Minimax-Optimal Off-Policy Evaluation with Linear Function Approximation

Yaqi Duan [1] **Zeyu Jia** [2] **Mengdi Wang** [3 4]

Abstract

This paper studies the statistical theory of off-policy policy evaluation with function approximation in batch data reinforcement learning problem. We consider a regression-based fitted Q iteration method, and show that it is equivalent to a model-based method that estimates a conditional mean embedding of the transition operator. We prove that this method is information-theoretically optimal and has nearly minimal estimation error. In particular, by leveraging contraction property of Markov processes and martingale concentration, we establish a finite-sample instance-dependent error upper bound and a nearly-matching minimax lower bound. The policy evaluation error depends sharply on a restricted χ^2-divergence over the function class between the long-term distribution of target policy and the distribution of past data. This restricted χ^2-divergence characterizes the statistical limit of off-policy evaluation, and is both instance-dependent and function-class-dependent. Further, we provide an easily computable confidence bound for the policy evaluator, which may be useful for optimistic planning and safe policy improvement.

1. Introduction

Batch data reinforcement learning (RL) is common in decision-making applications where rich experiences are available but new experiments are costly. A first-order question is how much one can learn from existing experiences to predict and improve the performance of new policies. This is known as the off-policy policy evaluation (OPE) problem, where one needs to estimate the cumulative rewards (aka value) to be earned by a new policy based on logged history.

In this paper, we study the off-policy evaluation using linear function approximation. We assume that the Q-functions of interests belong to a known function class $\mathcal{Q}$ with d basis functions. We adopt a direct regression-based approach and investigate the basic fitted Q iteration (FQI) (Bertsekas et al., 1995; Sutton & Barto, 2018). It works by iteratively estimating Q-functions via supervised learning using the batch data. This approach turns out to be equivalent to the model-based plug-in estimator where one estimates the conditional mean embedding of the unknown transition model and uses it to compute a plug-in value estimator. It is also related to variants of importance sampling methods (see discussions in Sections 1.1 and 3.3).

We provide a finite-sample error upper bound for this policy evaluator, as well as a nearly matching minimax-optimal lower bound. Putting them together, we see that the regression-based policy evaluator is nearly statistically optimal. For RL with horizon H, the minimax-optimal OPE error takes the form

$$|\widehat{v}^\pi - v^\pi| \asymp H^2 \sqrt{\frac{1+\chi^2_{\mathcal{Q}}(\mu^\pi, \bar{\mu})}{N}} + o(1/\sqrt{N}),$$

where μ^π is some long-term state-action occupancy measure of the target policy π and $\overline{\mu}$ is the data distribution, $\chi^2_{\mathcal{Q}}$ is a variant of χ^2-divergence *restricted to the family* $\mathcal{Q}$:

$$\chi^2_{\mathcal{Q}}(p_1, p_2) := \sup_{f \in \mathcal{Q}} \frac{\mathbb{E}_{p_1}[f(x)]^2}{\mathbb{E}_{p_2}[f(x)^2]} - 1.$$

The term $\chi^2_{\mathcal{Q}}(\mu^\pi, \overline{\mu})$ captures the distributional mismatch, between the behavior policy and the target policy, that is relevant to the function class $\mathcal{Q}$. It determines the theoretical limits of OPE within this function class. In the tabular case, it relates to the worst-case density ratio, which often shows up in importance sampling methods. However, when we use function approximation, this $\chi^2_{\mathcal{Q}}$ divergence term can be significantly smaller than the worst-case density ratio. In particular, our analysis shows that $\chi^2_{\mathcal{Q}}(\mu^\pi, \bar{\mu})$ is the condition number of a finite matrix, which can be reliably estimated. This result suggests that OPE could be more data-efficient with appropriate function approximation.

A summary of technical results of this paper:

*Equal contribution [1]Department of Operations Research and Financial Engineering, Princeton University, NJ, USA [2]School of Mathematics, Peking University, Beijing, China [3]Department of Operations Research and Financial Engineering, Princeton University, NJ, USA [4]DeepMind, London, UK. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

- A regression-based algorithm that unifies FQI and plug-in estimation. It does not require knowledge of the behavior policy $\overline{\pi}$, or try to estimate it. It uses iterative regression but does not require Monte Carlo sampling. In the case of linear models, the estimator can be computed easily using simple matrix-vector operations.
- Finite-sample error upper bound for the regression-based policy evaluator. Despite that regression may be biased for OPE, we show that the curse of horizon does not occur as long as $N = \Omega(dH^3)$. A key to the analysis is the use of contraction properties of a Markov process to show that estimation error accumulates linearly in multi-step policy evaluation, instead of exponentially.
- A minimax error lower bound that sets the statistical limit for OPE with function approximation. The lower bound nearly matches our upper bound, therefore proves the efficiency of regression-based FQI.
- A data-dependent confidence bound that can be computed as a byproduct of the FQI algorithm.

1.1. Related Literature

Off-policy policy evaluation (OPE) is often the starting point of batch reinforcement learning. A direct approach is to estimate the transition probability distributions and then execute the target policy on an estimated model. This has been studied in the tabular case with bias and variance analysis (Mannor et al., 2004). In real-world applications, in order to tackle MDPs with infinite or continuous state spaces, one often needs various forms of function approximation, and many methods like fitted Q-iteration and least square policy iteration were developed (Jong & Stone, 2007; Lagoudakis & Parr, 2003; Grunewalder et al., 2012; Fonteneau et al., 2013). Regression methods are often used to fit value functions and to satisfy the Bellman equation (Bertsekas et al., 1995; Sutton & Barto, 2018; Yang & Wang, 2019b).

A popular class of OPE methods use importance sampling (IS) to reweigh sample rewards to get unbiased value estimate of a new policy (Precup, 2000). Doubly robust technique blends IS with model-based estimators to reduce the high variance (Jiang & Li, 2016; Thomas & Brunskill, 2016). Liu et al. (2018) suggested that one should estimate the stationary state occupancy measure instead of the cumulative importance ratio in order to break the curse of horizon. Many IS methods only apply to tabular MDP and require knowledge of the behavior policy. Following these ideas, Nachum et al. (2019) proposed a minimax optimization problem that uses function approximation to learn the IS weights, without requiring knowledge of the bahavior policy. Dann et al. (2019) provided error bounds and certificates for the tabular case to achieve accountability. Liu et al. (2019) studied off-policy gradient method for batch data policy optimization.

