in the above equation cannot sum up over actions $\mathbf{a}$ not in the behavior policy in the second term (these actions will be pushed down to have $-\infty$ Q-values). If we just cancelled the π_β in the denominator, note that $g(\tau \cdot \pi(a|s))$ is not guaranteed to be 0 on actions $\mathbf{a}$, where $\pi_\beta(a|s) = 0$. Hence, we must retain the indicator function to explain this result properly. Instead, for the case of standard CQL, the divergence $D(\pi, \pi_\beta)(s)$ term prevents π from putting a non-zero density on actions where $\pi_\beta(a|s)$ in CQL, or else $D(\pi, \pi_\beta)(\mathbf{s}) = \infty$, meaning that the Q-value for such a π would be $-\infty$. □

C.7 Proof of Lemma 4.3

Proof. The central fact we use is that for any given value of α, if the temperature τ in Equation 9 is set to be extremely large, then $\exp(-A(\mathbf{s}, \mathbf{a})/\tau)$ approaches a constant value of 1 exponentially fast. Hence, $\rho(\mathbf{a}|\mathbf{s}) \to \pi_\beta(\mathbf{a}|\mathbf{s})$. When this happens, $\pi^{\text{re}}(\mathbf{a}|\mathbf{s})$ is given by:

$$\pi^{\text{re}}(\mathbf{a}|\mathbf{s}) = \frac{1}{2}\pi(\mathbf{a}|\mathbf{s}) + \frac{1}{2}(\pi_\beta(\mathbf{a}|\mathbf{s})). \tag{49}$$

Plugging in the value of $\pi^{\text{re}}(\mathbf{a}|\mathbf{s})$ into the expression for the regularizer $R(\theta; \rho)$, we note that the conservatism regularizer in ReDS is exactly equal to 0.5 times the standard CQL regularizer. Formally,

$$\begin{aligned} &\mathcal{R}_{\text{ReDS}}(\theta; \rho). = \mathbf{E}_{\mathbf{a}\sim\pi^{\text{re}}(\cdot|\mathbf{s})}[Q_\theta(\mathbf{s}, \mathbf{a})] - \mathbf{E}_{\mathbf{a}\sim\pi_\beta(\cdot|\mathbf{s})}[Q_\theta(\mathbf{s}, \mathbf{a})]. \\ &= \frac{1}{2}\left(\mathbf{E}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})}[Q_\theta(\mathbf{s}, \mathbf{a})] - \mathbf{E}_{\mathbf{a}\sim\pi_\beta(\cdot|\mathbf{s})}[Q_\theta(\mathbf{s}, \mathbf{a})]\right) = 0.5 \cdot \mathcal{R}_{\text{CQL}}(\theta). \end{aligned} \tag{50}$$

This means that ReDS (Equation 10) and CQL (Equation 2) will exhibit identical learning dynamics if $\alpha_{\text{ReDS}} = 2 \cdot \alpha_{\text{CQL}}$. As a result, the performance of ReDS for the best possible α and τ will strictly dominate the performance of CQL with the best possible α as CQL can be obtained as a special case of ReDS. Formally, this has shown that

$$\max_{\alpha,\tau} J(\pi_{\text{ReDS};\alpha,\tau}) \geq \max_{\alpha} J(\pi_{\text{CQL};\alpha}). \tag{51}$$

To conclude, the above guarantee shows that ReDS will never be worse than CQL in terms of performance.

□

D Implementation details of CQL (ReDS)

In this section, we will provide implementation details about our algorithm, CQL (ReDS). The pseudo-code in Algorithm 1 illustrates the different update steps of our algorithms. In addition, we provided a detailed python-like algorithm description for ease of implementation. This can be found below in Section D.2.

