Near-Optimal Offline Reinforcement Learning via Double Variance Reduction

Ming Yin [1 2] **Yu Bai** [3] **Yu-Xiang Wang** [1]

Abstract

We consider the problem of offline reinforcement learning (RL) — a well-motivated setting of RL that aims at policy optimization using only historical data. In this paper, we propose *Off-Policy Double Variance Reduction* (OPDVR), a new variance reduction based algorithm for offline RL. Our main result shows that OPDVR provably identifies an ϵ-optimal policy with $\widetilde{O}(H^2/d_m\epsilon^2)$ episodes of offline data in the finite-horizon *stationary transition* setting and this improves over the best known upper bound by a factor of H. Moreover, we establish an information-theoretic lower bound of $\Omega(H^2/d_m\epsilon^2)$ which certifies that OPDVR is optimal up to logarithmic factors. Lastly, we show that OPDVR also achieves rate-optimal sample complexity under alternative settings such as the finite-horizon MDPs with non-stationary transitions and the infinite horizon MDPs with discounted rewards.

1. Introduction

Offline reinforcement learning (offline RL) aims at learning the near-optimal policy by using a static offline dataset that is collected by a certain behavior policy μ (Lange et al., 2012). Despite its practical significance, a precise theoretical understanding of offline RL has been lacking. Previous sample complexity bounds for RL has primarily focused on the online setting (Azar et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019; Efroni et al., 2019; Cai et al., 2019) or the generative model (simulator) setting (Azar et al., 2013; Sidford et al., 2018a;b; Yang & Wang, 2019; Agarwal et al., 2020; Wainwright, 2019; Lattimore & Szepesvari, 2019), both of which assuming interactive access to the environment and not applicable to offline RL. On the other hand, the sample complexity of offline RL remains unsettled even for environments with finitely many state and actions.

Our Contributions In this paper, we propose an algorithm ***OPDVR*** (Off-Policy Doubled Variance Reduction) for offline reinforcement learning based on an extension of the variance reduction technique initiated in (Sidford et al., 2018a; Yang & Wang, 2019). ***OPDVR*** performs stochastic (minibatch style) value iterations using the available offline data, and can be seen as a version of stochastic optimal planning that interpolates value iteration and Q-learning.

- We show that ***OPDVR*** finds an ϵ-optimal policy with high probability using $\widetilde{O}(H^2/d_m\epsilon^2)$ episodes of offline data (Section 4.1). This improves upon the best known sample complexity by an H factor and to the best of our knowledge is the first that achieves an $O(H^2)$ horizon dependence.
- We establish a sample (episode) complexity lower bound $\Omega(H^2/d_m\epsilon^2)$ for offline RL in the finite-horizon stationary setting (Theorem 4.2), showing that the sample complexity of ***OPDVR*** is optimal up to logarithmic factors.
- In the finite-horizon non-stationary setting, and infinite horizon γ-discounted setting, we show that ***OPDVR*** achieves $\widetilde{O}(H^3/d_m\epsilon^2)$ sample (episode) complexity (Section 3) and $\widetilde{O}((1-\gamma)^{-3}/d_m\epsilon^2)$ sample complexity (Section 4.2) respectively. They are both optimal up to logarithmic factors.
- On the technical end, our algorithm presents a sharp analysis of offline RL with stationary transitions, and, *importantly*, the use of the doubling technique to resolve the initialization dependence defect which fails to make the original variance reduction algorithm of (Sidford et al., 2018a) to be optimal, see Appendix F.4. Running (Sidford et al., 2018a) may not yield the desired accuracy as they stated and our result is robust in preserving the optimality.

2. Preliminaries

We consider reinforcement learning problems modeled by finite Markov Decision Processes (MDPs) (we focus on

[1]Department of Computer Science, University of California, Santa Barbara, USA [2]Department of Statistics and Applied Probability, University of California, Santa Barbara, USA [3]Salesforce Research, CA, USA. Correspondence to: Ming Yin <ming_yin@ucsb.edu>.

