Augmenting Online RL with Offline Data is All You Need: A Unified Hybrid RL Algorithm Design and Analysis

Ruiquan Huang[*1] **Donghao Li**[*2] **Chengshuai Shi**[2] **Cong Shen**[2] **Jing Yang**[2]

[1]Electrical Engineering Dept., The Pennsylvania State University, State College, Pennsylvania, USA
[2]Computer Science Dept., University of Virginia, Charlottesville, Virginia, USA,

Abstract

This paper investigates a hybrid learning framework for reinforcement learning (RL) in which the agent can leverage both an offline dataset and online interactions to learn the optimal policy. We present a unified algorithm and analysis and show that augmenting confidence-based online RL algorithms with the offline dataset outperforms any pure online or offline algorithm alone and achieves state-of-the-art results under two learning metrics, i.e., sub-optimality gap and online learning regret. Specifically, we show that our algorithm achieves a sub-optimality gap $\tilde{O}(\sqrt{1/(N_0/\mathtt{C}(\pi^*|\rho)+N_1)})$, where $\mathtt{C}(\pi^*|\rho)$ is a new concentrability coefficient, N_0 and N_1 are the numbers of offline and online samples, respectively. For regret minimization, we show that it achieves a constant $\tilde{O}(\sqrt{N_1/(N_0/\mathtt{C}(\pi^-|\rho)+N_1)})$ speed-up compared to pure online learning, where $\mathtt{C}(\pi^-|\rho)$ is the concentrability coefficient over all sub-optimal policies. Our results also reveal an interesting separation on the desired coverage properties of the offline dataset for sub-optimality gap minimization and regret minimization. We further validate our theoretical findings in several experiments in special RL models such as linear contextual bandits and Markov decision processes (MDPs).

1 INTRODUCTION

Sequential decision making [Lattimore and Szepesvári, 2020, Sutton and Barto, 1998, Bubeck and Cesa-Bianchi, 2012] is often cast as an online learning problem, where an agent interacts with its environment and dynamically updates its policy based on the actions and feedback. The fundamental challenge lies in the exploration-exploitation tradeoff, requiring the agent to balance exploiting known high-reward actions with exploring potentially beneficial but uncertain alternatives. Although exploration is a must-have for sequential decision making, there are unavoidable costs, e.g., performance degradation, incurred by exploration during the online learning process, which are often undesirable in practical applications.

To overcome the drawbacks, offline policy learning has been studied in both bandits [Li et al., 2022, Wang et al., 2023, Oetomo et al., 2023, Zhang et al., 2019, Brandfonbrener et al., 2021, Nguyen-Tang et al., 2021] and reinforcement learning [Hester et al., 2018, Nair et al., 2018, 2020, Rajeswaran et al., 2017, Lee et al., 2022, Xie et al., 2021b, Song et al., 2022, Wagenmaker and Pacchiano, 2023, Agrawal et al., 2023, Li et al., 2023]. In this setting, the agent attempts to learn an optimal policy based solely on an offline dataset that was collected a priori by a behavior policy, without any online interaction with the environment. This setting has attracted growing interest mainly because in many practical applications such as recommendation systems [Thomas et al., 2017], healthcare [Gottesman et al., 2019], and wireless networking [Yang et al., 2023], logged data is often available from prior tasks while acquiring new data is costly. A critical challenge in offline policy learning, however, is that its performance depends critically on the quality of the dataset.

A natural solution that achieves the benefits of both online and offline settings is *hybrid learning*, where the agent has access to an offline dataset while also having the ability to interact with the environment in an online fashion. A number of works [Hester et al., 2018, Nair et al., 2018, 2020, Rajeswaran et al., 2017, Lee et al., 2022, Song et al., 2022] have empirically demonstrated that offline datasets can help online learning. However, there are only limited studies that theoretically investigate the efficiency of hybrid learning [Xie et al., 2021b, Song et al., 2022, Wagenmaker and Pacchiano, 2023, Agrawal et al., 2023, Li et al.,

[*]Equal contribution.

2023]. It has been shown that in tabular MDPs, hybrid learning can outperform pure offline RL and pure online RL algorithms in terms of the sample complexity required to identify an ϵ-optimal policy [Li et al., 2023]. Similar results have been obtained in linear MDPs [Wagenmaker and Jamieson, 2022] and stochastic K-armed bandits [Agrawal et al., 2023]. The benefit of utilizing offline datasets to reduce the online learning regret has been characterized for RL with general function approximation in Tan and Xu [2024]. A complete literature review can be found in Appendix A.

Despite the theoretical successes of hybrid learning in RL, to the best of our knowledge, there is a lack of a unified understanding regarding its benefits. The main question we aim to answer is:

Can we develop a unified algorithm for hybrid RL and characterize the fundamental impact of offline datasets?

In this work, we provide an affirmative answer to the above question through a generic hybrid RL framework and two lower bounds. In addition, by analyzing the benefits of hybrid RL through a unified analysis for sub-optimality gap and regret, which are two key metrics measuring RL algorithms, our findings also answer the following question affirmatively:

Do we need different offline datasets when we minimize sub-optimality gap and regret in hybrid RL?

We summarize our contributions as follows.

- We first establish a framework based on a commonly adopted notion in decision-making problems, namely the uncertainty level. This framework is used to derive a novel concentrability coefficient $\mathtt{C}(\pi|\rho)$ and to analyze the sub-optimality gap or regret of hybrid RL algorithms. We show that if a confidence-based online RL algorithm is augmented with the offline dataset, the sub-optimality gap scales in the order $\tilde{O}(1/\sqrt{N_0/\mathtt{C}_1 + N_1})$ and the regret scales in the order $\tilde{O}(N_1/\sqrt{N_0/\mathtt{C}_2 + N_1})$, where N_0 is the size of the offline dataset, N_1 is the number of episodes during online learning, and $\mathtt{C}_1$ and $\mathtt{C}_2$ are concentrability coefficients. Compared to the sub-optimality gap $\tilde{O}(1/\sqrt{N_1})$ and regret $\tilde{O}(\sqrt{N_1})$ of pure online learning, our results demonstrate a constant $\sqrt{N_1/(N_0/\mathtt{C} + N_1)}$ speed-up compared to pure online learning, where $\mathtt{C}$ is the concentrability coefficient that depends on the problem. We also specialize our general framework and results to linear contextual bandits and Markov decision processes (MDPs) for a better understanding. A full comparison of our results with existing results in the literature is provided in Table 1.
- Then, we derive lower bounds for both sub-optimality gap and regret minimization problems. Specifically, we show that any hybrid RL algorithm must incur a sub-optimality gap that scales in $\tilde{\Omega}(1/\sqrt{N_0/\mathtt{C}_1 + N_1})$ and regret that scales in $\tilde{\Omega}(N_1/\sqrt{N_0/\mathtt{C}_2 + N_1})$. These results show that initializing with an offline dataset as in the proposed hybrid RL framework is order-wisely optimal in terms of the number of samples and concentrability coefficient.
- Our upper bound reveals the fundamental impact of the behavior policy used to collect the offline dataset on the performance of hybrid learning. In particular, for *sub-optimality gap minimization*, our results show that if the behavior policy has good coverage on the *optimal policy*, the sub-optimality gap of hybrid learning can be very low. On the other hand, for *regret minimization*, as long as the behavior policy provides good coverage on *any sub-optimal policies*, hybrid learning can help reduce the regret. Such separation highlights the fundamental distinction between those two performance metrics, and invites further investigation in the hybrid learning setting.
- Finally, we validate our findings in classic MDP examples such as linear contextual bandits and tabular MDPs. The empirical results verify the theoretical benefit of hybrid learning and the impact of different offline behavior policies, particular the aforementioned separation performance of offline data under sub-optimality gap and regret minimization problems.

2 PROBLEM FORMULATION

Notations. Throughout this paper, we use $\|x\|_V$ to denote $\sqrt{x^\intercal V x}$. The set of all probability distributions over a set $\mathcal{X}$ is represented by $\Delta(\mathcal{X})$. $\mathbb{1}\{\cdot\}$ stands for the indicator function, and $[H] = \{1, 2, \ldots, H\}$ for $H \in \mathbb{N}$.

2.1 PRELIMINARIES

Reinforcement Learning. We consider episodic Markov decision processes in the form of $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P, H, R, q)$, where $\mathcal{X}$ is the state space and $\mathcal{A}$ is the action space, H is the number of time steps in each episode, $P = \{P_h\}_{h=1}^H$ is a collection of transition kernels, and $P_h(x_{h+1}|x_h, a_h)$ denotes the transition probability from the state-action pair (x_h, a_h) at step h to state s_{h+1} in the next step, $r = \{r_h\}_{h=1}^H$ is a collection of reward functions of state-action pairs, where $r_h : \mathcal{X} \times \mathcal{A} \to [0, 1]$, $q \in \Delta(\mathcal{X})$ is the initial state distribution.

A Markov policy π is a set of mappings $\{\pi_h : \mathcal{X} \to \Delta(\mathcal{A})\}_{h=1}^H$. In particular, $\pi_h(a|s)$ denotes the probability of selecting action a in state s at time step h. We denote the set of all Markov policies by Π. For an agent adopting policy π in an MDP $\mathcal{M}$, at each step $h \in [H]$, the agent observes state $x_h \in \mathcal{X}$, and takes an action $a_h \in \mathcal{A}$ according to π, after which the agent receives a random reward $r_t \in [0, 1]$ whose expectation is $r(x_t, a_t)$ and the environment transits

Table 1: Comparison of results on sub-optimality gap (SOG) and regret. All results omit the big-O notation and logarithm terms. $\mathtt{C}_w$ is an all-policy concentrability coefficient (CE). $\mathtt{C}_{\rm off}$ and $\mathtt{C}_{\rm on}$ are inversely related coefficients. $\mathtt{C}_l$ is a single-policy CE. $\mathtt{C}$ is a CE defined in multi-armed bandits. $\mathtt{C}(\pi|\rho)$ is defined in Definition 3.2. In the 4-th row, offline indicates pure offline learning algorithms [Rashidinejad et al., 2021, Xie et al., 2021b] and online indicates pure online algorithms [Lattimore and Szepesvári, 2020, Sutton and Barto, 1998]. In the 5-th row, the orders apply to both lower and upper bounds. Our results match or outperform the SOTA and show clear differences between SOG and regret minimization.

Algorithm	**Sub-optimality Gap**	**Algorithm**	**Regret**
FTPEDAL Wagenmaker and Pacchiano [2023]	$\frac{1}{\sqrt{N_0/\mathtt{C}_w(\cdot\|\rho)+N_1}}$	DISC-GOLF Tan and Xu [2024]	$\sqrt{N_1}\sqrt{\frac{\mathtt{C}_{\rm off}N_1}{N_0}}+\sqrt{\mathtt{C}_{\rm on}N_1}$
RAFT Li et al. [2023]	$\sqrt{\frac{\mathtt{C}_{\rm off}}{N_0+N_1}}+\sqrt{\frac{\mathtt{C}_{\rm on}}{N_1}}$	MIN-UCB Cheung and Lyu [2024]	$\frac{N_1}{\sqrt{N_0/\mathtt{C}+N_1}}$
Offline & Online Uehara et al. [2021]	$\sqrt{\frac{\mathtt{C}_l}{N_0}}$ & $\frac{1}{\sqrt{N_1}}$	Online Jin et al. [2020]	$\sqrt{N_1}$
Ours Theorems 3.1 and 5.1	$\frac{1}{\sqrt{N_0/\mathtt{C}(\pi^*\|\rho)+N_1}}$	Ours Theorems 3.2 and 5.1	$\frac{N_1}{\sqrt{N_0/\mathtt{C}(\pi^{-\varepsilon}\|\rho)+N_1}}$

to the next state x_{h+1} with probability $P_h(x_{h+1}|x_h, a_h)$. The episode ends after H steps.

Let V_P^π be the value function of policy π under the transition model P. Mathematically, $V_{\mathcal{M}}^\pi := \mathbb{E}\left[\sum_{h=1}^H R_h(x_h, a_h)\middle|P, \pi, q\right]$, where the expectation is taken over all random variables including reward R, state x_h and action a_h.

2.2 HYBRID LEARNING

Hybrid learning seeks to combine the advantages of both online and offline learning. Specifically, the learning agent has access to a finite offline dataset $\mathcal{D}_0 \subset (\mathcal{X})^H \times (\mathcal{A})^H \times [0,1]^H$ with size N_0. Each data point $\tau \in \mathcal{D}_0$ has the form $\tau = (x_1, a_1, r_1, \ldots, x_H, a_H, r_H)$, also called a trajectory, and is randomly sampled under a behavior policy ρ from the ground-truth MDP environment $\mathcal{M}^* = (\mathcal{X}, \mathcal{A}, P^*, H, R^*, q^*)$. Then, the agent performs online learning with the knowledge of offline data. Let π_t be the policy chosen by the agent at episode $t \geq 1$. Denote $\tau_t \sim \pi_t$ as the trajectory sampled from π_t. Let $\mathcal{D}_t = \{\tau_t\}_{s=1}^t$ be the data collected in the online learning procedure after episode t. In this paper, we consider two classical learning objectives, as elaborated below.

Sub-optimality Gap Minimization. For this learning objective, the goal of the agent is to learn a policy $\hat{\pi}$ from both the offline dataset $\mathcal{D}_0$ and the online dataset $\mathcal{D}_{N_1}$ such that the sub-optimality gap of $\hat{\pi}$, defined in Equation (1), is minimized.

$$\text{Sub-opt}(\hat{\pi}) = \max_{\pi\in\Pi} V_{\mathcal{M}^*}^\pi - V_{\mathcal{M}^*}^{\hat{\pi}}. \tag{1}$$

We remark that this objective is widely studied in both online and offline RL literature [Li et al., 2022, Uehara et al., 2021, Jin et al., 2021b].