Most of the components of Algorithm 1 are straightforward and follow the same convention, training update and, as we will discuss, hyperparameters as the CQL algorithm. This includes training the policy π_ϕ, and for the most part training the critic Q_θ. The main difference in the update for CQL (ReDS) is utilizing the mixture of π and ρ in the CQL regularizer. For obtaining ρ_ψ, we utilize a standard advantage-weighted training update, following the papers [27, 43, 44]. Following these prior works, we also clip the argument to the exponent between a minimum range and a maximum range to be numerically stable:

$$\mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}[\log \rho_\psi(\mathbf{a}|\mathbf{s}) \cdot \exp\left[\text{clip}\left(-A^\pi_\theta(\mathbf{s}, \mathbf{a})/\tau, \sigma_{\min}, \sigma_{\max}\right)\right]. \tag{52}$$

In our experiments, we chose $\sigma_{\min} = -10$ and $\sigma_{\max} = 5$ across all the tasks and domains we study. These details are standard in training advantage-weighted algorithms.

Hyperparameters	Values
CQL Lagrange	True
CQL Lagrange Target Action Gap	0.8
Critic Network	256-256-256-256
Actor Network	256-256
Reward Scale	10
Reward Bias	-5
Critic Learning Rate	3×10^{-4}
Actor Learning Rate	3×10^{-4}

Figure 10: Full range of hyperparameters that we use for CQL(ReDS) building off of JaxCQL

D.1 Hyperparameters for CQL (ReDS)

D.2 Detailed algorithm description for CQL (ReDS)

Algorithm 1 provides the pseudo-code for CQL (ReDS). We provide the detailed description of how each update step in Algorithm 1 is implemented using Python syntax based on the PyTorch Framework in this section. We include 3 code listings below, illustrating the update steps for the parametric Q-functions, the policy and the learnt distribution ρ.

Listing 1: Training Q networks given a batch of data, corresponding to step 3 in Algorithm 1

```
q_data = critic(batch['observations'], batch['actions'])

next_dist = actor(batch['next_observations'])
next_pi_actions, next_log_pis = next_dist.sample()

target_qval = target_critic(batch['observations'],
                            next_pi_actions)
target_qval = batch['rewards'] + \
    self.gamma * (1 - batch['dones']) * target_qval

td_loss = mse_loss(q_data, target_qval)

# importance sampling term
num_samples = 4

# assume env is normalized between [-1, 1]
random_actions = uniform_sample((num_samples,
    batch['actions'].shape[-1]), min=-1, max=1)
random_pi = 0.5 ** batch['actions'].shape[-1]

dist = actor(batch['observations'])
pi_actions, log_pis = dist.sample(num_samples)

rho_dist = rho(batch['observations'])
rho_actions, log_probs_rho = rho_dist.sample(num_samples)

q_rand_is = critic(batch['observations'],
                   random_actions) - random_pi
q_pi_is = critic(batch['observations'],
                 pi_actions) - log_pis
q_rho_is = critic(batch['observations'],
                  rho_actions) - log_probs_rho

cat_q = concatenate(q_rand_is, q_pi_is, new_axis=True)
cat_q = logsumexp(cat_q, axis=-1)
```

```
cat_q_rho = logsumexp(q_rho_is, axis=-1)

# average between rho and pi
push_down_term_reds = 0.5 * (cat_q + cat_q_rho)

reds_loss = td_loss + \\
    ((push_down_term_reds - q_data).mean() * cql_alpha)

critic_optimizer.zero_grad()
reds_loss.backward()
critic_optimizer.step()
```

Listing 2: Training the policy (or the actor) given a batch of data (step 4 in Algorithm 1)

```
# Identical to CQL
# return distribution of actions
dist = actor(batch['observations'])

# sample actions with associated log probabilities
pi_actions, log_pis = dist.sample()

# calculate q value of actor actions
qpi = critic(batch['observations'], actions)
qpi = qpi.min(axis=0)

# same objective as CQL (kumar et al.)
actor_loss = (log_pis * self.alpha - qpi).mean()

# optimize loss
actor_optimizer.zero_grad()
actor_loss.backward()
actor_optimizer.step()
```

