Conservative Data Sharing for Multi-Task Offline Reinforcement Learning

Tianhe Yu[*,1,2], **Aviral Kumar**[*,2,3], **Yevgen Chebotar**[2], **Karol Hausman**[1,2], **Sergey Levine**[2,3], **Chelsea Finn**[1,2]

[1]Stanford University, [2]Google Research, [3]UC Berkeley ([*]Equal Contribution)

tianheyu@cs.stanford.edu, aviralk@berkeley.edu

Abstract

Offline reinforcement learning (RL) algorithms have shown promising results in domains where abundant pre-collected data is available. However, prior methods focus on solving individual problems from scratch with an offline dataset without considering how an offline RL agent can acquire multiple skills. We argue that a natural use case of offline RL is in settings where we can pool large amounts of data collected in various scenarios for solving different tasks, and utilize all of this data to learn behaviors for all the tasks more effectively rather than training each one in isolation. However, sharing data across all tasks in multi-task offline RL performs surprisingly poorly in practice. Thorough empirical analysis, we find that sharing data can actually exacerbate the distributional shift between the learned policy and the dataset, which in turn can lead to divergence of the learned policy and poor performance. To address this challenge, we develop a simple technique for data-sharing in multi-task offline RL that routes data based on the improvement over the task-specific data. We call this approach conservative data sharing (CDS), and it can be applied with multiple single-task offline RL methods. On a range of challenging multi-task locomotion, navigation, and vision-based robotic manipulation problems, CDS achieves the best or comparable performance compared to prior offline multi-task RL methods and previous data sharing approaches.

1 Introduction

Recent advances in offline reinforcement learning (RL) make it possible to train policies for real-world scenarios, such as robotics [32, 60, 33] and healthcare [24, 67, 35], entirely from previously collected data. Many realistic settings where we might want to apply offline RL are inherently *multi-task* problems, where we want to solve multiple tasks using all of the data available. For example, if our goal is to enable robots to acquire a range of different behaviors, it is more practical to collect a modest amount of data for each desired behavior, resulting in a large but heterogeneous dataset, rather than requiring a large dataset for every individual skill. Indeed, many existing datasets in robotics [17, 11, 66] and offline RL [19] include data collected in precisely this way. Unfortunately, leveraging such heterogeneous datasets leaves us with two unenviable choices. We could train each task only on data collected for that task, but such small datasets may be inadequate for good performance. Alternatively, we could combine all of the data together and use data relabeled from other tasks to improve offline training, but this naïve data sharing approach can actually often degrade performance over simple single-task training in practice [33]. In this paper, we aim to understand how data sharing affects RL performance in the offline setting and develop a reliable and effective method for selectively sharing data across tasks.

A number of prior works have studied multi-task RL in the *online* setting, confirming that multi-tasking can often lead to performance that is worse than training tasks individually [56, 62, 90].

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

These prior works focus on mitigating optimization challenges that are aggravated by the online data generation process [64, 89, 88]. As we will find in Section 4, multi-task RL remains a challenging problem in the offline setting when sharing data across tasks, even when exploration is not an issue. While prior works have developed heuristic methods for reweighting and relabeling data [3, 16, 44, 33], they do not yet provide a principled explanation for why data sharing can hurt performance in the offline setting, nor do they provide a robust and general approach for selective data sharing that alleviates these issues while preserving the efficiency benefits of sharing experience across tasks.

In this paper, we hypothesize that data sharing can be harmful or brittle in the offline setting because it can exacerbate the distribution shift between the policy represented in the data and the policy being learned. We analyze the effect of data sharing in the offline multi-task RL setting, and present evidence to support this hypothesis. Based on this analysis, we then propose an approach for selective data sharing that aims to minimize distributional shift, by sharing only data that is particularly relevant to each task. Instantiating a method based on this principle requires some care, since we do not know a priori which data is most relevant for a given task before we've learned a good policy for that task. To provide a practical instantiation, we propose the conservative data sharing (CDS) algorithm. CDS reduces distributional shift by sharing data based on a learned conservative estimate of the Q-values that penalizes Q-values on out-of-distribution actions. Specifically, CDS relabels transitions when the conservative Q-value of the added transitions exceeds the expected conservative Q-values on the target task data. We visualize how CDS works in Figure 1.

The main contributions of this work are an analysis of data sharing in offline multi-task RL and a new algorithm, *conservative data sharing* (CDS), for multi-task offline RL problems. CDS relabels a transition into a given task only when it is expected to improve performance based on a conservative estimate of the Q-function. After data sharing, similarly to prior offline RL methods, CDS applies a standard conservative offline RL algorithm, such as CQL [39], that learns a conservative value function or BRAC [82], a policy-constraint offline RL algorithm. Further, we theoretically analyze CDS and characterize scenarios under which it provides safe policy improvement guarantees. Finally, we conduct extensive empirical analysis of CDS on multi-task locomotion, multi-task robotic manipulation with sparse rewards, multi-task navigation, and multi-task imaged-based robotic manipulation. We compare CDS to vanilla offline multi-task RL without sharing data, to naïvely sharing data for all tasks, and to existing data relabeling schemes for multi-task RL. CDS is the only method to attain good performance across all of these benchmarks, often significantly outperforming the best *domain-specific* method, improving over the next best method on each domain by **17.5%** on average.

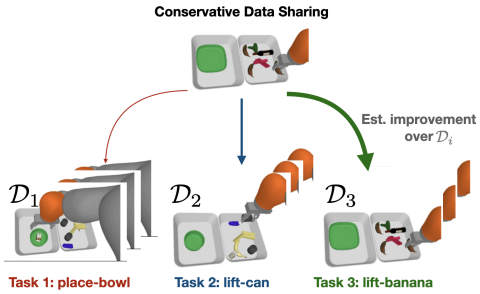

Figure 1: A visualization of CDS, which routes a transition to the offline dataset $\mathcal{D}_i$ for each task i with a weight based on the estimated improvement over the behavior policy $\pi_\beta(\mathbf{a}|\mathbf{s}, i)$ of $\mathcal{D}_i$ after sharing the transition.

2 Related Work

Offline RL. Offline RL [14, 61, 40, 43] has shown promise in domains such as robotic manipulation [32, 52, 60, 70, 33], NLP [29, 30], recommender systems & advertising [72, 22, 7, 78, 79], and healthcare [67, 80]. The major challenge in offline RL is distribution shift [20, 37, 39], where the learned policy might generate out-of-distribution actions, resulting in erroneous value backups. Prior offline RL methods address this issue by regularizing the learned policy to be "close" to the behavior policy [20, 50, 29, 82, 93, 37, 68, 57], through variants of importance sampling [59, 74, 49, 75, 54], via uncertainty quantification on Q-values [2, 37, 82, 43], by learning conservative Q-functions [39, 36], and with model-based training with a penalty on out-of-distribution states [34, 91, 53, 4, 76, 60, 42, 92]. While current benchmarks in offline RL [19, 25] contain datasets that involve multi-task structure, existing offline RL methods do not leverage the shared structure of multiple tasks and instead train each individual task from scratch. In this paper, we exploit the shared structure in the offline multi-task setting and train a general policy that can acquire multiple skills.

Multi-task RL algorithms. Multi-task RL algorithms [81, 56, 77, 15, 27, 89, 85, 88, 33, 71] focus on solving multiple tasks jointly in an efficient way. While multi-task RL methods seem to provide a

promising way to build general-purpose agents [33], prior works have observed major challenges in multi-task RL, in particular, the optimization challenge [27, 64, 89]. Beyond the optimization challenge, how to perform effective representation learning via weight sharing is another major challenge in multi-task RL. Prior works have considered distilling per-task policies into a single policy that solves all tasks [62, 77, 23, 85], separate shared and task-specific modules with theoretical guarantees [13], and incorporating additional supervision [71]. Finally, sharing data across tasks emerges as a challenge in multi-task RL, especially in the off-policy setting, as naïvely sharing data across all tasks turns out to hurt performance in certain scenarios [33]. Unlike most of these prior works, we focus on the offline setting where the challenges in data sharing are most relevant. Methods that study optimization and representation learning issues are complementary and can be readily combined with our approach.

Data sharing in multi-task RL. Prior works [3, 31, 58, 63, 16, 44, 33, 8] have found it effective to reuse data across tasks by recomputing the rewards of data collected for one task and using such relabeled data for other tasks, which effectively augments the amount of data available for learning each task and boosts performance. These methods perform relabeling either uniformly [33] or based on metrics such as estimated Q-values [16, 44], domain knowledge [33], the distance to states or images in goal-conditioned settings [3, 58, 55, 48, 73, 47, 28, 51, 87, 8], and metric learning for robust inference in the offline meta-RL setting [45]. All of these methods either require online data collection and do not consider data sharing in a fully offline setting, or only consider offline goal-conditioned or meta-RL problems [8, 45]. While these prior works empirically find that data sharing helps, we believe that our analysis in Section 4 provides the first analytical understanding of why and when data sharing can help in multi-task offline RL and why it hurts in some cases. Specifically, our analysis reveals the effect of distributional shift introduced during data sharing, which is not taken into account by these prior works. Our proposed approach, CDS, tackles the challenge of distributional shift in data sharing by intelligently sharing data across tasks and improves multi-task performance by effectively trading off between the benefits of data sharing and the harms of excessive distributional shift.

3 Preliminaries and Problem Statement

Multi-task offline RL. The goal in multi-task RL is to find a policy that maximizes expected return in a multi-task Markov decision process (MDP), defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, \{R_i, i\}_{i=1}^N)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, dynamics $P(\mathbf{s}'|\mathbf{s}, \mathbf{a})$, a discount factor $\gamma \in [0, 1)$, and a finite set of task indices $1, \cdots, N$ with corresponding reward functions $R_1, \cdots, R_N$. Each task i presents a different reward function R_i, but we assume that the dynamics P are shared across tasks. While this setting is not fully general, there are a wide variety of practical problem settings for which only the reward changes including various goal navigation tasks [19], distinct object manipulation objectives [83], and different user preferences [10]. In this work, we focus on learning a policy $\pi(\mathbf{a}|\mathbf{s}, i)$, which in practice could be modelled as independent policies $\{\pi_1(\mathbf{a}|\mathbf{s}), \cdots, \pi_N(\mathbf{a}|\mathbf{s})\}$ that do not share any parameters, or as a single task-conditioned policy, $\pi(\mathbf{a}|\mathbf{s}, i)$ with parameter sharing. Our goal in this paper is to analyze and devise methods for data sharing and the choice of parameter sharing is orthogonal, and can be made independently. We formulate the policy optimization problem as finding a policy that maximizes expected return over all the tasks: $\pi^*(\mathbf{a}|\mathbf{s}, \cdot) := \arg\max_\pi \mathbb{E}_{i\sim[N]}\mathbb{E}_{\pi(\cdot|\cdot,i)}[\sum_t \gamma^t R_i(\mathbf{s}_t, \mathbf{a}_t)]$. The Q-function, $Q^\pi(\mathbf{s}, \mathbf{a}, i)$, of a policy $\pi(\cdot|\cdot, i)$ is the long-term discounted reward obtained in task i by executing action $\mathbf{a}$ at state $\mathbf{s}$ and following policy π thereafter.

Standard offline RL is concerned with learning policies $\pi(\mathbf{a}|\mathbf{s})$ using only a given static dataset of transitions $\mathcal{D} = \{(\mathbf{s}_j, \mathbf{a}_j, \mathbf{s}'_j, r_j)\}_{j=1}^N$, collected by a behavior policy $\pi_\beta(\mathbf{a}|\mathbf{s})$, without any additional environment interaction. In the multi-task offline RL setting, the dataset $\mathcal{D}$ is partitioned into per-task subsets, $\mathcal{D} = \cup_{i=1}^N \mathcal{D}_i$, where $\mathcal{D}_i$ consists of experience from task i. While algorithms can choose to train the policy for task i (i.e., $\pi(\cdot|\cdot, i)$) only on $\mathcal{D}_i$, in this paper, we are interested in data-sharing schemes that correspond to relabeling data from a different task, $j \neq i$ with the reward function r_i, and learn $\pi(\cdot|\cdot, i)$ on the combined data. To be able to do so, we assume access to the functional form of the reward r_i, a common assumption in goal-conditioned RL [3, 16], and which often holds in robotics applications through the use of learned classifiers [83, 32], and discriminators [18, 9].