Regret Minimization. For this learning objective, the agent aims to minimize the regret during the online interactions with horizon N_1, as defined below:

$$\text{Regret}(N_1) = \sum_{t=1}^{N_1}\left(\max_{\pi\in\Pi} V_{\mathcal{M}^*}^\pi - V_{\mathcal{M}^*}^{\pi_t}\right). \tag{2}$$

Regret minimization has also been studied intensively [Abbasi-Yadkori et al., 2011, Lattimore and Szepesvári, 2020, Sharma et al., 2020, Silva et al., 2023, Shivaswamy and Joachims, 2012].

3 A UNIFIED HYBRID RL FRAMEWORK

In this section, we first present a unified framework for hybrid RL, and then analyze its performance under certain general assumptions. We would like to emphasize that both the learning framework and the analysis are quite universal and can be applied to various MDP settings.

3.1 A UNIFIED HYBRID RL FRAMEWORK

Oracle Algorithm. The core of the hybrid RL framework relies on an oracle algorithm, denoted as `Alg`. `Alg` takes a dataset $\mathcal{D}$ sampled from an unknown environment $\mathcal{M}^*$ as its input, and is able to output: (i) an estimator that estimates the value function $V_{\mathcal{M}^*}^\pi$ for any policy π, denoted as $\hat{V}_{\mathtt{Alg}}^\pi$; and (ii) an uncertainty function $\hat{\mathtt{U}}_{\mathtt{Alg}}^\pi$ that upper bounds the estimation error in $\hat{V}_{\mathtt{Alg}}^\pi$ with high probability, i.e.,

$$\hat{\mathtt{U}}_{\mathtt{Alg}}^\pi \geq \left|V_{\mathcal{M}^*}^\pi - \hat{V}_{\mathtt{Alg}}^\pi\right|$$

with probability at least $1-\delta$ for $\delta \in (0,1)$.

In the following, we use $(\hat{V}^{\pi}_{\texttt{Alg}}(\mathcal{D}), \hat{\texttt{U}}^{\pi}_{\texttt{Alg}}(\mathcal{D}))$ to denote the output of Alg for a given input $\mathcal{D}$. We sometimes omit $\mathcal{D}$ from the notation when it is clear from the context.

The Unified Hybrid RL Framework. With the pre-selected oracle algorithm Alg, we are ready to present the unified hybrid RL framework.

Specifically, at each online episode $t \in [N_1]$, we maintain an online dataset $\mathcal{D}_{t-1}$, which stores all trajectories collected during the online learning so far. Instead of using $\mathcal{D}_{t-1}$ to find the next online policy, we augment $\mathcal{D}_{t-1}$ with the offline dataset $\mathcal{D}_0$ and feed $\mathcal{D}_0 \cup \mathcal{D}_{t-1}$ to the oracle algorithm Alg. With the output $(\hat{V}^{\pi}_{\texttt{Alg}}, \hat{\texttt{U}}^{\pi}_{\texttt{Alg}})$, we then construct the online policy π_t following the optimism in face of uncertainty principle. I.e., we set π_t to be the policy that maximize the upper confidence bound (UCB) of the expected return defined as $\hat{V}^{\pi}_{\texttt{Alg}} + \hat{\texttt{U}}^{\pi}_{\texttt{Alg}}$. We then collect the new trajectory $\tau_t = (x_{t,1}, a_{t,1}, r_{t,1}, \dots, x_{t,H}, a_{t,H}, r_{t,H})$ and update $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{\tau_t\}$. Note that the **regret** during online learning is exactly the summation of the sub-optimality gaps of $\{\pi_t\}_{t \in [N_1]}$.

For the learning objective of **sub-optimality gap minimization**, the agent will need to output a near-optimal policy at the end of online learning phase. The policy is then obtained by utilizing the well-known pessimism principle. Specifically, the agent feeds $\mathcal{D}_0 \cup \mathcal{D}_{N_1}$ to Alg and obtains $(\hat{V}^{\pi}_{\texttt{Alg}}, \hat{\texttt{U}}^{\pi}_{\texttt{Alg}})$. Then, the lower confidence bound (LCB) of the expected return can be expressed as $\hat{V}^{\pi}_{\texttt{Alg}} - \hat{\texttt{U}}^{\pi}_{\texttt{Alg}}$, and the near-optimal policy is the one that maximizes the LCB. The pseudo-code is presented in Algorithm 1.

Algorithm 1 Hybrid RL Framework

```
1:  Input: Offline dataset D_0, total online steps N_1.
2:  for t = 1, ..., N_1 do
3:      (V̂^π_Alg, Û^π_Alg) ← Alg(D_0 ∪ D_{t-1})
4:      Execute policy π_t = argmax_π V̂^π_Alg + Û^π_Alg.
5:      Collect trajectory τ_t.
6:      Update D_t = D_{t-1} ∪ {τ_t}.
7:  end for
8:  if sub-optimality gap minimization: then
9:      (V̂^π_Alg, Û^π_Alg) ← Alg(D_0 ∪ D_{N_1}).
10:     Output: π̂ = argmax_π V̂^π_Alg − Û^π_Alg.
11: end if
```

We remark that Algorithm 1 enjoys a clean structure where we can utilize the online confidence-based algorithms by simply augmenting with offline data. Such an approach is more practically amenable compared with the much more complicated hybrid RL algorithm design in Li et al. [2023], Wagenmaker and Pacchiano [2023], Tan et al. [2024].

3.2 THEORETICAL ANALYSIS

In this section, we analyze the theoretical performance of the unified hybrid RL framework presented in Algorithm 1. Intuitively, the quality of the offline dataset $\mathcal{D}_0$ is of paramount importance for the hybrid learning performance. To assess the quality of the behavior policy ρ and the offline dataset $\mathcal{D}_0$, we first introduce several key concepts and properties, including concentrability coefficient, and Eluder-type condition.

Definition 3.1 (Uncertainty level)**.** Let $\texttt{Alg}_0$ be the best oracle algorithm that achieves the minimum estimation error in the worst case, i.e.,

$$\texttt{Alg}_0 = \arg\min_{\texttt{Alg}} \max_{\mathcal{M}} \mathbb{E}_{\mathcal{D}_0 \sim (\mathcal{M}, \rho)} \left[\left| V^{\pi}_{\mathcal{M}} - \hat{V}^{\pi}_{\texttt{Alg}} \right| \right].$$

The uncertainty level of a policy π, denoted as $\texttt{U}_{\mathcal{M}^*}(\pi) : \Pi \to \mathbb{R}$, is defined by $\texttt{U}_{\mathcal{M}^*}(\pi) = \mathbb{E}_{\mathcal{D}_0 \sim (\mathcal{M}^*, \rho)} \left[\left| V^{\pi}_{\mathcal{M}^*} - \hat{V}^{\pi}_{\texttt{Alg}_0} \right| \right]$.

The reason that we choose a minimax type definition is that $\mathcal{M}^*$ is unknown, and any learning algorithm should prepare for the worst case. Moreover, $\texttt{U}_{\mathcal{M}^*}(\pi)$ serves as a lower bound of $\mathbb{E}_{\mathcal{D}_0 \sim (\mathcal{M}^*, \rho)}[\hat{\texttt{U}}^{\pi}_{\texttt{Alg}}(\mathcal{D}_0)]$. Thus, $\texttt{U}_{\mathcal{M}^*}(\pi)$ is algorithm-independent and represents the essential hardness of estimating the value $V^{\pi}_{\mathcal{M}^*}$.

Definition 3.2 (Concentrability coefficient)**.** Given a behavior policy ρ, the concentrability coefficient of a target policy π is $\texttt{C}(\pi|\rho) = \left(\texttt{U}_{\mathcal{M}^*}(\pi) / \texttt{U}_{\mathcal{M}^*}(\rho)\right)^2 \in [1, \infty]$.

Intuitively, the definition describes how much more effort is needed to estimate $V^{\pi}_{\mathcal{M}^*}$ compared to estimating $V^{\rho}_{\mathcal{M}^*}$ from an offline dataset of size N_0 sampled under ρ.

To enable efficient learning, it is necessary to impose certain conditions on the oracle algorithm Alg. We adopt an Eluder-type condition, defined as follows. As we will show later, this condition plays a crucial role in controlling exploration and exploitation.

Definition 3.3 (Eluder-type condition)**.** Let N_1 be the total number of episodes during online learning. Fix an error probability δ. Let π_t and $\mathcal{D}_{t-1}$ be the policies and dataset generated by an oracle algorithm Alg at episode t. We say Alg satisfy Eluder-type condition if, with probability at least $1 - \delta$, $\sum_{t=1}^{N_1} \hat{\texttt{U}}^{\pi_t}_{\texttt{Alg}}(\mathcal{D}_{t-1})^2 \le C^2_{\texttt{Alg}}$.

At a high level, eluder-type condition is akin to the pigeonhole principle and the elliptical potential lemma widely used in tabular MDPs [Azar et al., 2017, Ménard et al., 2021] and linear bandits/MDPs [Abbasi-Yadkori et al., 2011, Jin et al., 2020], respectively. Intuitively, C_{alg} thus depends on the complexity of estimating V^{π} for all encountered policies and can be explicitly computed or upper-bounded in certain classes of RL problems, such as tabular

MDPs or linear bandits. The constant exists for most theoretical online reinforcement learning algorithms. As proved in Appendix C, this constant depends on the design of the algorithm and the complexity of the environment. We will show that the eluder-type condition holds for the specific RL algorithms considered in this work (See Appendix C).

We further assume that with probability at least $1-\delta$, $\hat{\mathtt{U}}^{\pi}_{\mathtt{Alg}}(\mathcal{D}_0) \le C_{\mathtt{Alg}} \mathtt{U}_{\mathcal{M}}(\pi)$ holds for any $\mathcal{M}$ and $\mathcal{D}_0$. This is a reasonable assumption, since $\hat{\mathtt{U}}^{\pi}_{\mathtt{Alg}}(\mathcal{D}_0) = O(1)$ and $\max_{\mathcal{M}} \mathtt{U}_{\mathcal{M}}(\pi)$ has a lower bound. In Appendix C, we also show how to find $C_{\mathtt{Alg}}$.

Theorem 3.1. *Let* `Alg` *satisfy the condition Definition 3.3, $\hat{\pi}$ be the output policy of Algorithm 1. Suppose π^* is an optimal policy. Then, with probability at least $1-O(\delta)$, the sub-optimality gap $\hat{\pi}$ is*

$$\textit{Sub-opt}(\hat{\pi}) = \tilde{O}\left(\frac{C_{\mathtt{Alg}}}{\sqrt{N_0/\mathtt{C}(\pi^*|\rho)+N_1}}\right),$$

where N_0 and N_1 are the number of offline and online trajectories, respectively, $\mathtt{C}(\pi^*|\rho)$ is the concentrability coefficient, and $C_{\mathtt{Alg}}$ is defined in Definition 3.3.

Remark 3.1. We elaborate on the performance of the hybrid RL framework with respect to different qualities of the behavior policy ρ and offline data as follows.

- When the behavior policy ρ is an optimal policy, we have $\mathtt{C}(\pi^*|\rho) = 1$. The sub-optimality gap of $\hat{\pi}$ is $\tilde{O}(\sqrt{1/(N_0+N_1)})$. This *strictly improves both pure online and offline learning algorithms* where the sub-optimal gap scales in $\tilde{O}(\sqrt{1/N_1})$ and $\tilde{O}(\sqrt{1/N_0})$, respectively.
- When the behavior policy ρ is extremely bad such that $\mathtt{C}(\pi^*|\rho) = \Omega(N_0)$, Theorem 3.1 states that the sub-optimality gap of $\hat{\pi}$ is $\tilde{O}(\sqrt{1/N_1})$, which *recovers the optimal pure online learning result*.
- When the behavior policy ρ has partial coverage on the optimal policy, we have $\mathtt{C}(\pi^*|\rho) \in (1, N_0)$. Theorem 3.1 suggests that Algorithm 1 is *equivalent to an online algorithm with $N_0/\mathtt{C}(\pi^*|\rho)+N_1$ episodes*, while it only runs N_1 episodes. Essentially, $N_0/\mathtt{C}^{\gamma}(\pi^*|\rho)$ serves as the number of effective episodes from the offline data.

Theorem 3.2. *Let* `Alg` *satisfy the conditions in Definition 3.1 and Definition 3.3. Then, the regret of Algorithm 1 scales as*

$$\textit{Regret}(N_1) = \tilde{O}\left(C_{\mathtt{Alg}}\sqrt{N_1}\sqrt{\frac{N_1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho)+N_1}}\right),$$

where $\mathtt{C}(\pi^{-\varepsilon}|\rho)$ is the maximum concentrability coefficient of the sub-optimal policies whose sub-optimality gap is at least ε, and $\varepsilon = \tilde{O}(1/\sqrt{N_0+N_1})$.

Remark 3.2. The key observation from Theorem 3.2 is that the regret does **not** depend on the concentrability coefficient over **the optimal policy** π^*. Rather, it depends on the concentrability coefficient over **sub-optimal policies**, which is *in contrast to the case in sub-optimality gap minimization problem*. Thus, a behavior policy that achieves the best sub-optimality gap may lead to poor performance for regret minimization. This phenomenon is confirmed by our experimental results (See Section 6).

