What can online reinforcement learning with function approximation benefit from general coverage conditions?

Fanghui Liu [1] Luca Viano [1] Volkan Cevher [1]

Abstract

In online reinforcement learning (RL), instead of employing standard structural assumptions on Markov decision processes (MDPs), using a certain coverage condition (original from offline RL) is enough to ensure sample-efficient guarantees (Xie et al., 2023). In this work, we focus on this new direction by digging more possible and general coverage conditions, and study the potential and the utility of them in efficient online RL. We identify more concepts, including the L^p variant of concentrability, the density ratio realizability, and trade-off on the partial/rest coverage condition, that can be also beneficial to sample-efficient online RL, achieving improved regret bound. Furthermore, if exploratory offline data are used, under our coverage conditions, both *statistically* and *computationally* efficient guarantees can be achieved for online RL. Besides, even though the MDP structure is given, e.g., linear MDP, we elucidate that, good coverage conditions are still beneficial to obtain faster regret bound beyond $\widetilde{\mathcal{O}}(\sqrt{T})$ and even a *logarithmic* order regret. These results provide a good justification for the usage of general coverage conditions in efficient online RL.

1. Introduction

Modern reinforcement learning (RL) algorithms modeled by Markov Decision Processes (MDPs) (Szepesvári, 2010), e.g., deep Q network (Mnih et al., 2015), Go (Silver et al., 2016), often work in an online setting under large (or even infinite) state space and action space. Here the terminology *online* means that the agent repeatedly interacts with the environment by executing a policy and observing the past trajectory. To tackle the large state/action space setting, function approximation (Sutton et al., 1999; Jin et al., 2020) is a powerful and indispensable technique in both theory and practice to approximate the true value function from a pre-given function class.

In online RL, much efforts are devoted to developing *sample-efficient* algorithms in a *general* function approximation class beyond linear MDP (Jin et al., 2020). By explicitly assuming some structural assumptions on MDPs, e.g., Eluder dimension (Russo and Van Roy, 2013), Bellman Eluder (BE) dimension (Jin et al., 2021a), typical algorithms including GOLF (Jin et al., 2021a), OPERA (Chen et al., 2022) enjoy sample efficient guarantees in online RL. Normally, these algorithms for general function approximation in online RL are not *computation-efficient* due to the constructed "global" confidence sets for exploration.

Instead of *explicitly* assuming structural assumptions as above-mentioned, the data coverage condition (Munos and Szepesvári, 2008), widely used in offline RL, in fact *implicitly* imposes structural assumptions on the MDP dynamics (Chen and Jiang, 2019). It asserts that a pre-given (even unknown) data distribution μ provides sufficient coverage over the state space. Recently, Xie et al. (2023) show that, even though no offline data are used, a good data coverage condition, w.r.t an underlying distribution, can ensure sample-efficient guarantees in online RL.

Accordingly, studying the coverage condition instead of classical structural assumption on MDPs in online RL is an alternative but promising way. This direction provides a natural connection between offline and online RL in both theory (Wang et al., 2021a; Foster et al., 2021; Zanette, 2021) and practice (Nair et al., 2020; Levine et al., 2020). Besides, it provides a new view to develop *sample-efficient* and even *computation-efficient* algorithms for general function approximation in online RL, as suggested by (Song et al., 2022).

In this work, we focus on this new direction by digging more general coverage conditions, and study the potential and the utility of them in efficient online RL under various scenarios as below.

[1] Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Correspondence to: Fanghui Liu <fanghui.liu@epfl.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

As a starting point, in Section 3, we identify more concepts of coverage conditions to ensure sample-efficient online RL, including the L^p variant of concentrability and the density ratio realizability. We take the $L^p(\mathrm{d}\mu)$ space with $p \geqslant 1$ measure as an example to combine them. Our general coverage conditions can ensure the typical GOLF algorithm (Jin et al., 2021a) to achieve sample-efficient guarantees in online RL. To further obtain computation-efficient efficiency, the offline data can be used for exploration, see a typical hybrid-Q algorithm (Song et al., 2022). Under this setting, our coverage condition is still useful to ensure both *statistically* and *computationally* efficient guarantees for online RL with general function approximation. By doing so, the required structure assumption in (Song et al., 2022), e.g., Bellman rank, BE dimension, can be substituted by our coverage condition. This utility of coverage conditions supports our target in this work.

In Section 4, based on our coverage condition, we decouple the all-policy coverage condition into a partial-policy coverage condition by some (unknown) data distribution and the rest-policy coverage condition, which is quite realistic in practice. We theoretically prove that the trade-off on the partial/rest coverage condition, are able to obtain a better regret bound than (Xie et al., 2023). This provides a good justification on the study of general coverage conditions.

In Section 5, we also identify that, even if the MDP structure is given, e.g., linear MDP, our coverage conditions are still useful. We demonstrate that, the typical LSVI-UCB algorithm in linear MDP (Jin et al., 2020) equipped with certain coverage conditions is able to obtain faster regret bound than $\mathcal{O}(\sqrt{T})$, and even $\mathcal{O}(\log T)$ regret.

Technical contributions: In this paper, we give an affirmative answer to identify more general coverage concepts for improved efficient online RL under several scenarios. We follow the proof framework of (Xie et al., 2023) on the regret analysis and the decomposition of the on-policy average Bellman error. The technical contributions of this work mainly lie in 1) under this framework, how to tackle the unbounded on-policy measure under the setting of partial/rest coverage trade-off in Section 4; 2) providing a new proof framework for linear MDP by building the connection between the on-policy measure and the underlying distribution for improved regret bounds in Section 5.

Goal of this paper: This paper does not contribute to design a new algorithm but provides a possibility to substitute structural assumptions by coverage conditions. We identify more general coverage conditions and dig several good examples for improved efficient online RL. Our analysis sheds light on the utility of coverage conditions in online RL, which could open the door to design new efficient algorithms from offline to online RL in practice motivated by our theoretical results. It bridges the study of offline and online RL and is important for the study of hybrid RL.

In fact, coverage condition is an intrinsic structural property of MDPs that describes the complexity of probability transitions, which does not involve additional information when compared to structural assumptions on MDP. Nevertheless, we do not claim that coverage conditions are better than certain structural assumptions on MDPs in online RL. The relationship between them requires a refined analysis but is beyond the scope of this work.

2. Preliminaries and related work

We start with introducing basic concepts of online and offline RL (Sutton and Barto, 2018), and then give an overview of function approximation in RL.

Notations: We use $[T]$ as a shorthand of $\{1, 2, \ldots, T\}$ for any positive integer T. Define the Lebesgue space $L^p(\mathbb{R}^d)$ with its norm $\|f\|_{L^p} = \int_{\mathbb{R}^d} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, and the $L^p(\mathrm{d}\mu)$ space with its norm $\|f\|^p_{L^p(\mathrm{d}\mu)} = \int [f(\boldsymbol{x})]^p \mathrm{d}\mu$ over the probability measure μ. Here we assume $p \geqslant 1$. A typical example is the $L^2(\mathrm{d}\mu)$ space, a Hilbert space, that is commonly used in learning theory. The notation $\widetilde{\mathcal{O}}$ omits the logarithmic factor.

2.1. Basic concepts

Markov decision processes (MDPs): In our work, we consider a finite-horizon episodic MDP, denoted as $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ setting to model reinforcement learning, where $\mathcal{S}$ is the state space with potentially infinite states; $\mathcal{A}$ is the finite action space; H is the number of steps in one episode; $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ is defined as the transition probability $\mathbb{P}_h(s_{h+1}|s_h, a_h)$ from the current state-action pair (s_h, a_h) to the next state $s_{h+1} \in \mathcal{S}$ for every $h \in [H]$; We use $r := \{r_h\}_{h=1}^H$ to denote the reward $r_h(s, a)$ received at each $h \in [H]$ when taking the action a at state s. For ease of description, we assume the reward is non-negative and $\sum_{h=1}^H r_h(s_h, a_h) \in [0, 1]$ for any possible trajectory.

A non-stationary policy π is a sequence of functions $\pi := \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=1}^H$, where π_h specifies a strategy at step h, and induces a distribution over trajectories $\{(s_h, a_h, r_h)\}_{h=1}^H$ by the following process: taking an action $a_h \sim \pi(\cdot|s_h)$, observing a reward $r_h(s_h, a_h)$, and obtaining $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$. We denote $\mathbb{E}_\pi[\cdot]$ as the expectation w.r.t the randomness of the trajectory $\{(s_h, a_h)\}_{h=1}^H$ generated by the policy π, and $\Pr^\pi[\cdot]$ as the probability under this process. Accordingly, the occupancy measure for a policy π is defined as

$$\rho_h^\pi(s, a) := \Pr^\pi[s_h = s, a_h = a], \quad \rho_h^\pi(s) := \Pr^\pi[s_h = s] \,.$$

The performance of the agent is captured by the *value function*. To be specific, given a policy π, the (state) value function $V_h^\pi \colon \mathcal{S} \to [0,1]$ is defined as the expected cumulative rewards of the MDP starting from step $h \in [H]$

$$V_h^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \big| s_h = s\right].$$

Similarly, the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ for a policy π is defined as

$$Q_h^\pi(s,a) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\big|\, s_h = s,\, a_h = a\right].$$

Since the episode length and the size of action space are both finite, there always exists an optimal policy $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ (Puterman, 2014) such that $V_h^{\pi^\star}(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$. For notational simplicity, we abbreviate $V_h^{\pi^\star}$ as $V_h^\star$ and $Q_h^{\pi^\star}$ as $Q_h^\star$. For a sequence of value functions $\{Q_h\}_{h=1}^H$, the Bellman operator at step h for a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is

$$(\mathcal{T}_h f)(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}[\max_{a' \in \mathcal{A}} f(s', a')].$$

We denote $f_h - \mathcal{T}_h f_{h+1}$ as the Bellman error (or Bellman residual).

The target of an RL algorithm is to find an ϵ-optimal policy such that $V_1^\star(s_1) - V_1^\pi(s_1) \leqslant \epsilon$. In **online RL**, suppose that an agent interacts with the environment for T episodes, the goal is to learn the optimal policy $\pi^\star$ by minimizing the cumulative regret

$$\texttt{Regret}(T) := \sum_{t=1}^{T} [V_1^\star(s_1) - V_1^{\pi^t}(s_1)].$$

In **offline RL**, the agent cannot interact with the environment. Instead, at each step h, what we have is an offline dataset with n_{off} samples $\{(s_h, a_h, r_h, s_{h+1})\}$: sampling $(s_h, a_h) \overset{iid}{\sim} \mu_h$, receiving the reward $r_h(s,a)$, and $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$, where offline data distributions are defined by a collection of data distribution $\mu := \{\mu_h\}_{h=1}^H$. The goal of offline RL is to use this offline dataset to learn an ϵ-optimal policy.

Function approximation: The target of function approximation in RL is to get rid of the size of the state space. A typical setting is under the value-based function approximation, where we approximate the value functions for the underlying MDP by a pre-given function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ with $\mathcal{F}_h \subset \{f \in \mathcal{S} \times \mathcal{A} \to [0,1]\}$. One can see that, a basic assumption in function approximation is to describe the size of $\mathcal{F}$, which aims to assert that $\mathcal{F}$ is large enough or complete to cover value functions under transition dynamics. For notational simplicity, we define $f := \{f_h\}_{h=1}^H$ and accordingly π^f to be the greedy policy w.r.t., f, which takes the action as $\pi_h^f(s) = \operatorname{argmax}_{a \in \mathcal{A}} f_h(s,a)$. Since no reward is collected at the $(H+1)$-th step, we always set $f_{H+1} = 0$.

For each episode t, we define the Bellman error $\delta_h^{(t)}(\cdot,\cdot) := f_h^{(t)}(\cdot,\cdot) - (\mathcal{T}_h f_{h+1}^{(t)})(\cdot,\cdot)$ at step h induced by $f^{(t)} \in \mathcal{F}$.

2.2. Related works on function approximation

Here we give an overview of recent works in function approximation under online RL, offline RL, and a hybrid setting, respectively.

Online RL: The agent under the online setting requires *exploration* schemes when interacting with the unknown environment. The simplest scheme is ϵ-greedy, i.e., randomly selecting new actions with ϵ probability. Though computational efficient, this scheme is demonstrated to be statistically inefficient in theory (Jin et al., 2015; Dann et al., 2022; Liu et al., 2022). Most literature work with "optimism in the face of uncertainty" principle for efficient exploration schemes, e.g., upper confidence bound (UCB)-type algorithms (Jin et al., 2020) and Thompson sampling (Russo et al., 2018; Agrawal and Goyal, 2012). They have been applied to linear MDP (Jin et al., 2020; Yang and Wang, 2020), kernel MDP (Yang et al., 2020), linear mixture MDP (Ayoub et al., 2020; Zhou et al., 2021). For general function approximation, under proper assumptions on MDP structure, e.g., Bellman rank (Jiang et al., 2017), Eluder dimension (Russo and Van Roy, 2013), Bilinear rank (Du et al., 2021), BE dimension (Jin et al., 2021a), admissible Bellman characterization class (Chen et al., 2022), decision-estimation coefficient class (Foster et al., 2021), and sequential exploration coefficient (Xie et al., 2023), sample-efficient algorithms based on optimistic principles are designed to ensure statistical efficiency but computational efficiency guarantees are often unattainable.

Offline RL: The agent under the offline RL setting (Levine et al., 2020) does not interact with the environment and just learns policies solely from a given offline dataset. Hence there is no possibility to do *exploration* but a *data coverage* condition over the offline dataset is required for statistical guarantees. It requires the dataset to contain any possible state, action pair or trajectory with a lower bounded probability. A typical example is all-policy concentrability (Munos and Szepesvári, 2008; Zhang et al., 2020), which requires the sufficient coverage of offline data over all (relevant) states and actions. Recent works focus on relaxation from such strong coverage condition to partial coverage (Uehara and Sun, 2022), and even single-policy concentrability (Rashidinejad et al., 2021; Zhan et al., 2022) by preventing the policy from visit states and actions where the offline data coverage is poor (Liu et al., 2020) or relying on the principle of "pessimism" (Xie et al., 2021a; Jin et al., 2021b).

Online RL with offline data: Empirical results work in this setting and have demonstrated the success of offline data (Rajeswaran et al., 2017), but under certain settings, offline data does not yield statistical improvements in tabular MDPs (Xie et al., 2021b). Recent work focus on digging the benefit of offline data in online RL, including computation efficiency (Song et al., 2022) and sample efficiency (Wagenmaker and Pacchiano, 2022).

Besides, Xie et al. (2023) demonstrate that, the data coverage condition is able to ensure sample-efficiency in online RL though no offline data is required to be accessed. This provides a bridge between the analysis techniques of offline and online RL.

2.3. Coverage conditions

Here we briefly introduce mathematical concepts of data coverage conditions.

A concept crucial to our discussions is the marginalized importance weights, which aims to measure the distribution shift from an arbitrary distribution (here we use the occupancy measure by any policy π) $\rho^\pi := \{\rho_h^\pi\}_{h=1}^H$ to the data distribution $\mu := \{\mu_h\}_{h=1}^H$. Define $w_{h,\pi/\mu}(s,a) := \frac{\rho_h^\pi(s,a)}{\mu_h(s,a)}$ if $\mu_h(s,a) \neq 0$, and then the commonly used concentrability coefficient for all policy in a policy class Π (Munos and Szepesvári, 2008; Chen and Jiang, 2019)

$$C_\infty := \max_{\pi\in\Pi, h\in[H]} \|w_{h,\pi/\mu}\|_\infty \leqslant \|w_{h,\pi/\mu}\|^2_{L^2(\mathrm{d}\mu)}\,,$$

where the $L^2(\mathrm{d}\mu)$ version is developed in (Xie and Jiang, 2020). For single-policy concentrability, only $\pi^\star$ instead of all possible $\pi \in \Pi$ is taken part in these concentrability coefficients (Uehara and Sun, 2022).

Concentrability coefficients can be also conducted by Bellman error, e.g., (Xie et al., 2021a). Here we give an example from (Song et al., 2022) by denoting $\delta_h := f_h - \mathcal{T}_h f_{h+1}$ such that

$$C_\pi := \max_{f\in\mathcal{F}} \frac{|[\mathbb{E}_{\rho_h^\pi}\delta_h(s,a)]|}{\sqrt{\mathbb{E}_{\mu_h}[\delta_h(s,a)]^2}}\,, \forall \pi \in \Pi\,, \qquad (1)$$

which can be upper bounded by the coverability coefficient C_∞. Recently another coverability coefficient is defined as below to ensure sample-efficient exploration in online RL.

