Submitted to Statistical Science

On the Statistical Complexity for Offline and Low-Adaptive Reinforcement Learning with Structures

Ming Yin, Mengdi Wang and Yu-Xiang Wang

Abstract. This article reviews the recent advances on the statistical foundation of reinforcement learning (RL) in the offline and low-adaptive settings. We will start by arguing why offline RL is the appropriate model for almost any real-life ML problems, even if they have nothing to do with the recent AI breakthroughs that use RL. Then we will zoom into two fundamental problems of offline RL: offline policy evaluation (OPE) and offline policy learning (OPL). It may be surprising to people that tight bounds for these problems were not known even for tabular and linear cases until recently. We delineate the differences between worst-case minimax bounds and instance-dependent bounds. We also cover key algorithmic ideas and proof techniques behind near-optimal instance-dependent methods in OPE and OPL. Finally, we discuss the limitations of offline RL and review a burgeoning problem of *low-adaptive exploration* which addresses these limitations by providing a sweet middle ground between offline and online RL.

Key words and phrases: Sample Complexity, Offline Reinforcement Learning, Low-Adaptive Exploration.

1. INTRODUCTION

Reinforcement learning (RL) has gained remarkable popularity lately. Most people would attribute the surge to its usage in AI milestones such as AlphaGo [17, 51, 56, 75, 76] and in instruction-tuning large language models [8, 12, 65, 77]. We, however, argue that it is caused by a more fundamental paradigm shift that places RL in the front and center of nearly every Machine Learning (ML) application in practice. Why? Training an accurate classifier is most likely not the end goal of an ML task. Instead, the predictions of the trained ML model is often used as interventions hence changing the distribution of future data. Real-world applications are usually sequential decision-making problems, and trained ML models need to be combined with RL methods to perform high-quality decision-making. We provide three examples.

Ming Yin is a Postdoc at the Department of Electrical and Computer Engineering at Princeton University (e-mail: my0049@princeton.edu). Mengdi Wang is an Associate Professor at the Department of Electrical and Computer Engineering at Princeton University (e-mail: mengdiw@princeton.edu). Yu-Xiang Wang is an Associate Professor at the Halıcıoğlu Data Science Institute at UC San Diego (e-mail: yuxiangw@ucsd.edu).

AI Diagnosis/Screening. In medical diagnosis, ML models are frequently used to predict the likelihood of a patient having a certain disease based on their symptoms and medical history. However, these predictions are not the final outcome; they often guide subsequent medical interventions, such as recommending further tests or treatments. These interventions, in turn, influence future patient states, creating a feedback loop that affects the data distribution. RL methods are essential in this context to optimize the sequence of decisions—like treatment plans—over time, improving patient's outcomes. For instance, [62] used RL to develop a model that assists in the management of ICU by recommending treatment strategies that adapt to the evolving condition of the patient.

Recommendation Systems. Traditional recommendation systems rely on ML models to predict user preferences based on historical data. However, when these recommendations are presented to users, they influence user behavior and preferences, which alters future data. This dynamic environment is well-suited to RL, where the goal is to maximize long-term user engagement by continuously adapting recommendations based on real-time feedback. For example, [104] applied RL to optimize a recommendation system for news articles, showing that it could significantly improve user click-through rates by considering the long-term effects of recommendations.

Video Streaming over Wireless Networks. In video streaming applications, ML models are used to predict network conditions and select appropriate streaming bitrates. These predictions directly influence the quality of the streaming experience and the subsequent network load, posing a challenging sequential decision-making problem. RL can be applied to adaptively adjust bitrates to optimize the trade-off between video quality and buffering. For instance, [52] introduced a system called *Pensieve*, which uses RL to optimize video streaming quality over wireless networks by learning from past streaming experiences and network conditions.

These examples not only demonstrate the fundamental applicability of RL across diverse domains but also bring to light the significant challenges it faces.

Notably, most real-life RL problems are *offline RL* problems. Unlike Chess or Go with unlimited access to simulators, it is often unsafe, illegal, or costly to conduct experiments in the task environment. Instead, we need to work with an offline dataset collected from the environment, which poses fundamental problems on *what can be learned offline* and *how (statistically) efficiently one can learn from the offline dataset*. Three critical aspects of the offline RL problems are:

- **Long horizon problem.** The long decision horizon in RL poses unique challenge for finding the optimal strategy. In particular, the undesired actions chosen at earlier phases will have long-lasting impact for the future, making the strategy suboptimal. Small deviations from the optimal policy early on can propagate and amplify over time, further complicating the learning process.
- **Distribution Shift and Coverage.** Distribution shift is a fundamental challenge in reinforcement learning that occurs when the distribution of data the agent encounters during training differs from the distribution of optimal policies. When the overlap (measured by certain distribution distance metric) between the two distributions is small, it would be hard to find optimal actions due to the insufficient data coverage, especially when the offline dataset is collected using a suboptimal policy.
- **Function Approximation and Generalization.** The state and action space of RL problems are often so large that a finite dataset cannot cover. In such cases, RL requires generalization across states through a certain feature representation of the states and a parametric approximation of the value functions. Learning such function approximations offline is more challenging.

This article aims to review recent advances of the statistical foundations for offline RL, covering both problems in *offline policy evaluation* and *offline policy learning*.

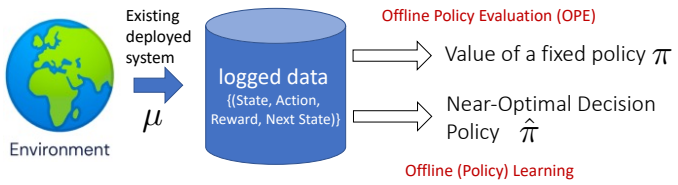

FIG 1. *Illustration of the offline reinforcement learning problem.*

Specifically, we review what the fundamental learning hardness/statistical limits for offline RL under different MDP (Markov Decision Processes) or function approximation structures are. By examining the statistical results, we reveal how factors such as distribution shift and horizon length affect the learning hardness of the problems. We also introduce the related algorithms and highlight the theoretical techniques to achieve these results.

Paper organization. We first introduce the mathematical notations and set up the problems of interest in Section 2. Then the remaining sections describe results in offline policy evaluation, offline policy learning and low-adaptive exploration under various assumptions (see Table 1).

Disclaimer. The literature of offline RL is gigantic. It is not our intention to provide thorough coverage. Instead, the topics and results covered in this paper focus on a niche that the coauthors studied in the past few years. Since our goal is pedagogical, we do not make any claims about novelty and precedence of scientific discovery. Please refer to the bibliography and the references therein for a more detailed discussion.

2. NOTATIONS AND PROBLEM SETUP

We first provide the background for different problem settings that we consider in this article.

2.1 Episodic time-inhomogenuous RL

A finite-horizon *Markov Decision Process* (MDP) is denoted by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, d_1)$ [78], where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. A time-inhomogenuous 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 $0 \le r \le R_{\max}$. 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 $s_1 \sim d_1, a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H]$ and r_h is a random realization given the observed s_h, a_h.

Bellman (optimality) equations. The value function $V_h^\pi(\cdot) \in \mathbb{R}^{\mathcal{S}}$ and Q-value function $Q_h^\pi(\cdot, \cdot) \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ for

Problem setup	Offline Evaluation	Offline Learning	Low-Adaptive Exploration
Tabular MDP	Section 3.3	Section 5	Section 7.1
Linear Approx.	Section 4.1	Section 6.1	Section 7.2
Parametric Approx.	Section 4.2	Section 6.2	Section 7.3

TABLE 1
Overview of the paper structure.

any policy π is defined as, $\forall s, a \in \mathcal{S}, \mathcal{A}, h \in [H]$:

$$V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^{H} r_t | s_h = s], \;\; Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^{H} r_t | s_h, a_h = s, a].$$

The Dynamic Programming principle follows [9, 67, 69] $\forall h \in [H]$:

(1)
$$Q_h^\pi(s,a) = r_h + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^\pi(s')], \;\; V_h^\pi = \mathbb{E}_{a \sim \pi_h}[Q_h^\pi],$$
$$Q_h^*(s,a) = r_h + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^*(s')], \; V_h^* = \max_a Q_h^*(\cdot, a).$$

The corresponding Bellman operators are defined as:

(2)
$$\mathcal{P}_h^\pi(f)(s,a) = r_h + \mathbb{E}_{s' \sim P_h(\cdot|s,a), a' \sim \pi(\cdot|s')}[f(s', a')],$$
$$\mathcal{P}_h(f)(s,a) = r_h + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[\max_{a'} f(s', a')].$$

We incorporate the standard marginal state-action occupancy $d_h^\pi(s,a)$ as: $d_h^\pi(s,a) := \mathbb{P}[s_h = s, a_h = a | s_1 \sim d_1, \pi]$. The performance per policy π is defined as

$$v^\pi := \mathbb{E}_{d_1}[V_1^\pi] = \mathbb{E}_{\pi, d_1}\left[\sum_{t=1}^{H} r_t\right].$$

2.2 Structured MDP models

In this article, we examine three fundamental yet representative MDP models (or related function approximation classes) that are well-structured. Despite their simplicity, as we will discuss in later sections, their statistical limits have not been well-understood until recently.

Tabular MDPs. Tabular MDP is arguably the most simple setting in RL. It is a Markov Decision Process with finite states $|\mathcal{S}| < \infty$ and finite actions $|\mathcal{A}| < \infty$. The most common tabular MDPs, such as Gridworlds, often have small state and action spaces. When the number of states and actions are large, they are generally not treated as discrete and are instead addressed using function approximators.

Linear MDPs. An episodic MDP $(\mathcal{S}, \mathcal{A}, P, r, H, d_1)$ is called a linear MDP with a known (unsigned) feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d unknown (unsigned) measures $\nu_h = (\nu_h^{(1)}, \ldots, \nu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that $\forall s', s \in \mathcal{S}, \; a \in \mathcal{A}, \; h \in [H]$,

$$P_h\left(s' \mid s, a\right) = \left\langle \phi(s,a), \nu_h\left(s'\right)\right\rangle, \; r_h\left(s,a\right) = \langle \phi(x,a), \theta_h \rangle$$

with $\int_{\mathcal{S}} \|\nu_h(s)\| \, ds \le \sqrt{d}$ and $\max(\|\phi(s,a)\|_2, \|\theta_h\|_2) \le 1$ for all $h \in [H]$ and $\forall s, a \in \mathcal{S} \times \mathcal{A}$.

When specify $d = |\mathcal{S}| \times |\mathcal{A}|$ and $\phi(x,a) = \mathbf{1}_{(x,a)}$ be the canonical basis in $\mathbb{R}^d$, linear MDPs recover tabular MDPs. Thus, linear MDPs strictly generalize tabular MDPs and allow continuous state-actions spaces.

Linear Functions approximation. By Bellman equation (1), Linear MDP aimplies that the value function Q_h^π for any policy π is a linear function in the feature vector ϕ, i.e.,

$$Q_h^\pi(\cdot,\cdot) \in \left\{ \langle \phi(\cdot,\cdot), \theta \rangle \mid \theta \in \mathbb{R}^d \right\}$$

for any $h \in [H]$ and π. It is sometimes sufficient to directly reason about these linear function approximations rather than relying on the stronger linear MDPs structures. There are various subtle differences in the various type of linear function approximation. For the purpose of this paper though, it suffices to just think about linear MDPs.

For more general MDPs, it is harder to impose tractable structures. Alternatively, we consider the following structured function class that is expressive enough to learn general MDPs.

Parametric Differentiable Functions. Let $\mathcal{S}, \mathcal{A}$ be arbitrary state, action spaces and a feature map $\phi(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \Psi \subset \mathbb{R}^m$. The parameter space $\Theta \in \mathbb{R}^d$. Both Θ and Ψ are compact spaces. Then the parametric function class (for a model $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$) is defined as

$$\mathcal{F} := \{ f(\theta, \phi(\cdot,\cdot)) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \theta \in \Theta \}$$

that satisfies differentiability/smoothness condition: 1. for any $\phi \in \mathbb{R}^m$, $f(\theta, \phi)$ is third-time differentiable with respect to θ; 2. $f, \partial_\theta f, \partial^2_{\theta,\theta} f, \partial^3_{\theta,\theta,\theta} f$ are jointly continuous for (θ, ϕ). Clearly, $\mathcal{F}$ generalizes linear function class (via choosing $f(\theta, \phi) = \langle \theta, \phi \rangle$).

2.3 Offline RL Tasks

The offline RL begins with a static offline data $\mathcal{D} = \left\{ \left(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau \right) \right\}_{\tau \in [n]}^{h \in [H]}$ rolled out from some behavior policy μ. In particular, the offline nature requires we cannot change μ and in particular we do not assume the functional knowledge of μ. There are two major tasks considered in offline RL.

- **Offline Policy Evaluation (OPE).** For a policy of interest π, the agent needs to evaluate its performance v^π using $\mathcal{D}$. In general, there is a distribution mismatch between π and μ. The goal is to construct an estimator $\widehat{v}^\pi$ such that $|v^\pi - \widehat{v}^\pi| < \epsilon$ or mean square error $\mathbb{E}_\mu[(v^\pi - \widehat{v}^\pi)^2] < \epsilon$).

- **Offline Policy Learning (OPL).** This requires the agent to find a reward-maximizing policy $\pi^* := \text{argmax}_\pi v^\pi$ given data $\mathcal{D}$. 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 $v^* - v^{\pi_{\text{alg}}} \leq \epsilon$.

Both OPE and OPL are essential to a real-world offline RL system since the decision maker should first run the offline learning algorithm to find a near optimal policy and then use OPE methods to check if the obtained policy is good enough. For instance, in finance, OPL can be applied for learning a strategy, but traders still need to run OPE for backtesting before deployment. On the other hand, they are also standalone research questions, *e.g.* doctors can be asked to evaluate a heuristic treatment plan that does not involve offline learning, which makes it a pure OPE problem.

