Hybrid Transfer Reinforcement Learning: Provable Sample Efficiency from Shifted-Dynamics Data

Chengrui Qu PKU
Laixi Shi Caltech
Kishan Panaganti Caltech
Pengcheng You PKU
Adam Wierman Caltech

Abstract

Online reinforcement learning (RL) typically requires online interaction data to learn a policy for a target task, but collecting such data can be high-stakes. This prompts interest in leveraging historical data to improve sample efficiency. The historical data may come from outdated or related source environments with different dynamics. It remains unclear how to effectively use such data in the target task to provably enhance learning and sample efficiency. To address this, we propose a hybrid transfer RL (HTRL) setting, where an agent learns in a target environment while accessing offline data from a source environment with shifted dynamics. We show that – without information on the dynamics shift – general shifted-dynamics data, even with subtle shifts, does not reduce sample complexity in the target environment. However, focusing on HTRL with prior information on the degree of the dynamics shift, we design HySRL, a transfer algorithm that outperforms pure online RL with problem-dependent sample complexity guarantees. Finally, our experimental results demonstrate that HySRL surpasses the state-of-the-art pure online RL baseline.

1 Introduction

In online reinforcement learning (RL), an agent learns by continuously interacting with an unknown environment. While this approach has led to remarkable successes across various domains, such as robotics (Espeholt et al., 2018), traffic control (He et al., 2023) and game playing (Silver et al., 2017), it often requires billions of data from interactions to develop an effective policy (Li et al., 2023b). Moreover, in many real-world scenarios, such interactions can be costly, time-consuming, or unsafe (Eysenbach et al., 2021), which significantly limits the broader application of RL in practice, highlighting the urgent need for more sample-efficient paradigms.

One promising direction to address sample inefficiency in RL is transfer learning (Zhu et al., 2023). When developing an effective policy for a target environment, it is often possible to leverage experiences from a similar source environment with shifted dynamics (Niu et al., 2024). These sources may include an imperfect simulator (Peng et al., 2018), historical operating data before external impacts (Luo et al., 2022), polluted offline datasets (Wang et al., 2023), or data from other tasks in a multi-task setting (Sodhani et al., 2021). This concept has led to various domains and pipelines, such as meta RL (Finn et al., 2017), cross-domain RL (Eysenbach et al., 2021; Liu et al., 2022), and distributionally robust RL (Shi et al., 2023), which demonstrate varying levels of effectiveness.

Despite recent practical progress, there are still no clear indications on how to perform transfer learning in a sample-efficient manner with guarantees. While some studies show that using shifted dynamics data can reduce the time required to achieve specific performance levels in the target environment (Liu et al., 2022; Serrano et al., 2023; Zhang et al., 2024a), others indicate that sometimes these transfers hinder rather than help learning (Ammar et al., 2015; You et al., 2022), a phenomenon known as negative transfer.

These practical challenges highlight the need for theoretical insights, which have not been addressed in existing frameworks. Recently, a new stream of research called hybrid RL (Xie et al., 2021) has emerged, showing that, theoretically, an offline dataset with no dynamics shift can facilitate more efficient online exploration. However, when the dataset is collected from a source environment with shifted dynamics, it remains unclear whether this dataset can still enable

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

more sample-efficient learning in the target environment. This brings us an interesting open question:

> *Can data from a shifted source environment be leveraged to provably enhance sample efficiency when learning in a target environment?*

To answer this question, we formulate a problem called hybrid transfer RL (HTRL), where an agent aims to learn an optimal policy for the target environment with minimal interactions, while having access to an offline dataset collected from a shifted source environment. The source and target environments differ in their environmental transition uncertainties (Doshi-Velez and Konidaris, 2013), which are typically unknown before exploring the target environment and thus refer to as an unknown dynamics shift.

Contributions. In this work, we propose a hybrid transfer RL setting, where the source and target environments share the same world structure, but differing in their dynamic transitions. We first investigate the inherent difficulty of general hybrid transfer RL by providing a minimax lower bound on the required sample complexity, and then demonstrate provable sample efficiency gains from the source environment dataset when additional prior information is available. To the best of our knowledge, we are the first to theoretically study the sample complexity of this transfer setting. Specifically:

- We formulate and focus on a new setting called hybrid transfer RL (HTRL). We find that, even with a subtle dynamics shift between the target MDP and the source MDP, datasets from the source MDP generally cannot reduce the sample complexity required for the target MDP without additional conditions, compared to state-of-the-art online RL sample complexity (Theorem 1). This result demonstrates that general HTRL is not feasible, motivating us to focus on more practical yet meaningful settings.
- We study HTRL with prior knowledge of the degree of the dynamics shift. We design a transfer algorithm, HySRL, which achieves problem-dependent sample complexity at least as good as state-of-the-art online methods, providing sample efficiency gains in many practical and meaningful scenarios (Theorem 2). The key technical contributions involve extending the current reward-free and bonus-based exploration techniques to accommodate more general rewards and incorporating shifted-dynamics data into the analysis. In addition, we conduct experiments in the GridWorld environment to evaluate the proposed algorithm HySRL, demonstrating that HySRL achieves superior sample efficiency than the state-of-the-art pure online RL baseline.

1.1 Related work

Finite-sample analysis of online, offline, and hybrid RL. Finite sample analysis in RL focuses on understanding the sample complexity – how many samples are required to achieve a desired policy with certain performance. In this line of research, a non-exhaustive list in online RL includes Dong et al. (2019); Zhang et al. (2021, 2020a); Jafarnia-Jahromi et al. (2020); Liu and Su (2021); Yang et al. (2021); Azar et al. (2017); Jin et al. (2018); Bai et al. (2019); Zhang et al. (2020b); Menard et al. (2021); Domingues et al. (2021b); He et al. (2021); Zanette and Brunskill (2019); Li et al. (2023a); Zhang et al. (2024b), while offline RL has seen advances such as Uehara et al. (2020); Li et al. (2014); Yang et al. (2020); Duan et al. (2020); Jiang and Li (2016); Jiang and Huang (2020); Kallus and Uehara (2020); Duan et al. (2021); Xu et al. (2021); Ren et al. (2021); Panaganti et al. (2025); Thomas and Brunskill (2016); Shi et al. (2022); Li et al. (2024a); Woo et al. (2024), and hybrid RL frameworks are explored in Song et al. (2023); Xie et al. (2021); Zhang and Zanette (2023); Li et al. (2024b). The most closely related setting is hybrid RL, in which an agent learns in a target environment with access to an offline dataset collected from the same environment. Our work extends hybrid RL by addressing cases where the offline dataset may come from an outdated or related environment with shifted dynamics relative to the target environment.

Transfer RL with dynamics shifts. One closely related setting is cross-domain RL with dynamics shifts, focusing on leveraging abundant samples from a source environment to reduce data requirements for a target environment (Eysenbach et al., 2021; Liu et al., 2022; Niu et al., 2022, 2023; Chen et al., 2024; Wen et al., 2024). To address the dynamics shift between the source and target environments, existing approaches often involve training a classifier to distinguish between source and target transitions, combined with techniques such as combining source and target datasets for policy training (Wen et al., 2024; Chen et al., 2024), and reshaping rewards by introducing a penalty term for dynamics shifts (Eysenbach et al., 2021; Liu et al., 2022). While these methods show promising empirical performance, a systematic study on sample complexity is missing. Our work fills this gap by offering a novel theoretical perspective on cross-domain RL.