We assume that relabeling data $\mathcal{D}_j$ from task j to task i generates a dataset $\mathcal{D}_{j\to i}$, which is then additionally used to train on task i. Thus, the effective dataset for task i after relabeling is given by $\mathcal{D}_i^{\text{eff}} := \mathcal{D}_i \cup (\cup_{j\neq i}\mathcal{D}_{j\to i})$. This notation simply formalizes data sharing and relabeling strategies

explored in prior work [16, 33]. Our aim in this paper will be to improve on this naïve strategy, which we will show leads to significantly better results.

Offline RL algorithms. A central challenge in offline RL is distributional shift: differences between the learned policy and the behavior policy can lead to erroneous target values, where the Q-function is queried at actions $\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})$ that are far from the actions it is trained on, leading to massive overestimation [43, 37]. A number of offline RL algorithms use some kind of regularization on either the policy [37, 20, 82, 29, 68, 57] or on the learned Q-function [39, 36] to ensure that the learned policy does not deviate too far from the behavior policy. For our analysis in this work, we will abstract these algorithms into a generic constrained policy optimization problem [39]:

$$\pi^*(\mathbf{a}|\mathbf{s}) := \arg\max_{\pi} \; J_{\mathcal{D}}(\pi) - \alpha D(\pi, \pi_\beta). \tag{1}$$

$J_{\mathcal{D}}(\pi)$ denotes the average return of policy π in the empirical MDP induced by the transitions in the dataset, and $D(\pi, \pi_\beta)$ denotes a divergence measure (e.g., KL-divergence [29, 82], MMD distance [37] or D_{CQL} [39]) between the learned policy π and the behavior policy π_β. In the multi-task offline RL setting with data-sharing, the generic optimization problem in Equation 1 for a task i utilizes the effective dataset $\mathcal{D}_i^{\text{eff}}$. In addition, we define $\pi_\beta^{\text{eff}}(\mathbf{a}|\mathbf{s}, i)$ as the effective behavior policy for task i and it is given by: $\pi_\beta^{\text{eff}}(\mathbf{a}|\mathbf{s}, i) := |\mathcal{D}_i^{\text{eff}}(\mathbf{s}, \mathbf{a})|/|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$. Hence, the counterpart of Equation 1 in the multi-task offline RL setting with data sharing is given by:

$$\forall i \in [N], \;\; \pi^*(\mathbf{a}|\mathbf{s}, i) := \arg\max_{\pi} \; J_{\mathcal{D}_i^{\text{eff}}}(\pi) - \alpha D(\pi, \pi_\beta^{\text{eff}}). \tag{2}$$

We will utilize this generic optimization problem to motivate our method in Section 5.

4 When Does Data Sharing Actually Help in Offline Multi-Task RL?

Our goal is to leverage experience from all tasks to learn a policy for a particular task of interest. Perhaps the simplest approach to leveraging experience across tasks is to train the task policy on not just the data coming from that task, but also relabeled data from all other tasks [6]. Is this naïve data sharing strategy sufficient for learning effective behaviors from multi-task offline data? In this section, we aim to answer this question via empirical analysis on a relatively simple domain, which will reveal interesting aspects of data sharing. We first describe the experimental setup and then discuss the results and possible explanations for the observed behavior. Using insights obtained from this analysis, we will then derive a simple and effective data sharing strategy in Section 5.

Experimental analysis setup. To assess the efficacy of data sharing, we experimentally analyze various multi-task RL scenarios created with the walker2d environment in Gym [5]. We construct different test scenarios on this environment that mimic practical situations, including settings where different amounts of data of varied quality are available for different tasks [33, 84, 69]. In all these scenarios, the agent attempts three tasks: `run forward`, `run backward`, and `jump`, which we visualize in Figure 3. Following the problem statement in Section 3, these tasks share the same state-action space and transition dynamics, differing only in the reward function that the agent is trying to optimize. Different scenarios are generated with varying size offline datasets, each collected with policies that have different degrees of suboptimality. This might include, for each task, a single policy with mediocre or expert performance, or a mixture of policies given by the initial part of the replay buffer trained with online SAC [26]. We refer to these three types of offline datasets as medium, expert and medium-replay, respectively, following Fu et al. [19].

We train a single-task policy $\pi_{\text{CQL}}(\mathbf{a}|\mathbf{s}, i)$ with CQL [39] as the base offline RL method, along with two forms of data-sharing, as shown in Table 1: no sharing of data across tasks (**No Sharing**)) and complete sharing of data with relabeling across all tasks (**Sharing All**). In addition, we also measure the divergence term in Equation 2, $D(\pi(\cdot|\cdot, i), \pi_\beta^{\text{eff}}(\cdot|\cdot, i))$, for $\pi = \pi_{\text{CQL}}(\mathbf{a}|\mathbf{s}, i)$, averaged across tasks by using the Kullback-Liebler divergence. This value quantifies the average divergence between the single-task optimal policy and the relabeled behavior policy averaged across tasks.

Analysis of results in Table 1. To begin, note that even naïvely sharing data is better than not sharing any data at all on **5/9** tasks considered (compare the performance across **No Sharing** and **Sharing All** in Table 1). However, a closer look at Table 1 suggests that data-sharing can significantly degrade performance on certain tasks, especially in scenarios where the amount of data available for the original task is limited, and where the distribution of this data is narrow. For example, when using

Dataset types / Tasks	Dataset Size	Avg Return		$D_{\text{KL}}(\pi, \pi_\beta)$	
		No Sharing	**Sharing All**	**No Sharing**	**Sharing All**
medium-replay / `run forward`	109900	**998.9**	966.2	**3.70**	10.39
medium-replay / `run backward`	109980	**1298.6**	1147.5	**4.55**	12.70
medium-replay / `jump`	109511	**1603.1**	1224.7	**3.57**	15.89
average task performance	N/A	**1300.2**	1112.8	**3.94**	12.99
medium / `run forward`	27646	297.4	**848.7**	**6.53**	11.78
medium / `run backward`	31298	207.5	**600.4**	**4.44**	10.13
medium / `jump`	100000	351.1	**776.1**	**5.57**	21.27
average task performance	N/A	285.3	**747.7**	**5.51**	14.39
medium-replay / `run forward`	109900	590.1	**701.4**	**1.49**	7.76
medium / `run backward`	31298	614.7	**756.7**	**1.91**	12.2
expert / `jump`	5000	**1575.2**	885.1	**3.12**	27.5
average task performance	N/A	**926.6**	781	**2.17**	15.82

Table 1: We analyze how sharing data across all tasks (**Sharing All**) compares to **No Sharing** in the multi-task walker2d environment with three tasks: run forward, run backward, and jump. We provide three scenarios with different styles of per-task offline datasets in the leftmost column. The second column shows the number of transitions in each dataset. We report the per-task average return, the KL divergence between the single-task optimal policy π and the behavior policy π_β after the data sharing scheme, as well as averages across tasks. **Sharing All** generally helps training while increasing the KL divergence. However, on the row highlighted in yellow, **Sharing All** yields a particularly large KL divergence between the single-task π and π_β and degrades the performance, suggesting sharing data for all tasks is brittle.

expert data for jumping in conjunction with more than 25 times as much lower-quality (mediocre & random) data for running forward and backward, we find that the agent performs poorly on the jumping task despite access to near-optimal jumping data.

Why does naïve data sharing degrade performance on certain tasks despite near-optimal behavior for these tasks in the original task dataset? We argue that the primary reason that naïve data sharing can actually hurt performance in such cases is because it exacerbates the distributional shift issues that afflict offline RL. Many offline RL methods combat distribution shift by implicitly or explicitly constraining the learned policy to stay close to the training data. Then, when the training data is changed by adding relabeled data from another task, the constraint causes the learned policy to change as well. When the added data is of low quality for that task, it will correspondingly lead to a lower quality learned policy for that task, unless the constraint is somehow modified. This effect is evident from the higher divergence values between the learned policy without any data-sharing and the effective behavior policy for that task *after* relabeling (e.g., expert+`jump`) in Table 1. Although these results are only for CQL, we expect that any offline RL method would, insofar as it combats distributional shift by staying close to the data, would exhibit a similar problem.

To mathematically quantify the effects of data-sharing in multi-task offline RL, we appeal to safe policy improvement bounds [41, 39, 92] and discuss cases where data-sharing between tasks i and j can degrade the amount of worst-case guaranteed improvement over the behavior policy. Prior work [39] has shown that the generic offline RL algorithm in Equation 1 enjoys the following guarantees of policy improvement on the actual MDP, beyond the behavior policy:

$$J(\pi^*) \geq J(\pi_\beta) - \mathcal{O}(1/(1-\gamma)^2)\mathbb{E}_{\mathbf{s},\mathbf{a}\sim d^\pi}\left[\sqrt{\frac{D(\pi(\cdot|\mathbf{s}), \pi_\beta(\cdot|\mathbf{s}))}{|\mathcal{D}(\mathbf{s})|}}\right] + \alpha/(1-\gamma)D(\pi, \pi_\beta). \quad (3)$$

We will use Equation 3 to understand the scenarios where data sharing can hurt. When data sharing modifies $\mathcal{D} = \mathcal{D}_i$ to $\mathcal{D} = \mathcal{D}_i^{\text{eff}}$, which includes $\mathcal{D}_i$ as a subset, it effectively aims at reducing the magnitude of the second term (i.e., sampling error) by increasing the denominator. This can be highly effective if the state distribution of the learned policy π^* and the dataset $\mathcal{D}$ overlap. However, an increase in the divergence $D(\pi(\cdot|\mathbf{s}), \pi^\beta(\cdot|\mathbf{s}))$ as a consequence of relabeling implies a potential increase in the sampling error, unless the increased value of $|\mathcal{D}^{\text{eff}}(\mathbf{s})|$ compensates for this. Additionally, the bound also depends on the quality of the behavior data added after relabeling: if the resulting behavior policy π_β^{eff} is more suboptimal compared to π_β, i.e., $J(\pi_\beta^{\text{eff}}) < J(\pi_\beta)$, then the guaranteed amount of improvement also reduces.

To conclude, our analysis reveals that while data sharing is often helpful in multi-task offline RL, it can lead to substantially poor performance on certain tasks as a result of exacerbated distributional shift between the optimal policy and the effective behavior policy induced after sharing data.

5 CDS: Reducing Distributional Shift in Multi-Task Data Sharing

The analysis in Section 4 shows that naïve data sharing may be highly sub-optimal in some cases, and although it often does improve over no data sharing at all in practice, it can also lead to exceedingly poor performance. Can we devise a conservative approach that shares data intelligently to not exacerbate distributional shift as a result of relabeling?

5.1 A First Attempt at Designing a Data Sharing Strategy

A straightforward data sharing strategy is to utilize a transition for training only if it reduces the distributional shift. Formally, this means that for a given transition $(\mathbf{s}, \mathbf{a}, r_j(\mathbf{s}, \mathbf{a}), \mathbf{s}') \in \mathcal{D}_j$ sampled from the dataset $\mathcal{D}_j$, such a scheme would prescribe using it for training task i (i.e., $(\mathbf{s}, \mathbf{a}, r_i(\mathbf{s}, \mathbf{a}), \mathbf{s}') \in \mathcal{D}_i^{\text{eff}}$) only if:

$$\textbf{CDS (basic)}: \quad \Delta^{\pi}(\mathbf{s}, \mathbf{a}) := D(\pi(\cdot|\cdot, i), \pi_\beta(\cdot|\cdot, i))(\mathbf{s}) - D(\pi(\cdot|\cdot, i), \pi_\beta^{\text{eff}}(\cdot|\cdot, i))(\mathbf{s}) \geq 0. \quad (4)$$

The scheme presented in Equation 4 would guarantee that distributional shift (i.e., second term in Equation 2) is reduced. Moreover, since sharing data can only increase the size of the dataset and not reduce it, this scheme is guaranteed to not increase the sampling error term in Equation 3. We refer to this scheme as the basic variant of conservative data sharing (**CDS (basic)**).