REMARK 1. The source of historical data $\mathcal{D}$ could be multilateral, and there are papers (e.g. [34, 74]) directly considers data distribution without specifying μ. We incorporate a specific behavior policy μ to manifest the distribution mismatch between μ and π.

2.4 Assumptions in offline RL

Due to the inherent distribution shift in offline RL, for both OPE and OPL, we revise different assumptions for different problem classes in Section 2.2. These assumptions are standard protocols for deriving provably efficient results.

Offline Policy Evaluation. For tabular OPE, it requires marginal state ratios and policy ratios to be finite, as stated below.

ASSUMPTION 1 (Tabular OPE [90, 94]). *Logging policy μ obeys that $d_m := \min_{t,s} d_t^\mu(s) > 0$. Also, $\tau_s := \max_{t,s} \frac{d_t^\pi(s)}{d_t^\mu(s)} < +\infty$ and $\tau_a := \max_{t,s,a} \frac{\pi(a|s)}{\mu(a|s)} < +\infty$.*

Having bounded weights is necessary for discrete state and actions, as otherwise the unbounded importance ratio would cause the estimation error become intractable.

ASSUMPTION 2 (Linear OPE [14, 27]). *Let the population feature covariance $\Sigma_h := \mathbb{E}_{\mu,h}\left[\phi(s,a)\phi(s,a)^\top\right]$. Then we assume $\min_h \lambda_{min}(\Sigma_h) > 0$ with λ_{min} being the minimal eigenvalue.*

This assumption ensures the behavior policy μ has good coverage over the state-action spaces. For instance, when $\phi(x,a) = \mathbf{1}_{(x,a)}$, the assumption above reduces to $\min_{s,a} d_h^\mu(s,a) > 0$.

ASSUMPTION 3 (Parametric OPE [100]). *Policy completeness: assume reward $r \in \mathcal{F}$ and for any $f \in \mathcal{F}$, we have $\mathcal{P}^\pi f \in \mathcal{F}$. Policy realizability: assume $Q^\pi(\cdot,\cdot) = f(\phi(\cdot,\cdot),\theta^\pi)$ for some $\theta^\pi \in \mathbb{R}^d$. Lastly, let the population feature covariance*

$$\Sigma_h := \mathbb{E}_{\mu,h}\left[\nabla f(\phi(s,a),\theta^\pi)\nabla f(\phi(s,a),\theta^\pi)^\top\right].$$

Then we assume $\min_h \lambda_{min}(\Sigma_h) > 0$.

Policy completeness and policy realizability ensure the policy class is rich enough to capture Q^π. Besides, the assumption on the population feature covariance generalizes the Linear OPE case.

Offline Policy Learning. Next, we summarize the common assumptions (from strong to weak) that can yield statistical sample efficiency for policy learning. After that, we introduce extra assumptions for offline learning in the function approximation settings.

ASSUMPTION 4 (Uniform data coverage [74, 96]). *For behavior policy, $d_m := \min_{h,s,a} d_h^\mu(s,a) > 0$. 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.*[1]

This is the strongest assumption in offline RL as it requires μ to explore each state-action pairs with positive probability at different time step h. For tabular RL, it mostly holds $1/SA \geq d_m$ under Assumption 4. This reveals offline learning is generically harder than *the generative model setting* [2, 43] in the statistical sense. On the other hand, for task where it needs to evaluate different policies simultaneously (such as *uniform OPE* task in [96]), this is required as the task considered is in general a harder task than offline learning.

ASSUMPTION 5 (Uniform concentrability [11, 42, 79, 89]). *For all policy π, $C_\mu := \sup_{\pi,h} ||d_h^\pi(\cdot,\cdot)/d_h^\mu(\cdot,\cdot)||_\infty$. The parameter $C_\mu < +\infty$ is commonly known as "concentrability efficient".*

This is a classical offline RL condition that is commonly assumed in the function approximation scheme (*e.g.* Fitted Q-Iteration in [42, 79], MSBO in [89]). Qualitatively, this is a uniform data-coverage assumption that is similar to Assumption 4, but quantitatively, the coefficient C_μ can be smaller than $1/d_m$ due the d_h^π term in the numerator. There are other variants of concentrability efficient [63, 91] that capture the data coverage of behavior policy slightly differently.

[1] Note here d_m is defined by minimizing over the state and action spaces. For Assumption 1, d_m only concerns state.

ASSUMPTION 6 (Single policy coverage [47, 95]). *There exists one optimal policy π^*, such that $\forall s_h, a_h \in \mathcal{S}, \mathcal{A}$, $d_h^\mu(s_h, a_h) > 0$ if $d_h^{\pi^*}(s_h, a_h) > 0$. We further denote the trackable set as $\mathcal{C}_h := \{(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$.*

Assumption 6 is arguably the weakest assumption needed for accurately learning the optimal value v^*. It only requires μ to trace the state-action space of one optimal policy and can be agnostic at other locations.

ASSUMPTION 7 (Realizability+Bellman Completeness [98]). *The parametric function class $\mathcal{F}$ in Section 2.2 satisfies: 1. Realizability: for optimal Q_h^*, there exists $\theta_h^* \in \Theta$ such that $Q_h^*(\cdot,\cdot) = f(\theta_h^*, \phi(\cdot))$ $\forall h$; 2. Bellman Completeness: let $\mathcal{G} := \{V(\cdot) \in \mathbb{R}^{\mathcal{S}} : s.t.\ \|V\|_\infty \le H\}$. Then in this case $\sup_{V\in\mathcal{G}} \inf_{f\in\mathcal{F}} \|f - \mathcal{P}_h(V)\|_\infty = 0$.*

Realizability and Bellman Completeness are widely adopted in the offline RL analysis with general function approximations [11, 91], and they are assumed to ensure class $\mathcal{F}$ is expressive enough to capture the Q-values of the problems and any bounded functions after Bellman updates.

Additional structural data coverage assumption. For linear OPL task, we adopt the Assumption 2 from linear OPE. As explained before, this assumption is a characterization of Assumption 4 with linear features.

ASSUMPTION 8 (Linear OPL, Identical to Assumption 4). *Let the population feature covariance $\Sigma_h := \mathbb{E}_{\mu,h}\left[\phi(s,a)\phi(s,a)^\top\right]$. Then we assume $\min_h \lambda_{min}(\Sigma_h) > 0$ with λ_{min} being the minimal eigenvalue.*

For parametric differentiable function class $\mathcal{F}$, we impose the following structural data coverage assumption to replace Assumption 4-6. The statistical limit is achieved due to properly leveraging the assumptions on the gradient covariance and the quadratic structure. It depends on both the MDPs and the function approximation class $\mathcal{F}$.

ASSUMPTION 9 (Uniform Coverage for $\mathcal{F}$). *We assume there exists $\kappa > 0$, such that $\forall h \in [H], \theta_1, \theta_2, \theta \in \Theta$,*

- $\mathbb{E}_{\mu,h}\left[\left(f(\theta_1, \phi(\cdot,\cdot)) - f(\theta_2, \phi(\cdot,\cdot))\right)^2\right] \ge \kappa \|\theta_1 - \theta_2\|_2^2$,
- $\mathbb{E}_{\mu,h}\left[\nabla f(\theta, \phi(s,a)) \cdot \nabla f(\theta, \phi(s,a))^\top\right] \succ \kappa I$,

In the linear function approximation regime, Assumption 9 reduces to Assumption 8. The first condition serves more for the "optimization" purpose as it can be cast as a variant of the *quadratic growth condition* [3]. For more discussion about this condition, please refer to [13, 98]. We will go through the statistical limits of offline policy learning with these assumptions.

3. OFFLINE POLICY EVALUATION IN CONTEXTUAL BANDITS AND TABULAR RL

Let's start by the problem of offline policy evaluation (OPE) — the problem of evaluating a fixed target policy π using data collected by executing a logging (or behavior) policy μ.

Readers may wonder why this is even a problem. Admittedly, in *supervised learning*, one can simply evaluate a classifier policy on a validation dataset. Similarly, in *online RL*, one can roll out policy π to see how well it works.

The problem starts to arise in offline problems because we do not have data directly associated with policy π.

3.1 OPE in contextual bandits

Let us build intuition by considering the contextual bandit problem. Contextual bandit (CB) problem is a special case of RL with horizon $H = 1$, where the initial state s is referred to as the "context". For short horizon problems such as CB, the main challenge is to handle *distribution shift*. Motivated by a change of measure formula

$$v_{\text{CB}}^\pi = \mathbb{E}_{\substack{s\sim d_1,\\ a\sim\pi(\cdot|s)}}[r(s,a)] = \mathbb{E}_{\substack{s\sim d_1,\\ a\sim\mu(\cdot|s)}}[\frac{\pi(a|s)}{\mu(a|s)} r(s,a)],$$

classical methods employ *importance sampling* (IS) [29, 45, 68] to corrects the mismatch in the distributions under the behavior policy μ and target policy π. Specifically, let the importance ratio be $\rho := \pi(a|s)/\mu(a|s)$, then the IS estimator is computed as:

$$\widehat{v}_{\text{IS-CB}}^\pi = \frac{1}{n}\sum_{i=1}^n \rho^{(i)} r^{(i)}.$$

It is known to be effective for real-world applications such as news article recommendations [15, 44].

The mean square estimation error (MSE) of the IS estimator $\widehat{v}_{\text{IS-CB}}^\pi$ decomposes into two terms

$$\frac{1}{n}(\mathbb{E}_\mu[\rho(s,a)^2 \text{Var}[r|s,a]] + \text{Var}_\mu[\rho(s,a)\mathbb{E}[r|s,a]]).$$

The first term comes from the noisy reward while the second term comes from the random (s,a) pair. Interestingly, if we make no assumption about $\mathbb{E}[r|s,a]$ and the size of the state-space is large, then IS is minimax optimal [86, Theorem 1]. On the contrary, if $\mathbb{E}[r|s,a]$ can be estimated sufficiently accurately, then there are methods that asymptotically do not depend on the $\text{Var}[\mathbb{E}[\cdot]]$.

Perhaps a bit surprising to some readers, the above conclusion implies that even for *on-policy* evaluation, i.e., $\pi = \mu$ and $\rho \equiv 1$, the naive value estimator of $\frac{1}{n}\sum_i r^{(i)}$ (IS with $\rho \equiv 1$) can be substantially improved using a good reward model.

3.2 "Curse of Horizon" in OPE for RL

The IS estimators are later adopted for long horizon sequential decision making (RL) problems. Concretely, denote the t-step importance ratio $\rho_t := \pi_t(a_t|s_t)/\mu_t(a_t|s_t)$ and the cumu-

lative importance ratio $\rho_{1:t} := \prod_{t'=1}^{t} \rho_{t'}$, the (stepwise) Importance Sampling estimators for RL are defined as:

$$\widehat{v}^{\pi}_{\text{IS}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{v}^{(i)}_{\text{IS}}, \quad \widehat{v}^{(i)}_{\text{IS}} := \rho^{(i)}_{1:H} \cdot \sum_{t=1}^{H} r^{(i)}_t;$$

$$\widehat{v}^{\pi}_{\text{step-IS}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{v}^{(i)}_{\text{step-IS}}, \quad \widehat{v}^{(i)}_{\text{step-IS}} := \sum_{t=1}^{H} \rho^{(i)}_{1:t} r^{(i)}_t,$$

where $\rho^{(i)}_{1:t} = \prod_{t'=1}^{t} \pi_{t'}(a^{(i)}_{t'}|s^{(i)}_{t'})/\mu_{t'}(a^{(i)}_{t'}|s^{(i)}_{t'})$. In addition, there are many works extend IS estimators and different variants such as *weighted IS estimators* and *doubly robust estimators* [15, 28, 32, 59] are proposed.

While IS-based OPE methods can correct the distribution shift and are statistically unbiased, the variance of the cumulative importance ratios $\rho_{1:t}$ may grow exponentially as the horizon goes long. We provide two concrete examples in Appendix A which demonstrate that IS-based methods suffer from exponential variance even in the simplest tabular RL problems.

To make matters worse, the exponential sample complexity in H cannot be improved in general in the large state-space regime unless we make additional assumptions [32]. This is known as the "curse of horizon" in offline RL. We refer readers to a sister article [33] that appears in the same issue of this journal for a quest to obtain sufficient and necessary conditions that enable $\text{poly}(H)$ sample complexity.

Instead of inspecting the exponential separation, we zoom into three well-established sufficient conditions (from Section 2) that circumvent the "curse of horizon" and focus on providing fine-grained statistical characterization of the optimal OPE error bound and design adaptive estimators that take advantage of individual problem instances. We will cover the case with small finite state spaces in Section 3.3 and then function approximation in Section 4.

3.3 OPE in Tabular MDPs

The most basic model of interest is the tabular MDP, namely, MDP when the state and action spaces are finite and that the policy μ gets to visit all states and actions that π visits. A statistical lower bound for OPE in the tabular MDP setting is established in [32].

THEOREM 3.1 (Cramer-Rao lower bound for tabular OPE [32]). *For discrete DAG MDPs with horizon H, the variance of any unbiased estimator $\hat{v}$ with n trajectories from policy μ satisfies*

$$n \cdot \text{Var}[\hat{v}] \geq \sum_{t=1}^{H} \mathbb{E}_{\mu}\left[\frac{d^{\pi}(s_t,a_t)^2}{d^{\mu}(s_t,a_t)^2}\text{Var}\left[V^{\pi}_{t+1}(s_{t+1}) + r_t \middle| s_t, a_t\right]\right].$$

The construction of the CR lower bound relies on computing the constrained version of Fisher Information Matrix. Under Assumption 1, this right hand side can be readily bounded by $O(\tau_s \tau_a H^3)$ (after a change of measure into $\mathbb{E}_{\pi}[\cdot]$)[2], which makes the IS estimators with a variance of $\exp(H)$ exponentially suboptimal.