Other related transfer RL settings include distributionally robust offline RL (Zhou et al., 2021; Panaganti and Kalathil, 2022; Xu et al., 2023; Panaganti

et al., 2022, 2024; Shi et al., 2023; Wang et al., 2024; Ma et al., 2023; Liu and Xu, 2024; Shi and Chi, 2024), which focuses on training a robust policy using only an offline dataset, without further exploration, to optimize performance in the worst-case scenario of the target environment. Another area, meta RL (Finn et al., 2017; Duan et al., 2016; Wang et al., 2017; Chen et al., 2022; Ye et al., 2023; Mutti and Tamar, 2024), trains an agent over a distribution of environments to enhance generalization capabilities. Our work contributes to distributionally robust offline RL by addressing the sample complexity when exploration in the target environment is allowed and complements meta RL by focusing on scenarios where the target environment lies outside the training distribution.

Notation. We denote by $[n]$ the set $\{1, \cdots, n\}$ for any positive integer n, and use $\mathbb{1}\{\cdot\}$ to represent the indicator function. For a function f defined on S, we define its expectation under the probability measure p as $pf \triangleq \mathbb{E}_{s\sim p} f(s)$ and its variance as $\mathrm{Var}_p(f) \triangleq \mathbb{E}_{s\sim p}(f(s) - \mathbb{E}_{s'\sim p} f(s'))^2 = p(f-pf)^2$. The total variation distance between probability measures p and q is defined as $\mathrm{TV}(p,q) \triangleq \sup_{A\subseteq S} |p(A) - q(A)|$. We use standard $O(\cdot)$ and $\Omega(\cdot)$ notation, where $f = O(g)$ means there exists some constant $C > 0$ such that $f \le Cg$ (similarly for $\Omega(\cdot)$), and use the tilde notation $\widetilde{O}(\cdot)$ to suppress additional log factors. We denote the cardinality of a set $\mathcal{X}$ by $|\mathcal{X}|$.

2 Hybrid Transfer RL

We begin by introducing the mathematical formulation of HTRL, benchmarking with standard online RL.

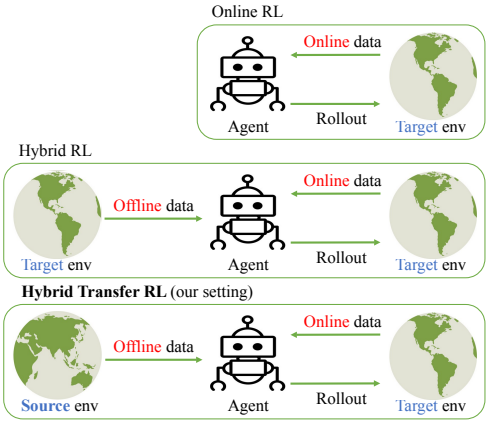


Figure 1: Comparison between different RL settings

Background: Markov decision process (MDP). We consider episodic Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, p, r, \rho)$, where $\mathcal{S}$ is the state space with size S, $\mathcal{A}$ is the action space with size A, H is the horizon length. $p(\cdot \mid s,a) : \mathcal{S}\times\mathcal{A} \mapsto \Delta(\mathcal{S})$ denotes the time-independent transition probability at each step, and the reward function is deterministic[1], given by $r : \mathcal{S}\times\mathcal{A} \mapsto [0,1]$. In this setting, a Markovian policy is given by $\pi := \{\pi_h\}_{h=1}^H$ where $\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})$. Additionally, we assume that each episode of the MDP starts from an initial state generated from an unknown distribution $\rho \in \Delta(\mathcal{S})$, namely, $s_1 \sim \rho$.

For a given transition p, the value function for state s at step h is defined as the expected cumulative future reward by executing policy π, which is given by $V_h^{p,\pi}(s) := \mathbb{E}_{p,\pi}[\sum_{i=h}^H r(s_i,a_i) \mid s_h = s]$. Similarly, the action-value function, or Q-function, is defined as $Q_h^{p,\pi}(s,a) = \mathbb{E}_{p,\pi}[\sum_{i=h}^H r(s_i,a_i) \mid s_h = s, a_h = a]$. We denote the expected value function of π with ρ as the initial state distribution by:

$$V_1^{p,\pi}(\rho) = \mathbb{E}_{s\sim\rho}[V_1^{p,\pi}(s)].$$

As is well known, there exists at least one deterministic policy that maximizes the value function and the Q-function simultaneously for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ (Bertsekas, 2007). Let $\pi^\star$ denote an optimal deterministic policy, and the corresponding optimal value function $V_h^\star$ and optimal Q-function $Q_h^\star$ are defined respectively by $V_h^{p,\star} \triangleq V_h^{p,\pi^\star}$, $Q_h^{p,\star} \triangleq Q_h^{p,\pi^\star}$, $\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.

2.1 Hybrid Transfer RL

In HTRL, the agent can interact with the target MDP $\mathcal{M}_{\mathrm{tar}} = (\mathcal{S}, \mathcal{A}, H, p_{\mathrm{tar}}, r, \rho)$ in a online manner. Specifically, in each episode, at each step $h \in [H]$, the agent observes a state $s_h \in \mathcal{S}$, chooses an action $a_h \in \mathcal{A}$, receives a reward $r(s_h, a_h)$ and transitions to the next state s_{h+1} at time step $h+1$, according to the underlying transition probability $p_{\mathrm{tar}}(\cdot \mid s_h, a_h)$, and so on so force.

Additionally, the agent has access to an offline dataset $\mathcal{D}_{\mathrm{src}} = \{(s_i, a_i, r_i, s_i')\}$ pre-collected from a source MDP $\mathcal{M}_{\mathrm{src}} = (\mathcal{S}, \mathcal{A}, H, p_{\mathrm{src}}, r, \rho)$. The target and source MDPs share the same structure except for the transition probabilities (i.e. $p_{\mathrm{tar}} \neq p_{\mathrm{src}}$). For simplicity, we assume the reward signals in $\mathcal{M}_{\mathrm{src}}$ and $\mathcal{M}_{\mathrm{tar}}$ are the same; however, our analysis still holds when the reward signals differ. We assume p_{src} and p_{tar} are both unknown to the agent.

Goal. With access to both $\mathcal{D}_{\mathrm{src}}$ and $\mathcal{M}_{\mathrm{tar}}$, the goal in HTRL is to find an ε-optimal policy for $\mathcal{M}_{\mathrm{tar}}$ with minimal online interactions with $\mathcal{M}_{\mathrm{tar}}$. Specifically,

[1]For simplicity, we consider deterministic rewards, as estimating rewards is not a significant challenge in deriving sample complexity results.

the agent aims to find a policy $\hat{\pi}$ for $\mathcal{M}_{\text{tar}}$, which satisfies that:

$$V_1^{p_{\text{tar}},\star}(\rho) - V_1^{p_{\text{tar}},\hat{\pi}}(\rho) \le \varepsilon.$$

Benchmarking with standard online RL. Online RL is a popular setting in which the agent learns from scratch by directly interacting with $\mathcal{M}_{\text{tar}}$ in episodes of length H. Different from HTRL, in online RL, the agent does not have the access to an offline dataset as additional information. Therefore, compared to online RL, the introduction of additional access to $\mathcal{D}_{\text{src}}$ in HTRL naturally raises the question: can we achieve better sample efficiency by leveraging $\mathcal{D}_{\text{src}}$? The answer is negative in general but potentially positive in many practical settings, which will be highlighted in the next two sections.