On the theoretical side, the sharpest OPE error bound to our best knowledge is given by Xie et al. (2019) and Yin & Wang (2020), which applies to time-inhomogeneous, tabular MDP. Jiang & Li (2016) provided a Cramer-Rao lower bound for discrete-tree MDP. To the authors' best knowledge, most existing theoretical results on OPE apply only to tabular MDP without function approximation. Le et al. (2019) studied batch RL and FQI for policy learning, evaluation and provides generalization bounds that depend on the VC dimension of the function class. Their results require a "concentration coefficient" assumption that the elementwise ratio between generating and target density functions are uniformly bounded across all states, actions and policies. In comparison, our results do not require such concentration condition, and appear to be the first and sharpest error bounds for OPE with linear function approximation.

2. Problem and Model

In this paper, we study off-policy policy evaluation of an Markov decision process (MDP) when we only have a fixed dataset of empirical transitions. An instance of MDP is a controlled random walk over a state space $\mathcal{S}$, where at each state s, if we pick action $a \in \mathcal{A}$, the system evolves to a random next state s' according to distribution $p(s' \mid s, a)$ and generates a reward $r' \in [0, 1]$ with $\mathbb{E}[r' \mid s, a] = r(s, a)$. A policy π specifies a distribution $\pi(\cdot \mid s)$ for choosing actions conditioned on the current state s.

Our objective is to evaluate the performance of a *target* policy π at a fixed initial distribution ξ_0, where the transition model p is unknown. The value to be estimated is the expected cumulative reward in an H-horizon episode, given by

$$v^\pi := \mathbb{E}^\pi \left[\sum_{h=0}^{H} r(s_h, a_h) \,\middle|\, s_0 \sim \xi_0 \right], \tag{1}$$

where $a_h \sim \pi(\cdot \mid s_h)$, $s_{h+1} \sim p(\cdot \mid s_h, a_h)$, $\mathbb{E}^\pi$ denotes expectation over the sample path generated under policy π.

Let $\mathcal{D} = \{(s_n, a_n, s'_n, r'_n)\}_{n=1}^N$ be a set of sample transitions, where each s'_n is sampled from distribution $p(\cdot \mid s_n, a_n)$. The sample transitions may be collected from multiple trajectories and under a possibly unknown *behavior* policy denoted as $\overline{\pi}$. Our goal is to estimate v^π from $\mathcal{D}$.

Given a *target* policy π and a reward function r, the state-action value functions, also known as Q functions, are defined as, for $h = 0, 1, \ldots, H$,

$$Q_h^\pi(s, a) := \mathbb{E}^\pi \left[\sum_{h'=h}^{H} r(s_{h'}, a_{h'}) \,\middle|\, s_h = s, a_h = a \right], \tag{2}$$

where $a_{h'} \sim \pi(\cdot \mid s_{h'}), s_{h'+1} \sim p(\cdot \mid s_{h'}, a_{h'})$. Let $\mathcal{X} := \mathcal{S} \times \mathcal{A}$. Define the *conditional transition operator* $\mathcal{P}^\pi : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ as

$$\mathcal{P}^\pi f(s,a) := \mathbb{E}^\pi\big[f(s',a') \,\big|\, s,a\big] \quad \text{for any } f:\mathcal{X}\to\mathbb{R},$$

where $s' \sim p(\cdot\,|\,s,a)$ and $a' \sim \pi(\cdot\,|\,s')$. Throughout the paper, we suppose that $\mathcal{P}^\pi$ operates in a function class $\mathcal{Q}$, such that we can approximate unknown Q functions within this family. Assume without loss of generality that $\mathbf{1} \in \mathcal{Q}$.

Assumption 1 (Function class)**.** *For any $f \in \mathcal{Q}$, $\mathcal{P}^\pi f \in \mathcal{Q}$, and $r \in \mathcal{Q}$. It follows that $Q_0^\pi, \ldots, Q_H^\pi \in \mathcal{Q}$, where $\mathcal{Q} \subseteq \mathbb{R}^{\mathcal{X}}$.*

In most parts of the paper, we assume that the transition data are collected from multiple independent episodes.

Assumption 2 (Data generating process)**.** *The dataset $\mathcal{D}$ consists of samples from K i.i.d. episodes $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_K$. Each $\boldsymbol{\tau}_k$ has H consecutive sample transitions generated by some policy on a single sample path, i.e., $\boldsymbol{\tau}_k = (s_{k,0}, a_{k,0}, r'_{k,0}, s_{k,1}, a_{k,1}, r'_{k,1}, \ldots, s_{k,H}, a_{k,H}, r'_{k,H})$.*

We will focus mainly on the case where $\mathcal{Q}$ is a linear space spanned by d feature functions $\phi_1, \ldots, \phi_d$. Also note that the behavior policy $\overline{\pi}$ is *not* known.

Notations Denote $\mathcal{X} = \mathcal{S} \times \mathcal{A}$. Let $\mathbb{R}^{\mathcal{X}}$ be the collection of all functions $f : \mathcal{X} \to \mathbb{R}$. For any $f \in \mathbb{R}^{\mathcal{X}}$, define $f^\pi : \mathcal{S} \to \mathbb{R}$ by $f^\pi(s) = \int_{\mathcal{A}} f(s,a)\pi(a\,|\,s)\mathrm{d}a$. If A is a positive symmetric semidefinite matrix, let $\sigma_{\min}(A)$ denote its smallest eigenvalue, and let $A^{1/2}$ denote the positive symmetric semidefinite matrix that $A^{1/2}A^{1/2} = A$. For nonnegative $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we denote $a_n \lesssim b_n$ if there exists $c > 0$ such that $a_n \le c b_n$ for $n = 1, 2, \ldots$. Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables and $\{a_n\}_{n=1}^\infty \subseteq \mathbb{R}$ be deterministic. We write $X_n = O_{\mathbb{P}}(a_n)$ if for any $\delta > 0$ there exists $M > 0$ such that $\mathbb{P}(|X_n| > a_n M) \le \delta$ for all n. If a distribution p is absolutely continuous with respect to distribution q, the Pearson χ^2-divergence is defined by $\chi^2(p,q) := \mathbb{E}_q\big[(\frac{\mathrm{d}p}{\mathrm{d}q} - 1)^2\big]$.

3. Regression-Based Off-Policy Evaluation

We consider a fitted Q-iteration method for new policy evaluation using linear function approximation. We show that it is equivalent to a model-based method that estimates a conditional mean operator that embeds the unknown p into the feature space. They admit a simple matrix-vector implementation when $\mathcal{Q}$ is a linear model with finite dimension.

3.1. Fitted Q-iteration (FQI)

The Q-functions satisfy the Bellman equation

$$Q_{h-1}^\pi(s,a) = r(s,a) + \mathbb{E}\big[V_h^\pi(s') \,\big|\, s,a\big] \tag{3}$$

for $h = 1, 2, \ldots, H$, where $s' \sim p(\cdot\,|\,s,a)$, $V_h^\pi : \mathcal{S} \to \mathbb{R}$ is the value function defined as $V_h^\pi(s) := \int_{\mathcal{A}} Q_h^\pi(s,a)\pi(a\,|\,s)\mathrm{d}a$.