Marginalized Importance Sampling. In [90, 94], we addressed the exponential gap by an idea that is now referred to as Marginalized Importance Sampling. If we re-examine the value objective with RL, by a change of measure formula,

$$v^{\pi} := \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} r_t\right] = \mathbb{E}_{\mu}\left[\sum_{t=1}^{H} \frac{d^{\pi}_t(s_t)}{d^{\mu}_t(s_t)} r^{\pi}_t(s_t)\right]$$

with $r^{\pi}_t(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[r_t(s,a)|s]$. This reformulation reveals, rather than applying $\rho_{1:t}$, we could instead estimate the marginal state density ratio d^{π}_t/d^{μ}_t. Inspired by this observation, the *Marginalized Importance Sampling* (MIS) estimator is defined as

$$\widehat{v}^{\pi}_{\text{MIS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H} \frac{\widehat{d}^{\pi}_t(s^{(i)}_t)}{\widehat{d}^{\mu}_t(s^{(i)}_t)} \widehat{r}^{\pi}_t(s^{(i)}). \tag{3}$$

Different design choices for $\widehat{d}^{\pi}, \widehat{d}^{\mu}, \widehat{r}^{\pi}$ in (3) yield different MIS estimators.

State MIS (SMIS [90]). For SMIS, $\widehat{d}^{\mu}_t(\cdot)$ is directly estimated using the empirical mean, *i.e.* $\widehat{d}^{\mu}_t(s_t) := \frac{1}{n}\sum_i \mathbf{1}(s^{(i)}_t = s_t) := \frac{n_{s_t}}{n}$ whenever $n_{s_t} > 0$ and $\widehat{d}^{\pi}_t(s_t)/\widehat{d}^{\mu}_t(s_t) = 0$ when $n_{s_t} = 0$. Marginal state distributions are estimated via recursion $\widehat{d}^{\pi}_t = \widehat{P}^{\pi}_t \widehat{d}^{\pi}_{t-1}$, followed by the estimations $P^{\pi}_t(s_t|s_{t-1})$ and state reward $r^{\pi}_t(s_t)$ as:

$$\begin{aligned}
\widehat{P}^{\pi}_t(s'|s) &= \frac{1}{n_s}\sum_{i=1}^{n} \frac{\pi(a^{(i)}|s)}{\mu(a^{(i)}|s)} \cdot \mathbf{1}\{(s^{(i)}_{t-1}, s^{(i)}_t) = (s,s')\}; \\
\widehat{r}^{\pi}_t(s) &= \frac{1}{n_s}\sum_{i=1}^{n} \frac{\pi(a^{(i)}|s)}{\mu(a^{(i)}|s)} r^{(i)}_t \cdot \mathbf{1}(s^{(i)}_t = s).
\end{aligned} \tag{4}$$

SMIS (3) explicitly gets rid of the cumulative importance ratio $\rho_{1:t}$ and provides the polynomial sample complexity for horizon under Mean Square Error.

THEOREM 3.2. *Under Assumption 1 and other mild regularity conditions, the MSE of state marginalized importance sampling satisfies*

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{v}^{\pi}_{\text{SMIS}} - v^{\pi})^2\right] \\
=&\frac{1}{n}\sum_{t=1}^{H} \mathbb{E}_{\mu}\left[\frac{d^{\pi}_t(s_t)^2}{d^{\mu}_t(s_t)^2}\text{Var}_{\mu}[\frac{\pi(a_t \mid s_t)}{\mu(a_t \mid s_t)}\left(V^{\pi}_{t+1}(s_{t+1}) + r_t\right)\middle| s_t]\right] \\
&\cdot \left(1 + O(\sqrt{\frac{\log n}{n}})\right) + O(\frac{1}{n^2}).
\end{aligned}$$

The big O notation hides universal constants.

The MSE of SMIS is $O(\tau_s \tau_a H^3/n)$ and the result holds even when the action space is continuous. This exponentially improves over the standard IS.

SMIS however, does not match the Cramer-Rao lower bound. In particular, the asymptotic MSE (modulo a $1 + O(n^{-1/2})$ multiplicative factor and an $O(1/n^2)$ additive factor) is

$$\frac{1}{n}\sum_{t=1}^{H} \mathbb{E}_{\mu}\left[\frac{d^{\pi}_t(s_t)^2}{d^{\mu}_t(s_t)^2}\text{Var}_{\mu}[\frac{\pi(a_t \mid s_t)}{\mu(a_t \mid s_t)}\left(V^{\pi}_{t+1}(s_{t+1}) + r_t\right)\middle| s_t]\right]$$

[2] The tightest bound is actually $O(\tau_s \tau_a H^2)$ using Lemma 3.3 which we describe later.

and is asymptotically bigger than the CR lower bound in Theorem 3.1 by an additive term

$$\frac{1}{n}\sum_{t=1}^{H}\mathbb{E}_\mu\left[\frac{d_t^\pi(s_t)^2}{d_t^\mu(s_t)^2}\mathrm{Var}_\mu[\frac{\pi_t(a_t\mid s_t)}{\mu_t(a_t\mid s_t)}Q_t^\pi(s_t,a_t)\mid s_t]\right]$$

due to the decomposition via *Law of total variance* that

$$\begin{aligned}&\mathrm{Var}_\mu\left[\frac{\pi(a_t\mid s_t)}{\mu(a_t\mid s_t)}[V_{t+1}^\pi(s_{t+1})+r_t]\middle|\ s_t\right]\\ =&\mathbb{E}_\mu\left[\frac{\pi(a_t\mid s_t)^2}{\mu(a_t\mid s_t)^2}\mathrm{Var}\left[V_{t+1}^\pi(s_{t+1})+r_t\mid s_t,a_t\right]\middle|\ s_t\right]\\ &+\mathrm{Var}_\mu\left[\frac{\pi(a_t\mid s_t)}{\mu(a_t\mid s_t)}Q_t^\pi(s_t,a_t)\middle|\ s_t\right].\end{aligned}\tag{5}$$

Not only does it miss the CR lower bound by an additive factor, it also has a worse dependence in horizon H. SMIS has an MSE that scales $O(H^3)$, but one can show that the Cramer-Rao lower bound in Theorem 3.1 scales only quadratically in H. This is a non-trivial fact that follows from the following lemma (iterative law of total variance).

LEMMA 3.3. *[[6, 22, 94]]For any policy π and MDP,*

$$\begin{aligned}\mathrm{Var}_\pi\left[\sum_{t=1}^H r_t\right]=&\sum_{t=1}^H\Big(\mathbb{E}_\pi\left[\mathrm{Var}\left[V_{t+1}^\pi(s_{t+1})+r_t\middle|s_t,a_t\right]\right]\\ &+\mathbb{E}_\pi\left[\mathrm{Var}\left[\mathbb{E}[V_{t+1}^\pi(s_{t+1})+r_t|s_t,a_t]\middle|s_t\right]\right]\Big).\end{aligned}$$

Observe that the first term on the RHS is the CR lower bound and the second term is non-negative, thus, by $|r_t|\le 1$ we get that the CR lower bound of $O(\tau_s\tau_a H^2)$.

The gap from the lower bound is rooted in the importance ratios applied for state transition estimations (4) which eventually propagate into the conditional variance terms of MSE. It is an open problem whether the $O(\tau_s\tau_a H^3/n)$ bound of SMIS can be improved in the setting of (exponentially) large action space $\mathcal{A}$. Our conjecture is inthe negative, similar to the contextual bandits results by [86].

When the action space is also finite, [94] proved that an alternative estimator, Tabular MIS, closes the gap.

Statistically optimal OPE — Tabular MIS. To remove importance weights (4), we need to go beyond state transitions and estimate state-action transitions $\widehat{P}_{t+1}(s'|s,a)$ and state-action reward $\widehat{r}_t(s,a)$ via:

$$\begin{aligned}\widehat{P}_{t+1}(s'|s,a)&=\frac{\sum_{i=1}^n\mathbf{1}[(s_{t+1}^{(i)},a_t^{(i)},s_t^{(i)})=(s',s,a)]}{n_{s,a}}\\ \widehat{r}_t(s,a)&=\frac{\sum_{i=1}^n r_t^{(i)}\mathbf{1}[(s_t^{(i)},a_t^{(i)})=(s,a)]}{n_{s,a}},\end{aligned}\tag{6}$$

with $\widehat{P}_{t+1}(s'|s,a)=0$ and $\widehat{r}_t(s,a)=0$ if $n_{s,a}=0$. The corresponding estimation of $\widehat{P}_t^\pi(s'|s)$ and $\widehat{r}_t^\pi(s)$ are defined by averaging $\widehat{P}_t$ and $\widehat{r}_t$ over π and then $\widehat{d}_t^\pi=\widehat{P}_t^\pi\widehat{d}_{t-1}^\pi$. Tabular MIS (TMIS) [94] is then defined via plugging $\widehat{P}_t^\pi$, $\widehat{r}_t^\pi$ and $\widehat{d}_t^\pi$ into (3). It differs from SMIS by leveraging the fact that each state-action pair is visited frequently under the tabular setting.

Estimator	MSE (realizable)	MSE (misspecified)
Import. Sampl. (IS)	$\exp(H)/n$	$\exp(H)/n$
Doubly Robust	$\exp(H)/n$	$\exp(H)/n$
State MIS	$H^3\tau_s\tau_a/n$	$H^3\tau_s\tau_a/n+\mathrm{bias}^2$
TMIS/FQE/Model-Based	$H^2\tau_s\tau_a/n$	$H^2\tau_s\tau_a/n+\mathrm{bias}^2$

TABLE 2
Summary of OPE methods for tabular RL and their squared estimation error.

MIS vs. Model-based estimators. Tabular marginalized importance sampling estimator has dual expressions

$$\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^H\frac{\widehat{d}_t^\pi(s_t^{(i)})}{\widehat{d}_t^\mu(s_t^{(i)})}\widehat{r}_t^\pi(s^{(i)})=\widehat{v}_{\mathrm{TMIS}}^\pi=\sum_{t=1}^H\sum_{s,a}\widehat{d}_t^\pi(s,a)\,\widehat{r}_t(s,a),$$

where the right-hand-side expression reveals TMIS is also a model-based estimator as it estimates model transitions $\widehat{P}_t$ and replaces the model with the estimated model for the evaluation purpose. Consequently, despite their differences in complex settings, marginalized importance sampling and model-based estimators can be unified through TMIS in tabular RL, using standard MLE estimators, similar to traditional statistical estimation problems [18]. More importantly, it is statistically optimal for the tabular OPE problem.

THEOREM 3.4. *Let $\mathcal{D}=\left\{(s_t^{(i)},a_t^{(i)},r_t^{(i)})\right\}_{i\in[n]}^{t\in[H]}$ be obtained by running a behavior policy μ and π is the target policy to evaluate. Under Assumption 1 and other mild regularity conditions, the MSE of tabular marginalized importance sampling satisfies*

$$\begin{aligned}&\mathbb{E}\left[(\widehat{v}_{\mathrm{TMIS}}^\pi-v^\pi)^2\right]\\ =&\frac{1}{n}\sum_{t=1}^H\mathbb{E}_\mu\left[\frac{d_t^\pi(s_t,a_t)^2}{d_t^\mu(s_t,a_t)^2}\mathrm{Var}_\mu[\left(V_{t+1}^\pi(s_{t+1})+r_t\right)\middle|\ s_t]\right]\\ &\cdot[1+O(\sqrt{\frac{\log n}{n}})]+O(\frac{1}{n^2}).\end{aligned}\tag{7}$$

The big O notation hides universal constants.

Asymptotic efficiency and local minimaxity. The error bound implies that $\lim_{n\to\infty}n\cdot\mathbb{E}[(\widehat{v}_{\mathrm{TMIS}}^\pi-v^\pi)^2]$ equals

$$\sum_{t=1}^H\mathbb{E}_\mu\left[\frac{d^\pi(s_t,a_t)^2}{d^\mu(s_t,a_t)^2}\mathrm{Var}\left[V_{t+1}^\pi(s_{t+1})+r_t\middle|s_t,a_t\right]\right].$$

This result exactly matches the CR-lower bound 3.1 and strictly improves *state MIS* estimator, indicating that both the lower bound 3.1 and upper bound 3.4 are tight. Modern estimation theory [82] establishes that CR-lower bound is the asymptotic minimax lower bound for the MSE of *all* estimators in every local neighborhood of the parameter space.[3] Therefore, Tabular marginalized importance sampling is asymptotically efficient,

[3] In classical statistical text, CR-lower bound is often used to lower bound the variance of the class of *unbiased* estimators.

and locally minimax optimal (i.e. optimal for every problem instance separately).

This result provides new insight even for on-policy evaluation. By default, on-policy evaluation is computed by averaging Monte Carlo returns and its MSE is $\mathrm{Var}_\pi\left[\sum_{t=1}^H r_t\right]$. As implied by Lemma 3.3, the surprising observation is that TMIS improves the efficiency even for the on-policy evaluation problem. This means the natural Monte Carlo estimator of the reward in the on-policy evaluation problem is in fact asymptotically inefficient.

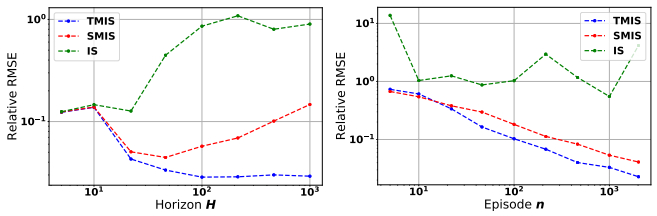

FIG 2. *Adopted from [94]. Different scaling law for TMIS, SMIS and IS for a time-inhomogenuous MDP. Relative RMSE ($\sqrt{MSE}/v^\pi$). For episode n, the right panel shows both TMIS and SMIS have a convergence rate of $n^{-1/2}$. For horizon H, the left panel shows the MSE of TMIS has the optimal dependence $O(H^2)$, while SMIS has the dependence $O(H^3)$.*

Discussion. The idea of MIS estimators goes well beyond tabular settings. MIS can be viewed as a dual form of the Bellman value decomposition. This view motivated researchers [21, 26, 46] to come up with alternative schemes for function approximations in deep RL, e.g., visitation measure $d^\pi(s,a)$ or the importance weights $\rho(s,a)=\frac{d^\pi}{d^\mu}(s,a)$ instead of the value functions. One can also approximates both the Q function and ρ functions, e.g., the double reinforcement learning approach [35, 36] and the DICE family [60, 81, 101]. It remains one of the active research areas in RL theory and algorithm design.