3 Minimax Lower Bound For HTRL

In this section, we establish a minimax lower bound on the sample complexity for general HTRL, formally demonstrating that sample complexity improvements for general HTRL are not feasible even with subtle shift.

Specifically, when p_{tar} is close to p_{src}, one might expect that fewer samples from $\mathcal{M}_{\text{tar}}$ are needed to reach a given performance level by leveraging additional information about $\mathcal{M}_{\text{src}}$ (Lobel and Parr, 2024, Section B). However, as demonstrated in Theorem 1, in the worst case, it still requires samples of same order from $\mathcal{M}_{\text{tar}}$ as in pure online RL. The proof of this theorem can be found in the full version Qu et al. (2024).

Theorem 1 (Minimax lower bound for HTRL)**.** *Suppose $S \ge 3$, $H \ge 3$, $A \ge 2$, $\varepsilon \le 1/48$. Consider any $\mathcal{M}_{\text{src}}$ and define the following set of possible MDPs:*

$$\mathcal{M}_\alpha \triangleq \{\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, p, r, \rho) \mid \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathrm{TV}(p(\cdot \mid s,a), p_{\text{src}}(\cdot \mid s,a)) \le \alpha\},$$

where $48\varepsilon/H^2 \le \alpha \le 1$. For any algorithm, there exists a $\mathcal{M}_{\text{src}}$ and a target MDP $\mathcal{M}_{\text{tar}} \in \mathcal{M}_\alpha$, if the number of samples n collected from the target MDP satisfies $n \le O(H^3SA/\varepsilon^2)$, then the algorithm suffers from an ε-suboptimality gap:

$$\mathbb{E}_{\text{tar}}\left[V_1^{p_{\text{tar}},\star}(\rho) - V_1^{p_{\text{tar}},\hat{\pi}}(\rho)\right] \ge \varepsilon,$$

where $\mathbb{E}_{\text{tar}}$ denotes the expectation with respect to the randomness during algorithm execution in the target MDP $\mathcal{M}_{\text{tar}}$.

Theorem 1 shows that the lower bound of sample complexity of general HTRL is $\Omega(H^3SA/\varepsilon^2)$, which matches the state-of-the-art sample complexity of pure online RL, $\widetilde{O}(H^3SA/\varepsilon^2)$ (e.g., Ménard et al. (2021)[2]; Wainwright (2019)). This demonstrates that, in general, practical transfer algorithms leveraging source environment data cannot reduce the sample complexity in the target environment. No matter what algorithms are used, there always exists a worst case where transfer learning cannot achieve better sample efficiency in the target environment, compared to pure online RL. That said, this lower bound is conservative, motivating us to explore practically meaningful and feasible settings to derive problem-dependent sample complexity bounds.

Comparisons to prior lower bounds. To the best of our knowledge, this is the first lower bound on the sample complexity when leveraging information from a source environment to explore a new target environment with an unknown dynamics shift. We highlight the novelty of our lower bound result by comparing it with prior lower bounds:

Lower bounds for transfer in RL: O' Donoghue (2021); Ye et al. (2023); Mutti and Tamar (2024) provide lower bounds on regret in settings where the agent is trained on N source tasks and is fine-tuned to the target task during testing. However, these lower bounds cannot be adapted to our setting as they assume the target task is one of the source tasks – which is stronger than ours.

Lower bounds for pure online RL: The existing lower bound on the sample complexity of pure online RL is also $\Omega(H^3SA/\varepsilon^2)$ (Gheshlaghi Azar et al., 2013). This demonstrates that the improvement in the sample complexity lower bound from the introducing additional information from a source environment is at most a constant factor. The construction of the lower bound for HTRL follows a procedure similar to existing lower bounds for online RL (Lattimore and Hutter, 2012; Yin et al., 2020), offline RL (Rashidinejad et al., 2021) and hybrid RL without dynamics shift (Xie et al., 2021). However, the new technical challenge in our setting is to bound the maximum information gain of the data from $\mathcal{M}_{\text{src}}$. We address this difficulty using a proper change-of-measure approach.

4 HTRL with Separable Shift

Although improved sample efficiency is not achievable for general HTRL in the worst case, practical tasks are typically more manageable than these difficult in-

[2] Ménard et al. (2021) considers time-dependent transitions and the sample complexity result is $\widetilde{O}(H^4SA/\varepsilon^2)$, which in our setting translates into $\widetilde{O}(H^3SA/\varepsilon^2)$ due to time-independent transitions.

stances. Inspired by practical tasks such as hierarchical RL (Chua et al., 2023) and meta RL (Chen et al., 2022), we instead focus on a class of HTRL with separable shift in the following. This setting allows us to leverage prior information about the degree of dynamics shift between the source and target environments. We then propose an algorithm, called HySRL, which achieves provably superior sample complexity compared to pure online RL.

4.1 β-separable shfits

We first introduce the definition of separable shift, characterized by the minimal degree of the dynamics shift between the source and target environments.

Definition 1 (β-separable shift)**.** *Consider a target MDP $\mathcal{M}_{\text{tar}} = (\mathcal{S}, \mathcal{A}, H, p_{\text{tar}}, r, \rho)$ and a source MDP $\mathcal{M}_{\text{src}} = (\mathcal{S}, \mathcal{A}, H, p_{\text{src}}, r, \rho)$. The shift between $\mathcal{M}_{\text{tar}}$ and $\mathcal{M}_{\text{src}}$ is β-separable if for some $\beta \in (0, 1]$, we have for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$p_{\text{src}}(\cdot \mid s, a) \neq p_{\text{tar}}(\cdot \mid s, a) \implies \text{TV}(p_{\text{src}}(\cdot \mid s, a), p_{\text{tar}}(\cdot \mid s, a)) \geq \beta.$$

In other words, for any state-action pair (s, a), the transitions in $\mathcal{M}_{\text{src}}$ and $\mathcal{M}_{\text{tar}}$ are either identical or different by at least the degree of β w.r.t the TV distance metric. This definition is widely used to characterize the "distance" between tasks in hierarchical RL (Chua et al., 2023), RL for latent MDPs (Kwon et al., 2024), multi-task RL Brunskill and Li (2013), and meta RL (Mutti and Tamar, 2024; Chen et al., 2022), serving the purpose of distinguishing different tasks with finite samples in practice.

Such a minimal degree of dynamic shift, β, can often be estimated beforehand as prior information for specific problems in practice (Brunskill and Li, 2013). Therefore, in this section, we design algorithms under the assumption that p_{tar} and p_{src} are β-separable.

Remark 1 (Separable shift makes HTRL feasible)**.** *The lower bound in Theorem 1 arises from potential challenging target MDPs that subtly differ from the source MDP, with a small "distance" associated with the required optimality gap ε. This subtlety requires extensive data to distinguish between them. However, in practice, the dynamics shift between source and target environments, characterized by β, is typically independent of ε. Therefore, we focus on HTRL with a β-separable shift, excluding over-conservative instances that are rare in practice.*

In addition to the aforementioned key definition — β-separable dynamics shifts, we introduce another assumption for the reachability of the target MDP. Note that it is not tailored for our Hybrid Transfer RL setting, but widely adopted in extensive RL tasks such as standard RL, meta RL and multi-task RL (Jaksch et al., 2010; Chen et al., 2022; Brunskill and Li, 2013), to ensure the problems are well-posed with agent's access to the entire environment (over all state-action pairs).