Specifically, when the behavior policy ρ is an *optimal policy*, and the support of optimal policy does not overlap with sub-optimal policies, i.e. $\mathtt{C}(\pi^*|\rho) = 1$, but $\mathtt{C}(\pi^{-\varepsilon}|\rho) = 0$, Theorem 3.2 suggests that the regret is $\tilde{O}(\sqrt{N_1})$, which *recovers the pure online learning result*. While the result seems surprising, it reflects the essential challenge of regret minimization: exploration-exploitation tradeoff. Because the offline policy $\rho = \pi^*$ encodes little exploration information, the agent still needs to explore sub-optimal policies to ensure there is no better policy. This procedure incurs the same regret as pure online learning. One may ask why we cannot use imitation learning in such case. It is because, we **do not** know if the offline policy is the best or not. Our goal is to develop a universal algorithm that is guaranteed to have sub-linear regret in **any** case and thus imitation learning would fail.

On the other hand, when the behavior policy ρ is exploratory such that $\mathtt{C}(\pi^{-\varepsilon}|\rho) = O(1)$, Theorem 3.2 states that the regret is $\tilde{O}(\sqrt{N_1}\sqrt{N_1/(N_0+N_1)})$, which *significantly improves the pure online learning by a factor of $\sqrt{N_1/(N_0+N_1)}$*.

Finally, in all cases, Theorem 3.2 proves that Algorithm 1 achieves a constant $\Theta(\sqrt{N_1/(N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho)+N_1)})$ speed-up compared with pure online learning.

4 EXAMPLES

In this section, we specialize Algorithm 1 to two classic examples, namely, tabular MDPs and linear contextual bandits, by specifying the corresponding oracle algorithm `Alg` to obtain the estimator of the value function and the uncertainty function.

4.1 TABULAR MDPS

Tabular MDPs assume that the state and action spaces are finite. Provided a dataset $\mathcal{D} = \{\tau_t\}$, where $\tau_t = (x_{t,1}, a_{t,1}, \ldots, x_{t,H}, a_{t,H})$ is a trajectory sampled from $\mathcal{M}^*$, a classic method to estimate the reward and transition kernel is as follows:

$$\begin{cases} \hat{r}_h(x_h, a_h) = \dfrac{\sum_t \mathbb{1}\{(x_{t,h}, a_{t,h}) = (x_h, a_h)\} r_{t,h}}{N_h(x_h, a_h)}, \\ \hat{P}_h(x_{h+1}|x_h, a_h) = \dfrac{N_h(x_{h+1}, x_h, a_h)}{N_h(x_h, a_h)}, \end{cases} \tag{3}$$

where $N_h(x_h, a_h) = \sum_t \mathbb{1}\{(x_{t,h}, a_{t,h}) = (x_h, a_h)\}$ and $N_h(x_{h+1}, x_h, a_h) = \sum_t \mathbb{1}\{(x_{t,h+1}, x_{t,h}, a_{t,h}) =$

$(x_{h+1}, x_h, a_h)\}$. Azar et al. [2017] has shown that the estimated model $\hat{\mathcal{M}} = \{\mathcal{X}, \mathcal{A}, \hat{P}, H, \hat{r}, \hat{q}\}$ satisfies

$$\left|V^{\pi}_{\mathcal{M}^*} - V^{\pi}_{\hat{\mathcal{M}}}\right| \leq \mathbb{E}\left[\sum_{h=1}^{H} \beta/\sqrt{N_h(x_h, a_h)}\middle|\hat{\mathcal{M}}, \pi\right], \quad (4)$$

for some $\beta = \tilde{O}(H)$. Thus, we can use the RHS of Equation (4) as the uncertainty function $\hat{\mathtt{U}}^{\pi}_{\mathtt{Alg}}$ and $V^{\pi}_{\hat{\mathcal{M}}}$ as the estimated value function. More importantly, the uncertainty function satisfies the eluder-type condition. Then, we have the following result.

Corollary 4.1. *For tabular MDPs, under the hybrid RL framework in Algorithm 1, using $\hat{\mathtt{U}}^{\pi}_{\mathtt{Alg}}$ defined in the RHS of Equation (4), the regret scales in*

$$\tilde{O}\left(\sqrt{H^4|\mathcal{X}||\mathcal{A}|N_1}\sqrt{\frac{N_1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right);$$

and the sub-optimality gap is

$$\tilde{O}\left(\sqrt{\frac{H^4|\mathcal{X}||\mathcal{A}|}{N_0/\mathtt{C}(\pi^{*}|\rho) + N_1}}\right).$$

Next, we analyze our concentrability coefficient in tabular MDPs. It is shown [Azar et al., 2017] that using the RHS of Equation (4) $\hat{\mathtt{U}}_{\mathtt{Alg}}$ achieves the optimal order of $|\mathcal{X}||\mathcal{A}|$ in learning regret, in the following, we use $\hat{\mathtt{U}}_{\mathtt{Alg}}$ as a proxy of the uncertainty level $\mathtt{U}_{\mathcal{M}^*}(\pi)$ defined in Definition 3.1.

Then, by defining the occupancy measure at step h as $d^{\pi}_h(x, a) = \mathbb{E}[\mathbb{1}\{x_h = x, a_h = a\}|\mathcal{M}^*, \pi]$, we have

$$\begin{aligned}\sqrt{\mathtt{C}(\pi|\rho)} &\approx \frac{\sum_h \sum_{x_h, a_h} d^{\pi}_h(x_h, a_h)\frac{1}{\sqrt{N_h(x_h, a_h)}}}{\sum_h \sum_{x_h, a_h} d^{\rho}_h(x_h, a_h)\frac{1}{\sqrt{N_h(x_h, a_h)}}} \\ &\leq \max_{h, x_h, a_h} \frac{d^{\pi}_h(x_h, a_h)}{d^{\rho}_h(x_h, a_h)},\end{aligned}$$

where the RHS is widely adopted concentrability coefficient in tabular MDPs [Xie et al., 2021b, Li et al., 2024]. This inequality indicates that the concentrability coefficient defined in definition 3.2 is lower than the existing definition for tabular MDPs. Thus, our upper bounds are tighter than the existing results [Xie et al., 2021b, Tan et al., 2024].

We remark that existing works [Li et al., 2023, Tan and Xu, 2024, Tan et al., 2024] typically show a sub-optimality gap scales in $\sqrt{\mathtt{C}_{\text{off}}/(N_0 + N_1)} + \sqrt{\mathtt{C}_{\text{on}}/N_1}$ and a regret scales in $\sqrt{N_1}\left(\sqrt{\mathtt{C}_{\text{off}} N_1/N_0} + \sqrt{\mathtt{C}_{\text{on}} N_1}\right)$ where $\mathtt{C}_{\text{off}}$ and $\mathtt{C}_{\text{on}}$ are concentrability coefficients for separate state-action spaces (e.g. $\mathtt{C}_{\text{on}} = \max_h \max_{(s_h, a_h)\in G} \frac{d^{\pi}_h(s_h, a_h)}{d^{\rho}_h(s_h, a_h)}$ for some set G). Besides it is hard to find the exact value of $\mathtt{C}_{\text{off}}, \mathtt{C}_{\text{on}}$, these results can only match with ours (otherwise are higher than ours) under a strict condition $\mathtt{C}_{\text{on}} = O(N_1/(N_0+N_1))$ and $\mathtt{C}_{\text{off}} = O(N_0/(N_0 + N_1))$, which is a rare occurrence. Hence, our results are tighter and easier to interpret.

4.2 LINEAR CONTEXTUAL BANDITS

Linear contextual bandits is a special case of MDPs when $H = 1$, and the reward admits a linear structure. While it simplifies the transition kernel, the linearity captures a core structure in many complex MDPs such as linear MDPs [Jin et al., 2020] and low-rank MDPs [Uehara et al., 2021]. Specifically, each state-arm or context-arm pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ is associated with a feature vector $\phi(x, a) \in \mathbb{R}^d$. At episode t, the learning agent observes a context x_t sampled from q^* and then pulls an arm a_t. By doing so, the agent receives a reward $r_t = \phi(x_t, a_t)^{\intercal}\theta^* + \xi_t$, where ξ_t is a random noise and $\theta^* \in \mathbb{R}^d$ is an unknown parameter. Throughout the paper, we assume that $\|\theta^*\|_2 \leq 1$ and $\|\phi(x, a)\|_2 \leq 1$, $\forall (x, a) \in \mathcal{X} \times \mathcal{A}$. We also assume that ξ_t is an independent zero-mean sub-Gaussian noise with parameter 1, i.e, $\mathbb{E}[\exp(\lambda\xi_t)] \leq \exp(\lambda^2/2)$.

Many classic algorithms of linear contextual bandits usually involve estimating the unknown parameter θ^* based on available data $\mathcal{D} := \{(x_t, a_t, r_t)\}_t$ through the linear regression defined as follows [Abbasi-Yadkori et al., 2011, Lattimore and Szepesvári, 2020]:

$$\hat{\theta} = \arg\min_{\theta} \sum_{(a_t, r_t)\in\mathcal{D}} (\phi(x_t, a_t)^{\intercal}\theta - r_t)^2 + \lambda\|\theta\|_2^2, \quad (5)$$

where $\lambda > 0$ is a given parameter. Let $\hat{\Lambda} := \lambda I_d + \sum_{(x_t, a_t)\in\mathcal{D}} \phi(x_t, a_t)\phi(x_t, a_t)^{\intercal}$. Then, the solution to Equation (5) can be expressed as $\hat{\theta} = \hat{\Lambda}^{-1}\sum_{(x_t, a_t, r_t)\in\mathcal{D}} r_t\phi(x_t, a_t)$. Furthermore, by choosing $\lambda = d$, with high probability, the following inequalities hold for any x, a (See Abbasi-Yadkori et al. [2011]):

$$|\phi(x, a)^{\intercal}\hat{\theta} - \phi(x, a)^{\intercal}\theta^*| \leq \beta\|\phi(x, a)\|_{\hat{\Lambda}^{-1}}, \quad (6)$$

where $\beta = \tilde{O}(\sqrt{d})$. Therefore, we can use $\mathbb{E}_{x\sim q^*, a\sim\pi(\cdot|x)}[\phi(x, a)^{\intercal}\hat{\theta}]$ as an estimated value function, and the RHS of Equation (6) as the uncertainty function $\hat{\mathtt{U}}_{\mathtt{Alg}}$.

Linear contextual bandits. If we use $\hat{V}^{\pi} = \mathbb{E}_{x\sim q^*, a\sim\pi(\cdot|x)}[\phi(x, a)^{\intercal}\hat{\theta}]$ as the estimator, and the RHS of Equation (6) as the uncertainty function $\hat{\mathtt{U}}_{\mathtt{Alg}}$, the corresponding algorithm is known as Lin-UCB [Abbasi-Yadkori et al., 2011], which satisfies the Eluder-type condition. Then, we have the following result.

Corollary 4.2. *For linear contextual bandits, under the hybrid RL framework in Algorithm 1, using $\hat{\mathtt{U}}_{\mathtt{Alg}}$ as defined in Equation (6), the regret is*

$$\tilde{O}\left(d\sqrt{N_1}\sqrt{\frac{N_1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right);$$

and the sub-optimality gap is

$$\tilde{O}\left(d\sqrt{\frac{1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right).$$

Since Lin-UCB is shown to be nearly minimax optimal [Chu et al., 2011, He et al., 2022], we can use $\hat{\mathtt{U}}_{\mathtt{Alg}}$ as an approximate of the uncertainty level: $\mathtt{U}_{\mathcal{M}^*}(\pi) \approx \beta\|\mathbb{E}_{x\sim q^*}\mathbb{E}_{a\sim\pi(x)}[\phi(x,a)]\|_{\hat{\Lambda}_0^{-1}}$. Therefore, we have

$$\begin{aligned}\mathtt{C}(\pi|\rho) &\approx \frac{\|\mathbb{E}_{x\sim q^*}\mathbb{E}_{a\sim\pi(x)}[\phi(x,a)]\|^2_{\hat{\Lambda}_0^{-1}}}{\|\mathbb{E}_{x\sim q^*}\mathbb{E}_{a\sim\rho(x)}[\phi(x,a)]\|^2_{\hat{\Lambda}_0^{-1}}} \\ &\le \max_{w:\|w\|=1}\frac{w^\top\mathbb{E}_{x\sim q^*}\mathbb{E}_{a\sim\pi(x)}[\phi(x,a)\phi(x,a)^\intercal]w}{w^\top\mathbb{E}_{x\sim q^*}\mathbb{E}_{a\sim\rho(x)}[\phi(x,a)\phi(x,a)^\intercal]w},\end{aligned}$$

where the RHS is widely adopted as a concentrability coefficient in linear MDPs or low-rank MDPs [Uehara et al., 2021, Tan and Xu, 2024]. This inequality indicates that our result is tighter than existing results, especially for problems with a linear structure.

We note that in multi-armed bandits, our regret can be further simplified to $\tilde{O}(N_1/\sqrt{N_0\min_a\rho(a)+N_1})$, which matches the result in Cheung and Lyu [2024] order-wisely.

5 LOWER BOUNDS

In this section, we provide a lower bound for hybrid RL. The lower bound shows the tightness of Theorem 3.1 and Theorem 3.2.

Theorem 5.1. *There exists an MDP instance such that any hybrid RL algorithm must incur a sub-optimality gap in* $\Omega\left(\frac{1}{\sqrt{N_0/\mathtt{C}(\pi^*|\rho)+N_1}}\right)$, *and regret in* $\Omega\left(\frac{N_1}{\sqrt{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho)+N_1}}\right)$.