Definition 1. (Xie et al., 2023, Coverability for online RL) The coverability coefficient C_{cov} is for a policy class Π

$$\begin{aligned} C_{\mathsf{cov}} &:= \inf_{\mu_1,\dots,\mu_H\in\Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\Pi, h\in[H]} \left\| \frac{\rho_h^\pi}{\mu_h} \right\|_\infty \\ &= \max_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sup_{\pi\in\Pi} \rho_h^\pi(s,a)\,. \end{aligned}$$

It is demonstrated to be equivalent to the cumulative reachability (see the second equality), refer to (Xie et al., 2023, Lemma 3) for details.

Besides, in online RL, the "uniformly excited feature" assumption (Abbasi-Yadkori et al., 2019; Lazic et al., 2020; Hao et al., 2021) is commonly used in reinforcement learning theory. It requires that every occupancy measure induced by a policy π yields a positive definite feature covariance matrix such that $\mathbb{E}_{\rho_h^\pi}[\phi(s,a)\phi(s,a)^\top] \succcurlyeq c\boldsymbol{I}$ for some constant $c > 0$ where $\phi(s,a)$ is the corresponding feature mapping. By doing so, each policy $\{\pi_h\}_{h=1}^H$ explores uniformly well in the feature space. This assumption is also used in offline RL but the expectation is taken as a data distribution μ, see feature coverage condition in (Wang et al., 2021a, Assumption 2).

2.4. Basic assumptions

Our work focuses on general function approximation in online RL, which is based on the following two standard and commonly-used assumptions in reinforcement learning theory (Wang et al., 2020; Jin et al., 2021a; Chen et al., 2022; Xie et al., 2023).

Assumption 1 (Realizability)**.** For a hypothesis class $\mathcal{F}$, we assume $Q_h^\star \in \mathcal{F}_h$ for any $h \in [H]$.

Define $\mathcal{T}_h\mathcal{F}_{h+1}$ as $\{\mathcal{T}_h f_{h+1} : f_{h+1} \in \mathcal{F}_{h+1}\}$, we require the function class $\mathcal{F}$ to be closed under the Bellman operator $\mathcal{T}_h$ as below.

Assumption 2 (Bellman completeness)**.** For a hypothesis class $\mathcal{F}$, we assume $\mathcal{T}_h\mathcal{F}_{h+1} \in \mathcal{F}_h$ for any $h \in [H]$.

If the function class $\mathcal{F}$ has finite elements, we can directly use its cardinality to measure its "size". If $\mathcal{F}$ has infinite elements, the covering number is needed to describe the "size" of $\mathcal{F}$.

Definition 2 (Covering number (Van Der Vaart et al., 1996))**.** The ϵ-covering number $\mathscr{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ for a function class $\mathcal{F}$ with respect to the metric $\|\cdot\|_\infty$ is the minimal number of balls with radius ϵ measured by $\|\cdot\|_\infty$-norm needed to cover the space $\mathcal{F}$. For short, we denote $\mathscr{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ as $\mathscr{N}_{\mathcal{F}}(\epsilon)$ by omitting $\|\cdot\|_\infty$.

3. Warm-up: Coverage conditions in L^p spaces

We give the definition of coverability coefficient in the L^p space, which covers the L^p variant of concentrability and the density ratio realizability. In Section 3.1, we demonstrate that, these coverage conditions are able to obtain better regret bound for sample-efficient online RL with general function approximation when compared to (Xie et al., 2023). Furthermore, under our coverage conditions, computational efficiency can be even achieved if exploratory offline data are used in Section 3.2.

3.1. Improved sample-efficient online RL

Definition 3 (L^p coverability coefficient)**.** Given a policy class Π, there exists a underlying distribution $\mu = \{\mu_h\}_{h=1}^H$ admitting $\sum_{(s,a)} \sqrt{\mu_h(s,a)} < \infty$, for any $p \geqslant 1$, the coverability coefficient $C_{\mathtt{cw}}$ defined in the L^p space is given by

$$C_{\mathtt{cw}} := \inf_{\mu_1,\dots,\mu_H \in \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\Pi, h\in[H]} \left\| \frac{\rho_h^\pi}{\mu_h} \right\|_{L^p(\mathrm{d}\mu_h)}^p .$$

Remark: This definition simply extends the application scope of $C_{\mathtt{cov}}$ from the L^∞ space to the L^p space. One interesting thing is, we only require $\sum_{(s,a)} \sqrt{\mu_h(s,a)} < \infty$ rather than $\sum_{(s,a)} [\mu_h(s,a)]^{1/p} < \infty$, which makes the underlying distribution μ more general.

It is clear that $C_{\mathtt{cw}} \leqslant |\mathcal{S}||\mathcal{A}|$ if we take μ is a uniform measure. The relationship between $C_{\mathtt{cw}}$ and $C_{\mathtt{cov}}$ can be built by the following lemma, deferred the proof to Appendix B.1.

Lemma 1. *Based on the definition of $C_{\mathtt{cw}}$ and $C_{\mathtt{cov}}$ in Definition 3 and Definition 1, respectively, we have*

$$C_{\mathtt{cw}}^{\frac{1}{p}} \leqslant C_{\mathtt{cov}}\,, \forall p \geqslant 1\,.$$

Lemma 1 can be used for demonstrating a better regret bound in online RL when compared to that of $C_{\mathtt{cov}}$ as below.

We take the GOLF algorithm (Jin et al., 2021a) as an example to demonstrate the sample-efficient guarantees of online RL. For self-completeness, we give a brief description on the GOLF algorithm (Jin et al., 2021a) in Algorithm 1, see Appendix A. This is a typical general function approximation algorithm in online RL, and yields sample-efficient guarantees if the BE dimension is small. Here we show that, under our coverage condition $C_{\mathtt{cw}}$, we can still achieve the sample-efficient guarantees for online RL, with the proof deferred to Appendix B.2.

Proposition 1. *Under Assumptions 1 and 2, there exists a constant c and the data coverage coefficient $C_{\mathtt{cw}}$ in Definition 3 such that for any $\delta \in (0,1)$, if we choose $\beta = c\log\left(\frac{\mathscr{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ in the GOLF algorithm 1, with probability at least $1-\delta$, we have*

$$\mathtt{Regret}(T) \lesssim \mathcal{O}\left(H\sqrt{C_{\mathtt{cw}}^{\frac{1}{p}} \beta T \log T} \right) .$$

Remark: We obtain a better regret bound than (Xie et al., 2023, Theorem 1) due to an improved data coverage coefficient in Lemma 1.

Our result in Proposition 1 demonstrates that if the coverage coefficient $C_{\mathtt{cw}}$ is small, the GOLF algorithm can achieve sublinear regret for sample-efficient guarantees without requiring the structure assumption of MDP. This is because, the coverage condition in fact implicitly imposes some structural assumptions on the MDP dynamics, see (Chen and Jiang, 2019, Theorem 4) for details. It is an intrinsic structural property of MDPs that describes the complexity of probability transitions. This shares a similar spirit with the sub-optimality gap (He et al., 2021) on describing the complexity of MDPs under probability transitions. Nevertheless, the condition of the sub-optimality gap is stronger because the reward feedback is also considered.

There appears a natural question on the relationship between coverage conditions and structural assumptions. Since coverage conditions do not involve additional information, they are often weaker than structural assumptions. For example, Sequential Exploration Coefficient (SEC) (Xie et al., 2023), as a structural assumption, is a general version of coverage condition, which admits

$$\mathtt{SEC} \lesssim C_{\mathtt{cw}}^{\frac{1}{p}} \log T \leqslant C_{\mathtt{cov}} \log T\,.$$

Apart from this, the relationship between various structural assumptions and coverage conditions requires a refined analysis but is beyond the scope of this work.

Nevertheless, coverage conditions are still more general than linear MDP (Jin et al., 2020). For example, in the Atari game, the state space (raw pixels) can be very large, but the dynamics is determined by a small number of unobserved latent states. This can be described as block MDP (Du et al., 2019), and accordingly the coverability coefficient can be small as it scales only with the number of latent states instead of the size of the whole state space.

3.2. Efficient online RL with exploratory offline data

As mentioned before, the GOLF algorithm is not computation efficient due to the constructed “global” confidence set. To avoid sophisticated exploration schemes, one typical way is to use offline data for exploration, which is recently popular both empirically (Ball et al., 2023) and theoretically (Song et al., 2022; Wagenmaker and Pacchiano, 2022).

Here we use the hybrid-Q algorithm (Song et al., 2022) to demonstrate the benefit of our data coverage condition when involving with offline data on the computation efficiency. This algorithm is based on the classical fitted Q-iteration (FQI) algorithm (Ernst et al., 2005) and uses offline data regarding the distribution $\nu := \{\nu_h\}_{h=1}^H$ for exploration, and thus the computation complexity of this algorithm is the same as FQI with a least squares regression oracle, refer to Appendix C.1 for details of the hybrid-Q algorithm.

Here we aim to demonstrate that without any structural assumption, the all-policy coverage conditions can ensure efficient online RL, both statistically and computationally if exploratory offline data are used. In the following, we take the all-policy concentrability coefficient C_π in Eq. (1) and our coverage condition $C_{\texttt{cw}}$ in Definition 3 as examples to illustrate this, with the proof deferred to Appendix C.2.

Proposition 2. *Under Assumptions 1 and 2, then for any $\delta \in (0,1)$, $T \in \mathbb{N}$, if we choose $n_{\text{off}} = T$ in Algorithm 2 and denote $\beta := \log\left(\frac{\mathcal{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$, with probability at least $1-\delta$,*

Case 1. *under the all-policy concentrability coefficient C_π in Eq. (1), we have*

$$\texttt{Regret} \lesssim \mathcal{O}\left(C_\pi H \sqrt{\beta T}\right) .$$

Case 2. *there exists a data distribution $\nu := \{\nu_h\}_{h=1}^H$ that provides a single-policy concentrability, $\max_{s,a,h} \frac{\mu_h^\star(s,a)}{\nu_h^2(s,a)} < \widetilde{C}$, where $\mu_h^\star(s,a)$ realizes the value of the coverage coefficient $C_{\texttt{cw}}$ endowed by $L^p(\mathrm{d}\mu)$ norm with $p \geqslant 1$ in Definition 3, we have*

$$\texttt{Regret}(T) \lesssim \mathcal{O}\left(C_{\texttt{cw}}^{\frac{1}{p}} H \sqrt{\beta \widetilde{C} T \log T}\right) .$$

Remark: We make the following remarks:
i): Song et al. (2022) achieve the regret bound $\widetilde{\mathcal{O}}(C_{\pi^\star} H \sqrt{d\beta T})$, where the single-policy concentrability coefficient $C_{\pi^\star}$ is defined in Eq. (1), and d is the Bilinear rank or BE dimension. Instead, in **Case 1**, the structure assumptions on MDP are not needed to ensure the same regret if the all-policy concentrability coefficient C_π is employed.
ii): In **Case 2**, if the all-policy concentrability coefficient $C_{\texttt{cw}}^{\frac{1}{p}}$ is used, an extra single-policy concentrability coefficient $\widetilde{C}$ is needed. As a single-policy version, it is often smaller than $C_{\texttt{cw}}^{\frac{1}{p}}$, and can be even a constant if we take ν to match $\mu^\star$. In this case, the $\widetilde{O}(\sqrt{T})$-regret can be still achieved without structural assumptions on MDP.

$$\begin{cases} \text{(Song et al., 2022)} \begin{cases} \text{single-policy coefficient } C_{\pi^\star} \\ \text{structural assumptions} \end{cases} \\ \textbf{Case 1}\text{: all-policy coefficient } C_\pi \\ \textbf{Case 2}\text{: } \begin{cases} \text{all-policy coefficient } C_{\texttt{cw}}^{\frac{1}{p}} \\ \text{single-policy coefficient } \widetilde{C} \end{cases} \end{cases}$$

Proofs techniques: To prove Propositions 1 and 2, we follow the proof framework in (Xie et al., 2023, Theorem 1) on the regret analysis and the decomposition of the on-policy average Bellman error. In Proposition 1, the difference lies in how to estimate the occupancy measure ratio by different coverage conditions. Further, in Proposition 2, since no exploration scheme is used, we need to build the connection between $\rho_h^{(t)}$ and ν_h by coverage conditions, which is used for the estimation of the in-sample squared Bellman error.

The results in this warm-up section provide a good justification of the usage of general coverage conditions for efficient online RL. This will motivate us to study partial/rest coverage trade-off and coverage conditions in linear MDP presented in the next two sections.

4. Partial/rest coverage trade-off

As we know, partial coverage or even single coverage conditions are more realistic in practice, and widely studied in offline RL (Xie et al., 2021a; Jin et al., 2021b; Zhan et al., 2022). However, Xie et al. (2023) point out that $C_{\texttt{cov}}$ under a single-policy coverage can not ensure sample-efficient online RL. Accordingly, in this section, based on our L^p coverage concepts in Section 3, we decouple the all-policy coverage condition into a partial-policy coverage condition by some underlying distribution and the rest-policy coverage condition, which is more realistic in practice. Under this setting, we aim to diagnose the effect of partial/rest coverage condition on the regret bound.

4.1. Definition of partial/rest coverage condition

Here we define the partial/rest coverage condition and then study the statistical guarantees of online RL algorithms.

Definition of partial policy class: Motivated by $C_{\texttt{cov}}$ in Definition 1 that can be regarded as a cumulative area over all possible ρ_h^π, we consider a possible policy class by evaluating how a policy is close to the *reference* policy $\bar{\pi} := \{\bar{\pi}_h\}_{h=1}^H$. A nature metric is the total variation (TV) distance[1], and accordingly, the candidate policy set $\mathcal{M} = \{\mathcal{M}_h\}_{h=1}^H$ is defined as

$$\mathcal{M}_h(\zeta) := \{\pi_h : |\mathrm{TV}(\rho_h^{\pi_h}, \rho_h^{\bar{\pi}_h}) \leqslant \zeta\} ,$$

where the reference policy $\bar{\pi}$ can be set to the optimal policy $\pi^\star$ or any possible policy that is controlled by some (unknown) data distribution. We can see that, $\bar{\pi}$ can be a high-quality policy or a low-quality policy, which is more realistic in practice. Clearly we have $\zeta \in [0,2]$ based on the definition of the TV distance. If $\zeta = 0$, we only have single policy concentrability (i.e., only the reference policy) and if $\zeta = 2$, we can recover the whole policy class Π. Hence this policy class $\mathcal{M}_h(\zeta)$ is a partial or incomplete policy class, and then the coverability coefficient defined over this policy class can be denoted as a partial coverage condition, introduced as below.

Definition 4. The partial coverability coefficient $P_{\texttt{cov}}(\zeta)$ is for a (partial) policy class $\mathcal{M}(\zeta)$

$$P_{\texttt{cov}}(\zeta) := \inf_{\mu_1,\dots,\mu_H \in \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi \in \mathcal{M}(\zeta), h \in [H]} \left\| \frac{\rho_h^\pi}{\mu_h} \right\|_\infty .$$

Remark: For notional simplicity, we denote $P_{\texttt{cov}}(\zeta)$, $\mathcal{M}(\zeta)$ by $P_{\texttt{cov}}$, $\mathcal{M}$ for short.

[1] Here the used metric between two distributions can be general, e.g., Wasserstein distance (Rüschendorf, 1985; Fournier and Guillin, 2015).

Denote $\hat{\mu}_h^\star := \operatorname{argmin}_{\mu_h \subseteq \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\mathcal{M}} \left\| \frac{\rho_h^\pi}{\mu_h} \right\|_\infty$, we can easily obtain the equivalent definition $P_{\texttt{cov}} = P_{\texttt{cr}} := \max_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sup_{\pi\in\mathcal{M}} \rho_h^\pi(s,a)$, refer to the proof in Appendix D.1.

Clearly $P_{\texttt{cov}} \geqslant 1$, and the single policy concentrability implies $P_{\texttt{cov}} = 1$ by taking $\mu_h := \rho_h^{\bar{\pi}_h}$. That means, $P_{\texttt{cov}}$ looses the ability to represent the complexity of state transition in MDPs, and thus is insufficient to ensure sample-efficient learning in online RL. In this case, we need to introduce extra conditions that aid for sufficient learning.