Assumption 1 (σ-reachability)**.** *We assume the target MDP $\mathcal{M}_{\text{tar}}$ has σ-reachability if there exists a constant $\sigma \in (0, 1]$ so that*

$$\max_{\pi} \max_{h \in [H]} d_h^{\pi}(s, a) \geq \sigma, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A},$$

where $d_h^{\pi}(s, a)$ is the probability of reaching (s, a) at step h by executing policy π in $\mathcal{M}_{\text{tar}}$.

4.2 Algorithm design: HySRL

Focusing on HTRL with β-separable shift, now we are ready to introduce our algorithm HySRL, outlined in Algorithm 1. To explicitly characterize the set of state-action pairs where p_{src} and p_{tar} differ, we introduce the following definition.

Definition 2 (Shifted region)**.** *We define the shifted region $\mathcal{B}$ as the set of state-action pairs where the transitions in $\mathcal{M}_{\text{src}}$ and $\mathcal{M}_{\text{tar}}$ differ:*

$$\mathcal{B} \triangleq \left\{(s, a) \in \mathcal{S} \times \mathcal{A} \mid p_{\text{src}}(\cdot \mid s, a) \neq p_{\text{tar}}(\cdot \mid s, a)\right\}.$$

Although p_{tar} is unknown in advance, it is possible to invest a small number of online samples to estimate p_{tar} and identify the shifted region $\mathcal{B}$. This helps determine which part of $\mathcal{D}_{\text{src}}$ can improve sample efficiency in $\mathcal{M}_{\text{tar}}$, allowing us to focus further exploration on the remaining areas to learn an effective policy. Since, in many practical applications, the dynamcis shift typically affects only a small portion of the state-action space (Chua et al., 2023), this approach can enable more sample-efficient exploration in $\mathcal{M}_{\text{tar}}$. This intuition drives the design of Algorithm 1.

Algorithm 1: Hybrid separable-transfer RL (HySRL). At a high level, given a desired optimality gap ε, if the reachability σ and the minimal shift level β is relatively small – implying that an excessive number of samples is required to identify the shifted region – Algorithm 1 chooses to ignore the offline dataset and instead relies on pure online learning. Otherwise, we proceed as follows: first, we run Algorithm 2 to obtain an estimated shifted region $\hat{\mathcal{B}}$, which, with high probability, matches the true shifted region $\mathcal{B}$. Next, in Algorithm 3, we further enhance exploration of $\hat{\mathcal{B}}$ by designing exploration bonuses that combine both the offline dataset $\mathcal{D}_{\text{src}}$ and online data, ultimately yielding a final policy for $\mathcal{M}_{\text{tar}}$. Below, we outline the key steps of Algorithm 1.

Algorithm 1 Hybrid separable-transfer RL

Require: Shift level β, confidence level δ, reachability σ, desired optimality gap ε, source dataset $\mathcal{D}_{\text{src}}$

if $\beta \le \sqrt{S/H}\varepsilon/\sigma$ **then**
  // If the shift is relatively hard to identify
  Set the estimated shifted region $\hat{\mathcal{B}}$ as $\mathcal{S}\times\mathcal{A}$
  // Abandon $\mathcal{D}_{\text{src}}$
else
  Execute Algorithm 2 to identify and estimate shifted region $\hat{\mathcal{B}}$
end if
Execute Algorithm 3 with $\hat{\mathcal{B}}$ to get π^{final}
// Collect data for $\hat{\mathcal{B}}$ and learn the policy through value iteration
return π^{final}

Step 1: Reward-free shift identification (Algorithm 2). Even with knowledge of β and σ, accurately estimating p_{tar} to identify the shifted region $\mathcal{B}$ is still challenging, as we need to control the errors in estimating high-dimensional transitions with finite samples.

Algorithm 2 Reward-free shift identification

Require: Shift level β, confidence level δ, reachability σ, source dataset $\mathcal{D}_{\text{src}}$

for $t = 0, 1, 2, \cdots$ **do**
  for $h = H, \cdots, 1$ **do**
    Update W_h^t using Eq. (1)
    Update $\pi_h^{t+1}(\cdot) = \arg\max_{a\in\mathcal{A}} W_h^t(\cdot, a)$
  end for
  Break if $3\sqrt{\mathbb{E}_{\rho,\pi_1^{t+1}}[W_1^t]} + \mathbb{E}_{\rho,\pi_1^{t+1}}[W_1^t] \le \sigma\beta/8$
  // sufficient data coverage is achieved
  Rollout π^{t+1} and observe new online samples
  for $(s,a)\in\mathcal{S}\times\mathcal{A}$ **do**
    Update visitation counts $n^t(s,a)$, $n^t(s,a,s')$ and empirical transitions $\hat{p}_{\text{tar}}^t(\cdot \mid s,a)$
  end for
end for
return Estimated shifted area $\hat{\mathcal{B}}$ by Eq. (2)

To this end, sufficient online data coverage is required for each (s,a), which aligns with the motivation behind reward-free exploration to collect enough data and achieve optimality for any reward signal $r : \mathcal{S}\times\mathcal{A}\mapsto[0,1]$. Inspired by RF-Express from Ménard et al. (2021), we propose Algorithm 2. Specifically, we first define an uncertainty function $W_h^t(s,a)$, which characterizes data sufficiency until the t^{th} episode, recursively (with $W_{H+1}^t(s,a) = 0$) for all $h\in[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$W_h^t(s,a) \triangleq \min\Bigg(1, \frac{4Hg_1(n^t(s,a),\delta)}{n^t(s,a)} + \sum_{s'}\hat{p}_{\text{tar}}^t(s'\mid s,a)\max_{a'\in\mathcal{A}} W_{h+1}^t(s',a')\Bigg), \tag{1}$$

where $g_1(n,\delta) \triangleq \log(6SAH/\delta) + S\log(8e(n+1))$, $n^t(s,a) \triangleq \sum_{\tau=1}^t\sum_{h=1}^H \mathbb{1}\{(s_h^\tau, a_h^\tau) = (s,a)\}$ denotes the visitation count for (s,a) in the first t episodes and $\hat{p}_{\text{tar}}^t(s,a)$ denotes the corresponding empirical transitions. Accordingly, we select $\pi_h^{t+1}(\cdot) = \arg\max_{a\in\mathcal{A}} W_h^t(\cdot,a)$ to collect online data from $\mathcal{M}_{\text{tar}}$, update $n^t(s,a)$, $\hat{p}_{\text{tar}}^t(s,a)$ and $W_h^t(s,a)$, and stop until:

$$3\sqrt{\mathbb{E}_{\rho,\pi_1^{t+1}}[W_1^t]} + \mathbb{E}_{\rho,\pi_1^{t+1}}[W_1^t] \le \sigma\beta/8,$$

where $\mathbb{E}_{\rho,\pi_1^{t+1}}[W_1^t] = \sum_s \rho(s) W_1^t(s,\pi_1^{t+1}(s))$. This stopping criterion is designed to ensure that sufficient data coverage is achieved when Algorithm 2 stops. Beyond reward-free exploration, our design further guarantees the confidence intervals

$$\text{TV}(p_{\text{tar}}(\cdot\mid s,a), \hat{p}_{\text{tar}}^t(\cdot\mid s,a)) \le \beta/4$$

is constructed for each (s,a), which is verified by Lemma 1. Then, we estimated the shifted region as:

$$\hat{\mathcal{B}} \triangleq \big\{(s,a)\in\mathcal{S}\times\mathcal{A} \mid \text{TV}(\hat{p}_{\text{src}}(\cdot\mid s,a), \hat{p}_{\text{tar}}^t(\cdot\mid s,a)) > \beta/2\big\}, \tag{2}$$

where $\hat{p}_{\text{src}}$ is the empirical transitions in $\mathcal{D}_{\text{src}}$, defined with the visitation count n_{src} in $\mathcal{D}_{\text{src}}$:

$$\hat{p}_{\text{src}}(\cdot\mid s,a) \triangleq \frac{n_{\text{src}}(s,a,\cdot)}{n_{\text{src}}(s,a)},\ \forall(s,a)\in\mathcal{S}\times\mathcal{A}.$$

We show that by executing Algorithm 2, the shifted region $\mathcal{B}$ can be identified with high probability within a sample size from $\mathcal{M}_{\text{tar}}$ that is independent of ε, as formally stated in Lemma 1. The proof of Lemma 1 can be found in the full version Qu et al. (2024).

