A Natural Extension To Online Algorithms For Hybrid RL With Limited Coverage

Kevin Tan*
kevtan@wharton.upenn.edu
Department of Statistics and Data Science
The Wharton School, University of Pennsylvania

Ziping Xu*
zipingxu@fas.harvard.edu
Department of Statistics
Harvard University

Abstract

Hybrid Reinforcement Learning (RL), leveraging both online and offline data, has garnered recent interest, yet research on its provable benefits remains sparse. Additionally, many existing hybrid RL algorithms (Song et al., 2023; Nakamoto et al., 2023; Amortila et al., 2024) impose a stringent coverage assumption called single-policy concentrability on the offline dataset, requiring that the behavior policy visits every state-action pair that the optimal policy does. With such an assumption, no exploration of unseen state-action pairs is needed during online learning. We show that this is unnecessary, and instead study online algorithms designed to "fill in the gaps" in the offline dataset, exploring states and actions that the behavior policy did not explore. To do so, previous approaches focus on estimating the offline data distribution to guide online exploration (Li et al., 2023b). We show that a natural extension to standard optimistic online algorithms – warm-starting them by including the offline dataset in the experience replay buffer – achieves similar provable gains from hybrid data even when the offline dataset does not have single-policy concentrability. We accomplish this by partitioning the state-action space into two, bounding the regret on each partition through an offline and an online complexity measure, and showing that the regret of this hybrid RL algorithm can be characterized by the best partition – despite the algorithm not knowing the partition itself. As an example, we propose DISC-GOLF, a modification of an existing optimistic online algorithm with general function approximation called GOLF used in Jin et al. (2021); Xie et al. (2022a), and show that it demonstrates provable gains over both online-only and offline-only reinforcement learning, with competitive bounds when specialized to the tabular, linear and block MDP cases. Numerical simulations further validate our theory that hybrid data facilitates more efficient exploration, supporting the potential of hybrid RL in various scenarios.

1 Introduction

Reinforcement Learning (RL) encompasses two main approaches: online and offline. Online RL involves agents learning to maximize rewards through real-time interactions with their environment, essentially learning by doing. Conversely, offline RL involves agents learning optimal actions by analyzing data collected by others, akin to learning by observation. However, learning by both watching and doing, or learning from both offline pre-collected data and online exploration, often called hybrid RL, remains under-explored.

Recent developments on hybrid RL theory have primarily focused on two aspects. The first line of work (Song et al., 2023; Nakamoto et al., 2023; Amortila et al., 2024) shows that hybrid RL, even without explicit exploration strategies like optimism during the online learning phase, can

*Equal contribution.

achieve the typical regret bounds of sample-efficient online algorithms that incorporate carefully designed exploration strategies. This is contingent upon the full single-policy concentrability of the offline dataset, highlighting hybrid RL's potential to simplify the design of the online learning component by eliminating the need for intricate exploration design. Our paper, however, follows another line of work (Wagenmaker & Pacchiano, 2023; Li et al., 2023b) that considers the case where the **offline dataset may not have full single-policy concentrability**[1]. Under partial coverage, the online algorithm could explore unseen states and actions not visited by the behavior policy, thereby demonstrating improvements over both pure offline and pure online learning approaches.

To analyze this case, Li et al. (2023b) suggest dividing the state and action space $\mathcal{X}$ within a tabular MDP into a disjoint partition $\mathcal{X}_{\text{off}} \oplus \mathcal{X}_{\text{on}} = \mathcal{X}$. The intuition is as follows. If the offline dataset has sufficient coverage of the state and action pairs in $\mathcal{X}_{\text{off}}$, a good algorithm should direct its online exploration to sufficiently explore $\mathcal{X}_{\text{on}}$. Previous approaches (Li et al., 2023b; Wagenmaker & Pacchiano, 2023) solve difficult optimization problems with the Frank-Wolfe algorithm to perform reward-free online exploration of the under-covered portion of the state and action space. These approaches are not generally applicable to existing state-of-the-art online algorithms for deep RL, and so we take a different approach.

Many online algorithms explore by maintaining an experience replay buffer, minimizing the empirical risk over it to sequentially update estimates about the unknown environment (Auer et al., 2008). One may trivially include the offline dataset in the experience buffer to obtain a hybrid RL algorithm, as others have previously noted (Song et al., 2023; Nakamoto et al., 2023; Amortila et al., 2024), under coverage assumptions on the offline dataset.[2]

Though being extensively applied in empirical studies, *it is not clear whether (1) simply appending the offline dataset to the experience replay buffer can lead to a provable improvement when the offline dataset is of poor quality, or (2) whether it ensures sufficient exploration for the portion of the state-action space without good coverage.* We seek to address this gap in our paper, tackling the more difficult setting where the offline data may be of arbitrarily poor quality without single-policy concentrability, in the context of regret-minimizing online RL with general function approximation.

Our Contributions. We address this gap by modifying an optimistic algorithm for general function approximation algorithm called GOLF (introduced in Jin et al. (2021) and used in Xie et al. (2022b)). We show that a hybrid version of GOLF (which we call DISC-GOLF) that simply includes an offline dataset in the parameter estimation achieves a provable improvement in the regret bound over pure online and offline learning, even when the offline dataset has poor coverage.

This is done through considering *arbitrary* (not necessarily disjoint) partitions of the state-action space $\mathcal{X}_{\text{off}} \cup \mathcal{X}_{\text{on}} = \mathcal{X}$. We bound the regret by the coverage of the behavior policy on the offline partition $\mathcal{X}_{\text{off}}$ and a complexity measure for online learning on the online partition $\mathcal{X}_{\text{on}}$. We then show that the overall regret of a hybrid algorithm can be characterized by the regret bound on the best possible partition – despite the algorithm not knowing the partition itself.[3] This analysis yields a general recipe for initializing generic online RL algorithms with offline data of arbitrarily poor quality, that we hope may be of use to other researchers seeking to derive similar algorithms.

We specialize this bound to the tabular, linear, and block MDP cases, achieving competitive sample complexities in each. Numerical simulations demonstrate that hybrid RL indeed encourages exploration of the region of the state-action space that is not well-covered by the offline dataset.

[1] An offline complexity measure that measures the coverage of the offline dataset (Zhan et al., 2022) with respect to the state-and-action pairs covered by a single reference policy.

[2] Unlike these, we are able to include the entire offline dataset – we do not need to discard any offline samples.

[3] This is similar in spirit to the adaptivity that Li et al. (2023b) showed for the tabular PAC RL case, but with a far more complicated algorithm that requires data splitting, behavior cloning, and reward-free exploration.

2 Problem Setup

We consider the situation where we are given access to a function class $\mathcal{F}$, and aim to model the optimal Q-function using it. Below, we introduce some notation that we use throughout the paper.