Proof sketch. Our proof is built upon a 2-arm linear contextual bandit instance specified in He et al. [2022]. It has been shown that the regret per episode or the sub-optimality gap is determined by the estimation error of parameter θ^*. Then, we show that the estimation error $\|\hat{\theta}-\theta^*\|_2^2 = \mathtt{U}^{\pi^*}(\mathcal{D})^2$ scales in the order of $\Theta(1/\mathbb{E}_{\mathcal{D}}[(\phi(x,a)^\intercal\theta^*_\perp)^2])$, where $\mathcal{D}$ is the available dataset, and $\theta^*_\perp$ is orthogonal to θ^*. By choosing $\mathcal{D}=\mathcal{D}_0$, we have $\mathtt{U}^{\pi^*}(\mathcal{D}_0) = 1/\mathbb{E}_{\mathcal{D}_0}[(\phi(x,a)^\intercal\theta^*_\perp)^2]$. Therefore, $\mathtt{C}(\pi^*|\rho) = N_0/\mathbb{E}_{\mathcal{D}_0}[(\phi(x,a)^\intercal\theta^*_\perp)^2]$. By choosing $\mathcal{D}=\mathcal{D}_0\cup\mathcal{D}_{t-1}$ at each episode t, we conclude that the regret per episode scales in $\Omega(1/\sqrt{N_0\mathtt{C}(\pi^*|\rho)+t})$.

6 EXPERIMENTAL RESULTS

In this section, we evaluate the performances of the proposed algorithms in synthetic environments. Additional experimental results are provided in Appendix E, including evaluations in a contextual linear bandit constructed from the MovieLens dataset [Harper and Konstan, 2015] and a tabular MDP discretized from the Mountain Car environment [Moore, 1990], implemented in Gymnasium [Towers et al., 2024]. All of the experiments are conducted on a server equipped with an AMD EPYC 7543 32-core processor and 256GB memory. No GPUs are used. We believe the experiments are also easy to replicate on common PCs.

Environment. We consider two types of environments as described in Section 4: linear contextual bandits and tabular MDPs.

For the linear contextual bandits, we set $|\mathcal{X}|=20$, $|\mathcal{A}|=100$ and $d=10$. At each episode $\{\phi(x,a)\}_{a\in\mathcal{A}}$ is randomly drawn from the unit sphere $\mathcal{S}^{d-1}$. We set θ^* as a unit vector with the first element being 1. In addition, the reward is given by $r_t = \phi(x_t,a_t)^\intercal\theta^* + \xi_t$, where $\xi_t\in[-1,1]$ is independently sampled from the uniform distribution.

For tabular MDPs, we set $H=3$, $|\mathcal{X}|=5$ and $|\mathcal{A}|=10$. The initial states are uniformly and randomly chosen at each episode. The transition probability $P_h(\cdot|s,a)$ at each step and state-action pair is uniformly sampled from the probability simplex. The reward $r_h(s,a)$ is uniformly generated from $[0,1]$ and is assumed to be known to the agent for simplicity.

Offline Dataset Collection. We adopt the Boltzmann policy [Szepesvári, 2022] as the behavior policy. Under the Boltzmann policy, actions are taken randomly according to $\rho_h(a|x) = \frac{\exp\{kQ_h(x,a)\}}{\sum_{a\in\mathcal{A}}\exp\{kQ_h(x,a)\}}$, where $k\in\mathbb{R}$, and $Q_h(x,a)$ is the optimal Q-value function starting from (x,a) at time step h. Note that a larger k makes ρ closer to the optimal policy, and therefore makes $\mathtt{C}(\pi^*|\rho)$ smaller. In particular, $Q_1(x,a)=r(x,a)$ in linear contextual bandits.

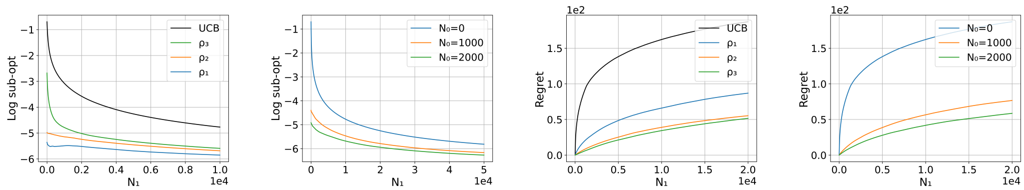
We consider three behavior policies, denoted as ρ_1, ρ_2, and ρ_3, by setting different k of the Boltzmann policy. In Figure 1(a) and (f), we list the values of k used to generate the Boltzmann policy and the concentrability coefficient $\mathtt{C}(\pi^{-\epsilon}|\rho)$ for the two environments. When $k=\infty$, ρ_1 is the optimal policy, so it has the best coverage of the optimal policy, and $\mathtt{C}(\pi^*|\rho)=1$. As k decreases, the policy becomes further away from the optimal policy, thus $\mathtt{C}(\pi^*|\rho)$ increases. In addition, in both environments, we ensure that ρ_2 and ρ_3 are sub-optimal polices and thus, $\mathtt{C}(\pi^{-\varepsilon}|\rho_1) > \mathtt{C}(\pi^{-\varepsilon}|\rho_2) > \mathtt{C}(\pi^{-\varepsilon}|\rho_3)$.

Finally, in Figure 1b and Figure 1g, we fix the offline dataset size N_0 as 2000 and 1000, respectively. In Figure 1c and Figure 1h, we fix the behavior policy as ρ_2.

Results. We present the experiment results in Figure 1. For both environments, we evaluate the sub-optimality gap and the regret with different offline behavior policies and varying numbers of offline trajectories. For each environment, we conduct 100 trials and plot the sample average sub-optimality gap or regret as a function of online time steps N_1. The baseline is the pure UCB algorithm without any offline dataset.

ρ	$\mathtt{C}(\pi^*\|\rho)$	k
ρ_1	1.0	∞
ρ_2	1.7	5
ρ_3	8.7	-10

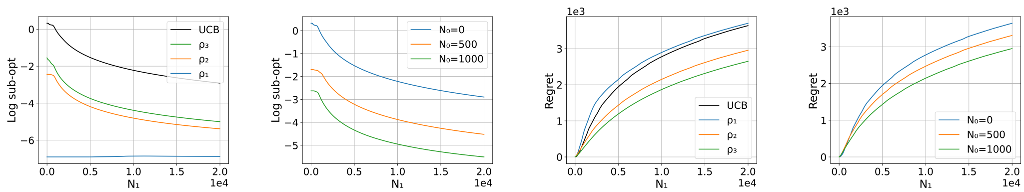
(a) CE in Bandits



(b) SOG v.s. ρ (c) SOG v.s. N_0 (d) Regret v.s. ρ (e) Regret v.s. N_0

ρ	$\mathtt{C}(\pi^*\|\rho)$	k
ρ_1	1.0	∞
ρ_2	2.2	2
ρ_3	4.8	0

(f) CE in MDPs



(g) SOG v.s. ρ (h) SOG v.s. N_0 (i) Regret v.s. ρ (j) Regret v.s. N_0

Figure 1: Experimental results on sub-optimality gap (SOG) and regret for different behavior policies and N_0. Figures (a) and (f) show the concentrability coefficients (CE) of three different behavior policies in linear contextual bandits and MDPs, respectively. Figures (b)-(e) are the results on linear contextual bandits. Figures (g)-(j) are results on tabular MDPs.

Sub-optimality gap. Our experimental results confirm our theoretical findings in Theorem 3.1. Specifically, augmentation using offline data exhibits superior performance than pure online results in both environments. More importantly, the smaller the concentrability coefficient $\mathtt{C}(\pi^*|\rho)$ (Figures 1b and 1g), or the larger the offline dataset size (Figures 1c and 1h), the smaller the sub-optimality gap under the same amount online episodes.

Regret. We then compare the results of regret minimization in both the contextual linear bandit and the tabular MDP. Recall that while $\mathtt{C}(\pi^*|\rho_1) < \mathtt{C}(\pi^*|\rho_2) < \mathtt{C}(\pi^*|\rho_3)$, we have $\mathtt{C}(\pi^{-\varepsilon}|\rho_1) > \mathtt{C}(\pi^{-\varepsilon}|\rho_2) > \mathtt{C}(\pi^{-\varepsilon}|\rho_3)$. As a result, Figures 1d and 1i shows that the regret decreases as the $\mathtt{C}(\pi^{-\varepsilon}|\rho)$ decreases and Figures 1e and 1j shows that larger offline dataset size leads to smaller regret. These experimental results align with our theoretical findings in Theorem 3.2. Finally, we remark that in Figure 1i, the pure online algorithm is slightly better than the hybrid algorithm with offline policy being the optimal policy. This outcome arises because the hybrid algorithm will prioritize exploring actions that are less explored in the offline dataset. When the offline dataset primarily consists of optimal actions, the hybrid algorithm takes sub-optimal actions more frequently than pure online algorithms, leading to slightly higher regret. Nevertheless, the performance is still comparable with the baseline and aligns with Theorem 3.2.

Key insight. The contrasting performances of the same behavior policy under sub-optimality gap minimization and regret minimization problems highlight the need for different offline datasets for these two tasks. Specifically, if the objective is to find a near-optimal policy and the cost of online exploration is negligible, then an offline dataset that focuses on covering the optimal policy is sufficient. However, if the goal is to minimize the regret, it is more effective to collect offline data using various sub-optimal policies rather than the optimal policy.

7 CONCLUSION

In our paper, we developed a general hybrid RL framework to minimize the sub-optimality gap and the online learning regret. The framework achieves performance bounds of $\tilde{O}(1/\sqrt{N_0/\mathtt{C}(\pi^*|\rho)+N_1})$ for the sub-optimality gap and $\tilde{O}(\sqrt{N_1}\sqrt{N_1/(N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho)+N_1)})$ for the regret, where $\mathtt{C}(\pi^*|\rho)$ and $\mathtt{C}(\pi^{-\varepsilon}|\rho)$ are two concentrability coefficients for optimal policy and sub-optimal policies, respectively. Our results demonstrate the benefits of integrating offline data with online interactions. More importantly, the same behavior policy ρ leads to different performances in sub-optimality gap and regret minimization. Our experimental results corroborated our theoretical findings. In addition, we particularized our framework to two specific settings: the contextual linear bandit setting and the tabular MDP setting. We also derived lower bounds for the hybrid RL problem, showing that our approach is nearly optimal. Our results highlight the advantages of leveraging offline datasets for more efficient online learning and provide insights into the selection of offline datasets and policies for different online tasks.

Acknowledgements

The work of R. Huang, D. Li and J. Yang was supported in part by the U.S. National Science Foundation (NSF) under the grants CNS-1956276 and CNS-2114542. The work of C. Shen was supported in part by NSF awards 2029978, 2002902, and 2143559

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

Shubhada Agrawal, Sandeep Juneja, Karthikeyan Shanmugam, and Arun Sai Suggala. Optimal best-arm identification in bandits with access to offline data. *arXiv preprint arXiv:2306.09048*, 2023.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47 (2-3):235–256, May 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR, 2017.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

David Brandfonbrener, William Whitney, Rajesh Ranganath, and Joan Bruna. Offline contextual bandits with overparameterized models. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2021.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.

Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*, 2020.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. *arXiv preprint arXiv:2405.02594*, 2024.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory*, pages 355–366, 2008.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Jiahao He, Jiheng Zhang, and Rachel Q Zhang. A reduction from linear contextual bandits lower bounds to estimations lower bounds. In *International Conference on Machine Learning*, pages 8660–8677. PMLR, 2022.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021b.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1): 4–22, March 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, pages 1702–1712. PMLR, 2022.

Gen Li, Wenhao Zhan, Jason D. Lee, Yuejie Chi, and Yuxin Chen. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=Nd3FennRJZ`.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1): 233–260, 2024.

Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using l_p confidence sets. *Advances in Neural Information Processing Systems*, 35: 20974–20987, 2022.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

Andrew William Moore. Efficient memory-based learning for robot control. Technical report, University of Cambridge, 1990.

Keval Morabia. Collaborative-filtering. `https://github.com/kevalmorabia97/Collaborative-Filtering`, 2019.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6292–6299. IEEE, 2018.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Thanh Nguyen-Tang, Sunil Gupta, A Tuan Nguyen, and Svetha Venkatesh. Offline neural contextual bandits: Pessimism, optimization and generalization. *arXiv preprint arXiv:2111.13807*, 2021.

Bastian Oetomo, R Malinga Perera, Renata Borovica-Gajic, and Benjamin IP Rubinstein. Cutting to the chase with warm-start contextual bandits. *Knowledge and Information Systems*, pages 1–33, 2023.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34: 11702–11716, 2021.

Stephane Ross and J Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. *arXiv preprint arXiv:1203.1007*, 2012.

Nihal Sharma, Soumya Basu, Karthikeyan Shanmugam, and Sanjay Shakkottai. Warm starting bandits with side information from confounded data. *arXiv preprint arXiv:2002.08405*, 2020.

Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pages 1046–1054. PMLR, 2012.

Nicollas Silva, Thiago Silva, Heitor Werneck, Leonardo Rocha, and Adriano Pereira. User cold-start problem in multi-armed bandits: When the first recommendations guide the users experience. *ACM Transactions on Recommender Systems*, 1(1):1–24, 2023.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998.

Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer nature, 2022.

Kevin Tan and Ziping Xu. A natural extension to online algorithms for hybrid rl with limited coverage. *arXiv preprint arXiv:2403.09701*, 2024.

Kevin Tan, Wei Fan, and Yuting Wei. Hybrid reinforcement learning breaks sample size barriers in linear mdps. *arXiv preprint arXiv:2408.04526*, 2024.

Philip Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4740–4745, 2017.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=tyrJsbKAe6`.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35:5968–5981, 2022.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning*, pages 35300–35338. PMLR, 2023.

Andrew Wagenmaker, Julian Katz-Samuels, and Kevin Jamieson. Experimental design for regret minimization in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3088–3096. PMLR, 2021.

Lequn Wang, Akshay Krishnamurthy, and Aleksandrs Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. *arXiv preprint arXiv:2306.07923*, 2023.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35: 12253–12266, 2022.

