Adaptive and Optimal Offline Reinforcement Learning Bound

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the *offline reinforcement learning* (offline RL) problem, where the goal is to learn a reward-maximizing policy in an unknown *Markov Decision Process* (MDP) using the data coming from a policy μ. In particular, we consider the sample complexity problems of offline RL for the finite horizon MDPs. Prior works derive the information-theoretical lower bounds based on different data-coverage assumptions and their upper bounds are expressed by the covering coefficients which lack the explicit characterization of system quantities. In this work, we analyze the *Adaptive Pessimistic Value Iteration* (APVI) algorithm and derive the suboptimality upper bound that nearly matches

$$O\left(\sum_{h=1}^{H}\sum_{s_h,a_h} d_h^{\pi^\star}(s_h,a_h)\sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^\star + r_h)}{d_h^\mu(s_h,a_h)}}\sqrt{\frac{1}{n}}\right). \quad (1)$$

We also prove an information-theoretical lower bound to show this quantity is required under the weak assumption that $d_h^\mu(s_h,a_h) > 0$ if $d_h^{\pi^\star}(s_h,a_h) > 0$. Here $\pi^\star$ is a optimal policy, μ is the behavior policy and $d(s_h,a_h)$ is the marginal state-action probability. We call this adaptive bound the *intrinsic offline reinforcement learning bound* since it directly implies all the existing optimal results: minimax rate under uniform data-coverage assumption, horizon-free setting, single policy concentrability, and the tight problem-dependent results. Later, we extend the result to the *assumption-free* regime (where we make no assumption on μ) and obtain the assumption-free intrinsic bound. Due to its generic form, we believe the intrinsic bound could help illuminate what makes a specific problem hard and reveal the fundamental challenges in offline RL.

1 Introduction

In *offline reinforcement learning* (offline RL Levine et al. [2020], Lange et al. [2012]), the goal is to learn a reward-maximizing policy in an unknown environment (*Markov Decision Process* or MDP for short) using the historical data coming from a (fixed) behavior policy μ. Unlike online RL, where the agent can keep interacting with the environment and gain new feedback by exploring unvisited state-action space, offline RL usually populates when such online interplays are expensive or even unethical. Due to its nature of without the access to interact with the MDP model, most of the literature that studies the sample complexity / provable efficiency of offline RL Le et al. [2019], Chen and Jiang [2019], Xie and Jiang [2020b,a], Yin et al. [2021a,b], Ren et al. [2021], Rashidinejad et al. [2021] relies on making different data-coverage assumptions to make their problems learnable and provide the near-optimal worst-case performance bounds that depend on their data-coverage coefficients. Those results are valuable in general as they do not depend on the structure of the particular problem, therefore, remain valid even for pathological MDPs. But is this good enough?

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In practice, the empirical performances of offline reinforcement learning (*e.g.* Gulcehre et al. [2020], Fu et al. [2020, 2021]) are often far better than what those non-adaptive / problem-independent bounds would indicate. Although empirical evidence can help explain why we may observe better or worse performances on different MDPs, a systematic understanding of what types of decision processes and what kinds of behavior policies are inherently easier or more challenging for offline RL is lacking. Besides, despite the fact that a non-adaptive bound can learn even the pathological examples within the assumption family, there is no guarantee for the instances outside the family. However, practical offline reinforcement learning problems are usually beyond the scope of certain data-coverage assumptions, which limits the applicability of those results. Can we make as few assumptions as possible? Or even more, what can we guarantee when no assumption is made about offline learning?

Those motivate us to derive the provably efficient bounds that are adaptive to the individual instances but only require minimal assumptions so they can be widely applied in most cases. Ideally, such bounds should characterize the system structures of the specific problems, hold even for peculiar instances that do not satisfy the standard data-coverage assumptions, and recover the worst-case guarantees when the assumptions are satisfied. As mentioned in Zanette and Brunskill [2019], a fully adaptive characterization in RL is important as it might bring considerable saving in the time spent designing domain-specific RL solutions and in training a human expert to judge and recognize the complexity of different problems.

1.1 Our contribution

In this work, we provide the analysis for our *adaptive pessimistic value iteration* (APVI) (Algorithm 1) in the finite horizon time-variant (non-stationary) MDP case and derive a strong adaptive bound that is near-optimal under the weak assumption $d_h^{\mu}(s_h, a_h) > 0$ if $d_h^{\pi^\star}(s_h, a_h) > 0$ (Theorem 4.1). Specifically, our bound (quantity (1)) explicitly depends on the marginal importance ratios (between the optimal policy $\pi^\star$ and the behavior policy μ) and the per-step conditional variances. In addition, we provide an adaptive minimax lower bound (Theorem 4.2) to show (1) is optimal for offline learning and call it *the intrinsic offline learning bound*. The intrinsic bound has the following consequences.

- In the non-adaptive domain (4.1-4.3), the intrinsic bound implies $\widetilde{O}(H^3/d_m\epsilon^2)$ complexity under the uniform data-coverage 2.1, $\tilde{O}(H^3SC^\star/\epsilon^2)$ complexity under the single policy concentrability assumption 2.3 and $\widetilde{O}(H/d_m\epsilon^2)$ complexity when the sum of rewards is bounded by 1. All of those are optimal in their respectively regimes [Yin et al., 2021a, Rashidinejad et al., 2021, Ren et al., 2021];
- In the adaptive domain (4.4), the intrinsic bound implies the tight problem-dependent counterpart of Zanette and Brunskill [2019], yields $\tilde{O}(H^3/nd_m)$ fast convergence in the deterministic systems, has improved complexity in the partially deterministic systems and a family of highly mixing problems, and remains optimal when reducing to the tabular contextual bandits.

Beyond the above, due to the generic form of the intrinsic bound, we could come up with as many problem instances (that are of our interests) as possible and study their properties. In this sense, the intrinsic bound helps illuminate the fundamental nature of offline RL.

