



Pessimistic value iteration for multi-task data sharing in Offline Reinforcement Learning

Chenjia Bai^{a,f}, Lingxiao Wang^b, Jianye Hao^c, Zhuoran Yang^d, Bin Zhao^{a,e},
Zhen Wang^{e,*}, Xuelong Li^{a,e,*}

^a Shanghai Artificial Intelligence Laboratory, Shanghai, China

^b Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA

^c Tianjin University, Tianjin, China

^d Department of Statistics and Data Science, Yale University, New Haven, CT, USA

^e School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China

^f Shenzhen Research Institute of Northwestern Polytechnical University, Shenzhen, China

ARTICLE INFO

Keywords:

Uncertainty quantification
Data sharing
Pessimistic value iteration
Offline Reinforcement Learning

ABSTRACT

Offline Reinforcement Learning (RL) has shown promising results in learning a task-specific policy from a fixed dataset. However, successful offline RL often relies heavily on the coverage and quality of the given dataset. In scenarios where the dataset for a specific task is limited, a natural approach is to improve offline RL with datasets from other tasks, namely, to conduct Multi-Task Data Sharing (MTDS). Nevertheless, directly sharing datasets from other tasks exacerbates the distribution shift in offline RL. In this paper, we propose an uncertainty-based MTDS approach that shares the entire dataset without data selection. Given ensemble-based uncertainty quantification, we perform pessimistic value iteration on the shared offline dataset, which provides a unified framework for single- and multi-task offline RL. We further provide theoretical analysis, which shows that the optimality gap of our method is only related to the expected data coverage of the shared dataset, thus resolving the distribution shift issue in data sharing. Empirically, we release an MTDS benchmark and collect datasets from three challenging domains. The experimental results show our algorithm outperforms the previous state-of-the-art methods in challenging MTDS problems.

1. Introduction

Deep Reinforcement Learning (DRL) [1] is a growing direction in Decision Making and plays an important role in broad Artificial Intelligence research. DRL has achieved remarkable success in a variety of tasks, including Atari games [2], Go [3], and StarCraft [4]. However, in most successful applications, DRL requires millions of interactions with the environment. In real-world applications such as navigation [5] and healthcare [6], acquiring a large number of samples by following a possibly suboptimal policy can be costly and dangerous. Alternatively, practitioners seek to develop *Offline RL* [7,8] algorithms that learn a policy based solely on an offline dataset, where the dataset is typically available. The most significant challenge in offline RL is the distribution shift issue, where the

* Corresponding authors.

E-mail addresses: baichenjia@pjlab.org.cn (C. Bai), w-zhen@nwpu.edu.cn (Z. Wang), li@nwpu.edu.cn (X. Li).

<https://doi.org/10.1016/j.artint.2023.104048>

Received 23 November 2022; Received in revised form 25 October 2023; Accepted 14 November 2023

Available online 20 November 2023

0004-3702/© 2023 Elsevier B.V. All rights reserved.

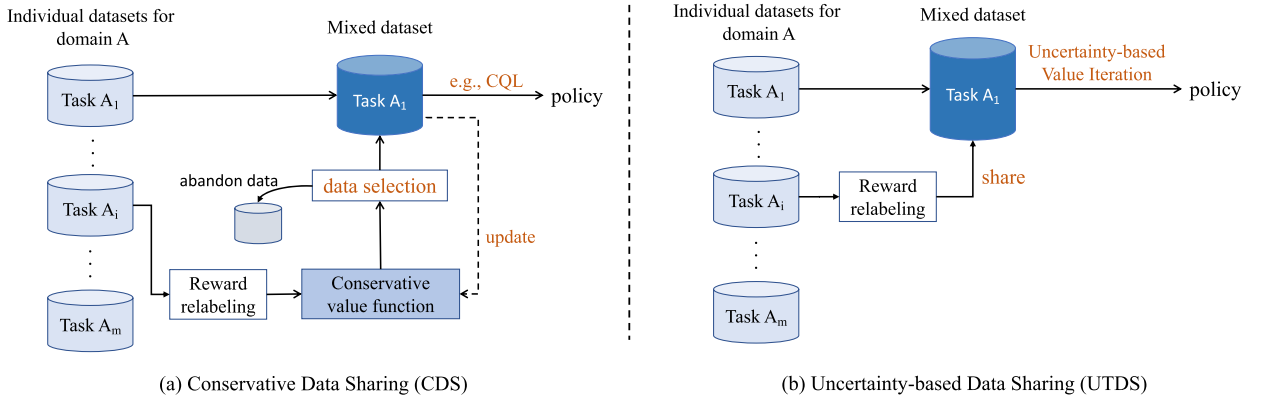


Fig. 1. The illustration of CDS and UTDS for MTDS in training task A_1 . (a) CDS includes a data selection process through the learned conservative value function. The selected data is added to the mixed dataset. (b) UTDS can share all data from other tasks without data selection. In policy training, UTDS performs pessimistic updates based on uncertainty in the large shared dataset.

offline dataset does not match the current policy in optimization. Such a distribution shift leads to inaccurate policy evaluation, which hinders the performance of offline RL. To tackle this problem, most of the previous offline RL methods enforce policy constraints between the learned policy and the behavior policy that collects the dataset [9,10]. As a consequence, the performance of such methods relies heavily on the quality of the behavior policy. Alternatively, previous methods also measure the uncertainty of state-action pairs and utilize such uncertainty measurement to conduct conservative policy evaluation. Nevertheless, such methods rely on the coverage of the given dataset in the state-action space [11,12]. As a result, if the offline dataset for a specific task is of low quality or has limited data coverage, the performance of offline RL algorithms is limited.

To this end, our research aims to develop a Multi-Task Data Sharing (MTDS) method, which enhances the offline RL on a specific task by utilizing datasets from other relevant tasks. The multi-task datasets are typically accessible in practical offline RL problems. For example, one can collect different datasets for various tasks with the same robot arm, where each task has a relatively small dataset. In such a scenario, directly training individual policies for each task with the corresponding dataset is insufficient for good performance since the data coverage for a single dataset is limited. Thus, a natural idea is to utilize offline datasets from other tasks in the same domain to help each task learn better (namely, MTDS). In particular, in learning a task $A_i \in A$ from domain A with dataset \mathcal{D}_{A_i} , an MTDS method constructs a mixed dataset $\hat{\mathcal{D}}_{A_i}$ by relabeling experience $e_j = (s, a, r_{A_j}(s, a), s')$ from task $A_j \in A$ to $e_{j \rightarrow i} = (s, a, r_{A_i}(s, a), s')$, where $i \neq j$. The shared tasks come from the same domain with the same dynamics; thus we only need to modify the rewards in the data sharing. We then train the offline policy for task A_i on the mixed dataset $\hat{\mathcal{D}}_{A_i}$.

Though such an MTDS process is easy to implement, naively sharing data can exacerbate the distribution shift between behavior and learned policies, which degrades performance compared to single-task training. In addition, when incorporating policy constraints, the different behavior policies of the mixed dataset drive the learned policy towards different directions, which results in conflicting gradient directions and instability in optimization. Previous methods propose Conservative Data Sharing (CDS) [13] to address this problem, which minimizes distribution shift by defining an effective behavior policy and only sharing data relevant to the main task. However, such a method hinges on policy constraints, while we insist policy coverage instead of the optimality of behavioral policy suits better in MTDS. In addition, the data selection process of CDS discards a large amount of shared data that could be potentially informative in training.

In this paper, we propose Uncertainty-based Data Sharing (UTDS) algorithm, which allows arbitrary data sharing without data selection mechanisms. In particular, we first perform standard MTDS to obtain a mixed dataset with other tasks, then train an ensemble of Q -functions [14] to provide uncertainty quantification for the mixed dataset. By measuring the uncertainty, we perform pessimistic value iteration to utilize the uncertainty as a penalty in offline training. Such an uncertainty quantifier considers the data coverage of the mixed dataset, which is less affected by the deviation between the behavior policies and the learned policy. In addition, we further penalize the out-of-distribution (OOD) actions within the support of the mixed dataset to improve the performance of the learned policy in the OOD region. A key factor for our proposed UTDS is that even if data sharing does not benefit the data coverage of the optimal policy, it does not degrade the UTDS learning results. Such a key factor makes UTDS inherently different from CDS [13], which may have degraded performance without appropriate data selection. We illustrate the difference between CDS and UTDS in Fig. 1.

A preliminary version of this work appeared as a spotlight conference paper [15], which mainly considers single-task offline RL. Nevertheless, we study the multi-task setting and data sharing problem in the present work. Our contributions are summarized as follows.

- The proposed UTDS algorithm provides a unified view for both single-task and multi-task offline RL learning without data selection. In particular, when the number of shared datasets is equal to one (namely $m = 1$), UTDS degenerates to uncertainty-based single-task offline learning.

- We provide a novel theoretical analysis in linear MDPs for MTDS. The analysis shows that the optimality gap of UTDS in data sharing is only related to the expected data coverage of the shared dataset on the optimal policy. Thus, UTDS is less affected by the change of the behavior policy in data sharing.
- Empirically, we conduct experiments on a suite of domains built on DeepMind Control Suite [16] and collect multi-task datasets to construct a benchmark for large-scale MTDS. We release the collected datasets and open-source code for the data-generation, and also reproduce several MTDS baselines for further utilization.¹

2. Preliminaries

2.1. Offline Reinforcement Learning

We consider an episodic MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, T, r, P)$, where \mathcal{S} is the state space, \mathcal{A} is the action space, $T \in \mathbb{N}$ is the length of episodes, r is the reward function, and P is the transition distribution. The goal of RL is to find a policy π that maximizes the expected cumulative rewards $\mathbb{E}[\sum_{i=0}^{T-1} \gamma^i r_i]$, where $\gamma \in [0, 1)$ is the discount factor in episodic settings. The corresponding Q -function of the optimal policy satisfies the following Bellman operator, as

$$\mathcal{T}Q(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim T(\cdot|s, a)} \left[\max_{a'} Q^-(s', a') \right], \quad (1)$$

where θ is the parameters of Q -network. We remark that relaxing the deterministic assumption of reward function to a stochastic one does not affect our method. Nevertheless, a deterministic reward $r(s, a)$ is quite standard in offline RL literature, and existing offline RL benchmarks also have deterministic reward functions. In Deep RL (DRL), the Q -value is updated by minimizing the TD-error, namely $\mathbb{E}_{(s, a, r, s')} [(Q - \mathcal{T}Q)^2]$. Empirically, the target $\mathcal{T}Q$ is typically calculated by a separate target-network parameterized by θ^- without gradient propagation [2]. In online RL, one typically samples the transitions (s, a, r, s') through iteratively interacting with the environment. The Q -network is then trained by sampling from the collected transitions.

In offline RL, the agent is not allowed to interact with the environment. The experiences are sampled from an offline dataset. Naive off-policy methods such as Q -learning suffer from the *distributional shift*, which is caused by different visitation distribution of the behavior policy and the learned policy. Specifically, the greedy action a' chosen by the target Q -network in s' can be an OOD-action since (s', a') is scarcely covered by the dataset \mathcal{D}_{in} . Thus, the value functions evaluated on such OOD actions typically suffer from significant extrapolation errors. Such errors can be further amplified through propagation and potentially diverges during training.

2.2. Multi-Task Data Sharing (MTDS)

A multi-task Markov Decision Process (MDP) is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, \{R_i, i\}_{i=1}^m)$, where m is the number of tasks from the same domain, and we omit the domain notation \mathcal{A} . We remark that all tasks share the same state space, action space, and transition function. Meanwhile, different tasks have different reward functions $\{R_i\}$ indexed by $i \in [m]$. Correspondingly, $Q_i^{\pi_i}(a|s)$ is the value function of task $i \in [m]$ with respect to the policy π_i .

The multi-task dataset \mathcal{D} contains m per-task dataset $\mathcal{D} = \cup_{i=1}^m \mathcal{D}_i$, where $\mathcal{D}_i = \{(s, a, r_i, s')\}$ is the dataset for the i -th task and the reward function is returned by R_i . We denote the behavior policy for each dataset by $\{\pi_{\beta_1}, \dots, \pi_{\beta_m}\}$. For the i -th task, MTDS shares the dataset \mathcal{D}_j from the j -th task and relabels the rewards in \mathcal{D}_j with the reward function R_i . We denote the relabeled dataset from task $j \rightarrow i$ by $\mathcal{D}_{j \rightarrow i}$. Then $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ denotes the shared dataset that contains all shared data from task j , where $j \neq i$. To conduct data sharing, we assume that the reward functions for all tasks are known to us, which is commonly achievable in robotic locomotion and manipulation tasks [16]. In special cases, if such reward functions are otherwise unknown, one can learn a parametric reward function based on the dataset.

2.3. Conservative Data Sharing (CDS)

Previous work shows that naively sharing data to construct $\hat{\mathcal{D}}_i$ can degrade performance [13]. The reason is that data sharing increases the policy divergence $D(\pi_i \| \hat{\pi}_{\beta_i})(s)$ between the learned policy and the behavior policy as the shared data modifies $\hat{\pi}_{\beta_i}$. To this end, CDS optimizes the following objective to select the shared data, as

$$J_{\text{cds}} = \max_{\hat{\pi}_{\beta_i}} \mathbb{E}_{s \sim \hat{\mathcal{D}}_i} \left[\mathbb{E}_{a \sim \pi_i} [Q(s, a)] - \alpha D(\pi_i \| \hat{\pi}_{\beta_i})(s) \right], \quad (2)$$

where CDS selects data to obtain a behavior policy with small policy divergence. Theoretically, let π_i^{cds} and $\pi_{\beta_i}^*$ be policies that maximize Eq. (2), then *w.p.* $\geq 1 - \delta$, the policy improvement bound can be quantified as $J(\pi_i^{\text{cds}}) \geq J(\pi_{\beta_i}^*) - \delta_{\text{cds}}$, where

$$\delta_{\text{cds}} \approx \mathcal{O}(1/(1-\gamma)^2) \mathbb{E}_s \left[\sqrt{(D(\pi_i^{\text{cds}} \| \pi_{\beta_i}^*)(s) + 1) / |\hat{\mathcal{D}}_i(s)|} \right] - [\alpha D(\pi_i^{\text{cds}} \| \pi_{\beta_i}^*) + J(\pi_{\beta_i}^*) - J(\pi_{\beta_i})]. \quad (3)$$

¹ See <https://github.com/Baichenjia/UTDS> for the datasets and code.

According to Eq. (3), the learned policy should be close to the behavior policy to reduce this bound. We remark that such a lower bound is closely related to imitation learning rather than policy optimization since it only cares about the relationship to the behavior policy $\pi_{\beta_i}^*$. In contrast, our analysis of UTDS provides the optimality gap between the learned policy and the optimal policy, which provides useful guidance for data sharing without policy constraints.

In CDS implementation, optimizing J_{cds} can be approximately achieved by selecting transitions with a conservative Q -value more than the top k -th quantiles. Nevertheless, since the shared dataset changes with the update of Q -function, the learning process becomes unstable with the change of the effective behavior policy per update. Choosing a small k (e.g., 10%) may overcome such a problem, while it drops 90% of the shared dataset that may contain helpful information for RL.

3. Method

In this section, we introduce our proposed UTDS method. In UTDS, we collect all data from the shared dataset without data selection and measure the uncertainty through an ensemble of Q -networks. UTDS then performs the pessimistic value update to penalize the state-action pairs with large uncertainties in the shared dataset, which results in a pessimistic data-sharing policy.

3.1. Uncertainty quantifier

Considering a shared dataset $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ that contains data from i -th task and relabeled data from the j -th task. We parameterize the Q -function for task i as $Q_i(s, a; \theta_i)$, where θ_i denotes the parameters. The value function can be obtained via the following target,

$$\hat{\tau} Q_i(s, a) = r_i(s, a) + \gamma \mathbb{E}_{s' \sim \hat{\mathcal{D}}_i, a' \sim \pi_i} [Q_i^-(s', a')], \quad (4)$$

where we denote $\hat{\tau} Q_i$ as the target corresponding to the experience (s, a, r, s') sampled from the shared dataset $\hat{\mathcal{D}}_i$, and we select a' following the policy π_i . The target-network Q_i^- is parameterized by θ_i^- , which is updated by $\theta_i^- \leftarrow (1 - \tau)\theta_i + \tau\theta_i$ without gradient propagation.

To acquire the uncertainty quantification, we train N independent Q -networks $\{Q_i^1, \dots, Q_i^N\}$ with different random initializations and independent target networks. We then measure the uncertainty of $(s, a) \sim \hat{\mathcal{D}}_i$ by the standard deviation of the ensemble- Q predictions, namely,

$$\Gamma_i(s, a) = \sqrt{\mathbb{V}[Q_i^j(s, a)]}, \quad j \in [1, \dots, N], \quad (5)$$

where \mathbb{V} is the variance calculator. The deviation among deep ensemble- Q networks trained with independent initializations tends to be large if the corresponding state-action pair scarcely occurs in the dataset. Such a deviation shrinks as the sample size increases with the data sharing. Deep ensemble- Q networks trained with independent initializations will typically find different solutions, which corresponds to a variety of Q -functions centered in large basins of attraction [17].

From the Bayesian perspective, the obtained ensemble- Q networks is a non-parametric approximation of the posterior of the Q -function [18,19], which yields similar value on areas with rich data and diversely on areas with scarce data. Thus, the deviation among the ensemble Q -functions yields an epistemic uncertainty estimation, which we aim to utilize as a penalization in estimating the Q -functions. Here Γ_i captures the radius of the confidence interval for the Q -function, which measures the uncertainty in the shared datasets. Recent works [20,21] also show that ensemble methods with stochastic gradient descent (SGD) and independent initialization are well generalized, thus capturing the epistemic uncertainty of the OOD data. We remark that although the ensemble-based method has been adopted in single-task offline RL [22,15], we present the first practical algorithm for uncertainty-based MTDS in offline RL.

We illustrate the uncertainty estimation with ensembles on a prediction task in \mathbb{R}^2 plane, as shown in Fig. 2. The data are drawn from the Gaussian distributions, and the response is output by a random network. Fig. 2(a) shows the uncertainty measured by the ensemble in the single-task dataset, and we find that the uncertainty is small for the state-action pairs near the data cluster. Nevertheless, since the single dataset has limited coverage, most areas still have large uncertainty. In contrast, in Fig. 2(b), we share other two datasets and the uncertainty distribution changes. More areas are ‘known’ and with low uncertainty, which helps reduce the uncertainty of the state-action pairs induced by the optimal policy. As we will discuss later, if the data sharing improves the data coverage on the trajectory induced by the optimal policy, UTDS enjoys a tighter optimality bound and better performance. In contrast, CDS that measures the policy divergence fails to adapt to the dataset similarity measured in reference to estimating the optimal value function, which means divergence quantification with trajectory distribution is inaccurate for uncertainty quantification.

3.2. Pessimistic value iteration

Based on the uncertainty quantifier Γ_i in Eq. (5), UTDS learns a pessimistic value function on the shared dataset. The Bellman target for each Q -function in the ensemble takes the following form,

$$\hat{\tau}^{\text{UTDS}} Q_i(s, a) = r_i(s, a) + \gamma \mathbb{E}_{s' \sim \hat{\mathcal{D}}_i, a' \sim \pi_i} [Q_i^-(s', a') - \beta_1 \Gamma_i(s', a')], \quad (6)$$

where β_1 is a tuning parameter. Importantly, the uncertainty Γ_i quantifies the deviation of a datapoint from the shared dataset, and the penalization enforces pessimism for the target- Q value.

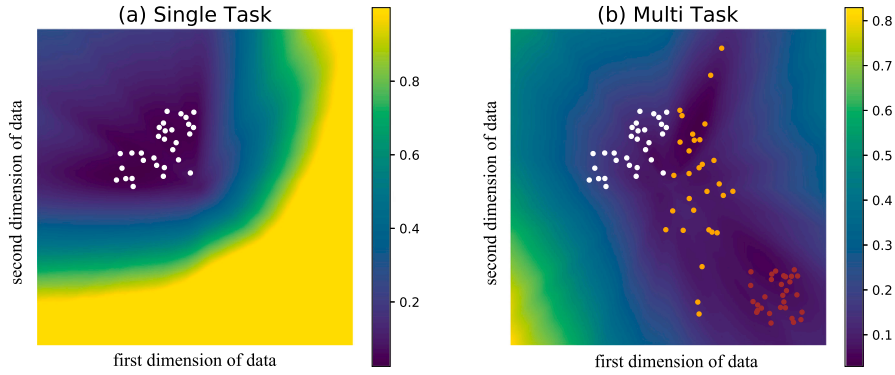


Fig. 2. An illustration of the uncertainty quantification of UTDS in (a) single-task dataset and (b) multi-task shared dataset. The multi-task datasets (i.e., the white, brown, and orange points) are generated by different Gaussian distributions. The uncertainty measured by the ensemble networks is represented by the color scales in the figures. The darker color means smaller uncertainty. As shown in the figures, sharing more data decreases the uncertainty. (For interpretation of the color(s), the reader is referred to the web version of this article.)