Notation. Let $\mathcal{N}_{\mathcal{F}}(\rho)$ be the ρ-covering number of function class $\mathcal{F}$ w.r.t the supremum norm. Let N_{off} and N_{on} (where $N = N_{\text{off}} + N_{\text{on}}$) be the number of episodes in the offline dataset and the number of online episodes respectively. We will use the notation $T = N_{\text{on}}$ interchangeably. For any set $\mathcal{X} \subset \mathcal{S} \times \mathcal{A} \times [H]$, let $\mathcal{X}_h = \{(s,a) \in \mathcal{S} \times \mathcal{A} : (s,a,h) \in \mathcal{X}\}$, and $\Delta(\mathcal{X})$ all distributions over $\mathcal{X}$.

Episodic MDPs. We consider episodic MDPs denoted by $\{\mathcal{S}, \mathcal{A}, H, P, R\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, H the horizon, $P = \{P_h\}_{h\in[H]}$ the collection of transition probabilities with each $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$, and $R = \{R_h\}_{h\in[H]}$ the collection of reward functions with each $R_h : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$. An agent interacts with the environment for H steps within each episode. On the each step $h \in [H]$, the agent observes the current state $s_h \in \mathcal{S}$ and chooses an action $a_h \in \mathcal{A}$, and the environment generates the next state $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ and the current reward $r_h = R_h(s_h, a_h)$. A policy π is a mapping from $\mathcal{S}$ to $\Delta(\mathcal{A})$, the set of distributions over the action space. The function class $\mathcal{F}$ induces a policy class $\Pi := \{\pi^f : f \in \mathcal{F}\}$ through the greedy policy with regard to each function π^f. Throughout the paper, we denote $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times [H]$.

Definition 1 (Occupancy Measure)**.** *The occupancy measure $d^\pi = \{d_h^\pi\}_{h=1}^H$ is the collection of state-action distributions induced by running policy π. We write $\mathbb{D}$ for the set of all possible d^π.*

Hybrid RL. We study the natural setting of online fine-tuning given access to an offline dataset, where an agent interacts with the environment for N_{on} steps given access to an offline dataset $\mathcal{D}_{\text{off}}$ consisting of N_{off} episodes. We assume that the offline dataset is collected through some fixed policy $\pi_{\text{off}} = \{\pi_{\text{off},h}\}_{h\in[H]}$. Let μ be the occupancy measure induced by π_{off}, and denote by $s_h^{(t)}$, $a_h^{(t)}$ and $r_h^{(t)}$ the state, action and reward on step $h \in [H]$ within episode $t \in [N_{\text{on}}]$. The goal of an online RL algorithm is to maximize the cumulative reward $\sum_{t=1}^{N_{\text{on}}} \sum_{h=1}^H r_h^{(t)}$.

We follow the standard definition of value functions for episodic MDPs. The value function of a policy π is $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'} \mid s_{h'} = s]$, where $\mathbb{E}_\pi$ denotes the expectation over trajectories induced by taking policy π. Let $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'} + V_{h'+1}^\pi(s_{h'+1}) \mid s_{h'} = s, a_{h'} = a]$, where we set $V_{H+1}^\pi(s) \equiv 0$. Write $V^\star$ and $Q^\star$ for the optimal value and Q-functions. The cumulative regret of an online algorithm $\mathcal{L}$ is $\mathrm{Reg}(N_{\text{on}}, \mathcal{L}) = \mathbb{E}_{\mathcal{L}}\left[\sum_{t=1}^{N_{\text{on}}} \left(V_1^\star(s_1^{(t)}) - \sum_{h=1}^H r_h^{(t)}\right)\right]$, where $\mathcal{L} : \mathcal{H} \to \Pi$ is any learning algorithm that maps all the previous observations, i.e. the history $\mathcal{H}$, to a policy, and $\mathbb{E}_{\mathcal{L}}$ denotes the expectation over all the trajectories generated by the interaction between algorithm $\mathcal{L}$ and the underlying MDP.

Function Approximation. We approximate the optimal Q-function with a function class $\mathcal{F} = \{\mathcal{F}_h\}_{h\in[H]}$, where each $\mathcal{F}_h \subseteq [0,H]^{\mathcal{S}\times\mathcal{A}}$. The Bellman operator for each $h \in [H-1]$ is $\mathcal{T}_h f_{h+1}(s,a) := R_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} f_{h+1}(s',a')\right]$. We further define the Bellman error w.r.t $f \in \mathcal{F}$ by $\mathcal{E}_h f = \mathcal{T}_h f_{h+1} - f_h$ and the squared Bellman error by $\mathcal{E}_h^2 f = (\mathcal{T}_h f_{h+1} - f_h)^2$. For a distribution $d \in \Delta(\mathcal{S}\times\mathcal{A})$, we write $\|f_h - \mathcal{T}_h f_{h+1}\|_{2,d}^2 = \mathbb{E}_{(s_h,a_h)\sim d}[\mathcal{E}_h^2 f]$. Below, we make the following routine assumption on the richness of the function class (Liu et al., 2020; Rajaraman et al., 2020; Rashidinejad et al., 2023; Uehara & Sun, 2023). This may be relaxed to the weaker related notion of realizability as in Zanette (2023) at the cost of an amplifying factor dependent on the metric entropy of the function class, dataset coverage, and the discrepancy between $\mathcal{F}$ and its image under the Bellman operator, but this is outside the scope of our analysis.

Assumption 1 (Bellman Completeness)**.** *We assume that for all $f_{h+1} \in \mathcal{F}_{h+1}$, $\mathcal{T}_h f_{h+1} \in \mathcal{F}_h$. Note that this implies realizability: $Q_h^* \in \mathcal{F}_h$.*

3 Measures of Complexity

In this section, we extend existing complexity measures for offline and online learning with general function approximation in order to use them to understand the complexity of hybrid RL. We will

use each on an arbitrary partition of the state-action space, with the intuition being that the offline complexity measure should characterize the difficulty of learning only on the portion that is well-covered by the behavior policy, and the online complexity measure for the difficulty of learning on the portion that has not been explored yet. We later show that a subsequent regret bound can be determined by the complexity measures over any partition, and so the regret is characterized by the infimum over the partitions of the complexity measures on them.