Furthermore, as a step towards *assumption-free* offline reinforcement learning, we build a modified AVPI and obtain an adaptive bound that could characterize the suboptimality gap in the state-action space that is agnostic to the behavior policy (Theorem 5.1). To the best of our knowledge, all of these results are the first of its kinds.

1.2 Related work

Finite sample analysis for offline reinforcement learning can be traced back to Szepesvári and Munos [2005], Antos et al. [2008a,b] for the *infinite horizon discounted setting* via Fitted Q-Iteration (FQI) type function approximation algorithms. [Chen and Jiang, 2019, Le et al., 2019, Xie and Jiang, 2020a,b] follow this line of research and include the information-theoretical considerations. Recently, Xie and Jiang [2020a] considers the offline RL with only the realizability assumption, Liu et al. [2020] considers the offline RL with limited exploration and Kidambi et al. [2020] uses the model-based

approach for addressing offline RL. In those challenging scenarios, their finite sample analysis are suboptimal (*e.g.* in terms of the effective horizon $(1-\gamma)^{-1}$). Recently, Yin et al. [2021a,b], Ren et al. [2021] study the finite horizon case. In the linear MDP case, Jin et al. [2020] studies the pessimistic algorithm for offline policy learning under the weak compliance assumption and Wang et al. [2020], Zanette [2020] prove some exponential lower bounds under their respective linear MDP assumptions.

Among them, there are a few works that achieve the sample optimality under their respective assumptions. Under the uniform data coverage (minimal state-action probability $d_m > 0$), Yin et al. [2021a] first proves the optimal $\tilde{O}(H^3/d_m\epsilon^2)$ complexity in the time-inhomogeneous MDP. Recently, Yin et al. [2021b] designs the offline variance reduction algorithm to achieve the optimal $\tilde{O}(H^2/d_m\epsilon^2)$ rate for the time-homogeneous case. Under the setting where the total cumulative reward is bounded by 1, Ren et al. [2021] obtains the horizon-free result with $\tilde{O}(1/d_m)$. More recently, Rashidinejad et al. [2021] considers the single concentrability coefficient $C^\star := \max_{s,a} d^{\pi^\star}(s,a)/d^\mu(s,a)$ and derives the upper bound $\tilde{O}[(1-\gamma)^{-5}SC^\star/\epsilon^2]$ in the infinite horizon setting. While those worst-case guarantees are desirable, none of them can explain the hardness of the individual problems.

Additional paper organization. We review the basic RL setting and problem setup in Section 2, discuss the Vanilla Pessmistic Value Iteration in Section 3 as a warm-up example, introduce the intrinsic offline reinforcement leaning bound in Section 4. We extend to the assumption-free offline RL in Section 5, put the proof sketch in Section 6, and conclude in Section 7. The complete proofs of the main paper are deferred to Appendix.

2 Preliminaries

Episodic non-stationary (time-variant) reinforcement learning. A finite-horizon *Markov Decision Process* (MDP) is denoted by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, H, d_1)$ [Sutton and Barto, 2018], where $\mathcal{S}$ is the finite state space and $\mathcal{A}$ is the finite action space with $S := |\mathcal{S}| < \infty, A := |\mathcal{A}| < \infty$. A non-stationary transition kernel $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0,1]$ maps each state action(s_h, a_h) to a probability distribution $P_h(\cdot|s_h, a_h)$ and P_h can be different across the time. Besides, $r : \mathcal{S} \times A \mapsto \mathbb{R}$ is the expected instantaneous reward function satisfying $0 \le r \le 1$. d_1 is the initial state distribution. H is the horizon. A policy $\pi = (\pi_1, \ldots, \pi_H)$ assigns each state $s_h \in \mathcal{S}$ a probability distribution over actions according to the map $s_h \mapsto \pi_h(\cdot|s_h)\,\forall h \in [H]$. An MDP together with a policy π induce a random trajectory $s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1}$ with $s_1 \sim d_1, a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_)$, $\forall h \in [H]$ and r_h is a random realization given the observed s_h, a_h.

Q-values, Bellman (optimality) equations. The value function $V_h^\pi(\cdot) \in \mathbb{R}^S$ and Q-value function $Q_h^\pi(\cdot,\cdot) \in \mathbb{R}^{S\times A}$ for any policy π is defined as: $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h = s]$, $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h, a_h = s, a]$, $\forall s, a \in \mathcal{S}, \mathcal{A}, h \in [H]$. The performance is defined as $v^\pi := \mathbb{E}_{d_1}[V_1^\pi] = \mathbb{E}_{\pi,d_1}\left[\sum_{t=1}^H r_t\right]$, where we denote V_h^π, Q_h^π as column vectors and $P_h \in \mathbb{R}^{SA\times S}$ the transition matrix, then the vector form Bellman (optimality) equations follow $\forall h \in [H]$: $Q_h^\pi = r_h + P_h V_{h+1}^\pi$, $V_h^\pi = \mathbb{E}_{a\sim\pi_h}[Q_h^\pi]$, $Q_h^\star = r_h + P_h V_{h+1}^\star$, $V_h^\star = \max_a Q_h^\star(\cdot, a)$.In addition, we denote the per-step marginal state-action occupancy $d_h^\pi(s,a)$ as: $d_h^\pi(s,a) := \mathbb{P}[s_h = s | s_1 \sim d_1, \pi] \cdot \pi_h(a|s)$, which is the marginal state-action probability at time h.

Offline setting and the goal. The offline RL requires the agent to find a policy π such that the performance v^π is maximized, given only the episodic data $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau\in[n]}^{h\in[H]}$ rolled out from some behavior policy μ. The offline nature requires we cannot change μ and in particular we do not assume the functional knowledge of μ. That is to say, given the batch data $\mathcal{D}$ and a targeted accuracy $\epsilon > 0$, the offline RL seeks to find a policy π_{alg} such that $v^\star - v^{\pi_{\text{alg}}} \le \epsilon$.

2.1 Assumptions in offline RL

We revise several types of assumptions proposed by existing studies that can yield provably efficient results. Recall $d_h^\mu(s_h, a_h)$ is the marginal state-action probability and μ is the behavior policy.