While this scheme can prevent the negative effects of increased distributional shift, this scheme is quite pessimistic. Even in our experiments, we find that this variant of CDS does not improve performance by a large margin. Additionally, as observed in Table 1 (medium-medium-medium data composition) and discussed in Section 4, data sharing can often be useful despite an increased distributional shift (note the higher values of $D_{\text{KL}}(\pi, \pi_\beta)$ in Table 1) likely because it reduces sampling error and potentially utilizes data of higher quality for training. **CDS (basic)** described above does not take into account these factors. Formally, the effect of the first term in Equation 2, $J_{\mathcal{D}^{\text{eff}}}(\pi)$ (the policy return in the empirical MDP generated by the dataset) and a larger increase in $|\mathcal{D}^{\text{eff}}(\mathbf{s})|$ at the cost of somewhat increased value of $D(\pi(\cdot|\mathbf{s}), \pi_\beta(\cdot|\mathbf{s})$ are not taken into account. Thus we ask: can we instead design a more complete version of CDS that effectively balances the tradeoff by incorporating all the discussed factors (distributional shift, sampling error, data quality)?

5.2 The Complete Version of Conservative Data Sharing (CDS)

Next, we present the complete version of our method. The complete version of CDS, which we will refer to as **CDS**, for notational brevity is derived from the following perspective: we note that a data sharing scheme can be viewed as altering the dataset $\mathcal{D}_i^{\text{eff}}$, and hence the effective behavior policy, $\pi_\beta^{\text{eff}}(\mathbf{a}|\mathbf{s}, i)$. Thus, we can directly *optimize* the objective in Equation 2 with respect to π_β^{eff}, in addition to π, where π_β^{eff} belongs to the set of all possible effective behavior policies that can be obtained via any form of data sharing. Note that unlike CDS (basic), this approach would not rely on only indirectly controlling the objective in Equation 2 by controlling distributional shift, but would aim to directly optimize the objective in Equation 2. We formalize this optimization below in Equation 5:

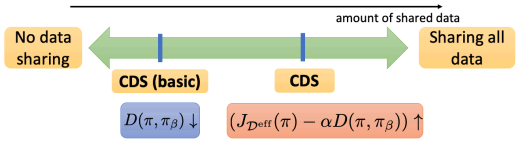

Figure 2: A schematic comparing **CDS** and **CDS (basic)** data sharing schemes relative to no sharing (left extreme) and full data sharing (right extreme). While p-CDS only shares data when distributional shift is strictly reduced, o-CDS is more optimistic and shares data when the objective in Equation 2 is larger. Typically, we would expect that CDS shares more transitions than CDS (basic).

$$\arg\max_{\pi} \max_{\pi_\beta^{\text{eff}} \in \Pi_{\text{relabel}}} \left[J_{\mathcal{D}_i^{\text{eff}}}(\pi) - \alpha D(\pi, \pi_\beta^{\text{eff}}; i) \right], \quad (5)$$

where Π_{relabel} denotes the set of all possible behavior policies that can be obtained via relabeling. The next result characterizes safe policy improvement for Equation 5 and discusses how it leads to improvement over the behavior policy and also produces an effective practical method.

Proposition 5.1 (Characterizing safe-policy improvement for CDS.)**.** *Let $\pi^*(\mathbf{a}|\mathbf{s})$ be the policy obtained by optimizing Equation 5, and let $\pi_\beta(\mathbf{a}|\mathbf{s})$ be the behavior policy for $\mathcal{D}_i$. Then, w.h.p.*

$\geq 1-\delta$, *π^* is a ζ-safe policy improvement over π_β, i.e., $J(\pi^*) \geq J(\pi_\beta) - \zeta$, where ζ is given by:*

$$\zeta = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right) \mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}} \left[\sqrt{\frac{D_{CQL}(\pi^*, \pi^*_\beta)(\mathbf{s}) + 1}{|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|}}\right] - \left[\alpha D(\pi^*, \pi^*_\beta) + \underbrace{J(\pi^*_\beta) - J(\pi_\beta)}_{(a)}\right],$$

where $\mathcal{D}_i^{\text{eff}} \sim d^{\pi^*_\beta}(\mathbf{s})$ and $\pi^*_\beta(\mathbf{a}|\mathbf{s})$ denotes the policy $\pi \in \Pi_{relabel}$ that maximizes Equation 5.

A proof and analysis of this proposition is provided in Appendix B, where we note that the bound in Proposition 5.1 is stronger than both no data sharing as well as naïve data sharing. We show in Appendix B that optimizing Equation 5 reduces the numerator $D_{\text{CQL}}(\pi^*, \pi^*_\beta)$ term while also increasing $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$, thus reducing the amount of sampling error. In addition, Lemma B.1 shows that the improvement term (a) is guaranteed to be positive if a large enough α is chosen in Equation 5. Combining these, we find data sharing using Equation 5 improves over both complete data sharing (which may increase $D_{\text{CQL}}(\pi, \pi_\beta)$) and no data sharing (which does not increase $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$). A schematic comparing the two variants of CDS and naïve and no data sharing schemes is shown in Figure 2.

Optimizing Equation 5 tractably. The next step is to effectively convert Equation 5 into a simple condition for data sharing in multi-task offline RL. While directly solving Equation 5 is intractable in practice, since both the terms depend on $\pi_\beta^{\text{eff}}(\mathbf{a}|\mathbf{s})$ (since the first term $J_{\mathcal{D}^{\text{eff}}}(\pi)$ depends on the empirical MDP induced by the effective behavior policy and the amount of sampling error), we need to instead solve Equation 5 approximately. Fortunately, we can optimize a *lower-bound approximation* to Equation 5 that uses the dataset state distribution for the policy update in Equation 5 similar to modern actor-critic methods [12, 46, 21, 26, 39] which only introduces an additional $D(\pi, \pi_\beta)$ term in the objective. This objective is given by: $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_i^{\text{eff}}}[\mathbb{E}_\pi[Q(\mathbf{s}, \mathbf{a}, i)] - \alpha' D(\pi(\cdot|\mathbf{s}, i), \pi_\beta^{\text{eff}}(\cdot|\mathbf{s}, i))]$, which is equal to the expected "conservative Q-value" $\hat{Q}^\pi(\mathbf{s}, \mathbf{a}, i)$ on dataset states, policy actions and task i. Optimizing this objective via a co-ordinate descent on π and π_β^{eff} dictates that π be updated using a standard update of maximizing the conservative Q-function, $\hat{Q}^\pi$ (equal to the difference of the Q-function and $D(\pi, \pi_\beta^{\text{eff}}; i)$). Moreover, π_β^{eff} should also be updated towards maximizing the same expectation, $\mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}_i^{\text{eff}}}[\hat{Q}^\pi(\mathbf{s}, \mathbf{a}, i)] := \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}_i^{\text{eff}}}[Q(\mathbf{s}, \mathbf{a}, i)] - \alpha D(\pi, \pi_\beta^{\text{eff}}; i)$. This implies that when updating the behavior policy during relabeling, we should prefer state-action pairs that maximize the conservative Q-function.

Deriving the data sharing strategy for CDS. Utilizing the insights for optimizing Equation 5 tractably as discussed above, we now present the effective data sharing rule prescribed by CDS. For any given task i, we want relabeling to incorporate transitions with the highest conservative Q-value into the resulting dataset $\mathcal{D}_i^{\text{eff}}$, as this will directly optimize the tractable lower bound on Equation 5. While directly optimizing Equation 5 will enjoy benefits of reduced sampling error since $J_{\mathcal{D}_i^{\text{eff}}}(\pi)$ also depends on sampling error, our tractable lower bound approximation does not enjoy this benefit. This is because optimizing the lower-bound only increases the frequency of a state in the dataset, $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$ by atmost 1. To encourage further reduction in sampling error, we modify CDS to instead share all transitions with a conservative Q-value more than the top k^{th} quantile of the original dataset $\mathcal{D}_i$, where k is a hyperparameter. This provably increases the objective value in Equation 5 still ensuring that term $(a) > 0$ in Proposition 5.1, while also reducing $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$ in the denominator. Thus, for a given transition $(\mathbf{s}, \mathbf{a}, \mathbf{s}') \in \mathcal{D}_j$,

$$\textbf{CDS: } (\mathbf{s}, \mathbf{a}, r_i, \mathbf{s}') \in \mathcal{D}_i^{\text{eff}} \text{ if } \Delta^\pi(\mathbf{s}, \mathbf{a}) := \hat{Q}^\pi(\mathbf{s}, \mathbf{a}, i) - P_{k\%}\left\{\hat{Q}^\pi(\mathbf{s}', \mathbf{a}', i) \colon \mathbf{s}', \mathbf{a}' \sim \mathcal{D}_i\right\} \geq 0, \tag{6}$$

where $\hat{Q}^\pi$ denotes the learned conservative Q-function estimate. If the condition in Equation 6 holds for the given $(\mathbf{s}, \mathbf{a})$, then the corresponding relabeled transition, $(\mathbf{s}, \mathbf{a}, r_i(\mathbf{s}, \mathbf{a}), \mathbf{s}')$ is added to $\mathcal{D}_i^{\text{eff}}$.

We summarize the pesudocode of CDS in Algorithm 1 in Appendix A and include the practical implementation details of CDS in Appendix C.

6 Experimental Evaluation

We conduct experiments to answer six main questions: **(1)** can CDS prevent performance degradation when sharing data as observed in Section 4?, **(2)** how does CDS compare to vanilla multi-task offline RL methods and prior data sharing methods? **(3)** can CDS handle sparse reward settings, where data sharing is particularly important due to scarce supervision signal? **(4)** can CDS handle goal-conditioned offline RL settings where the offline dataset is undirected and highly suboptimal? **(5)** Can CDS scale to complex visual observations? **(6)** Can CDS be combined with any offline RL algorithms? Besides these questions, we visualize CDS weights for better interpretation of the data sharing scheme learned by CDS in Figure 4 in Appendix D.2.

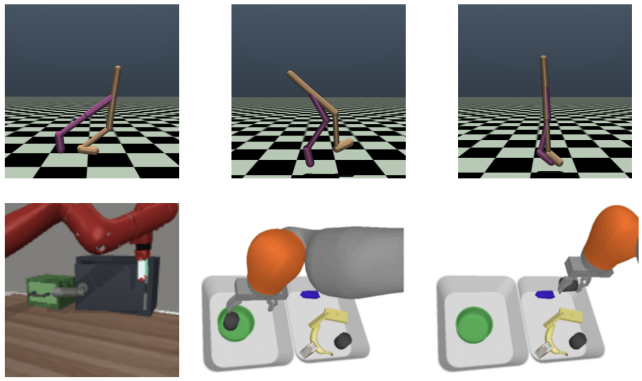

Figure 3: Environments (from left to right): walker2d run forward, walker2d run backward, walker2d jump, Meta-World door open/close and drawer open/close and vision-based pick-place tasks in [33].

Comparisons. To answer these questions, we consider the following prior methods. On tasks with low dimensional state spaces, we compare with the online multi-task relabeling approach **HIPI** [16], which uses inverse RL to infer for which tasks the datapoints are optimal and in practice routes a transition to task with the highest Q-value. We adapt HIPI to the offline setting by applying its data routing strategy to a conservative offline RL algorithm. We also compare to naïvely sharing data across all tasks (denoted as **Sharing All**) and vanilla multi-task offline RL method without any data sharing (denoted as **No Sharing**). On image-based domains, we compare CDS to the data sharing strategy based on human-defined skills [33] (denoted as **Skill**), which manually groups tasks into different skills (e.g. skill "pick" and skill "place") and only routes an episode to target tasks that belongs to the same skill. In these domains, we also compare to **HIPI**, **Sharing All** and **No Sharing**. Beyond these multi-task RL approaches with data sharing, to assess the importance of data sharing in offline RL, we perform an additional comparison to other alternatives to data sharing in multi-task offline RL settings. One traditionally considered approach is to use data from other tasks for some form of "pre-training" before learning to solve the actual task. We instantiate this idea by considering a method from Yang and Nachum [86] that conducts contrastive representation learning on the multi-task datasets to extract shared representation between tasks and then runs multi-task RL on the learned representations. We discuss this comparison in detail in Table 7 in Appendix D.3. To answer question (6), we use CQL [39] (a Q-function regularization method) and BRAC [82] (a policy-constraint method) as the base offline RL algorithms for all methods. We discuss evaluations of CDS with CQL in the main text and include the results of CDS with BRAC in Table 5 in Appendix D.1. For more details on setup and hyperparameters, see Appendix C.