Offline Complexity Measures. In offline RL, the sample complexity is bounded by the notion of concentrability (Xie et al., 2021). For a function class on Bellman error $\mathcal{G}$ and a reference policy π, the (L_2 Bellman-error) all-policy and single-policy concentrability coefficients (Zhan et al., 2022) are defined as:

$$c_{\text{off}}(\mathcal{F}, \pi) := \max_h \sup_{f \in \mathcal{F}} \frac{\|f_h - \mathcal{T}_h f_{h+1}\|^2_{2,d^\pi_h}}{\|f_h - \mathcal{T}_h f_{h+1}\|^2_{2,\mu_h}}, \text{ and } c_{\text{off}}(\mathcal{F}) := \sup_\pi c_{\text{off}}(\mathcal{F}, \pi).$$

We note that other variants exist, such as the L_∞ density-ratio single-policy concentrability which we define as $C^* = \sup_{h,s,a} d^{\pi^*}_h(s,a)/\mu_h(s,a)$. We clarify which variant of single-policy and all-policy concentrability we refer to whenever possible, but note that the L_2 Bellman-error concentrability is upper bounded by the L_∞ density-ratio concentrability Zhu et al. (2023). There is an algorithm (Xie et al., 2021) that finds an ϵ-optimal policy in $\tilde{\mathcal{O}}(c_{\text{off}}(\mathcal{F}, \pi^\star)/\epsilon^2)$ episodes.

Online Complexity Measures. To characterize the online complexity measure, we extend a recently proposed measure, the SEC (Sequential Extrapolation Coefficient) from Xie et al. (2022a):

$$c_{\text{on}}(\mathcal{F}, T) := \max_{h \in [H]} \sup_{\{f^{(1)}, \ldots, f^{(T)}\} \subseteq \mathcal{F}} \sup_{(\pi^{(1)}, \ldots, \pi^{(T)})} \left\{ \sum_{t=1}^{T} \frac{\mathbb{E}_{d^{\pi^{(t)}}_h}[f^{(t)}_h - \mathcal{T}_h f^{(t)}_{h+1}]^2}{H^2 \vee \sum_{i=1}^{t-1} \mathbb{E}_{d^{\pi^{(i)}}_h}[(f^{(t)}_h - \mathcal{T}_h f^{(t)}_{h+1})^2]} \right\}.$$

We note that in their paper, the SEC has a 1 in the denominator instead of H^2 because they assume $Q_h \in [0,1]$. Xie et al. (2022a) provide an online algorithm with a regret bound of the form $\tilde{\mathcal{O}}(H\sqrt{c_{\text{on}}(\mathcal{F}, T) \cdot T})$. Similar extensions can be proposed for other online complexity measures.

Reduced Complexity Through State-Action Space Partition. As previously mentioned, a hybrid algorithm can reduce its online learning complexity by exploring what has not been seen in the offline dataset. This motivates us to consider a partition on the state-action space $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times [H]$. We denote the offline and online partition by $\mathcal{X}_{\text{off}}$ and $\mathcal{X}_{\text{on}}$, respectively. We define the offline and online partial complexity measure on each partition by

$$c_{\text{off}}(\mathcal{F}, \mathcal{X}_{\text{off}}) := \max_h \sup_{f \in \mathcal{F}} \frac{\|(f_h - \mathcal{T}_h f_{h+1}) \mathbb{1}_{(\cdot,h) \in \mathcal{X}_{\text{off}}}\|^2_{2,d^\pi_h}}{\|(f_h - \mathcal{T}_h f_{h+1}) \mathbb{1}_{(\cdot,h) \in \mathcal{X}_{\text{off}}}\|^2_{2,\mu_h}},$$

$$c_{\text{on}}(\mathcal{F}, \mathcal{X}_{\text{on}}, T) := \max_{h \in [H]} \sup_{\{f^{(1)}, \ldots, f^{(T)}\} \subseteq \mathcal{F}} \sup_{(\pi^{(1)}, \ldots, \pi^{(T)})} \left\{ \sum_{t=1}^{T} \frac{\mathbb{E}_{d^{\pi^{(t)}}_h}[(f^{(t)}_h - \mathcal{T}_h f^{(t)}_{h+1}) \mathbb{1}_{(\cdot,h) \in \mathcal{X}_{\text{on}}}]^2}{H^2 \vee \sum_{i=1}^{t-1} \mathbb{E}_{d^{\pi^{(i)}}_h}[(f^{(t)}_h - \mathcal{T}_h f^{(t)}_{h+1})^2 \mathbb{1}_{(\cdot,h) \in \mathcal{X}_{\text{on}}}]} \right\}.$$

Viewing c_{on} and c_{off} as complexity measures on the function class $\mathcal{F}_h - \mathcal{T}_h \mathcal{F}_{h+1}$ induced by $\mathcal{F}$ and Bellman operator $\mathcal{T}$, our partial complexity measures can be seen as restricting this function class such that any function in this class is non-zero only when the input is in $\mathcal{X}_{\text{off}}$ or $\mathcal{X}_{\text{on}}$. This leads to smaller complexity measures for both online and online learning. This is not unique to our choices of complexity measures. Other measures in the literature, such as the Rademacher complexity and covering number, also indicate a reduced complexity for $\mathcal{F}_h - \mathcal{T}_h \mathcal{F}_{h+1}$.

Partial All-Policy Concentrability Is Less Stringent Than Single-Policy Concentrability. While Li et al. (2023b) successfully employ a notion of (L_∞ density ratio) partial single-policy concentrability in the tabular setting, our regret bound depends on the (L_2 Bellman error) partial all-policy concentrability. This falls short of the notion of partial single-policy concentrability that Li et al. (2023b) successfully employ in the tabular setting. We attribute this to our desire to work with the simple procedure of appending the offline dataset to the experience replay buffer in the

context of general function approximation – our algorithm is much simpler and their techniques, being specialized to the tabular case, cannot be extended to general function approximation.

However, as our regret bound utilizes the best partition of the state-action space, our result already obtains an improvement over the common requirement of single-policy concentrability *over the entire state-action space* in hybrid RL with general function approximation (Song et al., 2023; Nakamoto et al., 2023; Amortila et al., 2024). While the two are not directly comparable, the best partial all-policy concentrability coefficient, which our algorithm uses adaptively, is always finite (we can always take $\mathcal{X}_{\text{off}}$ to be a singleton) even when the single-policy concentrability coefficient is unbounded.

Main Result. Our main novel theoretical result is in showing that the overall regret of a hybrid algorithm (we first show this for DISC-GOLF, then for a general class of online algorithms) can be characterized by $c_{\text{off}}(\mathcal{F}, \mathcal{X}_{\text{off}})$ and $c_{\text{on}}(\mathcal{F}, \mathcal{X}_{\text{on}}, N_{\text{on}})$ for any (not necessarily disjoint) partition $\mathcal{X}_{\text{on}}$ and $\mathcal{X}_{\text{off}}$ – despite the algorithm not knowing the partition itself. As this holds for every partition, the guarantee we provide therefore incorporates the best possible split without the algorithm having to know or estimate it.