As a technical note, the analysis of SMIS and TMIS involves somewhat delicate calculations that leverage the Bellman recursion in both the estimated $\hat{d}^\pi$ and its covariance matrix. As a comparison — since TMIS is equivalent to the model-based plug-in estimator — we apply the classical "simulation lemma" [37] to it, which implies a bound of

$$|\hat{v}^\pi - v^\pi| \le H^2 \sup_{h,s,a} \|\hat{P}_h(\cdot|s,a) - P_h(\cdot|s,a)\|_1 = \tilde{O}(\sqrt{\frac{H^4 S^2}{n d_m}}).$$

Observe that our more delicate analysis improves the bound to $\sqrt{\frac{H^2\tau_s\tau_a}{n}} \le \sqrt{\frac{H^2}{n d_m}}$.

4. OFFLINE POLICY EVALUATION WITH FUNCTION APPROXIMATION

Next, we switch gears to consider OPE when the (state, action) pairs are described by a continuous feature vector $\phi(s,a)\in\mathbb{R}^d$. This covers most real-life RL problems (such as autonomous driving, robotic arm control, and health care).

The key challenge here is to generalize across unseen states while maintain the statistical optimality at the same time. In the discrete setting, empirical count (maximum likelihood estimate) is a natural algorithm that is optimal, but it cannot be generalized in the function approximation setting. However, to evaluate MDPs, the Bellman equations are universally true regardless of the setting. As a result, one can apply the approximate dynamic programming principles [67] for the given function class and data. This is realized by the following Fitted Q-Evaluation (FQE). For OPE with function approximation, we review the time-homogenuous RL (*i.e.* transition probabilities are identical across time $P_t = P$) and reformulate offline data $\mathcal{D} = \left\{\left(s_h^k, a_h^k, r_h^k\right)\right\}_{h\in[H],k\in[n]} = \{(s_i,a_i,r_i)\}_{i\in[N]}$ throughout the section ($N = nH$).

Fitted Q Evaluation. FQE is a variant of Fitted Q Iteration [4, 16] which is designed for policy optimization purpose. A brief history of FQI is discussed in Section 6. For a given function class $\mathcal{F}$ and data $\mathcal{D}$, FQE recursively estimates $Q_h^\pi, h\in[H]$ via ($\widehat{Q}_{H+1}^\pi = 0$)

$$\widehat{Q}_h^\pi = \operatorname*{argmin}_{f\in\mathcal{F}} \left\{ \frac{1}{N}\sum_{i=1}^N \left(f(s_i,a_i) - y_i\right)^2 + \lambda\rho(f) \right\} \tag{8}$$

with $y_n = r_i + \int_a \widehat{Q}_{h+1}^\pi(s_{i+1},a)\,\pi(a\,|\,s_{i+1})\,\mathrm{d}a$. Here $\rho(f)$ is a proper regularizer and is usually chosen as L_2, i.e. $\rho(f) = \|f\|_2^2$. The OPE estimator is

$$\widehat{v}^\pi = \mathbb{E}_{s\sim d_1, a\sim\pi(\cdot|s)}\left[\widehat{Q}_1^\pi(s,a)\right].$$

The squared loss function resembles the empirical approximation for Bellman question (1).

4.1 Linear function approximation

Linear function approximation considers the class $\mathcal{F}_{\text{lin}} = \{f : f(\cdot,\cdot) = \langle\phi(\cdot,\cdot),\theta\rangle, \theta\in\mathbb{R}^d\}$. Denote the shorthand $\phi_n := \phi(s_n,a_n)$ and $\phi^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[\phi(s,a)]$, then FQE can be computed recursively via $\widehat{Q}_h(s,a) = \phi(s,a)^\top \widehat{w}_h^\pi$ with

$$\widehat{w}_h^\pi = \widehat{R} + \widehat{M}_\pi \widehat{w}_{h+1}^\pi,$$

where $\widehat{M}_\pi = \widehat{\Sigma}^{-1}\sum_{n=1}^N \phi_n\cdot\phi^\pi(s_{n+1})^\top$, $\widehat{\Sigma} = \sum_{n=1}^N \phi_n\phi_n^\top + \lambda I_d$ and $\widehat{R} = \widehat{\Sigma}^{-1}\sum_{n=1}^N r_n\phi_n$. FQE does not learn the model/transition dynamics, and it is generally regraded as a model-free approach. Interestingly, by using the components $\widehat{M}_\pi, \widehat{w}_h^\pi$ to approximate the population counterparts M_π, w_h^π, linear FQE is equivalent to the model-based plug-in estimator [14, 27]. This phenomenon is similar to TMIS, which can be interpreted as a model-based estimator.

When the data coverage of the behavior policy μ spans the state-action space and the linear function class is expressive enough, FQE has the following efficiency guarantee.

THEOREM 4.1. *Suppose assumption 2 (good data coverage) and policy completeness of assumption 3 (linear function class is rich enough) are satisfied, then FQE is a consistent OPE estimator with $\sqrt{N}(\widehat{v}^\pi - v^\pi)$ is asymptotically distributed to normal $\mathcal{N}(0,\sigma^2)$. The asymptotic variance is given by*

$$\sigma^2 = \sum_{h_1,h_2=1}^H \left(\nu_{h_1}^\pi\right)^\top \Sigma^{-1}\Omega_{h_1,h_2}\Sigma^{-1}\nu_{h_2}^\pi,$$

where $\nu_h^\pi = \mathbb{E}^\pi\left[\phi\left(s_h, a_h\right) \mid s_1 \sim d_1\right]$, $\Sigma = \frac{1}{H}\sum_{h=1}^H \Sigma_h$, *and the cross-covariance*

$$\Omega_{h_1,h_2} = \mathbb{E}\Big[\frac{1}{H}\sum_{h'=1}^H \phi\left(s_{h'}, a_{h'}\right)\phi\left(s_{h'}, a_{h'}\right)^\top \varepsilon_{h_1,h'}\varepsilon_{h_2,h'}\Big]$$

and $\varepsilon_{h_1,h'} = Q_{h_1}^\pi\left(s_{h'}, a_{h'}\right) - \Big(r_{h'} + V_{h_1+1}^\pi\left(s_{h'+1}\right)\Big)$.

Critically, the above asymptotic variance is optimal for OPE with linear function approximation. In fact, for linear OPE with Assumption 2 and policy completeness, the variance of any unbiased estimator is lower bounded by σ^2 in Theorem 4.1. Such a lower bound is constructed via computing the influence function from semi-parametric statistics [80, 82]. As a special case, the linear optimality is also consistent with the tabular OPE. For the time-inhomogeneous MDPs, the cross-terms vanish, and the variance $\sigma^2 = \sum_{h=1}^H \left(\nu_h^\pi\right)^\top \Sigma^{-1}\Omega_{h,h}\Sigma^{-1}\nu_h^\pi$ coincides with L_{CR} in Theorem 3.1 when features are indicator functions for states and actions.

From offline policy evaluation to offline policy inference. In addition to the point estimator, Efron's bootstrap [57] is utilized in literature for distributional inference. By sampling episodes $\mathcal{D}^* = \{\tau_1^*, \ldots, \tau_n^*\}$ independently and with replacement from $\mathcal{D}$, the bootstrapped FQE is consistent in distribution, meaning

$$\sqrt{N}\left(\widehat{v}_{\text{Bootstrap}}^\pi - \widehat{v}_{\text{FQE}}^\pi\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

This implies the consistency of the moment estimations, and one particular example is for the second order, i.e.

$$\lim_{n\to\infty} \mathrm{Var}[\sqrt{N}\left(\widehat{v}_{\text{Bootstrap}}^\pi - \widehat{v}_{\text{FQE}}^\pi\right)] = \sigma^2.$$

Distribution shift characterization via Minimax-optimal OPE. The finite sample error bound for FQE provides a similar characterization for the hardness of OPE in the non-asymptotic way. By incorporating the Chi-square divergence $\chi^2_{\mathcal{F}_{\text{lin}}}(p,q) := \sup_{f\in\mathcal{F}_{\text{lin}}} \frac{\mathbb{E}_p[f(x)]^2}{\mathbb{E}_q[f(x)^2]} - 1$, there is a simplified finite error bound [14]:

$$|\widehat{v}^\pi - v^\pi| \lesssim H^2\sqrt{\frac{1+\chi^2_{\mathcal{F}_{\text{lin}}}(\pi,\mu)}{N}} + O\left(N^{-1}\right).$$

This shows the distribution divergence in an explicit way.

4.2 Parametric function approximation

Parametric models extends the linear representation $\langle\phi,\theta\rangle$ to the functional form $f(\theta,\phi)$, allowing for nonlinear or non-convex structures. Fitted Q-Evaluation over this generic class is less tractable since the regression objective (8) no longer yield a closed-form solution, and the optimal solution can only be characterized by the optimality condition through the lens of an *Z-estimator* [38]

$$\nabla_\theta\left\{\frac{1}{2N}\sum_{i=1}^N\left[f\left(\widehat{\theta}_h, \phi_i\right) - y_i\left(\widehat{\theta}_{h+1}\right)\right]^2 + \lambda\rho(\widehat{\boldsymbol{\theta}})\right\} = 0,$$

where $\phi_i = \phi(s_i, a_i)$, $y_j(\theta) := \mathbb{E}_{a'\sim\pi(\cdot|s_{j+1})}[f\left(\theta, \phi\left(s_{j+1}, a'\right)\right)]$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_H)$. Yet, FQE is still asymptotically efficient.

THEOREM 4.2. *Under Assumption 3 and mild conditions, when the number of episodes* $n \to \infty$ *and* $\lambda = o(n^{-1/2})$, *we have convergence in distribution* ($N = nH$) $\sqrt{N}\left(\widehat{v}_\pi - v_\pi\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$. *The asymptotic variance* σ^2 *is*

$$\sigma^2 = \sum_{h_1,h_2=1}^H [\nu_{h_1}^\pi]^T \Sigma_{h_1}^{-1}\Omega_{h_1,h_2}\Sigma_{h_2}^{-1}\nu_{h_2}^\pi.$$

Here $\Sigma_h = \mathbb{E}\left[\frac{1}{H}\sum_{j=1}^H \left(\nabla_{\theta_h} f\left(\theta_h^*, \phi_j\right)\right)^\top \left(\nabla_{\theta_h} f\left(\theta_h^*, \phi_j\right)\right)\right]$, $\nu_h^{\pi\top} = \mathbb{E}^\pi\left[\nabla_{\theta_h} f\left(\theta_h^*, \phi\left(s_h, a_h\right)\right)\right]$, *the cross covariance is*

$$\Omega_{i,j} = \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^H \left(\nabla_{\theta_i}^\top f\left(\theta_i^*, \phi_h\right)\right)\left(\nabla_{\theta_j} f\left(\theta_j^*, \phi_h\right)\right)\varepsilon_{i,h}\varepsilon_{j,h}\right]$$

with $\varepsilon_{j,h} = f(\theta_j^*, \phi_h) - r_h - \mathbb{E}^\pi[f(\theta_{j+1}^*, \phi_{h+1}) \mid s_{h+1}]$.

The parametric FQE strictly subsumes the linear FQE as a special case, and this can be seen by noticing $\nabla_{\theta_j} f(\theta_j^*, \phi_h) = \phi_h$ in the linear case. Besides, there is a matching Cramer Rao lower bound, showing that the asymptotic optimality is achieved [100].

On the analysis for OPE with function approximations. For both linear and parametric cases, the OPE error can be decomposed into two parts $v^\pi - \widehat{v}^\pi = E_1 + E_2$, and $E_1 = \frac{1}{N}\sum_{i=1}^N$ is the first order term with the form

$$e_i := \sum_{h=1}^H \left(\nu_h^\pi\right)^\top \Sigma^{-1}\zeta_i\left(Q_h^\pi\left(s_i, a_i\right) - \left(r_i' + V_{h+1}^\pi(s_i')\right)\right)$$

and E_2 is the higher order term. Depending on the setting, ζ_i is either ϕ_i or $\nabla_\theta f(\theta_h^*, \phi_i)$. Higher order terms are generally handled by the data coverage conditions, and the asymptotic normality can be proved by Martingale CLT [54] or Z-Estimator Master Theorem from empirical process theory [38].

5. OFFLINE POLICY LEARNING IN TABULAR RL: PESSIMISM AND INSTANCE-DEPENDENT BOUNDS

Policy learning differs from the policy evaluation in that it needs to optimize over a set of policies rather than just evaluating a given policy. Consider the case where there are finite number of policies $\pi_1, \pi_2, \ldots, \pi_K$, and the estimates for the respective policies are $\widehat{v}^{\pi_1}, \ldots, \widehat{v}^{\pi_K}$. If that is all the information provided, a natural algorithm for policy learning would be the ERM (Empirical Risk Minimizer)

$$\widehat{\pi}^* = \mathrm{argmax}_\pi \widehat{v}^\pi. \tag{9}$$

However, simply selecting policy via point estimators might not provide the best approach due to uncertainty, and the error in the estimators might cause incorrect prediction about the order of policies. In particular, the dataset could be biased toward certain states, contain many suboptimal actions, or even contain little information about the optimal policy, which poses significant challenges when trying to generalize beyond the observed data. If an agent is too optimistic in regions where it has little or no data, it may overestimate the value of actions in these regions, leading to poor policy performance.

The generic recipe for offline decision-making (not just for RL) is the so-called *pessimism in the face of uncertainty*, namely, to stay conservative

One should be biased towards the more conservative side when the value of a decision is uncertain.