We remark that there are two options to penalize the Q -functions through the operator $\hat{\mathcal{T}}^{\text{in}}$. In Eq. (6), we penalize the next- Q value $Q_i^-(s', a')$ with the corresponding uncertainty $\Gamma_i(s', a')$. Alternatively, we can also penalize the immediate reward by $\hat{r}(s, a) := r(s, a) - \Gamma_i(s, a)$ and use $\hat{r}(s, a)$ in place of $r(s, a)$ for the target. Nevertheless, since the datapoint $(s, a) \in \mathcal{D}_i$ lies on rich-data areas, the penalization $\Gamma_i(s, a)$ on the immediate reward is usually very small thus having less effect in training. In UTDS, we penalize the uncertainty of (s', a') in the next- Q value.

Nevertheless, the target in Eq. (6) only captures the uncertainty of experiences from the in-distribution data. Our empirical findings suggest that solely penalizing the uncertainty for in-distribution target is insufficient to control the OOD performance of the fitted Q -functions. In addition, we also need to penalize the value function for OOD actions. In UTDS, we sample OOD actions $a^{\text{ood}} \sim \pi_i(s)$ following the learned policy, and the OOD states are still sampled from the shared dataset as $s \sim \hat{\mathcal{D}}_i$. We highlight that such OOD sampling requires only the offline dataset $\hat{\mathcal{D}}_i$ and does not require additional generative models or access to the simulator.

It remains to design the target for OOD samples. Since the transition $P(\cdot|s, a^{\text{ood}})$ and reward $r(s, a^{\text{ood}})$ are unknown, the true target of OOD sample is inapplicable. In UTDS, we propose a pseudo-target for the OOD datapoints. We quantify the uncertainty $\Gamma_i(s, a^{\text{ood}})$ of the OOD data through the ensemble networks and use the following loss function to enforce pessimism for the OOD data,

$$\mathcal{L}_{\text{ood}}(s) = \mathbb{E}_{s \sim \hat{\mathcal{D}}_i, a^{\text{ood}} \sim \pi_i} \left[\left(\hat{\mathcal{T}}^{\text{ood}} Q_i(s, a^{\text{ood}}) - Q_i(s, a^{\text{ood}}) \right)^2 \right], \quad (7)$$

where $\hat{\mathcal{T}}^{\text{ood}} Q_i(s, a^{\text{ood}})$ is the pseudo-target for the OOD datapoints as,

$$\hat{\mathcal{T}}^{\text{ood}} Q_i(s, a^{\text{ood}}) := Q_i(s, a^{\text{ood}}) - \beta_2 \Gamma_i(s, a^{\text{ood}}), \quad (8)$$

where β_2 is a tuning parameter, and we remark $\hat{\mathcal{T}}^{\text{ood}} Q_i(s, a^{\text{ood}})$ is the target without gradient propagation. Our approach to OOD uncertainty quantification is inspired by PBRL [15], which we extend to the MTDS settings. As illustrated in Fig. 2, Γ_i provides well-calibrated uncertainty for both the in-distribution data and the OOD data.

We find that β_2 is important to the empirical performance. Specifically,

- At the beginning of training, both the Q -functions and the corresponding uncertainty quantifications are inaccurate. We use a large β_2 to enforce a strong regularization on OOD actions.
- We then decrease β_2 exponentially with a factor α since the value estimation and uncertainty quantification becomes more accurate in training, which also prevents overly pessimism in the training process. We remark that a smaller β_2 requires more accurate uncertainty estimate for the pessimistic target estimation $\hat{\mathcal{T}}^{\text{ood}} Q_i$.

We remark that applying the operator for infinite times to $Q_i(s, a^{\text{ood}})$ leads to $(\hat{\mathcal{T}}^{\text{ood}})^{\infty} Q_i(s, a^{\text{ood}})$, which equals

$$Q_i(s, a^{\text{ood}}) - \sum_{k=0}^{\infty} \alpha^k \beta_2 \Gamma_i(s, a^{\text{ood}}) = Q_i(s, a^{\text{ood}}) - \frac{\beta_2}{(1-\alpha)} \Gamma_i(s, a^{\text{ood}}).$$

Thus, iteratively applying the OOD operator does not lead to the negative infinity target. In addition, a decaying parameter β_2 stabilizes the convergence of $Q_i(s, a^{\text{ood}})$ in the training from the empirical perspective. Meanwhile, the operator $\hat{\mathcal{T}}^{\text{UTDS}} Q_i(s, a)$ is a contraction mapping (see Appendix A.5 for a proof), which ensures that $Q_i(s, a)$ also converges to a fixed point.

To conclude, when training the i -th task, the loss function for the n -th Q -function in the ensemble is

$$\mathcal{L}_{\text{UTDS}}^n = \mathbb{E}_{s, a, r, s' \sim \hat{\mathcal{D}}_i} \left[\left(\hat{\mathcal{T}}^{\text{UTDS}} Q_{\theta_i}^n(s, a) - Q_{\theta_i}^n(s, a) \right)^2 \right] + \mathcal{L}_{\text{ood}}(s), \quad (9)$$

where $\hat{\mathcal{T}}^{\text{UTDS}}$ and \mathcal{L}_{ood} are defined in Eq. (6) and Eq. (7), respectively. Each function $Q_{\theta_i}^n$ has its own target-network, where $n \in [N]$. The policy $\pi_{\phi_i}(s)$ for the i -th task is updated by minimizing the loss function as

Algorithm 1 UTDS for training the i -th task.

```

1: Initialize:  $N$  ensemble- $Q$  and target- $Q$  nets with parameter  $\theta$  and  $\theta^-$ , initialize the policy  $\pi_{\phi_i}$ ;
2: while not coverage do
3:   Sample a batch of transition  $(s, a, r, s')$  from  $\hat{\mathcal{D}}_i$  that contains shared data from other tasks;
4:   Perform OOD sampling based on the current policy to obtain  $(s, a^{\text{ood}})$  pairs;
5:   Calculate the uncertainty  $\Gamma_i(s, a)$  and  $\Gamma_i(s, a^{\text{ood}})$  based on the ensemble- $Q$  networks;
6:   Preform pessimistic update via  $\hat{\mathcal{T}}^{\text{UTDS}}$  and  $\hat{\mathcal{T}}^{\text{ood}}$  to train the critic parameters  $\theta_i$  via Eq. (9);
7:   Train the pessimistic policy to train the actor parameters  $\phi_i$  via Eq. (10);
8: end while

```

$$\mathcal{L}_{\text{policy}} = -\mathbb{E}_{s \sim \hat{\mathcal{D}}_i, a \sim \pi_{\phi_i}} \left[\min_{n=1, \dots, N} Q_{\theta_i^n}(s, a) \right], \quad (10)$$

which takes a greedy action based on the minimum of the ensemble Q -functions. See Algorithm 1 for a summary of UTDS.

4. Theoretical analysis

In this section, we consider linear MDPs [23,11] as a simplification, where we have access to a feature map of the state-action pair, $\phi(s, a) : S \times \mathcal{A} \rightarrow \mathbb{R}^d$. Meanwhile, the transition kernel and the reward function are linear in $\phi(s, a)$. We estimate the action value function by $Q_i(s, a) \approx \hat{w}_i^\top \phi(s, a)$. See Appendix A for the details.

4.1. UTDS in linear MDPs

From the theoretical perspective, an appropriate uncertainty quantification is essential to the provable efficiency in offline RL. Specifically, the ξ -uncertainty quantifier plays a central role in the analysis of both online and offline RL [24–26,23,11,27,12].

Definition 1 (ξ -Uncertainty Quantifier [11]). The set of penalization $\{\Gamma_t\}_{t \in [T]}$ forms a ξ -Uncertainty Quantifier if it holds with probability at least $1 - \xi$ that

$$|\hat{\mathcal{T}}V_{t+1}(s, a) - \mathcal{T}V_{t+1}(s, a)| \leq \Gamma_t(s, a)$$

for all $(s, a) \in S \times \mathcal{A}$, where \mathcal{T} is the Bellman equation and $\hat{\mathcal{T}}$ is the empirical Bellman equation that estimates \mathcal{T} based on the offline data.

In linear MDPs [23,26,11] where the transition kernel and reward function are assumed to be linear to the state-action representation $\phi : S \times \mathcal{A} \rightarrow \mathbb{R}^d$, The following lower confidence bound (LCB)-penalty [28,23] is known to be a ξ -uncertainty quantifier for appropriately selected $\{\beta_t\}_{t \in [T]}$,

$$\Gamma^{\text{lb}}(s_t, a_t) = \beta_t \cdot [\phi(s_t, a_t)^\top \Lambda_i^{-1} \phi(s_t, a_t)]^{1/2}, \quad (11)$$

where $\Lambda_i = \sum_{k=1}^{|\mathcal{D}_i|} \phi(s_t^k, a_t^k) \phi(s_t^k, a_t^k)^\top + \lambda \cdot \mathbf{I}$ is the covariance matrix corresponding to the dataset of task i , which plays the role of a pseudo-count intuitively. We remark that under such linear MDP assumptions, the penalty proposed in UTDS and $\Gamma^{\text{lb}}(s_t, a_t)$ in linear MDPs is equivalent under a Bayesian perspective. Specifically, we make the following claim.

Claim 1. In linear MDPs, the proposed bootstrapped uncertainty $\beta_t \cdot \Gamma(s_t, a_t)$ is an estimation to the LCB-penalty $\Gamma^{\text{lb}}(s_t, a_t)$ in Eq. (11) for an appropriately selected tuning parameter β_t .

The LCB in Eq. (11) can only be applied in linear cases. We refer to Appendix A for a detailed explanation and proof. Intuitively, the ensemble Q -functions estimates a non-parametric Q -posterior [14,29]. Correspondingly, the uncertainty quantifier $\Gamma(s_t, a_t)$ estimates the standard deviation of the Q -posterior, which scales with the LCB-penalty in linear MDPs. As an example, we show that under the tabular setting, $\Gamma^{\text{lb}}(s_t, a_t)$ is approximately proportional to the reciprocal pseudo-count of the corresponding state-action pair in the dataset. In offline RL, such uncertainty quantification measures how trustworthy the value estimations on state-action pairs are. A low LCB-penalty (or high pseudo-count) indicates that the corresponding state-action pair aligns with the support of offline data.

In the generic MDP settings, we estimate the uncertainty based on the non-parametric ensemble Q networks. Let us consider a scenario where we share a dataset \mathcal{D}_j to the i -th task and obtain $\mathcal{D}_{j \rightarrow i}$. Correspondingly, we solve the offline RL with the joint dataset $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$. The parameter of the Q -function Q_i learned from $\hat{\mathcal{D}}_i$ can be solved in a closed form following the least squares value iteration algorithm (LSVI), which iteratively minimizes the following loss function,

$$\tilde{w}_{ij} = \min_{w \in \mathbb{R}^d} \sum_{k=1}^{|\hat{\mathcal{D}}_i|} (\phi(s_t^k, a_t^k)^\top w - r(s_t^k, a_t^k) - V_{t+1}(s_{t+1}^k))^2 + \sum_{\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_{j \rightarrow i}^{\text{ood}}} (\phi(s^{\text{ood}}, a^{\text{ood}})^\top w - y^{\text{ood}})^2, \quad (12)$$

where the data in the first and second terms are based on the shared dataset and the OOD data, respectively. We use $\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_j^{\text{ood}}$ to represent the OOD data from the shared dataset for convenience. As discussed in §3, we sample s^{ood} from $\hat{\mathcal{D}}_i$, and sample a^{ood} by following the learned policy π_i . The target y^{ood} defined in Eq. (8) does not rely on the relabeled reward. The solution of \tilde{w}_{ij} is

$$\tilde{w}_{ij} = \tilde{\Lambda}_{ij}^{-1} \left(\sum_{k=1}^{|\hat{\mathcal{D}}_i|} \phi(s_i^k, a_i^k) y_i^k + \sum_{\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_j^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) y^{\text{ood}} \right), \quad (13)$$

where the covariance matrix $\tilde{\Lambda}_{ij}$ for the mixed data $\hat{\mathcal{D}}_i$ is

$$\tilde{\Lambda}_{ij} = (\Lambda_i + \Lambda_i^{\text{ood}}) + (\Lambda_j + \Lambda_j^{\text{ood}}) := \tilde{\Lambda}_i + \tilde{\Lambda}_j, \quad (14)$$

where $\Lambda_i = \sum_{k=1}^{|\mathcal{D}_i|} \phi(s_i^k, a_i^k) \phi(s_i^k, a_i^k)^\top$ is the covariance matrix of \mathcal{D}_i , and Λ_j is the covariance matrix of $\mathcal{D}_{j \rightarrow i}$. Meanwhile, $\Lambda_i^{\text{ood}} = \sum_{\mathcal{D}_i^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) \phi(s^{\text{ood}}, a^{\text{ood}})^\top$ is the covariance matrix of OOD data $\mathcal{D}_i^{\text{ood}}$, and Λ_j^{ood} is the covariance matrix of $\mathcal{D}_j^{\text{ood}}$. Based on the solution, the following theorem establishes the relationship between the uncertainty in single dataset and shared dataset.

Theorem 1. For a given state-action pair (s, a) , we denote the uncertainty for the single-task dataset \mathcal{D}_i and shared dataset $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ as $\Gamma_i(s, a; \mathcal{D}_i)$ and $\Gamma_i(s, a; \hat{\mathcal{D}}_i)$, respectively. Then the following inequality holds as

$$\Gamma_i^{\text{LCB}}(s, a; \mathcal{D}_i) \geq \Gamma_i^{\text{LCB}}(s, a; \hat{\mathcal{D}}_i), \text{ and } \Gamma_i(s, a; \mathcal{D}_i) \geq \Gamma_i(s, a; \hat{\mathcal{D}}_i), \quad (15)$$

where signifies that the shared data reduce the ensemble uncertainty.

We refer to Appendix A for a proof. Theorem 1 shows that the uncertainty decreases with more shared data, which is also illustrated in Fig. 2. If a (s, a) pair scarcely occurs in the single-task dataset, the LCB penalty $\Gamma_i(s, a)$ will be high and the corresponding $Q_i(s, a)$ function is pessimistic, which makes the agent have a low probability of choosing this action. Nevertheless, such pessimism comes from the lack of knowledge of the environment, which does not indicate a is actually a bad choice. In UTDS, the penalty $\Gamma_i(s, a)$ will gradually decrease with more shared (s, a) (or similar) samples, which makes the value function become less pessimistic and extends the agent's knowledge of the environment.

4.2. Optimality gap

Recent theoretical analysis shows that appropriate uncertainty quantification leads to provable efficiency in offline RL [27,15]. In particular, the pessimistic value iteration [11] with a general ξ -uncertainty quantifier as the penalty achieves provably efficient pessimism in offline RL. In linear MDPs, the LCB-penalty defined in Eq. (11) is known to be a ξ -uncertainty quantifier for the appropriately selected factor $\{\beta_t\}_{t \in [T]}$. In the following, we show that UTDS with the shared dataset $\hat{\mathcal{D}}_i$ also forms a valid ξ -uncertainty quantifier. We denote $\mathcal{T}V_{t+1}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim P_t(\cdot|s_t, a_t)} [V_{t+1}(s_{t+1}, a_{t+1})]$ as the true Bellman target.

Theorem 2. Let $\tilde{\Lambda}_{ij} \geq \lambda \cdot \mathbf{I}$, if we set the OOD target as $y^{\text{ood}} = \mathcal{T}V_{t+1}(s^{\text{ood}}, a^{\text{ood}})$ for the shared dataset $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$, then it holds for $\beta_i = O(T \cdot \sqrt{d} \cdot \log(T/\xi))$ that

$$\Gamma_i^{\text{LCB}}(s_t, a_t; \hat{\mathcal{D}}_i) = \beta_i [\phi(s_t, a_t)^\top \tilde{\Lambda}_{ij}^{-1} \phi(s_t, a_t)]^{1/2}$$

forms a valid ξ -uncertainty quantifier, where $\tilde{\Lambda}_{ij}$ is the covariance matrix given in Eq. (14).

We refer to Appendix A for a proof. Based on Theorem 2, we further characterize the optimality gap of UTDS based on the pessimistic value iteration as follows.

Corollary 1. Under the same conditions as Theorem 2, for the uncertainty quantification $\Gamma_i(s, a; \mathcal{D}_i)$ and $\Gamma_i(s, a; \hat{\mathcal{D}}_i)$ defined in \mathcal{D}_i and $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ respectively, we have

$$\text{SubOpt}(\pi_i^*, \tilde{\pi}_i) \leq \sum_{t=1}^T \mathbb{E}_{\pi_i^*} [\Gamma_i^{\text{LCB}}(s_t, a_t; \hat{\mathcal{D}}_i)] \leq \sum_{t=1}^T \mathbb{E}_{\pi_i^*} [\Gamma_i^{\text{LCB}}(s_t, a_t; \mathcal{D}_i)], \quad (16)$$

where $\tilde{\pi}_i$ and π_i^* are the learned policy and the optimal policy in $\hat{\mathcal{D}}_i$, respectively.

The first inequality can be directly obtained from [11] since the covariance matrix $\tilde{\Lambda}_{ij}$, the learned policy $\tilde{\pi}_i$ and the optimal policy π_i^* are all defined in the shared dataset $\hat{\mathcal{D}}_i$. The second upper bound follows from the fact that $\Gamma_i(s_t, a_t; \hat{\mathcal{D}}_i) \leq \Gamma_i(s_t, a_t; \mathcal{D}_i)$ (Theorem 1), namely, a larger dataset leads to tighter uncertainty. Corollary 1 shows that the optimality gap shrinks if the data coverage of the optimal policy π^* is better, in the sense that the expected uncertainty under the optimal policy π^* is smaller.

According to Corollary 1, we remark that having more data does not guarantee a strict decrease in the optimality gap. As an example, let the i -th task and the j -th task be irrelevant in the sense that the shared dataset $\mathcal{D}_{j \rightarrow i}$ does not contain any transitions induced by the optimal policy of the task i . Then, such data sharing does not help reduce the sub-optimality as the expected

uncertainty with respect to the optimal policy π_i^* in $\hat{\mathcal{D}}_i$ is the same as that in \mathcal{D}_i . Nevertheless, data sharing does not decrease the learning outcome in such a case, which makes UTDS different from previous methods [13,30]. In particular, in policy constraint-based methods, sharing data across irrelevant tasks may exacerbate the distribution shift and hinder learning performance.

5. Related work

Offline RL. Since the offline RL dataset often has limited coverage, the actions chosen by greedy policy could be OOD actions that are not contained in the dataset, and the corresponding value function typically suffers from a large extrapolation error. This problem is also known as the distribution shift problem [10]. Previous model-free offline RL algorithms typically rely on policy constraints to restrict the learned policy from producing the OOD actions. In particular, previous works add behavior cloning (BC) loss in policy training [9,31,32], measure the divergence between the behavior policy and the learned policy [10,33,34], apply advantage-weighted constraints to balance BC and advantages [35,36], penalize the prediction-error of a variational auto-encoder [37], and learn latent actions (or primitives) from the offline data [38,39]. Nevertheless, such methods may cause overly conservative value functions and are easily affected by the behavior policy [40,41]. We remark that the OOD actions that align closely with the support of offline data could also be trustworthy. CQL [42] directly minimizes the Q -values of OOD samples and thus casts an implicit policy constraint. Our method is related to CQL in the sense that both UTDS and CQL enforce conservatism in Q -learning. In contrast, we conduct explicit uncertainty quantification for OOD actions, while CQL penalizes the Q -values of all OOD samples equally. However, directly adopting such policy constraints to MTDS is challenging due to the possibly diverse behavioral policies from other tasks, yielding loose policy constraints. Another line of research adopts uncertainty-based pessimism to penalize state action pairs with large uncertainty [43,44,15,45]. Our algorithm is inspired by the uncertainty-based pessimism methods in single-task offline RL.

In contrast with model-free algorithms, model-based algorithms learn the dynamics model directly with supervised learning. Similar to our work, MOPO [44] and MOREL [46] incorporate ensembles of dynamics-models for uncertainty quantification, and penalize the value function through pessimistic updates. Other than the uncertainty quantification, previous model-based methods also attempt to constrain the learned policy through BC loss [47], advantage-weighted prior [48], CQL-style penalty [49], and Riemannian submanifold [50]. Decision Transformer [51] builds a transformer-style dynamic model and casts the problem of offline RL as conditional sequence modeling. However, such model-based methods may suffer from additional computation costs and may perform suboptimally in complex environments [52,53]. In contrast, UTDS conducts model-free learning and is less affected by such challenges.

Data Sharing in RL. Previous data sharing methods focus on solving online multi-task and multi-goal challenges. The intuition is using experiences from relevant tasks to solve the given task, where the relevant tasks are typically selected by human knowledge [54] or hindsight inverse RL [55,56]. Direct data sharing on offline RL fails due to distribution shift. CDS [13] designs the selection criteria and selects the data relevant to the main task. CDS-Zero [30] removes the relabeling process by setting the shared reward to zero to reduce the bias in the reward. The CDS-based methods discard a large amount of data and can be computationally inefficient, since the selection changes with the update of value functions. Other methods also study data sharing in meta-RL [57] and dynamics adaptation [58,59], where the main and shared tasks have different dynamics. On the contrary, we study the data sharing among tasks from the same domain.