Kun Yang, Cong Shen, Jing Yang, Shu-ping Yeh, and Jerry Sydir. Offline reinforcement learning for wireless network optimization with mixture datasets. *arXiv preprint arXiv:2311.11423*, 2023.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand N Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301*, 2019.

Augmenting Online RL with Offline Data is All You Need: A Unified Hybrid RL Algorithm Design and Analysis (Supplementary Material)

Ruiquan Huang[*1] **Donghao Li**[*2] **Chengshuai Shi**[2] **Cong Shen**[2] **Jing Yang**[2]

[1]Electrical Engineering Dept., The Pennsylvania State University, State College, Pennsylvania, USA
[2]Computer Science Dept., University of Virginia, Charlottesville, Virginia, USA,

A RELATED WORKS

Offline Reinforcement Learning (RL). It has been known that the performance of offline RL depends critically on the concept of "coverage". Early works [Munos and Szepesvári, 2008, Ross and Bagnell, 2012, Chen and Jiang, 2019, Duan et al., 2020] largely assume full coverage, which suggests that the data generated by offline policy can cover the data distribution generated by any policy. This restriction has been alleviated by recent results [Kumar et al., 2020, Jin et al., 2021b, Rashidinejad et al., 2021, Xie et al., 2021a, Zanette et al., 2021, Uehara and Sun, 2022]. These results provide a better understanding of the pessimism principle under partial coverage assumption, which only requires the offline data to cover the data distribution of the optimal policy or a comparator policy.

Hybrid RL. Several recent works have studied the hybrid policy learning problem for reinforcement learning. Xie et al. [2021b] focus on the sample complexity improvement in episodic tabular MDPs when both offline and online datasets are used. They show the single-policy concentrability coefficient C^* between offline behavior policy and optimal policy is crucial, and the sample complexity required to achieve an ϵ-optimal policy is $\tilde{O}(H^3 S \min(A, C^*)/\epsilon^2)$, where H, S, A are the episodic length, number of states, and number of actions of the MDPs, respectively. While it achieves the best sample complexity in both offline and online RL, it requires the knowledge of the concentrability and the access of the behavior policy. Li et al. [2023] also consider tabular MDPs with a fixed offline dataset and do not assume the knowledge of C^*. They introduce a new single-policy partial concentrability coefficient $C^*(\sigma)$ which generalizes the original C^* by allowing the behavior policy only cover a proportion σ of the state-action pairs. With the new concentrability coefficient and an imitation approach, the show that an ϵ-optimal policy can be obtained if $N_0 + N_1 \geq \tilde{O}(H^3 S C^*(\sigma)/\epsilon^2)$ and $N_1 \geq \tilde{O}(H^3 S \min(H\sigma, 1)/\epsilon^2)$ for a $\sigma \in [0, 1]$, where N_0 and N_1 are the numbers of offline and online samples, respectively. By choosing proper $\sigma \in [0, 1]$, such sample complexity outperforms pure online and offline RL algorithms.

Beyond the tabular setting, Wagenmaker and Pacchiano [2023] consider linear MDPs and also design a new online-to-offline concentrability coefficient C_{o2o}. Compared with the single-policy concentrability coefficient C^*, the coefficient C_{o2o} considers not only the coverage from the offline dataset but also the coverage of the potential online dataset. Under the assumptions that the offline dataset has good coverage and the number of online samples is not large, they show the leading term of the sample complexity can be reduced from $\tilde{O}(1/\epsilon^2)$ to $\tilde{O}(1/\epsilon^{8/5})$.

There are also works on hybrid RL with general function approximation. Song et al. [2022] propose hybrid Q-learning, which achieves regret in regret $\tilde{O}(\max\{C, 1\}\sqrt{dN_1})$ and shows empirical advantages in Atari environments. However, when the offline dataset does not have good coverage, the coverage coefficient C will be much greater than 1 and the regret could be worse than pure online learning, which is also mentioned in Wagenmaker and Pacchiano [2023]. Tan and Xu [2024] study a Global Optimism based on Local Fitting (GOLF) Jin et al. [2021a]-based algorithm for hybrid RL with general Q-function approximation, and show that it can achieve a regret of $\tilde{O}(\sqrt{dN_1^2/N_2} + \sqrt{dN_1})$ in stochastic linear bandits. We also note a concurrent work Tan et al. [2024] studied the sub-optimality gap and regret simultaneously in linear

[*]Equal contribution.
[*]Equal contribution.

MDPs. However, their results relies on all-policy concentrability coefficient and is less general than our results.

Online Bandits. The study of multi-armed bandit problems traces back to the original work by Thompson [1933] for adaptive clinical trials. Many classical algorithms have been proposed, including Thompson Sampling [Chapelle and Li, 2011, Agrawal and Goyal, 2012], and the family of Upper Confidence Bound (UCB) algorithms [Lai and Robbins, 1985, Auer et al., 2002, Audibert et al., 2009, Abbasi-Yadkori et al., 2011, Garivier and Cappé, 2011, Cappé et al., 2013]. These algorithms balance the intrinsic exploration-exploitation tradeoff and achieve the optimal learning regret. The linear bandit model, as a generalization of finite armed bandits with linear reward structure, has also been well studied. Auer et al. [2002] extend the UCB algorithm to stochastic linear bandits problem and achieve regret $\tilde{O}(\sqrt{N_1})$ over time horizon N_1. Dani et al. [2008], Abbasi-Yadkori et al. [2011] further match the lower bound $\Omega(d\sqrt{N_1})$ of Dani et al. [2008] up to logarithmic factors, where d is the feature dimension. The celebrated LinUCB has been proposed and analyzed in Li et al. [2010], which consider linear contextual bandits and achieve regret $\tilde{O}(\sqrt{KdN_1})$ for K-armed disjoint linear model. Other than UCB-type algorithms, Agrawal and Goyal [2013] use Thompson Sampling in linear contextual bandit and achieves regret $\tilde{O}(d^2\sqrt{N_1})$. Additionally, Soare et al. [2014], Jedra and Proutiere [2020] use optimal design for best-arm identification with $\tilde{O}(d/\epsilon^2)$ samples. Yang and Tan [2022] and Wagenmaker et al. [2021] utilize optimal design in linear bandits to achieve minimax optimal results for fixed-budget best-arm identification and regret minimization, respectively.

Offline Bandits. Research on offline bandits has been limited. Rashidinejad et al. [2021] develop a pessimism-based algorithm and match their information-theoretic lower bound on the sub-optimality gap $\Omega\left(\sqrt{SC^*/N_0}\right)$ for finite-arm finite-context bandits, where C^* is the concentrability coefficient C^* representing the coverage of the offline dataset on the optimal policy, S is the number of the contexts and N_0 is the size of the offline dataset. Li et al. [2022] propose a family of pessimistic learning rules for offline linear contextual bandits and pprove that the suboptimality gap scales in $\tilde{O}(\sqrt{dC^\star/N_0})$ for fixed contexts.

Hybrid Bandits. Hybrid policy learning in bandits is related to the *warm-start* bandits problem. Several works [Li et al., 2010, Sharma et al., 2020, Silva et al., 2023, Zhang et al., 2019] investigate utilizing the offline data to improve the online performance under different settings. Shivaswamy and Joachims [2012] study stochastic bandit with finite arms, where the offline dataset is utilized to approximate the confidence bound in UCB algorithms. It shows that the number of pulling a sub-optimal arm a with reward gap Δ_a scales as $\tilde{O}(\max(0, 8/\Delta_a^2 - N_{0,a}))$, where $N_{0,a}$ is the number of times pulling arm a in the offline dataset. Oetomo et al. [2023] augment the existing Thompson Sampling algorithms by initialing the covariance matrix and reward vector in linear contextual bandits using the offline data. They achieve regret $\tilde{O}(\sqrt{N_1 \log((\det(V_{N_0+N_1}))/\det(V_{N_0}))})$, where $V_{N_0+N_1}$ and V_{N_0} are the covariance matrices constructed with both online and offline datasets and offline dataset only, respectively. Since the improvement is logarithmic, the regret advantage is marginal. Agrawal et al. [2023] study the lower bound of the δ-correct best-arm identification in stochastic K-armed bandits given the offline dataset generated by an unknown policy, and further design an algorithm whose instant-dependent sample complexity matches the lower bound. Beyond these, Cheung and Lyu [2024] consider generalized hybrid learning in tabular stochastic bandits where the offline dataset can have different reward distributions than the online environment. When both distributions match, their result reduces to that of Shivaswamy and Joachims [2012], but the work demonstrates the transferability of hybrid learning. Our paper focuses on the hybrid stochastic linear bandit problem, and we develop algorithms that can simultaneously have better sub-optimality and regret than online or offline learning, which has not been studied before.

B MISSING PROOFS OF MAIN RESULTS

In this work, our analysis is built upon high-probability bound. That is, all inequalities and equations hold with probability at least $1 - \delta$, where $\delta > 0$ can be arbitrarily small with a $O(\log(1/\delta))$-factor blow-up in inequalities.

In particular, by taking $\hat{V}^\rho = \frac{1}{|\mathcal{D}_0|}\sum_{\tau\in\mathcal{D}_0}\sum_{h=1}^{H} r_h(\tau)$, we obtain a near-optimal upper bound for $\mathtt{U}(\rho|\mathcal{D}_0)$ through Azuma-Hoeffding inequality. Mathematically, the following inequality holds with probability at least $1 - \delta$.

$$\mathtt{U}(\rho|\mathcal{D}_0) \le \sqrt{\frac{2\log(1/\delta)}{N_0}} = \tilde{O}(1/\sqrt{N_0}).$$

Theorem B.1 (Restatement of Theorem 3.1)**.** *Let* Alg *be a confidence based algorithm and satisfy the conditions in*

Definition 3.3, $\hat{\pi}$ be the output policy of Algorithm 1. Suppose π^* is an optimal policy. Then, the sub-optimality gap $\hat{\pi}$ is

$$\textit{Sub-opt}(\hat{\pi}) = \tilde{O}\left(\frac{C_{\texttt{Alg}}}{\sqrt{N_0/\texttt{C}(\pi^*|\rho) + N_1}}\right),$$

where N_0 is the number of offline samples, N_1 is the number of online samples, $\texttt{C}(\pi^*|\rho)$ is the concentrability coefficient, and $C_{\texttt{Alg}}$ is defined in Definition 3.3.

Proof. We have

$$\begin{aligned}
\text{Sub-opt}(\hat{\pi}) &= V_{\mathcal{M}^*}^{\pi^*} - V_{\mathcal{M}^*}^{\hat{\pi}} \\
&\overset{(a)}{\leq} \hat{V}^{\pi^*} - \hat{V}^{\hat{\pi}} + \hat{\texttt{U}}_{\texttt{Alg}}(\pi^*|\mathcal{D}_0 \cup \mathcal{D}_{N_1}) + \hat{\texttt{U}}_{\texttt{Alg}}(\hat{\pi}|\mathcal{D}_0 \cup \mathcal{D}_{N_1}) \\
&\overset{(b)}{\leq} 2\hat{\texttt{U}}_{\texttt{Alg}}(\pi^*|\mathcal{D}_0 \cup \mathcal{D}_{N_1})
\end{aligned}$$

where (a) follows from the definition of confidence based algorithm, and (b) is due to the optimality of $\hat{\pi}$.

By Definition 3.2, we have

$$\begin{aligned}
\hat{\texttt{U}}_{\texttt{Alg}}(\pi^*|\mathcal{D}_0 \cup \mathcal{D}_{N_1}) &\leq \hat{\texttt{U}}_{\texttt{Alg}}(\pi^*|\mathcal{D}_0) \\
&\leq C_{\texttt{Alg}} \texttt{U}_{\mathcal{M}^*}(\pi^*) \\
&\leq C_{\texttt{Alg}} \sqrt{\texttt{C}(\pi^*|\rho)} \texttt{U}(\rho|\mathcal{D}_0) \leq \tilde{\Theta}\left(C_{\texttt{Alg}} \sqrt{\frac{\texttt{C}(\pi^*|\rho)}{N_0}}\right),
\end{aligned}$$

where the first inequality is due to the fact that $\hat{\texttt{U}}_{\texttt{Alg}}(\pi|\mathcal{D}) \leq \hat{\texttt{U}}_{\texttt{ALg}}(\pi|\mathcal{D}')$ if $\mathcal{D}' \subset \mathcal{D}$.

On the other hand, we have

$$\begin{aligned}
\hat{\texttt{U}}_{\texttt{Alg}}(\pi^*|\mathcal{D}_0 \cup \mathcal{D}_{N_1}) &= \frac{1}{N_1} \sum_{t=1}^{N_1} \hat{\texttt{U}}_{\texttt{Alg}}(\pi^*|\mathcal{D}_0 \cup \mathcal{D}_{N_1}) \\
&\overset{(a)}{\leq} \frac{1}{N_1} \sum_{t=1}^{N_1} \hat{\texttt{U}}_{\texttt{alg}}(\pi^*|\mathcal{D}_{t-1}) \\
&\overset{(b)}{\leq} \frac{1}{N_1} \sum_{t=1}^{N_1} \hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_{t-1}) \\
&\overset{(c)}{\leq} \tilde{O}\left(\frac{1}{N_1}\sqrt{N_1 C_{\texttt{Alg}}^2}\right) \\
&= \tilde{O}\left(C_{\texttt{Alg}}\sqrt{\frac{1}{N_1}}\right),
\end{aligned}$$

where (a) is due to the fact that the uncertainty level decreases as the available dataset increases, (b) follows from the optimality of π_t, and (c) is due to the combination of the Cauchy's inequality and eluder-type condition (Definition 3.3).