For the given target policy π, we apply regression recursively by letting $\widehat{Q}_{H+1}^\pi := 0$ and for $h = H, H-1, \ldots, 0$,

$$\widehat{Q}_h^\pi := \arg\min_{f\in\mathcal{Q}} \Bigg\{ \sum_{n=1}^N \Big(f(s_n,a_n) - r'_n - \int_{\mathcal{A}} \widehat{Q}_{h+1}^\pi(s'_n,a)\pi(a\,|\,s'_n)\mathrm{d}a \Big)^2 + \lambda\rho(f) \Bigg\}, \tag{4}$$

where $\lambda \ge 0$ and $\rho(\cdot)$ is a regularization function. The scheme above provides a recursive way to evaluate $\widehat{Q}_H^\pi, \widehat{Q}_{H-1}^\pi, \ldots, \widehat{Q}_0^\pi$ and v^π by regression using empirical data. It is essentially a fitted Q-iteration. The full algorithm is summarized in Algorithm 1.

Algorithm 1 Fitted Q-iteration for Off-Policy Evaluation (FQI-OPE)

Input: initial distribution ξ_0, target policy π, horizon H, function class $\mathcal{Q}$, sample transitions $\mathcal{D} = \{(s_n, a_n, s'_n, r'_n)\}_{n=1}^N$

Let $\widehat{Q}_{H+1}^\pi := 0$;
for $h = H, H-1, \ldots, 1$ **do**
  Calculate $\widehat{Q}_h$ by solving (4);
end for
Output: $\widehat{v}_{\mathsf{FQI}}^\pi := \int_{\mathcal{X}} \widehat{Q}_0^\pi(s,a)\xi_0(s)\pi(a\,|\,s)\mathrm{d}s\mathrm{d}a$

3.2. An equivalent model-based method using conditional mean operator

The preceding FQI method can be equivalently viewed as a model-based plug-in estimator. Recall the *conditional transition operator* $\mathcal{P}^\pi : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is

$$\mathcal{P}^\pi f(s,a) := \mathbb{E}^\pi\big[f(s',a') \,\big|\, s,a\big] \quad \text{for any } f:\mathcal{X}\to\mathbb{R}.$$

Under Assumption 1, it always holds that $\mathcal{P}^\pi Q_h^\pi \in \mathcal{Q}$. To this end, we are only interested in a "projection" of ground-truth $\mathcal{P}^\pi$ onto $\mathcal{Q}$. We estimate the conditional transition operator by $\widehat{\mathcal{P}}^\pi$: for any $f : \mathcal{X} \to \mathbb{R}$, let

$$\widehat{\mathcal{P}}^\pi f := \arg\min_{g\in\mathcal{Q}} \Bigg\{ \sum_{n=1}^N \Big(g(s_n,a_n) - \int_{\mathcal{A}} f(s'_n,a)\pi(a\,|\,s'_n)\mathrm{d}a \Big)^2 + \lambda\rho(g) \Bigg\}. \tag{5}$$

We can see that, if $N \to \infty$, $\widehat{\mathcal{P}}^\pi$ converges to a projected version of $\mathcal{P}^\pi$ onto $\mathcal{Q}$. Denote $\phi(\cdot) := [\phi_1(\cdot), \ldots, \phi_d(\cdot)]^\top : \mathcal{X} \to \mathbb{R}^d$. In the case where $\mathcal{Q}$ is a linear space given by $\mathcal{Q} = \big\{\phi(\cdot)^\top w \,\big|\, w \in \mathbb{R}^d\big\}$ and $\rho(\cdot)$ is taken as

$$\rho(f) := \|w\|_2^2 \qquad \text{for } f(\cdot) = \phi(\cdot)^\top w, \tag{6}$$

the constructed $\widehat{\mathcal{P}}^\pi$ in (5) corresponds to an estimated $\widehat{p}$ of the form

$$\widehat{p}(\cdot\,|\,s,a) := \phi(s,a)^\top \widehat{\Sigma}^{-1} \left(\sum_{n=1}^N \phi(s_n,a_n)\delta_{s'_n}(\cdot) \right),$$

where $\widehat{\Sigma} := \lambda I + \sum_{n=1}^N \phi(s_n,a_n)\phi(s_n,a_n)^\top$ is the empirical covariance matrix and $\delta_{s'}(\cdot)$ denotes the Dirac measure. Note this $\widehat{p}$ is not necessary a transition kernel.

We adopt a model-based approach and use $\widehat{\mathcal{P}}^\pi$ in the Bellman equation as a plug-in estimator. In particular, let

$$\widehat{r} := \arg\min_{f\in\mathcal{Q}} \left\{ \sum_{n=1}^{N} \big(f(s_n,a_n) - r'_n\big)^2 + \lambda\rho(f) \right\}, \quad (7)$$

and $\widehat{Q}^\pi_{H+1} := 0$,

$$\widehat{Q}^\pi_{h-1} := \widehat{r} + \widehat{\mathcal{P}}^\pi \widehat{Q}^\pi_h, \qquad h = H+1, H, \dots, 1.$$

Then we can estimate the policy value by

$$\widehat{v}^\pi_{\text{Plug-in}} := \int_{s,a} \widehat{Q}^\pi_0(s,a)\xi_0(s)\pi(a\,|\,s)\mathrm{d}s\mathrm{d}a.$$

It is easy to verify that this plug-in estimator is equivalent to the earlier FQI estimator. See the proof in Appendix A.

Theorem 1 (Equivalence between FQI and a model-based method)**.** *If $\mathcal{Q}$ is a linear space and ρ is given by (6), Algorithm 1 and the preceding model-based approach generate identical policy value estimators, i.e. $\widehat{v}^\pi := \widehat{v}^\pi_{\text{FQI}} = \widehat{v}^\pi_{\text{Plug-in}}$.*

When $\mathcal{Q}$ is a d-dimensional linear space with feature mapping ϕ, under Assumption 1, there exists a matrix $M^\pi \in \mathbb{R}^{d\times d}$ such that

$$\phi(s,a)^\top M^\pi = \mathbb{E}\big[\,\phi^\pi(s')^\top \,\big|\, s,a\,\big], \quad \forall (s,a)\in\mathcal{X},$$

where $\phi^\pi(s) := \int \phi(s,a)\pi(a|s)da$. We refer to M^π as the *matrix mean embedding* of the conditional transition operator $\mathcal{P}^\pi$. We can implement Algorithm 1 in simple vector forms. We embed the one-step reward function and conditional transition operator into a vector and a matrix, respectively:

$$\widehat{r}(\cdot) = \phi(\cdot)^\top \widehat{R} \quad \text{with } \widehat{R} := \widehat{\Sigma}^{-1}\left(\sum_{n=1}^{N} r'_n \phi(s_n,a_n)\right),$$

$$\widehat{M}^\pi := \widehat{\Sigma}^{-1}\left(\sum_{n=1}^{N} \phi(s_n,a_n)\phi^\pi(s'_n)^\top\right). \quad (8)$$

The corresponding conditional mean operator $\widehat{\mathcal{P}}^\pi$ is

$$\widehat{\mathcal{P}}^\pi f(s,a) = \phi(s,a)^\top \widehat{M}^\pi w, \ \text{ for } f(\cdot) = \phi(\cdot)^\top w. \quad (9)$$

We represent $\widehat{Q}^\pi_h$ in the form of $\widehat{Q}^\pi_h(s,a) = \phi(s,a)^\top \widehat{w}^\pi_h$. In this way, we can easily compute $\widehat{Q}^\pi_h$ using recursive compact vector-matrix operations, as given in Algorithm 2.