Lemma 1 (Sample-efficient shift identification)**.** *When Assumption 1 holds and $\delta\in(0,1)$, suppose the shift between $\mathcal{M}_{\text{tar}}$ and $\mathcal{M}_{\text{src}}$ is β-separable, and $\mathcal{D}_{\text{src}}$ contains at least $\widetilde{\Omega}(S/\beta^2)$ samples for $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$. With probability at least $1-\delta/2$, applying Algorithm 2 until $\widetilde{O}(H^2S^2A/(\sigma\beta)^2)$ samples are collected from $\mathcal{M}_{\text{tar}}$, the estimated empirical transition $\hat{p}_{\text{tar}}^t$ satisfies*

$$\text{TV}(p_{\text{tar}}(\cdot\mid s,a), \hat{p}_{\text{tar}}^t(\cdot\mid s,a)) \le \beta/4,\ \forall(s,a)\in\mathcal{S}\times\mathcal{A},$$

and the estimated shifted region $\hat{\mathcal{B}} = \mathcal{B}$.

The confidence interval for transitions with finite-sample guarantees in Lemma 1 is estabilished by extending reward-free exploration to accomodate more general reward functions $r : [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}\mapsto[0,1]$ in the analysis.

Step 2: Hybrid UCB value iteration (Algorithm 3). Once we have the estimated shifted region $\hat{\mathcal{B}}$, it is intuitive for the agent to focus more on exploring the estimated shifted region $\hat{\mathcal{B}}$. To achieve this, we introduce Algorithm 3 that incorporates the additional source dataset $\mathcal{D}_{\text{src}}$ in the design of the exploration bonus.

Algorithm 3 Hybrid UCB value iteration

Require: Shift level δ, desired optimality gap ε, estimated shifted region $\hat{\mathcal{B}}$, source dataset $\mathcal{D}_{\text{src}}$

for $t = 0, 1, 2, \cdots$ **do**
  for $h = H, \cdots, 1$ **do**
    Update upper confidence bounds $\overline{Q}_h^t$, G_h^t using Eqs. (3a) and (4)
    Update $\pi_h^{t+1}(\cdot) = \arg\max_{a\in\mathcal{A}} \overline{Q}_h^t(\cdot, a)$
  end for
  Break if $\mathbb{E}_{\rho,\pi_1^{t+1}}[G_1^t] \le \varepsilon$
  // ε-optimality is achieved
  Rollout π^{t+1} and observe new online samples
  for $(s,a) \in \hat{\mathcal{B}}$ **do**
    Update visitation counts $n^t(s,a)$, $n^t(s,a,s')$ and empirical transitions $\hat{p}_{\text{tar}}^t(\cdot \mid s,a)$
    // Only update n^t and $\hat{p}_{\text{tar}}^t$ inside $\hat{\mathcal{B}}$
  end for
end for
return $\pi^{\text{final}} = \pi^{t+1}$

This algorithm is inspired by BPI-UCBVI in Ménard et al. (2021); however, in our problem, we carefully design the exploration bonus to leverage the additional offline dataset $\mathcal{D}_{\text{src}}$ while controlling potential bias that it introduces. To effectively use $\mathcal{D}_{\text{src}}$ while avoiding potential bias, we define the upper confidence bounds of the optimal Q-functions and value functions for the estimated shifted region $\hat{\mathcal{B}}$ and its complement $\mathcal{S}\times\mathcal{A}\,/\,\hat{\mathcal{B}}$, respectively:

$$\overline{Q}_h^t(s,a) \triangleq \min\Bigg(H, r(s,a) + \frac{14H^2 g_1(\tilde{n}^t(s,a),\delta)}{\tilde{n}^t(s,a)} + 3\sqrt{\mathrm{Var}_{\tilde{p}^t}(\overline{V}_{h+1}^t)(s,a)\frac{g_2(\tilde{n}^t(s,a),\delta)}{\tilde{n}^t(s,a)}} + \frac{1}{H}\tilde{p}^t(\overline{V}_{h+1}^t - \underline{V}_{h+1}^t)(s,a) + \tilde{p}^t\overline{V}_{h+1}^t(s,a)\Bigg), \tag{3a}$$

$$\overline{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a),\ \overline{V}_{H+1}^t(s) \triangleq 0, \tag{3b}$$

where $g_2(n,\delta) \triangleq \log(6SAH/\delta) + \log(8e(n+1))$, $\underline{V}_{h+1}^t$ is the lower bounds of the optimal value functions defined similarly, which can be found in the full version Qu et al. (2024). and $\mathrm{Var}_{\tilde{p}^t}(\cdot)$ denotes the empirical variance under $\tilde{p}^t$. Here, for each $(s,a)\in\mathcal{S}\times\mathcal{A}$,

- If $(s,a) \in \hat{\mathcal{B}}$, then we choose $\tilde{n}^t(s,a) = n^t(s,a)$ and $\tilde{p}^t(\cdot \mid s,a) = \hat{p}_{\text{tar}}^t(\cdot \mid s,a)$;
- If $(s,a) \notin \hat{\mathcal{B}}$, then we choose $\tilde{n}^t(s,a) = n_{\text{src}}(s,a)$ and $\tilde{p}^t(\cdot \mid s,a) = \hat{p}_{\text{src}}(\cdot \mid s,a)$.

Aiming to achieve optimality in $\mathcal{M}_{\text{tar}}$, we choose $\pi_h^{t+1}(\cdot) = \arg\max_{a\in\mathcal{A}} \overline{Q}_h^t(\cdot,a)$ to collect samples from $\mathcal{M}_{\text{tar}}$ in Algorithm 3. Accordingly, we define the following function $G_h^t(s,a)$ to serve as an upper bound on the optimality gap $V_h^{p_{\text{tar}},\star} - V_h^{p_{\text{tar}},\pi^{t+1}}$ (with $G_{H+1}(s,a) = 0$):

$$G_h^t(s,a) \triangleq \min\Bigg(H, \frac{35H^2 g_1(\tilde{n}^t(s,a),\delta)}{\tilde{n}^t(s,a)} + 6\sqrt{\mathrm{Var}_{\tilde{p}^t}(\overline{V}_{h+1}^t)(s,a)\frac{g_2(\tilde{n}^t(s,a),\delta)}{\tilde{n}^t(s,a)}} + (1+\frac{3}{H})\tilde{p}^t\pi_{h+1}^{t+1}G_{h+1}^t(s,a)\Bigg), \tag{4}$$

Algorithm 3 stops when $\mathbb{E}_{\rho,\pi_1^{t+1}}[G_1^t] \le \varepsilon$, indicating that ε-optimality is achieved in the target domain $\mathcal{M}_{\text{tar}}$. This procedure requires at most $\widetilde{O}(H^3|\mathcal{B}|/\varepsilon^2)$ samples from $\mathcal{M}_{\text{tar}}$, as detailed in the final results in the next section.