Uncertainty Quantification. Our method is related to the previous online RL exploration algorithms based on uncertainty quantification, including bootstrapped Q -networks [60,61], ensemble dynamics [62], Bayesian NN [63,64], and distributional value functions [65–68]. Uncertainty quantification is more challenging in offline RL than its online counterpart due to the limited coverage of offline data and the distribution shift of the learned policy. In model-based offline RL, MOPO [44] and MOREL [46] incorporate ensemble dynamics-model for uncertainty quantification. BOPAH [69] and MOOSE [70] combine uncertainty penalization and behavior-policy constraints. In model-free offline RL, UWAC [43] adopts dropout-based uncertainty [71] while relying on policy constraints in learning value functions. In contrast, UTDS does not require additional policy constraints. In addition, according to the study in image prediction with data shift [20], the bootstrapped uncertainty is more robust to data shift than the dropout-based approach [72,73]. EDAC [22] is a concurrent work that uses the ensemble Q -network. Specifically, EDAC calculates the gradients of each Q -function and diversifies such gradients to ensure sufficient penalization for OOD actions. In contrast, UTDS penalizes the OOD actions through direct OOD sampling and the associated uncertainty quantification.

6. Experiments

In experiments, we provide a suite of domains built on DeepMind Control Suite [16] and collect multi-task datasets to construct a benchmark for large-scale MTDS. We do not use the standard D4RL [74] environment since D4RL has a collection of single task datasets.

6.1. Tasks and datasets

The environment contains 3 domains with 4 tasks per domain, resulting in 12 tasks in total. (i) **Walker** (*Stand, Walk, Run, Flip*) tasks aim to control a biped in a 2D vertical plane. Different tasks learn different balancing and locomotion skills. (ii) **Quadruped** (*Jump, Roll-Fast, Walk, Run*) tasks aim to control a quadruped within a 3D space. Quadruped learns different moving and balancing skills in a 3D space, which is harder than Walker because of the high-dimensional state-action space. (iii) **Jaco Arm** (*Reach top left*,

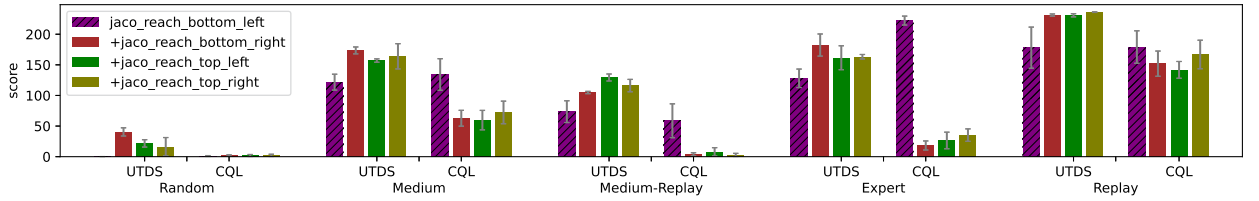


Fig. 3. The comparison between UTDS and Direct Sharing in *Jaco Arm*. The main task is *Reach-Bottom-Left*, and the shared data are replay datasets from the other three tasks (denoted as '+'). We show results of 5 dataset types of the main task. The shadow bars show the single-task scores.

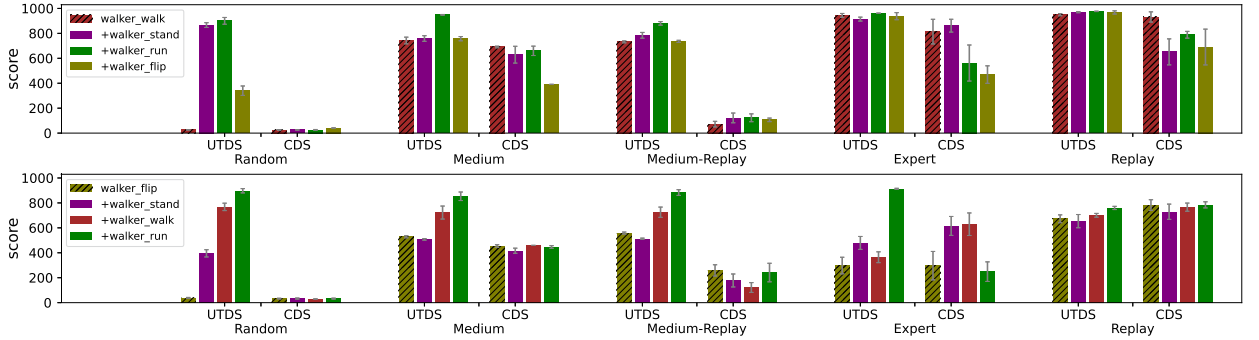


Fig. 4. The comparison between UTDS and CDS in *Walker*. The main tasks are *Walker-Walk* (Top) and *Walker-Flip* (Bottom), respectively. UTDS generally improves the performance compared to the single-task scores (i.e., the shadow bar), especially for non-expert datasets.

Reach top right, Reach bottom left, Reach bottom right) tasks aim to control a 6-DOF robotic arm with a three-finger gripper to move to different positions.

For each task, we run the TD3 algorithm [75] to collect five types of datasets (i.e., *random*, *medium*, *medium-replay*, *replay*, and *expert*) by following the standard settings in offline RL. Each dataset contains 10^3 episodes of interactions. For MTDS, we share the dataset of each task with the other 3 tasks of the same domain. Meanwhile, since all tasks are associated with 5 different types of data, the data sharing setup leads to a variety of combinations. To reduce the computation burden, we only share the *replay* dataset for each other tasks, which ensures sufficient coverage of the shared dataset. As a result, we have $4 \times 3 \times 5 = 60$ two-task sharing settings for each domain.

We follow the setting of D4RL to collect the multi-task datasets. (i) The ‘medium’ dataset is generated by training a policy online using TD3, early-stopping the training, and collecting samples from this partially-trained policy. (ii) The ‘random’ datasets are generated by unrolling a randomly initialized policy. (iii) The ‘medium-replay’ dataset consists of recording all samples in the replay buffer observed during training until the policy reaches the ‘medium’ level of performance. (iv) The ‘expert’ dataset is generated by an expert policy trained by a TD3 agent. (v) The ‘medium-expert’ dataset is collected by mixing equal amounts of expert demonstrations and suboptimal data, generated via a partially trained policy or by unrolling a uniform-at-random policy.

Baselines We compare the proposed UTDS with several baselines, including (i) **Direct Sharing** (i.e., CQL) that shares all relabeled data from \mathcal{D}_j and trains the policy through CQL [42]; (ii) **CDS** [13] that selects transitions with conservative Q -values within the highest 10% quantiles of $\mathcal{D}_{j \rightarrow i}$, where the conservative Q -value is learned by CQL; and (iii) **CDS-Zero** [30] that conservatively sets the shared reward to zero in CDS. Since the original work of CDS or CDS-Zero does not release the code, we implement the algorithms in our tasks and datasets based on their original papers. See Appendix B for implementation details of UTDS and baselines.

6.2. Experimental results

Directly sharing in Jaco Fig. 3 shows the result of single-task training, Direct Sharing, and UTDS in *Jaco Arm* domain. Since the different tasks in *Jaco Arm* require moving the robot arm toward different directions, the behavior policies among the tasks differs significantly. The diversity in behavior policies exacerbate the distribution shift in policy constraints, which makes Direct Sharing performs poor in the single-task training. In contrast, the diversity in behavior policies benefit the performance of UTDS. As the result, we observe improved performance of UTDS with data sharing in all settings. We defer the complete results to Appendix C.

Data sharing in Walker We show the result of comparisons between UTDS and CDS in *Walker-Walk* and *Walker-Flip* in Fig. 4. We find that UTDS generally improves the single-task performance via data sharing. The improvement is significant for non-expert datasets (i.e., *random*, *medium*, and *medium-replay*) generated by sub-optimal policies, which barely cover the optimal trajectories for the main task. As discussed in Corollary 1, since the shared data potentially contains trajectories induced by the optimal policy π_i^* , UTDS achieves a tighter sub-optimality bound $\mathbb{E}_{\pi_i^*}[\Gamma_i(s, a)]$ and attains better performance than CDS.

For *expert* and *replay* datasets that cover the optimal trajectories sufficiently well, the improvement of data sharing with UTDS

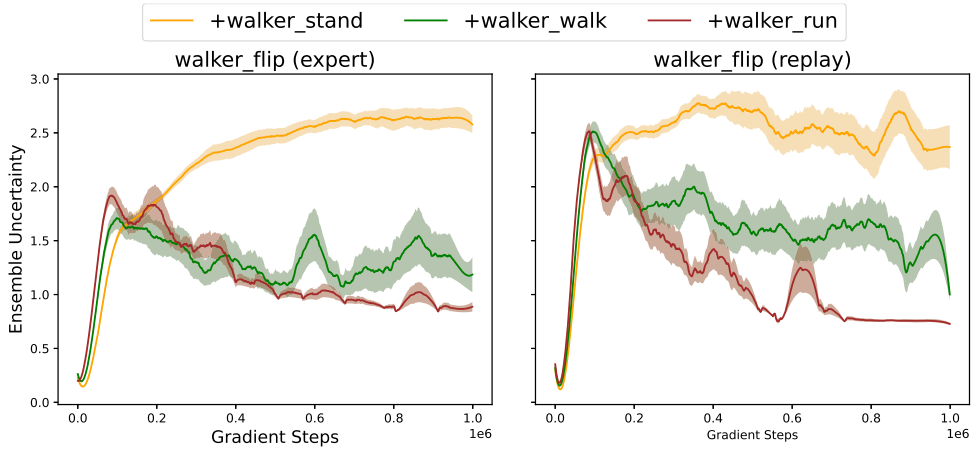


Fig. 5. The illustration of ensemble uncertainty of UTDS calculated on training batches. The main task is *Walker Flip* with *expert* (left) and *replay* (right) datasets.

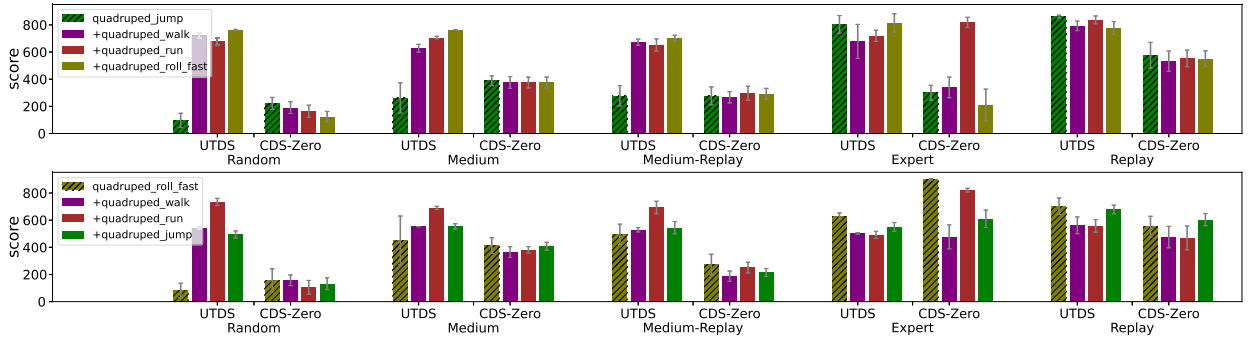


Fig. 6. The comparison between the proposed UTDS and CDS-Zero in *Quadruped*. The main tasks are *Quadruped-Jump* (Top) and *Quadruped-Roll-Fast* (Bottom), respectively.

is marginal for most cases since it is harder for UTDS to reduce the expected uncertainty $\mathbb{E}_{\pi^*}[\Gamma_i(s, a)]$ with the shared data. The only exception is *Walker-Flip* (*expert* and *replay*) tasks, where sharing the *Walker-Run* dataset improves the performance of UTDS significantly. Intuitively, since the two tasks are closely related, experiences from *Walker-Run* can potentially help the agent in learning complex locomotion skills and improve the learning of *Walker-Flip*. Fig. 5 shows that data sharing with *Walker-Run* leads to less uncertainty compared to other shared tasks, which results in tighter sub-optimality bound and better performance. We defer the complete results and the comparison of CDS-Zero in Appendix C. We find CDS-Zero performs very similarly to CDS.

Data sharing in Quadruped We show the comparison in *Quadruped Jump* and *Quadruped Roll-Fast* in Fig. 6. We find UTDS consistently improves the performance compared to single-task training in *random*, *medium*, and *medium-replay* datasets. In contrast, CDS-Zero performs poorly in most sharing tasks due to the significant distribution shift induced by the difference among tasks.

Nevertheless, we find that the performance of UTDS is slightly decreased when the main task is *expert* or *replay* comparing with the single-task training. We hypothesize that such a phenomenon is also caused by inaccurate feature representation in uncertainty estimation. In particular, our theory requires that the feature $\phi(s, a)$ is accurate for uncertainty quantification, which is not the case in our implementation since we use a deep neural network to learn the representation. Since the *Quadruped* domain has high-dimensional states, the representation learning in the *Quadruped* domain is more challenging. We leave this for future research. The complete results are given in Appendix C. One possible remedy is combining UTDS with representation learning methods such as contrastive learning [76] and bootstrapping [77], which we leave as future research directions. The complete results are given in Appendix C.

6.3. Visualization of the uncertainty

We visualize the change of uncertainty with data sharing in Walker and Quadruped domains, as shown in Fig. 7 and Fig. 8, respectively. We use PCA to project the state-action pairs to two-dimensional space. In both figures, the white dots denote data samples from the main task, and the orange dots denote data sampled from the shared task. We use the ensemble Q -networks trained by UTDS on the main task and shared dataset to illustrate the change of uncertainty in data sharing. According to the result, we find (i) in single-task results, the shared data can be considered as OOD data and have large uncertainty; (ii) in shared-task results, the uncertainty of the main task data slightly increases with data sharing, and the uncertainties of the shared data are significantly

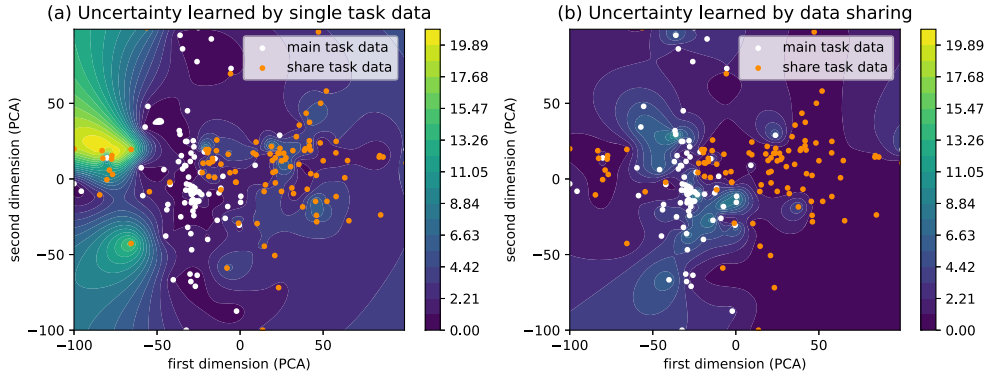


Fig. 7. Visualization of the uncertainty in data sharing for Walker domain. We share data from Walker-Run (replay) to Walker-Flip (medium-replay) task. (a) The ensemble Q -networks are trained in (a) single-task data and (b) shared dataset. We evaluate the uncertainty for both the main task data and the shared data.

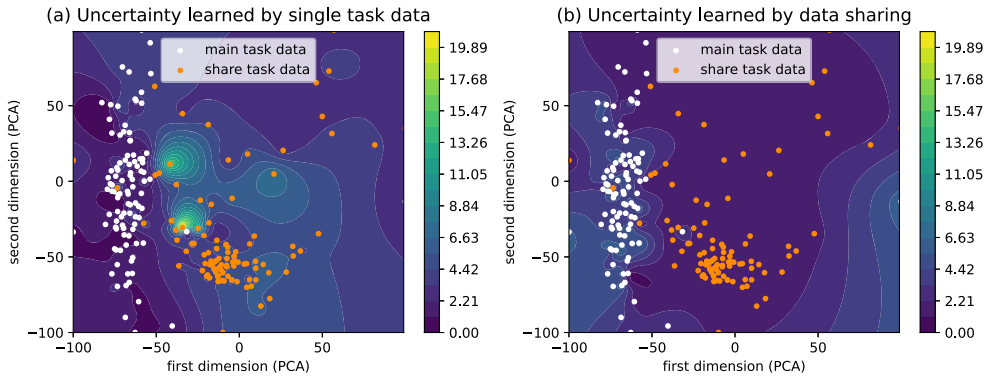


Fig. 8. Visualization of the uncertainty in data sharing for Quadruped domain. We share data from Quadruped Roll-Fast (replay) to Quadruped Jump (medium) task. The ensemble Q -networks are trained in (a) single-task data and (b) shared dataset. We find the uncertainties in for many areas are largely reduced with data sharing.

decreased. According to our analysis, the reduced uncertainty makes the value function less pessimistic and also shrinks the suboptimality bound, thus extending the agent's knowledge of the environment.

We remark that although the uncertainties quantifier $\Gamma^{\text{icb}}(s, a) = [\phi(s, a)^T \Lambda^{-1} \phi(s, a)]^{1/2}$ for any (s, a) pair is guaranteed to reduce with data sharing in linear and tabular MDPs, we find it is somewhat defective. According to the visualization, the uncertainties for some data points from the main task slightly increase. Such effect occurs for several reasons. (i) The linear and tabular MDP settings assume that $\phi(s, a)$ is known and fixed in learning. In contrast, for with high-dimensional problem, $\phi(s, a)$ is randomly initialized by a neural network and is often learned with the value function, which makes the uncertainty estimation inaccurate. (ii) We use ensemble Q -networks to estimate the uncertainty and prove that the LCB-penalty and ensemble uncertainty are equivalent under mild conditions. In practice, the ensemble uncertainty is accurate with infinite networks, while it is unachievable and we only use 5 ensembles in our experiments. The empirical result shows 5 ensembles are sufficient to provide strong performance.

7. Conclusion

In this paper, we propose UTDS algorithm for multi-task offline RL to resolve the distribution shift in data sharing. We apply ensemble Q -networks for uncertainty quantification and obtain a pessimistic value function for state-action pairs with low data coverage. Our theoretical result shows that the suboptimality gap of UTDS is related to the expected uncertainty of the shared dataset. Experiments on the proposed benchmark show that UTDS outperforms other data-sharing algorithms in various scenarios. We remark that although UTDS may be slightly weaker in expert or replay dataset, it is of less concern in practical applications as the single-task RL algorithms with uncertainty penalty or policy constraints are sufficient to establish strong performances.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62306242), the National Key R&D Program of China (Grant No. 2022ZD0160100), and Shanghai Artificial Intelligence Laboratory.

Appendix A. Theoretical analysis

In this section, we give a detailed theoretical analysis of UTDS in linear MDPs [23].

A.1. UTDS for single-task linear MDPs

In linear MDPs, we assume that the transition dynamics and reward function are linear to the state-action feature embedding $\phi : S \times \mathcal{A} \mapsto \mathbb{R}^d$, as

$$\mathbb{P}_t(s_{t+1} | s_t, a_t) = \langle \psi(s_{t+1}), \phi(s_t, a_t) \rangle, \quad r(s_t, a_t) = \theta^\top \phi(s_t, a_t), \quad \forall (s_{t+1}, a_t, s_t) \in S \times \mathcal{A} \times S, \quad (\text{A.1})$$

where we assume the feature is bounded by $\|\phi\|_2 \leq 1$, and the reward function $r : S \times \mathcal{A} \mapsto [0, 1]$ is also bounded. Since the transition function and reward function are linear to ϕ , the state-action value function for any policy π is also linear in ϕ , as $Q_t(s_t, a_t) = w_t^\top \phi(s_t, a_t)$. In the following, we omit the notation of time step t in the Q -function and the uncertainty. Our discussion is applicable to the solution for arbitrary time step.