Reinforcement Learning Theory Workshop at *International Conference on Machine Learning*, 2021. The full version of the paper can be found in https://arxiv.org/abs/2102.01748

Table 1. Comparison of sample complexities for tabular offline RL interpretation.

Method/Analysis	Setting	Assumptions	Sample complexity[a]
FQI (Chen & Jiang, 2019)	∞-horizon	Full Concentrability	$\widetilde{O}((1-\gamma)^{-6}C/\epsilon^2)$
MSBO/MABO (Xie & Jiang, 2020b)	∞-horizon	Full Concentrability	$\widetilde{O}((1-\gamma)^{-4}C_\mu/\epsilon^2)$
OPEMA (Yin et al., 2021)	H-horizon non-stationary	Full Concentrability	$\widetilde{O}(H^3/d_m\epsilon^2)$
OPDVR (Section 3)	H-horizon non-stationary	Weak Coverage	$\widetilde{O}(H^3/d_m\epsilon^2)$
OPDVR (Section 4)	H-horizon stationary	Weak Coverage	$\widetilde{O}(H^2/d_m\epsilon^2)$
OPDVR (Section 4.2)	∞-horizon	Weak Coverage	$\widetilde{O}((1-\gamma)^{-3}/d_m\epsilon^2)$

[a] Number of episodes in the finite horizon setting and number of steps in the infinite horizon.
[b] $\beta_\mu, C, C_\mu, 1/d_m$ are the concentrability-type coefficients that measure the state-action coverage. See Assumption 2.1 and also Section F.2 for discussions.

the finite-horizon episodic setting, and defer the infinite-horizon discounted setting to Section 4.2.) An MDP is denoted by a tuple $M = (\mathcal{S}, \mathcal{A}, r, T, d_1, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces with finite cardinality $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. $P_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition kernel with $P_t(s'|s,a)$ be the probability of entering state s' after taking action a at state s. We consider both the stationary and non-stationary transition setting: The stationary transition setting assumes $P_t \equiv P$ is identical for all $t \in [H]$, and the non-stationary transition setting allows P_t to be different for different t. $r_t : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function which we assume to be deterministic[1]. d_1 is the initial state distribution, and H is the time horizon. A (non-stationary) policy $\pi : \mathcal{S} \to \mathbb{P}_{\mathcal{A}}^H$ assigns to each state $s_t \in \mathcal{S}$ a distribution over actions at each time t. We use $d_t^\pi(s,a)$ or $d_t^\pi(s)$ to denote the marginal state-action/state distribution induced by policy π at time t, i.e. $d_t^\pi(s) := \mathbb{P}^\pi(s_t = s)$ and $d_t^\pi(s,a) := \mathbb{P}^\pi(s_t = s, a_t = a)$.

Offline learning problem. In this paper we investigate the offline learning problem, where we do not have interactive access to the MDP, and can only observe a static dataset $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)})\right\}_{i\in[n]}^{t\in[H]}$.We assume that $\mathcal{D}$ is obtained by executing a pre-specified *behavior policy* μ (also known as the *logging policy*) for n episodes and collecting the trajectories $\tau^{(i)} = (s_1^{(i)}, a_1^{(i)}, r_1^{(i)}, \ldots, s_H^{(i)}, a_H^{(i)}, r_H^{(i)}, s_{H+1}^{(i)})$, where each episode is rendered in the form: $s_1^{(i)} \sim d_1$, $a_t^{(i)} \sim \mu_t(\cdot|s_t^{(i)})$, $r_t^{(i)} = r(s_t^{(i)}, a_t^{(i)})$, and $s_{t+1}^{(i)} \sim P_t(\cdot|s_t^{(i)}, a_t^{(i)})$. Given the dataset $\mathcal{D}$, our goal is to find an ϵ-**optimal policy** π_{out}, in the sense that $||V_1^{\pi^\star} - V_1^{\pi_{\text{out}}}||_\infty < \epsilon$.