4.3 Theoretical guarantees: sample complexity

In this subsection, we discuss the total sample complexity of Algorithm 1, highlighting its sample efficiency gains compared to the state-of-the-art pure online RL sample complexity and connections to practical transfer algorithms. The proof of Theorem 2 can be found in the full version Qu et al. (2024).

Theorem 2 (Problem-dependent sample complexity)**.** *Let Assumption 1 hold, and $\delta \in (0,1)$ and $\varepsilon \in (0,1]$ be given. Suppose the shift between $\mathcal{M}_{\text{tar}}$ and $\mathcal{M}_{\text{src}}$ is β-separable, and $\mathcal{D}_{\text{src}}$ contains at least $\widetilde{\Omega}(H^3/\varepsilon^2 + S/\beta^2)$ samples for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. With probability at least $1-\delta$, the output policy π^{final} of Algorithm 1 satisfies*

$$V_1^{p_{\text{tar}},\star}(\rho) - V_1^{p_{\text{tar}},\pi^{\text{final}}}(\rho) \le \varepsilon, \tag{5}$$

if the total number of online samples collected from $\mathcal{M}_{\text{tar}}$ is

$$\widetilde{O}\left(\min\left(\frac{H^3SA}{\varepsilon^2}, \frac{H^3|\mathcal{B}|}{\varepsilon^2} + \frac{H^2S^2A}{(\sigma\beta)^2}\right)\right). \tag{6}$$

Theorem 2 provides a problem-dependent sample complexity of Algorithm 1 that is at least as good as the state-of-the-art $\widetilde{O}(H^3SA/\varepsilon^2)$ in pure online RL (Ménard et al., 2021; Wainwright, 2019), and improves upon it when the dynamics shift degree β is relatively large. Specifically, for a given β:

- When $\beta < \Omega(\sqrt{S/H} \cdot \varepsilon/\sigma)$: it captures the scenarios where the minimal degree of the dynamics shift β is smaller than the order of the desired optimality gap ε, implying the difficulty of distinguishing between the source and the target environments. In this case, the sample complexity of Algorithm 1 becomes $\widetilde{O}(H^3SA/\varepsilon^2)$, which matches the state-of-the-art pure online RL sample complexity, showing that our framework provably avoids negative transfer in terms of sample efficiency.
- When $\beta \geq \Omega(\sqrt{S/H} \cdot \varepsilon/\sigma)$: the comparisons between the sample complexity of Algorithm 1 in Eq. (6) and the state-of-the-art pure online RL is as follows:

$$\widetilde{O}\left(\frac{H^3|\mathcal{B}|}{\varepsilon^2}\right) \quad v.s. \quad \widetilde{O}\left(\frac{H^3SA}{\varepsilon^2}\right),$$

  where $|\mathcal{B}|$ represents the cardinality of the shifted region, a problem-dependent parameter in HTRL that is strictly no larger than SA. It indicates that Algorithm 1 provably achieves better sample efficiency than state-of-the-art pure online RL algorithms in HTRL tasks, as long as the shift does not cover the entire state-action space, as validated in Section 5. In many practical scenarios, such as training cooking agents (Beck et al., 2024) or autonomous driving (Xiong et al., 2016), environmental variations between source and target environments (e.g., different kitchen layouts or obstacle positions) typically affect only a small portion of the state-action space, meaning $|\mathcal{B}| \ll SA$, with a large separable shift. This enables significant sample efficiency gains from reusing the source dataset.

Our results demonstrate that for HTRL tasks with β-separable shift between source and target environments, Algorithm 1 provably avoids harmful information transfer and enhances sample efficiency compared to pure online RL. While Algorithm 1 relies on β, we evaluate Algorithm 1 in broader scenarios where an inaccurate β is used, as discussed in Section 5, demonstrating the robustness of Algorithm 1.

Connections with practical cross-domain transfer algorithms. Practical algorithms for cross-domain transfer RL often involve training a neural network classifier to distinguish between source and target transitions (Eysenbach et al., 2021; Liu et al., 2022; Niu et al., 2023; Wen et al., 2024) and reusing source data accordingly. Our sample complexity results provide theoretical insights for determining the data collection budget in the target domain. They also demonstrate that the estimated transition shift serves as an effective metric for utilizing the source data and can provably improve sample efficiency.

Extensions of Theorem 2: variants of source data. In Theorem 2, we assume abundant samples from the source domain, which is a common assumption since we primarily focus on sample complexity in the target domain $\mathcal{M}_{\text{tar}}$. However, even when the source dataset $\mathcal{D}_{\text{src}}$ is insufficient, similar results hold, which can be verified directly as a corollary of Theorem 2. In particular, we consider the set of state-action pairs where $\mathcal{D}_{\text{src}}$ lacks sufficient samples:

$$\mathcal{C} \triangleq \left\{(s,a) \in \mathcal{S} \times \mathcal{A} \mid n_{\text{src}}(s,a) < \widetilde{\Omega}(H^3/\varepsilon^2 + S/\beta^2)\right\}.$$

By adjusting the input of Algorithm 3 to $\hat{\mathcal{B}} \cup \mathcal{C}$, Algorithm 1 can still achieve the identical optimality with the sample complexity as below:

$$\widetilde{O}\left(\min\left(\frac{H^3SA}{\varepsilon^2}, \frac{H^3|\mathcal{B} \cup \mathcal{C}|}{\varepsilon^2} + \frac{H^2S^2A}{(\sigma\beta)^2}\right)\right).$$

Similarly, when N datasets from N different source MDPs are available, Algorithm 1 can still function by executing Algorithm 2 once to identify the shifts in the target transition relative to each source transition and selecting useful source data accordingly. Let $\mathcal{B}_i$ denote the corresponding shifted region for each source MDP i. Under the conditions of Theorem 2, the required sample complexity in this setting becomes

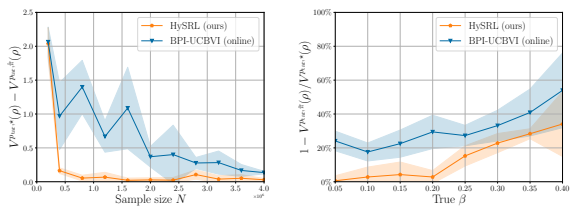
$$\widetilde{O}\left(\min\left(\frac{H^3SA}{\varepsilon^2}, \frac{H^3|\cap_{i\in[N]}\mathcal{B}_i|}{\varepsilon^2} + \frac{H^2S^2A}{(\sigma\beta)^2}\right)\right).$$

5 Experiments

We evaluate our proposed algorithm by comparing it to the state-of-the-art online RL baseline, BPI-UCBVI Ménard et al. (2021), in the GridWorld environment ($S = 16, A = 4, H = 20$).