For training a specific task i with a single dataset \mathcal{D}_i in offline RL, we perform Least-Squares Value Iteration (LSVI) to obtain the Q -function. The parameter w_i of Q_i -function can be solved by minimizing the following loss function as

$$\hat{w}_i = \min_{w \in \mathbb{R}^d} \sum_{k=1}^{|\mathcal{D}_i|} (\phi(s_t^k, a_t^k)^\top w - r(s_t^k, a_t^k) - V_{t+1}(s_{t+1}^k))^2 + \lambda \cdot \|w\|_2^2, \quad (\text{A.2})$$

where the experience $(s_t^k, a_t^k, r_t^k, s_{t+1}^k)$ is sampled from \mathcal{D}_i , and V_{t+1} is the estimated value function in the $(t+1)$ -th step. We denote $y_t^k = r(s_t^k, a_t^k) + V_{t+1}(s_{t+1}^k)$ as the target of LSVI, then the parameter \hat{w}_i can be solved in a closed-form as

$$\hat{w}_i = \Lambda_i^{-1} \sum_{k=1}^{|\mathcal{D}_i|} \phi(s_t^k, a_t^k) y_t^k, \quad \Lambda_i = \sum_{k=1}^{|\mathcal{D}_i|} \phi(s_t^k, a_t^k) \phi(s_t^k, a_t^k)^\top + \lambda \cdot \mathbf{I}, \quad (\text{A.3})$$

where the covariance matrix Λ_i accumulates the state-action features from the single-task dataset $|\mathcal{D}_i|$. The L_2 -norm of w_i provides regularizations to ensure that Λ_i is positive definite. Based on the solution of w_i , the action-value function can be estimated by $Q_i(s, a) \approx \hat{w}_i^\top \phi(s, a)$.

In offline RL with linear function assumption, a Lower-Confidence-Bound (LCB) measures the confidence interval of Q -function learned by the given dataset. When we learn value functions from the single \mathcal{D}_i , the LCB-penalty for a specific (s, a) pair is given as

$$\Gamma_i^{\text{LCB}}(s, a) = [\phi(s, a)^\top \Lambda_i^{-1} \phi(s, a)]^{1/2}, \quad (\text{A.4})$$

which forms an uncertainty quantification with the covariance matrix Λ_i^{-1} given the dataset \mathcal{D}_i [28,23,11]. Intuitively, $\Gamma_i^{\text{LCB}}(s, a)$ can be considered as a reciprocal pseudo-count of the state-action pair in the representation space.

Without loss of generality, we assume that w_i has a Gaussian prior as $w_i \sim \mathcal{N}(0, \mathbf{I}/\lambda)$, and denote the noise in regression as $\epsilon_t^k = y_t^k - w_i^\top \phi(s_t^k, a_t^k)$. Recall that the uncertainty quantification used in UTDS for a specific task i is defined as Eq. (5) as $\Gamma_i(s, a) = \sqrt{\mathbb{V}[Q_i^j(s, a)]}$, $j \in [1, \dots, N]$, then the following Lemma established the connection between $\Gamma_i(s, a)$ and the provable efficient LCB-penalty defined in Eq. (A.4).

Lemma 1 (Equivalence between LCB-penalty and Ensemble Uncertainty). *We assume that ϵ follows the standard Gaussian as $\mathcal{N}(0, 1)$ given the state-action pair (s, a) . It then holds for the posterior of w_i given \mathcal{D}_i that*

$$\mathbb{V}_{\hat{w}_i}[Q_i(s, a)] = \mathbb{V}_{\hat{w}_i}(\phi(s, a)^\top \hat{w}_i) = \phi(s, a)^\top \Lambda_i^{-1} \phi(s, a), \quad \forall (s, a) \in S \times \mathcal{A}. \quad (\text{A.5})$$

Proof. This lemma is similar to Lemma 1 in [15], and the proof follows standard Bayesian linear regression. Under the assumptions, we have that $y | (s, a), \hat{w}_i \sim \mathcal{N}(\hat{w}_i^\top \phi(s, a), 1)$. Since the prior distribution $w_i \sim \mathcal{N}(0, \mathbf{I}/\lambda)$. The posterior distribution of \hat{w}_i can be obtained following the Bayesian rule as

$$\log p(\hat{w}_i | \mathcal{D}_i) = \log p(\hat{w}_i) + \log p(\mathcal{D}_i | \hat{w}_i) + \text{Const}. \quad (\text{A.6})$$

Plugging the prior and likelihood distribution into Eq. (A.6) yields

$$\begin{aligned} \log p(\hat{w}_i | \mathcal{D}_i) &= -\|\hat{w}_i\|^2/2 - \sum_{k=1}^{|\mathcal{D}_i|} \|\hat{w}_i^\top \phi(s_t^k, a_t^k) - y_t^k\|^2/2 + \text{Const.} \\ &= -(\hat{w}_i - \mu_i)^\top \Lambda_i^{-1} (\hat{w}_i - \mu_i)/2 + \text{Const.}, \end{aligned} \quad (\text{A.7})$$

where we define $\mu_i = \Lambda_i^{-1} \sum_{k=1}^{|\mathcal{D}_i|} \phi(s_t^k, a_t^k) y_t^k$, and $\Lambda_i = \sum_{k=1}^{|\mathcal{D}_i|} \phi(s_t^k, a_t^k) \phi(s_t^k, a_t^k)^\top + \lambda \cdot \mathbf{I}$. Then we obtain that $\hat{w}_i = w | \mathcal{D}_i \sim \mathcal{N}(\mu_i, \Lambda_i^{-1})$. It then holds for all $(s, a) \in \mathcal{D}_i$ that

$$\mathbb{V}_{\hat{w}_i}(Q_i(s, a)) = \mathbb{V}_{\hat{w}_i}(\phi(s, a)^\top \hat{w}_i) = \phi(s, a)^\top \Lambda_i^{-1} \phi(s, a). \quad (\text{A.8})$$

Then we have $\Gamma_i = \sqrt{\mathbb{V}[Q_i(s, a)]} = [\phi(s, a)^\top \Lambda_i^{-1} \phi(s, a)]^{1/2} = \Gamma_i^{\text{lcb}}$, which concludes our proof. \square

In Lemma 1, we show that the uncertainty estimated by the standard deviation of the Q -posterior is equivalent to the LCB-penalty. LCB can only be applied in linear cases, while the ensemble uncertainty is more general with non-parametric ensemble Q -networks to handle problems with high-dimensional states and actions.

In UTDS, we further include an OOD sampling process in $\hat{\mathcal{T}}^{\text{ood}}$. Then the optimization objective becomes

$$\tilde{w}_i = \min_{w \in \mathbb{R}^d} \sum_{k=1}^{|\mathcal{D}_i|} (\phi(s_t^k, a_t^k)^\top w - r(s_t^k, a_t^k) - V_{t+1}(s_{t+1}^k))^2 + \sum_{s^{\text{ood}}, a^{\text{ood}} \sim \mathcal{D}_i^{\text{ood}}} (\phi(s^{\text{ood}}, a^{\text{ood}})^\top w - y^{\text{ood}})^2, \quad (\text{A.9})$$

where we denote $\mathcal{D}_i^{\text{ood}}$ as an OOD dataset for \mathcal{D}_i for simplicity. In practice, we sample s^{ood} from the original dataset \mathcal{D}_i , and sample a^{ood} by following the learned policy π . Such an OOD sampling process is easy to implement since it only relies on the current policy. In Eq. (A.9), we remove the \mathcal{L}_2 regularization $\lambda \|w\|_2^2$ since it is only suitable for linear cases. For deep offline RL, we find the OOD sampling process provides sufficient regularization for the extrapolation behavior of OOD data. The explicit solution of Eq. (A.9) in LSVI is given as

$$\tilde{w}_i = \tilde{\Lambda}_i^{-1} \left(\sum_{k=1}^{|\mathcal{D}_i|} \phi(s_t^k, a_t^k) y_t^k + \sum_{\mathcal{D}_i^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) y^{\text{ood}} \right), \quad (\text{A.10})$$

where the covariance matrix is

$$\tilde{\Lambda}_i = \sum_{k=1}^{|\mathcal{D}_i|} \phi(s_t^k, a_t^k) \phi(s_t^k, a_t^k)^\top + \sum_{\mathcal{D}_i^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) \phi(s^{\text{ood}}, a^{\text{ood}})^\top := \Lambda_i + \Lambda_i^{\text{ood}}, \quad (\text{A.11})$$

where we denote the second term that accumulates the OOD features as Λ_i^{ood} .

A.2. UTDS for multi-task data sharing

In MTDS, considering we share a dataset \mathcal{D}_j to task i as $\mathcal{D}_{j \rightarrow i}$, then we have $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ in training. We denote the parameter of Q -function learned in $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ as w_{ij} , then we have

$$\begin{aligned} \tilde{w}_{ij} &= \min_{w \in \mathbb{R}^d} \sum_{k_1=1}^{|\mathcal{D}_i|} (\phi(s_t^{k_1}, a_t^{k_1})^\top w - r(s_t^{k_1}, a_t^{k_1}) - V_{t+1}(s_{t+1}^{k_1}))^2 \\ &\quad + \sum_{k_2=1}^{|\mathcal{D}_{j \rightarrow i}|} (\phi(s_t^{k_2}, a_t^{k_2})^\top w - r(s_t^{k_2}, a_t^{k_2}) - V_{t+1}(s_{t+1}^{k_2}))^2 + \sum_{\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_{j \rightarrow i}^{\text{ood}}} (\phi(s^{\text{ood}}, a^{\text{ood}})^\top w - y^{\text{ood}})^2, \end{aligned} \quad (\text{A.12})$$

where the transitions in the first summation are sampled from the original dataset of task i , the transitions in the second summation are sampled from the relabeled dataset $\mathcal{D}_{j \rightarrow i}$ from task j , and the last term is sampled from the OOD data $\mathcal{D}_i^{\text{ood}}$ and $\mathcal{D}_{j \rightarrow i}^{\text{ood}}$. In addition, we remark that we do not need relabeled rewards for $(s^{\text{ood}}, a^{\text{ood}}) \sim \mathcal{D}_j^{\text{ood}}$ since the target y^{ood} defined in Eq. (C.1) is a pseudo-target that enforces pessimism based on the current value function and the uncertainty quantification. The calculation of y^{ood} does not rely on the reward function.

Following LSVI, the solution of parameter \tilde{w}_{ij} in Eq. (A.12) is

$$\tilde{w}_{ij} = \tilde{\Lambda}_{ij}^{-1} \left(\sum_{k_1=1}^{|\mathcal{D}_i|} \phi(s_t^{k_1}, a_t^{k_1}) y_t^{k_1} + \sum_{k_2=1}^{|\mathcal{D}_{j \rightarrow i}|} \phi(s_t^{k_2}, a_t^{k_2}) y_t^{k_2} + \sum_{\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_{j \rightarrow i}^{\text{ood}}} (\phi(s^{\text{ood}}, a^{\text{ood}}) y^{\text{ood}}) \right), \quad (\text{A.13})$$

where we accumulate all OOD data from $\mathcal{D}_i^{\text{ood}}$ and $\mathcal{D}_{j \rightarrow i}^{\text{ood}}$ in the last term. The covariance matrix $\tilde{\Lambda}_{ij}$ for the mixed data $\hat{\mathcal{D}}_i$ becomes

$$\begin{aligned}
\tilde{\Lambda}_{ij} &= \sum_{k_1=1}^{|\mathcal{D}_i|} \phi(s_i^{k_1}, a_i^{k_1}) \phi(s_i^{k_1}, a_i^{k_1})^\top + \sum_{\mathcal{D}_j^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) \phi(s^{\text{ood}}, a^{\text{ood}})^\top \\
&\quad + \sum_{k_2=1}^{|\mathcal{D}_{j \rightarrow i}|} \phi(s_i^{k_2}, a_i^{k_2}) \phi(s_i^{k_2}, a_i^{k_2})^\top + \sum_{\mathcal{D}_j^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) \phi(s^{\text{ood}}, a^{\text{ood}})^\top \\
&= \Lambda_i + \Lambda_i^{\text{ood}} + \Lambda_j + \Lambda_j^{\text{ood}} = \tilde{\Lambda}_i + \tilde{\Lambda}_j,
\end{aligned} \tag{A.14}$$

which contains the feature covariance matrix from both task i and task j . The definition of Λ_i and Λ_i^{ood} are given in Eq. (A.11).

Based on the covariance matrix given in Eq. (A.11) and Eq. (A.14), the following theorem establishes the relationship between the uncertainty estimation in the single dataset \mathcal{D}_i and the shared dataset $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$.

Theorem (Theorem 1 restate). For a given state-action pair (s, a) , we denote the uncertainty for the single-task dataset \mathcal{D}_i and shared dataset $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ as $\Gamma_i(s, a; \mathcal{D}_i)$ and $\Gamma_i(s, a; \hat{\mathcal{D}}_i)$, respectively. Then the following inequality holds as

$$\Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i) \geq \Gamma_i^{\text{lc}}(s, a; \hat{\mathcal{D}}_i), \text{ and } \Gamma_i(s, a; \mathcal{D}_i) \geq \Gamma_i(s, a; \hat{\mathcal{D}}_i), \tag{A.15}$$

where signifies that the shared data reduces the uncertainty.

Proof. In the following, we prove that $\Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i) \geq \Gamma_i^{\text{lc}}(s, a; \hat{\mathcal{D}}_i)$. The original relationship can be obtained by following Lemma 1 that $\Gamma_i(s, a; \mathcal{D}_i) = \Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i)$ and $\Gamma_i(s, a; \hat{\mathcal{D}}_i) = \Gamma_i^{\text{lc}}(s, a; \hat{\mathcal{D}}_i)$.

For UTDS in the shared dataset $\hat{\mathcal{D}}_i$, according to the solution of w_{ij} and covariance matrix given in Eq. (A.13) and Eq. (A.14), we have

$$\tilde{\Lambda}_{ij} = \tilde{\Lambda}_i + \tilde{\Lambda}_j, \tag{A.16}$$

where $\tilde{\Lambda}_j$ denotes the feature covariance matrix from shared task j with OOD sampling, as $\tilde{\Lambda}_j = \Lambda_j + \Lambda_j^{\text{ood}}$. Since both $\tilde{\Lambda}_{ij}$ and $\tilde{\Lambda}_i$ are positive semi-definite, by following the generalized Rayleigh quotient, we have

$$\frac{\phi^\top \tilde{\Lambda}_i^{-1} \phi}{\phi^\top (\tilde{\Lambda}_i + \tilde{\Lambda}_j)^{-1} \phi} \geq \lambda_{\min}((\tilde{\Lambda}_i + \tilde{\Lambda}_j) \tilde{\Lambda}_i^{-1}) = \lambda_{\min}(\mathbf{I} + \tilde{\Lambda}_j \tilde{\Lambda}_i^{-1}) = 1 + \lambda_{\min}(\tilde{\Lambda}_j \tilde{\Lambda}_i^{-1}), \tag{A.17}$$

where $\lambda_{\min}(\cdot)$ is the minimum eigenvalue of a matrix. Since $\tilde{\Lambda}_j$ and $\tilde{\Lambda}_i^{-1}$ are positive semi-definite matrices, the eigenvalues of $\tilde{\Lambda}_j \tilde{\Lambda}_i^{-1}$ are non-negative and $\lambda_{\min}(\tilde{\Lambda}_j \tilde{\Lambda}_i^{-1}) \geq 0$. Then we have

$$\phi(s, a)^\top \tilde{\Lambda}_i^{-1} \phi(s, a) \geq \phi(s, a)^\top (\tilde{\Lambda}_i + \tilde{\Lambda}_j)^{-1} \phi(s, a) = \phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \phi(s, a). \tag{A.18}$$

Then according to the definition of LCB-penalty as $\Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i) = [\phi(s, a)^\top \tilde{\Lambda}_i^{-1} \phi(s, a)]^{1/2}$, we have

$$\Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i) \geq \Gamma_i^{\text{lc}}(s, a; \hat{\mathcal{D}}_i). \tag{A.19}$$

Further, if we additionally share dataset $\mathcal{D}_{k \rightarrow i}$ from task k to task i , then the covariance matrix becomes $\tilde{\Lambda}_{ijk} = \tilde{\Lambda}_i + \tilde{\Lambda}_j + \tilde{\Lambda}_k$, then we have

$$\tilde{\Lambda}_{ijk} \geq \tilde{\Lambda}_{ij} \geq \tilde{\Lambda}_i, \tag{A.20}$$

and the corresponding LCB-penalties have the following relationship, as

$$\phi(s, a)^\top \tilde{\Lambda}_i^{-1} \phi(s, a) \geq \phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \phi(s, a) \geq \phi(s, a)^\top \tilde{\Lambda}_{ijk}^{-1} \phi(s, a). \tag{A.21}$$

Then we have

$$\Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i) \geq \Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}) \geq \Gamma_i^{\text{lc}}(s, a; \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i} \cup \mathcal{D}_{k \rightarrow i}), \tag{A.22}$$

and thus

$$\Gamma_i(s, a; \mathcal{D}_i) \geq \Gamma_i(s, a; \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}) \geq \Gamma_i(s, a; \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i} \cup \mathcal{D}_{k \rightarrow i}) \tag{A.23}$$

by Lemma 1, which concludes our proof. \square

Theorem 1 shows that the uncertainty for a specific (s, a) pair will decrease with more shared data, which has also been illustrated in Fig. 2. For example, if a (s, a) pair is scarcely occurred in the original dataset \mathcal{D}_i , the uncertainty-based penalty will be high and the agent will hardly choose this action since the corresponding $Q_i(s, a)$ function is pessimistic. However, such pessimism comes from the insufficient knowledge of the environment (i.e., epistemic uncertainty), which does not represent a is actually a bad choice

in state s . In MTDS, the uncertainty $\Gamma_i(s, a)$ will gradually decrease with more (s, a) pairs shared in $\widehat{\mathcal{D}}_i$, and the value function may become less pessimistic. Such a property is important since the agent will extend its knowledge in the state-action space with data sharing. In addition, since the neural network has generalization ability around near the input, $\Gamma_i(s, a)$ also decreases when we share similar state-action pairs of (s, a) .

In extreme cases, if the shared data contains sufficient transitions in the whole state-action space, the covariance matrix $\widetilde{\Lambda}_i$ is full rank and $\lambda_{\min}(\Lambda_i) \rightarrow \infty$. Then the uncertainty quantification $\phi(s, a)^\top \widetilde{\Lambda}_i^{-1} \phi(s, a) \rightarrow 0$. This case is similar to online RL with sufficient exploratory data, and the agent will choose actions based on the estimated value function without uncertainty penalty.

A.3. ξ -Uncertainty quantifier

For offline RL, we learn a pessimistic value function $\widehat{Q}_i(s_t, a_t)$ by penalizing the Q_i -function with the uncertainty quantification $\Gamma_i^{\text{lcb}}(s_t, a_t)$, as

$$\widehat{Q}_i(s_t, a_t) = Q_i(s_t, a_t) - \Gamma_i^{\text{lcb}}(s_t, a_t) = \widetilde{w}^\top \phi(s_t, a_t) - \Gamma_i^{\text{lcb}}(s_t, a_t), \quad (\text{A.24})$$

where the weight \widetilde{w} can be \widetilde{w}_i or \widetilde{w}_{ij} based on different datasets. Under the linear MDP setting, such pessimistic value iteration is known to be information-theoretically optimal [11]. In UTDS, we implement this pessimistic value function via uncertainty penalty in $\widehat{\mathcal{T}}_{\text{UTDS}}$ and $\widehat{\mathcal{T}}_{\text{ood}}$.

From the theoretical perspective, an appropriate uncertainty quantification is essential to provable efficiency in offline RL [27, 12, 15]. Definition 1 [11] defines a general ξ -uncertainty quantifier pessimistic value iteration as a penalty in Eq. (A.24) and achieves provable efficient pessimism in offline RL. In linear MDPs, the LCB-penalty defined in Eq. (A.4) is known to be a ξ -uncertainty quantifier for appropriately selected $\{\beta_t\}_{t \in [T]}$, as $\beta_t[\phi(s_t, a_t)^\top \Lambda_t^{-1} \phi(s_t, a_t)]^{1/2}$. In the following theorem, we show that the proposed UTDS with the shared dataset $\widehat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ also forms a valid ξ -uncertainty quantifier with the covariance matrix $\widetilde{\Lambda}_{ij}$ given in Eq. (A.14).

Theorem. (Theorem 2 restate) Let $\widetilde{\Lambda}_{ij} \geq \lambda \cdot \mathbf{I}$, if we set the OOD target as $y^{\text{ood}} = \mathcal{T}V_{t+1}(s^{\text{ood}}, a^{\text{ood}})$ for shared dataset $\widehat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$, then it holds for $\beta_t = O(T \cdot \sqrt{d} \cdot \log(T/\xi))$ that

$$\Gamma_i^{\text{lcb}}(s_t, a_t; \widehat{\mathcal{D}}_i) = \beta_t [\phi(s_t, a_t)^\top \widetilde{\Lambda}_{ij}^{-1} \phi(s_t, a_t)]^{1/2}$$

forms a valid ξ -uncertainty quantifier, where $\widetilde{\Lambda}_{ij}$ is the covariance matrix given in Eq. (A.14).