Listing 3: Training the ρ_ψ distribution given a batch of data (step 5 in Algorithm 1)

```
# AWR style update to find rho

# sample policy actions for advantage calculation
dist = actor(batch['observations'])
pi_actions, log_pis = dist.sample()

# calculate advantage
qdata = critic(batch['observations'], batch['actions'])
value = critic(batch['observations'], actions)
advantage = (qdata - value.min(0)).mean()

# awr style clipping
clipped_advantage = clip(advantage/self.temperature, \
    min=-10, max=5)

# find log rho(a|s)
rho_dist = _rho(batch['observations'])
log_prob_rho = rho_dist.log_prob(batch['actions'])

# Advantage Weighted Log Probabilities is the loss for rho
rho_loss = -(exp(-clipped_advantage) * log_prob_rho)
rho_loss = rho_loss.mean()

rho_optimizer.zero_grad()
```

```
rho_loss.backward()
rho_optimizer.step()
```

E Task and Dataset Descriptions

In this section, we will describe the various tasks we introduce in this paper. We also provide qualitative descriptions of these tasks here.

Heteroskedastic antmaze navigation. We introduce four new antmaze datasets which exhibit two different dataset distributions each for the medium and large mazes from D4RL [13]. We reuse the layouts of the mazes directly from D4RL, but modify the data collection protocol. For the `noisy` datasets, given an observation from the environment, we first compute the action that would have been taken by the D4RL behavior policy, and then add Gaussian noise to the action. While this alone is not much harder, crucially, note that the variance of this added Gaussian noise differs depending on the location of the Ant in the 2D Maze. In addition there is a small bias added to the D4RL behavior policy, but this bias is dominated by noise. We present the noise standard deviations (indicated "Noise") and the bias added (indicated "Bias") for this dataset as a function of different location intervals in the maze in the left part of the Figure 11 below.

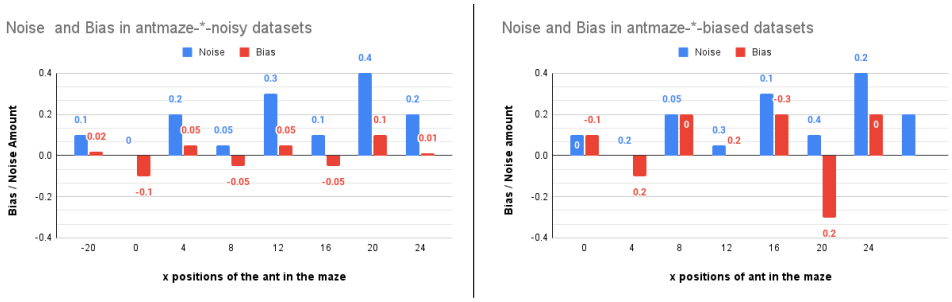

Figure 11: The distribution of noise and bias in the heteroskedastic antmaze datasets as a function of the x-position of the ant in the maze. While the noisy datasets primarily add noise, the biased datasets also add significant bias beyond the noise. The value of a given bar is the variance of the noise / bias added in the region between the x-position for that bar, and the next one.

For the `biased` datasets, in addition to adding location-dependent Gaussian noise to the action computed by D4RL behavior policies, we add a strong bias to the action (see Figure 11 (right)). Crucially note that the direction of this bias (i.e., the sign) changes based on the location of the Ant in the 2D maze, which mimics the scenario studied in our didactic navigation example in Section 3.1. In summary, because in some 2D regions of the maze, the values of the noise and bias added to the actions are larger, while in other 2D regions, they are smaller, the new offline datasets contain more heteroskedastic data distribution, where an optimal learned policy must deviate away from the data distribution much more in certain regions, whereas much lesser in other regions, which would correspond to an increase in the differential concentrability. This is demonstrated quantitatively in Table 3. Thus, we expect that learning well on these tasks modulating the strength of the distributional constraints per state.