Coverage condition outside $\mathcal{M}$: We give the definition of the rest coverage condition related to the policy $\hat{\mu}_h^\star$ over some policies outside $\mathcal{M}$, i.e., its complementary set $\bar{\mathcal{M}}$.

Definition 5. For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, the state-action pair set $\mathcal{B}^{\bar{\mathcal{M}}} := \{\mathcal{B}_h^{\bar{\mathcal{M}}}\}_{h=1}^H$ is denoted as

$$\mathcal{B}_h^{\bar{\mathcal{M}}} := \left\{ (s,a) \mid \rho_h^\pi(s,a) > c_1 P_{\texttt{cov}} \hat{\mu}_h^\star(s,a), \forall \pi \in \bar{\mathcal{M}} \right\} ,$$

with some constant $c_1 \geqslant 1$, then the partial coverage condition outside $\mathcal{M}$ is

$$P_{\texttt{out}}(\zeta) := \max_{h\in[H], \pi\notin\mathcal{M}} \left\| \frac{\rho_h^\pi}{\hat{\mu}_h^\star} \mathbb{1}_{\mathcal{B}_h^{\bar{\mathcal{M}}}} \right\|_{L^2}^{\frac{1}{2}} ,$$

defined in the L^2 space, and the indicator function $\mathbb{1}_{\mathcal{B}_h^{\bar{\mathcal{M}}}} = 1$ if $(s,a) \in \mathcal{B}_h^{\bar{\mathcal{M}}}$, and otherwise is zero.

Remark: This quantity $P_{\texttt{out}}$ defined in the L^2 space is also related to ζ due to $\mathcal{M}(\zeta)$, and we also omit it for notational simplicity.

Clearly $P_{\texttt{out}} \geqslant 0$, and there exists a trade-off between $P_{\texttt{cov}}(\zeta)$ and $P_{\texttt{out}}(\zeta)$ that depends on ζ. If ζ increases, $P_{\texttt{cov}}$ increases but $P_{\texttt{out}}$ decreases. For example, if $\zeta = 2$, we have $P_{\texttt{cov}} = C_{\texttt{cov}}$ and $P_{\texttt{out}} = 0$; if $\zeta = 0$, we have $P_{\texttt{cov}} = 1$. Accordingly, in this case $P_{\texttt{out}}$ is used to measure the structural information of MDPs. Here we explain this a bit.

In our proof (*c.f.* Appendix D.2), we also show that if we take the reference policy $\bar{\pi} := \pi^\star$ and c_1 large enough, e.g., $c_1 = \Omega(H)$, the probability that the optimal policy $\pi^\star$ visits this state-action pair set $\mathcal{B}^{\bar{\mathcal{M}}}$ is very small. That means, $P_{\texttt{out}}$ can be regarded as the distribution shift between a policy π and $\pi^\star$ on some low-probability set. We can see, it describes the ability that an algorithm overcomes the difficult state-action pairs in MDPs, which can be also regarded as an instance-based metric.

4.2. Sublinear regret bound

Based on our definition on $P_{\texttt{cov}}$ and $P_{\texttt{out}}$, we have the following theorem that demonstrates how the partial/rest coverage condition affects the regret bound, with the proof deferred to Appendix D.2.

Theorem 1. *Under Assumptions 1 and 2, there exists a constant c_1, the partial coverage coefficient $P_{\texttt{cov}}$ in Definition 4 and $P_{\texttt{out}}$ in Definition 5, then for any $\delta \in (0,1)$, $T \in \mathbb{N}$, if we choose $\beta = c\log\left(\frac{\mathcal{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ in GOLF, with probability at least $1-\delta$, we have*

$$\texttt{Regret} \lesssim \mathcal{O}\left(H\Big(\sqrt{c_1 P_{\texttt{cov}}} + \frac{P_{\texttt{out}}}{\sqrt{P_{\texttt{cov}}}}\Big)\sqrt{\beta T \log T} \right) .$$

Specifically, there always exists a proper $\zeta^\star \in [0,2]$ such that $P_{\texttt{out}}(\zeta^\star) = \sqrt{c_1}P_{\texttt{cov}}(\zeta^\star)$, the above regret bound can be improved to

$$\texttt{Regret} \lesssim \mathcal{O}\left(H\sqrt{c_1^{1/2}\beta T P_{\texttt{out}}(\zeta^\star)\log T} \right) . \tag{2}$$

which admits $P_{\texttt{out}}(\zeta^\star) \leq C_{\texttt{cov}}$.

Remark: One can choose c_1 to some constant up to H, so we remain c_1 in our bound. We make the following remarks. *i)* If we only consider the single policy in $\mathcal{M}$, which implies $P_{\texttt{cov}} = 1$, and our result is still applicable to ensure sample-efficient learning estimated by $P_{\texttt{out}}$. If we consider the whole policy class such that $P_{\texttt{cov}} = C_{\texttt{cov}}$ and then $P_{\texttt{out}} = 0$, we can recover the result of (Xie et al., 2023).
ii) Clearly, there exists a trade-off between $P_{\texttt{cov}}(\zeta)$ and $P_{\texttt{out}}(\zeta)$ that depends on ζ. That means, there always exists a proper $\zeta^\star$ such that Eq. (2) holds and $P_{\texttt{out}}(\zeta^\star) \leqslant C_{\texttt{cov}}$ by the property of the function $x + c/x$ for some constant c. This demonstrates a better regret bound than (Xie et al., 2023) by a good trade-off between $P_{\texttt{cov}}$ and $P_{\texttt{out}}$.

Proposition 2 extends the application scope of the hybrid-Q algorithm in the view of coverage conditions instead of structural assumptions, which provides a good justification on the study of coverage condition.

Proof sketch of Theorem 1: In our proof, the on-policy average Bellman error can be transformed to the occupancy measure ratio and the in-sample squared Bellman error. The technical difficulty is to control the ratio when the on-policy occupancy measure is unbounded. The in-sample squared Bellman error can be directly estimated by (Jin et al., 2021a). To handle the ratio, we split the on-policy occupancy measure into two cases: 1) $\rho_h^{(t)}(s,a) \leqslant c_1 P_{\texttt{cov}} \hat{\mu}_h^\star(s,a)$ and 2) $(s,a) \in \mathcal{B}^{\bar{\mathcal{M}}}$ which means $\rho_h^{(t)}$ is unbound by some (scaling) probability measure. In the first case, it is upper bounded by $P_{\texttt{cov}}\log T$; In the second case, $\rho_h^{(t)}$ cannot be controlled by previous occupancy measures $\{\rho_h^{(i)}\}_{i=1}^{t-1}$ in terms of Bellman residual. We build the connection between $\rho_h^{(t)}$ and $\hat{\mu}_h^\star$, introduce $P_{\texttt{out}}$ to control such distribution shift, and trade-off $P_{\texttt{cov}}$ and $P_{\texttt{out}}$ for a better regret bound.

5. Coverage conditions help linear MDP

Till now we have already demonstrated that, without explicit structure assumptions on MDP, the new devised coverage conditions are able to ensure sample-efficient online RL in Proposition 1 and Theorem 1, respectively. By general coverage conditions, we are able to achieve better regret bound than (Xie et al., 2023). Here we are also interested in

If the structural assumption is given, what can we still benefit from coverage conditions?

In this section, we take the classical linear MDP using the LSVI-UCB algorithm (Jin et al., 2020) as an example, and demonstrate that a faster regret bound than $\widetilde{\mathcal{O}}(\sqrt{T})$ or even $\mathcal{O}(\log T)$ can be achieved if extra coverage conditions are employed.

For ease of description, we give some notations here. Details about the LSVI-UCB algorithm can be found in Appendix E.1. Denote the feature mapping $\phi(s,a) \in \mathbb{R}^d$ in linear MDP (Jin et al., 2020) satisfying $\|\phi(s,a)\|_2 \leqslant 1$, and Λ_h^t constructed by the standard LSVI-UCB algorithm with the regularization parameter λ, i.e.,

$$\Lambda_h^t := \lambda I + \sum_{i=1}^{t-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top \,. \tag{3}$$

We assume that the underlying data distribution μ satisfies the following condition.

Assumption 3. (Wang et al., 2021a, feature coverage condition) There exists a underlying distribution $\mu := \{\mu_h\}_{h=1}^H$ such that $\lambda_{\min}(\mathbb{E}_\mu[\phi(s,a)\phi(s,a)^\top]) \geqslant \gamma > 0$.

Remark: We make the following remarks.
i): This assumption shares the similar spirit with the "uniformly excited feature" assumption (Abbasi-Yadkori et al., 2019; Papini et al., 2021) but is much weaker than them as they require the minimum eigenvalue lower bounded under *any* occupancy measure. Our assumption only requires the validity under the *single* measure.
ii): This assumption can be easily achieved, e.g., $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] \succ 0$ in statistics for linear feature mapping (Wainwright, 2019). Besides, another typical example is that, the minimum eigenvalue of neural tangent kernel (Jacot et al., 2018) can be lower bounded by a positive constant (Nguyen et al., 2021).

Based on our discussion, we can see Assumption 3 is much weaker than the "uniformly excited feature" assumption and can be easily achieved in practice. That means, this assumption might be not enough to ensure better results for linear MDP. In this case, we need to strength the condition on the underlying distribution μ as below.

Assumption 4 (low variance condition)**.** For the LSVI-UCB algorithm with the empirical covariance matrix Λ_h^t defined in Eq. (23), there exists the underlying distribution $\mu = \{\mu_h\}_{h=1}^H$ with $(s_h, a_h) \sim \mu_h$ such that

$$\mathbb{V}\left[\|\phi_h(s_h,a_h)\|^2_{(\Lambda_h^t)^{-1}}\right] \lesssim \frac{1}{\lambda^{2\alpha}} \,, \quad \alpha > 1 \,.$$

Remark: We make the following remarks.
i) Under the standard linear MDP setting, we always have the bounded random variable $\|\phi_h(s,a)\|^2_{(\Lambda_h^t)^{-1}} \leqslant 1/\lambda$, and thus its variance admits $\mathbb{V}[\|\phi_h(s_h,a_h)\|^2_{(\Lambda_h^t)^{-1}}] \leqslant \frac{1}{4\lambda^2}$. That means, Assumption 4 always holds with $\alpha = 1$ for any distribution.
ii) Our assumption requires $\alpha > 1$, and in fact requires that the data distribution μ make $\|\phi_h(s_h,a_h)\|^2_{(\Lambda_h^t)^{-1}}$ concentrate around its mean, i.e., a low variance condition. It shares similar spirit with (Du et al., 2019; Wang et al., 2021b) that characterizes the anti-concentration of a distribution μ.
iii) We give an example here by denoting $\boldsymbol{x} := (s_h, a_h)$ for short, and the upper/lower bound of $\|\phi_h(\boldsymbol{x})\|^2_{(\Lambda_h^t)^{-1}}$ as M and m. Accordingly, we have

$$\frac{1}{\lambda + t - 1} \leqslant m \leqslant \|\phi_h(\boldsymbol{x})\|^2_{(\Lambda_h^t)^{-1}} \leqslant M \leqslant \frac{1}{\lambda} \,, \tag{4}$$

by the Weyl inequality and $\boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{a} \geqslant \lambda_{\min}(\boldsymbol{A})\|\boldsymbol{a}\|_2^2$ for any PSD matrix $\boldsymbol{A}$. Since Eq. (4) holds for any distribution $\boldsymbol{x} \sim \mu$. There exists some certain distributions μ such that the random variable $\|\phi_h(\boldsymbol{x})\|^2_{(\Lambda_h^t)^{-1}}$ concentrates, i.e., $M - m$ is small. For example, taking $M := 1/\lambda$, $m := 1/\lambda - 1/\lambda^2$ such that $M - m \leq 1/\lambda^2$. That means, under a certain distribution, the feature mapping ϕ_h has the similar (semi)-norm in the $(\Lambda_h^t)^{-1}$-(semi)-norm based space. Then, by Popoviciu's inequality on variances, we have $\mathbb{V}[\|\phi_h(s_h,a_h)\|^2_{(\Lambda_h^t)^{-1}}] \leqslant \frac{1}{4(M-m)^2} \leqslant \frac{1}{4\lambda^4}$, which implies $\alpha = 2$, and thus our assumption holds.

Based on the above two assumptions, we are ready to improve the regret in linear MDP from $\widetilde{\mathcal{O}}(\sqrt{T})$ to faster rate and even in the logarithmic order by the following proposition, with the proof deferred to Appendix E.1.

Theorem 2. *For linear MDP using the LSVI-UCB algorithm, under Assumption 3 with $\gamma > 0$ and Assumption 4 with $\alpha > 1$, taking the regularization parameter $\lambda := T^\eta$ with $\eta \in (0,1]$ and the bonus parameter $\beta = \widetilde{\mathcal{O}}\left(\sqrt{\lambda}H(d + \sqrt{\log\frac{1}{\delta}})\right)$ for any $\delta \in (0,1)$, with probability at least $1 - \delta$, we have*

$$\begin{aligned}\texttt{Regret}(T) &\lesssim \left(\frac{H^2d^2}{\gamma}\log T + \frac{H^2 d\lambda\sigma}{\gamma}\sqrt{T}\right)\log\left(\frac{4}{\delta}\right) \\ &= \begin{cases} \mathcal{O}\left(\dfrac{d^2H^2}{\gamma}\log T\right), & \textit{if } \eta(\alpha-1) \geqslant 1/2 \\ \mathcal{O}\left(\dfrac{dH^2}{\gamma}T^{\frac{1}{2}-\eta(\alpha-1)}\right), & \textit{if } \eta(\alpha-1) \in (0, \frac{1}{2}) \,. \end{cases}\end{aligned}$$

Remark: We make the following remarks.
i) If we take $\alpha = 1$, Assumption 4 always holds. Since Assumption 3 easily holds for a underlying distribution μ, we can recover the $\widetilde{\mathcal{O}}(\sqrt{T})$-regret in (Jin et al., 2020).
ii) If $\eta(\alpha - 1) \geqslant 1/2$, that means, α can be large, the regret enjoys the logarithmic order of T. If $0 < \eta(\alpha - 1) < 1/2$, we have a sublinear $\mathcal{O}(T^{\frac{1}{2}-\eta(\alpha-1)})$ regret, faster than the classical $\mathcal{O}(\sqrt{T})$ regret.
iii) The regularization parameter λ decreases with the increasing T though we use $\lambda := \mathcal{O}(T^{\eta})$ with $\eta \in (0, 1]$. This is because, the "true" regularization parameter is λ/T as we need to scale LSVI with the number of the involved state-action pairs. The regularization parameter decaying with the number of samples is fair and commonly used in learning theory (Cucker and Zhou, 2007). Besides, taking $\eta = 0$ in the regularization parameter λ is able to improve the regret rate under a slight changes of Assumption 4. Detailed discussion can be found in Appendix E.2.
iv) Instance-dependent regret bound has been widely studied for linear MDP with the logarithmic-order regret (He et al., 2021) under the minimal sub-optimality gap (strictly larger than zero) and further improved to the constant regret (Papini et al., 2021). This requires a separation between the optimal action and the rest ones; while our assumptions focus on a "distinct" feature mapping under certain distributions.

Proof sketch: We provide a new proof framework on LSVI-UCB for linear MDPs to achieve faster regret bound. By a telescoping lemma (Jiang, 2022), the regret can be upper bounded by $\|\phi_h(s_h, a_h)\|^2_{(\Lambda_h^t)^{-1}}$ over the on-policy measure $\rho_h^{(t)}$. The key challenge is, if we directly apply change-of-measure: from $\rho_h^{(t)}$ to the underlying distribution μ_h, the elliptical potential lemma is invalid. In this case, in our analysis, we build the connection between $\mathbb{E}_{\rho_h^{(t)}} \|\phi_h(s_h, a_h)\|^2_{(\Lambda_h^t)^{-1}}$ and $\mathbb{E}_{\mu_h} \|\phi_h(s_h, a_h)\|^2_{(\Lambda_h^t)^{-1}}$ by our coverage condition in Assumption 3. Accordingly, the regret can be bounded by $\mathbb{E}_{\mu_h} \|\phi_h(s_h, a_h)\|^2_{(\Lambda_h^t)^{-1}}$ and thus improved if μ has a lower variance in Assumption 4.