Proof. The proof follows that of the analysis of PBRL [15] in linear MDPs [11]. We define the empirical Bellman operator of UTDS learned in $\widehat{\mathcal{D}}_i$ as $\widetilde{\mathcal{T}}$, then

$$\widetilde{\mathcal{T}}V_{t+1}(s_t, a_t) = \phi(s_t, a_t)^\top \widetilde{w}_{ij},$$

where the parameter \widetilde{w}_{ij} follows the solution in Eq. (A.13). Following the ξ -uncertainty quantifier defined in Definition 1, we upper bound the difference between the empirical Bellman operator of UTDS and the true Bellman operator as

$$\mathcal{T}V_{t+1}(s, a) - \widetilde{\mathcal{T}}V_{t+1}(s, a) = \phi(s, a)^\top (w_t - \widetilde{w}_{ij}).$$

Here we define w_t as follows

$$w_t = \theta + \int_{\mathcal{S}} V_{t+1}(s_{t+1}) \psi(s_{t+1}) ds_{t+1}, \quad (\text{A.25})$$

where θ and ψ are defined in Eq. (A.1). It then holds that

$$\begin{aligned} \mathcal{T}V_{t+1}(s, a) - \widetilde{\mathcal{T}}V_{t+1}(s, a) &= \phi(s, a)^\top (w_t - \widetilde{w}_{ij}) \\ &= \phi(s, a)^\top w_t - \phi(s, a)^\top \widetilde{\Lambda}_{ij}^{-1} \sum_{k_1=1}^{|\mathcal{D}_i|} \phi(s_t^{k_1}, a_t^{k_1}) (r(s_t^{k_1}, a_t^{k_1}) + V_{t+1}^i(s_{t+1}^{k_1})) \\ &\quad - \phi(s, a)^\top \widetilde{\Lambda}_{ij}^{-1} \sum_{k_2=1}^{|\mathcal{D}_{j \rightarrow i}|} \phi(s_t^{k_2}, a_t^{k_2}) (r(s_t^{k_2}, a_t^{k_2}) + V_{t+1}^i(s_{t+1}^{k_2})) \\ &\quad - \phi(s, a)^\top \widetilde{\Lambda}_{ij}^{-1} \sum_{\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_{j \rightarrow i}^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) y^{\text{ood}}, \end{aligned} \quad (\text{A.26})$$

where we use the solution of \widetilde{w}_{ij} in Eq. (A.13). By the definitions of $\widetilde{\Lambda}_i$ and w_t in Eq. (A.14) and Eq. (A.25), respectively, we have

$$\begin{aligned} \phi(s, a)^\top w_t &= \phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \tilde{\Lambda}_{ij} w_t = \phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \left(\sum_{k_1=1}^{|\mathcal{D}_i|} \phi(s_t^{k_1}, a_t^{k_1}) \mathcal{T} V_{t+1}(s_t^{k_1}, a_t^{k_1}) \right. \\ &\quad \left. + \sum_{k_2=1}^{|\mathcal{D}_{j \rightarrow i}|} \phi(s_t^{k_2}, a_t^{k_2}) \mathcal{T} V_{t+1}(s_t^{k_2}, a_t^{k_2}) + \sum_{\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_j^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) \mathcal{T} V_{t+1}(\phi(s^{\text{ood}}, a^{\text{ood}})) \right). \end{aligned} \quad (\text{A.27})$$

Plugging Eq. (A.27) into Eq. (A.26) yields

$$\mathcal{T} V_{t+1}(s, a) - \tilde{\mathcal{T}} V_{t+1}(s, a) = (\text{i}) + (\text{ii}) + (\text{iii}), \quad (\text{A.28})$$

where we define

$$\begin{aligned} (\text{i}) &= \phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \sum_{k_1=1}^{|\mathcal{D}_i|} \phi(s_t^{k_1}, a_t^{k_1}) (\mathcal{T} V_{t+1}(s_t^{k_1}, a_t^{k_1}) - r(s_t^{k_1}, a_t^{k_1}) - V_{t+1}^i(s_t^{k_1})), \\ (\text{ii}) &= \phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \sum_{k_2=1}^{|\mathcal{D}_{j \rightarrow i}|} \phi(s_t^{k_2}, a_t^{k_2}) (\mathcal{T} V_{t+1}(s_t^{k_2}, a_t^{k_2}) - r(s_t^{k_2}, a_t^{k_2}) - V_{t+1}^i(s_t^{k_2})), \\ (\text{iii}) &= \phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \sum_{\mathcal{D}_i^{\text{ood}} \cup \mathcal{D}_j^{\text{ood}}} \phi(s^{\text{ood}}, a^{\text{ood}}) (\mathcal{T} V_{t+1}(s^{\text{ood}}, a^{\text{ood}}) - y^{\text{ood}}). \end{aligned}$$

Following the standard analysis based on the concentration of self-normalized process [28,25,26,23,11] and the fact that $\tilde{\Lambda}_{ij} \succeq \lambda \cdot \mathbf{I}$, it holds that

$$|(\text{i})| \leq \beta_i \cdot [\phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \phi(s, a)]^{1/2}, \quad |(\text{ii})| \leq \beta_j \cdot [\phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \phi(s, a)]^{1/2}, \quad (\text{A.29})$$

with probability at least $1 - \xi$, where $\beta_{i(j)} = O(T \cdot \sqrt{d} \cdot \log(T/\xi))$. By following Eq. (A.21), we have

$$\begin{aligned} |(\text{i})| + |(\text{ii})| &\leq 2 \max\{\beta_i, \beta_j\} \cdot [\phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \phi(s, a)]^{1/2} \\ &= \beta_t \cdot [\phi(s, a)^\top \tilde{\Lambda}_{ij}^{-1} \phi(s, a)]^{1/2}, \end{aligned} \quad (\text{A.30})$$

where we denote $\beta_t = 2 \max\{\beta_i, \beta_j\}$. Meanwhile, by setting $y^{\text{ood}} = \mathcal{T} V_{t+1}(s^{\text{ood}}, a^{\text{ood}})$, it holds that $(\text{iii}) = 0$. Thus, we obtain from Eq. (A.28) that

$$|\mathcal{T} V_{t+1}(s, a) - \tilde{\mathcal{T}} V_{t+1}(s, a)| \leq \beta_t \cdot [\phi(s_t, a_t)^\top \Lambda_{ij}^{-1} \phi(s_t, a_t)]^{1/2} \quad (\text{A.31})$$

with probability at least $1 - \xi$, which concludes our proof. \square

Theorem 2 shows that the disagreement among ensemble Q -networks is a valid ξ -uncertainty quantifier with data sharing, while it needs the covariance matrix $\tilde{\Lambda}_{ij}$ is lower bounded. In practice, such an assumption can be achieved through (i) sharing more data to $\hat{\mathcal{D}}_i$ since we have $\tilde{\Lambda}_{ij} \succeq \tilde{\Lambda}_i$ according to Eq. (A.20), and (ii) randomly generating OOD actions to make the embeddings of the OOD sample isotropic, which ensures the eigenvalues of the covariate matrix Λ^{ood} are lower bounded.

In addition, since the transition function is unknown for OOD samples, the OOD target $\mathcal{T} V_{t+1}$ is impossible to obtain in practice as it requires knowing the transition at the OOD datapoint. In practice, if TD error is sufficiently minimized, then $Q_{t+1}(s, a)$ should well estimate the target $\mathcal{T} V_{t+1}$. Thus, in practice, the targets of OOD data are set to be $y_i^{\text{ood}} = Q_i(s^{\text{ood}}, a^{\text{ood}}) - \Gamma_i(s^{\text{ood}}, a^{\text{ood}})$ by assuming the TD-error is sufficiently minimized.

A.4. Suboptimality gap

Theorem 2 allows us to further characterize the optimality gap based on the pessimistic value iteration [11]. In particular, the following corollary holds,

Corollary (Corollary 1 restate). Under the same conditions as Theorem 2, for the uncertainty quantification $\Gamma_i(s, a; \mathcal{D}_i)$ and $\Gamma_i(s, a; \hat{\mathcal{D}}_i)$ defined in \mathcal{D}_i and $\hat{\mathcal{D}}_i = \mathcal{D}_i \cup \mathcal{D}_{j \rightarrow i}$ respectively, we have

$$\text{SubOpt}(\pi_i^*, \tilde{\pi}_i) \leq \sum_{t=1}^T \mathbb{E}_{\pi_i^*} [\Gamma_i^{\text{lb}}(s_t, a_t; \hat{\mathcal{D}}_i)] \leq \sum_{t=1}^T \mathbb{E}_{\pi_i^*} [\Gamma_i^{\text{lb}}(s_t, a_t; \mathcal{D}_i)], \quad (\text{A.32})$$

where $\tilde{\pi}_i$ and π_i^* are the learned policy and the optimal policy in $\hat{\mathcal{D}}_i$, respectively.

Proof. The first inequality holds since $\Gamma_i^{\text{lb}}(s_t, a_t; \hat{\mathcal{D}}_i)$ forms a valid ξ -uncertainty quantifier by Theorem 2. We refer to [11] for detailed proof.

The second inequality is induced by $\Gamma_i(s_t, a_t; \hat{\mathcal{D}}_i) < \Gamma_i(s_t, a_t; \mathcal{D}_i)$ in Theorem 1 with the fixed optimal policy π_i^* . The optimality gap is information-theoretically optimal under the linear MDP setup with finite horizon. \square

Consider an extreme case in tabular MDPs, where the state action feature $\phi(s, a)$ is a one-hot vector. Then we have $\Gamma^{\text{lb}}(s, a) = 1/\sqrt{N(s, a) + \lambda}$, where $N(s, a)$ is the pseudo-count of the (s, a) pair. In this case, even if we share data that are totally different from the main task, the optimality gap $\mathbb{E}_{\pi^*}[\Gamma^{\text{lb}}(s, a)]$ does not increase since the pseudo-count $N(s, a)$ cannot decrease with data sharing. As a result, even if data sharing does not benefit the data coverage of the optimal policy, it does not degrade performance. Such a key factor makes UTDS inherently different from the policy constraint methods that need appropriate data selection to select similar samples compared to the main task.

A.5. Supplementary lemma

According to Lemma 1 in our manuscript, the uncertainty term is equivalent to $\Gamma^{\text{lb}}(s, a)$ in linear MDPs. The value of $\Gamma^{\text{lb}}(s, a) = [\phi(s, a)^\top \Lambda^{-1} \phi(s, a)]^{\frac{1}{2}}$ only relies on the state-action embedding in the offline datasets. Then we have the following lemma.

Lemma 2 (contraction mapping). *For state-action pairs from the offline dataset, the UTDS operator \hat{T}^{UTDS} is a γ -contraction operator in the \mathcal{L}_∞ -norm.*

Proof. Let Q_1 and Q_2 be two arbitrary Q functions. Since $a \in \text{Support}(\mu(\cdot|s))$, then we have

$$\begin{aligned} & \|\hat{T}^{\text{UTDS}} Q_1 - \hat{T}^{\text{UTDS}} Q_2\|_\infty \\ &= \max_{s,a} \left| (r(s, a) + \gamma \mathbb{E}_{s' \sim P, a' \sim \pi} [Q_1(s', a') - \Gamma^{\text{lb}}(s', a')]) \right. \\ & \quad \left. - (r(s, a) + \gamma \mathbb{E}_{s' \sim P, a' \sim \pi} [Q_2(s', a') - \Gamma^{\text{lb}}(s', a')]) \right| \\ &= \gamma \max_{s,a} \left| \mathbb{E}_{s' \sim P, a' \sim \pi} [Q_1(s', a') - Q_2(s', a')] \right| \leq \gamma \max_{s,a} \|Q_1 - Q_2\|_\infty = \gamma \|Q_1 - Q_2\|_\infty. \end{aligned}$$

Hence, for some arbitrarily initialized Q -function, it is guaranteed to converge to a unique fixed point by repeatedly applying \hat{T}^{UTDS} . \square

Appendix B. Implementation details

In this section, we provide the hyper-parameters and experimental settings. We refer to <https://github.com/Baichenjia/UTDS> for the open-source implementation and the released multi-task datasets.

B.1. UTDS

The implementation of ensemble- Q networks in UTDS is based on PBRL and EDAC, while we use much fewer ensemble networks (i.e., with 5 networks) compared to PBRL (with 10 networks) and EDAC (with 10-50 networks). We use the *same* hyper-parameter settings for data-sharing tasks in all domains. The hyper-parameters are listed in Table B.1.

For the critic training, we use different factors (i.e., β_1 and β_2) for the offline data and the OOD data. The update of the Q -function in offline data relies on an ordinary Bellman target with a constant factor for uncertainty penalty. In contrast, since the OOD data does not have the reward and transition function, the pseudo-target is used for uncertainty penalty based on the current value estimation. According to our analysis, β_2 should exponentially decay to ensure $Q(s^{\text{ood}}, a^{\text{ood}})$ converges to a fixed point.

Table B.1
Hyper-parameters of UTDS.

Hyper-parameters	Value
The number of bootstrapped networks N	5
Policy network	FC(256,256,256,action_dim) with ReLU activations
Q -network	FC(256,256,256,1) with ReLU activations
Target network smoothing coefficient	$5e-3$
Discount factor γ	0.99
Learning rate (policy)	$1e-4$
Learning rate (Q -networks)	$1e-4$
Optimizer	Adam
Batch size	1024
Number of OOD samples for each state	3
Factor β_1 for uncertainty penalty in offline data	0.001
Factor β_2 for uncertainty penalty in OOD data	$3.0 \rightarrow 0.1$ (first half), 0.1 (other)
Factor α for exponentially decay of β_2 with training steps	0.99995
Training steps	1M

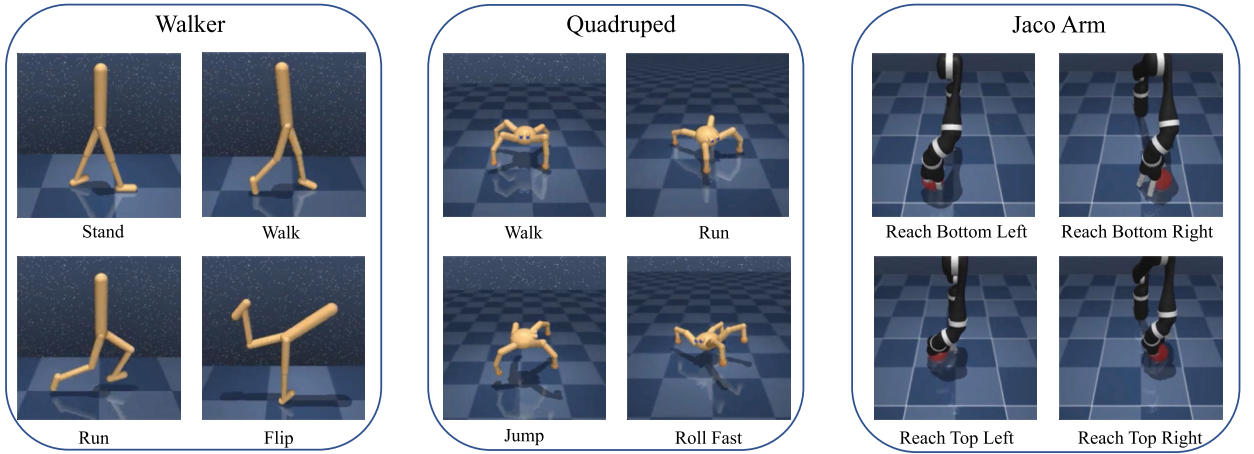


Fig. B.9. The illustration of 12 continuous control tasks from 3 domains in our experiments.

Table B.2
Comparison of computational costs.

	Runtime (hours / 1M steps)	GPU memory	Number of parameters
CDS	15.75	1.31G	0.37M
CDS-Zero	15.68	1.31G	0.37M
UTDS	6.56	1.34G	0.81M

For the actor training, we follow SAC-N [22] by using the minimum of ensemble Q -function $\min_{n=1,\dots,N} Q_n$ as the target, which is approximately equivalent to using $Q_n - \beta_0 \cdot \text{Std}(Q_n)$ with a fixed β_0 .

B.2. Baselines

Direct sharing For the shared dataset, we relabel the reward function based on the Mujoco ‘Physics’ class of DeepMind Control Suite. Then we combine the relabeled data with the original data to construct a mixed dataset for training. For direct sharing, we use CQL [42] to train a policy based on the mixed dataset.

CDS and CDS-Zero [13,30] These two methods are implemented based on CQL. For each training step, we sample experiences ten times the size of the batch size, and calculate the conservative value function for each state-action pair. Then we choose examples with conservative Q -values within the top 10% of the sampled batch in training. For CDS-Zero, we do not perform reward relabeling and set the shared reward to be zero. We remark that we additionally use layer-normalization for the actor-critic network in CDS-based methods since we find it helps CDS perform more stable. In contrast, we do not use layer-normalization for the proposed UTDS algorithm.

B.3. Experimental settings

Environments and training We use 12 tasks from *Walker*, *Quadruped*, and *Jaco Arm* domains in experiments, as shown in Fig. B.9. For *Walker* domain, the observation space has 24 dimensions, and the action space has 6 dimensions. For *Quadruped* domain, the observation space has much higher dimensions (i.e., 78), which contains egocentric state, torso velocity, torso upright, IMU, and force torque. The action space of *Quadruped* has 12 dimensions. For *Jaco Arm*, the observation space has 42 dimensions, which contains information of the arm, hand, and target. The action space has 9 dimensions.

According to the default settings of DeepMind Control Suite, the episode lengths for *Walker* and *Quadruped* domains are set to 1000, and the episode length for *Joco* domain is set to 250. As a result, the maximum episodic reward for *Walker* and *Quadruped* domains are 1000, and for *Joco* domain is 250. In experiments, we run each algorithm for 1M training steps with 3 random seeds. We evaluate the algorithm for 10 episodes every 1K training steps.

Computation cost comparison We compare the computational cost of UTDS, CDS, and CDS-Zero based on the average cost among *Walker* domain. We report the number of parameters, GPU memory, and runtime per sharing task in training. The experiment is conducted on a single RTX-2080Ti GPU. We give the result in Table B.2. The result shows that (i) UTDS requires extra parameters to handle the ensemble of Q -networks, while (ii) CDS and CDS-Zero need more training time to perform conservative Q -value calculation and data-selection per epoch. CDS-Zero is slightly faster since it does not perform reward relabeling. (iii) The GPU

memories for the two methods are similar since CDS and CDS-Zero need much larger training batches to first perform data selection and then perform RL training.

Appendix C. Additional experimental results

In this section, we report the complete results of data sharing in *Walker*, *Quadruped*, and *Jaco Arm* domains. For each domain, we report

- the single-task training score without data sharing.
- the direct data-sharing score with relabeled rewards, and the policy is trained by CQL.
- the result comparison between the proposed UTDS and CDS [13] that performs conservative data selection.
- the result comparison between the proposed UTDS and CDS-Zero [30] that performs conservative data selection with zero relabeled reward.

C.1. Single-task results

UTDS provides a unified view for single-task and multi-task training, as shown in Fig. 1. Thus, UTDS can be directly used for training in single-task datasets. For CDS and CDS-Zero, performing single-task training degenerates into the CQL algorithm [42] that does not perform data selection. Fig. C.10, Fig. C.11, and Fig. C.12 show the single-task training result for the *Walker*, *Quadruped*, and *Jaco Arm* domains, respectively.

C.2. Direct data-sharing results

For offline RL, direct sharing with policy constraint methods can exacerbate the distribution shift when the sharing tasks are very different from the main task. Fig. C.13, Fig. C.14, and Fig. C.15 show the result of the direct sharing for the *Walker*, *Quadruped*, and *Jaco Arm* domains, respectively. In each figure, we also show the single-task training results for the main task in the leftmost bar (i.e., the shadow bar).

1. In *Walker* domain, direct data sharing slightly improves the performance, the reason is the four tasks in *Walker* is closely related. For example, the agent should first *Stand* and then *Walk*, *Run* and *Flip*, which makes the dataset of *Walk*, *Run* and *Flip* also contain the experiences of *Stand*. As a result, sharing data from other three tasks often improves the performance of *Stand*.