Visual robotic pick and place. We introduce a pick and place dataset, which exhibits a unique dataset distribution for a robotic pick and place manipulation task, building on the framework from Singh et al. [50]. As shown in Figure 13, the robotic setup is a 6-DOF WidowX robot in front of a green bowl with 2 objects: a target object (the ball in this case) and a distractor object (the can). The objective is to place the target object into the bin. The reward function is a sparse, binary indicator of success, where a +1 reward is given when the object is placed in the bin. This task must be done from $128 \times 128 \times 3$ raw visual observations, without access to either the robot state, or the state of the objects, which can change as the objects can roll on the surface.

Visual robotic bin sorting. We introduce a bin sorting tasks, which are also built on the framework from Singh et al. [50]. As shown in Figure 13, the robotic setup is a 6-DOF WidowX robot in front of two identical white bins with 2 objects to sort. The objective is to sort each object into its respective

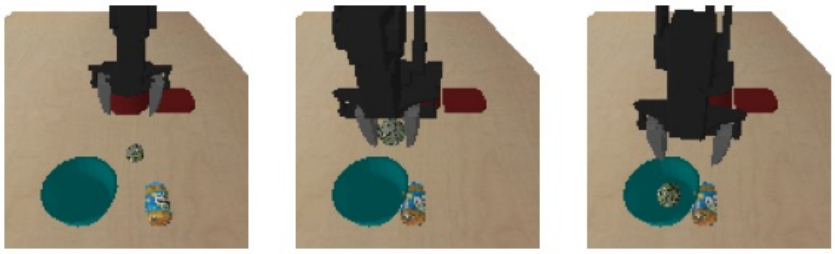

Figure 12: **Visualizing a sample trajectory for the visual pick-place robotic manipulation task.** Here is an example successful trajectory in the dataset collected using a scripted policy. Th robot reaches for the target object (the green ball), lifts it, and places it inside of the green container.

bins the target object in to the bin. The reward function is a sparse, binary indicator of success, where a "+1" reward is given when both objects are placed in their correct bins. This task must be done from $128 \times 128 \times 3$ raw visual observations, without access to either the robot state, or the state of the objects, which can change as the objects can roll on the surface.

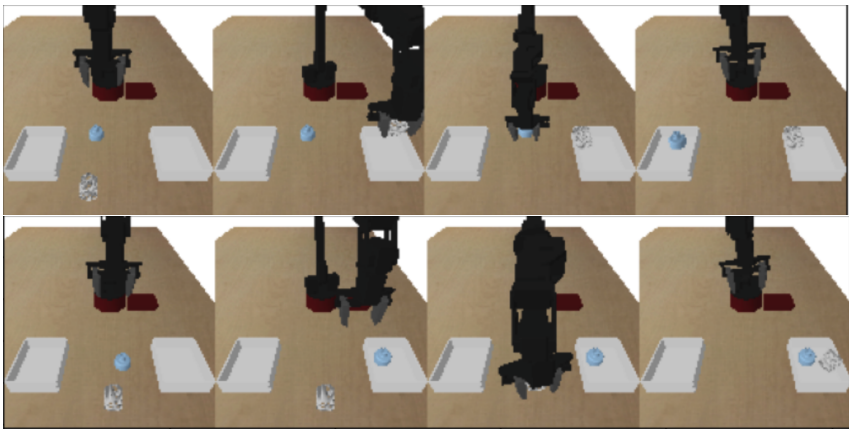

Figure 13: **Visualizing sample trajectories for the visual bin sorting robotic manipulation task.** Here are two sample trajectories for the binsorting domain. **Top:** A successful trajectory in the dataset. Here the robot places the white cylinder in the right bin and the blue ball in the left bin, succesfully sorting the objects into their respective bins. **Bottom:** A unsuccessful trajectory in the dataset. Here the robot places both objects in the same bin. This is unsuccessful as an object was placed in the incorrect bin thereby not sorting them in a correct manner.