Pessimism is particularly important in real-world offline applications, where safety and reliability are crucial. For example, in healthcare, an overly optimistic RL policy might recommend treatments that appear effective in limited data but are unproven or unsafe. Pessimism ensures that decisions are made conservatively, focusing on treatments with scientific evidence. Besides, offline RL can be used to train self-driving systems using logged data. A pessimistic approach ensures that the vehicle avoids risky maneuvers that haven't been sufficiently tested in the training data.

Consider the multi-arm bandit (MAB) problem as a simple example, where there are K decision arms. The goal is to identify the arm with the highest mean reward. Each arm has reward estimate and uncertainty, then the principle of pessimism will choose

$$\text{(10)} \quad \text{argmax}_{k\in[K]}\{\text{Reward\_Estimate}_k - \alpha \cdot \text{Uncertainty}_k\}$$

for some penalty parameter $\alpha > 0$. The above quantity

$$\text{Reward\_Estimate}_k - \alpha \cdot \text{Uncertainty}_k$$

is often termed as the *lower confidence bound*, which discourages the effect of uncertainty when no exploration is allowed (i.e. the offline case). In contrast, online setting optimizes the *upper confidence bound* $\text{Reward\_Estimate}_k + \alpha \cdot \text{Uncertainty}_k$ to encourage exploring region with high uncertainty (see Figure below).

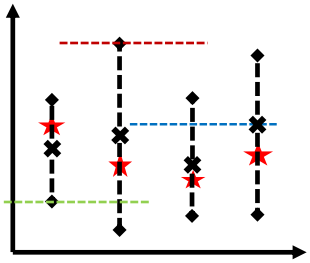

FIG 3. *An instance of MAB problem with pessimism being the right choice. Red choice: upper confidence bound. Blue choice: empirical risk minimizer. Green choice: lower confidence bound* (10)*. Red star denotes the true mean reward; black cross denotes the point estimator* (9)*.*

5.1 Pessimism is Minimax Optimal

An offline policy learning algorithm targets at identifying a policy π_{alg} such that its value differences with respect to the optimal policy π^* is small. The performance is theoretically characterized by the *probably approximately correct* (PAC) bound,

$$v^* - v^{\pi_{\text{alg}}} \le \text{Poly}(H, S, A, \frac{1}{\sqrt{n}}, C_\mu) \text{ with high probability,}$$

meaning the performance gap is polynomial in the planning horizon, number of states and actions, $1/\sqrt{n}$, and certain data coverage parameter C_μ. As the number of the episodes n goes to infinity, $v^* - v^{\pi_{\text{alg}}} \to 0$.

In addition, for the given n episodic data, the minimax risk (suboptimality gap) [41, 83]

$$\mathcal{R}_n := \inf_{\pi_{\text{alg}}} \sup_{\text{MDP } \mathcal{M}} \mathbb{E}_{\mathcal{M}}\left[v^* - v^{\pi_{\text{alg}}}\right]$$

measures the best possible performance (information-theoretical limit) for a class of MDP problems $\mathcal{M}$ in the worst-case-scenario sense. If for certain algorithm $\mathcal{A}$, the PAC bound of its suboptimality gap $v^* - v^{\pi_{\text{alg}}}$ matches $\mathcal{R}_n$, then we call algorithm $\mathcal{A}$ minimax optimal. The key feature for minimax lower bound is that the supremum is taken over the whole MDP class, making it instance independent. This is to say, the worst case optimality are optimal "globally".

- For the Uniform data coverage $d_m > 0$ (Assumption 4), the minimax optimal bound has rate $\Theta(\sqrt{\frac{H^3}{n \cdot d_m}})$, and it is attained by choosing the ERM estimator [74, 96].
- For the single policy coverage $C^* = \left\| \frac{d_h^{\pi^*}}{d_h^{\mu}} \right\|_\infty$ (Assumption 6), the minimax optimal bound has rate $\Theta(\sqrt{\frac{H^3 SC^*}{n}})$ by using the reference advantage techniques [73, 92].

The (near-)optimal worst-case performance bounds that depend on their data-coverage coefficients are valuable as they do not depend on the structure of the particular problem, therefore, remain valid even for pathological MDPs. However, the global optimal characterizations are unable to depict what types of decision processes and what kinds of behavior policies are inherently easier or more challenging for offline RL. In particular, the empirical performances of real applications are often far better than what those non-adaptive / problem-independent bounds would indicate. For example, city driving vs. highway driving in autonomous driving. In both cases, the state-action space could be defined by the position, velocity, orientation of the vehicle, but city driving is more complex due to a highly dynamic and unpredictable environment whereas highway driving is simpler in comparison because the environment is more structured.

Alternatively, rather than obtaining the PAC bound that depends on the global parameters H, S, A, we can delve into the instance level and consider the instance-dependent characterization via transition kernel P, reward r, and behavior policy μ. In general, instance dependent bounds should have the following properties:

- It adapts to the individual instances and only require minimal assumptions so they can be widely applied in most cases.
- It should characterize the system structures of the specific problems, hold even for peculiar instances that do not satisfy the standard data-coverage assumptions.
- It should recover the worst-case guarantees when the data-coverage assumptions are satisfied.

For the rest of the section, we review how these guidelines are accomplished for the policy learning tasks.

5.2 Towards instance optimality via Pessimism

In this section, we review instance-dependent offline RL in the tabular setting.

Pessimistic Value Iteration. To approximate the optimal Q-function in (1), one can perform the update

$$\widehat{Q}_h(\cdot,\cdot) \leftarrow (\widehat{\mathcal{P}}_h \widehat{Q}_{h+1})(\cdot,\cdot), \tag{11}$$

where $\widehat{\mathcal{P}}_h$ denotes the approximation for the Bellman operator in (2) and $\widehat{V}_h(s) := \max_a \widehat{Q}_h(s,a)$. For a uncertainty quantification $\Gamma_h(\cdot,\cdot)$, the principle of pessimism applies

$$\widehat{Q}_h(\cdot,\cdot) \leftarrow \widehat{Q}_h(\cdot,\cdot) - \Gamma_h(\cdot,\cdot). \tag{12}$$

This is the analogy to the lower confidence bound in the bandit setting (10). The learned policy and value function by pessimistic value iteration are defined as ($\forall h \in [H]$):

$$\pi_h^{\text{PVI}}(\cdot|s_h) \leftarrow \text{argmax}_{\pi_h} \langle \widehat{Q}_h(s_h,\cdot), \pi_h(\cdot|s_h)\rangle.$$
$$\widehat{V}_h(s_h) \leftarrow \max_{\pi_h} \langle \widehat{Q}_h(s_h,\cdot), \pi_h(\cdot|s_h)\rangle.$$

Model-based Estimators. Recall the approximated Bellman operator $\widehat{\mathcal{P}}_h$ in (11) is defined via the plug-in estimators:[4]

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

If $n_{s,a} = 0$, $\widehat{P}_h(s'|s,a) = 1/S, \widehat{r}_h(s,a) = 0$.

A Bernstein-style uncertainty. It turns out that the following variance-dependent quantity

$$\Gamma_h(s,a) = \widetilde{O}\left[\sqrt{\frac{\text{Var}_{\widehat{P}_{s,a}}(\widehat{r}_h + \widehat{V}_{h+1})}{n_{s,a}}} + \frac{H}{n_{s,a}}\right]$$

describes the uncertainty for (12) appropriately. The conditional variance $\text{Var}_{\widehat{P}_{s,a}}(\widehat{r}_h + \widehat{V}_{h+1})$ corresponds to the **aleatoric uncertainty** which measures the intrinsic uncertainty of the transition P, given the state-action (s,a). This is the uncertainty due to the natural variability of the system being modeled, which cannot be reduced by collecting more data. The term $1/n_{s,a}$ is the **epistemic uncertainty** that comes from incomplete knowledge and can be reduced by gathering more data [31].

The condition variance in Γ_h creates the Bernstein-style pessimism. Compared to the Hoeffding-style pessimism $\widetilde{O}(H/\sqrt{n_{s,a}})$ which is overly pessimistic (due to $\sqrt{\text{Var}_{\widehat{P}}(\widehat{r}_h + \widehat{V}_{h+1})} \le H$), Γ_h is more data-adaptive. Furthermore, for the fully deterministic environments where the transitions and rewards are deterministic, the conditional variances vanishes and the uncertainty has a faster scale $1/n_{s,a}$.

THEOREM 5.1. *Under the Assumption 6, 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* $0 < \delta < 1$, *such that when* $n > 1/\bar{d}_m \cdot \log(HSA/\delta)$, *with probability* $1-\delta$, *the output of Pessimistic Value Iteration satisfies*

$$\begin{aligned} 0 &\le v^* - v^{\pi^{PVI}} \\ &\lesssim \sum_{h=1}^H \sum_{(s,a)\in\mathcal{C}_h} d_h^{\pi^*}(s,a) \cdot \sqrt{\frac{\text{Var}_{P_{s,a}}(r_h + V_{h+1}^*)}{n \cdot d_h^\mu(s,a)}} \\ &\quad + \widetilde{O}\left(\frac{H^3}{n \cdot \bar{d}_m}\right). \end{aligned} \tag{13}$$

Unlike the worst-case bounds that rely on the data-coverage parameters, the instance bound requires the minimal assumption 6. The key distinction is the main term in (13) is expressed by the system quantities that admits no explicit dependence on H, S, A. It depicts the interrelations within the problem when the problem instance is a tuple $(\mathcal{M}, \pi^*, \mu)$: an MDP $\mathcal{M}$ (coupled with the optimal policy π^*) with the data rolling from an offline behavior policy μ, so it helps understand what type of problems are harder / easier than others in a *quantitative* way.

The complexity of the main term in (13) can be decomposed into $\frac{d_h^{\pi^*}(s,a)}{\sqrt{d_h^\mu(s,a)}} \cdot \sqrt{\text{Var}_{P_{s,a}}(r_h + V_{h+1}^*)}$, and this reveals the learning hardness of offline RL stems from two aspects.

- **Environmental variation**[5] $\sqrt{\text{Var}_{P_{s,a}}(r_h + V_{h+1}^*)}$ is jointly determined by the stochasticity of transition, reward, and optimal value function. A problem with lower environmental variation is easier than a problem with higher environmental variation. This theoretical characterization explains the intuition that stochastic environments are generally harder than the deterministic environments.
- **Distribution mismatch** $\frac{d_h^{\pi^*}(s,a)}{\sqrt{d_h^\mu(s,a)}}$ is the other factor that affects the learning hardness. When the behavior policy μ deviates far from the optimal policy π^*, the problem is intrinsically harder since the mismatch ratio becomes large. When $\pi^* = \mu$, the mismatch ratio is bounded by 1 for all states and actions.

Due to its fine-grained expression, (13) is named *Intrinsic offline learning bound* by recent literature [95]. It also subsumes the existing worst-case bounds that are optimal.

For uniform data-coverage 4, the optimal suboptimality is $\Theta(\sqrt{\frac{H^3}{nd_m}})$. The intrinsic RL bound can be upper bounded by this rate via *cauchy inequality* and Lemma 3.3:[6]

$$\begin{aligned} v^* - v^{\pi^{\text{PVI}}} &\lesssim \sum_{h=1}^H \langle d_h^{\pi^*}(\cdot), \sqrt{\frac{\text{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^*)}{n \cdot d_h^\mu(\cdot)}}\rangle \\ &= \sum_{h=1}^H \langle \sqrt{d_h^{\pi^*}(\cdot)}, \sqrt{\frac{d_h^{\pi^*}(\cdot) \odot \text{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^*)}{n \cdot d_m}}\rangle \\ &\le \sum_{h=1}^H \left\| \sqrt{d_h^{\pi^*}(\cdot)} \right\|_2 \left\| \sqrt{\frac{d_h^{\pi^*}(\cdot) \odot \text{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^*)}{n \cdot d_m}} \right\|_2 \end{aligned}$$

[4] $n_{s,a,h} := \sum_{\tau=1}^n \mathbf{1}[s_h^\tau, a_h^\tau = s,a]$ be the total counts that visit (s,a) pair at time h.

[5] [50] named a similar quantity environmental norm.

[6] Here $\odot$ denotes element-wise multiplication.

$$\leq\sqrt{\frac{H\cdot \mathrm{Var}_{\pi^*}(\sum_{h=1}^H r_h)}{n\cdot d_m}} \leq \sqrt{\frac{H^3}{n\cdot d_m}}$$

which recovers the optimal rate.

For the single policy coverage 6 with $C^* := \left\|\frac{d_h^{\pi^*}}{d_h^\mu}\right\|_\infty < \infty$, a similar computation using Lemma 3.3 can recover the optimal rate $\Theta(\sqrt{\frac{H^3SC^*}{n}})$ via

$$v^* - v^{\pi^{\mathrm{PVI}}} \lesssim \sum_{h=1}^H \langle d_h^{\pi^*}(\cdot), \sqrt{\frac{\mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^*)}{n\cdot d_h^\mu(\cdot)}}\rangle$$

$$\leq\sqrt{\frac{C^*}{n}} \sum_{h=1}^H \langle \sqrt{d_h^{\pi^*}(\cdot)}, \sqrt{\mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^*)}\rangle$$

$$\leq\sqrt{\frac{SC^*}{n}} \sum_{h=1}^H \sqrt{\sum_{s\in\mathcal{S}} d_h^{\pi^*}(s,\pi_h^*(s))\cdot \mathrm{Var}_{P_{s,\pi_h^*(s)}}(r_h + V_{h+1}^*)}$$

$$\leq\sqrt{\frac{SC^*}{n}}\sqrt{H}\cdot \sqrt{\mathrm{Var}_\pi\left[\sum_{t=1}^H r_t\right]} \leq \sqrt{\frac{H^3SC^*}{n}}.$$

Problem dependent domain. Similar to the online RL [99], if we denote $\mathbb{Q}_h^* = \max_{s_h,a_h} \mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^*)$ 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 5.1 implies:

$$v^* - v^{\pi^{\mathrm{PVI}}} \leq \min\left\{\widetilde{O}(\sum_{h=1}^H \sqrt{\frac{\mathbb{Q}_h^*}{n\bar{d}_m}}), \widetilde{O}(\sqrt{\frac{H\mathcal{B}^2}{n\bar{d}_m}})\right\} + \widetilde{O}(\frac{H^3}{n\bar{d}_m}).$$

For the problem instances with either small $\mathcal{B}$ or small $\mathbb{Q}_h^*$, the intrinsic bound yields much better performances, as discussed in the following.