4 Online Finetuning From Offline Data

Here is an example. In this section, we derive an efficient regret bound for an optimistic online algorithm with general function approximation that is warm-started with offline data of arbitrarily poor quality. This regret bound demonstrates provable gains over both online-only and offline-only reinforcement learning through splitting the state-action space.[4]

An Optimistic Hybrid RL Algorithm Warm-Started With Offline Data. We modify the GOLF algorithm from Xie et al. (2022a) to incorporate a dataset $\mathcal{D}_{\text{off}}$ collected by a behavior policy π_b with occupancy measure μ. We name the resulting algorithm DISC-GOLF.[5] The modification is simple and intuitive – we simply warm-start the online exploration by appending the offline data to the experience replay buffer at the beginning, and explore from there. Remarkably, this simple modification enables us to deal with an offline dataset that only has partial coverage. To our knowledge, this has only previously been accomplished in the tabular setting with a far more complicated algorithm (Li et al., 2023b).

Algorithm 1 DISC-GOLF

Input: Offline dataset $\mathcal{D}_{\text{off}}$, samples sizes N_{on}, N_{off}, function class $\mathcal{F}$ and confidence width $\beta > 0$

Initialize: $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} \leftarrow \emptyset, \forall h \in [H]$

for episode $t = 1, 2, \ldots, N_{on}$ **do**

Select policy $\pi^{(t)} \leftarrow \pi_{f^{(t)}}$, where $f^{(t)} := \operatorname{argmax}_{f \in \mathcal{F}^{(t-1)}} f_1\left(x_1, \pi_{f,1}\left(x_1\right)\right)$.

Execute $\pi^{(t)}$ for one episode and obtain trajectory $(s_1^{(t)}, a_1^{(t)}, r_1^{(t)}), \ldots, (s_H^{(t)}, a_H^{(t)}, r_H^{(t)})$.

Update dataset $\mathcal{D}_h^{(t)} \leftarrow \mathcal{D}_h^{(t-1)} \cup \{(s_h^{(t)}, a_h^{(t)}, r_h^{(t)}, s_{h+1}^{(t)})\}, \forall h \in [H]$.

Compute confidence set:

$$\mathcal{F}^{(t)} \leftarrow \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t)}\left(f_h, f_{h+1}\right) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}\left(f_h', f_{h+1}\right) \leq \beta \quad \forall h \in [H] \right\},$$

$$\text{where } \mathcal{L}_h^{(t)}\left(f, f'\right) := \sum_{(s,a,r,s') \in \mathcal{D}_h^{(t)} \cup \mathcal{D}_{\text{off},h}} \left(f(s,a) - r - \max_{a' \in \mathcal{A}} f'\left(s', a'\right) \right)^2, \forall f \in \mathcal{F}_h, f' \in \mathcal{F}_{h+1}.$$

end for

Main Result. The following result shows that the regret can be decomposed into two terms that depend on the offline and online complexity measures over the best possible partition of $\mathcal{X}$.

[4] The algorithm is never aware of the partition. The partition is only a convenient, but useful, theoretical construct.

[5] Data Informed Sequential Confidence-sets – Global Optimism based on Local Fitting.

Theorem 1 (Regret Bound for DISC-GOLF)**.** *Let $\mathcal{X}_{\text{off}}, \mathcal{X}_{\text{on}}$ be an arbitrary partition over $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times [H]$. Algorithm 1 satisfies the following regret bound with probability at least $1 - \delta$:*

$$\text{Reg}(N_{\text{on}}) = \mathcal{O}\left(\inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} \left(\sqrt{\beta H^4 N_{\text{on}} \left(\frac{N_{\text{on}}}{N_{\text{off}}}\right) c_{\text{off}}(\mathcal{F}, \mathcal{X}_{\text{off}})} + \sqrt{\beta H^4 N_{\text{on}} c_{\text{on}}(\mathcal{F}, \mathcal{X}_{\text{on}}, N_{\text{on}})}\right)\right),$$

where $\beta = c_1 \log\left(NH\mathcal{N}_{\mathcal{F}}(1/N)/\delta\right)$ for some constant c_1 with $N = N_{\text{on}} + N_{\text{off}}$.[6]

We defer the proof to Appendix A. This shows that an optimistic online RL algorithm can be adapted to the hybrid setting in a very natural way – initializing it with an offline dataset. Although the algorithm is completely unaware of the partition, the regret bound provides the best regret guarantee over all partitions of the state-action space.

The offline term depends on $N_{\text{on}}\left(\frac{N_{\text{on}}}{N_{\text{off}}}\right)$, and so depends on the ratio of the number of online and offline episodes. However, due to the infimum over partitions, the overall regret bound will always be no worse than $\tilde{\mathcal{O}}(\sqrt{N_{\text{on}}})$, as when $N_{\text{on}} \gg N_{\text{off}}$ we can simply take $\mathcal{X}_{\text{on}} = \mathcal{X}$ to find that $c_{\text{off}}(\mathcal{F}, \emptyset) = 0$. Conversely, in the few-shot learning setting where $N_{\text{off}} \gg N_{\text{on}}$, the regret bound is approximately $\tilde{\mathcal{O}}\left(\sqrt{\beta H^4 N_{\text{on}} c_{\text{on}}(\mathcal{F}, \mathcal{X}_{\text{on}}, N_{\text{on}})}\right)$, improving on the GOLF regret of $\tilde{\mathcal{O}}\left(\sqrt{\beta H^4 N_{\text{on}} c_{\text{on}}(\mathcal{F}, \mathcal{X}, N_{\text{on}})}\right)$.

This bound roughly matches that of Song et al. (2023); Nakamoto et al. (2023); Amortila et al. (2024) in terms of the dependence on horizon and log-covering number. However, unlike these, we do not require single-policy concentrability. The infimum over partitions gives us a finite partial all-policy concentrability coefficient $c_{\text{off}}(\mathcal{F}, \mathcal{X}_{\text{off}})$, even when the single-policy concentrability coefficient over the entire space C^* is unbounded. Additionally, these previous approaches discard any offline data beyond the size of the online dataset (i.e. offline datapoints $N_{\text{on}} + 1, ..., N_{\text{off}}$), and so obtain a guarantee that does not depend on N_{off}. We do not need to discard any offline samples, enabling us to use the offline data in our regret bound.

5 Case Studies

Theorem 1 established a regret bound for the general function approximation setting. Throughout this section, we examine case studies to demonstrate the exact improvement of hybrid RL algorithm over pure online and pure offline algorithms and characterize the set of good partitions. We defer all proofs in this section to Appendix C.