Therefore,

$$\begin{aligned}
\text{Sub-opt}(\hat{\pi}) &\leq \tilde{O}\left(C_{\texttt{Alg}} \min\left\{\sqrt{\frac{\texttt{C}(\pi^*|\rho)}{N_0}}, \sqrt{\frac{1}{N_1}}\right\}\right) \\
&= \tilde{O}\left(C_{\texttt{Alg}}\sqrt{\min\left\{\frac{\texttt{C}(\pi^*|\rho)}{N_0}, \frac{1}{N_1}\right\}}\right) \\
&\overset{(a)}{\leq} \tilde{O}\left(C_{\texttt{Alg}}\sqrt{\frac{2}{N_0/\texttt{C}(\pi^*|\rho) + N_1}}\right)
\end{aligned}$$

where (a) is the harmonic mean. $\square$

Theorem B.2 (Restatement of Theorem 3.2)**.** *Let* `Alg` *be a confidence-based algorithm satisfying and Definition 3.3. Then, the regret of Algorithm 1 is*

$$\textit{Regret}(N_1) = \tilde{O}\left(C_{\texttt{Alg}}\sqrt{N_1}\sqrt{\frac{N_1}{N_0/\texttt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right),$$

where $\texttt{C}(\pi^{-\varepsilon}|\rho)$ *is the maximum concentrability coefficient of the sub-optimal policies whose sub-optimality gap is at least* ε*, and* $\varepsilon = \tilde{O}(1/\sqrt{N_0+N_1})$*.*

Proof. Define $\mathrm{Reg}_t = V_{\mathcal{M}^*}^{\pi^*} - V_{\mathcal{M}^*}^{\pi_t}$, where π^* is an optimal policy. We mainly consider t such that $\mathrm{Reg}_t > \varepsilon$, where ε is decided later. By the definition of `Alg`, we have

$$\begin{aligned}
\mathrm{Reg}_t &= V_{\mathcal{M}^*}^{\pi^*} - V_{\mathcal{M}^*}^{\pi_t} \\
&\leq \hat{V}^{\pi^*} + \hat{\texttt{U}}_{\texttt{Alg}}(\pi^*|\mathcal{D}_0 \cup \mathcal{D}_{t-1}) - \hat{V}^{\pi_t} + \hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_0 \cup \mathcal{D}_{t-1}) \\
&\leq 2\hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_0 \cup \mathcal{D}_{t-1})
\end{aligned}$$

By Definition 3.2, we have

$$\begin{aligned}
\hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_0 \cup \mathcal{D}_{t-1}) &\leq \hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_0) \\
&\leq C_{\texttt{Alg}}\texttt{U}(\pi_t) \\
&\leq C_{\texttt{Alg}}\sqrt{\texttt{C}(\pi_t|\rho)}\texttt{U}(\rho|\mathcal{D}_0) \\
&\leq \tilde{\Theta}\left(C_{\texttt{Alg}}\sqrt{\frac{\texttt{C}(\pi^{-\varepsilon}|\rho)}{N_0}}\right),
\end{aligned}$$

where $\texttt{C}(\pi^{-\varepsilon}|\rho) = \max_{\pi: V_{P^*}^{\pi} < V_{P^*}^{\pi^*} - \varepsilon} \texttt{C}(\pi|\rho)$. Therefore, by choosing $\varepsilon = O(1/\sqrt{N_0+N_1})$, we have

$$\begin{aligned}
\mathrm{Reg} &= \sum_{t=1}^{N_1} \mathrm{Reg}_t \\
&\leq \sum_{t=1}^{N_1} \max\left\{\tilde{O}\left(\min\left\{C_{\texttt{ALg}}\sqrt{\frac{\texttt{C}(\pi^{-\varepsilon}|\rho)}{N_0}}, \hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_{t-1})\right\}\right), \varepsilon\right\} \\
&\leq \tilde{O}\left(\sum_{t=1}^{N_1} C_{\texttt{ALg}}\sqrt{\frac{2}{N_0/\texttt{C}(\pi^{-\varepsilon}|\rho) + C_{\texttt{Alg}}^2\hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_{t-1})^{-2}}}\right) + N_1\varepsilon \\
&= \tilde{O}\left(C_{\texttt{ALg}}\sqrt{\frac{\texttt{C}(\pi^{-\varepsilon}|\rho)}{N_0}}\sum_{t=1}^{N_1}\sqrt{1 - \frac{1}{C_{\texttt{Alg}}^{-2}\hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_{t-1})^2 N_0/\texttt{C}(\pi^{-\varepsilon}|\rho) + 1}}\right) + N_1\varepsilon \\
&\overset{(a)}{\leq} \tilde{O}\left(C_{\texttt{ALg}}\sqrt{\frac{\texttt{C}(\pi^{-\varepsilon}|\rho)}{N_0}}\sqrt{N_1}\sqrt{N_1 - \sum_{t=1}^{N_1}\frac{1}{C_{\texttt{Alg}}^{-2}\hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_{t-1})^2 N_0/\texttt{C}(\pi^{-\varepsilon}|\rho) + 1}}\right) + N_1\varepsilon \\
&\overset{(b)}{\leq} \tilde{O}\left(C_{\texttt{ALg}}\sqrt{\frac{\texttt{C}(\pi^{-\varepsilon}|\rho)}{N_0}}\sqrt{N_1}\sqrt{N_1 - \frac{N_1^2}{C_{\texttt{Alg}}^{-2}\sum_{t=1}^{N_1}\hat{\texttt{U}}_{\texttt{Alg}}(\pi_t|\mathcal{D}_{t-1})^2 N_0/\texttt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right) + N_1\varepsilon \\
&\overset{(c)}{\leq} \tilde{O}\left(C_{\texttt{ALg}}\sqrt{\frac{\texttt{C}(\pi^{-\varepsilon}|\rho)}{N_0}}\sqrt{N_1}\sqrt{N_1 - \frac{N_1^2}{N_0/\texttt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right) + N_1\varepsilon \\
&= \tilde{O}\left(C_{\texttt{ALg}}\sqrt{N_1}\sqrt{\frac{N_1}{N_0/\texttt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right),
\end{aligned}$$

where (a) and (b) follow from the Cauchy's inequality, and (c) is due to the eluder-type condition (Definition 3.3).

□

C MISSING PROOFS OF EXAMPLES

C.1 TABULAR MDPS

We first show that the algorithm proposed in Azar et al. [2017] satisfies the eluder-type condition.

Lemma C.1. *The algorithm proposed in Azar et al. [2017] satisfies the eluder-type condition.*

Proof. First, we introduce several additional notations. Given a dataset $\mathcal{D}_t$, Let $\hat{\mathcal{M}}_t$ be the estimated model from Equation (3) and the corresponding counters are $N_{t,h}(x_h, a_h)$. We define another uncertainty function $\tilde{\mathtt{U}}(\pi|\mathcal{D})$ as

$$\tilde{\mathtt{U}}(\pi|\mathcal{D}) = \mathbb{E}\left[\sum_{h=1}^{H} \frac{\beta}{\sqrt{N_h(x_h, a_h)}} \Big| \mathcal{M}^*, \pi\right],$$

where $\beta = \tilde{O}(H)$. According to Section B in Azar et al. [2017], the difference of value functions under the true model $\mathcal{M}^*$ and the estimated model $\hat{\mathcal{M}}$ is also upper bounded by $\tilde{\mathtt{U}}$. Therefore,

$$\begin{aligned}
\sum_{t=1}^{N_1} \tilde{\mathtt{U}}_{\mathtt{Alg}}(\pi_t|\mathcal{D}_t)^2 &\le \beta^2 \sum_{t=1}^{N_1} \mathbb{E}\left[H \sum_{h=1}^{H} \frac{1}{N_{t,h}(x_h, a_h)} \Bigg| \mathcal{M}^*, \pi_t\right] \\
&= \beta^2 H \sum_{h=1}^{H} \sum_{x_h, a_h} \mathbb{E}\left[\sum_{t=1}^{N_1} \frac{\mathbb{1}\{x_{t,h} = x_h, a_{t,h} = a_h\}}{N_{t,h}(x_h, a_h)}\right] \\
&= \beta^2 H \sum_{h=1}^{H} \sum_{x_h, a_h} \mathbb{E}\left[\sum_{t=1}^{N_1} \frac{N_{t+1,h}(x_h, a_h) - N_{t,h}(x_h, a_h)}{N_{t,h}(x_h, a_h)}\right] \\
&\le \beta^2 H^2 |\mathcal{X}||\mathcal{A}| \log N_1 \\
&= \tilde{\Theta}(H^4 |\mathcal{X}||\mathcal{A}|).
\end{aligned}$$

□

While Lemma C.1 shows that the eluder-type condition is satisfied, the result upper bound is loose. To obtain a tighter bound, we directly apply the inequalities in Azar et al. [2017] and prove Corollary 4.1.

Corollary C.1 (Restatement of Corollary 4.1)**.** *For tabular MDPs, under the hybrid RL framework in Algorithm 1, using $\hat{U}_{\mathtt{Alg}}(\pi|\mathcal{D})$ defined in the RHS of Equation* (4)*, the regret scales in*

$$\tilde{O}\left(\sqrt{H^4|\mathcal{X}||\mathcal{A}|N_1}\sqrt{\frac{N_1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho) + N_1}}\right);$$

and the sub-optimality gap is

$$\tilde{O}\left(\sqrt{\frac{H^4|\mathcal{X}||\mathcal{A}|}{N_0/\mathtt{C}(\pi^*|\rho) + N_1}}\right).$$

Proof. Due to Lemma C.1, we have that

$$\sum_{t=1}^{N_1} \hat{\mathtt{U}}_{\mathtt{Alg}}(\pi_t|\mathcal{D}_t) \le \tilde{O}(\sqrt{H^4|\mathcal{X}||\mathcal{A}|N_1}).$$

In addition, we show that $\frac{\hat{\mathtt{U}}_{\mathtt{Alg}}(\pi^*)}{\mathtt{U}_{\mathcal{M}^*}(\pi^*)} \le \sqrt{|\mathcal{X}||\mathcal{A}|}$.

By Theorem 3 in Xiong et al. [2022], there exists an MDP $\mathcal{M}^*$ such that

$$\min_{\mathtt{Alg}} \mathbb{E}_{\mathcal{D}_0}\left[|V_{\mathcal{M}}^{\pi^*} - V_{\mathtt{Alg}}^{\pi^*}\right] | \ge \Omega\left(\sqrt{|\mathcal{X}||\mathcal{A}|}\mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H} \frac{1}{\sqrt{N(s_h, a_h)}}\right]\right)$$

Thus, we have

$$\frac{\hat{\mathtt{U}}_{\mathtt{Alg}}(\pi^*)}{\mathtt{U}_{\mathcal{M}^*}(\pi^*)} \le \tilde{O}\left(\frac{\mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H}\frac{\beta}{\sqrt{N(s_h,a_h)}}\right]}{\sqrt{|\mathcal{X}||\mathcal{A}|}\mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H}\frac{1}{\sqrt{N(s_h,a_h)}}\right]}\right) = O(\frac{H}{\sqrt{|\mathcal{X}||\mathcal{A}|}}) = O(\sqrt{H^4|\mathcal{X}||\mathcal{A}|})$$

Therefore, the regret of Algorithm 1 is

$$\begin{aligned}\text{Regret}(N_1) &= \tilde{O}\left(\min\left\{N_1\sqrt{\frac{\mathtt{C}(\pi^{-\varepsilon}|\rho)}{N_0}}, \sqrt{H^4|\mathcal{X}||\mathcal{A}|N_1}\right\}\right)\\ &\le \tilde{O}\left(\sqrt{H^4|\mathcal{X}||\mathcal{A}|N_1}\sqrt{\frac{1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho)+N_1}}\right).\end{aligned}$$

For the sub-optimality gap, a straightforward modification of the proof in Azar et al. [2017] shows that $\frac{1}{N_1}\sum_{t=1}^{N_1}\hat{\mathtt{U}}_{\mathtt{Alg}}(\pi_t|\mathcal{D}_t) \le \tilde{O}(\sqrt{H^4|\mathcal{X}||\mathcal{A}|/N_1})$. Following the same proof in Theorem B.1 except without using the Cauchy's inequality, we complete the proof.

□

C.2 LINEAR CONTEXTUAL BANDITS

First, we show that Lin-UCB satisfies the eluder-type condition.

Lemma C.2. *Lin-UCB satisfies the eluder-type condition.*

Proof. Note that $\hat{\mathtt{U}}_{\mathtt{Alg}}(\pi|\mathcal{D}) \le \beta\mathbb{E}_{x\sim q^*}\|\mathbb{E}_{a\sim\pi(x)}[\phi(x,a)]\|_{\hat{\Lambda}^{-1}}$ according to Equation (6), where $\beta = \tilde{O}(\sqrt{d})$

Therefore, by the elliptical potential lemma [Carpentier et al., 2020], we have

$$\begin{aligned}\hat{\mathtt{U}}_{\mathtt{Alg}}(\pi_t|\mathcal{D}_t)^2 &\le \beta^2\mathbb{E}\left[\sum_{t=1}^{N_1}\|\phi(x_t,a_t)\|^2_{\hat{\Lambda}_{t-1}^{-1}}\right]\\ &\le \beta^2 d\log N_1\\ &= \tilde{\Theta}(d^2).\end{aligned}$$

□

Combining Lemma C.2 and Theorems 3.1 and 3.2, we obtain the following corollary.