Deterministic systems. For systems equipped with low stochasticity, *e.g.* robotics, or even deterministic dynamics, *e.g.* the game of GO, 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 $\mathbb{Q}_h^* = 0$ for all h. This enables a faster convergence rate of order $\frac{H^3}{nd_m}$ and significantly improves over the existing worst-case results that have order $\frac{1}{\sqrt{n}}$.

Partially deterministic systems. Sometimes, practical applications can 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 $t\cdot\sqrt{\max Q_h^*/n\bar{d}_m}$ suboptimality gap, which could be much smaller than $H\cdot\sqrt{\max Q_h^*/n\bar{d}_m}$ when $t \ll H$.

Fast mixing domains. Consider a class of highly mixing non-stationary MDPs 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 $\bar{s}_t := \arg\max V_t^*(s)$ and $\underline{s}_t := \arg\min V_t^*(s)$. Also, denote $\mathrm{rng}V_h^*$ to be the range of V_h^*. In such cases, Bellman optimality equations have the form

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

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

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

5.3 Assumption-Free Offline RL

We now review the scenario where the behavior policy can be arbitrary in this section. In this case, μ might not cover any optimal policy π^* (*i.e.* there might be high reward location (s,a) that μ can never visit). This can happen when a mediocre doctor only uses one treatment for certain patient all the time. Statistically, even with the infinite amount of episodic data, algorithms might not learn the optimal policy exactly.

To better characterize the discrepancy, an augmented MDP $\mathcal{M}^\dagger$ is defined with one extra state $s_h^\dagger$ for all $h\in\{2,\ldots,H+1\}$ with the augmented state space $\mathcal{S}^\dagger = \mathcal{S}\cup\{s_h^\dagger\}$. Compared to the original MDP $\mathcal{M}$, the transition and the reward are modified as follows:

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

$$r^\dagger(s_h,a_h) = \begin{cases} r(s_h,a_h), & n_{s_h,a_h} > 0, \\ 0, & s_h = s_h^\dagger \text{ or } n_{s_h,a_h} = 0. \end{cases}$$

here δ_s is the Dirac measure and we denote $V_h^{\dagger\pi}$ and $v^{\dagger\pi}$ to be the values under $\mathcal{M}^\dagger$. In this case, the pessimistic value iteration guarantees with high probability that, for any behavior policy μ,

$$v^* - v^{\pi^{\mathrm{PVI}}} \lesssim \sum_{h=2}^{H+1} d_h^{\dagger\pi^*}(s_h^\dagger)$$

$$+\sum_{h=1}^H \sum_{(s,a)\in\mathcal{C}_h} d_h^{\dagger\pi^*}(s,a)\cdot\sqrt{\frac{\mathrm{Var}_{P_{s,a}^\dagger}(r_h^\dagger + V_{h+1}^{\dagger\pi^*})}{n\cdot d_h^\mu(s,a)}}$$

$$+\widetilde{O}\left(\frac{H^3}{n\bar{d}_m}\right).$$

The off-support gap $\sum_{h=2}^{H+1} d_h^{\dagger\pi^*}(s_h^\dagger)$ satisfies $d_h^{\dagger\pi^*}(s_h^\dagger) = \sum_{t=1}^{h-1}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}\backslash\mathcal{C}_t} d_t^{\dagger\pi^*}(s,a)$ and $\mathcal{C}_h := \{(s,a): d_h^\mu(s,a) > 0\}$. When assumption 4 or 6 is satisfied, this gap vanishes since $\mathcal{S}\times\mathcal{A}\backslash\mathcal{C}_h = \emptyset$, and the assumption-free generalization reduces to (13).

Beyond the tabular setting, assumption-free RL is also considered in the function approximation setting. [48] uses ϵ_ζ, the probability under a policy of escaping to state-actions with insufficient data during an episode, to measure the state-action region that is agnostic to the behavior policy, then it incurs

an off-support gap $\frac{V_{\max}\epsilon_\zeta}{1-\gamma}$.[7] The other study relies the condition *Compliance of Dataset* which only requires the data tuples (s_i, a_i, r_i, s_i') to follow the same MDP transition P that might not cover any good policy, and the data agnostic region is handled by regularization to avoid singularity [34]. For general function approximation, the off-support gap is characterized by Theorem 3.1 of [91].

6. OFFLINE POLICY LEARNING WITH FUNCTION APPROXIMATIONS

Fitted Q-Iteration (FQI) [16], which is initially named as *fitted value iteration* (FVI) [24], makes it possible to take full advantage of any regression algorithm for achieving generalization for reinforcement learning. In particular, it is widely adopted for offline RL when only historical data are provided [4, 58]. In the previous section 4, we have seen that its variants Fitted Q-Evaluation are the statistically optimal estimator for the *offline policy evaluation* task. For policy learning with function approximation, we review FQI/FVI as it still yields strong instance-dependent guarantees.

Pessimism remains effective for function approximation. The general prototype of pessimism combined with FVI in (11), (12) remains valid for any MDPs. If the point-wise condition [34]

$$|(\widehat{\mathcal{P}}_h \widehat{Q}_{h+1})(\cdot,\cdot) - (\mathcal{P}_h \widehat{Q}_{h+1})(\cdot,\cdot)| \le \Gamma(\cdot,\cdot)$$

holds true, then the suboptimality gap can be bounded by

$$v^* - v^{\pi^{\text{PFVI}}} \le 2\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim\pi^*}\left[\Gamma_h(s_h,a_h)\right].$$

6.1 OPL with Linear Function Approximation

When Linear MDP models (c.f. section 2.2) are instantiated, the FVI solves

(14)
$$\widehat{w}_h = \text{argmax}_{w\in\mathbb{R}^d}\left\{\sum_{\tau=1}^{n}\left(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau) - \phi(x_h^\tau, a_h^\tau)^\top w\right)^2 + \lambda\|w\|_2^2\right\},$$

and $\widehat{\mathcal{P}}_h\widehat{Q}_{h+1}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{w}_h$ has a closed-form solution. The pessimism $\Gamma_h(s,a) = dH\sqrt{\phi(s,a)^\top \Lambda_h^{-1}\phi(s,a)}$, and $\phi(s,a)^\top\Lambda_h^{-1}\phi(s,a)$ represents the effective number of samples observed in offline data along the ϕ direction, and thus represents the uncertainty along the ϕ direction. Here $\Lambda_h = \sum_{\tau=1}^n \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda\cdot I$ is the Gram matrix. The resulting bound scales as

(15)
$$v^* - v^{\pi^{\text{PFVI}}} \lesssim dH\sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim\pi^*}\left[\sqrt{\phi(s_h,a_h)^\top\Lambda_h^{-1}\phi(s_h,a_h)}\right].$$

Is FQI/FVI itself sufficient for optimality? When reducing to the tabular MDPs with $\phi(s,a) = \mathbf{1}_{s,a}$, PFVI has the form $\widetilde{O}(dH\cdot\sum_{h,s,a} d_h^{\pi^*}(s,a)\sqrt{\frac{1}{n\cdot d_h^\mu(s,a)}})$, and this deviates from Theorem 5.1 $\widetilde{O}(\sum_{h,s,a} d_h^{\pi^*}(s,a)\sqrt{\frac{\text{Var}_{P_{s,a}}(r+V_{h+1}^*)}{n\cdot d_h^\mu(s,a)}})$ by a factor of $H^{1/2}$. By direct comparison, it can be seen that PFVI cannot get rid of the explicit H factor due to missing the variance information (*w.r.t* V^*).

Intuitively, it might not be ideal to put equal weights on all the training samples in the FQI/FVI objectives, as different data pieces carry different "amount" of information. The term $\text{Var}_{P_{s,a}}(r + V_{h+1}^*)$ happens to measure the aleatoric uncertainty at location (s,a). If $\text{Var}_{P_{s_1,a_1}}(r + V_{h+1}^*) \ll \text{Var}_{P_{s_2,a_2}}(r + V_{h+1}^*)$, then the information contained in sample piece (s_1, a_1, s_1', r_1) is more certain than the sample (s_2, a_2, s_2', r_2). To address this, existing literature deployed variance reweighting [55, 93, 97] for FQI/FVI.

Variance-weighted FVI. Instead of regressing via (14), Variance-weighted FVI reweights each sample via an estimated conditional variance $\widehat{\sigma}^2$

$$\widehat{w}_h := \text{argmin}_{w\in\mathbb{R}^d}\sum_{k=1}^{n}\frac{\left[\langle\phi(s_h^k,a_h^k),w\rangle - r_h^k - \widehat{V}_{h+1}(s_{h+1}'^k)\right]^2}{\widehat{\sigma}_h^2(s_h^k,a_h^k)} + \lambda\|w\|_2^2$$

where $\widehat{\sigma}^2$ approximates $\text{Var}_{P_{s,a}}(r + V_{h+1}^*)$ and can be computed by estimating the first and second order moments separately. The pessimism in this case is modified as:

$$\Gamma_h \approx O\left(\sqrt{d}\cdot(\phi(\cdot,\cdot)^\top\widehat{\Lambda}_h^{-1}\phi(\cdot,\cdot))^{1/2}\right) + \frac{H^4\sqrt{d}}{n}$$

with $\widehat{\Lambda}_h = \sum_{k=1}^n \phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^\top/\widehat{\sigma}_h^2(s_h^k,a_h^k) + \lambda I_d$ being the reweighted Gram matrix. With the update $Q_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top\widehat{w}_h - \Gamma_h(\cdot,\cdot)$, we have the following.

THEOREM 6.1. *For linear MDPs, under assumption 8 and some mild conditions, with high probability, for all policy π simultaneously, $v^* - v^{\pi^{\text{Vw-PFVI}}}$ is bounded by*

$$\widetilde{O}(\sqrt{d}\sum_{h=1}^{H}\mathbb{E}_\pi\left[\sqrt{\phi(\cdot,\cdot)^\top\Lambda_h^{-1}\phi(\cdot,\cdot)}\right]) + \frac{2H^4\sqrt{d}}{n},$$

where $\Lambda_h = \sum_{k=1}^K \frac{\phi(s_h^k,a_h^k)\cdot\phi(s_h^k,a_h^k)^\top}{\sigma^2_{\widehat{V}_{h+1}(s_h^k,a_h^k)}} + \lambda I_d$. Moreover, $v^* - v^{\pi^{\text{Vw-PFVI}}}$ is also bounded by

(16)
$$\widetilde{O}\left(\sqrt{d}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\sqrt{\phi(\cdot,\cdot)^\top\Lambda_h^{*-1}\phi(\cdot,\cdot)}\right]\right) + \frac{2H^4\sqrt{d}}{n},$$

where $\Lambda_h^* = \sum_{k=1}^K \frac{\phi(s_h^k,a_h^k)\cdot\phi(s_h^k,a_h^k)^\top}{\sigma^2_{V_{h+1}^*(s_h^k,a_h^k)}} + \lambda I_d$ and $\widetilde{O}$ hides universal constants and the Polylog terms.

Theorem 6.1 extends the instance-dependent characterization for offline RL in 5.2 to the linear case. Compared to FVI (15), the main term in Theorem 6.1 replaces the explicit dependence on H with a more adaptive/instance-dependent characterization. For instance, if we ignore the technical treatment by taking

[7] In the discounted setting, $1/(1-\gamma)$, the effective horizon, is similar to H in the finite horizon setting.

$\lambda = 0$ and $\sigma_h^* \approx \mathrm{Var}_P(V_{h+1}^*)$, then for the partially deterministic systems (where there are t stochastic P_h's and $H - t$ deterministic P_h's), the main term diminishes to

$$\sqrt{d}\sum_{i=1}^{t}\mathbb{E}_{\pi^*}\left[\sqrt{\phi(\cdot,\cdot)^\top \Lambda_{h_i}^{*-1}\phi(\cdot,\cdot)}\right]$$

with $h_i \in \{h : \ s.t. \ P_h \text{ is stochastic}\}$ and can be a much smaller quantity when $t \ll H$. Furthermore, for the fully deterministic system, 6.1 automatically provides faster convergence rate $O(\frac{1}{n})$, given that the main term degenerates to 0.

6.2 OPL with Parametric Function Approximation

The parametric function class $\mathcal{F} := \{f(\theta, \phi(\cdot,\cdot)) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \theta \in \Theta\}$ provides the flexibility of selecting model f, making it possible for handling a variety of tasks. For instance, when f is instantiated to be neural networks, θ corresponds to the weights of each network layers and $\phi(\cdot,\cdot)$ corresponds to the state-action representations (which is induced by the network architecture). When facing with easier tasks, we can deploy simpler model f such as polynomials or even linear function $f(\theta, \phi) = \langle \theta, \phi \rangle$.