6. Conclusion

Our work focuses on the question: *what can online RL benefit from coverage conditions?* In our setting, the standard structural assumptions on MDPs are substituted by coverage conditions in online RL. We answer this question in three folds: sample efficient guarantees of GOLF by various coverage conditions, the sample- and computation- efficiency guarantees of hybrid-Q, and faster regret bound of LSVI-UCB in linear MDP. Our results provide more possibilities of digging the potential and the utility of various coverage conditions. We believe that the relationship between coverage conditions and structural assumptions is always an interesting and important direction in general function approximation in RL, both empirically and theoretically, which requires more refined analysis in the future.

Acknowledgement

The authors would like to thank anonymous reviewers for their constructive suggestions.

Fanghui is supported by SNF project – Deep Optimisation of the Swiss National Science Foundation (SNSF) under grant number 200021_205011; Luca is funded in part through a PhD fellowship of the Swiss Data Science Center, a joint venture between EPFL and ETH Zurich; Volkan is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n°725594 - time-data).

References

Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 1–39. JMLR Workshop and Conference Proceedings, 2012.

A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

P. J. Ball, L. Smith, I. Kostrikov, and S. Levine. Efficient online reinforcement learning with offline data. *arXiv preprint arXiv:2302.02948*, 2023.

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

Z. Chen, C. J. Li, A. Yuan, Q. Gu, and M. I. Jordan. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

F. Cucker and D. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

C. Dann, Y. Mansour, M. Mohri, A. Sekhari, and K. Sridharan. Guarantees for epsilon-greedy reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 4666–4689, 2022.

S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

S. S. Du, Y. Luo, R. Wang, and H. Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, 2019.

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

B. Hao, T. Lattimore, C. Szepesvári, and M. Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

J. He, D. Zhou, and Q. Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

N. Jiang. Notes on exploration in linear MDPs. 2022.

N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

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

C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021a.

T. Jin, J. Yu, J. You, K. Zeng, C. Li, and Z. Yu. Low-rank matrix factorization with multiple hypergraph regularizer. *Pattern Recognition*, 48(3):1011–1022, 2015.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021b.

N. Lazic, D. Yin, M. Farajtabar, N. Levine, D. Gorur, C. Harris, and D. Schuurmans. A maximum-entropy approach to off-policy evaluation in average-reward MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 12461–12471, 2020.

S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

F. Liu, L. Viano, and V. Cevher. Understanding deep neural function approximation in reinforcement learning via ϵ-greedy exploration. In *Advances in Neural Information Processing Systems*, 2022.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems*, pages 1264–1274, 2020.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

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

A. Nair, M. Dalal, A. Gupta, and S. Levine. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Q. Nguyen, M. Mondelli, and G. F. Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning*, pages 8119–8129. PMLR, 2021.

M. Papini, A. Tirinzoni, A. Pacchiano, M. Restelli, A. Lazaric, and M. Pirotta. Reinforcement learning in linear MDPs: Constant regret and representation selection. In *Advances in Neural Information Processing Systems*, volume 34, pages 16371–16383, 2021.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, pages 11702–11716, 2021.

L. Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70 (1):117–129, 1985.

D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Y. Song, Y. Zhou, A. Sekhari, J. A. Bagnell, A. Krishnamurthy, and W. 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, 2018.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Nneural Information Processing Systems*, pages 1–7, 1999.

C. Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

M. Uehara and W. Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

A. W. Van Der Vaart, A. W. van der Vaart, A. van der Vaart, and J. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

A. Wagenmaker and A. Pacchiano. Leveraging offline data in online reinforcement learning. *arXiv preprint arXiv:2211.04974*, 2022.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

R. Wang, R. R. Salakhutdinov, and L. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, pages 6123–6135, 2020.

R. Wang, D. P. Foster, and S. M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021a.

Y. Wang, R. Wang, and S. Kakade. An exponential lower bound for linearly realizable MDP with constant suboptimality gap. In *Advances in Neural Information Processing Systems*, pages 9521–9533, 2021b.

T. Xie and N. Jiang. $q^\star$ approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559, 2020.

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 6683–6694, 2021a.

T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in neural information processing systems*, pages 27395–27407, 2021b.

T. Xie, D. J. Foster, Y. Bai, N. Jiang, and S. M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023.

L. Yang and M. Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Z. Yang, C. Jin, Z. Wang, M. Wang, and M. I. Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. In *Advances in Neural Information Processing Systems*, 2020.

A. Zanette. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *International Conference on Machine Learning*, pages 12287–12297. PMLR, 2021.

W. Zhan, B. Huang, A. Huang, N. Jiang, and J. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, pages 4572–4583, 2020.

D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

Appendix

Table of Contents

A. Flowchart of GOLF

For self-completeness, we include the flowchart of GOLF (Jin et al., 2021a) in Algorithm 1 here. This is a typical general function approximation algorithm in online RL, and yields sample-efficient guarantees if the BE dimension is small. The key step is line 7: optimization based exploration under the constraint of an identified confidence region $\mathcal{F}^{(t)}$ with a confidence parameter β. The quantity $\mathcal{L}_h^{(t)}(f, f')$ can be regarded as an approximation of the squared Bellman error at step h.

Algorithm 1 GOLF (Jin et al., 2021a)

```
1: **Input:** Function class: $\mathcal{F}$, confidence parameter $\beta$
2: Initialize $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} = \emptyset \;\; \forall h \in [H]$
3: **for** $t = 1, \ldots, T$ **do**
4:     Let $\pi^t$ be the greedy policy w.r.t. $f^t$ i.e., $f^t = \operatorname{argmax}_{f \in \mathcal{F}^{(t-1)}} f(s_1, \pi_{f,1}(s_1))$.
5:     For each $h \in [H]$, execute $\pi^t$ and obtain a trajectory $\{(s_h^t, a_h^t, r_h^t)\}_{h=1}^H$
6:     For each $h \in [H]$, dataset augment: $\mathcal{D}_h^{(t)} \leftarrow \mathcal{D}_h^{(t-1)} \bigcup \{(s_h^t, a_h^t, r_h^t, s_{h+1}^t)\}$.
7:     Update the confidence set with $f_{H+1} = 0$:
           $$\mathcal{F}^{(t)} \leftarrow \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t)}(f_h, f_{h+1}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f_h', f_{h+1}) \leq \beta, \quad \forall h \in [H] \right\}$$
       where $\mathcal{L}_h^{(t)}(f, f') := \sum_{(s,a,r,s') \in \mathcal{D}_h^{(t)}} \left[ f(s,a) - r - \max_{a' \in \mathcal{A}} f'(s', a') \right]^2, \forall f, f' \in \mathcal{F}$.
8: **end for**
9: **Output:** a policy uniformly sampled from $\{\pi^t\}_{t=1}^T$
```

B. Proofs for Section 3.1

In this section, we provide the proofs in Section 3 that the L^p coverage conditions are identified to ensure sample efficient online RL. Appendix B.1 gives the proof of Lemma 1, and the proof of Proposition 1 can be found in Appendix B.2.

B.1. Proof of Lemma 1

Proof. According to Definition 3, the formulation of $C_{\mathtt{cw}}$ endowed by the $L^p(\mathrm{d}\mu_h)$ norm implies

$$\begin{aligned} C_{\mathtt{cw}} &:= \inf_{\mu_1, \cdots, \mu_H \in \Delta(\mathcal{S} \times \mathcal{A})} \sup_{\pi \in \Pi, h \in [H]} \left\| \frac{\rho_h^\pi}{\mu_h} \right\|_{L^p(\mathrm{d}\mu_h)}^p \leqslant \inf_{\mu_1, \cdots, \mu_H \in \Delta(\mathcal{S} \times \mathcal{A})} \sum_{(s,a)} \sup_{\pi \in \Pi, h \in [H]} \frac{[\rho_h^\pi(s,a)]^p}{[\mu_h(s,a)]^{p-1}} \\ &:= \inf_{\mu_{\tilde{h}} \in \Delta(\mathcal{S} \times \mathcal{A})} \sum_{(s,a)} \sup_{\pi \in \Pi} \frac{[\rho_{\tilde{h}}^\pi(s,a)]^p}{[\mu_{\tilde{h}}(s,a)]^{p-1}} \quad \text{for a certain } \tilde{h} \in [H] \\ &= \inf_{\mu_{\tilde{h}} \in \Delta(\mathcal{S} \times \mathcal{A})} \sum_{(s,a)} \left(\frac{\sup_\pi \rho_{\tilde{h}}^\pi(s,a)}{\mu_{\tilde{h}}(s,a)} \right)^{p-1} \sup_\pi \rho_{\tilde{h}}^\pi(s,a) \\ &\leqslant \inf_{\mu_{\tilde{h}} \in \Delta(\mathcal{S} \times \mathcal{A})} \left(\max_{(s,a)} \frac{\sup_\pi \rho_{\tilde{h}}^\pi(s,a)}{\mu_{\tilde{h}}(s,a)} \right)^{p-1} \sum_{(s,a)} \sup_\pi \rho_{\tilde{h}}^\pi(s,a) \,, \end{aligned} \tag{5}$$

where the first inequality holds by Jensen inequality for a convex function sup.

Based on the formulation of $C_{\mathtt{cov}}$ in Definition 1, we have

$$C_{\mathtt{cov}} = \max_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sup_{\pi \in \Pi} \rho_h^\pi(s,a) \geqslant \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sup_{\pi \in \Pi} \rho_{\tilde{h}}^\pi(s,a) \,,$$

which implies

$$
\begin{aligned}
C_{\texttt{cw}} &\leqslant \inf_{\mu_{\tilde{h}} \in \Delta(\mathcal{S}\times\mathcal{A})} \left(\max_{(s,a)} \frac{\sup_\pi \rho^\pi_{\tilde{h}}(s,a)}{\mu_{\tilde{h}}(s,a)} \right)^{p-1} C_{\texttt{cov}} \\
&= C_{\texttt{cov}} \left(\inf_{\mu_{\tilde{h}} \in \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\Pi} \left\| \frac{\rho^\pi_{\tilde{h}}}{\mu_{\tilde{h}}} \right\|_\infty \right)^{p-1} \\
&\leqslant C_{\texttt{cov}}^p \,,
\end{aligned}
$$

where the second equality uses Definition 1 for $C_{\texttt{cov}}$ and the involved functions are monotonic w.r.t p. Finally we finish the proof. □

B.2. Proof of Proposition 1

Our proof framework follows (Xie et al., 2023, Theorem 1), and there is only one slight difference involved with the weaker data coverage coefficient $C_{\texttt{cw}}$, which leads to a different "exploration" phase based on $C_{\texttt{cw}}$. For self-completeness, we present the detailed proof here, which is also helpful to our remaining results.

Proof of Proposition 1. For every step h, denote

$$
\mu_h^\star := \operatorname*{argmin}_{\mu_h \subseteq \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\Pi} \left\| \frac{\rho_h^\pi}{\mu_h} \right\|^p_{L^p(\mathrm{d}\mu_h)} ,
$$

we have

$$
C_{\texttt{cw}} = \sup_{\pi\in\Pi, h\in[H]} \sum_{(s,a)} \frac{[\rho_h^\pi(s,a)]^p}{[\mu_h^\star(s,a)]^{p-1}} \geqslant \frac{[\rho_h^{(t)}(s,a)]^p}{[\mu_h^\star(s,a)]^{p-1}} \,, \forall t, h, (s,a) \,. \tag{6}
$$

For notational simplicity, we adopt the shorthand $\rho_h^{(t)} := \rho_h^{\pi^{(t)}}$, and define

$$
\tilde{\rho}_h^{(t)}(s,a) := \sum_{i=1}^{t-1} \rho_h^{(i)}(s,a) \,.
$$

which is the summation of all previous occupancy measure before episode t. Note that $\tilde{\rho}_h^{(t)}$ is not a probability measure because it is unnormalized. Accordingly, we introduce the notion of an "exploration" phase for each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ based on $C_{\texttt{cw}}\mu_h^\star$ such that

$$
\tau_h(s,a) = \min\left\{ t \mid \tilde{\rho}_h^{(t)}(s,a) \geqslant [C_{\texttt{cw}}\mu_h^\star(s,a)]^p \right\} \,, \tag{7}
$$

which describes the earliest time at which (s,a) has been explored. We refer to $t < \tau_h(s,a)$ as the exploration phase for (s,a).

In the next, following (Xie et al., 2023) on the regret decomposition, denoting $\delta_h^{(t)}(s,a) := f_h^{(t)}(s,a) - (\mathcal{T}_h f_{h+1}^{(t)})(s,a)$, we have

$$
\begin{aligned}
\texttt{Regret} \leqslant \sum_{t=1}^T \left(f_1^{(t)}(s_1, \pi_{f_1^{(t)},1}(s_1)) - J(\pi^{(t)}) \right) &= \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{(s,a)\sim\rho_h^{(t)}} \left[f_h^{(t)}(s,a) - (\mathcal{T}_h f_{h+1}^{(t)})(s,a) \right] \\
&= \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{(s,a)\sim\rho_h^{(t)}} \left[\delta_h^{(t)}(s,a) \mathbb{1}[t < \tau_h(s,a)] \right] + \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{(s,a)\sim\rho_h^{(t)}} \left[\delta_h^{(t)}(s,a) \mathbb{1}[t \geqslant \tau_h(s,a)] \right] \,,
\end{aligned}
$$

where the first term is the "exploration" phase and the second term is the stable phase.

In particular, for the "exploration" phase, we use $|\delta_h^{(t)}| \leqslant 1$ to bound

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{(s,a)\sim\rho_h^{(t)}} \left[\delta_h^{(t)}(s,a)\mathbb{1}[t < \tau_h(s,a)]\right] &\leqslant \sum_{(s,a)} \sum_{t<\tau_h(s,a)} \rho_h^{(t)}(s,a) = \sum_{(s,a)} \tilde{\rho}_h^{(\tau_h(s,a))}(s,a) \\
&= \sum_{(s,a)} [\tilde{\rho}_h^{(\tau_h(s,a)-1)}(s,a) + \rho_h^{(\tau_h(s,a)-1)}(s,a)] \\
&\leqslant \sum_{(s,a)} [C_{\texttt{cw}}\mu_h^\star(s,a)]^p + \sum_{(s,a)} C_{\texttt{cw}}^{\frac{1}{p}} [\mu_h^\star(s,a)]^{\frac{p-1}{p}} \\
&\leqslant C_{\texttt{cw}}^p + C_{\texttt{cw}}^{\frac{1}{p}} \sum_{s,a} [\mu_h^\star(s,a)]^{\frac{p-1}{p}} \\
&\lesssim C_{\texttt{cw}}^p \,,
\end{aligned}$$

where the second inequality holds by Eqs. (6), (7), and the last inequality holds by

$$\sum_{(s,a)} [\mu_h^\star(s,a)]^{\frac{p-1}{p}} \leqslant \sum_{(s,a)} \sqrt{\mu_h^\star(s,a)} < C \,, \tag{8}$$

for some constant C.

For the stable phase, by change-of-measure, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{(s,a)\sim\rho_h^{(t)}} \left[\delta_h^{(t)}(s,a)\mathbb{1}[t \geqslant \tau_h(s,a)]\right] \\
&= \sum_{t=1}^{T} \sum_{(s,a)} \rho_h^{(t)}(s,a) \left(\frac{\tilde{\rho}_h^{(t)}(s,a)}{\tilde{\rho}_h^{(t)}(s,a)}\right)^{\frac{1}{2}} \delta_h^{(t)}(s,a)\mathbb{1}[t \geqslant \tau_h(s,a)] \\
&\leqslant \sqrt{\underbrace{\sum_{t=1}^{T} \sum_{(s,a)} \frac{\left(\mathbb{1}[t \geqslant \tau_h(s,a)]\rho_h^{(t)}(s,a)\right)^2}{\tilde{\rho}_h^{(t)}(s,a)}}_{:=\texttt{I}_\texttt{A}}} \cdot \sqrt{\underbrace{\sum_{t=1}^{T} \sum_{(s,a)} \tilde{\rho}_h^{(t)}(s,a) \left(\delta_h^{(t)}(s,a)\right)^2 \mathbb{1}[t \geqslant \tau_h(s,a)]}_{:=\texttt{I}_\texttt{B}}},
\end{aligned} \tag{9}$$

where the last inequality is an application of Cauchy-Schwarz inequality.