Assumption 2.1 (Weak coverage)**.** *The behavior policy μ satisfies the following: There exists* some *optimal policy $\pi^\star$ such that $d_{t'}^\mu(s_{t'}, a_{t'}) > 0$ if there exists $t < t'$ such that $d_{t:t'}^{\pi^\star}(s_{t'}, a_{t'}|s_t, a_t) > 0$, where $d_{t:t'}^{\pi^\star}(s_{t'}, a_{t'}|s_t, a_t)$ is the conditional multi-step transition probability from step t to t'. Furthermore, we define $d_m := \min_{t,s_t,a_t}\{d_t^\mu(s_t,a_t) : d_t^\mu(s_t,a_t) > 0\}$.*

Intuitively, Assumption 2.1 requires μ to "cover" certain optimal policy $\pi^\star$, in the sense that any $s_{t'}, a_{t'}$ is reachable by μ if it is attainable from a previous state-action pair by $\pi^\star$. It is similar to (Liu et al., 2019, Assumption 1). Note that this is weaker than the standard "concentrability" assumption (Munos, 2003; Le et al., 2019; Chen & Jiang, 2019): Concentrability defines $\beta_\mu := \sup_{\pi\in\Pi} ||d^\pi(s_t,a_t)/d^\mu(s_t,a_t)||_\infty < \infty$ (cf. (Le et al., 2019, Assumption 1 & Example 4.1)), which requires the sufficient exploration for tabular case[2] since we optimize over all policies (see Section F.2 for a discussion). In contrast, our assumption only requires μ to "trace" one single optimal policy.

3. Variance reduction for offline RL

3.1. Review: variance reduction for RL

In the case of policy optimization, VR is an algorithm that approximately iterating the Bellman optimality equation, using an inner loop that performs an approximate value (or Q-value) iteration using fresh interactive data to estimate $V^\star$, and an outer loop that performs multiple steps of such iterations to refine the estimates. Concretely, to obtain an reliable $Q_t(s,a)$ for some step $t \in [H]$, by the Bellman equation $Q_t(s,a) = r(s,a) + P_t^\top(\cdot|s,a)V_{t+1}$, we need to estimate $P_t^\top(\cdot|s,a)V_{t+1}$ with sufficient accuracy. VR handles this by decomposing:

$$P_t^\top(\cdot|s,a)V_{t+1} = P_t^\top(\cdot|s,a)(V_{t+1} - V_{t+1}^{\text{in}}) + P_t^\top(\cdot|s,a)V_{t+1}^{\text{in}}, \quad (1)$$

where V_{t+1}^{in} is a *reference* value function obtained from previous calculation (See line 4,13 in the inner loop of Algorithm 1) and $P_t^\top(\cdot|s,a)(V_{t+1} - V_{t+1}^{\text{in}})$, $P_t^\top(\cdot|s,a)V_{t+1}^{\text{in}}$ are estimated separately at different stages. This technique can help in reducing the "effective variance" along the learning process (see Wainwright (2019) Section 2 for a discussion).

[1] This is commonly assumed in the RL literature. The randomness in the reward will only cause a lower order error (than the randomness in the transition) for learning.

[2] Note Xie & Jiang (2020b) has a tighter concentration coefficient with $C_\mu := \max_{\pi\in\Pi} \left\|w_{d_\pi/\mu}\right\|_{2,\mu}^2$ but it still requires full exploration when Π contains all policies.

In addition, in order to translate the guarantees from learning values to learning policies, we build on the following "monotonicity property": For any policy π that satisfies the monotonicity condition $V_t \le \mathcal{T}_{\pi_t} V_{t+1}$ for all $t \in [H]$, the performance of π is sandwiched as $V_t \le V_t^{\pi} \le V_t^{\star}$, i.e. π is guaranteed to perform the same or better than V_t. This property is first captured by (Sidford et al., 2018a) (for completeness we provide a proof in Lemma B.1), and later reused by Yang & Wang (2019); Sidford et al. (2020) under different settings. We rely on this property in our offline setting as well for providing policy optimization guarantees.