5.1 Tabular MDPs.

The most commonly considered MDP family is that of the Tabular MDPs, with a finite number of states and actions. As each Q function at the step h can be represented as a $|\mathcal{S}| \times |\mathcal{A}|$ dimensional vector, we consider the function class $\mathcal{F}_h = [0, H]^{|\mathcal{S}||\mathcal{A}|}$. For a constant $\rho > 0$, an intuitive choice of partition that corresponds closely to the choice of Li et al. (2023b) is $\mathcal{X}_{\text{off}}(\rho) := \{(s, a, h) : \sup_\pi d_h^\pi(s, a)/\mu_h(s, a) \leq \rho\}$. As such, the partial offline concentrability coefficient reduces to the supremum of density ratios over the offline partition, allowing us to bound the partial SEC by the cardinality of the online partition.

Proposition 1. *We can bound $c_{\text{off}}(\mathcal{F}, \mathcal{X}_{\text{off}}) \leq \sup_\pi \sup_{(s,a,h) \in \mathcal{X}_{\text{off}}} \frac{d_h^\pi(s,a)}{\mu_h^\pi(s,a)} = \sup_\pi \left\| \frac{d_h^\pi \mathbb{1}_{\mathcal{X}_{\text{off}}}}{\mu_h^\pi} \right\|_\infty$ and $c_{\text{on}}(\mathcal{F}, \mathcal{X}_{\text{on}}) \lesssim \max_{h \in [H]} |\mathcal{X}_{\text{on},h}| \log(N_{\text{on}})$. As such, with probability at least $1 - \delta$,*

$$\text{Reg}(N_{\text{on}}) = \tilde{\mathcal{O}}\left(\inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} \left(\sqrt{H^5 S A N_{\text{on}} \left(\frac{N_{\text{on}}}{N_{\text{off}}}\right) \sup_\pi \left\| \frac{d_h^\pi \mathbb{1}_{\mathcal{X}_{\text{off}}}}{\mu_h^\pi} \right\|_\infty} + \sqrt{H^5 S A \max_{h \in [H]} |\mathcal{X}_{\text{on}}| N_{\text{on}}}\right)\right).$$

[6] The online-only bound in Xie et al. (2022a) is of the form $\sqrt{\beta H^2 N_{\text{on}} c_{\text{on}}(\mathcal{F}, \mathcal{X}, N_{\text{on}})}$, as they assume Q-functions are bounded by $[0, 1]$, accounting for the remaining H^2 dependence.

Therefore, if the offline dataset has good coverage on a subset $\mathcal{X}_{\text{off}}$, the complexity of online learning complexity can be reduced to the cardinality of its complement $\mathcal{X}_{\text{on}}$. We then obtain a regret bound that is at most a factor of H^2SA off from the minimax-optimal results in the offline-only and online-only cases (Rashidinejad et al., 2023; Shi et al., 2022; Azar et al., 2017; Xie et al., 2022b), even though (1) DISC-GOLF is a very general model-free function-approximation algorithm, and (2) we did not perform a specialized analysis of this case beyond simply bounding the partial SEC in this setting. We anticipate that analyzing specialized versions of DISC-GOLF can achieve tighter sample complexities in the same sense that Li et al. (2023a) accomplish for Q-learning. Note that in a few shot learning setting, where $N_{\text{off}} \gg N_{\text{on}}$, the regret is approximately $\tilde{\mathcal{O}}\left(\sqrt{H^5SA \max_h |\mathcal{X}_{\text{on},h}| N_{\text{on}} \log(N_{\text{on}})}\right)$, where $\mathcal{X}_{\text{on}}$ is the set of state, action and step tuples where the offline occupancy measure μ is unsupported.

5.2 Linear MDPs.

The family of Linear MDPs is a common MDP family that generalizes the tabular case, defined in Definition 2. It can be shown that the linear function class for action-value function approximation: $\mathcal{F}_h = \{\langle \phi(\cdot), w_h \rangle : w_h \in \mathbb{R}^d, \|w_h\| \le 2H\sqrt{d}\}$ is Bellman complete (Jin et al., 2020).

Definition 2 (Linear MDP)**.** *An episodic MDP is a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown (signed) measures $\boldsymbol{\nu}_h = (\nu_h^{(1)}, \ldots, \nu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $P_h(\cdot \mid s,a) = \langle \phi(s,a), \boldsymbol{\nu}_h(\cdot) \rangle$ and $r_h(s,a) = \langle \phi(s,a), \boldsymbol{\theta}_h \rangle$, where $\|\phi(s,a)\|_2 \le 1$ for all s,a and $\max\{\|\boldsymbol{\nu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_h(\mathcal{S})\| \le \sqrt{d}\}$ for all $h \in [H]$.*

We can define a partition of the state-action space $\mathcal{X}$ as follows. For any subset $\mathcal{X}' \subset \mathcal{S} \times \mathcal{A}$, consider the image of the feature map $\phi(\mathcal{X}') = \{\phi(s,a) : (s,a) \in \mathcal{X}'\}$. We can choose $\Phi_{\text{off}} \subseteq \mathbb{R}^d$ and $\Phi_{\text{on}} \subseteq \mathbb{R}^d$ to be the subspaces spanned by $(\phi(\mathcal{X}_{\text{on},h}))_{h\in[H]}$ and $(\phi(\mathcal{X}_{\text{off},h}))_{h\in[H]}$, with dimensions d_{off} and d_{on} respectively. That is, any partition of the state-action space $\mathcal{X}$ induces two subspaces of $\mathbb{R}^d$ through the feature map ϕ. Let $\mathcal{P}_{\text{off}}$ and $\mathcal{P}_{\text{on}}$ be the orthogonal projection operators onto Φ_{off} and Φ_{on}. We can then upper bound the complexity measures over each partition, as we show in Proposition 2.

Proposition 2. *Let $\phi_{\text{off}} = \mathcal{P}_{\text{off}}\phi$. We have $c_{\text{off}}(\mathcal{F}, \mathcal{X}_{\text{off}}) \le \max_h 1/\lambda_{d_{\text{off}}}(\mathbb{E}_{\mu_h}[\phi_{\text{off}}\phi_{\text{off}}^\top])$ and $c_{\text{on}}(\mathcal{G}_{\text{on}}) = \mathcal{O}(d_{\text{on}} \log(HN_{\text{on}}) \log(N_{\text{on}}))$, where λ_n is the n-th largest eigenvalue. Then, with probability at least $1-\delta$, the regret $\mathrm{Reg}(N_{\text{on}})$ is bounded by*

$$\mathrm{Reg}(N_{\text{on}}) = \tilde{\mathcal{O}}\left(\inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} \left(\sqrt{dH^5 N_{\text{on}} \left(\frac{N_{\text{on}}}{N_{\text{off}}}\right) \max_h \frac{1}{\lambda_{d_{\text{off}}}(\mathbb{E}_{\mu_h}[\phi_{\text{off}}\phi_{\text{off}}^\top])}} + \sqrt{d_{\text{on}} dH^5 N_{\text{on}}}\right)\right).$$

We can compare this result to the $\sqrt{d^2H^3N_{\text{on}}}$ minimax lower bound from Zhou et al. (2021), and the best known upper bound from Zanette et al. (2020) of $\sqrt{d^2H^4N_{\text{on}}}$, for online RL in linear MDPs. It is exciting to note that by incorporating offline data into an online algorithm, we can improve the dependence on dimension of the regret incurred on the online partition from d^2 to $d_{\text{on}}d$. We accomplish this by bounding the SEC in the linear MDP case by d_{on}, up to logarithmic factors. This therefore demonstrates another example of provable gains from hybrid RL.