Multi-task environments. We consider a number of multi-task reinforcement learning problems on environments visualized in Figure 3. To answer questions (1) and (2), we consider the walker2d locomotion environment from OpenAI Gym [5] with dense rewards. We use three tasks, `run forward`, `run backward` and `jump`, as proposed in prior offline RL work [91]. To answer question (3), we also evaluate on robotic manipulation domains using environments from the Meta-World benchmark [90]. We consider four tasks: `door open`, `door close`, `drawer open` and `drawer close`. Meaningful data sharing requires a consistent state representation across tasks, so we put both the door and the drawer on the same table, as shown in Figure 3. Each task has a sparse reward of 1 when the success condition is met and 0 otherwise. To answer question (4), we consider maze navigation tasks where the temporal "stitching" ability of an offline RL algorithm is crucial to obtain good performance. We create goal reaching tasks using the ant robot in the medium and hard mazes from D4RL [19]. The set of goals is a fixed discrete set of size 7 and 3 for large and medium mazes, respectively. Following Fu et al. [19], a reward of +1 is given and the episode terminates if the state is within a threshold radius of the goal. Finally, to explore how CDS scales to image-based manipulation tasks (question (5)), we utilize a simulation environment similar to the real-world setup presented in [33]. This environment, which was utilized by Kalashnikov et al. [33]

as a representative and realistic simulation of a real-world robotic manipulation problem, consists of 10 image-based manipulation tasks that involve different combinations of picking specific objects (banana, bottle, sausage, milk box, food box, can and carrot) and placing them in one of the three fixtures (bowl, plate and divider plate) (see example task images in Fig. 3). More environment details are in the appendix. We report the average return for locomotion tasks and success rate for AntMaze and both manipluation environments, averaged over 6 and 3 random seeds for environments with low-dimensional inputs and image inputs respectively.

Environment	Dataset types / Tasks	$D_{\text{KL}}(\pi, \pi_\beta)$			
		No Sharing	**Sharing All**	**CDS (basic) (ours)**	**CDS (ours)**
walker2d	medium-replay / run forward	**1.49**	7.76	14.31	**1.49**
	medium / run backward	**1.91**	12.2	8.26	6.09
	expert / jump	3.12	27.5	13.25	**2.91**

Table 2: Measuring $D_{\text{KL}}(\pi, \pi_\beta)$ on the walker2d environment. **Sharing All** degrades the performance on task jump with limited expert data as discussed in Table 1. CDS manages to obtain a π_β after data sharing that is closer to the single-task optimal policy in terms of the KL divergence compared to **No Sharing** and **Sharing All** on task `jump` (highlighted in yellow). Since CDS also achieves better performance, this analysis suggests that reducing distribution shift is important for effective offline data sharing.

Multi-task datasets. Following the analysis in Section 4, we intentionally construct datasets with a variety of heterogeneous behavior policies to test if CDS can provide effective data sharing to improve performance while avoiding harmful data sharing that exacerbates distributional shift. For the locomotion domain, we use a large, diverse dataset (medium-replay) for `run forward`, a medium-sized dataset for `run backward`, and an expert dataset with limited data for `run jump`. For Meta-World, we consider medium-replay datasets with 152K transitions for task `door close` and `drawer open` and expert datasets with only 2K transitions for task `door open` and `drawer close`. For AntMaze, we modify the D4RL datasets for antmaze-*-play environments to construct two kinds of multi-task datasets: an "undirected" dataset, where data is equally divided between different tasks and the rewards are correspondingly relabeled, and a "directed" dataset, where a trajectory is associated with the goal closest to the final state of the trajectory. This means that the per-task data in the undirected setting may not be relevant to reaching the goal of interest. Thus, data-sharing is crucial for good performance: methods that do not effectively perform data sharing and train on largely task-irrelevant data are expected to perform worse. Finally, for image-based manipulation tasks, we collect datasets for all the tasks individually by running online RL [32] until the task reaches medium-level performance (40% for picking tasks and 80% placing tasks). At that point, we merge the entire replay buffers from different tasks creating a final dataset of 100K RL episodes with 25 transitions for each episode.

Results on domains with low-dimensional states. We present the results on all non-vision environments in Table 3. CDS achieves the best average performance across all environments except that on walker2d, it achieves the second best performance, obtaining slightly worse the other variant CDS (basic). On the locomotion domain, we observe the most significant improvement on task `jump` on all three environments. We interpret this as strength of conservative data sharing, which mitigates the distribution shift that can be introduced by routing large amount of other task data to the task with limited data and narrow distribution. We also validate this by measuring the $D_{\text{KL}}(\pi, \pi_\beta)$ in Table 2 where π_β is the behavior policy after we perform CDS to share data. As shown in Table 2, CDS achieves lower KL divergence between the single-task optimal policy and the behavior policy after data sharing on task `jump` with limited expert data, whereas **Sharing All** results in much higher KL divergence compared to **No Sharing** as discussed in Section 4 and Table 1. Hence, CDS is able to mitigate distribution shift when sharing data and result in performance boost.

On the Meta-World tasks, we find that the agent without data sharing completely fails to solve most of the tasks due to the low quality of the medium replay datasets and the insufficient data for the expert datasets. **Sharing All** improves performance since in the sparse reward settings, data sharing can introduce more supervision signal and help training. CDS further improves over **Sharing All**, suggesting that CDS can not only prevent harmful data sharing, but also lead to more effective multi-task learning compared to **Sharing All** in scenarios where data sharing is imperative. It's worth noting that CDS (basic) performs worse than CDS and **Sharing All**, indicating that relabeling data that only mitigates distributional shift is too pessimistic and might not be sufficient to discover the shared structure across tasks.

In the AntMaze tasks, we observe that CDS performs better than **Sharing All** and significantly outperforms HIPI in all four settings. Perhaps surprisingly, **No Sharing** is a strong baseline, however,

Environment	Tasks / Dataset type	CDS (ours)	CDS (basic)	HIPI [16]	Sharing All	No Sharing
walker2d	run forward / medium-replay	**1057.9**±121.6	968.6±188.6	695.5±61.9	701.4±47.0	590.1±48.6
	run backward / medium	564.8±47.7	594.5±22.7	626.0±48.0	**756.7**±76.7	614.7±87.3
	jump / expert	1418.2±138.4	1501.8±115.1	**1603.7**±146.8	885.1±152.9	1575.2±70.9
	average	1013.6±71.5	**1021.6**±76.9	975.1±45.1	781.0±100.8	926.6±37.7
Meta-World [90]	door open / expert	**58.4%**±9.3%	30.1%±16.6%	26.5%±20.5%	34.3%±17.9%	14.5%±12.7
	door close / medium-replay	**65.3%**±27.7%	41.5%±28.2%	1.3%±5.3%	48.3%±27.3%	4.0%±6.1%
	drawer open / medium-replay	**57.9%**±16.2%	39.4%±16.9%	41.2%±24.9%	55.1%±9.4%	16.0%±17.5%
	drawer close / expert	98.8%±0.7%	86.3%±0.9%	62.2%±33.4%	**100.0%**±0%	99.0%±0.7%
	average	**70.1%**±8.1%	49.3%±16.0%	32.8%±18.7%	59.4%±5.7%	33.4%±8.3%
AntMaze [19]	large maze (7 tasks) / undirected	**22.8**% ± 4.5%	10.0% ± 5.9%	1.3% ± 2.3%	16.7% ± 7.0%	13.3% ± 8.6%
	large maze (7 tasks) / directed	**24.6%** ± 4.7%	0.0% ± 0.0%	11.8% ± 5.4%	20.6% ± 4.4%	19.2% ± 8.0%
	medium maze (3 tasks) / undirected	**36.7%** ± 6.2%	0.0% ± 0.0%	8.6% ± 3.2%	22.9% ± 3.6%	21.6% ± 7.1%
	medium maze (3 tasks) / directed	**18.5**% ± 6.0%	0.0% ± 0.0%	8.3% ± 9.1%	12.4% ± 5.4%	**17.0%** ± 3.2%

Table 3: Results for multi-task locomotion (walker2d), robotic manipulation (Meta-World) and navigation environments (AntMaze) with low-dimensional state inputs. Numbers are averaged across 6 seeds, ± the 95%-confidence interval. We include per-task performance for walker2d and Meta-World domains and the overall performance averaged across tasks (highlighted in gray) for all three domains. We bold the highest score across all methods. CDS achieves the best or comparable performance on all of these environments.

Task Name	CDS (ours)	HIPI [16]	Skill [33]	Sharing All	No Sharing
`lift-banana`	**53.1%**±3.2%	48.3%±6.0%	32.1%±9.5%	41.8%±4.2%	20.0%±6.0%
`lift-bottle`	**74.0%**±6.3%	64.4%±7.7%	55.9%±9.6%	60.1%±10.2%	49.7%±8.7%
`lift-sausage`	**71.8%**±3.9%	71.0%±7.7%	68.8%±9.3%	70.0%±7.0%	60.9%±6.6%
`lift-milk`	**83.4%**±5.2%	79.0%±3.9%	68.2%±3.5%	72.5%±5.3%	68.4%±6.1%
`lift-food`	61.4%±9.5%	**62.6%**±6.3%	41.5%±12.1%	58.5%±7.0%	39.1%±7.0%
`lift-can`	65.5%±6.9%	**67.8%**±6.8%	50.8%±12.5%	57.7%±7.2%	49.1%±9.8%
`lift-carrot`	**83.8%**±3.5%	78.8%±6.9%	66.0%±7.0%	75.2%±7.6%	69.4%±7.6%
`place-bowl`	**81.0%**±8.1%	77.2%±8.9%	80.8%±6.9%	70.8%±7.8%	80.3%±8.6%
`place-plate`	85.8%±6.6%	83.6%±7.9%	78.4%±9.6%	78.7%±7.6%	**86.1**%±7.7%
`place-divider-plate`	**87.8%**±7.6%	78.0%±10.5%	80.8%±5.3%	79.2%±6.3%	85.0%±5.9%
average	**74.8%**±6.4%	71.1%±7.5%	62.3%±8.9%	66.4%±7.2%	60.8%±7.5%

Table 4: Results for multi-task vision-based robotic manipulation domains in [33]. Numbers are averaged across 3 seeds, ± the 95% confidence interval. We consider 7 tasks denoted as `lift-object` where the goal of each task is to lift a different object and 3 tasks denoted as `place-fixture` that aim to place a lifted object onto different fixtures. CDS outperforms both a skill-based data sharing strategy [33] (**Skill**) and other data sharing methods on the average task success rate (highlighted in gray) and 7 out of 10 per-task success rates.

is outperformed by CDS with the harder undirected data. Moreover, CDS performs on-par or better in the undirected setting compared to the directed setting, indicating the effectiveness of CDS in routing data in challenging settings.

Results on image-based robotic manipulation domains. Here, we compare CDS to the hand-designed **Skill** sharing strategy, in addition to the other methods. Given that CDS achieves significantly better performance than CDS (basic) on low-dimensional robotic manipulation tasks in Meta-World, we only evaluate CDS in the vision-based robotic manipulation domains. Since CDS is applicable to any offline multi-task RL algorithm, we employ it as a separate data-sharing strategy in [33] while keeping the model architecture and all the other hyperparameters constant, which allows us to carefully evaluate the influence of data sharing in isolation. The results are reported in Table 4. CDS outperforms both **Skill** and other approaches, indicating that CDS is able to scale to high-dimensional observation inputs and can effectively remove the need for manual curation of data sharing strategies.

7 Conclusion

In this paper, we study the multi-task offline RL setting, focusing on the problem of sharing offline data across tasks for better multi-task learning. Through empirical analysis, we identify that naïvely sharing data across tasks generally helps learning but can significantly hurt performance in scenarios where excessive distribution shift is introduced. To address this challenge, we present conservative data sharing (CDS), which relabels data to a task when the conservative Q-value of the given transition is better than the expected conservative Q-value of the target task. On multitask locomotion, manipulation, navigation, and vision-based manipulation domains, CDS consistently outperforms or achieves comparable performance to existing data sharing approaches. While CDS attains superior results, it is not able to handle data sharing in settings where dynamics vary across tasks and requires functional forms of rewards. We leave these as future work.

Acknowledgements

We thank Kanishka Rao, Xinyang Geng, Avi Singh, other members of RAIL at UC Berkeley, IRIS at Stanford and Robotics at Google and anonymous reviewers for valuable and constructive feedback on an early version of this manuscript. This research was funded in part by Google, ONR grants N00014-20-1-2675 and N00014-21-1-2685, Intel Corporation and the DARPA Assured Autonomy Program. CF is a CIFAR Fellow in the Learning in Machines and Brains program.

References

[1] Abbas Abdolmaleki, Jost Tobias Springenberg, Y. Tassa, R. Munos, N. Heess, and Martin A. Riedmiller. Maximum a posteriori policy optimisation. *ArXiv*, abs/1806.06920, 2018.