3.2. OPDVR: variance reduction for offline RL

We begin with non-stationary setting for the ease of explaining algorithmic design. We let $\iota := \log(HSA/\delta)$.

Prototypical offline VR. We first describe a prototypical version of our offline VR algorithm in Algorithm 1, which we will instantiate with different parameters *twice* (hence the name"Double") in each of the three settings of interest.

Algorithm 1 takes estimators $\mathbf{z}_t$ and $\mathbf{g}_t$ that produce lower confidence bounds (LCB) of the two terms in (1) using offline data. Specifically, we assume $\mathbf{z}_t, \mathbf{g}_t$ are both available in *function forms* in that they take an offline dataset (with an arbitrary size), fixed value function $V_{t+1}, V_{t+1}^{\text{in}}$ and an external scalar input u then return $z_t, g_t \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. z_t, g_t satisfies that

$$z_t(s_t,a_t) \le P^{\top}(\cdot|s_t,a_t)V_{t+1}^{\text{in}}, \quad g_t(s_t,a_t) \le P^{\top}(\cdot|s_t,a_t)[V_{t+1}-V_{t+1}^{\text{in}}],$$

uniformly for all s_t, a_t with high probability.

Algorithm 1 then proceeds by taking the input offline dataset as a stream of iid sampled trajectories and use an exponentially increasing-sized batches of independent data to pass in $\mathbf{z}_t$ and $\mathbf{g}_t$ while updating the estimated Q value function by applying the Bellman backup operator except that the update is based on a conservative and variance reduced estimated values. Each inner loop iteration backs up from the last time-step and update all Q_t for $t = H, ..., 1$; and each outer loop iteration passes a new batch of data into the inner loop while ensuring reducing the suboptimality gap from the optimal policy by a factor of 2 in each outer loop iteration, provided that the estimators $\mathbf{z}_t + \mathbf{g}_t$ are increasingly more accurate estimates of (1) as the suboptimality gap gets smaller.

Plug-in estimators and high-confidence LCBs. The estimators $\mathbf{z}_t$ and $\mathbf{g}_t$ we use for the three different settings are provided in Figure A. They are essentially the natural plug-in estimators of $P_t^{\top}(\cdot|s,a)(V_{t+1} - V_{t+1}^{\text{in}})$ and $P_t^{\top}(\cdot|s,a)V_{t+1}^{\text{in}}$ as well as their standard deviation by replacing P_t with $\widehat{P}_t$ except that we use two disjoint splits $\mathcal{D}_1$ and $\mathcal{D}_2$ for $\mathbf{z}_t$ and $\mathbf{g}_t$ so they remain statistically independent. A key difference from the generative model setting is that these estimators are dependent across t, thus it requires new technical steps to establish the convergence of these estimators as well as putting them together to show that Algorithm 1 works.

The doubling procedure. It turns out that Algorithm 1 alone does not yield a tight sample complexity guarantee, due to its *suboptimal dependence on the initial optimality gap* $u^{(0)} \ge \sup_t \|V_t^{\star} - V_t^{(0)}\|_{\infty}$ (recall $u^{(0)}$ is the initial parameter in the outer loop of Algorithm 1). This is captured in the following (for the non-stationary case):

Proposition 3.1 (Informal version of Lemma B.10)**.** *Suppose $\epsilon \in (0,1]$ is the final target accuracy. Algorithm 1 outputs the ϵ-optimal policy with episode complexity:*

$$\bullet\ \tilde{O}(H^4/d_m\epsilon^2), \quad \text{If } u^{(0)} > \sqrt{H}; \quad \bullet\ \tilde{O}(H^3/d_m\epsilon^2), \quad \text{If } u^{(0)} \le \sqrt{H}.$$