Assumption 2.1 (Uniform data coverage Yin et al. [2021a])**.** *The behavior policy obeys that* $d_m := \min_{h,s_h,a_h} d_h^\mu(s_h, a_h) > 0$. *Here the infimum is over all the states satisfying there exists certain policy so that this state can be reached by the current MDP with this policy.*

This is the strongest assumption in offline RL as it requires μ to explore each state-action pairs with positive probability. Under 2.1, it mostly holds $1/d_m \geq SA$. This reveals offline learning is generically harder than *the generative model setting* [Agarwal et al., 2020] in the statistical sense. On the other hand, this is required for their *uniform OPE* task as it seeks to simultaneously evaluate all the policies within the policy class and could be a much harder problem than the offline learning task.

Assumption 2.2 (Uniform concentrability Szepesvári and Munos [2005], Chen and Jiang [2019])**.** *For all the policies,* $C_\mu := \sup_{\pi,h} ||d_h^\pi(\cdot,\cdot)/d_h^\mu(\cdot,\cdot)||_\infty < \infty$.

This is a classical offline RL condition that works well with the function approximation scheme (*e.g.* Fitted Q-Iteration). Qualitatively, this is a uniform data-coverage assumption that is similar to Assumption 2.1, but quantitatively, the coefficient C_μ can be smaller than $1/d_m$ due the d_h^π term in the numerator. We point out Xie and Jiang [2020a] also follow this assumption (due to their Proposition 1) even though they only assume realizability for function approximation.

Assumption 2.3. *There exists at least one optimal policy* $\pi^\star$*, s.t.* $\forall s_h, a_h \in \mathcal{S}, \mathcal{A}$, $d_h^\mu(s_h, a_h) > 0$ *if* $d_h^{\pi^\star}(s_h, a_h) > 0$*. We further denote the trackable set as* $\mathcal{C}_h := \{(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$.

This is assumed in Liu et al. [2019]. Assumption 2.3 is (arguably) the weakest assumption needed for accurately learning the optimal value $v^\star$ and we will use 2.3 for most parts of this paper. It only requires μ to trace the state-action space of one optimal policy and can be agnostic at other locations. Rashidinejad et al. [2021] essentially considers this assumption but their analysis is based on the single concentrability coefficient $C^\star := \max_{s,a} d^{\pi^\star}(s,a)/d^\mu(s,a)$. The dependence on $C^\star$ makes their result less adaptive since (usually) there are a lot of locations that have the ratio $d^{\pi^\star}(s,a)/d^\mu(s,a)$ to be much smaller than $C^\star$. Furthermore, what could we end up with when 2.3 is not met? We will provide our answers for all the assumptions in the subsequent sections.

3 A warm-up case study: Vanilla Pessimistic Value Iteration

As a step towards the optimal and strong adaptive offline RL bound, we analyze *the vanilla pessimistic value iteration* (VPVI), a tabular counterpart of *pessimistic value iteration* (PEVI initiated in Jin et al. [2020]), to understand what is missing for achieving the fully adaptivity. In particular, VPVI relies on the model-based construction.

Model-based Components. Given data $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau\in[n]}^{h\in[H]}$, we denote $n_{s_h,a_h} := \sum_{\tau=1}^n \mathbf{1}[s_h^\tau, a_h^\tau = s_h, a_h]$ be the total counts that visit (s_h, a_h) pair at time h, then we use the offline plug-in estimator constructs the estimators for P_h and r_h as:

$$\widehat{P}_h(s'|s_h,a_h) = \frac{\sum_{\tau=1}^n \mathbf{1}[(s_{h+1}^\tau, a_h^\tau, s_h^\tau) = (s', s_h, a_h)]}{n_{s_h,a_h}},\ \widehat{r}_h(s_h,a_h) = \frac{\sum_{\tau=1}^n \mathbf{1}[(a_h^\tau, s_h^\tau) = (s_h, a_h)] \cdot r_h^\tau}{n_{s_h,a_h}}, \tag{2}$$

if $n_{s_h,a_h} > 0$ and $\widehat{P}_h(s'|s_h,a_h) = 1/S, \widehat{r}_h(s_h,a_h) = 0$ if $n_{s_h,a_h} = 0$. In particular, we use the word "vanilla" as it directly mirrors Jin et al. [2020] with a pessimistic penalty of order $O(H/\sqrt{n_{s_h,a_h}})$.[1]

With $\widehat{P}_h, \widehat{r}_h$ in Algorithm 2 (which we defer to Appendix), VPVI guarantees the following:

Theorem 3.1. *Under the Assumption 2.3, denote* $\bar{d}_m := \min_{h\in[H]}\{d_h^\mu(s_h,a_h) : d_h^\mu(s_h,a_h) > 0\}$*. For any* $0 < \delta < 1$*, there exists absolute constants* $c_0, C' > 0$*, such that when* $n > c_0 \cdot 1/\bar{d}_m \cdot \iota$ *(*$\iota = \log(HSA/\delta)$*), with probability* $1-\delta$*, the output policy* $\widehat{\pi}$ *of VPVI satisfies*

$$0 \leq v^\star - v^{\widehat{\pi}} \leq C' H \sum_{h=1}^H \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\frac{\iota}{n \cdot d_h^\mu(s_h,a_h)}}. \tag{3}$$

The full proof can be found in Appendix B. Theorem 3.1 makes some improvements over the existing works. First, it is more adaptive than the regime with uniform data-coverage Assumption 2.1 (Yin et al. [2021a], Ren et al. [2021]). In addition, by straightforward calculation (3) can be bounded by $\tilde{O}(\sqrt{H^4SC^\star/n})$ which improves VI-LCB [Rashidinejad et al., 2021] by a factor of H.[2] Besides,

[1] This is due to $\sqrt{\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)}$ reduces to $\sqrt{1/n_{s_h,a_h}}$ when setting $\phi(s_h,a_h) = \mathbf{1}(s_h,a_h)$.