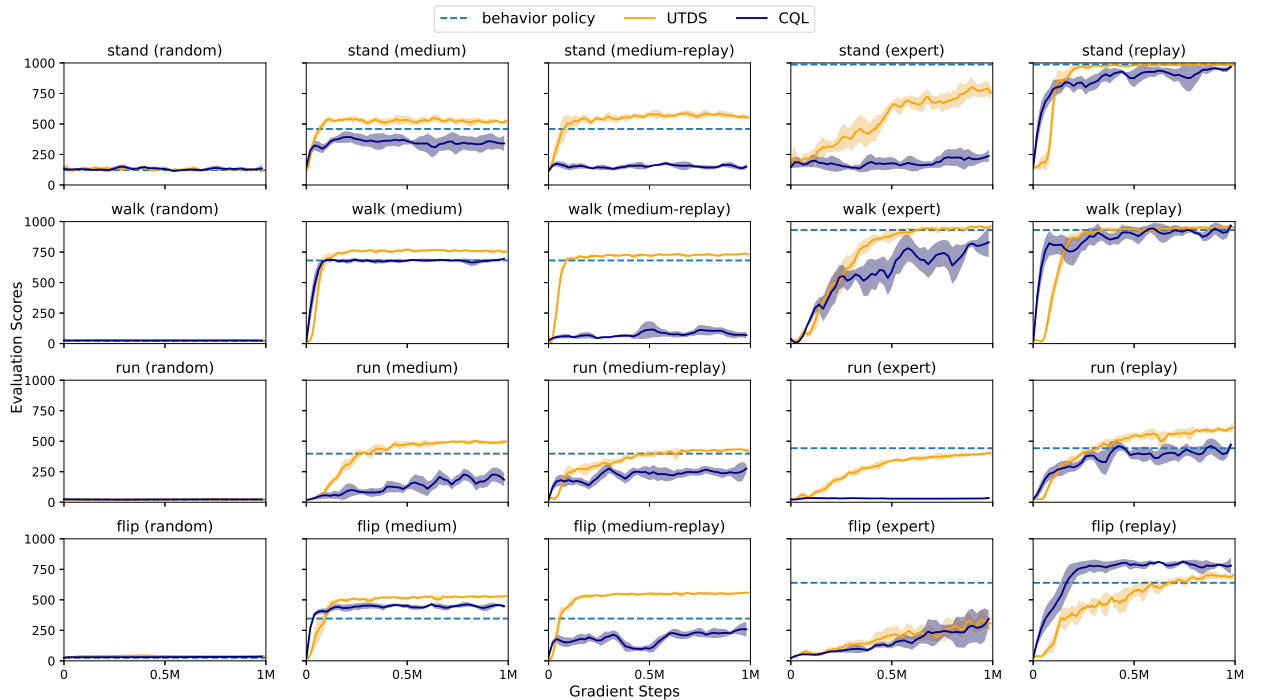


Fig. C.10. Result comparison of single-task training in the *Walker* domain. UTDS outperforms CQL in most tasks, especially for *medium*, *medium-replay*, and *expert* datasets. Both methods perform poorly in random datasets.

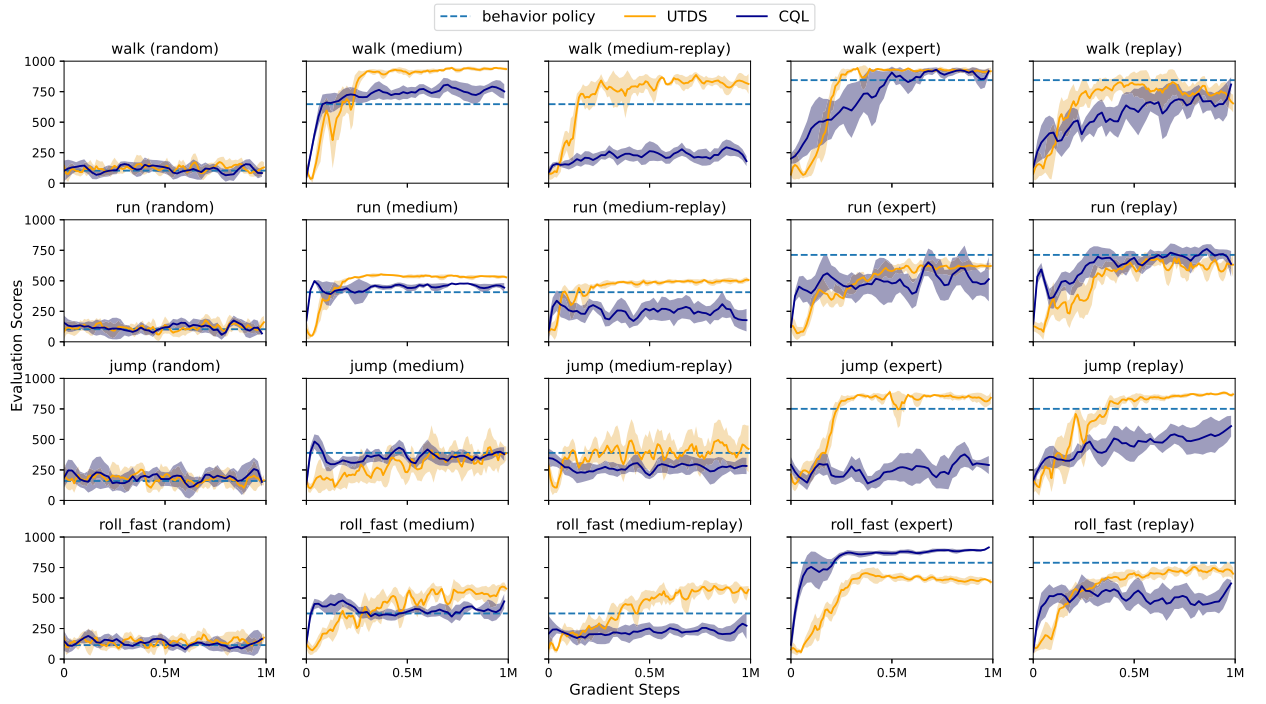


Fig. C.11. Result comparison of single task training in *Quadraped* domain. UTDS outperforms CQL in most tasks, except for the *Roll-Fast* expert dataset. Both methods perform poorly in random datasets.

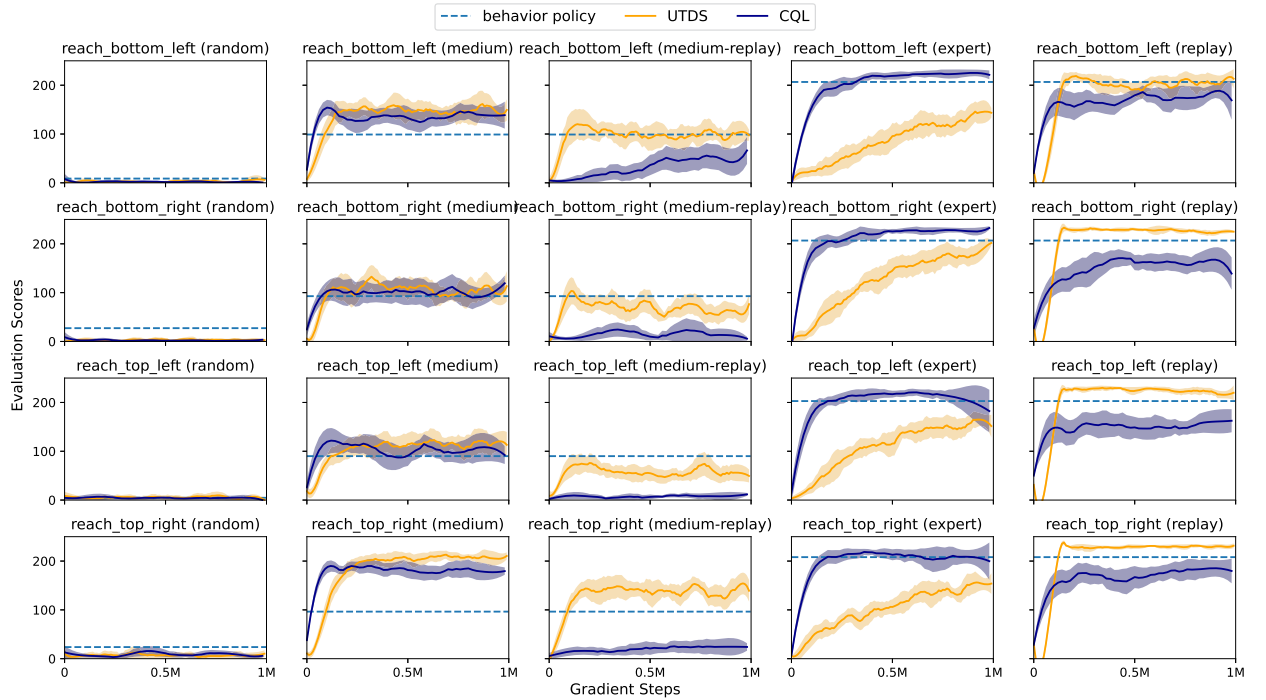


Fig. C.12. Result comparison of single-task training in *Jaco Arm* domain. We find CQL outperforms UTDS in expert dataset. The policy constraints direct imitate the expert policies, which make CQL converge faster with optimal trajectories.

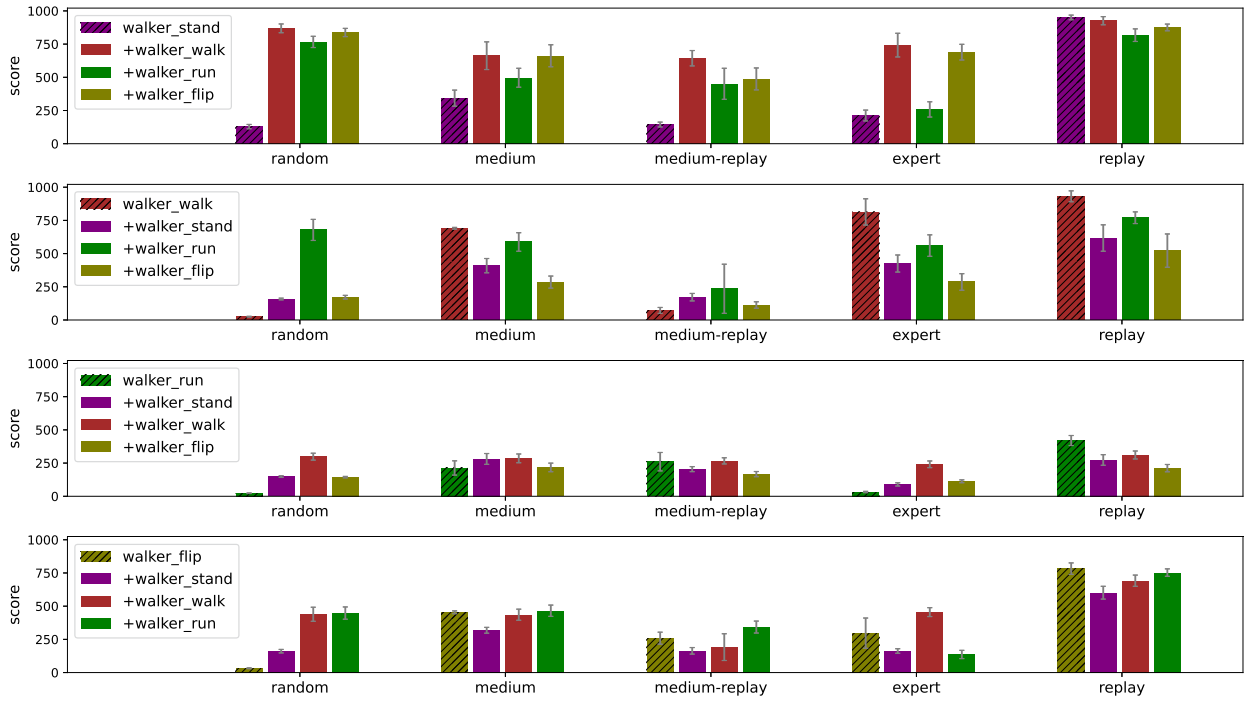


Fig. C.13. Result comparison of single-task training and direct sharing with CQL in Walker domain.

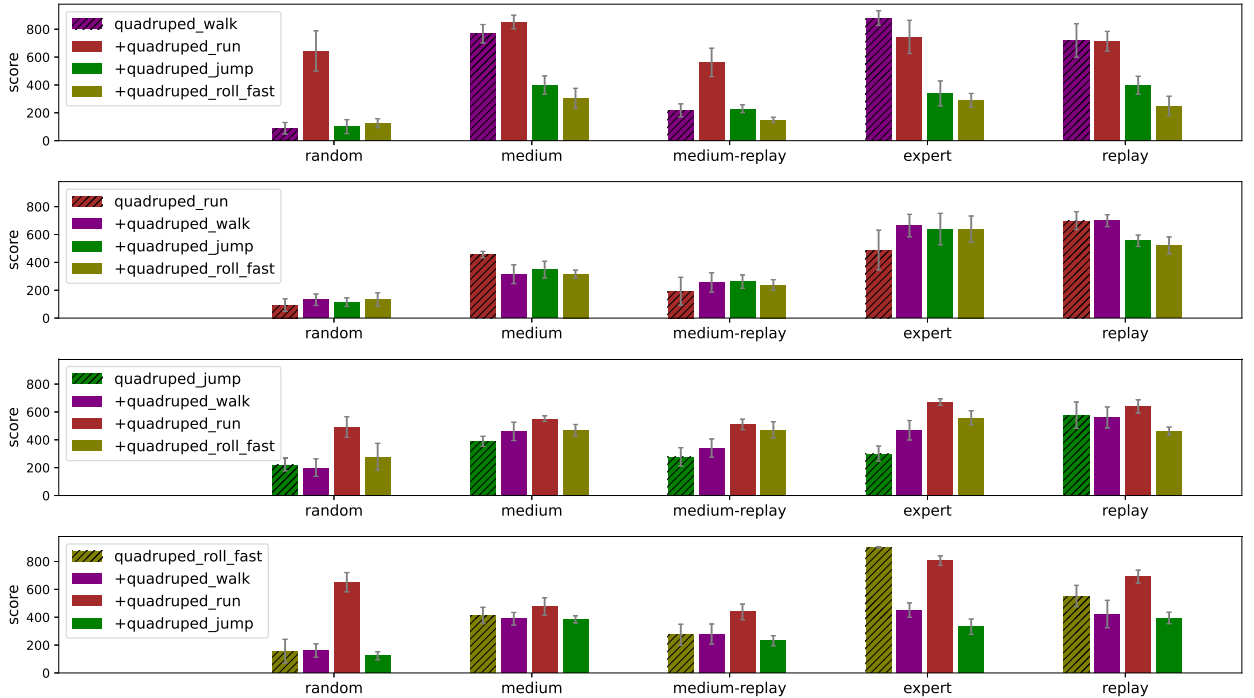


Fig. C.14. Result comparison of single-task training and direct sharing with CQL in Quadruped domain.

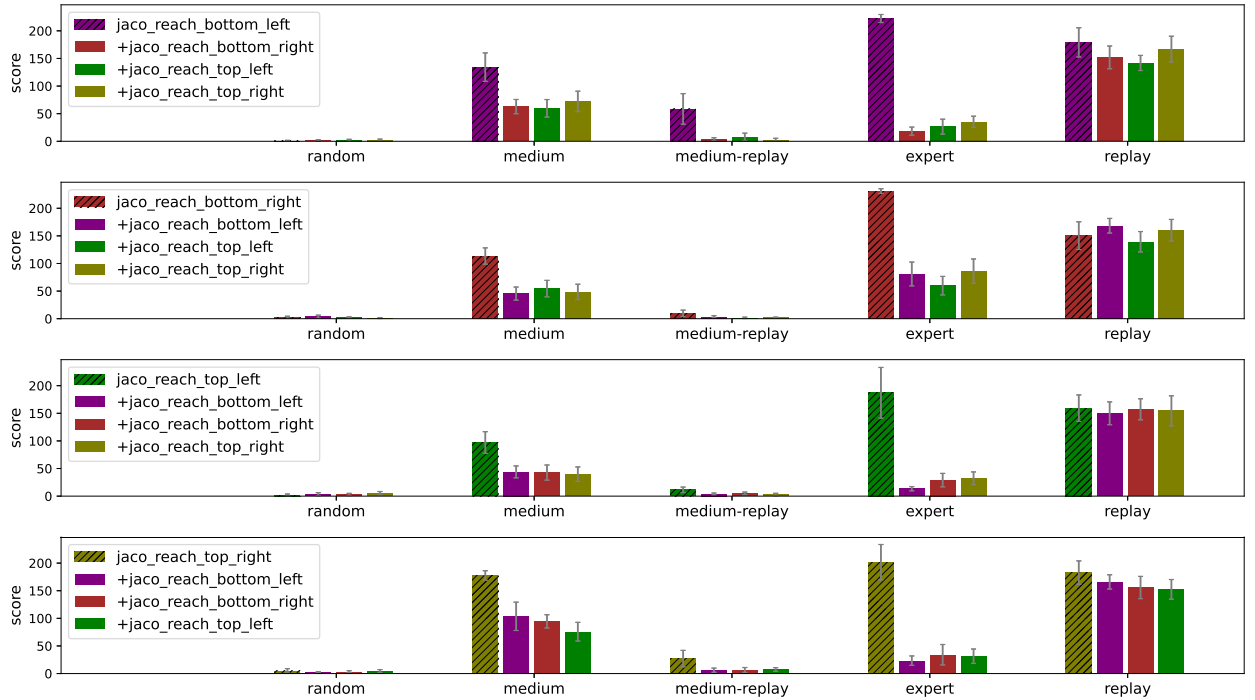


Fig. C.15. Result comparison of single-task training and direct sharing with CQL in *Jaco Arm* domain. Performing direct data sharing with policy constraints methods degrades the performance in most sharing settings.

Table C.3

Comparison of single-task training and direct sharing in three domains. The maximum scores of the *Walker* and *Quadruped* domains are 1000, and for *Jaco Arm* is 250.

	Walker		Quadruped		Jaco Arm	
	Mean	Median	Mean	Median	Mean	Median
Behavior policy	497.96	450.75	515.33	526.98	123.46	97.7
Single-task training (CQL)	363.57	267.13	437.10	411.88	106.66	128.94
Single-task training (PBRL)	522.17	542.33	561.75	599.58	125.43	141.37
Direct share (CQL)	419.35	357.97	435.96	357.97	54.64	20.47

2. In *Quadruped* domain, *Jump* task can benefit from data sharing from other tasks since first *Walk*, *Run* or *Roll* make it easier for the agent to *Jump*. Since the behavior policies of *Jump* and other tasks are related, performing direct data sharing also helps learning. For other sharing tasks, direct sharing data degenerates the performance compared to single-task training in most cases.
3. In *Jaco Arm* domain, since the different tasks drive agents to different directions, the behavior policies between tasks can be very different. Direct data sharing significantly exacerbates the distribution shift in offline RL and deteriorates performance.

We summarize the mean and medium scores for single-task training (with CQL [42] and PBRL [15]), and direct data sharing with CQL (i.e., CQL-share) in each domain. The results are shown in Table C.3. The result shows direct sharing data degenerates the performance compared to single-task training in most cases. Meanwhile, PBRL outperforms the behavior policy in all domains.

C.3. Result comparison between UTDS and CDS

We compare the performance of UTDS and CDS in three domains, and the results are shown in Fig. C.16, Fig. C.17, and Fig. C.18. In each figure, we show the single-task training results for the main task in the shadow bar.

1. In *Walker* domain, we find that UTDS generally improves performance through data sharing in the *Stand*, *Walk*, and *Flip* tasks. An exception is the *Run* task, where data sharing does not significantly improve performance. We hypothesize that the *Run* task is the most difficult task in the *Walker* domain, thus the shared data do not contain state-action pairs of the optimal trajectories in the *Run* task. As a result, sharing data in *Run* task does not bring tighter optimality bound but makes the feature representation learn slower.

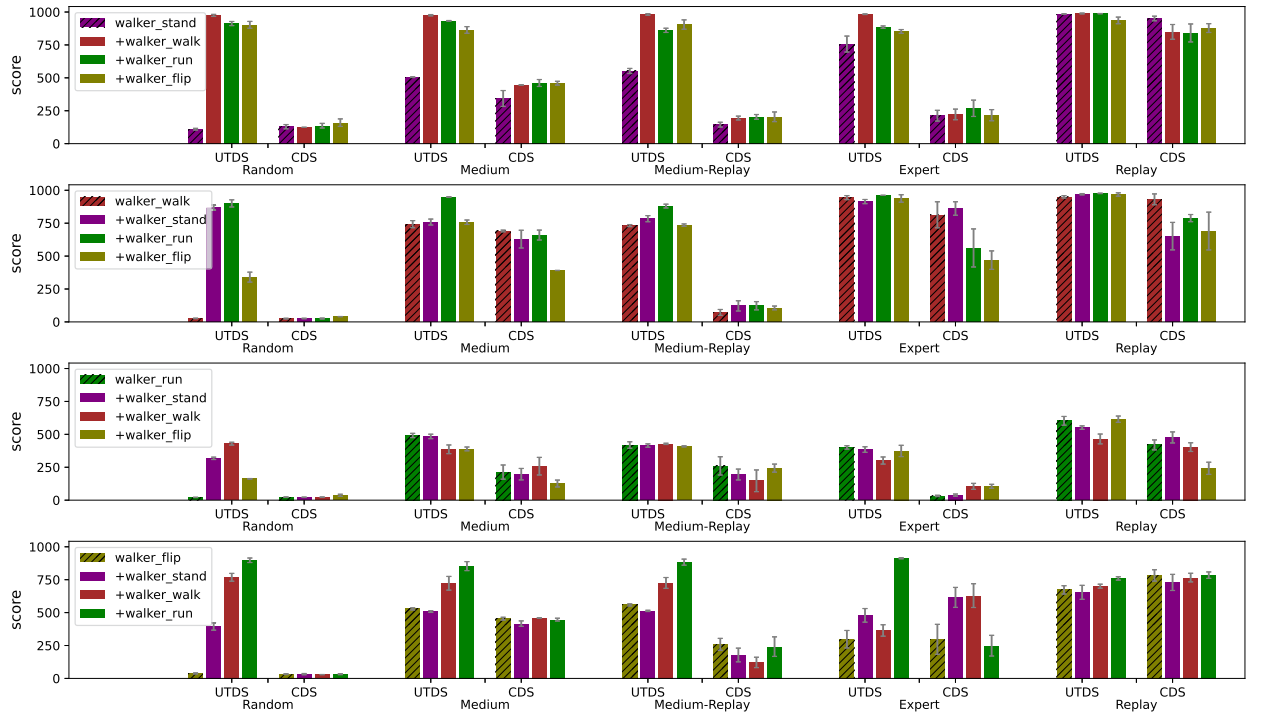


Fig. C.16. Result comparison of the proposed UTDS and CDS in Walker domain.