Proposition 3.1 suggests that Algorithm 1 may have a suboptimal sample complexity when the initial optimality gap $u^{(0)} > \sqrt{H}$. Unfortunately, this is precisely the case for standard initializations such as $V_t^{(0)} := \mathbf{0}$, for which we must take $u^{(0)} = H$. We overcome this issue by designing a two-stage *doubling* procedure: At stage 1, we use Algorithm 1 to obtain $V_t^{\text{intermediate}}, \pi^{\text{intermediate}}$ that are $\epsilon' = \sqrt{H}\epsilon$ accurate; At stage 2, we then use Algorithm 1 again with $V_t^{\text{intermediate}}, \pi^{\text{intermediate}}$ as the input and further reduce the error from ϵ' to ϵ. The main take-away of this doubling procedure is that the episode complexity of both stage is only $\tilde{O}(H^3/d_m\epsilon^2)$, therefore the total sample complexity optimality is preserved. The pseudo-code of the two-stage procedure ***OPDVR*** is summarized in Algorithm 2.

3.3. *OPDVR* for non-stationary transition settings

Theorem 3.2 (***OPDVR*** in episodic non-stationary setting)**.** *For the H-horizon non-stationary setting, there exist universal constants $c_1, c_2, c_3 > 0$ such that if we set $m_1' = c_1H^4/d_m$ for Stage 1, $m_2' = c_2H^3/d_m$ for Stage 2, set $K_1 = K_2 = \log_2(\sqrt{H}/\epsilon)$, take $\mathbf{g}_t$ and $\mathbf{z}_t$ according to Figure A, then* **OPDVR** *(Algorithm 2) with probability $1-\delta$ outputs an ϵ-optimal policy $\hat{\pi}$ provided that the number of episodes in the offline data $\mathcal{D}$ exceeds (below can be readily simplified as $\widetilde{O}\left(H^3/d_m\epsilon^2\right)$):*

$$\frac{c_3 \max[\frac{m_1'}{H}, m_2']}{\epsilon^2}\left(\iota + \log\log_2(\frac{\sqrt{H}}{\epsilon})\right)\log_2(\frac{\sqrt{H}}{\epsilon}),$$

Optimality of sample complexity. Theorem 3.2 shows that our ***OPDVR*** algorithm can find an ϵ-optimal policy with $\widetilde{O}(H^3/d_m\epsilon^2)$ episodes of offline data. Compared with the sample complexity lower bound $\Omega(H^3/d_m\epsilon^2)$ for offline learning (Theorem G.2. in Yin et al. (2021)), we see that our ***OPDVR*** algorithm matches the lower bound up to logarithmic factors. The same rate was achieved previously by the local uniform convergence argument of Yin et al. (2021) under a stronger assumption of full data coverage.

4. *OPDVR* for stationary transition settings

In this section, we switch gears to the *stationary* transition setting, in which the transition probabilities are identical at all time steps: $P_t(s'|s,a) :\equiv P(s'|s,a)$. We will consider both the (a) finite-horizon case where each episode is consist of H steps; and (b) the infinite-horizon case where the reward at the t-th step is discounted by γ^t, where $\gamma \in (0,1)$ is a discount factor. These settings encompass additional challenges compared with the non-stationary case, as in theory the transition probabilities can now be estimated more accurately due to the shared information across time.

4.1. Finite-horizon stationary setting

Theorem 4.1 (Sample complexity of ***OPDVR*** in finite-horizon stationary setting)**.** *In the H-horizon stationary transition setting, there exists universal constants c_1', c_2', c_3' such that if we set $m_1' = c_1' H^3/d_m$, $m_2' = c_2' H^2/d_m$ for Stage 1 and 2, set $K_1 = K_2 = \log_2(\sqrt{H}/\epsilon)$, and take $\mathbf{z}_t$ and $\mathbf{g}_t$ according to Figure A, then with probability $1-\delta$, Practical* **OPDVR** *finds an ϵ-optimal policy provided that the number of episodes in the offline data $\mathcal{D}$ exceeds (which is of order $\widetilde{O}\left(\frac{H^2}{d_m\epsilon^2}\right)$):*

$$\frac{c_3' \max[\frac{m_1'}{H}, m_2']}{\epsilon^2}\left(\iota + \log\log_2(\frac{\sqrt{H}}{\epsilon})\right)\log_2(\frac{\sqrt{H}}{\epsilon}),$$