Algorithm 2 Conditional Mean Embedding for Policy Evaluation (CME-PE)

Input: initial distribution ξ_0, target policy π, horizon H, a basis $\{\phi_1,\dots,\phi_d\}$ of $\mathcal{Q}$, sample transitions $\mathcal{D} = \{(s_n,a_n,s'_n,r'_n)\}_{n=1}^N$,

Estimate $\widehat{R}$ and $\widehat{M}^\pi$ according to (8);
Let $\widehat{w}^\pi_{H+1} := 0$;
Let $\nu^\pi_0 := \int_{\mathcal{X}} \phi(s,a)\xi_0(s)\pi(a\,|\,s)\mathrm{d}s\mathrm{d}a$;
for $h = H, H-1, \dots, 0$ **do**
  Calculate $\widehat{w}^\pi_h := \widehat{R} + \widehat{M}^\pi \widehat{w}^\pi_{h+1}$;
end for
Output: $\widehat{v}^\pi := (\nu^\pi_0)^\top \widehat{w}^\pi_0$

3.3. Relations to other methods

Our method turns out to be closely related to variants of importance sampling method for OPE. For examples:

- *Marginalized importance sampling:* Our FQI estimator takes the form

  $$\widehat{v}^\pi = \frac{1}{N}\sum_{n=1}^{N} \widehat{w}_{\pi/\mathcal{D}}(s_n,a_n) r'_n$$

  where

  $$\widehat{w}_{\pi/\mathcal{D}}(s,a) := N \sum_{h=0}^{H} (\nu^\pi_0)^\top (\widehat{M}^\pi)^h \widehat{\Sigma}^{-1}\phi(s,a).$$

  By viewing $\widehat{w}_{\pi/\mathcal{D}}(s,a)$ as weights, our estimator can be obtained equivalently by importance sampling. In the special tabular case, our $\widehat{v}^\pi$ is *equivalent* to the marginalized importance sampling (MIS) estimator in (Yin & Wang, 2020).
- *DualDICE:* Nachum et al. (2019) proposed a minimax formulation to find the stationary state occupancy measure and residue (weight for importance sampling) with function approximation . We observe that, if those function classes are taken to be $\mathcal{Q}$, a version of DualDICE produces the same estimator as the FQI estimator. The two methods can be viewed as dual to each other.

These intriguing relations permit a unified view of OPE methods. See Appendix A for more discussions.

4. Finite-Sample Error Bound

Recall that $\mathcal{D}$ is a collection of K independent H-horizon trajectories. Let Σ be the uncentered covariance matrix of the data distribution:

$$\Sigma = \mathbb{E}\left[\frac{1}{H}\sum_{h=0}^{H-1} \phi(s_{1,h},a_{1,h})\phi(s_{1,h},a_{1,h})^\top\right],$$

which is determined by the unknown behavior policy $\overline{\pi}$. Given a target policy π, let ξ^π be an invariant distribution of the Markov chain with transition kernel $p^\pi(s'\,|\,s) = \int_{\mathcal{A}} p(s'\,|\,s,a)\pi(a\,|\,s)\mathrm{d}a$. Define

$$\Sigma^\pi := \mathbb{E}\big[\phi^\pi(s)\phi^\pi(s)^\top \,\big|\, s\sim\xi^\pi\big].$$

We assume $\phi(s,a)^\top\Sigma^{-1}\phi(s,a) \le C_1 d$ without loss of generality. Theorem 2 provides an instance-dependent policy evaluation upper bound. Its complete proof is given in Appendix B.

Theorem 2 (Upper bound)**.** *Let $\delta\in(0,1)$. Under Assumptions 1 and 2, if $N \ge 20\kappa_1(2+\kappa_2)^2\ln(12dH/\delta)C_1dH^3$ and $\lambda \le \ln(12dH/\delta)C_1dH\sigma_{\min}(\Sigma)$, then with probability at least $1-\delta$,*

$$\begin{aligned}&\left|v^\pi-\widehat{v}^\pi\right|\\ \le&\sum_{h=0}^{H}(H-h+1)\sup_{f\in\mathcal{Q}}\frac{\mathbb{E}^\pi\left[f(s_h,a_h)\,\middle|\,s_0\sim\xi_0\right]}{\sqrt{\mathbb{E}\left[\frac{1}{H}\sum_{h=0}^{H-1}f^2(s_{1,h},a_{1,h})\right]}}\\ &\cdot\sqrt{\frac{\ln(12/\delta)}{2N}}+\frac{C\ln(12dH/\delta)dH^{3.5}}{N},\end{aligned}\tag{10}$$

where $C:=15\kappa_1C_1(3+\kappa_2)\sqrt{(\nu_0^\pi)^\top\Sigma^{-1}\nu_0^\pi}$,
$\kappa_1:=\mathrm{cond}\big(\Sigma^{-1/2}\Sigma^\pi\Sigma^{-1/2}\big)$,
$\kappa_2:=\left\|\Sigma^{-1/2}\mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H}\phi^\pi(s_{1,h})\phi^\pi(s_{1,h})^\top\right]\Sigma^{-1/2}\right\|_2\vee 1$.