We bound the first term $\texttt{I}_\texttt{A}$ in Eq. (9) with

$$\begin{aligned}
\texttt{I}_\texttt{A} := \sum_{t=1}^{T} \sum_{(s,a)} \frac{\left(\mathbb{1}_{\{t \geqslant \tau_h(s,a)\}}\rho_h^{(t)}(s,a)\right)^2}{\tilde{\rho}_h^{(t)}(s,a)} &\leqslant 2 \sum_{t=1}^{T} \sum_{(s,a)} \frac{\left(\mathbb{1}_{\{t \geqslant \tau_h(s,a)\}}\rho_h^{(t)}(s,a)\right)^2}{[C_{\texttt{cw}}\mu_h^\star(s,a)]^p + \tilde{\rho}_h^{(t)}(s,a)} \\
&\lesssim \sum_{t=1}^{T} \sum_{(s,a)} \rho_h^{(t)}(s,a) \frac{\rho_h^{(t)}(s,a)}{[C_{\texttt{cw}}\mu_h^\star(s,a)]^p + \tilde{\rho}_h^{(t)}(s,a)} \\
&\leqslant \sum_{t=1}^{T} \sum_{(s,a)} C_{\texttt{cw}}^{\frac{1}{p}} [\mu_h^\star(s,a)]^{\frac{p-1}{p}} \frac{\rho_h^{(t)}(s,a)}{[C_{\texttt{cw}}\mu_h^\star(s,a)]^p + \tilde{\rho}_h^{(t)}(s,a)} \\
&\lesssim C_{\texttt{cw}}^{\frac{1}{p}} \sum_{(s,a)} [\mu_h^\star(s,a)]^{\frac{p-1}{p}} \log T \quad \text{[using Lemma 7]} \\
&\lesssim C_{\texttt{cw}}^{\frac{1}{p}} \log T \quad \text{[using Eq. (8)]} \,,
\end{aligned} \tag{10}$$

where the first inequality uses $\widetilde{\rho}_h^{(t)}(s,a) \geqslant \frac{1}{2}\widetilde{\rho}_h^{(t)}(s,a) + \frac{1}{2}[C_{\texttt{cw}}\mu^\star(s,a)]^p$ and the third inequality holds by Eq. (6).

Algorithm 2 The Hybrid-Q algorithm using both offline and online data (Song et al., 2022)

```
Input: Value function class: $\mathcal{F}$, offline dataset $\mathcal{D}_h^{\nu}$ of size $n_{\text{off}}$ for $h \in [H-1]$
Initialize $f_h^1(s,a) = 0$.
for episode $t = 1, \dots, T$ do
    Let $\pi^t$ be the greedy policy w.r.t. $f^t$ i.e., $\pi_h^t(s) = \operatorname{argmax}_a f_h^t(s,a)$.
    For each $h$, sample $s_h \sim \rho_h^{\pi^t}$, $a_h \sim \pi^t(\cdot|s_h, a_h)$, and $\mathcal{D}_h^{(t)} \leftarrow \mathcal{D}_h^{(t-1)} \bigcup \{(s_h^t, a_h^t, r_h^t, s_{h+1}^t)\}$. // Online collection
    Set $f_H^{t+1}(s,a) = 0$.
    for $h = H-1, \dots, 0$ do
        Estimate $f_h^{t+1}$ using FQI on both offline and online data by defining $\varrho_h^t := [f(s,a) - r - \max_{a' \in \mathcal{A}} f_{h+1}^{t+1}(s',a')]^2$:
            $f_h^{t+1} \leftarrow \operatorname*{argmin}_{f \in \mathcal{F}_h} \left\{ \sum_{(s,a,r,s') \in \mathcal{D}_h^{\mu}} \varrho_h^t + \sum_{(s,a,r,s') \in \mathcal{D}_h^{(t)}} \varrho_h^t \right\}.$
    end for
end for
```

For the second term $\mathtt{I}_{\mathtt{B}}$ in Eq. (9), we can directly employ the result of (Jin et al., 2021a), see Lemma 5. By taking $\beta = c \log\left(\frac{\mathscr{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ for some constant c and $\delta \in (0,1)$, the quantity $\mathtt{I}_{\mathtt{B}}$ holds with probability at least $1-\delta$

$$\mathtt{I}_{\mathtt{B}} \lesssim \mathcal{O}(\beta T) \,.$$

Combining the results of the "exploration" phase and the stable phase, our regret bound holds with probability at least $1-\delta$

$$\mathtt{Regret} \leqslant \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim\rho_h^{(t)}}[\delta_h^{(t)}(s,a)] \lesssim \mathcal{O}\left(HC_{\mathtt{cw}}^{p} + H\sqrt{C_{\mathtt{cw}}^{\frac{1}{p}}\beta T \log T}\right) = \mathcal{O}\left(H\sqrt{C_{\mathtt{cw}}^{\frac{1}{p}}\beta T \log T}\right) \,,$$

which concludes the proof. □

C. Proof for Section 3.2

In this section, we firstly include the flowchart of the hybrid-Q algorithm in Appendix C.1 for self-completeness, and then present the proof of Proposition 2 in Appendix C.2.

C.1. Flowchart of the hybrid-Q algorithm

We include the flowchart of hybrid-Q in Algorithm 2 here for self-completeness. The idea of the hybrid-Q algorithm is intuitive. It is based on the classical fitted Q-iteration (FQI) algorithm on the offline dataset $\mathcal{D}_h^{\nu}$ and on-policy trajectory generated by the current policy interacting with the environment. This algorithm avoids sophisticated exploration schemes in online RL but uses offline data for exploration, and thus the computation complexity of this algorithm is the same as FQI with a least square regression oracle.

C.2. Proof of Proposition 2

In this section, we aim to prove that, using C_π in Eq. (1) or $C_{\mathtt{cw}}$ is able to ensure Algorithm 2 statistically and computationally efficient.

Proof of Proposition 2 . We firstly prove **Case 1** and then **Case 2**.

Proof of Case 1:

Lemma 6 implies that, for any $\delta \in (0,1)$, by taking $\beta = c \log\left(\frac{\mathscr{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ for some constant c, with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T} \mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2 \lesssim \frac{\beta T}{n_{\text{off}}} \,.$$

Accordingly, we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{(s,a)\sim\rho_h^{(t)}}\left[\delta_h^{(t)}(s,a)\right] &\leqslant \sum_{t=1}^{T}\mathbb{E}_{\rho_h^{(t)}}\delta_h^{(t)}(s,a)\left(\frac{\mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2}{\mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2}\right)^{\frac{1}{2}}\\ &\leqslant \sqrt{\sum_{t=1}^{T}\frac{[\mathbb{E}_{\rho_h^{(t)}}\delta_h^{(t)}(s,a)]^2}{\mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2}}\sqrt{\sum_{t=1}^{T}\mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2}\\ &\lesssim \sqrt{\sum_{t=1}^{T}\frac{[\mathbb{E}_{\rho_h^{(t)}}\delta_h^{(t)}(s,a)]^2}{\mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2}}\sqrt{\frac{\beta T}{n_{\text{off}}}}\\ &\leqslant \sqrt{T\max_{t\leqslant T}\frac{[\mathbb{E}_{\rho_h^{(t)}}\delta_h^{(t)}(s,a)]^2}{\mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2}}\sqrt{\frac{\beta T}{n_{\text{off}}}}\\ &\leqslant C_\pi\sqrt{\frac{\beta T^2}{n_{\text{off}}}}\,,\end{aligned}$$

where the second inequality uses the Cauchy-Schwartz inequality and the last inequality holds by the following result

$$\forall\pi\in\Pi,\quad \sqrt{\max_{t\leqslant T}\frac{[\mathbb{E}_{\rho_h^{(t)}}\delta_h^{(t)}(s,a)]^2}{\mathbb{E}_{\nu_h}[\delta_h^{(t)}(s,a)]^2}}\leqslant \max_{f\in\mathcal{F}}\frac{|[\mathbb{E}_{\rho_h^{\pi}}\delta_h(s,a)]|}{\sqrt{\mathbb{E}_{\nu_h}[\delta_h(s,a)]^2}}=C_\pi\,,$$

defined by Eq. (1). In our setting, we take $n_{\text{off}}=T$ for achieving $\mathcal{O}(\sqrt{T})$ regret.

Finally, by taking $\beta=c\log\left(\frac{\mathcal{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ for some constant c and $\delta\in(0,1)$, the regret bound of Algorithm 2 holds with probability at least $1-\delta$

$$\texttt{Regret}\leqslant\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{(x,a)\sim\rho_h^{(t)}}[\delta_h^{(t)}(s,a)]\lesssim\mathcal{O}\left(C_\pi H\sqrt{\beta T}\right)\,.$$

Proof of Case 2:

Our proof differs from that of Proposition 1 in how to estimate the in-sample squared Bellman error under Algorithm 2 without the structural assumption. This is also the technical challenge in this work when compared to (Song et al., 2022).

Recall the definition of $\tau_h(s,a)$ in Eq. (7), we have

- if $t\leqslant\tau_h(s,a)$, we have $\tilde{\rho}_h^{(t)}(s,a)\leqslant[C_{\texttt{cw}}\mu_h^\star(s,a)]^p$.
- if $t>\tau_h(s,a)$, we have $\tilde{\rho}_h^{(t)}(s,a)>[C_{\texttt{cw}}\mu_h^\star(s,a)]^p$ and $\rho^{(t)}(s,a)<C_{\texttt{cw}}^{\frac{1}{p}}[\mu_h^\star(s,a)]^{\frac{p-1}{p}}$ in Eq. (6).

Based on this, when $t>\tau_h(s,a)$, the unnormalized measure $\tilde{\rho}_h^{(t)}$ can be upper bounded by

$$\begin{aligned}\tilde{\rho}_h^{(t)}=\tilde{\rho}_h^{(t)}\mathbb{1}_{\{t\leqslant\tau_h(s,a)\}}+\tilde{\rho}_h^{(t)}\mathbb{1}_{\{t>\tau_h(s,a)\}}&\leqslant[C_{\texttt{cw}}\mu_h^\star(s,a)]^p+\sum_{i=\tau_h(s,a)+1}^{t-1}\rho_h^{(i)}\\ &\leqslant[C_{\texttt{cw}}\mu_h^\star(s,a)]^p+\sum_{i=\tau_h(s,a)+1}^{t-1}C_{\texttt{cw}}^{\frac{1}{p}}[\mu_h^\star(s,a)]^{\frac{p-1}{p}}\,.\end{aligned}$$

Following Eq. (9), the result on $\texttt{I}_\texttt{A}$ can be directly obtained by Eq. (10) in the proof of Proposition 1 such that $\texttt{I}_\texttt{A}\lesssim C_{\texttt{cw}}^{\frac{1}{p}}\log T$, and our main effort here is to estimate the in-sample squared Bellman error related to $\mathtt{I}_{\mathtt{B}}$. We split it into two terms

$$\begin{aligned}
&\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\tilde{\rho}_h^{(t)}(s,a)\left(\delta_h^{(t)}(s,a)\right)^2\mathbb{1}[t\geqslant\tau_h(s,a)]}\\
&\leqslant\underbrace{\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}[C_{\mathtt{cw}}\mu_h^\star(s,a)]^p\left(\delta_h^{(t)}(s,a)\right)^2}}_{:=\mathtt{I}_{\mathtt{B1}}}+\underbrace{\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\sum_{i=\tau_h(s,a)+1}^{t}C_{\mathtt{cw}}^{\frac{1}{p}}[\mu_h^\star(s,a)]^{\frac{p-1}{p}}\left(\delta_h^{(i)}(s,a)\right)^2}}_{:=\mathtt{I}_{\mathtt{B2}}}\,,
\end{aligned}\tag{11}$$

where we use $\sqrt{a+b}\leqslant\sqrt{a}+\sqrt{b}$ for $a,b\geqslant 0$.

For the first term $\mathtt{I}_{\mathtt{B1}}$ in Eq. (11), using Lemma 6, for any $\delta\in(0,1)$, by taking $\beta=c\log\left(\frac{\mathscr{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ for some constant c, with probability at least $1-\delta$, we have

$$\begin{aligned}
\mathtt{I}_{\mathtt{B1}}&\leqslant C_{\mathtt{cw}}^{\frac{p}{2}}\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\mu_h^\star(s,a)\left(\delta_h^{(t)}(s,a)\right)^2}=C_{\mathtt{cw}}^{\frac{p}{2}}\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\nu_h(s,a)\frac{\mu_h^\star(s,a)}{\nu_h(s,a)}\left(\delta_h^{(t)}(s,a)\right)^2}\\
&\lesssim C_{\mathtt{cw}}^{\frac{p}{2}}\sqrt{\frac{\widetilde{C}\beta T}{n_{\text{off}}}}\,,
\end{aligned}$$

where we use the coverage condition $\max_{s,a,h}\frac{\mu_h^\star(s,a)}{\nu_h(s,a)}\leqslant\widetilde{C}$. At the end of the proof, we discuss the choice of the offline distribution ν.

Similarly, for the second term $\mathtt{I}_{\mathtt{B2}}$ in Eq. (11), we have

$$\begin{aligned}
\mathtt{I}_{\mathtt{B2}}&\leqslant C_{\mathtt{cw}}^{\frac{1}{2p}}\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\sum_{i=1}^{t}[\mu_h^\star(s,a)]^{\frac{p-1}{p}}\left(\delta_h^{(i)}(s,a)\right)^2}=C_{\mathtt{cw}}^{\frac{1}{2p}}\sqrt{\sum_{t=1}^{T}t\sum_{(s,a)}[\mu_h^\star(s,a)]^{\frac{p-1}{p}}\left(\delta_h^{(i)}(s,a)\right)^2}\\
&\leqslant C_{\mathtt{cw}}^{\frac{1}{2p}}\sqrt{\sum_{t=1}^{T}t\sum_{(s,a)}\sqrt{\mu_h^\star(s,a)}\left(\delta_h^{(i)}(s,a)\right)^2}\,.
\end{aligned}$$

Using Lemma 6 and the coverage condition $\max_{s,a,h}\frac{\mu_h^\star(s,a)}{\nu_h^2(s,a)}\leqslant\widetilde{C}$, with the same probability as conducted in $\mathtt{I}_{\mathtt{B1}}$, we have

$$\sum_{(s,a)}\sqrt{\mu_h^\star(s,a)}\left(\delta_h^{(t)}(s,a)\right)^2\lesssim\sqrt{\widetilde{C}}\sum_{(s,a)}\nu_h(s,a)\left(\delta_h^{(t)}(s,a)\right)^2\lesssim\frac{\beta\sqrt{\widetilde{C}}}{n_{\text{off}}}\,.$$

which implies

$$\mathtt{I}_{\mathtt{B2}}\lesssim C_{\mathtt{cw}}^{\frac{1}{2p}}T\sqrt{\frac{\beta\widetilde{C}^{\frac{1}{2}}}{n_{\text{off}}}}\leqslant C_{\mathtt{cw}}^{\frac{1}{2p}}T\sqrt{\frac{\beta\widetilde{C}}{n_{\text{off}}}}\,,$$

where we use $\widetilde{C}\geqslant 1$.

Combining the estimation of $\mathtt{I}_{\mathtt{B1}}$ and $\mathtt{I}_{\mathtt{B2}}$ into Eq. (11), the in-sample squared Bellman error related to $\mathtt{I}_{\mathtt{B}}$ can be estimated with probability at least $1-\delta$

$$\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\tilde{\rho}_h^{(t)}(s,a)\left(\delta_h^{(t)}(s,a)\right)^2\mathbb{1}[t\geqslant\tau_h(s,a)]}\lesssim C_{\mathtt{cw}}^{\frac{p}{2}}\sqrt{\frac{\beta T\widetilde{C}}{n_{\text{off}}}}+C_{\mathtt{cw}}^{\frac{1}{2p}}T\sqrt{\frac{\beta\widetilde{C}}{n_{\text{off}}}}\,.$$

Accordingly, for any $\delta\in(0,1)$, by taking $\beta=c\log\left(\frac{\mathscr{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ for some constant c, the regret bound of Algorithm 2 holds with probability at least $1 - \delta$

$$\begin{aligned} \texttt{Regret} &\leqslant \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{(x,a)\sim d_h^{(t)}}[\delta_h^{(t)}(s,a)] \lesssim \mathcal{O}\left(HC_{\texttt{cw}}^{p} + HC_{\texttt{cw}}^{\frac{1}{2p}}\sqrt{\log T}\left[C_{\texttt{cw}}^{\frac{p}{2}}\sqrt{\frac{\beta T\widetilde{C}}{n_{\text{off}}}} + C_{\texttt{cw}}^{\frac{1}{2p}} T\sqrt{\frac{\beta\widetilde{C}}{n_{\text{off}}}}\right]\right) \\ &\lesssim \mathcal{O}\left(C_{\texttt{cw}}^{\frac{1}{p}} H \sqrt{\frac{\beta T^2 \widetilde{C}\log T}{n_{\text{off}}}}\right) . \end{aligned}$$

If taking $n_{\text{off}} := T$, we conclude the proof.