Corollary C.2 (Restatement of Corollary 4.2)**.** *For linear contextual bandits, under the hybrid RL framework in Algorithm 1, using $\hat{\mathtt{U}}_{\mathtt{Alg}}$ as defined in Equation* (6)*, the regret is*

$$\tilde{O}\left(d\sqrt{N_1}\sqrt{\frac{N_1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho)+N_1}}\right);$$

and the sub-optimality gap is

$$\tilde{O}\left(d\sqrt{\frac{1}{N_0/\mathtt{C}(\pi^{-\varepsilon}|\rho)+N_1}}\right).$$

D MISSING PROOFS IN SECTION 5

In this section, we provide the full analysis for lower bounds.

D.1 HARD INSTANCE

First, we introduce the notation of truncated Gaussian distributions. If a Gaussian random variable $X \sim N(0, I_d)$ is truncated to $\{x : \|x\|_2 \leq r\}$, then we denote the truncated Gaussian distribution of X as $N(0, I_d|r)$.

In the lower bound analysis, we follow the setting in He et al. [2022], as specified below.

Arms and dimension: There are 2 arms: $\{1, 2\}$, and the feature dimension is 2.

Feature vectors and the context distribution: The feature vector of the second arm is always 0. For the feature vector of the first arm, let the distribution of the context x satisfy that each $\phi(x, 1) \in \mathbb{R}^2$ is sampled from a truncated normal $N(0, I_2|1)$.

Model parameter: The model parameter $\theta^* \in \mathbb{R}^2$ is sampled uniformly from a sphere $\mathbb{S}_r = \{x \in \mathbb{R}^2 : \|x\| = r\}$, where $r \in [0, 1/\sqrt{d}]$. The constraint on r is due to the boundedness assumption that $\|\theta^*\| \leq 1$.

Additional notations: Recall that $\mathcal{D}_t = \{x_\tau, a_\tau, r_\tau\}_{\tau<t}$ is the online dataset. Here r_t is sampled from a sub-Gaussian distribution with mean $\phi(x_t, a_t)^\intercal\theta^*$ and variance 1. We further denote that $\phi(x_t, a) = x_{t,a}$. We re-parameterize θ^* by its angle, i.e. $\theta^* = r(\cos\gamma^*, \sin\gamma^*)^\intercal$, and γ^* is sampled uniformly from the interval $[0, 2\pi)$. We further denote $e_1 = (1, 0)$ and $e_2 = (0, 1)$, which form the canonical basis of $\mathbb{R}^2$.

With the aforementioned setting, we present the generic regret lower bound modified from Proposition 3.5 in He et al. [2022].

Theorem D.1. *For the hard instance described in Appendix D.1, if available dataset is $\mathcal{D}$, the sub-optimality gap is lower bounded by*

$$\Omega\left(\inf_{\theta\in\mathcal{F}(\mathcal{D})} \frac{1}{r}\mathbb{E}_v\left[\|\theta^* - \theta\|^2\right]\right);$$

if $\mathcal{D}_0 \cup \mathcal{D}_t$ is the available dataset at episode t, regret can be lower bounded by

$$\Omega\left(\sum_{t\in[T]} \inf_{\theta_t\in\mathcal{F}(\mathcal{D}_0\cup\mathcal{D}_t)} \frac{1}{r}\mathbb{E}_v\left[\|\theta^* - \theta_t\|^2\right]\right).$$

D.2 PROOF OF LOWER BOUND

Equipped with Theorem D.1, we are able to lower bound the regret by the estimation error.

Proof Outline: **Step 1** is to decompose the estimation error to the expectations a random variable Z which capture the covariance of the estimator. **Step 2** upper bounds $\mathbb{E}[Z]$. **Step 3** combines the previous steps to prove the results.

Step 1: Decompose the Estimation Error.

Lemma D.1 (Fingerprinting Lemma)**.** *Define random variables Z as follows.*

$$Z = (\hat{\theta} - \theta^*)^\intercal(-e_1\sin\gamma^* + e_2\cos\gamma^*)(-e_1\sin\gamma^* + e_2\cos\gamma^*)^\intercal\bar{V}(\bar{\theta} - \theta^*),$$

where $\bar{V} = \Lambda_0 + \sum_{\tau<t}\phi(x_\tau, 1)\phi(x_\tau, 1)^\intercal$, and $\bar{\theta} = \bar{V}^\dagger\left(\sum_{(x,a,r)\in\mathcal{D}_0}\phi(x, a)r + \sum_{\tau<t}\phi(x_\tau, 1)r_\tau\mathbb{1}\{a_\tau = 1\}\right)$, and recall that r_τ is sampled from $\mathcal{N}(\phi(x_\tau, a_\tau)^\intercal\theta^*, 1)$.

Then, we have

$$\mathbb{E}\left[\left\|\theta^* - \hat{\theta}\right\|^2\right] = 2r^2 - 2r^2\mathbb{E}[Z].$$

Proof. Due to $\|\theta^*\| = \|\hat{\theta}\| = r$, it suffices to analyze the term $\mathbb{E}\left[\hat{\theta}^\intercal \theta^*\right]$. Note that $\theta^* = r(\cos\gamma^*, \sin\gamma^*)^\intercal$. Then, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}^\intercal \theta^*\right] &= \frac{r}{2\pi}\int_0^{2\pi} e_1^\intercal \mathbb{E}[\hat{\theta}|\gamma^*]\cos\gamma^* + e_2^\intercal \mathbb{E}[\hat{\theta}|\gamma^*]\sin\gamma^* d\gamma^* \\
&= \frac{r}{2\pi}\left(e_1^\intercal \mathbb{E}[\hat{\theta}|\gamma^*]\sin\gamma^* - e_2^\intercal \mathbb{E}[\hat{\theta}|\gamma^*]\cos\gamma^*\right)\Big|_{\gamma^*=0}^{\gamma^*=2\pi} \\
&\quad - \frac{r}{2\pi}\int_0^{2\pi} e_1^\intercal \frac{\partial}{\partial\gamma^*}\mathbb{E}[\hat{\theta}|\gamma^*]\sin\gamma^* + e_2^\intercal \frac{\partial}{\partial\gamma^*}\mathbb{E}[\hat{\theta}|\gamma^*]\cos\gamma^* d\gamma^* \\
&= r\mathbb{E}_{\gamma^*}\left[(-e_1\sin\gamma^* + e_2\cos\gamma^*)^\intercal \frac{\partial}{\partial\gamma^*}\mathbb{E}[\hat{\theta}|\gamma^*]\right].
\end{aligned}$$

For the derivative, it is worth noting that $\mathbb{E}[\hat{\theta}|\gamma^*] = \mathbb{E}\left[\mathbb{E}\left[\hat{\theta}|\mathcal{D}_0 \cup \mathcal{D}_t\right]|\gamma^*\right]$. We have

$$\begin{aligned}
\frac{\partial}{\partial\gamma^*}\mathbb{E}[\hat{\theta}|\gamma^*] &= \int_{\mathcal{D}_0\cup\mathcal{D}_t} \mathbb{E}\left[\hat{\theta}|\mathcal{D}_0 \cup \mathcal{D}_t\right] \frac{1}{(2\pi)^{(t-1)/2}} \frac{\partial}{\partial\gamma^*}\exp\left(-\frac{1}{2}\sum_{(x,a,r)\in\mathcal{D}_0\cup\mathcal{D}_t}(r - \phi(x,a)^\intercal\theta^*)^2\right) \\
&= r\mathbb{E}\left[\mathbb{E}\left[\hat{\theta}|\mathcal{D}_0 \cup \mathcal{D}_t\right](-e_1\sin\gamma^* + e_2\cos\gamma^*)^\intercal \sum_{(x,a,r)\in\mathcal{D}_0\cup\mathcal{D}_t}\phi(x,a)(r - \phi(x,a)^\intercal\theta^*)\Big|\theta^*\right] \\
&= r\mathbb{E}\left[\hat{\theta}(-e_1\sin\gamma^* + e_2\cos\gamma^*)^\intercal \bar{V}(\bar{\theta} - \theta^*)\big|\theta^*\right].
\end{aligned}$$

Combining with the fact that $\mathbb{E}[\bar{V}(\bar{\theta} - \theta^*)|\theta^*, \bar{V}] = 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}^\intercal \theta^*\right] &= r^2\mathbb{E}\left[(-e_1\sin\gamma^* + e_2\cos\gamma^*)^\intercal\left(\hat{\theta} - \theta^*\right)(-e_1\sin\gamma^* + e_2\cos\gamma^*)^\intercal \bar{V}(\bar{\theta} - \theta^*)\right] \\
&= r^2\mathbb{E}[Z].
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\left\|\theta^* - \hat{\theta}\right\|^2\right] = 2r^2 - 2\mathbb{E}\left[\hat{\theta}^\intercal\theta^*\right] = 2r^2 - 2r^2\mathbb{E}[Z],$$

which completes the proof. ◻

Step 2: Upper Bound Each $\mathbb{E}[Z]$.

Lemma D.2. *Under the same setting as in Lemma D.1, we have*

$$\mathbb{E}[Z] \le \sqrt{(t - 1 + N_0\mathbb{E}_{x\sim q^*, a\sim\rho(\cdot|x)}[(\phi(x,a)^\intercal\theta_\perp^*)^2])\mathbb{E}\left[\|\hat{\theta} - \theta^*\|^2\right]}. \tag{7}$$

Proof. Recall that

$$Z_i = (\hat{\theta} - \theta^*)^\intercal(-e_1\sin\gamma^* + e_2\cos\gamma^*)(-e_1\sin\gamma^* + e_2\cos\gamma^*)^\intercal \sum_{(x,a,r)\in\mathcal{D}_0\mathcal{D}_t}\phi(x,a)(r - \phi(x,a)^\intercal\theta^*).$$

By the Cauchy's inequality, we have

$$\begin{aligned}
\mathbb{E}[Z]^2 &\le \mathbb{E}\left[\|\hat{\theta}-\theta^*\|^2\right]\mathbb{E}\left[\left(\big(-e_1\sin\gamma^*+e_2\cos\gamma^*\big)^\intercal \sum_{(x,a,r)\in\mathcal{D}_0\cup\mathcal{D}_t}\phi(x,a)(r-\phi(x,a)^\intercal\theta^*)\right)^2\right]\\
&= \mathbb{E}\left[\|\hat{\theta}-\theta^*\|^2\right]\mathbb{E}\left[\sum_{(x,a,r)\in\mathcal{D}_0\cup\mathcal{D}_t}\big(\big(-e_1\sin\gamma^*+e_2\cos\gamma^*\big)^\intercal\phi(x,a)\big)^2\right]\\
&\le (t-1+N_0\mathbb{E}_{x\sim q^*,a\sim\rho(\cdot|x)}[(\phi(x,a)^\intercal\theta^*_\perp)^2])\mathbb{E}\left[\|\hat{\theta}-\theta^*\|^2\right].
\end{aligned}$$

□

Step 3: Lower Bound the Total Regret.

Theorem D.2 (Restatement of Theorem 5.1)**.** *Under the instance described in Appendix D.1, any hybrid RL algorithm must incur a sub-optimality gap in* $\Omega\left(\frac{1}{\sqrt{N_0/\mathsf{C}(\pi^*|\rho)+N_1}}\right)$, *and regret in* $\Omega\left(\frac{N_1}{\sqrt{N_0/\mathsf{C}(\pi^{-\varepsilon}|\rho)+N_1}}\right)$.

Proof. Combine Step 1 (Lemma D.1) and Step 2 (Lemma D.2), we have

$$\begin{aligned}
2r^2 &= \mathbb{E}\left[\|\hat{\theta}-\theta^*\|^2\right] + 2r^2\mathbb{E}\left[Z\right]\\
&\le \mathbb{E}\left[\|\hat{\theta}-\theta^*\|^2\right] + 2r^2\sqrt{(N_0\alpha+t-1)\mathbb{E}\left[\|\hat{\theta}-\theta^*\|^2\right]},
\end{aligned}$$

where $\alpha = \mathbb{E}_{x\sim q^*,a\sim\rho(\cdot|x)}[(\phi(x,a)^\intercal\theta^*_\perp)^2]$.

Therefore,

$$r^2 \ge \mathbb{E}\left[\|\hat{\theta}-\theta^*\|^2\right] \ge \frac{r^2}{4r^2(N_0\alpha+t-1)+4}.$$

Substituting the above result into the generic lower bound in Theorem D.1, and selecting $r = 1/\sqrt{N_0\alpha+N_1}$, we conclude that

$$\begin{aligned}
\text{Regret}(T) &\ge \Theta\left(\sum_{t\in[T]}\frac{r}{4r^2(N_0\alpha+t-1)+4}\right)\\
&= \Theta\left(\frac{1}{4r}\log\left(1+\frac{r^2N_1}{1+r^2N_0\alpha}\right)\right)\\
&= \Theta\left(\frac{rN_1}{1+r^2N_0\alpha}\right)\\
&= \Theta\left(\frac{N_1}{\sqrt{N_0\alpha+N_1}}\right).
\end{aligned}$$

To establish the relationship between α and the concentrability coefficient, we repeatedly apply Lemma D.1 and Lemma D.2 on dataset $\mathcal{D}_0$. Then, we obtain

$$r^2 \ge \mathbb{E}[\|\hat{\theta}-\theta^*\|^2] \ge \frac{r^2}{4r^2N_0\alpha+4}.$$

By choosing $r = 1/\sqrt{N_0\alpha}$, we have $\frac{1}{r}\mathbb{E}[\|\hat{\theta} - \theta^*\|^2] = \Theta(1/\sqrt{N_0\alpha})$. Thus,

$$\begin{aligned}\alpha^{-1} &= \Theta\left(\frac{\frac{1}{r}\mathbb{E}[\|\hat{\theta} - \theta^*\|^2]}{1/\sqrt{N_0}}\right)^2 \\ &= \Theta\left(\frac{\mathtt{U}(\pi^*|\mathcal{D}_0)}{\mathtt{U}(\rho|\mathcal{D}_0)}\right)^2 \\ &= \mathtt{C}(\pi^*|\rho).\end{aligned}$$

The proof is completed by noting that $\mathtt{U}(\pi^*|\mathcal{D}_0) = \Theta(\mathtt{U}(\pi^{-\varepsilon}|\mathcal{D}_0))$ for any $\varepsilon = O(1/\sqrt{N_0 + N_1})$ in the setting described in Appendix D.1. $\square$

E ADDITIONAL EXPERIMENTS

In this section, we provide additional experimental results evaluating the performance of algorithms instantiated within our proposed framework in more realistic environments. Specifically, we consider a contextual linear bandit constructed from the MovieLens dataset [Harper and Konstan, 2015] and a tabular MDP discretized from the Mountain Car environment [Moore, 1990] implemented in Gymnasium [Towers et al., 2024].