In the source and the target environments, the agent may fail to take an action and go to a wrong direction. Compared with the source environment, the target environment includes three absorbing states. The source dataset is collected by running Algorithm 2 in the source environment for $T = 1 \times 10^5$ episodes (satisfies the conditions for $\mathcal{D}_{\text{src}}$ in Theorem 2). We implement both algorithms in the benchmark rlberry (Domingues et al., 2021a), similar to Ménard et al. (2021). The results are averaged over 5 random seeds with a 95% confidence interval, presented in Fig. 2. See Appendix A for details on the experiment setup. The code is available at `https://github.com/crqu/hybrid-transfer-rl`.

As shown in Fig. 2a, Algorithm 1 learns the optimal policy for the target environment with fewer interactions (samples) inside the target environment, outperforming the pure online RL baseline BPI-UCBVI. This



(a) (b)

Figure 2: Fig. 2a shows the optimality gap of HySRL (ours) and BPI-UCBVI as the sample size varies. Fig. 2b presents the percentage optimality gap of HySRL (ours) and BPI-UCBVI as the true β varies.

demonstrates that transferring shifted-dynamics data from a source environment can significantly improve sample efficiency. To study whether exact knowledge of β is necessary, we conduct an ablation study with input $\beta = 0.45$, while the true β ranges from 0.05 to 0.4. As shown in Fig. 2b, even with an approximate β that violates Definition 1, the performance degradation of the output policy from Algorithm 1 is minor and still outperforms BPI-UCBVI within finite samples, demonstrating the robustness of our algorithm.

6 Conclusion

This paper introduces Hybrid Transfer RL, a learning framework designed to evaluate the sample efficiency of practical hybrid transfer algorithms. We establish a worst-case lower bound for general HTRL, highlighting the inherent difficulty of achieving sample efficiency gains from shifted-dynamic source data to outperform pure online RL in general. However, we demonstrate that in many practical scenarios, where prior knowledge of the dynamics shift degree is available, transferring shifted-dynamic data can provably reduce the sample complexity in the target environment, providing valuable theoretical insights for practical algorithm design.

Acknowledgment

The work of C. Qu is supported in part by NSFC through 723B1001 and by the Summer Undergraduate Research Fellowships at California Institute of Technology. The work of L. Shi is supported in part by the Resnick Institute and Computing, Data, and Society Postdoctoral Fellowship at California Institute of Technology. K. Panaganti is supported in part by the Resnick Institute and the 'PIMCO Postdoctoral Fellow in Data Science' fellowship at the California Institute of Technology. The work of P. You is supported in part from NSFC through 72431001, 72201007. The work of A. Wierman is supported in part from the NSF through CCF-2326609, CNS-2146814, CPS-2136197, CNS-2106403, NGSDI-2105648, and the Resnick Institute. We extend our sincere appreciation to Prof. Yuejie Chi for her insightful and valuable discussions.

References

Ammar, H. B., Eaton, E., Luna, J. M., and Ruvolo, P. (2015). Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3345–3351. AAAI Press.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pages 8002–8011.

Beck, J., Vuorio, R., Liu, E. Z., Xiong, Z., Zintgraf, L., Finn, C., and Whiteson, S. (2024). A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 3rd edition.

Brunskill, E. and Li, L. (2013). Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 122–131, Arlington, Virginia, USA. AUAI Press.

Chen, W., Mishra, S., and Paternain, S. (2024). Domain adaptation for offline reinforcement learning with limited samples. *arXiv preprint arXiv:2408.12136*.

Chen, X., Hu, J., Jin, C., Li, L., and Wang, L. (2022). Understanding domain randomization for sim-to-real transfer. *arXiv preprint arXiv:2110.03239*.

Chua, K., Lei, Q., and Lee, J. (2023). Provable hierarchy-based meta-reinforcement learning. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10918–10967. PMLR.

Domingues, O. D., Flet-Berliac, Y., Leurent, E., Ménard, P., Shang, X., and Valko, M. (2021a). rlberry - A Reinforcement Learning Library for Research and Education.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021b). Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR.

Dong, K., Wang, Y., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*.

Doshi-Velez, F. and Konidaris, G. D. (2013). Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. *IJCAI : proceedings of the conference*, 2016:1432–1440.

Duan, Y., Jia, Z., and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. (2016). RI^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Duan, Y., Wang, M., and Wainwright, M. J. (2021). Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. (2018). IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1407–1416. PMLR.

Eysenbach, B., Asawa, S., Chaudhari, S., Levine, S., and Salakhutdinov, R. (2021). Off-dynamics reinforcement learning: Training for transfer with domain classifiers. *arXiv preprint arXiv:2006.13916*.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.*, 91(3):325–349.

He, J., Zhou, D., and Gu, Q. (2021). Nearly minimax optimal reinforcement learning for discounted mdps. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22288–22300. Curran Associates, Inc.

He, S., Wang, Y., Han, S., Zou, S., and Miao, F. (2023). A robust and constrained multi-agent

reinforcement learning electric vehicle rebalancing method in amod systems. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5637–5644.

Jafarnia-Jahromi, M., Wei, C.-Y., Jain, R., and Luo, H. (2020). A model-free learning algorithm for infinite-horizon average-reward MDPs with near-optimal regret. *arXiv preprint arXiv:2006.04354*.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. (2024). Rl for latent mdps: regret guarantees and a lower bound. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Lattimore, T. and Hutter, M. (2012). Pac bounds for discounted mdps. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, page 320–334, Berlin, Heidelberg. Springer-Verlag.

Li, G., Shi, L., Chen, Y., and Chi, Y. (2023a). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Information and Inference: A Journal of the IMA*, 12(2):969–1043.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2024a). Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260.

Li, G., Zhan, W., Lee, J. D., Chi, Y., and Chen, Y. (2024b). Reward-agnostic fine-tuning: provable statistical benefits of hybrid reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Li, L., Munos, R., and Szepesvári, C. (2014). On minimax optimal offline policy evaluation. *arXiv preprint arXiv:1409.3653*.

Li, Q., Zhai, Y., Ma, Y., and Levine, S. (2023b). Understanding the complexity gains of single-task RL with a curriculum. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 20412–20451. PMLR.

Liu, J., Zhang, H., and Wang, D. (2022). Dara: Dynamics-aware reward augmentation in offline reinforcement learning. *arXiv preprint arXiv:2203.06662*.

Liu, S. and Su, H. (2021). Regret bounds for discounted mdps. *arXiv preprint arXiv:2002.05138*.

Liu, Z. and Xu, P. (2024). Minimax optimal and computationally efficient algorithms for distributionally robust offline reinforcement learning. *arXiv preprint arXiv:2403.09621*.

Lobel, S. and Parr, R. (2024). An optimal tightness bound for the simulation lemma. *Reinforcement Learning Journal*, 2:785–797.

Luo, F.-M., Jiang, S., Yu, Y., Zhang, Z., and Zhang, Y.-F. (2022). Adapt to environment sudden changes by learning a context sensitive policy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7637–7646.

Ma, X., Liang, Z., Blanchet, J., Liu, M., Xia, L., Zhang, J., Zhao, Q., and Zhou, Z. (2023). Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*.

Menard, P., Domingues, O. D., Shang, X., and Valko, M. (2021). Ucb momentum q-learning: Correcting the bias without forgetting. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7609–7618. PMLR.

Mutti, M. and Tamar, A. (2024). Test-time regret minimization in meta reinforcement learning. *arXiv preprint arXiv:2406.02282*.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. (2021). Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7599–7608. PMLR. ISSN: 2640-3498.