Theorem 4.1 encompasses our main technical contribution, as the compact data aggregation among different time steps make analyzing the estimators (60) and (61) knotty due to data-dependence (unlike the non-stationary transition setting where estimators are designed using data at specific time). In particular, we need to fully exploit the property that transition P is identical across different times in a pinpoint way to obtain the H^2 dependence in the sample complexity bound.

Improved dependence on H. Theorem 4.1 shows that ***OPDVR*** achieves a sample complexity upper bound $\widetilde{O}(H^2/d_m\epsilon^2)$ in the stationary setting. To the best of our knowledge, this is the first result that achieves an H^2 dependence for offline RL with stationary transitions.

Optimality of $\widetilde{O}(H^2/d_m\epsilon^2)$. We accompany Theorem 4.1 by a establishing a sample complexity lower bound for this setting, showing that our algorithm achieves the optimal dependence of all parameters up to logarithmic factors.

Theorem 4.2 (Information-theoretic lower bound)**.** *For all $0 < d_m \le \frac{1}{SA}$, let the family of problem be $\mathcal{M}_{d_m} := \left\{(\mu, M) \;\middle|\; \min_{t,s_t,a_t} d_t^\mu(s_t,a_t) \ge d_m\right\}$. There exists universal constants c_1, c_2, c, p (with $H, S, A \ge c_1$ and $0 < \epsilon < c_2$) such that when $n \le cH^2/d_m\epsilon^2$, $\inf_{v^{\pi_{alg}}} \sup_{(\mu,M)\in\mathcal{M}_{d_m}} \mathbb{P}_{\mu,M}\left(v^* - v^{\pi_{alg}} \ge \epsilon\right) \ge p$.*

4.2. Infinite-horizon discounted setting

Finally, we consider the infinite-horizon discounted setting. The setting is slightly different to the finite horizon case as we adopt the same assumption of (Chen & Jiang, 2019; Xie & Jiang, 2020b) that data $\mathcal{D} = \{s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)}\}_{i\in[n]}$ are i.i.d off-policy pieces with $(s,a) \sim d^\mu$ and $s' \sim P(\cdot|s,a)$. The infinite horizon-versions of ***OPDVR*** (Algorithm 3 and 4) are stated in the Appendix due to the space limit.

Theorem 4.3 (Sampe complexity of ***OPDVR*** in infinite-horizon discounted setting)**.** *Consider Algorithm 4. There are constants c_1', c_2', c_3', such that if we set $m_1' = O((1-\gamma)^{-4}/d_m), m_2' = O((1-\gamma)^{-3}/d_m)$ (see more precise expressions in Lemma D.7), $K_1 = \log_2((1-\gamma)^{-1}/\epsilon), K_2 = \log_2(\sqrt{(1-\gamma)^{-1}}/\epsilon)$, $R = \log(4/\epsilon(1-\gamma))$, and choose LCB estimators $\mathbf{z}$ and $\mathbf{g}$ as in Figure A, then with probability $1-\delta$, the infinite horizon version of **OPDVR** (Algorithm 4) outputs an ϵ-optimal policy provided that in offline data $\mathcal{D}$ has number of samples exceeding*

$$\frac{c_3' \max[\frac{m_1'}{(1-\gamma)^{-1}}, m_2']}{\epsilon^2}\cdot\iota' = \widetilde{O}\left[(1-\gamma)^{-3}/d_m\epsilon^2\right].$$

where $\iota' := R\cdot(\log(32(1-\gamma)^{-1}RSA/\delta) + \log\log_2(\sqrt{(1-\gamma)^{-1}}/\epsilon))\cdot\log_2(\sqrt{(1-\gamma)^{-1}}/\epsilon)$.