[2] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR, 2020.

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *arXiv preprint arXiv:1707.01495*, 2017.

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

[5] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[6] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[7] Denis Charles, Max Chickering, and Patrice Simard. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14, 2013.

[8] Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, and Sergey Levine. Actionable models: Unsupervised offline reinforcement learning of robotic skills. *arXiv preprint arXiv:2104.07749*, 2021.

[9] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[10] Paul Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017.

[11] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning, 2020.

[12] Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

[13] Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*, 2019.

[14] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[15] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, 2018.

[16] Benjamin Eysenbach, Xinyang Geng, Sergey Levine, and Ruslan Salakhutdinov. Rewriting history with inverse rl: Hindsight inference for policy improvement. *arXiv preprint arXiv:2002.11089*, 2020.

[17] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.

[18] Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational inverse control with events: A general framework for data-driven reward definition. *arXiv preprint arXiv:1805.11686*, 2018.

[19] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

[20] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. *arXiv preprint arXiv:1812.02900*, 2018.

[21] Scott Fujimoto, Herke Van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

[22] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. Offline and online evaluation of news recommender systems at swissinfo. ch. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 169–176, 2014.

[23] Dibya Ghosh, Avi Singh, Aravind Rajeswaran, Vikash Kumar, and Sergey Levine. Divide-and-conquer reinforcement learning. *arXiv preprint arXiv:1711.09874*, 2017.

[24] Arthur Guez, Robert D Vincent, Massimo Avoli, and Joelle Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *AAAI*, pages 1671–1678, 2008.

[25] Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gómez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, et al. Rl unplugged: Benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*, 2020.

[26] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

[27] Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019.

[28] Zhiao Huang, Fangchen Liu, and Hao Su. Mapping state space using landmarks for universal goal reaching. *Advances in Neural Information Processing Systems*, 32:1942–1952, 2019.

[29] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

[30] Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*, 2020.

[31] Leslie Pack Kaelbling. Learning to achieve goals. In *IJCAI*, pages 1094–1099. Citeseer, 1993.

[32] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673. PMLR, 2018.

[33] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[34] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

[35] Taylor W Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare. *arXiv preprint arXiv:2011.11235*, 2020.

[36] Ilya Kostrikov, Jonathan Tompson, Rob Fergus, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. *arXiv preprint arXiv:2103.08050*, 2021.

[37] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11761–11771, 2019.

[38] Aviral Kumar, Abhishek Gupta, and Sergey Levine. Discor: Corrective feedback in reinforcement learning via distribution correction. *arXiv preprint arXiv:2003.07305*, 2020.

[39] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

[40] Sascha Lange, Thomas Gabel, and Martin A. Riedmiller. Batch reinforcement learning. In *Reinforcement Learning*, volume 12. Springer, 2012.

[41] Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

[42] Byung-Jun Lee, Jongmin Lee, and Kee-Eung Kim. Representation balancing offline model-based reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QpNz8r_Ri2Y.

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

[44] Alexander C Li, Lerrel Pinto, and Pieter Abbeel. Generalized hindsight for reinforcement learning. *arXiv preprint arXiv:2002.11708*, 2020.

[45] Jiachen Li, Quan Vuong, Shuang Liu, Minghua Liu, Kamil Ciosek, Keith Ross, Henrik Iskov Christensen, and Hao Su. Multi-task batch reinforcement learning with metric learning. *arXiv preprint arXiv:1909.11373*, 2019.

[46] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[47] Xingyu Lin, Harjatin Singh Baweja, and David Held. Reinforcement learning without ground-truth state. *arXiv preprint arXiv:1905.07866*, 2019.

[48] Hao Liu, Alexander Trott, Richard Socher, and Caiming Xiong. Competitive experience replay. *arXiv preprint arXiv:1902.00528*, 2019.

[49] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *CoRR*, abs/1904.08473, 2019.

[50] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

[51] Corey Lynch and Pierre Sermanet. Grounding language in play. *arXiv preprint arXiv:2005.07648*, 2020.

[52] Ajay Mandlekar, Fabio Ramos, Byron Boots, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Dieter Fox. Iris: Implicit reinforcement without interaction at scale for learning control from offline robot manipulation data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4414–4420. IEEE, 2020.

[53] Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. *arXiv preprint arXiv:2006.03647*, 2020.

[54] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[55] Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *arXiv preprint arXiv:1807.04742*, 2018.

[56] Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*, 2015.

[57] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

[58] Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal difference models: Model-free deep rl for model-based control. *arXiv preprint arXiv:1802.09081*, 2018.

[59] Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pages 417–424, 2001.

[60] Rafael Rafailov, Tianhe Yu, A. Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. *Learning for Decision Making and Control (L4DC)*, 2021.

[61] Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pages 317–328. Springer, 2005.

[62] Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

[63] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320. PMLR, 2015.

[64] Tom Schaul, Diana Borsa, Joseph Modayil, and Razvan Pascanu. Ray interference: a source of plateaus in deep reinforcement learning. *arXiv preprint arXiv:1904.11455*, 2019.

[65] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[66] Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (mime): Large scale demonstrations data for imitation. In *Conference on robot learning*, pages 906–915. PMLR, 2018.

[67] Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1-2):109–136, 2011.

[68] Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

[69] Avi Singh, Huihan Liu, Gaoyue Zhou, Albert Yu, Nicholas Rhinehart, and Sergey Levine. Parrot: Data-driven behavioral priors for reinforcement learning. *arXiv preprint arXiv:2011.10024*, 2020.

[70] Avi Singh, Albert Yu, Jonathan Yang, Jesse Zhang, Aviral Kumar, and Sergey Levine. Cog: Connecting new skills to past experience with offline reinforcement learning. *arXiv preprint arXiv:2010.14500*, 2020.

[71] Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. *arXiv preprint arXiv:2102.06177*, 2021.

[72] Alex Strehl, John Langford, Sham Kakade, and Lihong Li. Learning from logged implicit exploration data. *arXiv preprint arXiv:1003.0120*, 2010.

[73] Hao Sun, Zhizhong Li, Xiaotong Liu, Dahua Lin, and Bolei Zhou. Policy continuation with hindsight inverse dynamics. *arXiv preprint arXiv:1910.14055*, 2019.

[74] Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1):2603–2631, 2016.

[75] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *J. Mach. Learn. Res*, 16:1731–1755, 2015.

[76] Phillip Swazinna, Steffen Udluft, and Thomas Runkler. Overcoming model bias for robust offline deep reinforcement learning. *arXiv preprint arXiv:2008.05533*, 2020.

[77] Yee Whye Teh, Victor Bapst, Wojciech Marian Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. *arXiv preprint arXiv:1707.04175*, 2017.

[78] Georgios Theocharous, Philip S Thomas, and Mohammad Ghavamzadeh. Ad recommendation systems for life-time value optimization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1305–1310, 2015.

[79] Philip S Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *AAAI*, pages 4740–4745, 2017.

[80] L. Wang, Wei Zhang, Xiaofeng He, and H. Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[81] Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*, pages 1015–1022, 2007.

[82] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[83] Annie Xie, Avi Singh, Sergey Levine, and Chelsea Finn. Few-shot goal inference for visuomotor learning and planning. In *Conference on Robot Learning*, pages 40–52. PMLR, 2018.

[84] Annie Xie, Frederik Ebert, Sergey Levine, and Chelsea Finn. Improvisation through physical understanding: Using novel objects as tools with visual foresight. *Robotics: Science and Systems (RSS)*, 2019.

[85] Zhiyuan Xu, Kun Wu, Zhengping Che, Jian Tang, and Jieping Ye. Knowledge transfer in multi-task deep reinforcement learning for continuous control. 2020.

[86] Mengjiao Yang and Ofir Nachum. Representation matters: Offline pretraining for sequential decision making. *arXiv preprint arXiv:2102.05815*, 2021.

[87] Rui Yang, Jiafei Lyu, Yu Yang, Jiangpeng Ya, Feng Luo, Dijun Luo, Lanqing Li, and Xiu Li. Bias-reduced multi-step hindsight experience replay. *arXiv preprint arXiv:2102.12962*, 2021.

[88] Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. *arXiv preprint arXiv:2003.13661*, 2020.

[89] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

[90] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100. PMLR, 2020.

[91] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[92] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*, 2021.

[93] Wenxuan Zhou, Sujay Bajracharya, and David Held. Plas: Latent action space for offline reinforcement learning. *arXiv preprint arXiv:2011.07213*, 2020.

Appendices

A Pseudocode of CDS

We present the summary of the pseudocode of CDS in Algorithm 1.

Algorithm 1 CDS: Conservative Data Sharing

Require: Multi-task offline dataset $\cup_{i=1}^{N} \mathcal{D}_i$.
1: Randomly initialize policy $\pi_\theta(\mathbf{a}|\mathbf{s}, i)$.
2: **for** $k = 1, 2, 3, \cdots,$ **do**
3: Initialize $\mathcal{D}^{\text{eff}} \leftarrow \{\}$
4: **for** $i = 1, \cdots, N$ **do**
5: $\mathcal{D}_i^{\text{eff}} = \mathcal{D}_i \cup \{(\mathbf{s}_j, \mathbf{a}_j, \mathbf{s}'_j, r_i) \in \mathcal{D}_{j \to i} : \Delta^{\pi}(\mathbf{s}, \mathbf{a}) \geq 0\}$ using Eq. 6 (CDS) or Eq. 18 (CDS(basic)).
6: Improve policy by solving eq. 2 using samples from $\mathcal{D}^{\text{eff}}$ to obtain π_θ^{k+1}.

B Analysis of CDS

In this section, we will analyze the key idea behind our method CDS (Section 5) and show that the abstract version of our method (Equation 5) provides better policy improvement guarantees than naïve data sharing and that the practical version of our method (Equation 6) approximates Equation 5 resulting in an effective practical algorithm.

B.1 Analysis of the Algorithm in Equation 5

We begin with analyzing Equation 5, which is used to derive the practical variant of our method, CDS. We build on the analysis of safe-policy improvement guarantees of conventional offline RL algorithms [41, 39] and show that data sharing using CDS attains better guarantees in the worst case. To begin the analysis, we introduce some notation and prior results that we will directly compare to.

Notation and prior results. Let $\pi_\beta(\mathbf{a}|\mathbf{s})$ denote the behavior policy for task i (note that index i was dropped from $\pi_\beta(\mathbf{a}|\mathbf{s}; i)$ for brevity). The dataset, $\mathcal{D}_i$ is generated from the marginal state-action distribution of π_β, i.e., $\mathcal{D} \sim d^{\pi_\beta}(\mathbf{s})\pi_\beta(\mathbf{a}|\mathbf{s})$. We define $d^{\pi}_{\mathcal{D}}$ as the state marginal distribution introduced by the dataset $\mathcal{D}$ under π. Let $D_{\text{CQL}}(p, q)$ denote the following distance between two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ with equal support $\mathcal{X}$:

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

Unless otherwise mentioned, we will drop the subscript "CQL" from D_{CQL} and use D and D_{CQL} interchangeably. Prior works [39] have shown that the optimal policy π_i^* that optimizes Equation 1 attains a high probability safe-policy improvement guarantee, i.e., $J(\pi_i^*) \geq J(\pi_\beta) - \zeta_i$, where ζ_i is:

$$\zeta_i = \mathcal{O}\left(\frac{1}{(1-\gamma)^2} \right) \mathbb{E}_{\mathbf{s} \sim d^{\pi_i^*}_{\mathcal{D}_i}} \left[\sqrt{\frac{D_{\text{CQL}}(\pi_i^*, \pi_\beta)(\mathbf{s}) + 1}{|\mathcal{D}_i(\mathbf{s})|}} \right] + \alpha D(\pi_i^*, \pi_\beta). \tag{7}$$

The first term in Equation 7 corresponds to the decrease in performance due to sampling error and this term is high when the single-task optimal policy π_i^* visits rarely observed states in the dataset $\mathcal{D}_i$ and/or when the divergence from the behavior policy π_β is higher under the states visited by the single-task policy $\mathbf{s} \sim d^{\pi_i^*}_{\mathcal{D}_i}$.