To collect a heteroskedastic dataset, we run data collection using hardcoded scripted control policies, whose success rate and variance can be controlled. Each trajectory in the dataset was collected in the following manner. For the first phase, the robot reaches towards the object without any bias and grasps it with a reasonable success rate. Here, though, the noise and stochasticity in the scripted data collection and the inaccuracies in the scripted policy make it not succeed for every trial. During the second phase, where the robot places the object in the bin, there was a bias towards placing the object in a position in the workspace that does not correspond to the target location of the bin for which the robot can attain a reward, and which the robot will observe during evaluation. For our experiments, this bias was 85%. This forces the data distribution to be heteroskedastic: for the picking segment of this task, the behavior policy is centered around the desired optimal behavior i.e., grasping the object successfully, whereas for the placing segment, the behavior is biased towards carrying the object to the incorrect regions, requiring significant deviations from the behavior policy to succeed. An algorithm is now required to have a non-uniform amount of closeness to the behavior policy.

Atari game playing. For the Atari tasks we consider in the paper, we devised a heteroskedastic data composition based on the DQN Replay dataset [2] which comprises of transitions found in the replay buffer of a run of an online DQN. Since this dataset consist of all the policies that the DQN agent produced over the course of training, and since Atari typically uses ϵ-greedy exploration, where the

value of ϵ decays over time, different trajectories of this dataset are generated from different behavior policies, that all have different levels stochasticity. Naturally, since the value of ϵ decays over training and the performance of online DQN increases, the trajectories with higher return generally correlate with having lower stochasticity.

Given this information, we attempted to subsample a dataset that is heteroskedastic. For this purpose, we first divide the full replay buffer into N equal chunks, where chunk 0 consists to experience observed earliest in training, while the chunk $N-1$ consists of experience seen near the final parts of training. Then, we subsample 20% of the trajectories from each of these chunks independently to obtain an intermediate dataset that comes from multiplies policies, observed at different times while training online DQN. Then, for any given trajectory τ of length L in the replay chunk i, we only retain the transitions occurring between time steps $\lfloor \frac{(L-i)\times N}{L} \rfloor : \lfloor \frac{(L-i+1)\times N}{L} \rfloor$ in our final dataset and discard all the remaining transitions. This essentially means that the data closer to the initial states of the game comes from a good, less stochastic policy, whereas the data close to the final states of the game from a worse, highly stochastic policy. We develop two such datasets corresponding to $N = 2$ and $N = 5$ chunks. These chunks concatenated together construct the replay buffer of transitions that correspond to the 2 and 5 policy experiments seen in Section 5.

To see why this data is heteroskedastic, note that at different states of the game, we observe actions with different amounts of stochasticity and bias. This is because, as the game progresses, the effective behavior policy induced by the offline dataset exhibits a bias towards suboptimal actions (from the chunks that are earlier in DQN training) while also exhibiting substantial noise. The states that are closer to the initial states of the game, on the other hand, have an effective behavior policy that is primarily centered around a good action, with very little noise. In order to succeed, an offline RL algorithm must have different amount of conservatism at different states.

In our experiments, we considered 10 games including several standard games, and this is a subset of games studied in prior work [34]. The games we considered are: ASTERIX, BREAKOUT, Q*BERT, SEAQUEST, SPACEINVADERS, BEAMRIDER, MSPACMAN, WIZARDOFWOR, JAMESBOND, PONG.

F Experimental Details

For our experiments on the AntMaze domains, we built on the following open-source implementation of CQL: https://github.com/young-geng/JaxCQL, for our visual robotic experiments, we utilized our own port of the following implementation from Singh et al. [50] in Jax: https://github.com/avisingh599/cog, and for our Atari experiments, we use the official implementation of CQL built on Dopamine [7]: https://github.com/aviralkumar2907/CQL/tree/master/atari. For certain baselines (e.g EDAC, BEAR), we utilize the source implementation to stay consistent with the author's tested and tuned implementation. We additionally verified the results for D4RL benchmark for these tasks. We will summarize the hyperparameters in the next sections.