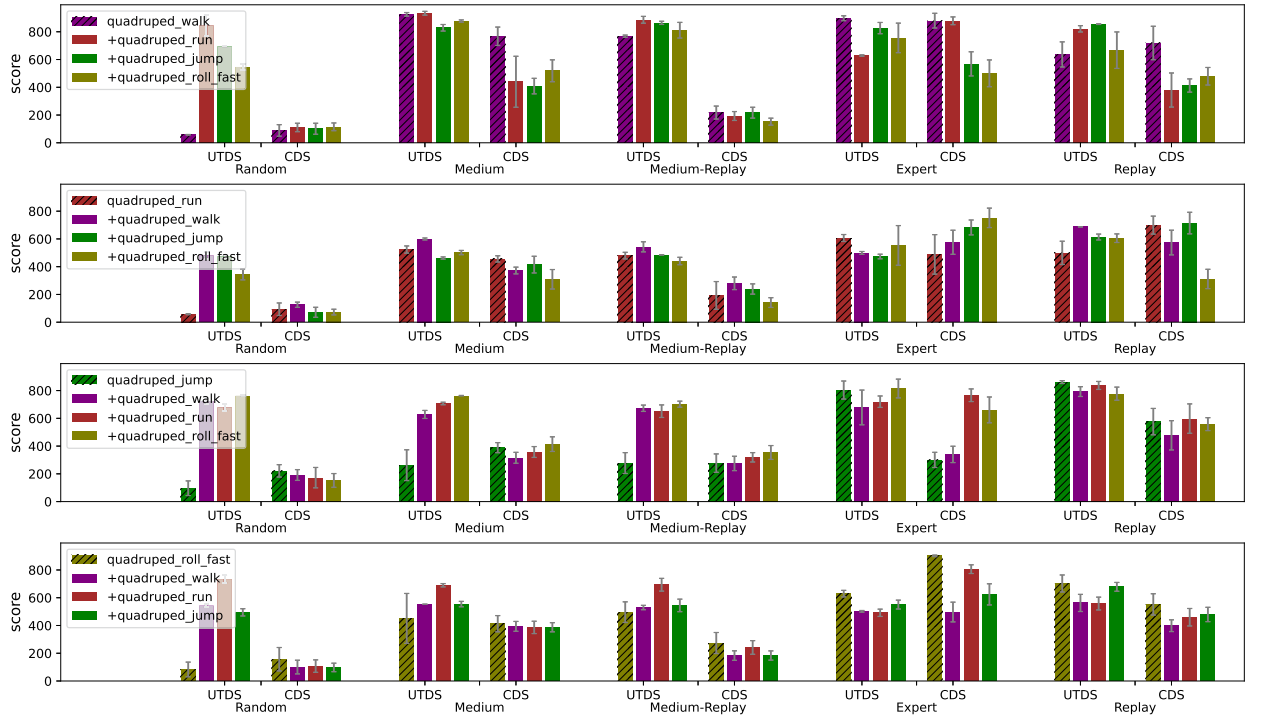


Fig. C.17. Result comparison of the proposed UTDS and CDS in Quadruped domain.

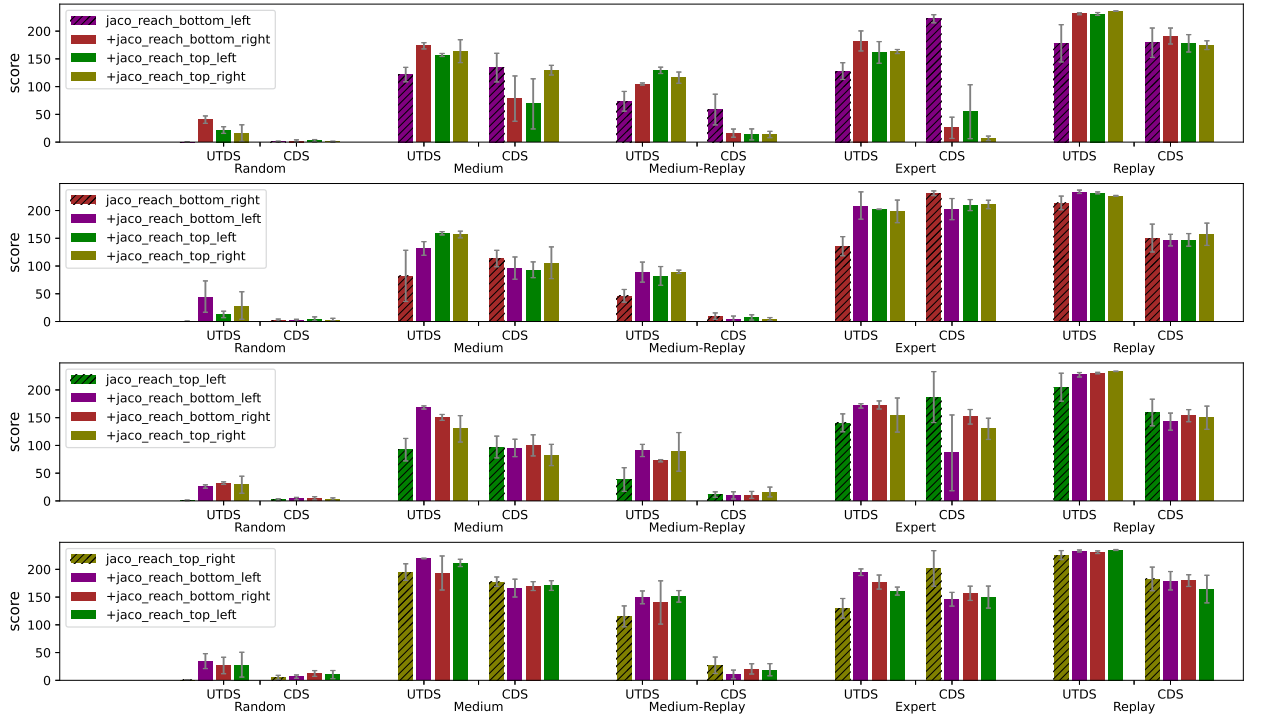


Fig. C.18. Result comparison of the proposed UTDS and CDS in *Jaco Arm* domain.

2. In *Quadruped* domain, we find that UTDS generally improves performance in the *Stand*, *Jump*, and *Roll fast* tasks, especially for non-expert datasets. Similar to *Walker* domain, since *Run* task is more difficult than other tasks, data sharing cannot help reduce the expected uncertainty of trajectories induced by the optimal policy. Compared to UTDS, CDS cannot obtain significant improvement compared to single-task training.
3. In *Jaco Arm* domain, the behavior policies between tasks can be very different since different tasks drive agents to different directions. CDS based on policy constraints cannot obtain improvements in most cases. In contrast, we find that UTDS improves performance compared to single-task training in most sharing tasks.

C.4. Result comparison between UTDS and CDS-Zero

We compare the performance of UTDS and CDS-Zero in three domains, and the results are shown in Fig. C.19, Fig. C.20, and Fig. C.21. The performance of CDS and CDS-Zero in data sharing are similar. We refer to Appendix C.3 for the analysis.

C.5. Aggregate evaluation

We follow the reliable principles [78] to evaluation the statistical performance. Since the ordinary aggregate measures like *mean* can be easily dominated by a few outlier scores, [78] presents several robust alternatives that are not unduly affected by outliers and have small uncertainty even with a handful of runs. Based on bootstrap Confidence Intervals (CIs), we can extract aggregate metrics from score distributions, including median, mean, interquartile mean (IQM), and optimality gap. IQM discards the bottom and top 25% of the runs and calculates the mean score of the remaining 50% runs. Optimality gap calculates the amount of runs that fail to meet a minimum score of $\eta = 50.0$. We also give performance profiles that reveal performance variability through score distributions. A score distribution shows the fraction of runs above a certain score and is given by $\hat{F}(\tau) = \hat{F}(\tau; x_{1:M,1:N}) = \frac{1}{M} \sum_{m=1}^M \frac{1}{N} \sum_{n=1}^N \mathbb{1}[x_{m,n} \geq \tau]$. The aggregated results and performance profiles for three domains are given in Fig. C.22 and Fig. C.23, respectively.

C.6. Ablation study

We remark that β_2 is used for constructing the pseudo-target for the OOD datapoint, as

$$\hat{\tau}^{\text{ood}} Q_i(s, a^{\text{ood}}) := Q_i(s, a^{\text{ood}}) - \beta_2 \Gamma_i(s, a^{\text{ood}}). \quad (\text{C.1})$$

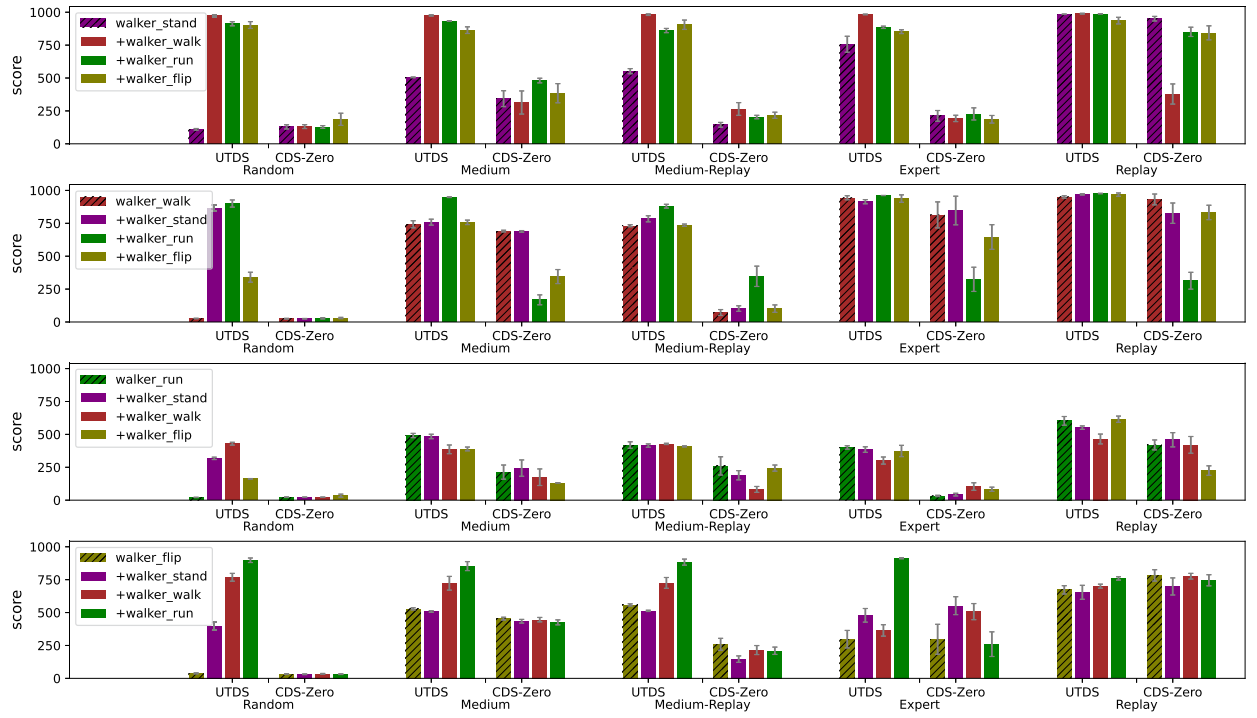


Fig. C.19. Result comparison of the proposed UTDS and CDS-Zero in Walker domain.

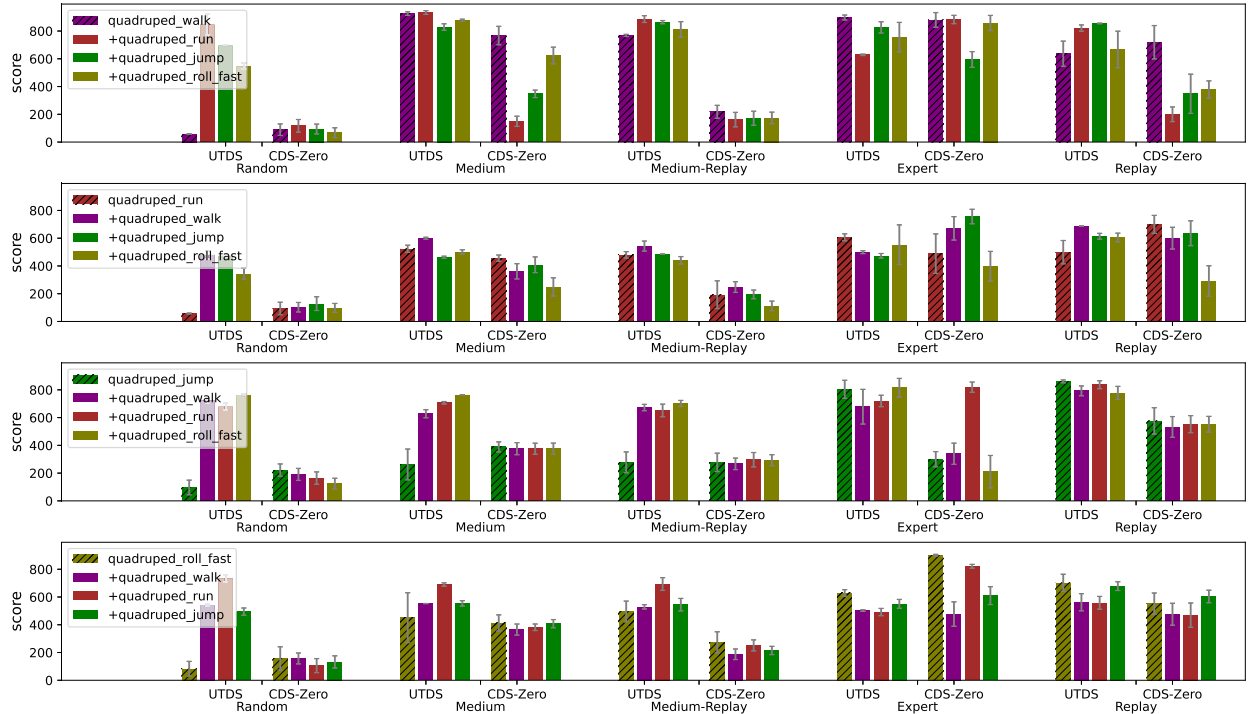


Fig. C.20. Result comparison of the proposed UTDS and CDS-Zero in Quadruped domain.

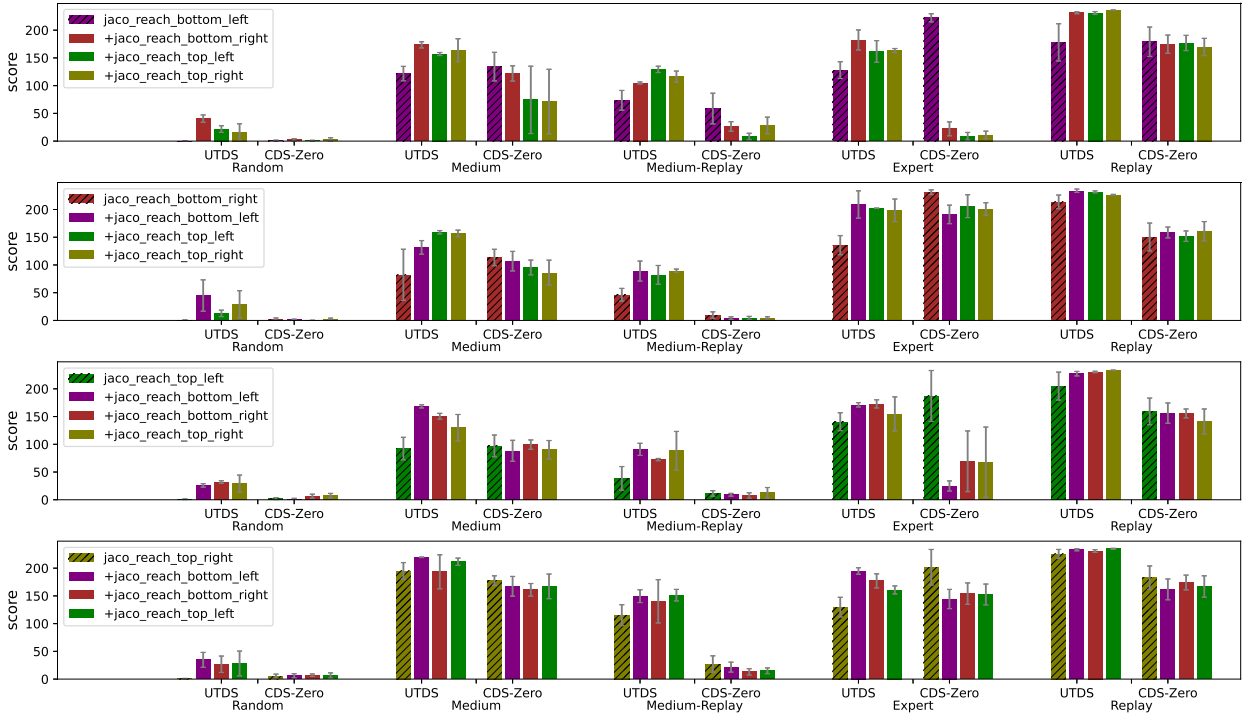


Fig. C.21. Result comparison of the proposed UTDS and CDS-Zero in Jaco Arm domain.

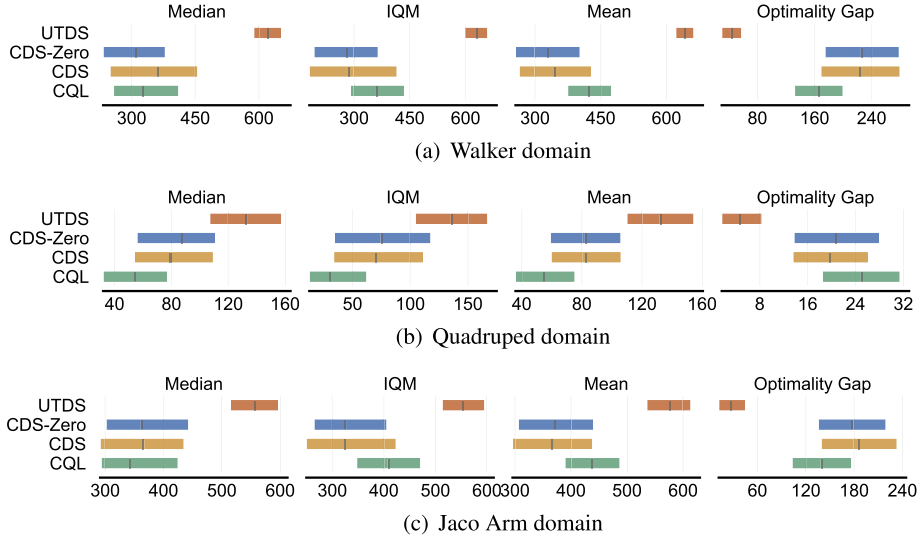


Fig. C.22. Aggregate metrics with 95% CIs based on 12 shared tasks among 5 dataset types for each task. Higher mean, median and IQM scores, and lower optimality gap are better. The CIs are estimated using the percentile bootstrap with stratified sampling.

We adopt an exponentially decay schedule for β_2 . Empirically, we use 3 factors to control the decay process of β_2 , including the initial value, the end value, and the decay factor α . In UTDS, we set the initial value to 3.0, the end value to 0.01, and the decay factor to 0.99995. We conducted ablation studies and the result is given in Fig. C.24, Fig. C.25, and Fig. C.26.

We conduct experiments on several data-sharing tasks from the Quadruped domain. According to Fig. 1, we find a large initial value (i.e., 3 or 5) is important for a stable performance. Since the uncertainty quantifier for OOD states and actions is inaccurate initially, a large penalty is required to enforce a strong regularization for the value function of OOD data. In contrast, a small factor will cause over-estimation of OOD data.