Niu, H., Hu, J., Zhou, G., and Zhan, X. (2024). A comprehensive survey of cross-domain policy transfer for embodied agents. In Larson, K., editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8197–8206. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Niu, H., Ji, T., Liu, B., Zhao, H., Zhu, X., Zheng, J., Huang, P., Zhou, G., Hu, J., and Zhan, X. (2023). H2o+: An improved framework for hybrid offline-and-online rl with dynamics gaps. *arXiv preprint arXiv:2309.12716*.

Niu, H., sharma, s., Qiu, Y., Li, M., Zhou, G., HU, J., and Zhan, X. (2022). When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36599–36612. Curran Associates, Inc.

O' Donoghue, B. (2021). Variational bayesian reinforcement learning with regret bounds. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28208–28221. Curran Associates, Inc.

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 9582–9602.

Panaganti, K., Wierman, A., and Mazumdar, E. (2024). Model-free robust ϕ-divergence reinforcement learning using both offline and online data. *ICML, arXiv preprint arXiv:2405.05468*.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2022). Robust reinforcement learning using offline data. *Advances in Neural Information Processing Systems (NeurIPS)*.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2025). Bridging distributionally robust learning and offline rl: An approach to mitigate distribution shift and partial data coverage. *Learning for Dynamics and Control Conference*.

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3810.

Qu, C., Shi, L., Panaganti, K., You, P., and Wierman, A. (2024). Hybrid transfer reinforcement learning: Provable sample efficiency from shifted-dynamics data. *arXiv preprint arXiv:2411.03810*.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11702–11716. Curran Associates, Inc.

Ren, T., Li, J., Dai, B., Du, S. S., and Sanghavi, S. (2021). Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34.

Serrano, S. A., Martinez-Carranza, J., and Sucar, L. E. (2023). Similarity-based knowledge transfer for cross-domain reinforcement learning. *arXiv preprint arXiv:2312.03764*.

Shi, L. and Chi, Y. (2024). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning (ICML)*.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 79903–79917. Curran Associates, Inc.

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

Sodhani, S., Zhang, A., and Pineau, J. (2021). Multi-task reinforcement learning with context-based representations. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9767–9779. PMLR.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. (2023). Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR.

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR.

Wainwright, M. J. (2019). Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*.

Wang, H., Shi, L., and Chi, Y. (2024). Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. (2017). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wang, Y., Zheng, Z., and Shen, M. (2023). Online pricing with polluted offline data. *SSRN Electronic Journal*.

Wen, X., Bai, C., Xu, K., Yu, X., Zhang, Y., Li, X., and Wang, Z. (2024). Contrastive representation for data filtering in cross-domain offline reinforcement learning. *arXiv preprint arXiv:2405.06192*.

Woo, J., Shi, L., Joshi, G., and Chi, Y. (2024). Federated offline reinforcement learning: Collaborative single-policy coverage suffices. In *International Conference on Machine Learning*, pages 53165–53201. PMLR.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27395–27407. Curran Associates, Inc.

Xiong, X., Wang, J., Zhang, F., and Li, K. (2016). Combining deep reinforcement learning and safety based control for autonomous driving. *arXiv preprint arXiv:1612.00147*.

Xu, T., Yang, Z., Wang, Z., and Liang, Y. (2021). A unified off-policy evaluation approach for general value function. *arXiv preprint arXiv:2107.02711*.

Xu, Z., Panaganti, K., and Kalathil, D. (2023). Improved sample complexity bounds for distributionally robust reinforcement learning. In *Artificial Intelligence and Statistics*.

Yang, K., Yang, L., and Du, S. (2021). Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized Lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561.

Ye, H., Chen, X., Wang, L., and Du, S. S. (2023). On the power of pre-training for generalization in RL: Provable benefits and hardness. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39770–39800. PMLR.

Yin, M., Bai, Y., and Wang, Y.-X. (2020). Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*.

You, H., Yang, T., Zheng, Y., Hao, J., and Taylor, Matthew, E. (2022). Cross-domain adaptive transfer reinforcement learning based on state-action correspondence. In Cussens, J. and Zhang, K., editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2299–2309. PMLR.

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR.

Zhang, G., Feng, L., Wang, Y., Li, M., Xie, H., and Tan, K. C. (2024a). Reinforcement learning with adaptive policy gradient transfer across heterogeneous problems. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3):2213–2227.

Zhang, K., Kakade, S., Basar, T., and Yang, L. (2020a). Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33.

Zhang, R. and Zanette, A. (2023). Policy finetuning in reinforcement learning via design of experiments using offline data. In *Advances in Neural Information Processing Systems*, volume 36, pages 59953–59995. Curran Associates, Inc.

Zhang, Z., Chen, Y., Lee, J. D., and Du, S. S. (2024b). Settling the sample complexity of online reinforcement learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5213–5219. PMLR.

Zhang, Z., Ji, X., and Du, S. (2021). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134

of *Proceedings of Machine Learning Research*, pages 4528–4531. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020b). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33.

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3331–3339. PMLR.

Zhu, Z., Lin, K., Jain, A. K., and Zhou, J. (2023). Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Experiment Setup

We compare our algorithm with the state-of-the-art pure online RL algorithm BPI-UCBVI in Ménard et al. (2021) on GridWorld environment ($S = 16, A = 4, H = 20$). The goal is to navigate in a room to collect rewards. In the source and the target environments, the same structure includes:

- state-action space: the state space is a 4×4 room, and the action space is to go up/down/left/right.
- horizon: each episode has a horizon length 20.
- success probability, the agent may fail in taking an action and go to the wrong direction with uniform probalibities. The success probability is set to be 0.95 in experiment 1.
- reward: $r = 1$ at state $(1, 4)$, $r = 0.1$ at state $(2, 3)$, $r = 0.01$ at state $(3, 2)$, and $r = 1.5$ at state $(3, 4)$. The state $(1, 4)$ is an absorbing state, where the agent cannot escape once steps in and the reward can only be obtained once.
- initial state: the agent starts from state $(3, 2)$ in each episode.

Compared with the source environment, the target environment includes additional "traps" (absorbing states), at states $(2, 2)$, $(2, 4)$ and $(3, 3)$, where the agent cannot escape once steps in. For experiment 1 and 2, the source dataset is collected by running Algorithm 2 in the source environment for $T = 1 \times 10^5$ episodes, which satisfies the condition in Theorem 2. For both algorithms, $\varepsilon = 0.1$, $\delta = 0.1$. We re-scale the exploration bonus in BPI-UCBVI and Algorithm 3 with the same constant 2×10^{-3} to mitigate the effect of the large hidden constant within $\widetilde{O}(\cdot)$ (similarlily for Algorithm 2 with 1×10^{-6}). The optimality gap of a policy in the target environment is evaluated by running the policy for 100 episodes and calculating the average the results.

For experiment 1, we run both algorithms in the target environment for $T = 2 \times 10^5$ episodes to examine the relationship between optimality gaps and the sample size from the target environment. We set $\beta = 0.45$ and $\sigma = 0.25$ for Algorithm 1, satisfying Definition 1 and Assumption 1.

For experiment 2, we vary the success probability of taking an action in the target environment (not accounted in the implementation of Algorithm 1) to examine the effect of maximum unidentified shift degree. The real success probability is set from 0.9 to 0.55 with a step size 0.05. Because β is still set to be 0.45 for Algorithm 1, the maximum unidentified shift degree (real β) ranges from 0.05 to 0.4 with a step size 0.05. For each success probability, we run the algorithms for 5 runs, each run contains $T = 1 \times 10^5$ episodes.