E.1 EXPERIMENTAL RESULTS WITH MOVIELENS DATASET

Environment. We construct our linear contextual bandit environment using the MovieLens-100K dataset [Harper and Konstan, 2015], which provides sparse ratings from 943 users on 1682 movies. Following Bogunovic et al. [2021], we first apply collaborative filtering [Morabia, 2019] to complete the partially observed rating matrix. We then factorize the resulting rating matrix $R = [r_{i,a}] \in \mathbb{R}^{943\times 1682}$ using non-negative matrix factorization with 3 latent factors, yielding $R = XH$. Here, $X \in \mathbb{R}^{943\times 3}$ represents user feature vectors, and $H \in \mathbb{R}^{3\times 1682}$ represents movie feature vectors. In the linear contextual bandit framework, we treat each row in X (i.e., each users feature vector) as the context, denoted by $x_i \in \mathbb{R}^3$ for the i-th user. The contexts are known in the contextual bandits. Meanwhile, we randomly select 20 columns of H to serve as the arms (i.e., 20 movies for chosen), where each arms unknown parameter vector corresponds to a movies feature vector that must be estimated from data, denoted as $\theta_a \in \mathbb{R}^3$. At each decision point, the linear contextual bandit model randomly provides a user context and the agent predicts the expected reward for choosing arm a (i.e., recommending a movie) based on observed context x.

Offline Dataset Collection. Similar to the offline data collection method in the main text, we adopt the Boltzmann policy [Szepesvári, 2022] as our behavior policy. Specifically, the policy chooses an action a according to

$$\rho(a|x) = \frac{\exp\{kr(x,a)\}}{\sum_{b\in\mathcal{A}}\exp\{kr(x,b)\}},$$

where $k \in \mathbb{R}$, $\mathcal{A}$ is the set of all arms (movies), and $r(x, a)$ denotes the true reward function. Under the Boltzman policy, a larger k makes ρ closer to the greedy policy (i.e., always selecting the highest reward arm), whereas a smaller k makes the policy more exploratory.

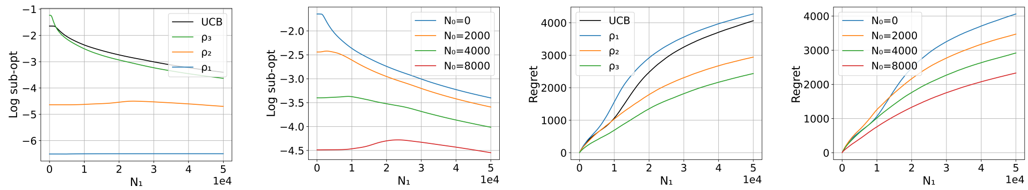
The three different behavior policies, ρ_1, ρ_2, and ρ_3, are constructed in the same way introduced in the main text, each defined by a distinct value of k in the Boltzmann distribution. As k decreases, the policys action deviates from the optimal choice, increasing the concentrability coefficient $\mathtt{C}(\pi^*|\rho)$ and thus degrading coverage of the optimal policy. Figure 2(a) provide the exact k values and the approximated $\mathtt{C}(\pi^*|\rho)$ in the environment.

In Figure 2b and Figure 2d, we fix the offline dataset size N_0 to be 4000 and vary k to illustrate the impact of different levels of coverage. In constrast, in Figure 2c and Figure 2e, we fix the behavior policy to be ρ_2 while varying the offline dataset size. This setup allows us to systematically examine how both the behavior policys level of optimality and the sample size affect the performance of various learning algorithms.

Results. Figure 2 compares sub-optimality gaps and regrets under various offline behavior policies and dataset sizes, using the pure UCB algorithm without offline data as the baseline. Overall, the results mirror the main texts theoretical and empirical findings.

ρ	$\mathtt{C}(\pi^*\|\rho)$	k
ρ_1	1.0	∞
ρ_2	2.43	5
ρ_3	4.05	0

(a) CE in bandits



(b) SOG v.s. ρ (c) SOG v.s. N_0 (d) Regret v.s. ρ (e) Regret v.s. N_0

Figure 2: Experimental results on sub-optimality gap (SOG) and regret for different behavior policies and N_0. Figure (a) shows the concentrability coefficients (CE) of three different behavior policies in MovieLens linear contextual bandits.

For gap minimization, policies with smaller concentrability coefficients $\mathtt{C}(\pi^*|\rho)$ or larger offline samples lead to tighter gaps, reaffirming that offline data focused on covering optimal actions can greatly enhance efficiency.

In regret minimization, the algorithm benefits from the offline dataset. It benefit from the offline data more, if the offline data has diverse offline coverage generated by larger $\mathtt{C}(\pi^*|\rho)$ and larger offline data size, as shown in Figure 2d and Figure 2e. A slight exception occurs when the offline dataset is collected by optimal actions. In this case, the offline data has very big $\mathtt{C}(\pi^{-\epsilon}|\rho)$ and may encourage the algorithm to explore the sub-optimal arms first. Nonetheless, the outcomes are still consistent with theoretical predictions, highlighting the distinct offline data requirements for gap minimization versus regret minimization.

E.2 EXPERIMENTAL RESULTS WITH MOUNTAIN CAR ENVIRONMENT

Environment. The Mountain Car environment[Moore, 1990] is a classic benchmark task in which an underpowered car must drive up a steep slope, featuring a continuous state space over position $[-1.2, 0.6]$ and velocity $[-0.07, 0.07]$ and a discrete action set for accelerating forward, backward and no acceleration. To model the Mountain Car environment and implement a UCB-type algorithm, we first discretize the state space and apply the UCB algorithm on the tabular MDP. Specifically, we designate any state with position exceeding 0.5 as the goal state, while all other states are formed by uniformly discretizing the position range $[-1.2, 0.5]$ and velocity range $[-0.07, 0.07]$ into 30 equal intervals each, yielding 901 discrete states in total. The agent receives a reward of 1 only upon taking an action from the goal state; otherwise, the reward is 0. After an action is taken in the goal state, the environment is reset to its start configuration, then follows the original Mountain Car transition dynamics.

Offline Data Collection. Different from Boltzmann policy-based offline data collection, we use Algorithm 2, which iteratively interleaves exploration and exploitation to generate an offline dataset for the Mountain Car environment. Specifically, at each iteration, a model $\hat{P}$ is used to estimate two Q functions, $\hat{Q}_b(s,a)$ and $\hat{Q}_r(s,a)$. Here, $b(s,a)$ is an exploration bonus function akin to a UCB term [Auer et al., 2002], encouraging broader exploration, whereas $r(s,a)$ is the known reward function driving exploitation. These estimates yield two policies, exploration-focused $\hat{\pi}_b$ and exploitation-focused $\hat{\pi}_r$. Trajectories collected under these policies populate two datasets, D and D', respectively. After each round of data collection, both the model $\hat{P}$ and the bonus function $b(s,a)$ are updated, reflecting the optimism-in-the-face-of-uncertainty principle characteristic of UCB-based methods. After 10,000 iterations, trajectories from D and D' are combined using the offline coefficient α, thus balancing exploration and exploitation in the final offline dataset.

The motivation for this approach is that, in the Mountain Car environment, a purely uniform policy tends to remain confined to the valley, failing to explore higher positions effectively. This leads to inadequate coverage of the state-action space. By contrast, the pure online exploration policy $\hat{\pi}_b$ naturally seeks out all state-action pairs and thereby achieves broader coverage, populating $\mathcal{D}$ with a wide range of trajectories. Meanwhile, the pure exploitation policy $\hat{\pi}_r$ focuses on maximizing rewards and populates $\mathcal{D}'$ with near-optimal behavior. Finally, the offline coefficient α determines how these two datasets are combined into the final offline dataset $\mathcal{D}_0$, so Algorithm 2 can simulate the offline dataset with different coverage for our hybrid learning. Higher α generates better coverage on the optimal policy and lower α generates better coverage on all policies. Remarkably, this offline dataset collection method does not contradict our setting that the offline data should be collected by one fixed policy. The offline dataset can be viewed as being collected by a mixture policy that randomly samples $\{\hat{\pi}_b^i\}_i$ and $\{\hat{\pi}_r^i\}_i$ for N_0 times. However, it is difficult to provide the estimated $\mathtt{C}(\pi^*|\rho)$ for each policy or data distribution. Intuitively, a greater offline coefficient α results in smaller $\mathtt{C}(\pi^*|\rho)$ and greater $\mathtt{C}(\pi^{-\epsilon}|\rho)$; a smaller offline coefficient α results in greater $\mathtt{C}(\pi^*|\rho)$ and smaller $\mathtt{C}(\pi^{-\epsilon}|\rho)$.

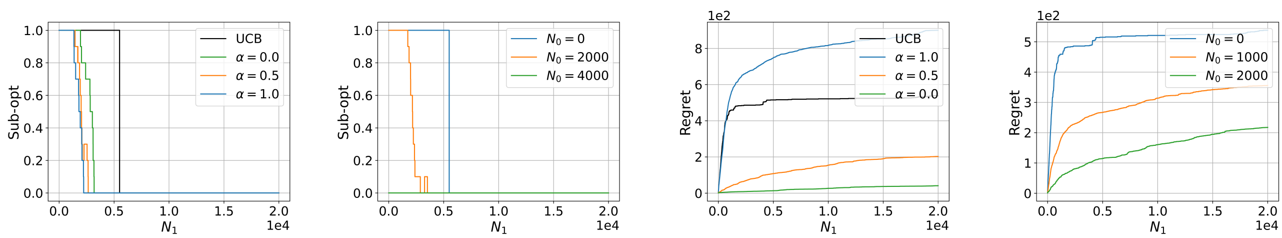
Algorithm 2 Mountain Car Offline Data Collection

Input: Coefficient $\alpha \in [0, 1]$, number of fffline trajectories $N_0 \leq 10000$, discount factor $\gamma = 0.99$.
Initialization: $\mathcal{D} \leftarrow \emptyset$, $\mathcal{D}' \leftarrow \emptyset$, $\hat{P}$ as a uniform transition model, $b(s, a) \leftarrow 1$
for $i = 1$ **to** 10000 **do**
  Estimate

$$\hat{Q}_b(s_0, a_0) = \hat{\mathbb{E}}_{\hat{P}}\Big[\sum_t \gamma^t b(s_t, a_t) \mid s_0, a_0\Big],$$

$$\hat{Q}_r(s_0, a_0) = \hat{\mathbb{E}}_{\hat{P}}\Big[\sum_t \gamma^t r(s_t, a_t) \mid s_0, a_0\Big]$$

  Derive policies $\hat{\pi}_b^i$ from $\hat{Q}_b$ and $\hat{\pi}_r^i$ from $\hat{Q}_r$
  Collect trajectories τ_i under $\hat{\pi}_b^i$ and τ_i' under $\hat{\pi}_r^i$
  Update $\mathcal{D} \leftarrow D \cup \{\tau_i\}$ and $\mathcal{D}' \leftarrow D' \cup \{\tau_i'\}$
  Update $\hat{P}$ and $b(s, a)$ using $\mathcal{D}$
end for
Sample $\alpha\, N_0$ trajectories from $\mathcal{D}'$ and $(1 - \alpha)\, N_0$ trajectories from $\mathcal{D}$ to form $\mathcal{D}_0$
Output: Offline dataset $\mathcal{D}_0$



(a) SOG v.s. α (b) SOG v.s. N_0 (c) Regret v.s. α (d) Regret v.s. N_0

Figure 3: Experimental results on sub-optimality gap (SOG) and regret for different offline coefficient α and number of offline trajectories N_0.

Results. We summarize our findings in Figure 3, where we experiment with various values of the offline coefficient α and different sizes of the offline dataset in the Mountain Car environment. For each configuration, we conduct 10 independent runs and track the mean sub-optimality gap or regret as a function of the online time horizon N_1. The pure UCB method without any offline data serves as our baseline. In Figure 3a and Figure 3c, the number of offline trajectories are 2000. In Figure 3b and Figure 3d, α is 0.5.

For sub-optimality gap minimization, our results in Figure 3a and Figure 3b indicate that incorporating offline data significantly improves performance compared to the pure online approach, validating the theoretical insights. In Figure 3a, when α is larger, the offline dataset emphasizes more trajectories generated by the near-optimal policy, reducing the sub-optimality gap more quickly. Also, in Figure 3b, increasing the size of the offline dataset further accelerates this reduction in sub-optimality.

For regret minimization, the results also show the benefit of incorporating the offline dataset. As in other settings, an offline dataset that prioritizes a near-optimal policy (i.e., higher α) can sometimes lead to slightly higher regret in the experiments, since the hybrid algorithm devotes exploration efforts to less-visited actions, which may be sub-optimal. Conversely, when α is smaller, the offline dataset covers a broader range of behaviors, leading to more informed exploration for regret minimization as Figure 3c. Also, in Figure 3d, enlarging the offline dataset size lowers the regret curve in each case, aligning with our theoretical predictions.