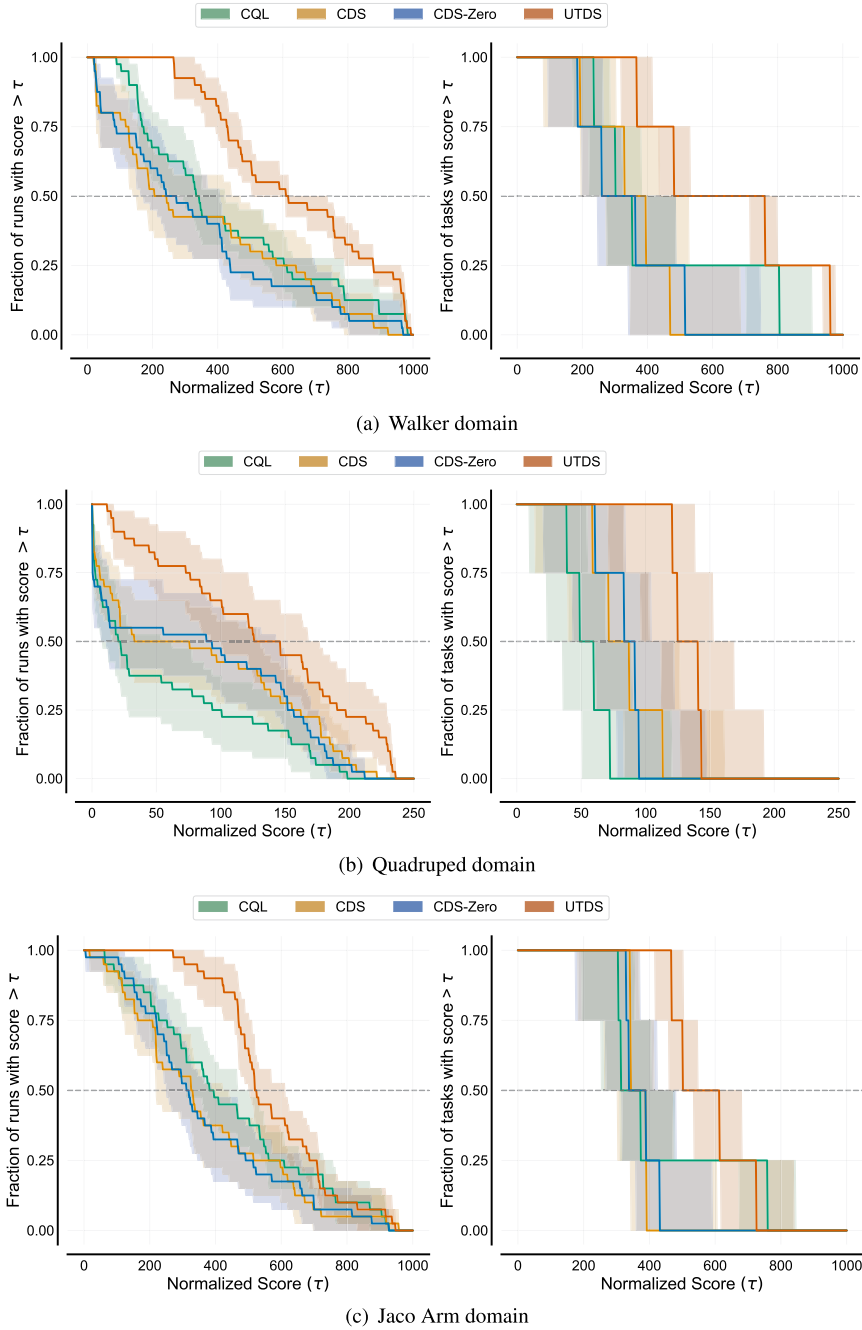


Fig. C.23. Performance profiles based on score distributions (left), and average score distributions (right). Shaded regions show pointwise 95% confidence bands based on percentile bootstrap with stratified sampling. The τ value where the profiles intersect $y = 0.5$ shows the median, and the area under the performance profile corresponds to the mean.

Fig. C.25 shows the experiment result of different end values in the Quadrupe domain. We find that setting the end value larger than 0.01 is suitable, which provides a small penalty in the final training process. In contrast, decreasing β_2 to 0 may cause over-estimation and lead to slightly worse performance.

According to our theoretical analysis, an exponential decay factor has well theoretical property to make the OOD operator not lead to negative infinity, which makes a contraction mapping for the OOD target. We try different decay factors in the experiment, the result in Fig. C.26 shows a slow decay process is important to obtain a stable process. As the uncertainty quantifier becomes more accurate in the training process, we decay the penalty factor to make the pseudo-target dependent on the estimated uncertainty gradually.

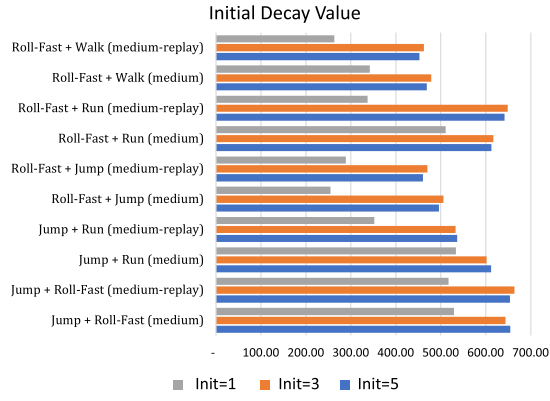


Fig. C.24. Initial decay value.

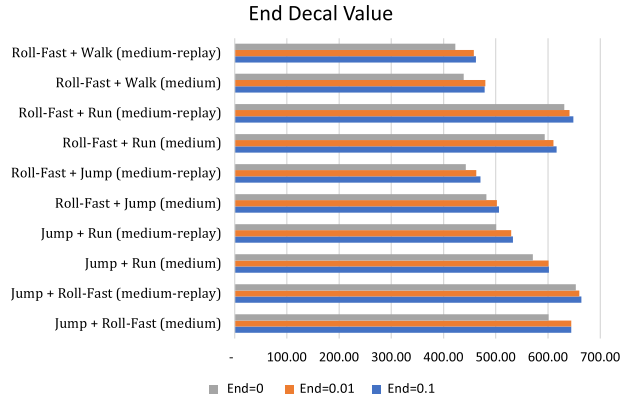
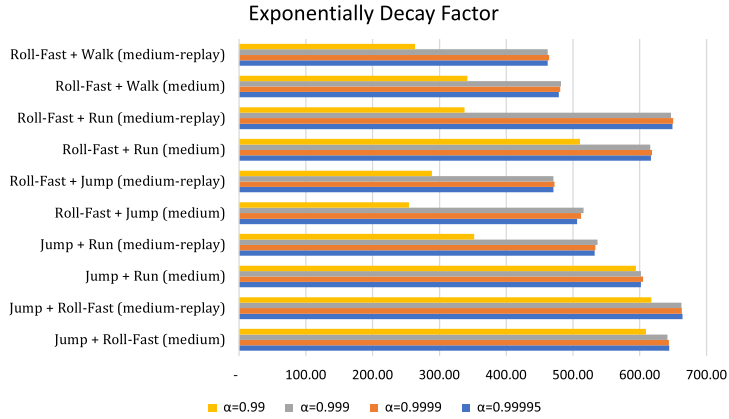


Fig. C.25. End decay value.

Fig. C.26. Ablation study of the exponentially decay factor of β_2 in OOD target.

References

- [1] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.
- [2] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M.A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, *Nature* 518 (2015) 529–533.
- [3] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., A general reinforcement learning algorithm that masters chess, shogi, and go through self-play, *Science* 362 (2018) 1140–1144.
- [4] O. Vinyals, I. Babuschkin, W.M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al., Grandmaster level in starcraft ii using multi-agent reinforcement learning, *Nature* (2019) 1–5.
- [5] P. Mirowski, M. Grimes, M. Malinowski, K.M. Hermann, K. Anderson, D. Teplyashin, K. Simonyan, A. Zisserman, R. Hadsell, et al., Learning to navigate in cities without a map, *Adv. Neural Inf. Process. Syst.* 31 (2018) 2419–2430.
- [6] C. Yu, J. Liu, S. Nemati, Reinforcement learning in healthcare: a survey, *arXiv preprint, arXiv:1908.08796*, 2019.

- [7] S. Levine, A. Kumar, G. Tucker, J. Fu, Offline reinforcement learning: tutorial, review, and perspectives on open problems, arXiv preprint, arXiv:2005.01643, 2020.
- [8] S. Lange, T. Gabel, M. Riedmiller, Batch reinforcement learning, in: Reinforcement Learning: State-of-the-Art, Springer, 2012, pp. 45–73.
- [9] S. Fujimoto, D. Meger, D. Precup, Off-policy deep reinforcement learning without exploration, in: International Conference on Machine Learning, PMLR, 2019, pp. 2052–2062.
- [10] A. Kumar, J. Fu, M. Soh, G. Tucker, S. Levine, Stabilizing off-policy q-learning via bootstrapping error reduction, in: Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 11784–11794.
- [11] Y. Jin, Z. Yang, Z. Wang, Is pessimism provably efficient for offline rl?, in: International Conference on Machine Learning, PMLR, 2021, pp. 5084–5096.
- [12] T. Xie, N. Jiang, H. Wang, C. Xiong, Y. Bai, Policy finetuning: bridging sample-efficient offline and online reinforcement learning, arXiv preprint, arXiv:2106.04895, 2021.
- [13] T. Yu, A. Kumar, Y. Chebotar, K. Hausman, S. Levine, C. Finn, Conservative data sharing for multi-task offline reinforcement learning, Adv. Neural Inf. Process. Syst. 34 (2021) 11501–11516.
- [14] I. Osband, C. Blundell, A. Pritzel, B. Van Roy, Deep exploration via bootstrapped dqn, in: Advances in Neural Information Processing Systems, vol. 29, 2016, pp. 4026–4034.
- [15] C. Bai, L. Wang, Z. Yang, Z.-H. Deng, A. Garg, P. Liu, Z. Wang, Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning, in: International Conference on Learning Representations, 2022.
- [16] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D.d.L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, et al., Deepmind control suite, arXiv preprint, arXiv:1801.00690, 2018.
- [17] A.G. Wilson, P. Izmailov, Bayesian deep learning and a probabilistic perspective of generalization, Adv. Neural Inf. Process. Syst. 33 (2020) 4697–4708.
- [18] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, R. Zecchina, Entropy-sgd: biasing gradient descent into wide valleys, J. Stat. Mech. Theory Exp. 2019 (12) (2019) 124018.
- [19] F. D'Angelo, V. Fortuin, Repulsive deep ensembles are Bayesian, Adv. Neural Inf. Process. Syst. 34 (2021) 3451–3465.
- [20] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, Adv. Neural Inf. Process. Syst. 32 (2019) 13991–14002.
- [21] Y. Jiang, V. Nagarajan, C. Baek, J.Z. Kolter, Assessing generalization of SGD via disagreement, in: International Conference on Learning Representations, 2022.
- [22] G. An, S. Moon, J.-H. Kim, H.O. Song, Uncertainty-based offline reinforcement learning with diversified q-ensemble, in: Advances in Neural Information Processing Systems, 2021.
- [23] C. Jin, Z. Yang, Z. Wang, M.I. Jordan, Provably efficient reinforcement learning with linear function approximation, in: Conference on Learning Theory, PMLR, 2020, pp. 2137–2143.
- [24] T. Jaksch, R. Ortner, P. Auer, Near-optimal regret bounds for reinforcement learning, J. Mach. Learn. Res. (2010).
- [25] M.G. Azar, I. Osband, R. Munos, Minimax regret bounds for reinforcement learning, in: International Conference on Machine Learning, 2017.
- [26] R. Wang, S.S. Du, L.F. Yang, R. Salakhutdinov, On reward-free reinforcement learning with linear function approximation, in: Advances in Neural Information Processing Systems, 2020.
- [27] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, A. Agarwal, Bellman-consistent pessimism for offline reinforcement learning, in: Advances in Neural Information Processing Systems, 2021.
- [28] Y. Abbasi-Yadkori, D. Pál, C. Szepesvári, Improved algorithms for linear stochastic bandits, in: Advances in Neural Information Processing Systems, vol. 24, 2011, pp. 2312–2320.
- [29] I. Osband, J. Aslanides, A. Cassirer, Randomized prior functions for deep reinforcement learning, Adv. Neural Inf. Process. Syst. 31 (2018).
- [30] T. Yu, A. Kumar, Y. Chebotar, K. Hausman, C. Finn, S. Levine, How to leverage unlabeled data in offline reinforcement learning, in: International Conference on Machine Learning, PMLR, 2022.
- [31] S. Fujimoto, S.S. Gu, A minimalist approach to offline reinforcement learning, in: Advances in Neural Information Processing Systems, 2021.
- [32] S.K.S. Ghasemipour, D. Schuurmans, S.S. Gu, Emaq: expected-max q-learning operator for simple yet effective offline and online rl, in: International Conference on Machine Learning, PMLR, 2021, pp. 3682–3691.
- [33] Y. Wu, G. Tucker, O. Nachum, Behavior regularized offline reinforcement learning, arXiv preprint, arXiv:1911.11361, 2019.
- [34] I. Kostrikov, R. Fergus, J. Tompson, O. Nachum, Offline reinforcement learning with Fisher divergence critic regularization, in: International Conference on Machine Learning, PMLR, 2021, pp. 5774–5783.
- [35] N.Y. Siegel, J.T. Springenberg, F. Berkenkamp, A. Abdolmaleki, M. Neunert, T. Lampe, R. Hafner, N. Heess, M. Riedmiller, Keep doing what worked: behavioral modelling priors for offline reinforcement learning, in: International Conference on Learning Representations, 2020.
- [36] Z. Wang, A. Novikov, K. Zolna, J.S. Merel, J.T. Springenberg, S.E. Reed, B. Shahriari, N. Siegel, C. Gulcehre, N. Heess, et al., Critic regularized regression, Adv. Neural Inf. Process. Syst. 33 (2020).
- [37] S. Rezaeifar, R. Dadashi, N. Vieillard, L. Hussenot, O. Bachem, O. Pietquin, M. Geist, Offline reinforcement learning as anti-exploration, in: AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 8106–8114.
- [38] W. Zhou, S. Bajracharya, D. Held, Plas: latent action space for offline reinforcement learning, in: Conference on Robot Learning, 2020.
- [39] A. Ajay, A. Kumar, P. Agrawal, S. Levine, O. Nachum, Opal: offline primitive discovery for accelerating offline reinforcement learning, in: International Conference on Learning Representations, 2021.
- [40] A. Nair, M. Dalal, A. Gupta, S. Levine, Accelerating online reinforcement learning with offline datasets, arXiv preprint, arXiv:2006.09359, 2020.
- [41] S. Lee, Y. Seo, K. Lee, P. Abbeel, J. Shin, Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble, in: Annual Conference on Robot Learning, 2021.
- [42] A. Kumar, A. Zhou, G. Tucker, S. Levine, Conservative q-learning for offline reinforcement learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1179–1191.
- [43] Y. Wu, S. Zhai, N. Srivastava, J.M. Susskind, J. Zhang, R. Salakhutdinov, H. Goh, Uncertainty weighted actor-critic for offline reinforcement learning, in: International Conference on Machine Learning, vol. 139, 2021, pp. 11319–11328.
- [44] T. Yu, G. Thomas, L. Yu, S. Ermon, J.Y. Zou, S. Levine, C. Finn, T. Ma, Mopo: model-based offline policy optimization, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 14129–14142.
- [45] S.K.S. Ghasemipour, S.S. Gu, O. Nachum, Why so pessimistic? Estimating uncertainties for offline rl through ensembles, and why their independence matters, arXiv preprint, arXiv:2205.13703, 2022.
- [46] R. Kidambi, A. Rajeswaran, P. Netrapalli, T. Joachims, Morel: model-based offline reinforcement learning, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 21810–21823.
- [47] T. Matsushima, H. Furuta, Y. Matsuo, O. Nachum, S. Gu, Deployment-efficient reinforcement learning via model-based offline optimization, in: International Conference on Learning Representations, 2020.
- [48] C. Cang, A. Rajeswaran, P. Abbeel, M. Laskin, Behavioral priors and dynamics models: improving performance and domain transfer in offline rl, arXiv preprint, arXiv:2106.09119, 2021.
- [49] T. Yu, A. Kumar, R. Rafailov, A. Rajeswaran, S. Levine, C. Finn, Combo: conservative offline model-based policy optimization, in: Advances in Neural Information Processing Systems, 2021.

- [50] G. Tennenholtz, N. Baram, S. Mannor, GELATO: geometrically enriched latent model for offline reinforcement learning, CoRR, arXiv:2102.11327, 2021.
- [51] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, I. Mordatch, Decision transformer: reinforcement learning via sequence modeling, in: *Advances in Neural Information Processing Systems*, 2021.
- [52] K. Chua, R. Calandra, R. McAllister, S. Levine, Deep reinforcement learning in a handful of trials using probabilistic dynamics models, *Adv. Neural Inf. Process. Syst.* 31 (2018).
- [53] M. Janner, J. Fu, M. Zhang, S. Levine, When to trust your model: model-based policy optimization, *Adv. Neural Inf. Process. Syst.* 32 (2019) 12519–12530.
- [54] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, K. Hausman, Mt-opt: continuous multi-task robotic reinforcement learning at scale, arXiv preprint, arXiv:2104.08212, 2021.
- [55] B. Eysenbach, X. Geng, S. Levine, R.R. Salakhutdinov, Rewriting history with inverse rl: hindsight inference for policy improvement, *Adv. Neural Inf. Process. Syst.* 33 (2020) 14783–14795.
- [56] A. Li, L. Pinto, P. Abbeel, Generalized hindsight for reinforcement learning, *Adv. Neural Inf. Process. Syst.* 33 (2020) 7754–7767.
- [57] E. Mitchell, R. Rafailov, X.B. Peng, S. Levine, C. Finn, Offline meta-reinforcement learning with advantage weighting, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 7780–7791.
- [58] J. Liu, Z. Hongyin, D. Wang, Dara: Dynamics-aware reward augmentation in offline reinforcement learning, in: *International Conference on Learning Representations*, 2021.
- [59] P.J. Ball, C. Lu, J. Parker-Holder, S. Roberts, Augmented world models facilitate zero-shot dynamics generalization from a single offline environment, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 619–629.
- [60] C. Bai, L. Wang, L. Han, J. Hao, A. Garg, P. Liu, Z. Wang, Principled exploration via optimistic bootstrapping and backward induction, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 577–587.
- [61] C. Bai, L. Wang, L. Han, A. Garg, J. Hao, P. Liu, Z. Wang, Dynamic bottleneck for robust self-supervised exploration, *Adv. Neural Inf. Process. Syst.* 34 (2021) 17007–17020.
- [62] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, D. Pathak, Planning to explore via self-supervised world models, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 8583–8592.
- [63] B. O'Donoghue, I. Osband, R. Munos, V. Mnih, The uncertainty Bellman equation and exploration, in: *International Conference on Machine Learning*, 2018, pp. 3836–3845.
- [64] K. Azizzadenesheli, E. Brunskill, A. Anandkumar, Efficient exploration through Bayesian deep q-networks, in: *2018 Information Theory and Applications Workshop (ITA)*, IEEE, 2018, pp. 1–9.
- [65] B. Mavrin, H. Yao, L. Kong, K. Wu, Y. Yu, Distributional reinforcement learning for efficient exploration, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 4424–4434.
- [66] N. Nikolov, J. Kirschner, F. Berkenkamp, A. Krause, Information-directed exploration for deep reinforcement learning, in: *International Conference on Learning Representations*, 2019.
- [67] C. Bai, T. Xiao, Z. Zhu, L. Wang, F. Zhou, A. Garg, B. He, P. Liu, Z. Wang, Monotonic quantile network for worst-case offline reinforcement learning, *IEEE Trans. Neural Netw. Learn. Syst.* (2022).
- [68] J. Shi, C. Bai, H. He, L. Han, D. Wang, B. Zhao, X. Li, X. Li, Robust quadrupedal locomotion via risk-averse policy learning, arXiv preprint, arXiv:2308.09405, 2023.
- [69] B. Lee, J. Lee, P. Vrancx, D. Kim, K.-E. Kim, Batch reinforcement learning with hyperparameter gradients, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 5725–5735.
- [70] P. Swazinna, S. Udluft, T. Runkler, Overcoming model bias for robust offline deep reinforcement learning, *Eng. Appl. Artif. Intell.* 104 (2021) 104366.
- [71] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: *International Conference on Machine Learning*, PMLR, 2016, pp. 1050–1059.
- [72] R. Yang, C. Bai, X. Ma, Z. Wang, C. Zhang, L. Han, Rorl: robust offline reinforcement learning via conservative smoothing, *Adv. Neural Inf. Process. Syst.* 35 (2022) 23851–23866.
- [73] S. Qiu, L. Wang, C. Bai, Z. Yang, Z. Wang, Contrastive ucb: provably efficient contrastive self-supervised learning in online reinforcement learning, in: *International Conference on Machine Learning*, PMLR, 2022, pp. 18168–18210.
- [74] J. Fu, A. Kumar, O. Nachum, G. Tucker, S. Levine, D4rl: datasets for deep data-driven reinforcement learning, arXiv preprint, arXiv:2004.07219, 2020.
- [75] S. Fujimoto, H. Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: *International Conference on Machine Learning*, PMLR, 2018, pp. 1587–1596.
- [76] A. Stooke, K. Lee, P. Abbeel, M. Laskin, Decoupling representation learning from reinforcement learning, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 9870–9879.
- [77] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, *Adv. Neural Inf. Process. Syst.* 33 (2020) 21271–21284.
- [78] R. Agarwal, M. Schwarzer, P.S. Castro, A. Courville, M.G. Bellemare, Deep reinforcement learning at the edge of the statistical precipice, in: *Advances in Neural Information Processing Systems*, 2021.