□

D. Proofs for Section 4

In this section, we mainly focus on the proof of Theorem 1 that provides the sample-efficient guarantees of the GOLF algorithm under our partial/rest coverage condition. The key difficulty is how to tackle the issue that some occupancy measures cannot be upper bounded by some (scaling) distribution. Before our proof, we require the following result on the equivalence for $P_{\texttt{cov}}$.

D.1. Proof on the equivalence

Based on our definition, it can be easily found that $P_{\texttt{cov}} = P_{\texttt{cr}} := \max_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sup_{\pi\in\mathcal{M}} \rho_h^{\pi}(s,a)$. The proof can be easily given from (Xie et al., 2023), and we present it here just for self-completeness.

Proof. For every step h, denote

$$\hat{\mu}_h^{\star} := \operatorname*{argmin}_{\mu_h \subseteq \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\mathcal{M}} \left\| \frac{\rho_h^{\pi}}{\mu_h} \right\|_{\infty} , \tag{12}$$

we have, one hand

$$\begin{aligned} \sum_{(s,a)} \sup_{\pi\in\mathcal{M}} \rho_h^{\pi}(s,a) &= \sum_{(s,a)} \frac{\max_{\pi\in\mathcal{M}} \rho_h^{\pi}(s,a)}{\hat{\mu}_h^{\star}(s,a)} \hat{\mu}_h^{\star}(s,a) \\ &\leqslant \sum_{(s,a)} \left(\max_{(s,a)} \frac{\max_{\pi\in\mathcal{M}} \rho_h^{\pi}(s,a)}{\hat{\mu}_h^{\star}(s,a)} \right) \hat{\mu}_h^{\star}(s,a) \\ &\leqslant \sum_{(s,a)} P_{\texttt{cov}} \hat{\mu}_h^{\star}(s,a) = P_{\texttt{cov}} . \end{aligned} \tag{13}$$

On the other hand, for any $\pi \in \mathcal{M}$, take $\mu_h \propto \max_{\pi\in\mathcal{M}} \rho_h^{\pi}$, we have

$$\frac{\rho_h^{\pi}(s,a)}{\mu_h(s,a)} = \frac{\rho_h^{\pi}(s,a)\sum_{(s',a')} \max_{\pi'\in\mathcal{M}} \rho_h^{\pi'}(s',a')}{\max_{\pi''\in\mathcal{M}} \rho_h^{\pi''}(s,a)} \leqslant \sum_{(s',a')} \max_{\pi'\in\mathcal{M}} \rho_h^{\pi'}(s',a') = P_{\texttt{cr}} ,$$

which implies $P_{\texttt{cov}} \leqslant P_{\texttt{cr}}$. Combining with Eq. (13), we conclude $P_{\texttt{cov}} = P_{\texttt{cr}}$. □

D.2. Proof of Theorem 1

Here we give the proof of sample-efficient guarantees of the GOLF algorithm under the coverage condition regarding the partial/rest policy class. The key difficulty is how to tackle the issue that some occupancy measures cannot be upper bounded by $P_{\texttt{cov}}\hat{\mu}_h^{\star}$. We need to build the connection between $\rho_h^{(t)}$ and $\hat{\mu}_h^{\star}$ and introduce $P_{\texttt{out}}$ to control such distribution shift.

Proof of Theorem 1. Similar to Proposition 1, the "exploration" phase for each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ based on our partial coverage $P_{\texttt{cov}}$ is defined as

$$\hat{\tau}_h(s,a) = \min\left\{ t \mid \tilde{\rho}_h^{(t)}(s,a) \geqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a) \right\} . \tag{14}$$

Regarding the “exploration” phase, we use that $|\delta_h^{(t)}| \leqslant 1$ to bound

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{(s,a)\sim\rho_h^{(t)}}\left[\delta_h^{(t)}(s,a)\mathbb{1}_{\{t<\hat{\tau}_h(s,a)\}}\right] &\leqslant \sum_{(s,a)}\sum_{t<\hat{\tau}_h(s,a)} \rho_h^{(t)}(s,a) = \sum_{(s,a)} \tilde{\rho}_h^{(\hat{\tau}_h(s,a))}(s,a)\\ &= \sum_{(s,a)} [\tilde{\rho}_h^{(\hat{\tau}_h(s,a)-1)}(s,a) + \rho_h^{(\hat{\tau}_h(s,a)-1)}(s,a)]\\ &\leqslant \sum_{(s,a)} P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a) + 1\\ &\leqslant 2P_{\texttt{cov}}\,,\end{aligned}$$

where we use Eq. (14) in the second inequality.

In the stable phase, we have $\widetilde{\rho}_h^{(t)}(s,a) \geqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)$. Similar to Eq. (9), we aim to estimate the following quantity

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}_{(s,a)\sim\rho_h^{(t)}}\left[\delta_h^{(t)}(s,a)\mathbb{1}[t \geqslant \tau_h(s,a)]\right]\\ &\leqslant \sqrt{\underbrace{\sum_{t=1}^{T}\sum_{(s,a)} \frac{\left(\mathbb{1}[t \geqslant \tau_h(s,a)]\rho_h^{(t)}(s,a)\right)^2}{\tilde{\rho}_h^{(t)}(s,a)}}_{:=\texttt{I}_{\texttt{A}}}} \cdot \sqrt{\underbrace{\sum_{t=1}^{T}\sum_{(s,a)} \tilde{\rho}_h^{(t)}(s,a)\left(\delta_h^{(t)}(s,a)\right)^2 \mathbb{1}[t \geqslant \tau_h(s,a)]}_{:=\texttt{I}_{\texttt{B}}}}\,,\end{aligned} \tag{15}$$

where the inequality holds by the Cauchy-Schwarz inequality and $\texttt{I}_{\texttt{B}} \lesssim \mathcal{O}(\beta T)$ w.h.p by Lemma 5 from the result of (Jin et al., 2021a). Our main effort in this proof is to bound $\texttt{I}_{\texttt{A}}$.

Bound $\texttt{I}_{\texttt{A}}$: If the current policy $\pi_h^{(t)}$ generating $\rho_h^{(t)}(s,a)$ belongs to $\mathcal{M}_h$, according to Eq. (12), we have

$$P_{\texttt{cov}} = \sup_{\pi\in\mathcal{M},h\in[H]} \left\|\frac{\rho_h^{\pi}}{\hat{\mu}_h^{\star}}\right\|_{\infty} \geqslant \max_{(s,a),h\in[H]} \frac{\rho_h^{(t)}(s,a)}{\hat{\mu}_h^{\star}(s,a)}\,,$$

which implies that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $\rho_h^{(t)}(s,a) \leqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)$ if $\pi_h^{(t)} \in \mathcal{M}_h$. Nevertheless, in online RL, we can not ensure $\pi_h^{(t)} \in \mathcal{M}_h$ such that $\rho_h^{(t)}(s,a) \leqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)$, which leads to the main difficulty: how to bound the first term in Eq. (15) if $\pi_h^{(t)} \notin \mathcal{M}_h$. Accordingly, we split $\texttt{I}_{\texttt{A}}$ into two cases: $\rho_h^{(t)}(s,a) \leqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)$ and $\rho_h^{(t)}(s,a) > P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)$ as below

$$\begin{aligned}\texttt{I}_{\texttt{A}} &:= \sum_{t=1}^{T}\sum_{(s,a)} \frac{\left(\mathbb{1}_{\{t\geqslant\hat{\tau}_h(s,a)\}}\rho_h^{(t)}(s,a)\right)^2}{\tilde{\rho}_h^{(t)}(s,a)} \leqslant 2\sum_{t=1}^{T}\sum_{(s,a)} \frac{\left(\mathbb{1}_{\{t\geqslant\hat{\tau}_h(s,a)\}}\rho_h^{(t)}(s,a)\right)^2}{P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a) + \tilde{\rho}_h^{(t)}(s,a)}\\ &= \underbrace{2\sum_{t=\hat{\tau}_h}^{T}\sum_{(s,a)} \rho_h^{(t)}(s,a)\frac{\rho_h^{(t)}(s,a)\mathbb{1}_{\left\{\rho_h^{(t)}(s,a)\leqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)\right\}}}{P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a) + \tilde{\rho}_h^{(t)}(s,a)}}_{\texttt{I}_{\texttt{A1}}} + \underbrace{2\sum_{t=\hat{\tau}_h}^{T}\sum_{(s,a)} \rho_h^{(t)}(s,a)\frac{\rho_h^{(t)}(s,a)\mathbb{1}_{\left\{\rho_h^{(t)}(s,a)> P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)\right\}}}{P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a) + \tilde{\rho}_h^{(t)}(s,a)}}_{\texttt{I}_{\texttt{A2}}}\,.\end{aligned} \tag{16}$$

Bound $\texttt{I}_{\texttt{A1}}$: Since $\rho_h^{(t)}(s,a) \leqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)$ satisfies the condition in Lemma 7, similar to Eq. (10), term $\texttt{I}_{\texttt{A1}}$ can be estimated by

$$\begin{aligned}\texttt{I}_{\texttt{A1}} &\lesssim \sum_{(s,a)} \max_{i\leqslant T}\rho_h^{(i)}(s,a)\sum_{t=1}^{T} \frac{\rho_h^{(t)}(s,a)\mathbb{1}_{\left\{\rho_h^{(t)}(s,a)\leqslant P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)\right\}}}{P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a) + \tilde{\rho}_h^{(t)}(s,a)}\\ &\lesssim \sum_{(s,a)} P_{\texttt{cov}}\hat{\mu}_h^{\star}(s,a)\log T\\ &= P_{\texttt{cov}}\log T\,.\end{aligned}$$

Bound $\mathtt{I}_{\mathtt{A2}}$: We cast the regime $\rho_h^{(t)}(s,a) > P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)$ into two cases:

- **Case 1:** $\rho_h^{(t)}(s,a) \leqslant c_1 P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and some constant $c_1 \geqslant 1$.
- **Case 2:** $\rho_h^{(t)}(s,a) > c_1 P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)$ for all potential (s,a).

Recall the definition of $\mathcal{B}^{\bar{\mathcal{M}}}$ in Definition 5, we consider a special case

$$\mathcal{B}_h^{(t)} := \left\{(s,a)\in\mathcal{S}\times\mathcal{A} \mid \rho_h^{(t)}(s,a) > c_1 P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)\right\}\,,\quad h\in[H]\,.$$

Case 1: $\rho_h^{(t)}(s,a) \leqslant c_1 P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)$.
We split term $\mathtt{I}_{\mathtt{A2}}$ into two parts

$$\mathtt{I}_{\mathtt{A2}} = \underbrace{\sum_{t=\hat{\tau}_h}^{T}\sum_{(s,a)}\rho_h^{(t)}(s,a)\frac{\rho_h^{\bar{\pi}}(s,a)\mathbb{1}_{\left\{\rho_h^{(t)}(s,a)>P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)\right\}}}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)}}_{(II_1)} + \underbrace{\sum_{t=\hat{\tau}_h}^{T}\sum_{(s,a)}\rho_h^{(t)}(s,a)\frac{[\rho_h^{(t)}(s,a)-\rho_h^{\bar{\pi}}(s,a)]\mathbb{1}_{\left\{\rho_h^{(t)}(s,a)>P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)\right\}}}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)}}_{(II_2)}\,.$$

For (II_1), we know $\rho_h^{\bar{\pi}}(s,a) \leqslant P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)$ due to $\bar{\pi}\in\mathcal{M}$, we have

$$\begin{aligned}\sum_{t=1}^{T}\frac{\rho_h^{\bar{\pi}}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)} &\leqslant 2\sum_{t=1}^{T}\log\left(1+\frac{\rho_h^{\bar{\pi}}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)}\right) \leqslant 2\sum_{t=1}^{T}\log\left(1+\frac{\rho_h^{(t)}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)}\right)\\ &= 2\log\left(\prod_{t=1}^{T}\frac{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\sum_{i=1}^{t}\rho_h^{(i)}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\sum_{i=1}^{t-1}\rho_h^{(i)}(s,a)}\right) = 2\log\left(1+\frac{\sum_{i=1}^{T}\rho_h^{(i)}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)}\right)\\ &\leqslant 2\log(1+c_1T)\,,\end{aligned} \tag{17}$$

where in the first inequality we use $x \leqslant 2\log(1+x)$ for any $x\in[0,1]$; the second inequality holds by $\rho_h^{\bar{\pi}}(s,a) < P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a) < \rho_h^{(t)}(s,a)$ and the last inequality uses the condition of **Case 1**. Based on this result, we can upper bound term (II_1) such that

$$\begin{aligned}(II_1) &\lesssim \sum_{(s,a)}\max_{i\leqslant T}\rho_h^{(i)}(s,a)\sum_{t=1}^{T}\frac{\rho_h^{\bar{\pi}}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)}\\ &\lesssim \sum_{(s,a)}[c_1P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)]\log(1+c_1T)\\ &\lesssim c_1P_{\mathsf{cov}}\log T\,.\end{aligned}$$

For (II_2), since $\rho_h^{(t)}(s,a)-\rho_h^{\bar{\pi}}(s,a) \leqslant c_1P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)$, similar to Eq. (17), we have

$$\sum_{t=1}^{T}\frac{[\rho_h^{(t)}(s,a)-\rho_h^{\bar{\pi}}(s,a)]}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)} \leqslant \sum_{t=1}^{T}\frac{c_1P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)+\tilde{\rho}_h^{(t)}(s,a)} \leqslant 2c_1\log(1+c_1T)\,,$$

which implies $(II_2) \lesssim c_1P_{\mathsf{cov}}\log T$.

Case 2: $(s,a)\in\mathcal{B}_h^{(t)}$. Note that if we choose the reference policy $\bar{\pi} := \pi^\star$, according to the definition of $\mathcal{B}_h^{(t)}$ under this case, we have

$$1 \geqslant \sum_{(s,a)\in\mathcal{B}_h^{(t)}}\rho_h^{(t)}(s,a) > \sum_{(s,a)\in\mathcal{B}_h^{(t)}}c_1P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a) \geqslant \sum_{(s,a)\in\mathcal{B}_h^{(t)}}c_1\rho_h^{\pi^\star}(s,a)\,, \tag{18}$$

which implies that the probability that $\pi_h^\star$ visits this state-action pair set $\mathcal{B}_h^{(t)}$ is smaller than $1/c_1$. That means, we can still identify the optimal policy $\pi^\star$ with probability at least $(1-\frac{1}{c_1})^H \geqslant 1-\frac{H}{c_1}$ for a proper c_1 even though we do not consider $\mathcal{B}^{(t)} := \{\mathcal{B}_h^{(t)}\}_{h=1}^H$.

For general reference policy $\bar{\pi}$, we have the following result. According to the definition of $\tilde{\rho}_h^{(t)}$ for any $t > \hat{\tau}_h$, each component at episode t in term (II) admits

$$\sum_{(s,a)} \rho_h^{(t+1)}(s,a) \frac{\rho_h^{(t+1)}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a) + \tilde{\rho}_h^{(t+1)}(s,a)} \leqslant \sum_{(s,a)} \rho_h^{(t+1)}(s,a) \frac{\rho_h^{(t+1)}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a) + c_1 P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a) + \tilde{\rho}_h^{(t)}(s,a)}\,,$$

due to $\rho_h^{(t)}(s,a) > c_1 P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)$.