Similar to FQE for the parametric function approximation, FQI perform the update with pessimism ($\phi_{h,k} = \phi(s_h^k, a_h^k)$)

$$\widehat{\theta}_h \leftarrow \operatorname*{argmin}_{\theta\in\Theta} \sum_{k=1}^{n} \frac{\left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right]^2}{\widehat{\sigma}_h^2(s_h^k, a_h^k)} + \lambda \|\theta\|_2^2$$

$$\Gamma_h(\cdot,\cdot) \leftarrow \widetilde{O}\left(d\sqrt{\nabla_\theta f(\widehat{\theta}_h, \phi(\cdot,\cdot))^\top \Lambda_h^{-1} \nabla_\theta f(\widehat{\theta}_h, \phi(\cdot,\cdot))} + \frac{1}{K}\right),$$

where $\widehat{\sigma}_h^2$ approximates $\sigma_h^2(s,a) := \mathrm{Var}_{P(\cdot|s,a)}(r + V_{h+1}^*)$, and the reweighted Gram matrix has the form

$$\Lambda_h \leftarrow \sum_{k=1}^{n} \nabla f(\widehat{\theta}_h, \phi_{h,k}) \nabla f(\widehat{\theta}_h, \phi_{h,k})^\top / \widehat{\sigma}^2(s_h^k, a_h^k) + \lambda \cdot I.$$

Compared to Linear function approximation, the feature representation ϕ is replaced with $\nabla f(\widehat{\theta}, \phi)$, and regression objective admits no closed-form solution. This design generalizes the results in linear function approximation as follows:

THEOREM 6.2 ([98]). *Suppose Assumption 7,9 and other mild conditions, with probability $1 - \delta$, for all policy π simultaneously, it holds ($\phi_h = \phi(s_h, a_h)$)*

$$v^\pi - v^{\widehat{\pi}} \lesssim d\sum_{h=1}^{H} \mathbb{E}_\pi \left[\left\|\nabla_\theta f(\widehat{\theta}_h, \phi_h)\right\|_{\Lambda_h^{-1}}\right] + \frac{1}{n}.$$

In particular, it has

$$v^* - v^{\widehat{\pi}} \lesssim d\sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\left\|\nabla_\theta^\top f(\theta_h^*, \phi_h)\right\|_{\Lambda_h^{*-1}}\right] + \frac{1}{n}.$$

Here $\Lambda_h^* = \sum_{k=1}^{K} \frac{\nabla_\theta f(\theta_h^*, \phi_{h,k}) \nabla_\theta^\top f(\theta_h^*, \phi_{h,k})}{\sigma_h^*(s_h^k, a_h^k)^2} + \lambda I_d$ *and the* $\sigma_h^*(\cdot,\cdot)^2 := \max\{1, \mathrm{Var}_{P_h} V_{h+1}^*(\cdot,\cdot)\}$.

From a technical perspective, the key tool for finite-sample analysis in function approximation is the *Self-Normalized Concentration for Vector-Valued Martingales* [1], originally developed for analyzing stochastic linear bandits. This tool provides a Hoeffding-style concentration bound that does not rely on variance or second-order information. Recently, Zhou et al. [105] extended this by proving a Bernstein version of Self-Normalized Concentration for linear mixture MDPs. This approach is well-suited for analyzing variance reweighting mechanisms in offline RL, applicable to both linear MDPs [93, 97] and parametric models [98] as discussed above.

6.3 Pessimism in the wild

Beyond the theoretical focus, the aim of pessimism is to explicitly account for uncertainty in state-action value estimation and "penalize" actions in areas of high uncertainty. This approach generally involves modifying the value function (or policy optimization procedure) to discourage actions that have high uncertainty. There are several ways to implement pessimism:

Lower Confidence Bound (LCB). Instead of using the point estimate of the value function, the agent computes a lower bound based on the confidence interval around the estimate. If the agent is uncertain about the true value $Q(s,a)$, it will use a pessimistic estimate such as:

$$Q_{\mathrm{LCB}}(s,a) = \hat{Q}(s,a) - \lambda \cdot U(s,a)$$

with λ controlling the level of pessimism. This encourages the agent to favor actions with more reliable estimates, avoiding overestimated, risky actions, and is celebrated by theoretical research [13, 64, 73, 85, 92] and other research mentioned in the previous sections.

Penalty on Critic. Another approach is to add a divergence penalty term to the critic objective to make conservative value estimates [61]. For instance, [39] uses Fisher divergence with respect to the Boltzmann policy and behavior policy, and *conservative Q-learning* [40, 49] uses Kullback–Leibler divergence for the Boltzmann policy and the behavior policy.

Policy Regularization. Regularization is often used to prevent the learned policy from deviating too much from the behavior policy. This can be viewed as a pessimistic strategy because the learned policy is constrained to stay close to what has been observed, reducing the risk of taking untested actions. For instance, BRAC [87] adds a regularization term $\mathbb{E}_{s\sim\mathcal{D}}\left[D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid s) \| \pi_b(\cdot \mid s)\right)\right]$ to prevent the learned policy from deviating too much from the behavior policy, thus ensuring pessimistic behavior in uncertain areas.

Optimism vs Pessimism? While, under the offline setting, the pessimistic algorithm is consistent with rational decision-making using preferences that satisfy uncertainty aversion [23], it remains intriguing whether pessimism is uniformly better than optimism in the instance-dependent scenarios. For multi-armed bandit problems, [88] demonstrated that greedy, optimistic, and pessimistic approaches are all (globally) minimax optimal for offline optimization, with each potentially outperforming the others in specific instances. For example, in case 1, where batch data frequently includes good arms, pessimism performs better. Conversely, in case 2, if the behavior policy pulls good arms infrequently, optimism may be advantageous due to the higher uncertainty associated with good arms. This suggests that existing instance-dependent offline RL studies primarily address case 1 (echo Assumption 6), leaving open the question of whether section 5.3 could be further enhanced by incorporating an optimistic perspective, as in case 2.

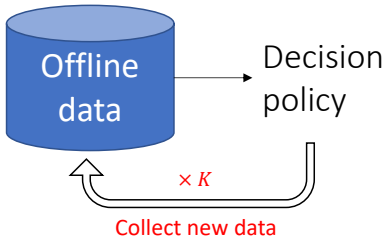

FIG 4. *Illustration of the problem of low-adaptive RL.*

7. LOW-ADAPTIVE EXPLORATION IN RL

So far, we have focused on offline RL which aims at doing the best one can with the given data in learning a new policy. The resulting algorithm is based on the "pessimism" principle that discourages exploration.

If an offline RL agent ends up finding a near-optimal policy, that is because we are lucky to have observed data that covered all states/actions that that optimal policy has taken. Alternatively, we can change the goal-post (in "assumption-free" offline RL) by declaring that we will only learn that part that is "observable" based on the offline data. Statistical lower bounds indicate that both these results cannot be substantially improved.

This conclusion is quite pessimistic indeed in that it does not take into account the common real-life scenario that the learned policy may get deployed and that a new batch of data will eventually be collected, nor how to make use of the new data.

This is a much weaker claim than the online RL, which involves algorithmically ensuring that the exploration policies to have good coverage, hence allowing the learner to identify the optimal policy.

One way to think about this is that offline RL is a problem with no adaptivity allowed, while online RL allows adaptively choosing a new policy after every trajectory. This motivates us to consider the problem *in between* by asking:

Can we learn as well as the best online RL agent while using only a few batches?

One can also think about the problem as a sequence of offline RL problem, but the learner can decide on the exploration policy μ to run for the next batch. All three examples that we considered as motivation of offline RL in the introduction may actually allow some limited exploration. Minimizing the number of batches help to alleviate all of the following issues.

- **Deployment Costs:** Updating policies in distributed systems, such as autonomous vehicles or network routers, can be computationally expensive.
- **Testing and Approval Overheads:** Policies in sensitive domains (e.g., healthcare) require extensive testing, ethical approvals, and regulatory compliance.
- **Concurrency Challenges:** Running experiments in parallel to identify optimal policies is limited by physical and logistical constraints.

Low-adaptive RL addresses these challenges by limiting the number of policy changes (K) during the learning process, where $K \ll T$ (the total number of rounds). This problem is well-studied in multi-armed bandits and linear bandits which shows that no-regret learning with $\tilde{O}(\sqrt{T})$ regret can be achieved with only $O(\log\log T)$ batches of exploration [10, 20, 66], but the same problem on RL is only getting started recently [30, 53, 70, 71].

There are two closely related settings.

RL with low switching cost The learner must limit the number of times the deployed policy changes to K (for historical reason the policy is often confined to deterministic policies).

RL with low batch complexity The learner must schedule K batches of exploration (i.e., experiments) ahead of time and only look at the collected data in the completed batch and decide on the (sequence of) policies to use for the next batch at pre-determined checkpoints.

Both settings could make sense in practice with the second setting being qualitatively stronger[8] as monitoring certain statistics might be much cheaper than deploying new policies.

Regret and sample complexity in online RL. Considering low-switching or low batch complexity in isolation does not make sense, the algorithm must also be able to find a near-optimal policy. To quantify the performance of an RL algorithm it is typical that we consider (cumulative regret) of the sequence of policies being played $\pi^{(t)}$

$$\text{Regret} := \sum_{t=1}^{T} v^{\pi^*} - v^{\pi^{(t)}}$$

or the number of samples T as a function of $\epsilon > 0$ such that we can identify $\hat{\pi}$ that satisfies

$$v^{\pi^*} - v^{\hat{\pi}} \le \epsilon.$$

In the remainder of this section, we survey the existing work on reinforcement learning for both the tabular case and under function approximation.

7.1 Learning Tabular RL in $O(\log\log T)$ batches

Let us first state the known information-theoretic lower bounds in this problem.

THEOREM 7.1. *Consider the tabular RL problems (defined in Section 2). Assume $S < A^{H/2}$* [9].

1. *Any algorithms with a regret of $\tilde{O}(\sqrt{T\text{poly}(H,S,A)})$ must incur a switching cost of $\Omega(HSA\log\log T)$ and use $\Omega(H/\log T + \log\log T)$ batches of exploration.*
2. *Moreover, any algorithms with a regret of $o(T)$ must incur a switching cost of $\Omega(HSA)$ and use $\Omega(H/\log T)$ batches of exploration.*

[8] It is stronger when we allow randomized policies, and not compatible if we restrict to deterministic policies.

[9] This is without loss of generality because otherwise uniform exploration and IS-based OPE with curse-of-horizon suffices to solve the problem with one batch.

The lower bounds of for the batch complexity is due to Theorem B.3 of [30] and Corollary 3 of [20]. The lower bounds of the switching costs are due to [71, Theorem 4.2 and 4.3].

Now let us inspect the algorithmic techniques in this space. First, a doubling schedule of exploration due to the UCB2 algorithm [5] can be combined with optimistic exploration Q-learning to obtain a near-optimal regret while using only $\log(T)$ switching cost[7], but since it requires monitoring the exploration to decide when to change the policy, its batch complexity remains T.

Can this algorithm be improved? Qiao et al. [71] designed a policy elimination-based method called *Adaptive Policy Elimination by Value Estimation* (APEVE) that achieves the following guarantees:

- **Near-Optimal Regret** $\tilde{O}(\sqrt{H^4S^2AT})$ which is optimal up to a factor of HS.
- **Switching Costs** of $O(HSA\log\log T)$ matching the information-theoretic lower bound.
- **Batch complexity** of $O(H\log\log T)$, which matches the information-theoretic lower bound in T [10]

These results highlight that low-adaptive RL can achieve comparable performance to traditional online RL while using only a small number of batches.

APEVE is a *policy elimination* method, which iteratively narrows down the set of candidate policies by eliminating those deemed suboptimal. The method combines:

1. **Crude Layer-Wise Exploration:** A coarse-grained exploration scheme that explores each h, s, a layer by layer. This provides a crude approximation of the useful part of the MDP's transition kernel.
2. **Fine Stagewise Exploration:** Use the crude transition kernel estimate to plan and identify HSA policies that each visits a particular triplets h, s, a most frequently among all policies in the remaining set of policies, then execute these policies to collect more data.
3. **Confidence-Bound Based Elimination:** Use the dataset with good coverage to conduct OPE on all policies that remains to be contenders, then eliminate those policies with their upper confidence bound lower than the highest lower confidence bound.

Let the total number of stages be K, and the k^{th} stage have length $T^{(k)} = K^{1-1/k}$, one can work out that the smallest K such that $\sum_{k=1}^{K} T^{(k)} > T$ is $K = O(\log\log T)$. The total number of stages is only $O(\log\log T)$ and in each stage, it requires deterministically changing policies for HSA times per stage.

Reward-free exploration with $O(H)$-batches. Qiao et al. [71] also presented a reward-free exploration method (LARFE) with a sample complexity of $O(H^5S^2A/\epsilon^2)$ for identifying any policies while using only $2H$ rounds of adaptivity. LARFE does not need to perform the $\log\log T$ stages of exploration since it does not care about regret, so the crude-layerwise exploration can reach a reasonable approximation and the HSA exploration policies can be identified at one shot for driving the error down.

[10] A minor variation of APEVE called APEVE+ achieves **Batch complexity** of $O(H + \log\log T)$ [71].

These results demonstrate that there are algorithms that can achieve nearly the same regret or sample complexity as the best online algorithm even if we only give a very small room for adaptively updating the policies. The result is further improved in [102], who improved the regret bound to the optimal $\tilde{O}(\sqrt{H^3SAT})$ while retaining the same batch complexity.

7.2 Linear function approximation and Reward-Free Exploration in $O(H)$ batches

The natural next question is whether APEVE-like algorithms can be derived for RL under linear function approximation. The lower bounds are in place,

THEOREM 7.2 (Theorem 7.2 and 7.2 of [70]). *Under linear MDPs setting, any algorithm that achieves $\tilde{O}(\sqrt{T\mathrm{poly}(d,H)})$ regret must incur a switching cost of $\Omega(dH\log\log T)$ and a batch complexity of $\Omega(H/\log d + \log\log T)$*

Unfortunately, there are technical challenges and the best low-adaptive learner of linear MDPs for regret minimization still requires $O(\log T)$ batches from the doubling trick [19, 84] using the doubling trick from Abbasi-Yadkori et al. [1].

On the other hand, in the reward-free exploration setting, a policy elimination approach [70] with merely H batches of exploration while achieving a sample-complexity bound of $O(d^2H^5/\epsilon^2)$. This improves over a related result [30] that obtains $O(d^3H^5/\epsilon^2\nu_{\min}^2)$ where $\nu_{\min}$ is an (arbitrarily small) problem-specific reachability parameter. The algorithm of [70] is also more satisfying as it does not need to know $\nu_{\min}$ and the result does not deteriorate as $\nu_{\min}$ gets smaller.

The key algorithmic ideas are closely related to the reward-free exploration algorithm (LARFE) for the tabular case that uses layer-wise exploration (which gives rise to H batches of exploration), with a carefully chosen batch of exploration policy for the next layer after knowing the MDP parameters for the current layer.