5.3 Block MDPs.

A block MDP (BMDP) refers to an environment with a finite but unobservable latent state space $\mathcal{U}$, a finite action space $\mathcal{A}$, and a possibly infinite but observable state space $\mathcal{S}$ (Dann et al., 2019; Misra et al., 2019; Du et al., 2021). At each step, the environment generates a current state $s_h \sim q(\cdot \mid u_h)$ given the underlying latent state $u_h \in \mathcal{U}$. This is described by the block structure outlined below.

Definition 3 (Block Structure)**.** *A block MDP is an MDP where each context $x \in \mathcal{X}$ uniquely determines its generating state $u \in \mathcal{U}$, i.e. there is a decoding function $f^* : \mathcal{S} \mapsto \mathcal{U}$ such that $q(\cdot \mid u)$ is supported on $(f^*)^{-1}(u)$.*

Any partition $\mathcal{X}_{\text{off}}, \mathcal{X}_{\text{on}}$ induces a partition on the latent state-action space $\bar{\mathcal{X}}_{\text{off}} = \{(f^*(s), a, h) : (s, a, h) \in \mathcal{X}_{\text{off}}\}$ and $\bar{\mathcal{X}}_{\text{on}} = \{(f^*(s), a, h) : (s, a, h) \in \mathcal{X}_{\text{on}}\}$, and the offline behavior policy and a given policy π induce measures $\bar{\mu}_h$ and $\bar{d}_h^\pi$ on $\mathcal{U} \times \mathcal{A}$. Then, Proposition 3 shows that the offline and online learning complexities are determined by the cardinalities of the induced partitions of the latent state space. This bound is also dependent on β, but we omit it in the main text for brevity.

Proposition 3. *In a block MDP,* $c_{\text{off}}(\mathcal{F}, \mathcal{X}_{\text{off}}) \leq \sup_\pi \sup_{(u,a,h) \in \bar{\mathcal{X}}_{\text{off}}} \frac{\bar{d}_h^\pi(u,a)}{\bar{\mu}_h^\pi(u,a)}$ *and* $c_{\text{on}}(\mathcal{F}, \mathcal{X}_{\text{on}}, T) = \mathcal{O}(\max_h |\bar{\mathcal{X}}_{\text{on},h}| \log(N_{\text{on}}))$ *if* $\mathcal{F}$ *is Bellman-complete. Then, with probability at least* $1 - \delta$*,*

$$\text{Reg}(N_{\text{on}}) = \tilde{\mathcal{O}}\left(\inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} \left(\sqrt{H^4 N_{\text{on}} \left(\frac{N_{\text{on}}}{N_{\text{off}}}\right) \sup_\pi \sup_{(u,a,h) \in \bar{\mathcal{X}}_{\text{off}}} \frac{\bar{d}_h^\pi(u,a)}{\bar{\mu}_h^\pi(u,a)}} + \sqrt{H^4 N_{\text{on}} \max_h |\bar{\mathcal{X}}_{\text{on},h}|}\right)\right).$$

6 A Recipe for General Algorithms

The analysis and techniques used above are by no means applicable only to DISC-GOLF. In Proposition 4 below, we provide a general recipe that can be used to analyze how a general online algorithm $\mathcal{L}$ can benefit from being initialized with access to an offline dataset.

We define $d_h^{(t)}$ to be the measure over $\mathcal{S} \times \mathcal{A}$ induced by running algorithm $\mathcal{L}$ for t iterations at horizon h. This bound depends on a set of error terms δ_h^t, which for example is (1) the Bellman error $f_h^t - \mathcal{T}_h f_{h+1}^t$ in the case of general function approximation with DISC-GOLF, (2) the sum of upper confidence bonus terms, estimation errors, and two martingale terms with UCBVI (Azar et al., 2017) for the tabular setting, and (3) the gap multiplied by the probability each arm is pulled in the bandit case with UCB (Auer, 2003). We then have the following result below that provides a guarantee for the procedure of "hybridifying" general online algorithms by initializing them with offline datasets. We defer the proof of Proposition 4 to Appendix D.

Proposition 4. *Let* $\mathcal{L}$ *be a general online learning algorithm that satisfies the following conditions:*

1. $\mathcal{L}$ *admits the regret decomposition* $\text{Reg}_{\mathcal{L}}(T) \leq \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{(s,a) \sim d_h^{(t)}}[\delta_h^t(s,a)]$ *for some collection of random functions*[7] $(\delta_h^t)_{h=1}^H$ *with each* δ_h^t *a mapping from* $\mathcal{X} \mapsto \mathbb{R}$*;*
2. $\sum_{t=1}^T \sum_{h=1}^H \left(N_{\text{off}} \mathbb{E}_{(s,a) \sim \mu_h}[\delta_h^t(s,a)^2] + \sum_{i=1}^{t-1} \mathbb{E}_{(s,a) \sim d_h^i}[\delta_h^t(s,a)^2]\right) \leq \beta(\delta, H)$ *w.p.* $1 - \delta$*;*
3. *there exists a function* $c_{\text{on}} : \mathcal{P}(\mathcal{X}) \times \mathbb{N}$ *such that for any* $\mathcal{X}' \subset \mathcal{X}$*, it holds with a probability at least* $1 - \delta$ *that* $\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{(s,a) \sim d_h^{(t)}}[\delta_h^t(s,a) \mathbb{1}(x,a,h) \in \mathcal{X}'] = \mathcal{O}(c_{\text{on}}(\mathcal{X}', T) H^\gamma \beta(\delta, H) T)^\xi$*, for some* $\xi \in (0,1)$*,* $\gamma \in \mathbb{Z}_{\geq 0}$*, and where* $\beta : (0,1) \mapsto \mathbb{R}$ *is some measure of complexity of the algorithm and its dependence on the probability of failure* δ*;*
4. *a coverage measure on any* $\mathcal{X}' \subset \mathcal{X}$ *of* $c_{\text{off}}(\mathcal{X}') := \sup_{h \in [H]} \sup_\pi \frac{\mathbb{E}_{d_h^\pi}[\delta_h^t(s,a) \mathbb{1}(s,a,h \in \mathcal{X}')]}{\mathbb{E}_{\mu_h}[\delta_h^t(s,a) \mathbb{1}(s,a,h \in \mathcal{X}')]}$[8]*.*