Accordingly, for **Case 2**, $\forall (s,a) \in \mathcal{B}_h^{(t)}$, term $\mathtt{I}_{\mathtt{A2}}$ can be estimated by

$$\begin{aligned}\mathtt{I}_{\mathtt{A2}} &\leqslant \sum_{t=\hat{\tau}_h}^{T} \sum_{(s,a)\in\mathcal{B}_h^{(t)}} \rho_h^{(t)} \frac{\rho_h^{(t)}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a) + \tilde{\rho}_h^{(t)}(s,a)} \leqslant \sum_{t=\hat{\tau}_h}^{T} \sum_{(s,a)\in\mathcal{B}_h^{(t)}} \rho_h^{(t)} \frac{\rho_h^{(t)}(s,a)}{P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a) + \tilde{\rho}_h^{(\hat{\tau}_h)}(s,a) + (t-\hat{\tau}_h)P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)} \\ &\leqslant \sum_{t=\hat{\tau}_h}^{T} \sum_{(s,a)\in\mathcal{B}_h^{(t)}} \rho_h^{(t)} \frac{\rho_h^{(t)}(s,a)}{(t-\hat{\tau}_h+1)P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)}\,.\end{aligned} \tag{19}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\sum_{(s,a)\in\mathcal{B}_h^{(t)}} \rho_h^{(t)} \frac{\rho_h^{(t)}(s,a)}{(t-\hat{\tau}_h+1)P_{\mathsf{cov}}\hat{\mu}_h^\star(s,a)} &\leqslant \frac{1}{P_{\mathsf{cov}}} \sqrt{\sum_{(s,a)\in\mathcal{B}_h^{(t)}} \frac{[\rho_h^{(t)}(s,a)]^2}{[\hat{\mu}_h^\star(s,a)]^2}} \cdot \sqrt{\sum_{(s,a)\in\mathcal{B}_h^{(t)}} \frac{[\rho_h^{(t)}(s,a)]^2}{[t-\hat{\tau}_h+1]^2}} \\ &\leqslant \frac{1}{P_{\mathsf{cov}}(t-\hat{\tau}_h+1)} \sqrt{\sum_{(s,a)\in\mathcal{B}_h^{(t)}} \frac{[\rho_h^{(t)}(s,a)]^2}{[\hat{\mu}_h^\star(s,a)]^2}} \\ &= \frac{1}{P_{\mathsf{cov}}(t-\hat{\tau}_h+1)} \left\| \frac{\rho_h^{(t)}}{\hat{\mu}_h^\star} \mathbb{1}_{\mathcal{B}_h^{(t)}} \right\|_{L^2}\,,\end{aligned} \tag{20}$$

where the indicator function $\mathbb{1}_{\mathcal{B}_h^{(t)}} = 1$ if $(s,a) \in \mathcal{B}_h^{(t)}$, and otherwise is zero. Accordingly, taking this equation back to Eq. (19), we have

$$\mathtt{I}_{\mathtt{A2}} \leqslant \sum_{t=\hat{\tau}_h}^{T} \frac{1}{P_{\mathsf{cov}}(t-\hat{\tau}_h+1)} \left\| \frac{\rho_h^{(t)}}{\hat{\mu}_h^\star} \mathbb{1}_{\mathcal{B}_h^{(t)}} \right\|_{L^2} \lesssim \frac{\log T}{P_{\mathsf{cov}}} \max_{t\leqslant T} \left\| \frac{\rho_h^{(t)}}{\hat{\mu}_h^\star} \mathbb{1}_{\mathcal{B}_h^{(t)}} \right\|_{L^2}\,.$$

Accordingly, combining the results of $\mathtt{I}_{\mathtt{A1}}$ and $\mathtt{I}_{\mathtt{A2}}$ into Eq. (16), under the definition of $\mathcal{B}_h^{(t)}$, we have

$$\mathtt{I}_{\mathtt{A}} = \sum_{t=1}^{T} \sum_{(s,a)} \frac{\left(\mathbb{1}_{\{t\geqslant\hat{\tau}_h(s,a)\}}\rho_h^{(t)}(s,a)\right)^2}{\tilde{\rho}_h^{(t)}(s,a)} \lesssim \left(c_1 P_{\mathsf{cov}} + \frac{1}{P_{\mathsf{cov}}} \max_{\hat{\tau}_h\leqslant t\leqslant T} \left\| \frac{\rho_h^{(t)}}{\hat{\mu}_h^\star} \mathbb{1}_{\mathcal{B}_h^{(t)}} \right\|_{L^2} \right) \log T\,.$$

Following Eq. (15), by taking $\beta = c\log\left(\frac{\mathcal{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ for some constant c and $\delta \in (0,1)$, combining the results of $\mathtt{I}_{\mathtt{A}}$ and

$\mathtt{I_A}$, the result for the stable phase holds with probability at least $1-\delta$

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{(s,a)\sim\rho_h^{(t)}}\left[\delta_h^{(t)}(s,a)\mathbb{1}[t\geqslant\hat{\tau}_h(s,a)]\right]\\
&\leqslant\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\frac{\left(\mathbb{1}[t\geqslant\hat{\tau}_h(s,a)]\rho_h^{(t)}(s,a)\right)^2}{\tilde{\rho}_h^{(t)}(s,a)}}\cdot\sqrt{\sum_{t=1}^{T}\sum_{(s,a)}\tilde{\rho}_h^{(t)}(s,a)\left(\delta_h^{(t)}(s,a)\right)^2\mathbb{1}[t\geqslant\hat{\tau}_h(s,a)]}\\
&\lesssim\mathcal{O}\left(\sqrt{c_1P_{\mathtt{cov}}+\frac{1}{P_{\mathtt{cov}}}\max_{\hat{\tau}_h\leqslant t\leqslant T}\left\|\frac{\rho_h^{(t)}}{\hat{\mu}_h^{\star}}\mathbb{1}_{\mathcal{B}_h^{(t)}}\right\|_{L^2}}\sqrt{\beta T\log T}\right)\\
&\lesssim\mathcal{O}\left(\left(\sqrt{c_1P_{\mathtt{cov}}}+\frac{1}{\sqrt{P_{\mathtt{cov}}}}\max_{\hat{\tau}_h\leqslant t\leqslant T}\left\|\frac{\rho_h^{(t)}}{\hat{\mu}_h^{\star}}\mathbb{1}_{\mathcal{B}_h^{(t)}}\right\|_{L^2}^{\frac{1}{2}}\right)\sqrt{\beta T\log T}\right)\,,
\end{aligned}
$$

where the last inequality uses $\sqrt{a+b}\leqslant\sqrt{a}+\sqrt{b}$ for any $a,b\geqslant 0$. Besides, we can set the quantity to $\max\left\{1,\max_{\hat{\tau}_h\leqslant t\leqslant T}\left\|\frac{\rho_h^{(t)}}{\hat{\mu}_h^{\star}}\mathbb{1}_{\mathcal{B}_h^{(t)}}\right\|_{L^2}\right\}$ such that the square root operator can be taken into the $\max$. If this quantity is smaller than 1, that means $\sqrt{c_1P_{\mathtt{cov}}}$ dominates the result and thus the second can be omitted. Recall the definition of $P_{\mathtt{out}}$ in Definition 5, we have

$$
\max_{h\in[H],\pi\notin\mathcal{M}}\left\|\frac{\rho_h^{(t)}}{\hat{\mu}_h^{\star}}\mathbb{1}_{\mathcal{B}_h^{(t)}}\right\|_{L^2}^{\frac{1}{2}}\leqslant P_{\mathtt{out}}\,.
$$

Accordingly, our regret bound holds with probability at least $1-\delta$

$$
\begin{aligned}
\mathtt{Regret}&\leqslant\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{(x,a)\sim\rho_h^{(t)}}[\delta_h^{(t)}(s,a)]\lesssim\mathcal{O}\left(HP_{\mathtt{cov}}+H\Big(\sqrt{c_1P_{\mathtt{cov}}}+\frac{P_{\mathtt{out}}}{\sqrt{P_{\mathtt{cov}}}}\Big)\sqrt{\beta T\log T}\right)\\
&=\mathcal{O}\left(H\Big(\sqrt{c_1P_{\mathtt{cov}}}+\frac{P_{\mathtt{out}}}{\sqrt{P_{\mathtt{cov}}}}\Big)\sqrt{\beta T\log T}\right)\,.
\end{aligned}
\tag{21}
$$

If $\mathcal{B}_h$ is an empty set for some h, it means that $\rho_h^{(t)}(s,a)<c_1P_{\mathtt{cov}}\hat{\mu}_h^{\star}(s,a)$ always holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, which falls into the $\mathcal{M}=\Pi$ case. In this case, we have $P_{\mathtt{cov}}=C_{\mathtt{cov}}$, and the second term with $P_{\mathtt{out}}=0$ in the above equation is discarded. Hence we can recover the result of (Xie et al., 2023).

Clearly, there exists a trade-off between $P_{\mathtt{cov}}(\zeta)$ and $P_{\mathtt{out}}(\zeta)$ that depends on ζ. That means, there exists a proper $\zeta^{\star}$ such that $P_{\mathtt{out}}(\zeta)=\sqrt{c_1}P_{\mathtt{cov}}(\zeta^{\star})$ by the property of the function $x+c/x$ for some constant c. Accordingly, the regret bound in Eq. (21) can be improved to

$$
\mathtt{Regret}\lesssim\mathcal{O}\left(H\sqrt{c_1^{\frac{1}{2}}\beta TP_{\mathtt{out}}(\zeta^{\star})\log T}\right)
$$

which admits $P_{\mathtt{out}}(\zeta^{\star})\leq C_{\mathtt{cov}}$. This demonstrates a better regret bound than (Xie et al., 2023) by a good trade-off between $P_{\mathtt{cov}}$ and $P_{\mathtt{out}}$. Finally, we conclude the proof. □

E. Proof for Section 5

In this section, we first prove Theorem 2 in Appendix E.1 and then discuss the choice of the regularization parameter in Appendix E.2.

E.1. Proof of Theorem 2

To prove our result, we need the following notations and lemmas to aid our proof. For self-completeness, we include the LSVI-UCB algorithm (Jin et al., 2020) for linear MDP, see Algorithm 3 for details.

Algorithm 3 LSVI-UCB for linear MDP (Jin et al., 2020)

```
**Input:** The regularization parameter $\lambda$ and confidence parameter $\beta$.
**for** episode $t = 1, \dots, T$ **do**
    Receive the initial state $s_1^t$ and set $V_{H+1}^t$ as the zero function.
    **for** step $h = H, \dots, 1$ **do**
        Obtain $\Lambda_h^t \leftarrow \sum_{\tau=1}^t [\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top] + \lambda I$
        Obtain $\widehat{\boldsymbol{w}}_h^t \leftarrow (\Lambda_h^t)^{-1} \sum_{\tau=1}^t \phi(s_h^\tau, a_h^\tau)[r_h(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^t(s_{h+1}^\tau, a)]$ and $\widehat{Q}_h^t(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \widehat{\boldsymbol{w}}_h^t \rangle$
        Obtain $Q_h^t(\cdot,\cdot) \leftarrow \min\{\widehat{Q}_h^t(\cdot,\cdot) + \beta[\phi(\cdot,\cdot)^\top (\Lambda_h^t)^{-1} \phi(\cdot,\cdot)]^{1/2}, H\}$
    **end for**
    **for** step $h = 1, \dots, H$ **do**
        Take action $a_h^t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h^t(s_h^t, a)$ and obtain $V_h^t(\cdot) = \max_{a \in \mathcal{A}} Q_h^t(\cdot, a)$.
        Observe the reward $r_h(s_h^t, a_h^t)$ and the next state $s_{h+1}^t$.
    **end for**
**end for**
```

In LSVI-UCB, the estimator is given by solving a regularized least squares problem as below.

$$\widehat{\boldsymbol{w}}_h^t \leftarrow \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{\tau=1}^{t-1} [r_h(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^t(s_{h+1}^\tau, a) - \langle \boldsymbol{w}, \phi(s_h^\tau, a_h^\tau) \rangle]^2 + \lambda \|\boldsymbol{w}\|_2^2 \,, \tag{22}$$

where the feature mapping $\phi(s,a) \in \mathbb{R}^d$ satisfies $\|\phi(s,a)\|_2 \leqslant 1$ and $\lambda \geqslant 1$ is the regularization parameter. For notational simplicity, denote

$$\Lambda_h^t := \lambda I + \sum_{i=1}^{t-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top := \lambda I + (\Phi_h^t)^\top \Phi_h^t \,, \quad \text{with } (s_h^i, a_h^i) \sim \rho_h^{(t)} \,, \tag{23}$$

where $\Phi_h^t = [\phi_h(s_h^1, a_h^1), \cdots, \phi_h(s_h^{t-1}, a_h^{t-1})]^\top \in \mathbb{R}^{(t-1)\times d}$, and accordingly we can easily obtain an estimation of eigenvalues of $(\Lambda_h^t)^{-1}$ such that

$$\frac{1}{\lambda} \geqslant \lambda_{\max}[(\Lambda_h^t)^{-1}] \geqslant \lambda_{\min}[(\Lambda_h^t)^{-1}] = \frac{1}{\lambda_{\max}[(\Phi_h^t)^\top \Phi_h^t + \lambda I]} \geqslant \frac{1}{\lambda_{\max}[(\Phi_h^t)^\top \Phi_h^t] + \lambda} \geqslant \frac{1}{d + \lambda} \,, \tag{24}$$

where the last inequality holds by $\|\phi(s,a)\|_2 \leqslant 1$ and the fact $\|\boldsymbol{A}\|_2 \leqslant \sqrt{mn} \max_{i,j} A_{ij}$ where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.

In the next, we have the following lemmas.

Lemma 2. *For the intermediate quantity $\phi(s_h^i, a_h^i)^\top (\Lambda_h^t)^{-1} \phi(s_h^i, a_h^i)$ with $i \in [T]$, where Λ_h^t defined by Eq.* (23) *realized by the occupancy measure $\rho_h^{(t)}$, and the feature mapping $\phi(s_h^i, a_h^i)$ is assumed to admit $(s_h^i, a_h^i) \overset{\text{i.i.d}}{\sim} \mu_h$ for a underlying distribution μ_h, then we have*

$$\frac{1}{T} \sum_{i=1}^{T} [\phi(s_h^i, a_h^i)^\top (\Lambda_h^t)^{-1} \phi(s_h^i, a_h^i)] \leqslant \frac{2d^2}{T\lambda} \log(T+1) \,.$$

Proof. We introduce an auxiliary variable $\widetilde{\Lambda}_h^t \in \mathbb{R}^{d \times d}$ such that

$$\widetilde{\Lambda}_h^t = \lambda I + \sum_{j=1}^{t-1} \phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top \quad \text{with } (s_h^j, a_h^j) \overset{\text{i.i.d}}{\sim} \mu_h \,,$$

then we have

$$
\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T}[\phi(s_h^i,a_h^i)^\top(\Lambda_h^t)^{-1}\phi(s_h^i,a_h^i)] &= \frac{1}{T}\sum_{i=1}^{T}\left(\phi(s_h^i,a_h^i)^\top(\widetilde{\Lambda}_h^{(i-1)})^{-1}[\widetilde{\Lambda}_h^{(i-1)}(\Lambda_h^t)^{-1}]\phi(s_h^i,a_h^i)\right)\\
&= \frac{1}{T}\sum_{i=1}^{T}\mathrm{Tr}\left(\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^\top(\widetilde{\Lambda}_h^{(i-1)})^{-1}[\widetilde{\Lambda}_h^{i}(\Lambda_h^t)^{-1}]\right)\\
&\overset{(a)}{\leqslant} \frac{1}{T}\sum_{i=1}^{T}\mathrm{Tr}\left(\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^\top(\widetilde{\Lambda}_h^{(i-1)})^{-1}\right)\|\widetilde{\Lambda}_h^{(i-1)}(\Lambda_h^t)^{-1}\|_2\\
&\overset{(b)}{\leqslant} \frac{d}{T\lambda}\sum_{i=1}^{T}\left(\phi(s_h^i,a_h^i)^\top(\widetilde{\Lambda}_h^{(i-1)})^{-1}\phi(s_h^i,a_h^i)\right)\\
&\overset{(c)}{\leqslant} \frac{2d^2}{T\lambda}\log(T+1)\,,
\end{aligned}
$$

where (a) uses $\mathrm{Tr}(\boldsymbol{AB}) \leqslant \mathrm{Tr}(\boldsymbol{A})\|\boldsymbol{B}\|_2$; (b) uses $\|(\Lambda_h^t)^{-1}\|_2 \leqslant \frac{1}{\lambda}$, $\|\widetilde{\Lambda}_h^i\|_2 \leqslant d$ via $\|\boldsymbol{A}\|_2 \leqslant \sqrt{mn}\max_{i,j} A_{ij}$ where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$; and (c) uses the elliptical potential lemma with $U_t = U_{t-1} + X_t X_t^\top \in \mathbb{R}^{d\times d}$, $U_0 = \lambda I$, and $\|X_t\|_2 \leqslant 1$ such that

$$
\sum_{t=1}^{T} X_t^\top U_{t-1} X_t \leqslant 2d\log\left(1+\frac{T}{\lambda d}\right)\,.
$$