The main difference from the tabular case is that instead of estimating the transition kernels as discrete probability distributions, we now solve linear regression problems. Instead of identifying the policies that maximizes the visitation measure to every (h, s, a), we identify a set of policies $\Pi_{h,\epsilon}$ that maximizes the visitation to every direction of features $\phi(h, s, a)$ that is relevant to learning while still keeping the set relatively small. Then the batched exploration policy π that can be obtained using a variant of G-optimal experiment design that minimizes the maximum "misalignment" of the covariance matrix, namely, $\max_{\pi'\in\Pi_{h,\epsilon}} \mathbb{E}_{\pi'}[\phi(s,a)^T\Sigma_\pi\phi(s,a)]$. This is still infeasible because π' is not executed, but we can estimate the $\mathbb{E}[\cdot]$ uniformly for every $\pi' \in \Pi_{h,\epsilon}$ and showed that the approximate G-optimal design still works.

7.3 Beyond Linear MDPs

Low-adaptive RL beyond linear function approximation is more open-ended. Most existing work settles with $O(\log T)$-style switching cost bounds that generalizes the "doubling trick" to more abstract settings such as linear Bellman-complete MDPs with low inherent Bellman error [72] or low Bellman Eluder-dimension [103]. There hasn't been any algorithm that

achieves no regret learning with either $O(\log\log T)$ switching cost or $O(\log\log T)$ batches of exploration. This is a major open problem in this space. The best-policy identification problem is likely to be easier. We believe reward-free exploration in the low-adaptive case is tractable by combining techniques from [70] and [98].

8. CONCLUSION AND OPEN PROBLEMS

In this paper, we have surveyed recent advances in the statistical theory of offline reinforcement learning as well as the related problem of low-adaptive exploration. Both problems are well-motivated by the emerging applications of reinforcement learning for real-life sequential decision-making problems. We covered results that characterize the optimal statistical complexity of each problem family as well as algorithms that are not only minimax optimal but also adaptive to individual problem instances across a hierarchy of coverage assumptions and structural conditions. We described not only the technical results but also theoretical insights on how these algorithms work and where the technical challenges are.

We conclude the paper by highlighting a few open directions of research in this rich problem space.

- **Agnostic Offline RL with function approximation.** Most provable offline RL algorithms in the function approximation settings require strong assumptions on the realizability and self-consistency (i.e., Bellman completeness) of the given function class. In practice, it is observed that even when linear function approximation is a poor approximation, the resulting policy that one can learn with it under a realistic exploration budget is still very impressive. At the moment there is no appropriate theoretical framework that satisfactorily quantifies this behavior. It will be nice to understand how much we can push the theoretical limit towards achieving similar levels of agnostic learning for offline (and online) RL comparable to supervised learning.
- $O(\log\log T)$**-adaptive RL with function approximation** As we described in Section 7.2 it remains open even under linear MDP how to achieve the optimal $O(\log\log T)$ batch complexity or switching cost while achieving a $\tilde{O}(\sqrt{T})$ regret. This is a concrete open problem that we hope to see resolved in the next few years.
- **Efficient computation** The paper focuses on the information-theoretical aspects of the problems and does not distinguish whether the OPE estimators, offline RL algorithms or the low-adaptive online learners are efficiently computable. For offline RL, anything beyond linear MDPs are computationally intractable. For low-adaptive RL, the algorithms are inefficient even for the tabular case (except in some cases when there are linear-program reformulations of the experiment-design).
- **Theory-inspired algorithms in offline Deep RL** Despite the widely-recognized importance of offline RL problems, the theory and practice remain pretty disjoint. The principle of "pessimism" is independently discovered but the theoretically approaches for implementing "pessimism" and deep RL heuristics for implementing "pessimism" are very different. The Deep RL heuristics are often overly optimized to the specific test cases in popular benchmarks and do not work well in new problems. This was demonstrated in the context of RL for computer networking [25] and that an simple alternative algorithm inspired by the pessimistic bonus of [98] turns out to work significantly better than state-of-the-art deep RL counterparts. We believe it is a productive avenue of research to bring some of the theoretical ideas from offline and low-adaptive RL to practice in different problem domains.

APPENDIX A: EXAMPLES OF "CURSE OF HORIZON" FOR IMPORTANCE SAMPLING ESTIMATORS

In this Appendix, we provide two concrete examples where the IS estimators suffer from the "Curse of Horizon".

Example 1.[46] Consider a "ring MDP" with n (an odd number) states $\mathcal{S} = \{0, 1, \cdots, n-1\}$, arranged on a circle (see the figure on the right). There are two actions for all states, "L" and "R". The L action moves the agent from the current state counterclockwise to the next state, and the R action does the opposite direction. This can be equivalently written as:

$$P(s' \mid s,\ \mathrm{L}) = \mathbb{I}(s' = s - 1 \bmod n)$$
$$P\left(s' \mid s, \mathrm{R}\right) = \mathbb{I}\left(s' = s + 1 \bmod n\right).$$

Let $\eta \in [0,1]$ and $\eta \neq 1/2$. We choose the behavior policy μ and target policy π as follows: $\pi(\mathrm{R}|s) = \mu(\mathrm{L}|s) = 1-\eta,\ \ \mu(\mathrm{R}|s) = \pi(\mathrm{L}|s) = \eta$.

PROPOSITION 1. Variance of cumulative ratio $\rho_{1:H}$ grows exponentially in H. Formally, $\mathrm{Var}_\mu[\rho_{1:H}] = A_\eta^H - 1$ with $A_\eta = \frac{\eta^3 + (1-\eta)^3}{(1-\eta)\eta} > 1$. Similarly, it further holds $\mathrm{Var}_\mu[\widehat{v}_{\mathrm{IS}}^\pi] = \Theta(A_\eta^H)$.

PROOF. Denote $C = (1-\eta)/\eta$ and $\boldsymbol{\tau}$ to be the random trajectory, then $F(\boldsymbol{\tau}) = \sum_{t=1}^H \mathbb{I}(a_t = \mathrm{R})$ follows a Binomial distribution $Binomial(H, \eta)$. Furthermore, the relation holds that

$$\rho_{1:H}(\boldsymbol{\tau}) = \prod_{t=1}^H \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} = \left(\frac{1-\eta}{\eta}\right)^{2F(\boldsymbol{\tau})-H} = C^{2F(\boldsymbol{\tau})-H}.$$

Note $F(\boldsymbol{\tau}) \sim Bin(H, \eta)$ implies $\mathbb{E}_{\boldsymbol{\tau}\sim\mu}[\rho_{1:H}(\boldsymbol{\tau})] = 1$, and the second order moment

$$\mathbb{E}_{\boldsymbol{\tau}\sim\mu}\left[\rho_{1:H}(\boldsymbol{\tau})^2\right] = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[(C^{2F(\boldsymbol{\tau})-H})^2\right] =$$
$$\Phi(4\log C)\cdot C^{-2H} = \left[\left(1-\eta+\eta C^4\right)C^{-2}\right]^H = A_\eta^H.$$

Here Φ is the moment generating generating function of Binomial distribution ($\forall \lambda \in \mathbb{R}$):

$$\Phi(\lambda) := \mathbb{E}_{\boldsymbol{\tau}\sim\mu}[\exp(\lambda F(\boldsymbol{\tau}))] = (1-\eta+\eta\exp(\lambda))^H$$

Therefore, the variance is $A_\eta^H - 1$ which is exponential in H. Besides, $\mathrm{Var}_\mu[\widehat{v}_{\mathrm{IS}}^\pi] = \Theta(A_\eta^H)$ can be proved similarly. □

Example 2. [90] For the second example, we can consider an MDP with i.i.d. state transition and constant sparse reward 1 shown at the last step. The IS estimator becomes

$\widehat{v}^{\pi}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n}[\prod_{t=1}^{H}\frac{\pi\left(a_t^{(i)}|s_t^{(i)}\right)}{\mu\left(a_t^{(i)}|s_t^{(i)}\right)}]$. Suppose $\log\frac{\pi_t}{\mu_t}$ is bounded (or equivalently $\frac{\pi_t}{\mu_t}$ is bounded from both sides) with $E_{\log} = \mathbb{E}[\log\frac{\pi_t}{\mu_t}]$ and $V_{\log} = \mathrm{Var}[\log\frac{\pi_t}{\mu_t}]$. By Central limit theorem, random variable $\sum_{t=1}^{H}\frac{\pi_t}{\mu_t} \sim \mathcal{N}(HE_{\log}, HE_{\log})$ asymptotically, and this is the same as $\prod_{t=1}^{H}\frac{\pi_t}{\mu_t} \sim \mathrm{LogNormal}(HE_{\log}, HV_{\log})$. This comes from the state transitions are i.i.d. The variance of $\prod_{t=1}^{H}\frac{\pi_t}{\mu_t}$ is again exponential in horizon $\Theta(\exp(HV_{\log}))$.

Both examples have finite number of states and actions, which demonstrates that IS-based estimators suffer from exponential variance even for the simplest tabular RL.

As we discussed, there are other estimators that do not suffer from the curse of horizon for these problems, but they all require the value functions to be easily estimable (with a small state space being a special case).

REFERENCES

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] 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. PMLR, 2020.

[3] Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10 (4):1116–1135, 2000.

[4] András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

[5] P Auer. Finite-time analysis of the multiarmed bandit problem, 2002.

[6] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR, 2017.

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

[8] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Technical Report*, 2022.

[9] Richard Bellman. Dynamic programming. *science*, 153(3731): 34–37, 1966.

[10] Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems*, 26, 2013.

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

[12] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[13] Qiwei Di, Heyang Zhao, Jiafan He, and Quanquan Gu. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. *arXiv preprint arXiv:2310.01380*, 2023.

[14] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[15] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[16] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

[17] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610 (7930):47–53, 2022.

[18] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pages 700–725. Cambridge University Press, 1925.

[19] Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

[20] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655, 2019.

[22] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349, 2013.

[23] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18 (2):141–153, 1989.

[24] Geoffrey J Gordon. *Approximate solutions to Markov decision processes*. Carnegie Mellon University, 1999.

[25] Momin Haider, Ming Yin, Menglei Zhang, Arpit Gupta, Jing Zhu, and Yu-Xiang Wang. Networkgym: Reinforcement learning environments for multi-access traffic management in network simulation. *Advances in Neural Information Processing Systems (NeurIPS 2024)-Dataset and Benchmark*, 2024.

[26] Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pages 1372–1383. PMLR, 2017.

[27] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

[28] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[29] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[30] Jiawei Huang, Jinglin Chen, Li Zhao, Tao Qin, Nan Jiang, and Tie-Yan Liu. Towards deployment-efficient reinforcement learning: Lower bound and optimality. In *International Conference on Learning Representations*, 2022.

[31] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[32] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

[33] Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

[34] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[35] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

[36] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

[37] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

[38] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

[39] Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pages 5774–5783. PMLR, 2021.

[40] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[41] John Lafferty, Han Liu, and Larry Wasserman. Minimax theory. *Lecture notes on Statistical Machine Learning*, 2008. URL http://www.stat.cmu.edu/~larry/=sml/Minimax.pdf.

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

[43] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

[44] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

[45] Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

[46] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

[47] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with stationary distribution correction. In *Uncertainty in artificial intelligence*, pages 1180–1190. PMLR, 2020.

[48] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

[49] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1711–1724, 2022.

[50] 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, 2014.

[51] Daniel J Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263, 2023.

[52] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. Neural adaptive video streaming with pensieve. In *Proceedings of the conference of the ACM special interest group on data communication*, pages 197–210, 2017.

[53] Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. In *International Conference on Learning Representations*, 2021.

[54] Donald L McLeish. Dependent central limit theorems and invariance principles. *the Annals of Probability*, 2(4):620–628, 1974.

[55] Yifei Min, Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34:7598–7610, 2021.

[56] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[57] Christopher Z Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. sage, 1993.

[58] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

[59] Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

[60] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

[61] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[62] Shamim Nemati, Mohammad M Ghassemi, and Gari D Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2978–2981. IEEE, 2016.

[63] Thanh Nguyen-Tang and Raman Arora. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. *Advances in neural information processing systems*, 36, 2024.

[64] Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, and Raman Arora. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9310–9318, 2023.

[65] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[66] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660 – 681, 2016.

[67] Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

[68] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[69] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

[70] Dan Qiao and Yu-Xiang Wang. Near-optimal deployment efficiency in reward-free reinforcement learning with linear function approximation. In *International Conference on Learning Representations*, 2023.

[71] Dan Qiao, Ming Yin, Ming Min, and Yu-Xiang Wang. Sample-efficient reinforcement learning with loglog (t) switching cost. In *International Conference on Machine Learning*, pages 18031–18061. PMLR, 2022.

[72] Dan Qiao, Ming Yin, and Yu-Xiang Wang. Logarithmic switching cost in reinforcement learning beyond linear mdps. *ISIT-2024*, 2024.

[73] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[74] Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34:15621–15634, 2021.

[75] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[76] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[77] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

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

[79] 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.

[80] Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[81] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[82] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[83] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[84] Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34:13524–13536, 2021.

[85] Xinqi Wang, Qiwen Cui, and Simon S Du. On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:14865–14877, 2022.

[86] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

[87] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[88] Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pages 11362–11371. PMLR, 2021.

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

[90] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

[91] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[92] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[93] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

[94] 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.

[95] Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

[96] 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, 2021.

[97] Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

[98] Ming Yin, Mengdi Wang, and Yu-Xiang Wang. Offline reinforcement learning with differentiable function approximation is provably efficient. *International Conference on Learning Representations*, 2023.

[99] 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.

[100] Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *International Conference on Machine Learning*, pages 26713–26749. PMLR, 2022.

[101] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020.

[102] Zihan Zhang, Yuhang Jiang, Yuan Zhou, and Xiangyang Ji. Near-optimal regret bounds for multi-batch reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24586–24596, 2022.

[103] Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *Advances in Neural Information Processing Systems*, 2024.

[104] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pages 95–103, 2018.

[105] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.