Let $J_{\mathcal{D}}(\pi)$ denote the return of a policy π in the empirical MDP induced by the transitions in the dataset $\mathcal{D}$. Further, let us assume that optimizing Equation 5 gives us the following policies:

$$\pi^*(\mathbf{a}|\mathbf{s}), \pi^*_\beta(\mathbf{a}|\mathbf{s}) := \arg \max_{\pi, \pi_\beta \in \Pi_{\text{relabel}}} \underbrace{J_{\mathcal{D}_i^{\text{eff}}}(\pi) - \alpha D(\pi, \pi_\beta)}_{:= f(\pi, \pi_\beta; \mathcal{D}_i^{\text{eff}})}, \tag{8}$$

where the optimized behavior policy π^*_β is constrained to lie in a set of all policies that can be obtained via relabeling, Π_{relabel}, and the dataset, $\mathcal{D}_i^{\text{eff}}$ is sampled according to the state-action marginal

distribution of π^*_β, i.e., $\mathcal{D}^{\text{eff}}_i \sim d^{\pi^*_\beta}(\mathbf{s}, \mathbf{a})$. Additionally, for convenience, define, $f(\pi_1, \pi_2; \mathcal{D}) := J_{\mathcal{D}}(\pi_1) - \alpha D(\pi_1, \pi_2)$ for any two policies π_1 and π_2, and a given dataset $\mathcal{D}$.

We now show the following result for CDS:

Proposition B.1 (Proposition 5.1 restated)**.** *Let $\pi^*(\mathbf{a}|\mathbf{s})$ be the policy obtained by optimizing Equation 5, and let $\pi_\beta(\mathbf{a}|\mathbf{s})$ be the behavior policy for $\mathcal{D}_i$. Then, w.h.p. $\geq 1-\delta$, π^* is a ζ-safe policy improvement over π_β, i.e., $J(\pi^*) \geq J(\pi_\beta) - \zeta$, where ζ is given by:*

$$\zeta = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right) \mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}^{\text{eff}}_i}} \left[\sqrt{\frac{D_{CQL}(\pi^*, \pi^*_\beta)(\mathbf{s}) + 1}{|\mathcal{D}^{\text{eff}}_i(\mathbf{s})|}}\right] - \left[\alpha D(\pi^*, \pi^*_\beta) + \underbrace{J(\pi^*_\beta) - J(\pi_\beta)}_{(a)}\right],$$

where $\mathcal{D}^{\text{eff}}_i \sim d^{\pi^*_\beta}(\mathbf{s})$ and $\pi^*_\beta(\mathbf{a}|\mathbf{s})$ denotes the policy $\pi \in \Pi_{relabel}$ that maximizes Equation 5.

Proof. To prove this proposition, we shall quantify the lower-bound on the improvement in the policy performance due to Equation 8 in the empirical MDP, and the potential drop in policy performance in the original MDP due to sampling error, and combine the terms to obtain our bound. First note that for any given policy π, and a dataset $\mathcal{D}^{\text{eff}}_i$ with effective behavior policy $\pi_\beta(\mathbf{a}|\mathbf{s})$, the following bound holds [39]:

$$J(\pi) \geq J_{\mathcal{D}^{\text{eff}}_i}(\pi) - \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right) \mathbb{E}_{\mathbf{s}\sim d^{\pi}_{\mathcal{D}^{\text{eff}}_i}} \left[\sqrt{\frac{D_{\text{CQL}}(\pi, \pi^*_\beta)(\mathbf{s}) + 1}{|\mathcal{D}^{\text{eff}}_i(\mathbf{s})|}}\right], \tag{9}$$

where the $\mathcal{O}(\cdot)$ notation hides constants depending upon the concentration properties of the MDP [41] and $1-\delta$, the probability with which the statement holds. Next, we provide guarantees on policy improvement in the empirical MDP. To see this, note that the following statements on $f(\pi_1, \pi_2; \mathcal{D})$ are true:

$$\forall \pi' \in \Pi_{\text{relabel}}, \quad f(\pi^*, \pi^*_\beta; \mathcal{D}^{\text{eff}}_i) \geq f(\pi', \pi', \mathcal{D}^{\text{eff}}_i) \tag{10}$$

$$\implies \forall \pi' \in \Pi_{\text{relabel}}, \quad J_{\mathcal{D}^{\text{eff}}_i}(\pi^*) - \alpha D(\pi^*, \pi^*_\beta) \geq J_{\mathcal{D}^{\text{eff}}_i}(\pi'). \tag{11}$$

And additionally, we obtain:

$$\forall \pi' \in \Pi_{\text{relabel}}, \quad f(\pi^*, \pi^*_\beta; \mathcal{D}^{\text{eff}}_i) \geq f(\pi^*, \pi'; \mathcal{D}^{\text{eff}}_i), \tag{12}$$

$$\implies \forall \pi' \in \Pi_{\text{relabel}}, \quad D(\pi^*, \pi^*_\beta) \leq D(\pi^*, \pi'). \tag{13}$$

Utilizing 11, we obtain that:

$$J_{\mathcal{D}^{\text{eff}}_i}(\pi^*) - J_{\mathcal{D}^{\text{eff}}_i}(\pi_\beta) \geq \alpha D(\pi^*, \pi^*_\beta) + \left(J_{\mathcal{D}^{\text{eff}}_i}(\pi^*_\beta) - J_{\mathcal{D}^{\text{eff}}_i}(\pi_\beta)\right) \approx \alpha D(\pi^*, \pi^*_\beta) + \left(J(\pi^*_\beta) - J(\pi_\beta)\right), \tag{14}$$

where $\approx$ ignores sampling error terms that do not depend on distributional shift measures like D_{CQL} because π^*_β and π_β are behavior policies which generated the complete and part of the dataset, and hence these terms are dominated by and subsumed into the sampling error for π^*. Combining Equations 9 (by setting $\pi = \pi^*$) and 14, we obtain the following safe-policy improvement guarantee for π^*: $J(\pi^*) - J(\pi_\beta) \geq \zeta$, where ζ is given by:

$$\zeta = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right) \mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}^{\text{eff}}_i}} \left[\sqrt{\frac{D_{\text{CQL}}(\pi^*, \pi^*_\beta)(\mathbf{s}) + 1}{|\mathcal{D}^{\text{eff}}_i(\mathbf{s})|}}\right] - \left[\alpha D(\pi^*, \pi^*_\beta) + \underbrace{J(\pi^*_\beta) - J(\pi_\beta)}_{\text{(a)}}\right],$$

which proves the desired result. □

Proposition B.1 indicates that when optimizing the behavior policy with Equation 5, we can improve upon the conventional safe-policy improvement guarantee (Equation 7) with standard single-task offline RL: not only do we improve via $D_{\text{CQL}}(\pi^*, \pi^*_\beta)$, since, $D_{\text{CQL}}(\pi^*, \pi^*_\beta) \leq D_{\text{CQL}}(\pi^*, \pi_\beta)$, which reduces sampling error, but utilizing this policy π^*_β also allows us to improve on term (a), since Equation 8 optimizes the behavior policy to be close to the learned policy π^* and maximizes the learned policy return $J_{\mathcal{D}^{\text{eff}}_i}(\pi^*)$ on the effective dataset, thus providing us with a high lower bound on $J(\pi^*_\beta)$. We formalize this insight as Lemma B.1 below:

Lemma B.1. *For sufficiently large α, $J_{\mathcal{D}_i^{\text{eff}}}(\pi^*_\beta) \geq J_{\mathcal{D}_i^{\text{eff}}}(\pi_\beta)$ and thus $(a) \geq 0$.*

Proof. To prove this, we note that using standard difference of returns of two policies, we get the following inequality: $J_{\mathcal{D}_i^{\text{eff}}}(\pi^*_\beta) \geq J_{\mathcal{D}_i^{\text{eff}}}(\pi^*) - C\frac{R_{\max}}{1-\gamma} D_{\text{TV}}(\pi^*, \pi^*_\beta)$. Moreover, from Equation 11, we obtain that: $J_{\mathcal{D}_i^{\text{eff}}}(\pi^*) - \alpha D(\pi^*, \pi^*_\beta) \geq J_{\mathcal{D}_i^{\text{eff}}}(\pi_\beta)$. So, if α is chosen such that:

$$\frac{CR_{\max}}{1-\gamma} D_{\text{TV}}(\pi^*, \pi^*_\beta) \leq \alpha D(\pi^*, \pi^*_\beta), \tag{15}$$

we find that:

$$J_{\mathcal{D}_i^{\text{eff}}}(\pi^*_\beta) \geq J_{\mathcal{D}_i^{\text{eff}}}(\pi^*) - C\frac{R_{\max}}{1-\gamma} D_{\text{TV}}(\pi^*, \pi^*_\beta) \geq J_{\mathcal{D}_i^{\text{eff}}}(\pi^*) - \alpha D(\pi^*, \pi^*_\beta) \geq J_{\mathcal{D}_i^{\text{eff}}}(\pi_\beta),$$

implying that $(a) \geq 0$. For the edge cases when either $D_{\text{TV}}(\pi^*, \pi^*_\beta) = 0$ or $D_{\text{CQL}}(\pi^*, \pi^*_\beta) = 0$, we note that $\pi^*(\mathbf{a}|\mathbf{s}) = \pi^*_\beta(\mathbf{a}|\mathbf{s})$, which trivially implies that $J_{\mathcal{D}_i^{\text{eff}}}(\pi^*_\beta) = J_{\mathcal{D}_i^{\text{eff}}}(\pi^*) \geq J_{\mathcal{D}_i^{\text{eff}}}(\pi_\beta)$, because π^* improves over π_β on the dataset. Thus, term (a) is positive for large-enough α and the bound in Proposition B.1 gains from this term additionally. □

Finally, we show that the sampling error term is controlled when utilizing Equation 5. We will show in Lemma B.2 that the sampling error in Proposition B.1 is controlled to be not much bigger than the error just due to variance, since distributional shift is bounded with Equation 5.

Lemma B.2. *If π^* and π^*_β obtained from Equation 5 satisfy, $D_{CQL}(\pi^*, \pi^*_\beta) \leq \varepsilon \ll 1$, then:*

$$(\$) := \mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}}\left[\sqrt{\frac{D_{CQL}(\pi^*, \pi^*_\beta)(\mathbf{s}) + 1}{|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|}}\right] \leq (1+\varepsilon)^{\frac{1}{2}} \underbrace{\mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}}\left[\sqrt{\frac{1}{|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|}}\right]}_{:=\textit{sampling error w/o distribution shift}}. \tag{16}$$

Proof. This lemma can be proved via a simple application of the Cauchy-Schwarz inequality. We can partition the first term as a sum over dot products of two vectors such that:

$$\begin{aligned}
(\$) &= \sum_{\mathbf{s}} \sqrt{d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}(\mathbf{s})(D_{\text{CQL}}(\pi^*, \pi^*_\beta)(\mathbf{s}) + 1)}\sqrt{\frac{d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}(\mathbf{s})}{|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|}} \\
&\leq \sqrt{\left(\sum_{\mathbf{s}} d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}(\mathbf{s})(D_{\text{CQL}}(\pi^*, \pi^*_\beta)(\mathbf{s}) + 1)\right) \cdot \left(\sum_{\mathbf{s}} \frac{d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}(\mathbf{s})}{|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|}\right)} \\
&= \sqrt{\mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}}\left[D_{\text{CQL}}(\pi^*, \pi^*_\beta)(\mathbf{s}) + 1\right] \mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}}\left[\frac{1}{|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|}\right]} \leq (1+\varepsilon)^{0.5}\mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}}\left[\sqrt{\frac{1}{|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|}}\right],
\end{aligned}$$

where we note that $\mathbb{E}_{\mathbf{s}\sim d^{\pi^*}_{\mathcal{D}_i^{\text{eff}}}}\left[D_{\text{CQL}}(\pi^*, \pi^*_\beta)(\mathbf{s})\right] = D_{\text{CQL}}(\pi^*, \pi^*_\beta) \leq \varepsilon$ (based on the given information in the Lemma) and that $\sqrt{\sum_i w_i \frac{1}{x_i}} \leq \sum_i w_i \frac{1}{\sqrt{x_i}}$ for $x_i, w_i > 0$ and $\sum_i w_i = 1$, via Jensen's inequality for concave functions. □

To summarize, combining Lemmas B.1 and B.2 with Proposition B.1, we conclude that utilizing Equation 5 controls the increase in sampling error due to distributional shift, and provides improvement guarantees on the learned policy beyond the behavior policy of the original dataset. We also briefly now discuss the comparison between CDS and complete data sharing. Complete data sharing would try to reduce sampling error by increasing $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$, but then it can also increase distributional shift, $D_{\text{CQL}}(\pi^*, \pi^*_\beta)$ as discussed in Section 4. On the other hand, CDS increases the dataset size while also controlling for distributional shift (as we discussed in the analysis above), making it enjoy the benefits of complete data sharing and avoiding its pitfalls, intuitively. On the other hand, no data sharing will just incur high sampling error due to limited dataset size.