F.1 Hyperparameters for cql (ReDS)

Antmaze domains. For the AntMaze domains, we utilized a temperature parameter $\tau = 0.3$ in our experiments (found by sweeping over $\tau \in [0.1, 0.3, 1.0, 5.0]$), for all the four dataset types in Table 2. Every other hyperparameter was kept identical to CQL, which for the case of antmaze corresponds to applying the CQL regularizer $\mathcal{R}(\theta)$ with the dual version of CQL, where the threshold on the CQL regularizer is specified to be 1.0. Following CQL, we used 3-hidden layer critic and actor networks with layers of size 256, a critic learning rate of 3e-4 and an actor learning rate of 1e-4. We utilized the Bellman backup that computes the target value by performing a maximization over target values computed for $k = 10$ actions sampled from the policy at the next state.

Atari domains. For our Atari experiments, we tuned the value of α in CQL (Equation 2) between two values $[0.1, 0.2]$, and present the sensitivity results in Figure 5, and found that $\alpha = 0.1$ work better for CQL. We swept the value of $\tau \in [2.0, 5.0, 7.0]$ and report the sensitivity sweep in Figure 14.

Visual Robotic Domains. For the visual pick and place domains, we follow exactly the same hyperparameters as the CQL implementation from COG [50]: a critic learning rate of 3e-4, an actor learning rate of 1e-4, using $k = 4$ actions from the policy for computing the target values for computing the TD error, and using $k = 4$ actions to compute the log-sum-exp in CQL. For the value of τ, we swept over $\tau \in [0.1, 1.0, 10.0, 100.0]$, and used a $\tau = 1.0$ for our experiments.

G Additional Ablation Studies

In this section, we present some results of an ablation study of the performance of CQL (ReDS) with respect to the temperature hyperparameter τ that appears in Equation 7. Before discussing the results, let us intuitively aim to understand the significance of this hyperparameter. When τ is extremely small we would expect ρ_ψ to be a distribution centered at the worst possible action, within the support of the behavior policy. When τ is large, we would expect the learned ρ_ψ to be close to the behavior policy, since the exponentiated advantage term would essentially behave as a constraint against a uniform distribution. Neither of these extremes are desirable, while the former does not behave much differently than a distributional constraint (except that the Q-value at the action with the smallest Q-value in the dataset support is not pushed up anymore), the latter also behaves like a distributional constraint, but with just half the effective multiplier α on the CQL regularizer. We would therefore expect an intermediate τ to perform the best.

To verify these insights, we study the sensitivity of the performance of CQL (ReDS) with respect to α on the Atari datasets. Our results shown in Figure 15 confirm that indeed an intermediate value of $\tau = 5.0$ out of the tested values, $\tau \in [2.0, 5.0, 7.0]$ works the best.

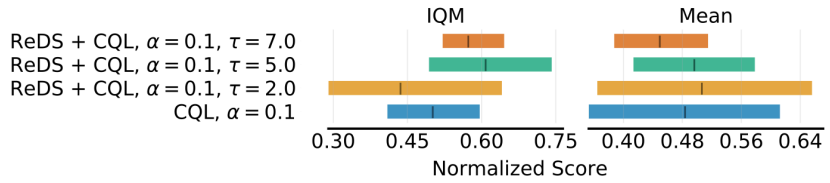

Figure 14: **Sensitivity of CQL (ReDS) to the temperature hyperparameter τ in Equation 7 evaluated on the Atari game experiments with 5 policies.** Observe that an intermediate value of temperature $\tau = 5.0$ works best

In addition, we study the sensitivity of the performance of CQL (ReDS) with respect to α on the Atari datasets. We report the performance for two different values of $\alpha \in \{0.1, 0.2\}$ from CQL (Equation 2) in Figure 15. Observe that CQL (ReDS) with a given α outperforms base CQL for the corresponding α. Additionally note that the degradation in performance of CQL (ReDS) as α increases is lesser than base CQL.