Additionally, if either one of the following holds:

- $\phi(s,a)^\top\Sigma^{-1}\phi(s',a')\ge 0$ *for any* $(s,a),(s',a')\in\mathcal{X}$;
- *the MDP is time-inhomogeneous,*

the upper bound can be improved to

$$\begin{aligned}&\left|v^\pi-\widehat{v}^\pi\right|\\ \le&\sup_{f\in\mathcal{Q}}\frac{\mathbb{E}^\pi\left[\sum_{h=0}^{H}(H-h+1)f(s_h,a_h)\,\middle|\,s_0\sim\xi_0\right]}{\sqrt{\mathbb{E}\left[\frac{1}{H}\sum_{h=0}^{H-1}f^2(s_{1,h},a_{1,h})\right]}}\\ &\cdot\sqrt{\frac{\ln(12/\delta)}{2N}}+\frac{C\ln(12dH/\delta)dH^{3.5}}{N}.\end{aligned}\tag{11}$$

Distributional mismatch as a $\mathcal{Q}$-χ^2-divergence.
Let $\overline{\mu}$ be the expected occupancy measure of observation $\{(s_n,a_n)\}_{n=1}^N$. Let μ^π be the weighted occupancy distribution of (s_h,a_h) under policy π and ξ_0, given by

$$\mu^\pi(s,a):=\frac{\mathbb{E}^\pi\left[\sum_{h=0}^{H}(H-h+1)\mathbf{1}(s_h=s,a_h=a)\right]}{\sum_{h=0}^{H}(H-h+1)}.$$

The upper bound (11) can be simplified to

$$|\widehat{v}^\pi-v^\pi|\le CH^2\sqrt{\frac{1+\chi^2_{\mathcal{Q}}(\mu^\pi,\bar{\mu})}{N}}+O(N^{-1}).$$

Moreover, each mismatch term in (10) has a vector form

$$\frac{\mathbb{E}^\pi[f(s_h,a_h)\,|\,s_0\sim\xi_0]}{\sqrt{\mathbb{E}[\frac{1}{H}\sum_{h=0}^{H-1}f^2(s_{1,h},a_{1,h})]}}=\sqrt{(\nu_h^\pi)^\top\Sigma^{-1}\nu_h^\pi},$$

where $\nu_h^\pi:=\mathbb{E}^\pi[\phi(s_h,a_h)\,|\,s_0\sim\xi_0]$, so it can be estimated tractably.

The case of tabular MDP.
In the tabular case, the condition $\phi(s,a)^\top\Sigma^{-1}\phi(s',a')\ge 0$ holds for all $(s,a),(s',a')\in\mathcal{X}$. It can be easily seen that the error bound (11) has a strong connection with the χ^2-divergence between the state-action distributions under the behavior and target policies.

Corollary 1 (Upper bound in tabular case)**.** *In the tabular case with $\mathcal{Q}=\mathbb{R}^{\mathcal{X}}$, if N is sufficiently large and $\lambda=0$, then with probability at least $1-\delta$,*

$$\left|v^\pi-\widehat{v}^\pi\right|\le 3H^2\sqrt{1+\chi^2(\mu^\pi,\overline{\mu})}\sqrt{\frac{\ln(12/\delta)}{2N}}+O(N^{-1}),\tag{12}$$

where $\chi^2(\cdot,\cdot)$ denotes the Pearson χ^2-divergence. If the MDP is also time-inhomogeneous, then

$$\begin{aligned}|v^\pi-\widehat{v}^\pi|\le&\sqrt{H\sum_{h=0}^{H}\sum_{s,a}\frac{\mu_h^\pi(s,a)^2}{\overline{\mu}_h(s,a)}\mathrm{Var}\left[r'+V_{h+1}^\pi(s')\,\middle|\,s,a\right]}\\ &\cdot\sqrt{\frac{2\ln(12/\delta)}{N}}+o(N^{-1/2}),\end{aligned}\tag{13}$$

where $\overline{\mu}_h$ is the marginal distribution of $(s_{1,h},a_{1,h})$ and μ_h^π is the marginal distribution of (s_h,a_h) under policy π and ξ_0.

The tabular-case upper bound (13) has the same form with Theorem 3.1 in Yin & Wang (2020). The proof of Corollary 1 is deferred to Appendix B.7.

4.1. Proof Outline

We decompose the error into three terms: $v^\pi-\widehat{v}^\pi=E_1+E_2+E_3$, where E_1 is a linear function of $\widehat{\mathcal{P}}^\pi-\mathcal{P}^\pi$, E_2 is a high-order function of $\widehat{\mathcal{P}}^\pi-\mathcal{P}^\pi$ and $E_3=O(\lambda)$. In the following, we outline the analysis of E_1 and E_2.

First-order term E_1: This linear error term takes the form $E_1=\frac{1}{N}\sum_{n=1}^{N}e_n$, where

$$\begin{aligned}e_n:=&\sum_{h=0}^{H}(\nu_h^\pi)^\top\Sigma^{-1}\phi(s_n,a_n)\\ &\cdot\Big(Q_h^\pi(s_n,a_n)-\big(r_n'+V_{h+1}^\pi(s_n')\big)\Big).\end{aligned}$$

Define a filtration $\{\mathcal{F}_n\}_{n=1,\dots,N}$ where $\mathcal{F}_n$ is generated by $(s_1,a_1,s_1',r_1'),\dots,(s_{n-1},a_{n-1},s_{n-1}',r_{n-1}')$ and (s_n,a_n). Then $e_1,e_2,\dots,e_N$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}_{n=1,\dots,N}$. In what is next, we analyze $\mathrm{Var}[e_n\,|\,\mathcal{F}_n]$ and apply the Freedman's inequality (Freedman, 1975) to derive a finite sample upper bound for E_1.

Consider the conditional variance $\mathrm{Var}[e_n\,|\,\mathcal{F}_n]$. By using the Cauchy-Schwarz inequality and the relation $\mathrm{Var}\left[r_n'+V_{h+1}^\pi(s_n')\,\middle|\,s_n,a_n\right]\le\frac{1}{4}(H-h+1)^2$, we have

$$\begin{aligned}&\mathrm{Var}\left[e_n\,\middle|\,\mathcal{F}_n\right]=\mathbb{E}\left[e_n^2\,|\,s_n,a_n\right]\\ \le&\frac{1}{4}\left(\sum_{h=0}^{H}(H-h+1)\sqrt{(\nu_h^\pi)^\top\Sigma^{-1}\nu_h^\pi}\right)\\ &\cdot\left(\sum_{h=0}^{H}\frac{H-h+1}{\sqrt{(\nu_h^\pi)^\top\Sigma^{-1}\nu_h^\pi}}\big((\nu_h^\pi)^\top\Sigma^{-1}\phi(s_n,a_n)\big)^2\right).\end{aligned}\tag{14}$$

We learn from matrix-form Bernstein inequality that $\frac{1}{N}\sum_{n=1}^{N}\phi(s_n,a_n)\phi(s_n,a_n)^\top$ concentrates around Σ with high probability. It follows that

$$\begin{aligned}&\sum_{n=1}^{N}\left((\nu_h^\pi)^\top\Sigma^{-1}\phi(s_n,a_n)\right)^2\\ =&(\nu_h^\pi)^\top\Sigma^{-1}\left(\sum_{n=1}^{N}\phi(s_n,a_n)\phi(s_n,a_n)^\top\right)\Sigma^{-1}\nu_h^\pi \qquad (15)\\ =&(\nu_h^\pi)^\top\Sigma^{-1}\nu_h^\pi\big(N+\sqrt{dH}\cdot O_{\mathbb{P}}(\sqrt{N})\big).\end{aligned}$$