Then, the algorithm $\mathcal{L}$ *satisfies the following regret bound w.p. at least* $1 - \delta$*:*

$$\text{Reg}_{\mathcal{L}}(T) = \mathcal{O}\left(\inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} (c_{\text{on}}(\mathcal{X}_{\text{on}}, T) \beta(\delta, H) H^\gamma T)^\xi + H \sqrt{\beta(\delta, H) \cdot c_{\text{off}}(\mathcal{X}_{\text{off}}) \cdot \frac{N_{\text{on}}^2}{N_{\text{off}}}}\right).$$

Informally, Proposition 4 states that given (1) a regret decomposition over the errors at each timestep, (2) a bound on the in-sample error (or just the error under the behavior policy measure), (3) an online-only regret bound for the original algorithm, and (4) an offline coverage measure, we can provide a similar guarantee to what we showed for DISC-GOLF in Theorem 1. We anticipate that one can use this or similar arguments to improve upon the minimax-optimal online-only and offline-only regret bounds when analyzing more specialized algorithms.

[7] This is often the Bellman error in the case of MDPs.

[8] We set 0/0 as 0.

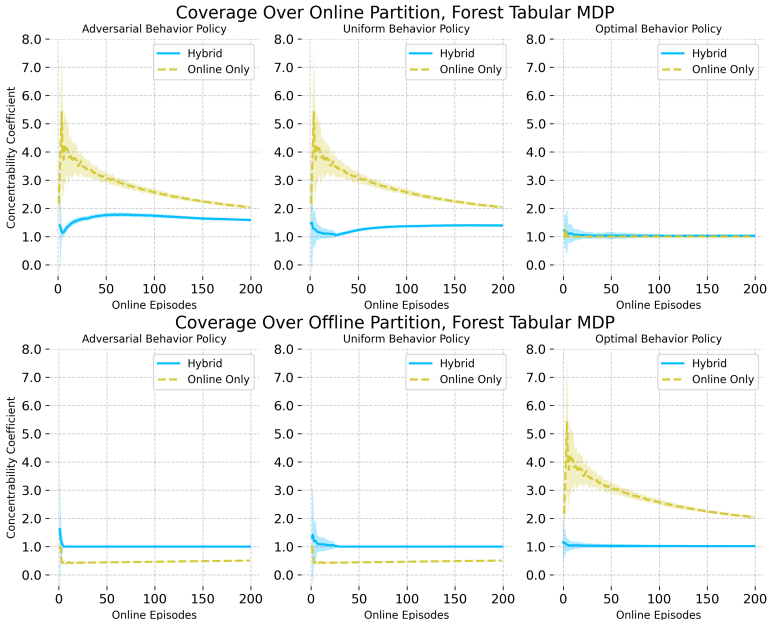

Figure 1: Coverage of the online samples averaged over 30 trials, with $1.96\hat{\sigma}$ confidence intervals. Hybrid RL explores more of the online partition and less of the offline partition than online RL when the behavior policy is poor, and vice-versa when the behavior policy is good. Lower is better.

7 Numerical Experiments

To illustrate the notion that appending the offline dataset to the experience replay buffer can encourage sufficient exploration for the portion of the state-action space that does not have good coverage, we perform two simulation studies in the tabular and linear MDP settings respectively.[9]

7.1 Forest, Tabular MDP.

We used a simple forest management simulator from the `pymdptoolbox` package of Cordwell et al. (2015). This environment has 4 states and 2 actions, and we used a horizon of 20 years. Every year, the agent can choose to wait and let the forest grow, earning a reward of 4 if the forest is 3 years old and 0 otherwise, or cut the forest down, earning a reward of 1 if the forest is between $1-2$ years old, 2 if the forest is 3 years old, and 0 otherwise. The forest burns down with 0.1 probability each year (making it 0 years old).

We examine how an optimistic model-based algorithm, UCBVI (Azar et al., 2017), behaves when warm-started with an offline dataset. We considered three behavior policies – adversarial, uniform, and optimal. The adversarial behavior policy does the opposite of the optimal policy 60% of the time, and takes a random action 40% of the time. Each offline dataset consisted of 100 trajectories. The offline partition was chosen to be the state-action pairs with occupancy at least $1/SA$, and the online partition was defined as its complement. In Figure 1, we plot the full and partial single-policy concentrability coefficients from running UCBVI on each partition and for each behavior policy. Between this and Figure 3 in Appendix F, which depicts the cumulative visits to each partition, we see that when the behavior policy is poor or middling, hybrid RL explores more of the online partition to fill in the gaps in the offline dataset than online RL does. However, when the behavior policy is optimal, hybrid RL sticks to the online partition due to the warm-started model estimation.

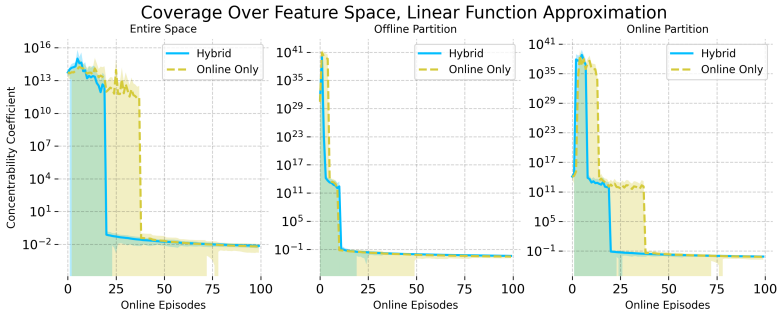

Figure 2: Plot of the full and partial all-policy concentrability coefficients of the online samples from 100 online episodes. The solid line represents the mean over 30 trials, and the shaded areas represent confidence intervals generated by 1.96 times the sample standard deviation. We see that hybrid RL takes fewer online episodes than online-only RL to achieve a lower concentrability coefficient.

7.2 Tetris, Linear MDP.

In another experiment, we consider a scaled-down version of Tetris with pieces of shape at most 2×2, where the game board has a width of 6. The agent can take four actions, corresponding to the degree of rotation in 90 degree intervals, at each timestep. The reward is the negative of any additional increase in the height of the stack beyond 2. We examine the extent to which an optimistic RL algorithm, LSVI-UCB from Jin et al. (2020), explores the feature space more effectively when initialized with an offline dataset of 200 trajectories of length 40 from a uniform behavior policy.