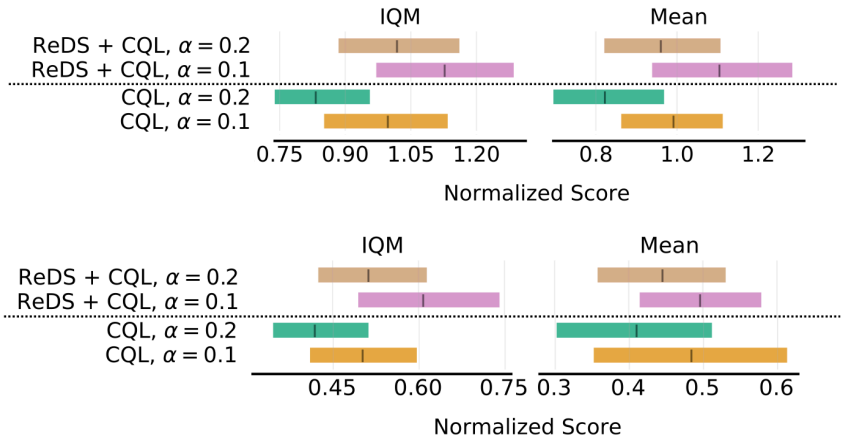

Figure 15: **Sensitivity of ReDS + CQL to the temperature hyperparameter α in Equation 2** We report the performance of CQL (ReDS) vs CQL on the IQM normalized score and the mean normalized score over ten Atari games, for the case of **two** (top) and **five** (bottom) policies. We consider this performance for two different values of $\alpha \in \{0.1, 0.2\}$ in CQL (Equation 2). Observe that CQL (ReDS) with a given α outperforms base CQL for the corresponding α. Additionally note that the degradation in performance of ReDS (CQL) as α increases is lesser than base CQL.

H Additional Baseline Comparison for Heteroskedastic Antmaze Navigation

In this section we will provide additional baseline comparison for REDS with two additional Offline RL methods: EDAC [3] and BEAR [31].

Table 5: EDAC Hyperparameters

Param	Values
N	10, 20, 50, 100
η	0, 1, 5, 10, 50, 100, 1000

Table 6: BEAR Hyperparameters

Param	Values
Kernel	Laplacian, Gaussian
σ	1, 10, 20, 50

H.1 Hyperparameters for EDAC

As done in An et al. [3], we tune the method over two hyperparameters. The first hyperparameter is the ensemble size N which specifies the number of Q functions. The second parameter we consider is η, the weight of the ensemble gradient diversity term. Below in table 5, we show the values considered for each hyperparameter. There is significant overlap to these parameters with the ones used in the Mujoco Gym and Adroit Domains that the authors used. We utilized the publicly available code (`https://github.com/snu-mllab/EDAC`) released by the authors of EDAC and were able to replicate the results they reported for the D4RL MuJoCo Gym environments in An et al. [3].

H.2 Hyperparameters for BEAR

As done in Kumar et al. [31], we tuned this method over two hyperparameters. The first is the Kernel Type of the MMD between the behavior policy π_β and the actor π, and found that Laplacian performed better. The second parameter considered is σ, which is needed for the Laplacian kernel as defined. Below in table 6, we show the values considered for each hyperparameter. There is significant overlap to these parameters with the ones used in the Mujoco Gym and Adroit Domains that the authors used. We utilized the publicly available code (`https://github.com/rail-berkeley/d4rl_evaluations`) released by the authors of BEAR and were able to replicate the results they reported for the D4RL MuJoCo Gym environments in Kumar et al. [31].