Plugging (15) into (14) and taking the summation, we obtain

$$\sum_{n=1}^{N}\mathrm{Var}[e_n\,|\,\mathcal{F}_n]\le\frac{1}{4}\left(\sum_{h=0}^{H}(H-h+1)\sqrt{(\nu_h^\pi)^\top\Sigma^{-1}\nu_h^\pi}\right)^2\cdot\big(N+\sqrt{dH}\cdot O_{\mathbb{P}}(\sqrt{N})\big).$$

It follows from the Freedman's inequality that with high probability,

$$|E_1|\lesssim\frac{1}{\sqrt{N}}\sum_{h=0}^{H}(H-h+1)\sqrt{(\nu_h^\pi)^\top\Sigma^{-1}\nu_h^\pi}+\frac{\sqrt{dH}}{N}.$$

High-order term E_2 (bias-inducing term): The high-order term E_2 involves powers of $\widehat{\mathcal{P}}^\pi-\mathcal{P}^\pi$. We use the contraction property of Markov process with respect to its invariant measure, in particular,f

$$\left\|(\Sigma^\pi)^{1/2}M^\pi(\Sigma^\pi)^{-1/2}\right\|_2\le 1. \qquad (16)$$

where $\Sigma^\pi=\mathbb{E}\big[\phi^\pi(s)\phi^\pi(s)^\top\,\big|\,s\sim\xi^\pi\big]$, ξ^π is an invariant distribution under policy π. Assume Σ^π has full rank for simplicity.

By using the contraction property, we will see that the value error will not grow exponentially in H for large N. We have:

$$|E_2|\le\sum_{h=0}^{H}\sqrt{(\nu_0)^\top(\Sigma^\pi)^{-1}\nu_0^\pi}\cdot Err(Q_h^\pi)\cdot\left(\big(1+Err(\widehat{M}^\pi)\big)^h\big(1+Err(N\widehat{\Sigma}^{-1})\big)-1\right), \qquad (17)$$

where the explicit definitions of errors $Err(\widehat{M}^\pi)$, $Err(N\widehat{\Sigma}^{-1})$ and $Err(Q_h^\pi)$ can be found in Lemma B.7, Appendix B.4. By concentration arguments, we can show $Err(\widehat{M}^\pi)$, $Err(N\widehat{\Sigma}^{-1})\lesssim\sqrt{dH/N}$ and $Err(Q_h^\pi)\lesssim(H-h+1)\sqrt{d/N}$ with high probability. According to (17), as long as $Err(\widehat{M}^\pi)\lesssim H^{-1}$, the policy evaluation error will not grow exponentially in H. As a result, if $N\gtrsim dH^3$, we have $|E_2|\lesssim dH^{3.5}/N$. □

5. Minimax Lower Bound

In this section, we establish a minimax lower bound that characterizes the hardness of off-policy evaluation using linear function approximators. Theorem 3 nearly matches with the finite-sample upper bound given by Theorem 2. The complete proof of Theorem 3 is given in Appendix C.

Theorem 3 (Minimax lower bound)**.** *Suppose that an MDP instance $M=(p,r)$ satisfies:*

- *There exists a set of high-value states $\overline{\mathcal{S}}\subseteq\mathcal{S}$ and a set of low-value states $\underline{\mathcal{S}}\subseteq\mathcal{S}$ under the target policy π such that $V_h^\pi(s)\ge\frac{3}{4}(H-h+1)$ if $s\in\overline{\mathcal{S}}$ and $V_h^\pi(s)\le\frac{1}{4}(H-h+1)$ if $s\in\underline{\mathcal{S}}$;*
- *$\overline{p}:=\int_{\overline{\mathcal{S}}}\min_{s\in\mathcal{S}}p^{\overline{\pi}}(s'\,|\,s)\mathrm{d}s'\ge c$ and $\underline{p}:=\int_{\underline{\mathcal{S}}}\min_{s\in\mathcal{S}}p^{\overline{\pi}}(s'\,|\,s)\mathrm{d}s'\ge c$ for $c>0$.*[1]

For any behavior policy $\overline{\pi}$, when N is sufficiently large, one has

$$\inf_{\widehat{v}^\pi}\sup_{M'\in\mathcal{N}(M)}\mathbb{P}_{M'}\left(\left|v^\pi-\widehat{v}^\pi(\mathcal{D})\right|\ge\frac{\sqrt{c}}{24\sqrt{N}}\cdot\sup_{f\in\mathcal{Q}}\frac{\mathbb{E}^\pi_{M'}\big[\sum_{h=0}^{H-1}(H-h)f(s_h,a_h)\,\big|\,s_0\sim\xi_0\big]}{\sqrt{\mathbb{E}_{M'}\big[\frac{1}{H}\sum_{h=0}^{H-1}f^2(s_{1,h},a_{1,h})\big]}}\right)\ge\frac{1}{6}, \qquad (18)$$

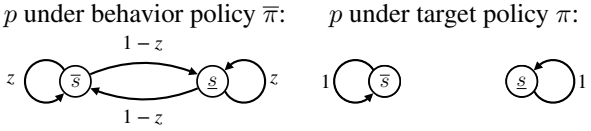
where $\mathcal{N}(M)$ is a small neighborhood of M given by $\mathcal{N}(M):=\big\{M'=(p',r)\;\big|\;\sup_{(s,a)\in\mathcal{X}}\big\|p'(\cdot\,|\,s,a)-p(\cdot\,|\,s,a)\big\|_{\mathrm{TV}}\le\varepsilon\big\}$ ($\|\cdot\|_{\mathrm{TV}}$ denotes the total variation, $\varepsilon\gtrsim\sqrt{cd/N}$). $\mathbb{P}_{M'}$ is the probability space of M', $\widehat{v}^\pi(\mathcal{D})$ is the output of some algorithm $\widehat{v}^\pi$ when $\mathcal{D}$ is given as the input.

Remark. The minimax lower bound is a worst-case error lower bound that applies to *any estimator*, biased or unbiased. Typical minimax lower bound takes the form of $\inf_{\widehat{v}}\sup_{\mathcal{M}}$ where the sup is taken over the entire class of MDP $\mathcal{M}$. Our lower bound is much stronger and can be easily relaxed to the typical form.

Compare Theorems 2 and 3. They nearly match each other, implying that the $\mathcal{Q}$-χ^2-divergence term $\chi^2_{\mathcal{Q}}(\overline{\mu},\mu^\pi)$ determines the statistical complexity of OPE.