[2] To be rigorous, translating the result from the infinite horizon setting to the finite horizon setting requires explanation. We add this discussion in Appendix C.

the analysis of VPVI also improves the direct reduction of PEVI [Jin et al., 2020] in the tabular case by a factor SA since their $\beta = SAH$.

Despite all that, VPVI is not optimal as the dependence on horizon is H^4 which does not match the optimal worst case guarantee H^3 [Yin et al., 2021a] in the nonstationary setting. Also, the explicit dependence on H in (3) possibly hides some key features of the specific offline RL instances. For example, no improvement can be made if the system has the deterministic transition.

Algorithm 1 Adaptive (*assumption-free*) Pessimistic Value Iteration

Input: Offline dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau,h=1}^{n,H}$. Set $C_1 = 2, C_2 = 14$, failure probability δ.
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$. Set $\iota = \log(HSA/\delta)$. (if assumption-free, set $M^\dagger, \widehat{M}^\dagger$ as in Section 5.)
for time $h = H, H-1, \ldots, 1$ **do**
  Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \widehat{r}_h(\cdot,\cdot) + (\widehat{P}_h \cdot \widehat{V}_{h+1})(\cdot,\cdot)$ (use $\widehat{r}_h^\dagger + (\widehat{P}_h^\dagger \cdot \widehat{V}_{h+1})$ if assumption-free)
  $\forall s_h, a_h$, set $\Gamma_h(s_h,a_h) = C_1\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{s_h,a_h}}(\widehat{r}_h + \widehat{V}_{h+1})\cdot\iota}{n_{s_h,a_h}}} + \frac{C_2 H\cdot\iota}{n_{s_h,a_h}}$ if $n_{s_h,a_h} \geq 1$, o.w. set to $\frac{CH\iota}{1}$.
  (If assumption-free, use $C_1\sqrt{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1})\cdot\iota / n_{s_h,a_h}} + \frac{C_2 H\cdot\iota}{n_{s_h,a_h}}$ if $n_{s_h,a_h} \geq 1$, o.w. use 0.)
  Set $\widehat{Q}_h^p(\cdot,\cdot) \leftarrow \widehat{Q}_h(\cdot,\cdot) - \Gamma_h(\cdot,\cdot)$. Set $\overline{Q}_h(\cdot,\cdot) \leftarrow \min\{\widehat{Q}_h^p(\cdot,\cdot), H-h+1\}^+$. // Pessmistic update
  $\forall s_h$, Select $\widehat{\pi}_h(\cdot|s_h) \leftarrow \mathrm{argmax}_{\pi_h}\langle \overline{Q}_h(s_h,\cdot), \pi_h(\cdot|s_h)\rangle$. Set $\widehat{V}_h(s_h) \leftarrow \langle \overline{Q}_h(s_h,\cdot), \widehat{\pi}_h(\cdot|s_h)\rangle$.
end for
Output: $\{\widehat{\pi}_h\}$.

4 Intrinsic Offline Reinforcement Learning bound

Now we go deeper to understand what is the more intrinsic characterization of the offline reinforcement learning. From the study of VPVI, penalizing the Q-function by $\widetilde{O}(H/\sqrt{n_{s_h,a_h}})$ is crude as it estimates the confidence width of $\widehat{Q}_h$ in Algorithm 2 too conservatively therefore loses the accuracy (the bound is suboptimal). This motivates us to use empirical standard deviation instead to create a more adaptive (and also less conservative) Bernstein-type confidence width as the pessimistic penalty:

$$\Gamma_h(s_h,a_h) = \widetilde{O}\left[\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{s_h,a_h}}(\widehat{r}_h + \widehat{V}_{h+1})}{n_{s_h,a_h}}} + \frac{H}{n_{s_h,a_h}}\right] \text{ (if } n_{s_h,a_h} > 0); \; = \widetilde{O}(H) \text{ (if } n_{s_h,a_h} = 0). \quad (4)$$

and update $\widehat{Q}_h \leftarrow \widehat{Q}_h - \Gamma_h$. On one hand, $\sqrt{\mathrm{Var}_{\widehat{P}_{s_h,a_h}}(\widehat{r}_h + \widehat{V}_{h+1})/n_{s_h,a_h}}$ is a "less pessimistic" penalty than VPVI due to $\sqrt{\mathrm{Var}_{\widehat{P}}(\widehat{r}_h + \widehat{V}_{h+1})} \leq H$ and critically this design is more data-adaptive since it holds negative view towards the locations with high uncertainties and recommends the locations that we are confident about, as opposed to the online RL (which encourages exploration in the uncertain locations). Such principles are not reflected by the isotropic design in VPVI. On the other hand, it carries the extremely negative view towards fully agnostic locations $\widetilde{O}(H)$ which in turn causes the agent unlikely to choose them. We summarized the this *adaptive pessimistic value iteration* (APVI) into the Algorithm 1, with $\widehat{P}_h, \widehat{r}_h$ defined in (2). APVI has the following guarantee. A sketch of the analysis is presented in Section 6 and Appendix E includes the full proof.

Theorem 4.1 (Intrinsic offline RL bound)**.** *Under the Assumption 2.3, denote* $\bar{d}_m := \min_{h\in[H]}\{d_h^\mu(s_h,a_h) : d_h^\mu(s_h,a_h) > 0\}$. *For any* $0 < \delta < 1$, *there exists absolute constants* $c_0, C' > 0$, *such that when* $n > c_0 \cdot 1/\bar{d}_m \cdot \iota$ ($\iota = \log(HSA/\delta)$), *with probability* $1-\delta$, *the output policy* $\widehat{\pi}$ *of APVI (Algorithm 1) satisfies ($\widetilde{O}$ hides log factor and higher order terms)*

$$0 \leq v^\star - v^{\widehat{\pi}} \leq C' \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)\cdot\iota}{n \cdot d_h^\mu(s_h,a_h)}} + \widetilde{O}\left(\frac{H^3}{n\cdot\bar{d}_m}\right) \quad (5)$$