□

Lemma 3. *Under Assumption 3 with $\gamma > 0$ and the feature mapping $\phi(s,a) \in \mathbb{R}^d$ in linear MDP satisfies $\|\phi(s,a)\|_2 \leqslant 1$, we have*

$$
\mathbb{E}_{\rho_h^{(t)}}\left[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)\right] \leqslant \frac{(d+\lambda)^2}{d^2\gamma^2\lambda}\left(\mathbb{E}_{\mu_h}[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)]\right)^2\,.
$$

Proof. Assumption 3 yields $\mathbb{E}_\mu[\|\phi(s,a)\|_2^2] \geqslant d\gamma$, by taking $C_e := \frac{1}{d^2\gamma^2}$, we have

$$
\mathbb{E}_{\rho_h^{(t)}}[\|\phi(s,a)\|_2^2] \leqslant 1 \leqslant C_e\left(\mathbb{E}_\mu[\|\phi(s,a)\|_2^2]\right)^2\,. \tag{25}
$$

Using the linearity of the trace operator and expectation, we have

$$
\begin{aligned}
\mathbb{E}_{\rho_h^{(t)}}\left[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)\right] &= \mathrm{Tr}\left(\mathbb{E}_{\rho_h^{(t)}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\right]\right)\\
&\leqslant \frac{1}{\lambda}\mathrm{Tr}\left(\mathbb{E}_{\rho_h^{(t)}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\top\right]\right) = \frac{1}{\lambda}\mathbb{E}_{\rho_h^{(t)}}[\|\phi(s,a)\|_2^2]\,,
\end{aligned}
$$

where we use $\|(\Lambda_h^t)^{-1}\|_2 \leqslant 1/\lambda$. Accordingly, we have

$$
\begin{aligned}
\mathbb{E}_{\rho_h^{(t)}}\left[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)\right] &\leqslant \frac{C_e}{\lambda}\left(\mathbb{E}_{\mu_h}[\|\phi(s_h,a_h)\|_2^2]\right)^2 \quad \text{[using Eq. (25)]}\\
&= \frac{(d+\lambda)^2C_e}{\lambda}\left(\frac{1}{d+\lambda}\mathbb{E}_{\mu_h}[\|\phi(s_h,a_h)\|_2^2]\right)^2\\
&\overset{(a)}{\leqslant} \frac{(d+\lambda)^2C_e}{\lambda}\left(\mathbb{E}_{\mu_h}[\lambda_{\min}[(\Lambda_h^t)^{-1}]\|\phi(s_h,a_h)\|_2^2]\right)^2\\
&\overset{(b)}{\leqslant} \frac{(d+\lambda)^2C_e}{\lambda}\left(\mathbb{E}_{\mu_h}\mathrm{Tr}[\phi(s_h,a_h)\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}]\right)^2\\
&= \frac{(d+\lambda)^2}{d^2\gamma^2\lambda}\left(\mathbb{E}_{\mu_h}[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)]\right)^2\,,
\end{aligned}
$$

where (a) uses Eq. (24) and (b) uses the fact that $\mathrm{Tr}(\boldsymbol{AB}) \geqslant \lambda_{\min}(\boldsymbol{A})\mathrm{Tr}(\boldsymbol{B})$ for two PSD matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. □

Lemma 4 (regret decomposition)**.** *Consider linear MDP with the feature mapping $\phi(s,a) \in \mathbb{R}^d$ satisfying $\|\phi(s,a)\|_2 \leqslant 1$, under Assumption 3 with $\gamma > 0$, using LSVI-UCB with the regularization parameter λ and a bonus parameter $\beta := \widetilde{\mathcal{O}}\left(\sqrt{\lambda}H(d+\sqrt{\log\frac{1}{\delta}})\right)$ with $0<\delta<1$, then with probability at least $1-\delta$, for a underlying distribution μ, the regret admits*

$$\texttt{Regret}(T) \leqslant \frac{2\beta(d+\lambda)}{d\gamma\sqrt{\lambda}} \sum_{h=1}^{H}\sum_{t=1}^{T} \mathbb{E}_{\mu_h}[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)]\,,$$

where (s_h,a_h) is iid sampled from μ_h.

Proof. Recall the definition of β in LSVI-UCB (Jin et al., 2020) with $0<\delta<1$

$$\beta := \widetilde{\mathcal{O}}\left(\sqrt{\lambda}H\left(d+\sqrt{\log\frac{1}{\delta}}\right)\right)\,,$$

then according to (Jiang, 2022), with probability at least $1-\delta$, we have the following regret decomposition

$$\texttt{Regret}(T) \leqslant \sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\rho_h^{(t)}}\left[2\beta\|\phi(s_h,a_h)\|_{(\Lambda_h^t)^{-1}}\right]\,,$$

where (s_h,a_h) is sampled from the occupancy measure $\rho_h^{(t)}$. In the next, we conduct the change-of-measure from $\rho_h^{(t)}$ to μ_h, i.e.

$$\begin{aligned}
\texttt{Regret}(T) &\leqslant \sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\rho_h^{(t)}}\left[2\beta\|\phi(s_h,a_h)\|_{(\Lambda_h^t)^{-1}}\right]\\
&= \sum_{h=1}^{H}\sum_{t=1}^{T}2\beta\mathbb{E}_{\rho_h^{(t)}}\sqrt{\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)}\\
&\overset{(a)}{\leqslant} \sum_{h=1}^{H}\sum_{t=1}^{T}2\beta\sqrt{\mathbb{E}_{\rho_h^{(t)}}\left[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)\right]}\\
&= \sum_{h=1}^{H}\sum_{t=1}^{T}2\beta\sqrt{\operatorname{Tr}\left(\mathbb{E}_{\rho_h^{(t)}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\right]\right)}\\
&\overset{(b)}{\leqslant} \frac{2\beta(d+\lambda)}{d\gamma\sqrt{\lambda}}\sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\mu_h}[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)]\,,
\end{aligned}$$

where (a) uses Jensen inequality for the square-root function (concave); (b) uses Lemma 3. □

Now we are ready to prove Theorem 2.

Proof. Considering the iid sampling $(s_h^i,a_h^i) \sim \mu_h$ and $0 \leqslant \phi(s_h^i,a_h^i)^\top(\Lambda_h^t)^{-1}\phi(s_h^i,a_h^i) \leqslant \frac{1}{\lambda}$, denote $\sigma^2 := \mathbb{V}[\phi(s_h^i,a_h^i)^\top(\Lambda_h^t)^{-1}\phi(s_h^i,a_h^i)] \leqslant \frac{1}{4\lambda^2}$, then by Bernstein inequality (Wainwright, 2019), we have

$$\Pr\left[\left|\frac{1}{T}\sum_{i=1}^{T}[\phi(s_h^i,a_h^i)^\top(\Lambda_h^t)^{-1}\phi(s_h^i,a_h^i)] - \mathbb{E}_{\mu_h}[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)]\right| \geqslant \epsilon\right] \leqslant 2\exp\left(-\frac{T\epsilon^2}{2(\sigma^2+\epsilon/\lambda)}\right)\,.$$

That means, with probability at least $1-\delta_1$, we have

$$\mathbb{E}_{\mu_h}[\phi(s_h,a_h)^\top(\Lambda_h^t)^{-1}\phi(s_h,a_h)] \leqslant \frac{1}{T}\sum_{i=1}^{T}[\phi(s_h^i,a_h^i)^\top(\Lambda_h^t)^{-1}\phi(s_h^i,a_h^i)] + 4\sqrt{\frac{\sigma^2\log(2/\delta_1)}{T}} + \frac{4\log(2/\delta_1)}{T\lambda}\,. \tag{26}$$

Combining Eq. (26) and Lemma 2 into Lemma 4, for any $\delta \in (0,1)$ and taking $\delta_1 := \delta/2$ and $\beta := \widetilde{\mathcal{O}}(\sqrt{\lambda} dH \log(2/\delta))$, with probability at least $1-\delta$, we have

$$
\begin{aligned}
\texttt{Regret}(T) &\lesssim \frac{\beta(d+\lambda)}{d\gamma\sqrt{\lambda}} \sum_{h=1}^{H}\sum_{t=1}^{T} \left(\frac{d^2}{T\lambda}\log(T+1) + 4\sqrt{\frac{\sigma^2 \log(4/\delta)}{T}} + \frac{4\log(4/\delta)}{T\lambda} \right) \\
&\lesssim \left(\frac{d+\lambda}{\gamma\lambda} d^2 H^2 \log T + \frac{d+\lambda}{\gamma} H^2 \sum_{t=1}^{T} \sqrt{\frac{\sigma^2}{T}} \right) \log\left(\frac{4}{\delta}\right) .
\end{aligned}
\tag{27}
$$

Using $\sigma^2 \lesssim \frac{1}{\lambda^{2\alpha}}$ with $\alpha > 1$ in Assumption 4 and taking $\lambda := T^{\eta}$ with $\eta \in (0,1]$ back to the above regret bound, with probability at least $1-\delta$, we have

$$
\begin{aligned}
\texttt{Regret}(T) &\lesssim \left(\frac{H^2 d^2}{\gamma} \log T + \frac{H^2 \lambda \sigma}{\gamma} \sqrt{T} + \frac{H^2 d}{\gamma} \sigma \sqrt{T} \right) \log\left(\frac{4}{\delta}\right) \\
&\lesssim \mathcal{O}\left(\frac{dH^2}{\gamma} \left(d\log T + T^{\frac{1}{2}-\eta(\alpha-1)} \right) \right) \\
&= \begin{cases} \mathcal{O}\left(\dfrac{d^2H^2}{\gamma} \log T \right), & \text{if } \eta(\alpha-1) \geqslant 1/2 \\ \mathcal{O}\left(\dfrac{dH^2}{\gamma} T^{\frac{1}{2}-\eta(\alpha-1)} \right), & \text{if } \eta(\alpha-1) \in (0, \frac{1}{2}) \end{cases}
\end{aligned}
$$

which concludes the proof. □

E.2. Discussion on the regularization parameter

Recall the regularized least squares in Eq. (22), it is equivalent to

$$
\widehat{\boldsymbol{w}}_h^t \leftarrow \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{t-1} \sum_{\tau=1}^{t-1} [r_h(s_h^\tau, a_h^\tau) + \max_{a\in\mathcal{A}} Q_{h+1}^t(s_{h+1}^\tau, a) - \langle \boldsymbol{w}, \phi(s_h^\tau, a_h^\tau) \rangle]^2 + \lambda' \|\boldsymbol{w}\|_2^2 \,,
$$

where $\lambda' = \frac{\lambda}{t-1}$. The first term is the empirical risk minimization and the second term is the regularizer as Tikhonov regularization. The regularization parameter $\lambda' \equiv \lambda'(t) > 0$ admits $\lim_{t\to\infty} \lambda'(t) = 0$. In learning theory, one typically assumes that $\lambda' = \mathcal{O}(t^{-\tau})$ with $\tau \in (0,1]$, decaying with the number of samples (Cucker and Zhou, 2007), which implies $\lambda = \mathcal{O}(t^{1-\tau})$ in Eq. (22). This verifies that our assumption on the regularization parameter makes sense. In LSVI-UCB (Jin et al., 2020), the regularization parameter is chosen as $\lambda = 1$, which implies $\lambda' = 1/t$.

In our problem, we denote $\eta := 1-\tau$ and directly choose $\lambda = \mathcal{O}(T^{\eta})$ with $\eta \in (0,1]$, independent of the number of state-action pairs $t-1$. We need to remark that, if we choose a more reasonable $\lambda = \mathcal{O}(t^{\eta})$ with $\eta \in (0,1]$, depending on the number of samples, we can still obtain the same regret as Theorem 2. To be specific, the regret bound in Eq. (27) is reformulated as (w.h.p)

$$
\begin{aligned}
\texttt{Regret}(T) &\lesssim \frac{d+\lambda}{\gamma\lambda} d^2 H^2 \log T + \frac{d+\lambda}{\gamma} H^2 \sum_{t=1}^{T} \sqrt{\frac{\sigma^2}{T}} \\
&\lesssim \frac{d^2H^2}{\gamma} \log T + \frac{dH^2}{\gamma} T^{-\frac{1}{2}} \int_1^T t^{-\eta(\alpha-1)} \mathrm{d}t \\
&= \begin{cases} \mathcal{O}\left(\dfrac{d^2H^2}{\gamma} \log T \right), & \text{if } \eta(\alpha-1) \geqslant 1/2 \\ \mathcal{O}\left(\dfrac{dH^2}{\gamma} T^{\frac{1}{2}-\eta(\alpha-1)} \right), & \text{if } \eta(\alpha-1) \in (0, \frac{1}{2}) \,. \end{cases}
\end{aligned}
$$

That means, there is no difference between these two regularization schemes whether it varies with the number of state-action pairs.

Besides, it appears that if we take $\eta = 0$, the regularization parameter λ is in a constant order, i.e., $\lambda' = \mathcal{O}(1/t)$, decaying fast, we cannot improve the regret rate beyond $\widetilde{\mathcal{O}}(1/\sqrt{T})$. It does not make sense in practice. Here we illustrate this to resolve this issue.

The main reason is, our low variance assumption 4 is based on λ. In our theorem, we require $\lambda = T^{\eta}$ with $\eta \in (0, 1]$, which makes the feature mapping $\|\phi_h(s_h, a_h)\|^2_{(\Lambda_h^t)^{-1}}$ concentrate around its mean and decay with the episode T. If we take $\eta = 0$, the constant order of λ does not make $\|\phi_h(s_h, a_h)\|^2_{(\Lambda_h^t)^{-1}}$ decaying with the episode T, and accordingly Assumption 4 does not work. In this case, there is no need to use λ as a bridge in our assumption. Instead, we can directly set $M - m$ small, decaying with T under some certain distribution.

F. Auxiliary lemma

In this section, we list some auxiliary lemmas that are needed for our proof.

Lemma 5. *(Jin et al., 2021a, Lemmas 39 and 40) Under Assumptions 1 and 2, for any $\delta \in (0, 1)$, if we choose $\beta = c \log\left(\frac{\mathcal{N}_{\mathcal{F}}(1/T)TH}{\delta}\right)$ in the GOLF algorithm 1 for some large constant c, with probability at least $1 - \delta$, we have*

- *$Q^\star \in \mathcal{F}^{(t)}$.*
- *$\sum_{i<t} \mathbb{E}_{(s,a)\sim\rho_h^{(i)}}[f_h(s, a) - \mathcal{T}_h f_{h+1}(s, a)]^2 \lesssim \mathcal{O}(\beta)$ for any $f \in \mathcal{F}^{(t)}$.*

Lemma 6. *(Song et al., 2022, Bellman error bound for FQI, Lemma 7) Let $\delta \in (0, 1)$, for any $h \in [H]$ and $t \in [T]$, f_h^{t+1} be the estimated value function computed by the least square regression using samples from $\mathcal{D}_h^\nu \bigcup \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)_{\tau=1}^t\}$ in Algorithm 2, then with probability at least $1 - \delta$, for any $h \in [H - 1]$ and $t \in [T]$, we have*

$$\mathbb{E}_{\mu_h}\left(\delta_h^{(t)}(s, a)\right)^2 \lesssim \frac{1}{n_{\text{off}}} \log\left(\frac{\mathcal{N}_{\mathcal{F}}(1/T)TH}{\delta}\right) .$$

Lemma 7. *(Xie et al., 2023, Per-state-action elliptic potential lemma, modified version) Let $\rho^{(1)}, \rho^{(2)}, \ldots, \rho^{(T)}$ be an arbitrary sequence of distributions over a set $\mathcal{Z}$ (e.g., $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$), and let $\mu \in \Delta(\mathcal{Z})$ be a distribution such that $\rho^{(t)}(z) \leqslant [C\mu(z)]^p$ for some $p \geqslant 1$ and all $(z, t) \in \mathcal{Z} \times [T]$. Then for all $z \in \mathcal{Z}$, we have*

$$\sum_{t=1}^{T} \frac{d^{(t)}(z)}{\sum_{i<t} d^{(i)}(z) + C \cdot \mu(z)} \leq \mathcal{O}\left(\log\left(T\right)\right) .$$