An example. Suppose that there is a high-value state $\overline{s}$ and a low-value state $\underline{s}$, which are two absorbing states under the target policy π, with reward 1 and 0 respectively.

We construct ϕ, π and $\overline{\pi}$ such that $\phi^{\overline{\pi}}(\overline{s})=[z,1-z]^\top$, $\phi^{\overline{\pi}}(\underline{s})=[1-z,z]^\top$; and $\phi^\pi(\overline{s})=[1,0]^\top$, $\phi^\pi(\underline{s})=[0,1]^\top$. Here $z\in[0,1]$ is a parameter. We construct the transition model as:



Suppose that the behavior policy $\overline{\pi}$ initiates at either one of the states with probability $1/2$, and the target policy π

[1] We assume the bahavior policy $\overline{\pi}$ is deterministic only for the sake of notational simplicity.

always initiates at state $\bar{s}$. We can see that

$$\Sigma = \begin{bmatrix} z^2 - z + \frac{1}{2} & z(1-z) \\ z(1-z) & z^2 - z + \frac{1}{2} \end{bmatrix},$$

and $\nu_0^\pi = \nu_1^\pi = \ldots = \nu_{H-1}^\pi = [1, 0]^\top$. For $z \in [\frac{1}{4}, \frac{3}{4}]$, the distributional mismatch term controlling the lower bound becomes

$$\Theta(H^2)\sqrt{1 + \frac{1}{(2z-1)^2}},$$

where z quantifies how much one can tell apart the two states under the target policy π using data generated by $\bar{\pi}$. When $z \approx 1/2$, one can not distinguish $\bar{s}$ and $\underline{s}$ from data generated by $\bar{\pi}$, where the lower bound becomes unbounded.

5.1. Proof Outline

We start with an arbitrary MDP M with transition kernel p that satisfies the assumption. We will construct a perturbed instance $\widetilde{p} = p + \Delta p$ so that the two transition models are similar but have a gap in their policy values, denoted by v^π and $\widetilde{v}^\pi$.

Construct the perturbation Δp such that $\Delta p(s' \mid s, a) \geq 0$ if $s' \in \overline{\mathcal{S}}$, $\Delta p(s' \mid s, a) \leq 0$ if $s' \in \underline{\mathcal{S}}$ and $\Delta p(s' \mid s, a) = 0$ elsewhere. In particular, we construct the perturbation as

$$\begin{aligned} &\Delta p(s' \mid s, a) = \phi(s, a)^\top \Delta q(s'), \\ &\Delta q(s') := \mathbf{x} \cdot \min_{s \in \mathcal{S}} p^{\overline{\pi}}(s' \mid s) \cdot \big(\underline{p}\mathbb{1}_{\overline{\mathcal{S}}}(s') - \overline{p}\mathbb{1}_{\underline{\mathcal{S}}}(s')\big), \end{aligned} \tag{19}$$

where $\overline{p}$ and $\underline{p}$ are picked such that $\int_{\mathcal{S}} \Delta p(s' \mid s, a) \mathrm{d}s' = 0$ for any s, a, $\mathbf{x}$ is a vector to be picked later.

Reduction to likelihood test We define likelihood functions $\mathcal{L}(\mathcal{D})$ and $\widetilde{\mathcal{L}}(\mathcal{D})$ of transition kernels p and $\widetilde{p}$. The likelihood ratio $\frac{\widetilde{\mathcal{L}}(\mathcal{D})}{\mathcal{L}(\mathcal{D})} = \prod_{n=1}^N \frac{\widetilde{p}(s_n' \mid s_n, a_n)}{p(s_n' \mid s_n, a_n)}$ reflects how likely the observation $\mathcal{D}$ comes from model $\widetilde{p}$ rather than p. When $p \approx \widetilde{p}$, with high probability, the dataset $\mathcal{D}$ generated by model p has a relatively large likelihood ratio, so that it is hard to distinguish p and $\widetilde{p}$ based on observation $\mathcal{D}$. We prove by a martingale concentration argument that, when N is sufficiently large,

$$\ln\left(\frac{\widetilde{\mathcal{L}}(\mathcal{D})}{\mathcal{L}(\mathcal{D})}\right) \gtrsim -\sqrt{N}\sqrt{\mathbf{x}^\top \Sigma \mathbf{x}} - N \cdot \mathbf{x}^\top \Sigma \mathbf{x}$$

with high probability. In particular, we have

$$\mathbb{P}\left(\frac{\widetilde{\mathcal{L}}(\mathcal{D})}{\mathcal{L}(\mathcal{D})} \geq \frac{1}{2}\right) \geq \frac{1}{2}. \tag{20}$$

when $\sqrt{\mathbf{x}^\top \Sigma \mathbf{x}} \lesssim N^{-1/2}$. If we further have $|v^\pi - \widetilde{v}^\pi| \geq \rho + \widetilde{\rho}$ for some constant gaps $\rho, \widetilde{\rho} \geq 0$, condition (20) implies that for an arbitrary algorithm $\widehat{v}^\pi$, only one of the following must hold: either $\mathbb{P}\big(|v^\pi - \widehat{v}^\pi(\mathcal{D})| \geq \rho\big) \geq \frac{1}{6}$ or $\widetilde{\mathbb{P}}\big(|\widetilde{v}^\pi - \widehat{v}^\pi(\mathcal{D})| \geq \widetilde{\rho}\big) \geq \frac{1}{6}$. In other words, no algorithm can achieve small OPE error for both p and $\widetilde{p}$.

Constructing similar instances with a gap in values We have

$$\widetilde{v}^\pi - v^\pi = \sum_{h=0}^H \xi_0^\top (\widetilde{\mathcal{P}}^\pi)^h (\widetilde{\mathcal{P}}^\pi - \mathcal{P}^\pi) Q_{h+1}^\pi. \tag{21}$$

By first-order Taylor expansion and our construction, if the perturbation Δp is sufficiently small, we have

$$\begin{aligned} \widetilde{v}^\pi - v^\pi &\approx \sum_{h=0}^H \xi_0^\top (\mathcal{P}^\pi)^h (\widetilde{\mathcal{P}}^\pi - \mathcal{P}^\pi) Q_{h+1}^\pi \\ &\gtrsim \sum_{h=0}^{H-1} (H-h)(\nu_h^\pi)^\top \mathbf{x}. \end{aligned} \tag{22}$$

For a given N, we maximize the above value over $\mathbf{x}$ under the constraint $\sqrt{\mathbf{x}^\top \Sigma \mathbf{x}} \lesssim N^{-1/2}$. Then we obtain $\mathbf{x}^* = \frac{c_0 \mathbf{x}_0}{\sqrt{N}\sqrt{\mathbf{x}_0^\top \Sigma \mathbf{x}_0}}$ where $c_0 > 0$ is a constant and $\mathbf{x}_0 = \Sigma^{-1} \sum_{h=0}^{H-1} (H-h)\nu_h^\pi$. In this way, we have shown that $\widetilde{v}^\pi - v^\pi \gtrsim \frac{1}{\sqrt{N}} \big\| \sum_{h=0}^{H-1} (H-h)\nu_h^\pi \big\|_{\Sigma^{-1}}^2$ using the above construction of $\mathbf{x}^*$.