APVI makes significant improvements in the a lot of aspects. First and foremost, the dominate term is fully expressed by the system quantities that admits no explicit dependence on H, S, A. To the best of our knowledge, this is the first offline RL bound that concretely depicts the interrelations

within the problem when the problem instance is a tuple $(M, \pi^\star, \mu)$: an MDP M (coupled with the optimal policy $\pi^\star$) with the data rolling from an offline logging policy μ. As we will discuss later, this result indicates (nearly) all the optimal worst-case non-adaptive bounds (and clearly also the VPVI) under their respective regimes / assumptions. Thus, (5) is generic. More interestingly, Theorem 4.1 caters to the specific MDP structures and adaptively yields improved sample complexities (*e.g.* faster convergence in deterministic systems) that existing works cannot imply. Such features are crucial as it helps us to understand what type of problems are harder / easier than others, and even more, in a *quantitative* way. Also, to illustrate this bound exhibits some intrinsic nature of offline RL, we provide the following adaptive minimax lower bound that matches the main term of Theorem 4.1. The proof and the sketch can be found in Appendix F and Section 6 respectively.

Theorem 4.2 (Adaptive minimax lower bound)**.** *Recall for each individual instance (μ, M), $\bar{d}_m := \min_{h\in[H]}\{d_h^\mu(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$ and $\mathcal{D}$ consists of n episodes. Now consider a class of problem family $\mathcal{G} := \{(\mu, M) : \exists \pi^\star \ s.t. \ d_h^\mu(s, a) > 0 \ if \ d_h^{\pi^\star}(s, a) > 0\}$. Let $\widehat{\pi}$ be the output of any algorithm on $\mathcal{D}$. Then there exists universal constants $c_0, p, C > 0$, such that if $n \geq c_0 \cdot 1/\bar{d}_m \cdot \log(HSA/p)$,*

$$\inf_{\widehat{\pi}} \sup_{(\mu,M)\in\mathcal{G}} \frac{\mathbb{E}_{\mu,M}\left[v^\star - v^{\widehat{\pi}}\right]}{\sum_{h=1}^H \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h, a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)}{n \cdot d_h^\mu(s_h,a_h)}}} \geq C. \tag{6}$$

Theorem 4.2 states the denominator in (6) is the statistical hardness of offline reinforcement learning and this certifies Theorem 4.1 is not only adaptive but also optimal. Hence, we call the quantity $\sum_{h=1}^H \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h, a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)}{n \cdot d_h^\mu(s_h,a_h)}}$ *intrinsic offline reinforcement learning bound*. In the sequel, we provide thorough discussions to explain the intrinsic bound embraces the fundamental challenges in offline RL and the strong adaptivity. The formal technical derivations that are missing in Section 4.1-4.4 are deferred to Appendix G.

Remark 4.3. *In the tabular case, our lower bound improves over the direct reduction of Jin et al. [2020] Theorem 4.6 as their lower bound is $\sum_{h=1}^H \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h, a_h) \cdot \sqrt{\frac{1}{n \cdot d_h^\mu(s_h,a_h)}}$, which deviates Theorem 3.1 by a factor of H and does not achieve the minimaxity. Our Theorem 4.1,4.2 are minimax optimal. In this regard, the tight information-theoretical limit for the linear MDP setting remains elusive.*

4.1 Optimality under Uniform data-coverage assumption

Under the uniform exploration Assumption 2.1 with parameter $d_m := \min_{h,s_h,a_h} d_h^\mu(s_h, a_h) > 0$, Yin et al. [2021a] uses a purely model-based plug-in approach to obtain the optimal sample complexity $\widetilde{O}(H^3/d_m\epsilon^2)$ and shows $\Omega(H^3/d_m\epsilon^2)$ is also the lower bound. Indeed, this can be directly implied by the intrinsic RL bound via *Cauchy inequality* and *the Sum of Total Variance* (Lemma H.6):[3]

$$\begin{aligned}
&\sum_{h=1}^H \langle d_h^{\pi^\star}(\cdot), \sqrt{\frac{\mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^\star)}{n \cdot d_h^\mu(\cdot)}}\rangle \leq \sum_{h=1}^H \langle \sqrt{d_h^{\pi^\star}(\cdot)}, \sqrt{\frac{d_h^{\pi^\star}(\cdot) \odot \mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^\star)}{n \cdot d_m}}\rangle \\
&\leq \sum_{h=1}^H \left\|\sqrt{d_h^{\pi^\star}(\cdot)}\right\|_2 \left\|\sqrt{\frac{d_h^{\pi^\star}(\cdot) \odot \mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^\star)}{n \cdot d_m}}\right\|_2 \leq \sqrt{\frac{H \cdot \mathrm{Var}_{\pi^\star}(\sum_{h=1}^H r_h)}{n \cdot d_m}} \leq \sqrt{\frac{H^3}{n \cdot d_m}}
\end{aligned} \tag{7}$$

which translates to $\widetilde{O}(H^3/d_m\epsilon^2)$ complexity. Our result maintains the optimal worst-case guarantee when μ has the uniform data-coverage:

Proposition 4.4. *Under Assumption 2.1 and apply Theorem 4.1, APVI achieves the sample complexity of minimax-rate $\widetilde{O}(H^3/d_m\epsilon^2)$ (Theorem 4.1 and Theorem G.2 in Yin et al. [2021a]).*

Remark 4.5. *We believe if the MDP is time-invariant, then by a modified construction of $\widehat{P}$, $\widehat{r}$ in (2) our result will imply the minimax-rate of $\widetilde{O}(H^2/d_m\epsilon^2)$ as studied in Yin et al. [2021b]. Due to the space constraint, we include this discussion in Appendix G.*

[3] Here $\odot$ denotes element-wise multiplication. Also note under 2.1, our $\bar{d}_m = d_m$.