We note that for the infinite horizon case, the sample-complexity measures the number of steps, thus $(1-\gamma)^{-3}$ is comparable to the H^2 dependence. To the best of our knowledge, Theorem 4.1 and Theorem 4.3 are the first results that achieve H^2, $(1-\gamma)^{-3}$ dependence in the offline regime respectively for stationary transition and infinite horizon setting, see Table 1.

5. Discussions

Estimating d_m. It is worth mentioning that the input of ***OPDVR*** depends on unknown system quantity d_m. Nevertheless, d_m is only one-dimensional scalar and thus it is plausible (from a statistical perspective) to leverage standard parameter-tuning tools (*e.g.* cross validation (Varma & Simon, 2006)) for obtaining a reliable estimate in practice. On the theoretical side, we provide the following result to show plug-in on-policy estimator $\widehat{d}_t^\mu(s_t,a_t) = n_{s_t,a_t}/n$ and $\widehat{d}_m := \min_{t,s_t,a_t}\{n_{s_t,a_t}/n : n_{s_t,a_t} > 0\}$, is sufficient for accurately estimating d_t^μ, d_m simultaneously.

Lemma 5.1. *For the finite-horizon setting (either stationary or non-stationary), there exists universal constant c, s.t. when $n \ge c\cdot 1/d_m\cdot\log(HSA/\delta)$, then w.p. $1-\delta$, we have $\forall t, s_t, a_t$, $\frac{1}{2}d_t^\mu(s_t,a_t) \le \widehat{d}_t^\mu(s_t,a_t) \le \frac{3}{2}d_t^\mu(s_t,a_t)$ and, in particular, $\frac{1}{2}d_m \le \widehat{d}_m \le \frac{3}{2}d_m$. See Appendix F.1 for proof.*

Improvement over VR in the generative model setting. First, the data collected in the offline case are highly dependent (in contrast in the generative model setting each simulator call is independent), and disentangling the dependent structure makes the offline setting inherently more challenging. Second, our doubling mechanism always guarantee the minimax rate with any initialization and the single

VR procedure does not have this property (see Appendix F.4 for a more detailed discussion), which could be a critical issue when (Sidford et al., 2018a) claims the optimality.

References

Agarwal, A., Jiang, N., and Kakade, S. M. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Antos, A., Szepesvári, C., and Munos, R. Fitted q-iteration in continuous action-space mdps. In *Advances in neural information processing systems*, pp. 9–16, 2008a.

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

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

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

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

Chernoff, H. 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.

Chung, F. and Lu, L. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1): 79–127, 2006.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2818–2826, 2015.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. *arXiv preprint arXiv:1811.03056*, 2018.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. *arXiv preprint arXiv:2010.03531*, 2020.

Duan, Y. and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*, 2020.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pp. 12203–12213, 2019.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111. PMLR, 2020.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 652–661. JMLR. org, 2016.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1704–1713. JMLR. org, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

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

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*, 2019a.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019b.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pp. 1840–1848, 2016.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Lattimore, T. and Szepesvari, C. Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

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

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. 2015.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. In *Uncertainty in Artificial Intelligence*, 2019.

Liu, Y., Bacon, P.-L., and Brunskill, E. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *International Conference on Machine Learning*, pp. 6184–6193. PMLR, 2020a.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020b.

Munos, R. Error bounds for approximate policy iteration. In *ICML*, volume 3, pp. 560–567, 2003.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018a.

Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 770–787. SIAM, 2018b.

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002. PMLR, 2020.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, pp. 1151–1160, 2019.

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

Uehara, M. and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Varma, S. and Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.

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

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

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

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. *arXiv preprint arXiv:2003.03924*, 2020b.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, 2019.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004, 2019.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS-20*, 2020.

Yin, M., Bai, Y., and Wang, Y.-X. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. In *AISTATS-21*, 2021.

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

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020.