B.2 From Equation 5 to Practical CDS (Equation 6)

The goal of our practical algorithm is to convert Equation 5 to a practical algorithm while retaining the policy improvement guarantees derived in Proposition B.1. Since our algorithm does not utilize any estimator for dataset counts $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$, and since we operate in a continuous state-action space, our goal is to retain the guarantees of increased return of π^*_β, while also avoiding sampling error.

With this goal, we first need to relax the state-distribution in Equation 5: while both $J_{\mathcal{D}_i^{\text{eff}}}(\pi)$ and $D_{\text{CQL}}(\pi, \pi_\beta)$ are computed as expectations under the marginal state-distribution of policy $\pi(\mathbf{a}|\mathbf{s})$ on the MDP defined by the dataset $\mathcal{D}_i^{\text{eff}}$, for deriving a practical method we relax the state distribution to use the dataset state-distribution $d^{\pi^*_\beta}$ and rewrite the objective in accordance with most practical implementations of actor-critic algorithms [12, 1, 26, 21, 46] below:

$$\text{(Practical Equation 5)} \quad \max_{\pi} \max_{\pi_\beta \in \Pi_{\text{relabel}}} \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_i^{\text{eff}}}[\mathbb{E}_{\mathbf{a}\sim\pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})] - \alpha D(\pi(\cdot|\mathbf{s}), \pi_\beta(\cdot|\mathbf{s}))] \quad (17)$$

This practical approximation in Equation 17 is even more justified with conservative RL algorithms when a large α is used, since a larger α implies a smaller value for $D(\pi^*, \pi^*_\beta)$ found by Equation 5, which in turn means that state-distributions $d^{\pi^*_\beta}$ and d^{π^*} are close to each other [65]. Thus, our policy improvement objective optimizes the policies π and π_β by maximizing the conservative Q-function: $\hat{Q}^\pi(\mathbf{s},\mathbf{a}) = Q(\mathbf{s},\mathbf{a}) - \alpha\left(\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi_\beta(\mathbf{a}|\mathbf{s})} - 1\right)$, that appears inside the expectation in Equation 17. While optimizing the policy π with respect to this conservative Q-function $\hat{Q}^\pi(\mathbf{s},\mathbf{a})$ is equivalent to a standard policy improvement update utilized by most actor-critic methods [21, 26, 39], we can optimize $\hat{Q}^\pi(\mathbf{s},\mathbf{a})$ with respect to $\pi_\beta \in \Pi_{\text{relabel}}$ by relabeling only those transitions $(\mathbf{s},\mathbf{a},r'_i,\mathbf{s}') \in \mathcal{D}_{j\to i}$ that increase the expected conservative Q-value $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_i^{\text{eff}}}\left[\mathbb{E}_{\mathbf{a}\sim\pi_\beta(\cdot|\mathbf{s})}\left[\hat{Q}^\pi(\mathbf{s},\mathbf{a})\right]\right]$. Note that we relaxed the expectation $\mathbf{a}\sim\pi(\mathbf{a}|\mathbf{s})$ to $\mathbf{a}\sim\pi_\beta(\mathbf{a}|\mathbf{s})$ in this expectation, which can be done upto a lower-bound of the objective in Equation 17 for a large α, since the resulting policies π and π_β are close to each other.

The last step in our practical algorithm is to modify the solution of Equation 17 to still retain the benefits of reduced sampling error as discussed in Proposition B.1. To do so, we want to relabel as many points as possible, thus increasing $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$, which leads to reduced sampling error. Since quantifying $|\mathcal{D}_i^{\text{eff}}(\mathbf{s})|$ in continuous state-action spaces will require additional machinery such as density-models, we avoid these for the sake of simplicity, and instead choose to relabel every datapoint $(\mathbf{s},\mathbf{a}) \in \mathcal{D}_{j\to i}$ that satisfies $Q^\pi(\mathbf{s},\mathbf{a};i) \geq \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}_i}[\hat{Q}^\pi(\mathbf{s},\mathbf{a};i)] \geq 0$ to task i. These datapoints definitely increase the conservative Q-value and hence increase the objective in Equation 17 (though do not globally maximize it), while also enjoying properties of reduced sampling error (Proposition B.1). This discussion motivates our practical algorithm in Equation 6.

C Experimental details

In this section, we provide the training details of CDS in Appendix C.1 and also include the details on the environment and datasets that we use for the evaluation in Appendix C.2. Finally, we include the discussion on the compute information in Appendix C.3. We also compare CDS to an offline RL with pretrained representations from multi-task datasets method [86].

C.1 Training details

The pseudocode of CDS is summarized in Algorithm 1 in Appendix A. The complete variant of CDS can be directly implemented using the rule in Equation 6 with conservative Q-value estimates obtained via any offline RL method that constrains the learned policy to the behavior policy. For implementing CDS (basic), we reparameterize the divergence D in Equation 4 to use the learned conservative Q-values. This is especially useful for our implementation since we utilize CQL as the base offline RL method, and hence we do not have access to an explicit divergence. In this case, $\Delta^\pi(\mathbf{s},\mathbf{a})$ can be redefined as, $\Delta^\pi(\mathbf{s},\mathbf{a}) :=$

$$\mathbb{E}_{\mathbf{s}'\sim\mathcal{D}^i}\left[\mathbb{E}_{\mathbf{a}'\sim\pi}[\hat{Q}(\mathbf{s}',\mathbf{a}',i)] - \mathbb{E}_{\mathbf{a}''\sim\mathcal{D}_i}[\hat{Q}(\mathbf{s}',\mathbf{a}'',i)]\right] - \left(\mathbb{E}_{\mathbf{a}'\sim\pi}[\hat{Q}(\mathbf{s},\mathbf{a}',i)] - Q(\mathbf{s},\mathbf{a},i)\right), \quad (18)$$

Equation 18 can be viewed as the difference between the CQL [39] regularization term on a given $(\mathbf{s}, \mathbf{a})$ and the original dataset for task i, $\mathcal{D}_i$. This CQL regularization term is equal to the divergence between the learned policy $\pi(\cdot|\mathbf{s})$ and the behavior policy $\pi_\beta(\cdot|\mathbf{s})$, therefore Equation 18 practically computes Equation 4.

Note that both variants of CDS train a policy, $\pi(\mathbf{a}|\mathbf{s}; i)$, either conditioned on the task i (i.e., with weight sharing) or a separate $\pi(\mathbf{a}|\mathbf{s})$ policy for each task with no weight sharing, using the resulting relabeled dataset, $\mathcal{D}_i^{\text{eff}}$. Next, we discuss the training details of the complete version of CDS.

Our practical implementation of CDS optimizes the following objectives for training the critic and the policy:

$$\hat{Q}^{k+1} \leftarrow \arg\min_{\hat{Q}} \mathbb{E}_{i\sim[N]} \Big[\beta \Big(\mathbb{E}_{j\sim[N]} \Big[\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_j, \mathbf{a}\sim\mu(\cdot|\mathbf{s},i)} \Big[w_{\text{CDS}}(\mathbf{s}, \mathbf{a}; j \to i)\hat{Q}(\mathbf{s}, \mathbf{a}, i)\Big] - \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}_j} \Big[w_{\text{CDS}}(\mathbf{s}, \mathbf{a}; j \to i)\hat{Q}(\mathbf{s}, \mathbf{a}, i)\Big]\Big]\Big) + \frac{1}{2}\mathbb{E}_{j\sim[N],(\mathbf{s},\mathbf{a},\mathbf{s}')\sim\mathcal{D}_j} \Big[w_{\text{CDS}}(\mathbf{s}, \mathbf{a}; j \to i) \Big(\hat{Q}(\mathbf{s}, \mathbf{a}, i) - \widehat{\mathcal{B}}^\pi \hat{Q}^k(\mathbf{s}, \mathbf{a}, i)\Big)^2\Big]\Big],$$

and

$$\pi \leftarrow \arg\max_{\pi'} \mathbb{E}_{i\sim[N]} \Big[\mathbb{E}_{j\sim[N],\mathbf{s}\sim\mathcal{D}_j,\mathbf{a}\sim\pi'(\cdot|\mathbf{s},i)} \Big[w_{\text{CDS}}(\mathbf{s}, \mathbf{a}; j \to i)\hat{Q}^\pi(\mathbf{s}, \mathbf{a}, i)\Big]\Big],$$

where β is the coefficient of the CQL penalty on distribution shift, μ is a wide sampling distribution as in CQL and $\widehat{\mathcal{B}}$ is the sample-based Bellman operator.

To compute the relabeling weight $w_{\text{CDS}}(\mathbf{s}, \mathbf{a}; j \to i) := \sigma\left(\frac{\Delta(\mathbf{s},\mathbf{a};j\to i)}{\tau}\right)$, we need to pick the value of the temperature term τ. Instead of tuning τ manually, we follow the the adaptive temperature scaling scheme from [38]. Specifically, we compute an exponential running average of $\Delta(\mathbf{s}, \mathbf{a}; j \to i)$ with decay 0.995 for each task and use it as τ. We additionally clip the adaptive temperature term with a minimum and maximum threshold, which we tune manually. For multi-task halfcheetah, walker2d and ant, we clip the adaptive temperature such that it lies within $[10, \infty]$, $[5, \infty]$ and $[10, 25]$ respectively. For the multi-task Meta-Wold experiment, we use $[1, 50]$ for the clipping. For multi-task Antmaze, we used a range of $[10, \infty]$ for all the domains. We do not clip the temperature term on vision-based domains.

For state-based experiments, we use a stratified batch with 128 transitions for each task for the critic and policy learning. For each task i, we sample 64 transitions from $\mathcal{D}_i$ and another 64 transitions from $\cup_{j\neq i}\mathcal{D}_{j\to i}$, i.e. the relabeled datasets of all the other tasks. When computing $\Delta(\mathbf{s}, \mathbf{a}; j \to i)$, we only apply the weight to relabeled data on multi-task Meta-World environments and multi-task vision-based robotic manipulation tasks while also applying the weight to the original data drawn from $\mathcal{D}_i$ with 50% chance for each task $i \in [N]$ in the remaining domains.

We use CQL [39] as the base offline RL algorithm. On state-based experiments, we mostly follow the hyperparameters provided in prior work [39]. One exception is that on the multi-task ant domain, we set $\beta = 5.0$ and on the other two locomotion environments and the multi-task Meta-World domain, we use $\beta = 1.0$. On multi-task AntMaze, we use the Lagrange version of CQL, where the multiplier β is automatically tuned against a pre-specific constraint value on the CQL loss equal to $\tau = 5.0$. We use a policy learning rate $1e-4$ and a critic learning rate $3e-4$ as in [39]. On the vision-based environment, instead of using the direct CQL algorithm, we follow [8] and sample unseen actions according to the soft-max distritbution of the Q-values and set its Q target value to 0. This algorithm can be viewed the version of CQL with $\beta = 1.0$ in Eq.1 in [39], i.e. removing the term of negative expected Q-values on the dataset. We follow the other hyperparameters from prior work [32, 8, 33].

For the choice architectures, in the domains with low-dimensional state inputs, we use 3-layer feedforward neural networks with 256 hidden units for both the Q-networks and the policy. We append a one-hot task vector to the state of each environment. For the vision-based experiment, our Q-network architecture follows from multi-headed convolutional networks used in MT-Opt [33]. For the observation input, we use images with dimension $472 \times 472 \times 3$ along with additional state features $(g_{\text{status}}, g_{\text{height}})$ as well as the one-hot task vector as in [33]. For the action input, we use Cartesian space control of the end-effector of the robot in 4D space (3D position and azimuth angle) along with two discrete actions for opening/closing the gripper and terminating the episode respectively. More details can be found in [32, 33].

C.2 Environment and dataset details

In this subsection, we discuss the details of how we set up the multi-task environment and how we collect the offline datasets. We want to acknowledge that all datasets with state inputs use the MIT License.