4.2 Bounded sum of total rewards and the Horizon-Free case

There is another thread of studies that follow the bounded sum of total rewards assumption: *i.e.* $r_h \geq 0, \sum_{h=1}^H r_h \in [0,1]$ [Krishnamurthy et al., 2016, Jiang et al., 2017, Zhang et al., 2020]. Such a setting is much weaker than the uniform bounded instantaneous reward condition, as explained in Jiang and Agarwal [2018]. In offline RL, Ren et al. [2021] derives the nearly horizon-free worst case bound $\widetilde{O}(\sqrt{1/nd_m})$ for the time-invariant MDPs, under the Assumption 2.1. As a comparison, our Theorem 4.1 achieves the following guarantee for the time-variant (non-stationary) MDPs.

Proposition 4.6. *Assume $r_h \geq 0$, $\sum_{h=1}^H r_h \leq 1$. Then in the time-variant case AVPI (Theorem 4.1) outputs a policy $\widehat{\pi}$ such that the suboptimality gap $v^\star - v^{\widehat{\pi}}$ is bounded by $\widetilde{O}(\sqrt{H/nd_m})$ with high probability under the Assumption 2.1.*

The derivation is straightforward by using $\mathrm{Var}_{\pi^\star}(\sum_{h=1}^H r_h) \leq 1$ in (7). This proposition is interesting since it indicates when the MDP is non-stationary, $\widetilde{O}(H/d_m\epsilon^2)$ is required in the worst case even under $\sum_{h=1}^H r_h \leq 1$.[4] The extra H factor resembles the challenge that we have H transitions $(P_1, \ldots, P_H)$ to learn, as opposed to the bandit-type $1/d_m\epsilon^2$ result due to there is only one P throughout (time-invariant). This reveals that one hardness in solving the MDP is in proportion to the number of different transition kernels within the MDP. Such a finding could help researchers understand the special settings like *low switching cost in transitions* [Bai et al., 2019] or *non-stationarity* [Cheung et al., 2020].

4.3 Optimality with Single Concentrability

In the finite horizon discounted setting, Rashidinejad et al. [2021] proposes the single policy concentrability assumption which should be defined as $C^\star := \max_{h,s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} < \infty$ in the current episodic time-variant (non-stationary) MDP setting. As discussed in Appendix C, their lower bound translates to $\Omega(\sqrt{\frac{H^3SC^\star}{n}})$ and their VI-LCB algorithm yields $\widetilde{O}(\sqrt{\frac{H^5SC^\star}{n}})$ suboptimality gap in our case. Since single policy concentrability is strictly weaker than its uniform version (Assumption 2.2), we only discuss this set up. In particular, we have the following implication from our Theorem 4.1 (whose derivation can be found in Appendix G):

Proposition 4.7. *Let $\pi^\star$ be a deterministic policy such that $C^\star := \max_{h,s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} < \infty$. Then by Theorem 4.1, with high probability the output policy of APVI satisfies the suboptimality gap $\widetilde{O}(\sqrt{\frac{H^3SC^\star}{n}})$ in the time-variant (non-stationary) MDPs.*

This can computed similar to (7) except we use $\frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} \leq C^\star$. Our implication is optimal and improves the VI-LCB by the factor H^2 (in terms of sample complexity). Qualitatively, single concentrability is the same as Assumption 2.3, but the use of $C^\star$ makes the bound highly problem independent and limits the adaptivity. Problem dependent bound is a more interesting domain as it tailors to different MDPs specifically. We discuss it now.

4.4 Problem dependent domain

We define the *pre-step environmental norm* (the finite horizon counterpart of Maillard et al. [2014]) as: $\mathbb{Q}_h^\star = \max_{s_h,a_h} \mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)$ for all $h \in [H]$, and relax the total sum of rewards to be bounded by any arbitrary value $\mathcal{B}$ (*i.e.* $\sum_{h=1}^H r_h \leq \mathcal{B}$), then Theorem 4.1 implies:

Proposition 4.8. *Under Assumption 2.1, with high probability, subopmality of AVPI is bounded by*

$$\min\left\{\widetilde{O}\Big(\sum_{h=1}^H \sqrt{\frac{\mathbb{Q}_h^\star}{n\bar{d}_m}}\Big), \widetilde{O}\Big(\sqrt{\frac{H\cdot\mathcal{B}^2}{n\bar{d}_m}}\Big)\right\} + \widetilde{O}\Big(\frac{H^3}{n\bar{d}_m}\Big).$$

[4] Suppose in this case we can achieve $\widetilde{O}(1/d_m\epsilon^2)$ just like Ren et al. [2021], then by a rescaling we obtain the $\widetilde{O}(H^2/d_m\epsilon^2)$ under the usual $0 \leq r_h \leq 1$ assumption which violates the $\Omega(H^3/d_m\epsilon^2)$ lower bound.

Such a result mirrors the online version of the tight problem-dependent bound Zanette and Brunskill [2019] but with a more general *pre-step environmental norm* for the non-stationary MDPs.[5] For the problem instances with either small $\mathcal{B}$ or small $\mathbb{Q}_h^\star$, our result yields much better performances, as discussed in the following.

Deterministic systems. For many practical applications of interest, the systems are equipped with low stochasticity, *e.g.* robotics, or even deterministic dynamics, *e.g.* the game of GO. In those scenarios, the agent needs less experience for each state-action therefore the learning procedure could be much faster. In particular, when the system is fully deterministic (in both transitions and rewards) then $\mathbb{Q}_h^\star = 0$ for all h. This enables a faster convergence rate of order $\frac{H^3}{n\bar{d}_m}$ and significantly improves over the existing non-adaptive results. The convergence rate $\frac{1}{n}$ matches Wen and Van Roy [2013] by translating their constant (in T) regret into the PAC bound.

Partially deterministic systems. Practical worlds are complicated and we could sometimes have a mixture model which contains both deterministic and stochastic steps. In those scenarios, the main complexity is decided by the number of stochastic stages: suppose there are t stochastic P_h, r_h's and $H-t$ deterministic $P_{h'}, r_{h'}$'s, then completing the offline learning guarantees $t \cdot \sqrt{\max Q_h^\star / n\bar{d}_m}$ suboptimality gap, which could be much smaller than $H \cdot \sqrt{\max Q_h^\star / n\bar{d}_m}$ when $t \ll H$.