Due to combinatorial blowup, this environment is rather difficult to explore. We therefore chose to focus on the portion of the environment that was covered by the uniform behavior policy within the 8000 simulated timesteps in the offline dataset. This was accomplished through projecting the 640-dimensional one-hot state-action encoding into a 60-dimensional subspace estimated through performing SVD on the offline dataset. The offline partition was chosen to be the span of the top 5 eigenvectors, while the online partition was the span of the remaining 55. Without the projection, the results are qualitatively similar to what we have observed, except with concentrability coefficients that are orders of magnitudes higher.

In Figure 2, we plot the all-policy concentrability coefficients from $n = 1, ..., N_{\text{on}}$, given by the largest, k-th largest, and $d - k$-th largest eigenvalues of the data covariance matrix and its projections onto the offline and online partitions respectively. We see that the concentrability coefficients on the entire space, as well as the offline and online partitions, decrease much faster with the hybrid algorithm than that of the online-only algorithm. This further confirms that an online algorithm initialized with a precollected offline dataset can explore more effectively.

8 Conclusion and Discussion

We have answered through theoretical results and numerical simulations that simply appending the offline dataset to the experience replay buffer can (1) lead to an improvement when the offline dataset is of poor quality, and (2) encourage sufficient exploration for the portion of the state-action space without good coverage. This yields a general recipe for modifying existing online algorithms to incorporate offline data, and we propose DISC-GOLF, a modification of an existing optimistic online algorithm, as an example, with promising theoretical guarantees demonstrating provable gains over both offline-only and online-only learning.

Limitations and Future Work. Due to our desire to work with the simple procedure of appending the offline dataset to the experience replay buffer with general function approximation, our regret bound depends on partial all-policy concentrability. This is not bad, as the best partial

[9] All code can be found at https://github.com/hetankevin/hybridcov.

all-policy concentrability coefficient is always finite (as we can always take $\mathcal{X}_{\text{off}} = \emptyset$) even when the single-policy concentrability coefficient is unbounded. Still, improving this to a guarantee based on partial single-policy concentrability would be valuable. Potential approaches include the clipped single policy concentrability coefficient of Amortila et al. (2024) and the analysis in Theorem 3.1 of Xie et al. (2023). In particular, it is possible that instantiating the result in Theorem F.6 of Amortila et al. (2024) in our setting will lad to a similar tradeoff between the error on analogues of the offline and online partitions discussed in our analysis, though we leave such an approach to future work.

As GOLF, and therefore DISC-GOLF, uses the squared Bellman error, we (1) require completeness (Xie et al., 2022a), and (2) incur a total H^4 dependence before any additional penalties from the log-covering number of the function class.[10] We and Xie et al. (2022a) use this instead of the average Bellman error to facilitate change-of-measure arguments. If one could work with the average Bellman error without a change-of-measure, one could potentially reduce the dependence to H^3 while only requiring realizability, but it is not clear whether this can be accomplished.

Practical and computationally tractable adaptations of DISC-GOLF can be developed in the same sense as (Cheng et al., 2022; Nakamoto et al., 2023), including approaches to optimism in deep RL such as the optimistic actor-critic of Ciosek et al. (2019). One could extend the theoretical analyses in this paper to practical algorithms in deep RL.

Hybrid RL poses a unique opportunity to bypass the pitfalls of offline reinforcement learning. We address the issue of coverage in this work, but strategically collected online data may also help to solve other pertinent issues in offline RL such as distribution shift (Song et al., 2023; Cheng et al., 2022; Kumar et al., 2020), or confounding and partial observability (Wang et al., 2020; Kausik et al., 2023; Bruns-Smith & Zhou, 2023; Lu et al., 2023).

Finally, while DISC-GOLF uses optimistic online exploration, previous work and our general recipe in Proposition 4 shows it is possible to be pessimistic (Nakamoto et al., 2023), or neither (Song et al., 2023). We conjecture that in the presence of single-policy concentrability, or any other situation where the agent does not need to explore any unseen actions online beyond what was already observed in the collected dataset, as in Song et al. (2023); Nakamoto et al. (2023); Amortila et al. (2024), exploration during online learning, and therefore optimism, is not necessary. Otherwise, optimism can be helpful in aiding exploration. Further analysis on the relative merits of each, or even switching between them as Moskovitz et al. (2022) do, is welcomed.

Acknowledgments

We thank Susan Murphy for helpful discussion and comments on the manuscript. We acknowledge support from the Murphy lab through grants NIH/NIDA P50DA054039, NIH/NIDCR, UH3DE028723, NIH/NIBIB and OD P41EB028242.

References

Philip Amortila, Dylan J. Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning, 2024.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3(null):397–422, mar 2003. ISSN 1532-4435.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning, 2017.

David Bruns-Smith and Angela Zhou. Robust fitted-q-evaluation and iteration under sequentially exogenous unobserved confounders, 2023.

[10] Xie et al. (2022a) work with Q-functions bounded in $[0, 1]$ instead of $[0, H]$, so their bound depends on H^2.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning, 2022.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor-critic, 2019.

Steven Cordwell, Yasser Gonzales, and Theja. pymdptoolbox. https://github.com/sawcordwell/pymdptoolbox, 2015.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. On oracle-efficient pac rl with rich observations, 2019.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient rl with rich observations via latent state decoding, 2021.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms, 2021.

Chinmaya Kausik, Yangyi Lu, Kevin Tan, Maggie Makar, Yixin Wang, and Ambuj Tewari. Offline policy evaluation and optimization under confounding, 2023.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning, 2020.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis, 2023a.

Gen Li, Wenhao Zhan, Jason D Lee, Yuejie Chi, and Yuxin Chen. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. *arXiv preprint arXiv:2305.10282*, 2023b.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration, 2020.

Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes, 2023.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning, 2019.

Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano, Michael Arbel, and Michael I. Jordan. Tactical optimism and pessimism for deep reinforcement learning, 2022.

Mitsuhiko Nakamoto, Yuexiang Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning, 2023.

Nived Rajaraman, Lin F. Yang, Jiantao Jiao, and Kannan Ramachandran. Toward the fundamental limits of imitation learning, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism, 2023.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity, 2022.

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, 2023.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage, 2023.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning, 2023.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data, 2020.

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, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022a.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning, 2022b.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning, 2023.

Andrea Zanette. When is realizability sufficient for off-policy reinforcement learning?, 2023.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes, 2021.

Hanlin Zhu, Paria Rashidinejad, and Jiantao Jiao. Importance weighted actor-critic for optimal conservative offline reinforcement learning. 2023.