Similarly, one can show that for N sufficiently large, $v^\pi - \widetilde{v}^\pi \geq \rho + \widetilde{\rho}$ for $\rho = \frac{\sqrt{c}}{24\sqrt{N}} \big\| \sum_{h=0}^{H-1} (H-h)\nu_h^\pi \big\|_{\Sigma^{-1}}^2$ and $\widetilde{\rho} = \frac{\sqrt{c}}{24\sqrt{N}} \big\| \sum_{h=0}^{H-1} (H-h)\widetilde{\nu}_h^\pi \big\|_{\widetilde{\Sigma}^{-1}}^2$, where $\widetilde{\nu}_h^\pi$ and $\widetilde{\Sigma}$ are counterparts of ν^π and Σ under the perturbed model $\widetilde{p}$. Finally, we apply the result of the likelihood test and complete the proof. □

6. A Computable Confidence Bound

Next we study how to quantify the uncertainty in the policy evaluator given by Algorithm 1. In this section, we assume that the dataset is an arbitrary set of experiences, not necessarily independent episodes. We only assume that the transition samples $\mathcal{D} = \{(s_n, a_n, s_n', r_n')\}_{n=1,\ldots,N}$ are collected in time order.

Assumption 3. *The dataset $\mathcal{D}$ consists of sample transitions $\{(s_t, a_t, s_t', r_t')\}_{t=1}^N$ generated in time order, i.e. adapted to a filtration $\{\mathcal{F}_t\}_{t=1}^N$, where $\{(s_\tau, a_\tau, s_\tau', r_\tau')\}_{\tau=1}^t$ are $\mathcal{F}_t$-measurable.*

Assumption 3 is much weaker than Assumption 2. It allows the samples to be generated from a long single path possibly under a nonstationary adaptive policy, as is typical in online reinforcement learning.

Under this mildest assumption, we provide a confidence bound for the policy evaluation error $|v^\pi - \widehat{v}^\pi|$, which can be analytically computed from the data $\mathcal{D}$.

Theorem 4 (Computable confidence bound). *Let Assumptions 1 and 3 hold. Let $\omega := \max\big\{\|w\|_2 \,\big|\, 0 \leq \phi(s, a)^\top w \leq 1, \forall (s, a) \in \mathcal{X}\big\}$[2]. Assume $\|\phi(s, a)\|_2 \leq 1$*

[2]Such ω always exists and can be computed priorly

for any $(s,a) \in \mathcal{X}$. For a target policy π, with probability at least $1-\delta$, we have

$$\left|v^\pi - \widehat{v}^\pi\right| \le \sum_{h=0}^{H}(H-h+1)\sqrt{(\widehat{\nu}_h^\pi)^\top \widehat{\Sigma}^{-1}\widehat{\nu}_h^\pi} \cdot \left(\sqrt{2\lambda}\omega + 2\sqrt{2d\ln\Big(1+\frac{N}{\lambda d}\Big)\ln\Big(\frac{3N^2H}{\delta}\Big)} + \frac{4}{3}\ln\Big(\frac{3N^2H}{\delta}\Big)\right), \tag{23}$$

where $\widehat{\nu}_h^\pi$ is given by $(\widehat{\nu}_h^\pi)^\top := (\nu_0^\pi)^\top (\widehat{M}^\pi)^h$.

The proof begins with a decomposition of error given by $v^\pi - \widehat{v}^\pi = \sum_{h=0}^{H}(\widehat{\nu}_h^\pi)^\top\big(w_h^\pi - (\widehat{R} + \widehat{M}^\pi w_{h+1}^\pi)\big)$, from which we derive

$$\left|v^\pi - \widehat{v}^\pi\right| \le \sum_{h=0}^{H}\sqrt{(\widehat{\nu}_h^\pi)^\top \widehat{\Sigma}^{-1}\widehat{\nu}_h^\pi} \cdot \left\|\widehat{\Sigma}^{1/2}\big(w_h^\pi - (\widehat{R} + \widehat{M}^\pi w_{h+1}^\pi)\big)\right\|_2. \tag{24}$$

We analyze the concentration of $\Theta_h := \left\|\widehat{\Sigma}^{1/2}\big(w_h^\pi - (\widehat{R} + \widehat{M}^\pi w_{h+1}^\pi)\big)\right\|_2^2$ using a martingale argument that is similar to the bandit literature (e.g., proof of Theorem 5 in (Dani et al., 2008)). The complete proof is given in Appendix D.

The confidence bound given in Theorem 4 can be easily calculated as a byproduct of FQI-OPE (Algorithm 1), since $\widehat{\nu}_h^\pi, \widehat{\Sigma}$ were already computed in the iterations. In practice, one can tune the value of λ to get the smallest possible confidence bound.

7. Summary

This paper studies the statistical limits of off-policy evaluation using linear function approximation. We establish a minimax error lower bound that depends on a function class-restricted χ^2-divergence between the data distribution and the target policy's occupancy measure. We prove that a regression-based FQI method, which can be viewed as a plug-in estimator based on a conditional mean embedding of the transition operator, nearly achieves the minimax lower bound. We also provide a computable confidence bound as a byproduct of the algorithm.

Acknowledgments

Mengdi Wang gratefully acknowledges funding from the U.S. National Science Foundation (NSF) grant CMMI-1653435, Air Force Office of Scientific Research (AFOSR) grant FA9550-19-1-020, and C3.ai DTI.

References

Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, 2019.

Fonteneau, R., Murphy, S. A., Wehenkel, L., and Ernst, D. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208(1):383–416, 2013.

Freedman, D. A. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Grunewalder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. Modelling transition dynamics in mdps with rkhs embeddings. 2012.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Jong, N. K. and Stone, P. Model-based function approximation in reinforcement learning. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pp. 1–8, 2007.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec): 1107–1149, 2003.

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

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. 5356–5366, 2018.

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

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 72, 2004.

Nachum, O., Chow, Y., Dai, B., and Li, L. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32*. 2019.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

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

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

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

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*, pp. 9665–9675, 2019.

Yang, L. F. and Wang, M. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019a.

Yang, L. F. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *36th International Conference on Machine Learning, ICML 2019*, pp. 12095–12114. International Machine Learning Society (IMLS), 2019b.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. *arXiv preprint arXiv:2001.10742*, 2020.