Fast mixing domains. Consider a class of highly mixing non-stationary MDPs (a variant of Zanette and Brunskill [2018]) that satisfies the transition $P_h(\cdot|s_h, a_h) := \nu_h(\cdot)$ depends on neither the state s_h nor the action a_h. Define $\bar{s}_t := \arg\max V_t^\star(s)$ and $\underline{s}_t := \arg\max V_t^\star(s)$. Also, denote $\mathrm{rng}V_h^\star$ to be the range of $V_h^\star$. In such cases, Bellman optimality equations have the form

$$V_h^\star(\bar{s}_h) = \max_a \left(r_h(\bar{s}_h, a) + \nu_h^\top V_{h+1}^\star\right), \quad V_h^\star(\underline{s}_h) = \max_a \left(r_h(\underline{s}_h, a) + \nu_h^\top V_{h+1}^\star\right),$$

which yields $\mathrm{rng}V_h^\star = V_h^\star(\bar{s}_h) - V_h^\star(\underline{s}_h) = \max_a r_h(\bar{s}_h, a) - \min_a r_h(\underline{s}_h, a) \le 1$, and this in turn gives $\mathbb{Q}_h^\star \le 1 + (\mathrm{rng}V_h^\star)^2 = 2$. As a result, the suboptimality is bounded by $\widetilde{O}(\sqrt{H^2/nd_m})$ in the worst case. This result reveals, although this is a family of stochastic non-stationary MDPs, but it is only as hard as the family of stationary MDPs in the minimax sense ($\Omega(H^2/d_m\epsilon)$).

Tabular contextual bandits. Our result also implies $\widetilde{O}(\sum_{x_1,a_1} d_1^{\pi^\star}(x_1,a_1)\sqrt{\frac{\mathrm{Var}(r_1)}{n \cdot d_1^\mu(x_1,a_1)}})$ gap for the *offline tabular contextual bandit* problem and improves to $\widetilde{O}(1/nd_m)$ when the reward is deterministic. In either cases, the result is optimal and this is due to: when r_1 is deterministic, the agent only needs one sample at every location (see Bubeck and Cesa-Bianchi [2012] for a survey).

5 Towards Assumption-Free Offline RL

While assumption 2.3 is (arguably) the weakest assumption for correctly learning the optimal value, for the real-world problems even this might not be guaranteed. Can we still learn something? Here we consider this new and most general setting where the behavior policy μ can be arbitrary. To tackle this problem, we create a fictitious augmented MDP $M^\dagger$ that can help characterize the discrepancy of the values between the original MDP M and the estimated MDP $\widehat{M}^\dagger$. In particular, $M^\dagger$ is negative towards agnostic state-actions s_h, a_h by setting $r_h^\dagger = 0$ and transitions to an absorbing state $s_{h+1}^\dagger$.

Pessimistic augmented MDP. $M^\dagger$ is defined with one extra state $s_h^\dagger$ for all $h \in \{2, \dots, H\}$ with the augmented state space $\mathcal{S}^\dagger = \mathcal{S} \cup \{s_h^\dagger\}$. The transition and the reward are defined as follows:

$$P_h^\dagger(\cdot \mid s_h, a_h) = \begin{cases} P_h(\cdot \mid s_h, a_h), & n_{s_h,a_h} > 0, \\ \delta_{s_{h+1}^\dagger}, & s_h = s_h^\dagger \text{ or } n_{s_h,a_h} = 0. \end{cases} \quad r^\dagger(s_h, a_h) = \begin{cases} r(s_h, a_h), & n_{s_h,a_h} > 0, \\ 0, & s_h = s_h^\dagger \text{ or } n_{s_h,a_h} = 0. \end{cases}$$

here δ_s is the Dirac measure and we denote $V_h^{\dagger\pi}$ and $v^{\dagger\pi}$ to be the values under $M^\dagger$. $\widehat{M}^\dagger$ is the empirical counterpart of $M^\dagger$ with $\widehat{P}, \widehat{r}$ (the same as (2)) replacing P, r. By Algorithm 1, we have

Theorem 5.1 (Assumption-free offline reinforcement learning)**.** *Let us make no assumption for μ and still denote $\bar{d}_m := \min_{h\in[H]}\{d_h^\mu(s_h,a_h) : d_h^\mu(s_h,a_h) > 0\}$. For any $0 < \delta < 1$, there exists*

[5] Zanette and Brunskill [2019] uses the maximal version by maximizing over h.

absolute constants $c_0, C' > 0$, such that when $n > c_0 \cdot 1/\bar{d}_m \cdot \iota$ ($\iota = \log(HSA/\delta)$), with probability $1-\delta$, the output policy $\widehat{\pi}$ of APVI satisfies (recall $\mathcal{C}_h := \{(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$)

$$v^\star - v^{\widehat{\pi}} \leq \sum_{h=1}^{H} d_h^{\dagger\pi^\star}(s_h^\dagger) + C' \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\dagger\pi^\star}(s_h, a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(r_h^\dagger + V_{h+1}^{\dagger\pi^\star}) \cdot \iota}{n \cdot d_h^\mu(s_h, a_h)}} + \widetilde{O}\left(\frac{H^3}{n\bar{d}_m}\right), \quad (8)$$

where $d_h^{\dagger\pi^\star}(s_h, a_h) \leq d_h^{\pi^\star}(s_h, a_h)$, $V_h^{\dagger\pi^\star}(s_h) \leq V_h^\star(s_h)$ for all $s_h, a_h \in \mathcal{S} \times \mathcal{A}$, and for all $h \in [H]$, $d_h^{\dagger\pi^\star}(s_h^\dagger) = \sum_{t=1}^{h-1} \sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus\mathcal{C}_t} d_t^{\dagger\pi^\star}(s_t, a_t)$. The proof is in Appendix D.