Multi-task locomotion domains. We construct the environment by changing the reward function in [5]. On the halfcheetah environment, we follow [91] and set the reward functions of task `run forward`, `run backward` and `jump` as $r(s,a) = \max\{v_x, 3\} - 0.1 * \|a\|_2^2$, $r(s,a) = -\max\{v_x, 3\} - 0.1 * \|a\|_2^2$ and $r(s,a) = -0.1 * \|a\|_2^2 + 15 * (z - \text{init z})$ respectively where v_x denotes the velocity along the x-axis and z denotes the z-position of the half-cheetah and init z denotes the initial z-position. Similarly, on walker2d, the reward functions of the three tasks are $r(s,a) = v_x - 0.001 * \|a\|_2^2$, $r(s,a) = -v_x - 0.001 * \|a\|_2^2$ and $r(s,a) = -\|v_x\| - 0.001 * \|a\|_2^2 + 10 * (z - \text{init z})$ respectively. Finally, on ant, the reward functions of the three tasks are $r(s,a) = v_x - 0.5 * \|a\|_2^2 - 0.005 * \text{contact-cost}$, $r(s,a) = -v_x - 0.5 * \|a\|_2^2 - 0.005 * \text{contact-cost}$ and $r(s,a) = -\|v_x\| - 0.5 * \|a\|_2^2 - 0.005 * \text{contact-cost} + 10 * (z - \text{init z})$.

On each of the multi-task locomotion environment, we train each task with SAC [26] for 500 epochs. For medium-replay datasets, we take the whole replay buffer after the online SAC is trained for 100 epochs. For medium datasets, we take the online single-task SAC policy after 100 epochs and collect 500 trajectories with the medium-level policy. For expert datasets, we take the final online SAC policy and collect 5 trajectories with it for walker2d and halfcheetah and 20 trajectories for ant.

Meta-World domains. We take the `door open`, `door close`, `drawer open` and `drawer close` environments from the open-sourced Meta-World [90] repo[1]. We put both the door and the drawer on the same scene to make sure the state space of all four tasks are shared. For offline training, we use sparse rewards for each task by replacing the dense reward defined in Meta-World with the success condition defined in the public repo. Therefore, each task gets a reward of 1 if the task is fully completed and 0 otherwise.

For generating the offline datasets, we train each task with online SAC using the dense reward defined in Meta-World for 500 epochs. For medium-replay datasets, we take the whole replay buffer of the online SAC until 150 epochs. For the expert datasets, we run the final online SAC policy to collect 10 trajectories.

AntMaze domains. We take the `antmaze-medium-play` and `antmaze-large-play` datasets from D4RL [19] and convert the datasets into multi-task datasets in two ways. In the undirected version of these tasks, we split the dataset randomly into equal sized partitions, and then assign each partition to a particular randomly chosen task. Thus, the task data observed in the data for each task is largely unsuccessful for the particular task it is assigned to and effective data sharing is essential for obtaining good performance. The second setting is the directed data setting where a trajectory in the dataset is marked to belong to the task corresponding to the actual end goal of the trajectory. A sparse reward equal to +1 is provided to an agent when the current state reaches within a 0.5 radius of the task goal as was used default by Fu et al. [19].

Vision-based robotic manipulation domains. Following MT-Opt [33], we use sparse rewards for each task, i.e. reward 1 for success episodes and 0 otherwise. We define successes using the success detectors defined in [33]. To collect data for vision-based experiments, we train a policy for each task individually by running QT-Opt [32] with default hyperparameters until the task reaches 40% success rate for picking skills and 80% success rate for placing skills. We take the whole replay buffer of each task and combine all of such replay buffers to form the multi-task offline dataset with total 100K episodes where each episode has 25 transitions.

C.3 Computation Complexity

For all the state-based experiments, we train CDS on a single NVIDIA GeForce RTX 2080 Ti for one day. For the image-based robotic manipulation experiments, we train it on 16 TPUs for three days.

[1]The Meta-World environment can be found at the public repo https://github.com/rlworkgroup/metaworld

D Visualizations, Comparisons and Additional Experiments

In this section, we perform diagnostic and ablation experiments to: **(1)** understand the efficacy of CDS when applied with other base offline RL algorithms, such as BRAC [82], **(2)** visualize the weights learned by CDS to understand if the weighting scheme induced by CDS corresponds to what we would intuitively expect on different tasks, and **(3)** compare CDS to a prior approach that performs representation learning from offline multi-task datasets and then runs vanilla multi-task RL algorithm on top of the learned representations. We discuss these experiments next.

D.1 Applying CDS with BRAC [82], A Policy-Constraint Offline RL Algorithm

We implemented CDS on top of BRAC which is different from CQL that penalizes Q-functions. BRAC computes the divergence $D(\pi, \pi_\beta)$ in Equation 1 explicitly and penalizes the reward function $r(\mathbf{s}, \mathbf{a})$ with this value in the Bellman backups. To apply CDS to BRAC, we need to compute a conservative estimate of the Q-value as discussed in Section 5.2. While the Q-function from CQL directly provides us with this conservative estimate, BRAC does not directly learn a conservative Q-function estimator. Therefore, for BRAC, we compute this conservative estimate by explicitly subtracting KL divergence between the learned policy $\pi(a|s)$ and the behavior policy π^β on state-action tuples (s, a) from the learned Q-function's prediction. Formally, this means that we utilize $\hat{Q}(s, a) := Q(s, a) - \alpha \mathrm{D}_{\mathrm{KL}}(\pi(a|s), \pi^\beta(a|s))$ as our conservative Q-value estimate for BRAC. Given these conservative Q-value estimate, CDS weights can be computed directly using Equation 6.

Environment	Tasks / Dataset type	BRAC + CDS (ours)	BRAC + No Sharing	BRAC + Sharing All
Meta-World [90]	door open / expert	**44.0%**±3.0%	35.0%±25.9%	38.0%±2.2%
	door close / medium-replay	**32.5%** ± 5.0%	5.0% ± 8.6%	8.6% ± 3.4%
	drawer open / medium-replay	**28.5%**±3.5%	21.8%±5.6%	0.0% ± 0.0%
	drawer close / expert	100%±0.0%	100.0%±0.0%	99.0%±0.7%
	average	**52.5%**±7.4%	22.5%±13.3%	40.0%±5.0%

Table 5: Applying CDS on top of BRAC. Note that CDS + BRAC imrpoves over both BRAC + Sharing All and BRAC + No sharing, indicating that CDS is effective over other offline RL algorithms such as BRAC as well. The ± values indicate the value of the standard deviation of runs, and results are averaged over three seeds.

We evaluated BRAC + CDS on the Meta-World tasks and compared it to BRAC + Sharing All and BRAC + No Sharing. We present the results in Table 5. We use ± to denote the 95%-confidence interval. As observed below, BRAC + CDS significantly outperforms BRAC with Sharing All and BRAC with No sharing. This indicates that CDS is effective on top of BRAC.

D.2 Analyzing CDS weights for Different Scenarios

Next, to understand if the weights assigned by CDS align with our expectation for which transitions should be shared between tasks, we perform diagnostic analysis studies on the weights learned by CDS on the Meta-World and Antmaze domains.

On the Meta-World environment, we would expect that for a given target task, say Drawer Close, transitions from a task that involves a different object (door) and a different skill (open) would not be as useful for learning. To understand if CDS weights reflect this expectation, we compare the average CDS weights on transitions from all the other tasks to two target tasks, Door Open and Drawer Close, respectively and present the results in Table 6. We sort the CDS weights in the descending order. As shown, indeed CDS assigns higher weights to more related tasks and thus shares data from those tasks. In particular, the CDS weights for relabeling data from the task that handles the same object as the target task are much higher than the weights for tasks that consider a different object.

For example, when relabeling to the target task Door Open, datapoints from task Door Close are assigned with much higher weights than those from either task Drawer Open or task Drawer Close. This suggests that CDS filters the irrelevant transitions for learning a given task.

On the AntMaze-large environment, with undirected data, we visualize the CDS weight for the various tasks (goals) in the form of a heatmap and present the results in Figure 4. To generate this plot, we sample a set of state-action pairs from the entire dataset for all tasks, and then plot the weights assigned by CDS as the color of the point marker at the (x, y) locations of these state-action pairs in the maze. Each plot computes the CDS weight corresponding to the target task (goal) indicated by the

Relabeling Direction	CDS weight
door close → door open	0.46
drawer open → door open	0.10
drawer close → door open	0.02
drawer open → drawer close	0.35
door open → drawer close	0.26
door close → drawer close	0.22

Table 6: On the Meta-World domain, we visualize the CDS weights of data relabeled from other tasks to the two target tasks door open and drawer close shown in the second row and third row respectively. We sort the CDS weights for relabeled tasks to a particular target task in the descending order. As shown in the table, CDS upweights tasks that are more related to the target task, e.g. manipulating the same object.

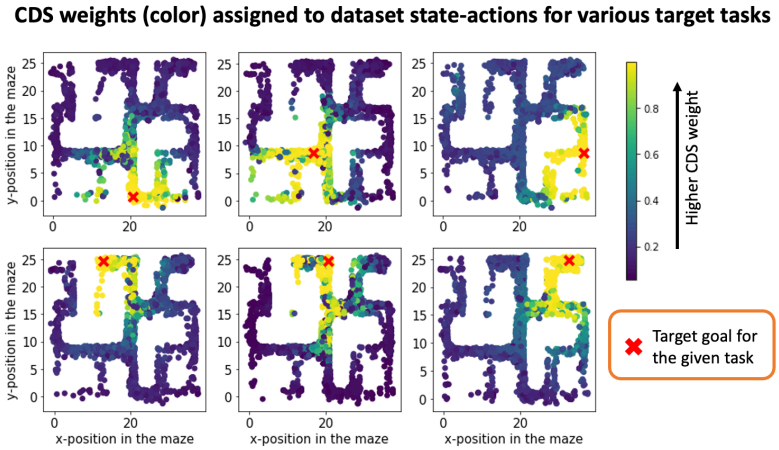

Figure 4: A visualization of the weights assigned by CDS to various transitions in the antmaze dataset for six target goals (indicated by $\times$ clustered by their spatial location. Note that CDS up-weights transitions spatially close to the target goal (indicated in the brighter yellow color), matching our expectation.

red $\times$ in the plot. As can be seen in Figure 4, CDS assigns higher weights to transitions from nearby goals as compared to transitions from farther away goals. This matches our expectation: transitions from nearby (x, y) locations are likely to be the most useful in learning a particular target task and CDS chooses to share these transitions to the target task.

D.3 Comparison of CDS with Other Alternatives to Data Sharing: Utilizing Multi-Task Datasets for Learning Pre-Trained Representations

Finally, we aim to empirically verify how other alternatives to data sharing perform on multi-task offline RL problems. One simple approach to utilize data from other tasks is to use this data to learn low-dimensional representations that capture meaningful information about the environment initially in a pre-training phase and then utilize these representations for improved multi-task RL without any specialized data sharing schemes. To assess the efficacy of this alternate approach of using multi-task offline data, in Table 7, we performed an experiment on the Meta-World domain that first utilizes the data from all the tasks to learn a shared representation using the best method, ACL [86] and then runs standard offline multi-task RL on top of this representation. We denote the method as **Offline Pretraining**. We include the average task success rates of all tasks in the table below. While the representation learning approach improves over standard multi-task RL without representation learning (**No Sharing**) consistent with the findings in [86], we still find that CDS with no representation learning outperforms this representation learning approach by a large margin on multi-task performance, which suggests that conservative data sharing is more important than pure pretrained representation from multi-task datasets in the offline multi-task setting. We finally remark that in principle, we could also utilize representation learning approaches in conjunction with data sharing strategies and systematically characterizing this class of hybrid approaches is a topic of future work.

Environment	Tasks / Dataset type	CDS (ours)	No Sharing	Offline Petraining [86]
Meta-World [90]	door open / expert	**58.4%**±9.3%	14.5%±12.7%	48.0%±40.0%
	door close / medium-replay	**65.3%**±27.7%	4.0%±6.1%	9.5%±8.4%
	drawer open / medium-replay	**57.9%**±16.2%	16.0%±17.5%	1.3% ± 1.4%
	drawer close / expert	98.8%±0.7%	99.0%±0.7%	96.0%±0.9%
	average	**70.1%**±8.1%	33.4%±8.3%	38.7%±11.1%

Table 7: Comparison between CDS and Offline Pretraining [86] that pretrains the representation from the multi-task offline data and then runs multi-task offline RL on top of the learned representation on the Meta-World domain. Numbers are averaged across 6 seeds, ± the 95%-confidence interval. CDS significantly outperforms Offline Pretraining.