Take-aways of Theorem 5.1. In $M^\dagger$, there is no agnostic location any more since the original unknown spaces now all have *known* deterministic transitions to $s^\dagger$ in $M^\dagger$. At a price, the algorithm has to suffer the constant suboptimality $\sum_{h=1}^{H} d_h^{\dagger\pi^\star}(s_h^\dagger)$ due to no data in the region. $\sum_{h=1}^{H} d_h^{\dagger\pi^\star}(s_h^\dagger)$ helps characterize the hardness when nothing is assumed about μ: it is always less than H (cannot suffer more than H suboptimality); under Assumption 2.1, it is 0 since $M^\dagger = M$ with high probability (by Chernoff bound) and this causes $\mathcal{S} \times \mathcal{A} \setminus \mathcal{C}_h = \emptyset$; under Assumption 2.3, it is also 0 and 5.1 reduces to Theorem 4.1 (see Appendix E).

6 Sketch of the Analysis

Due to the space limit, we could only briefly sketch the key proving ideas in Section 4. We provide an extended proof overview in Appendix A. Our analysis of the intrinsic learning bound in Section 4 leverage the key design feature of APVI that $\widehat{V}_{h+1}$ only depends on the transition data from time $h+1$ to H while $\widehat{P}_h$ only uses transition pairs at time h. This enables concentration inequalities due the *conditional* independence.[6] To cater for the data-adaptive bonus (4), we need to use *Empirical* Bernstein inequality to get $(\widehat{P}_h - P_h)\widehat{V}_{h+1} \lesssim \sqrt{\mathrm{Var}_{\hat{P}}(\widehat{V}_{h+1})/n_{s_h,a_h}}$. Especially, to recover the $\sqrt{\mathrm{Var}_P(V_{h+1}^\star)}$ structure to we uses a self-bounding reduction as follows. First, $\sqrt{\mathrm{Var}_{\hat{P}}(\widehat{V}_{h+1})} - \sqrt{\mathrm{Var}_P(\widehat{V}_{h+1})} \lesssim H/\sqrt{n\bar{d}_m}$ and $\sqrt{\mathrm{Var}_P(\widehat{V}_{h+1})} - \sqrt{\mathrm{Var}_P(V_{h+1}^\star)} \leq ||\widehat{V}_{h+1} - V_{h+1}^\star||_\infty$. Next, we use (3) as the intermediate step to crude bounding $||\widehat{V}_{h+1} - V_{h+1}^\star||_\infty \lesssim H^2/\sqrt{n\bar{d}_m}$ (where "the use of (3)" is the more intricate self-bounding Lemma D.7 in the actual proof) and this yields the desired structure of $\sqrt{\mathrm{Var}_P(V_{h+1}^\star)} + H^2/\sqrt{n\bar{d}_m}$. Lastly, we can combine this with *the extended value difference lemma* in Cai et al. [2020] to bound $V_1^\star - \widehat{V}_1$ and leverage the pessimistic design for bounding $\widehat{V}_1 - V_1^{\widehat{\pi}}$. For the lower bound, we build upon Jin et al. [2020] to create a family of hard instances and mirror the standard *Le Cam method* [Le Cam, 2012] to lower bound the risk of a randomized test function via *Pinkser's inequality*. Our proof shows it is possible recover the $\sqrt{\mathrm{Var}_P(V^\star)}$ term within those hard instances instead of the default value 1 (see Remark 4.3).

7 Conclusion

This work studies the offline reinforcement learning problem and contributes the intrinsic offline learning bound which is a optimal and strong adaptive bound that subsumes most worst-case bounds. The adaptive characterization of the intrinsic bound reveals the fundamental hardness of each individual instances and draws a clearer picture of what offline reinforcement learning looks like. Nevertheless, there are in-depth questions that remain unresolved and we conclude the section by mentioning two of them. First, the ultimate adaptivity for offline learning should possess the *per-instance* optimality. While the intrinsic learning bound **is** a per-instance bound, our lower bound Theorem 4.2 is in the minimax sense (with the hard instances being a specific class). Thus, our result does not achieve the per-instance optimality. A matching *local-minimaxity* guarantee would complete the story. Second, what is the "correct" (tight) information-theoretical limit for the linear MDP case remains elusive. The current results in Jin et al. [2020] deviate by a factor of dH, and it is unclear to us whether a structure like $\sqrt{\mathrm{Var}_P(V^\star)}$ should replace the factor H given the linear MDP structure. Both of them are interesting questions and we leave those as the future works.

[6] This trick is also leveraged in Yin et al. [2021a], but they consider the empirical optimal value $\widehat{V}^{\widehat{\pi}^\star}$ instead.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83, 2020.

Andras Antos, Remi Munos, and Csaba Szepesvari. Fitted q-iteration in continuous action-space mdps. In *Advances in Neural Information Processing Systems*, pages 9–16, 2008a.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71 (1):89–129, 2008b.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

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

Herman Chernoff et al. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pages 1843–1854. PMLR, 2020.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gómez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, et al. Rl unplugged: Benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*, 2020.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398, 2018.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning-Volume 70*, pages 1704–1713, 2017.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020.

Tobias Jung and Peter Stone. Gaussian processes for sample efficient reinforcement learning with rmax-like exploration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 601–616. Springer, 2010.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712, 2019.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. In *Uncertainty in Artificial Intelligence*, 2019.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor. How hard is my mdp?" the distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, 27:1835–1843, 2014.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *arXiv preprint arXiv:2103.14077*, 2021.

Karthik Sridharan. A gentle introduction to concentration inequalities. *Dept. Comput. Sci., Cornell Univ., Tech. Rep*, 2002.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pages 880–887, 2005.

Joel Tropp et al. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Zheng Wen and Benjamin Van Roy. Efficient exploration and value function generalization in deterministic systems. *Advances in Neural Information Processing Systems*, 2013.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020a.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Uncertainty in Artificial Intelligence*, pages 550–559, 2020b.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021a.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021b.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. *arXiv preprint arXiv:2012.08005*, 2020.

Andrea Zanette and Emma Brunskill. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *International Conference on Machine Learning*, pages 5747–5755. PMLR, 2018.

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

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , see Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not have any negative social impact since this is a theoretical work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have read them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The complete proofs can be found in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]