Pessimistic Algorithms for Provably Efficient Offline Reinforcement Learning: Addressing Spurious Correlations and Uncertainty

Chris Junchi Li$^\diamond$

Department of Electrical Engineering and Computer Sciences$^\diamond$
University of California, Berkeley

October 7, 2024

Abstract

Offline reinforcement learning (RL) aims to learn an optimal policy from a fixed dataset without further interactions with the environment, presenting unique challenges compared to online RL, particularly in domains where data collection is expensive or risky. In this paper, we propose a novel pessimistic value iteration algorithm (PEVI) that incorporates a penalty function to address the spurious correlations and intrinsic uncertainty arising from the offline setting. We establish theoretical guarantees for PEVI in general Markov Decision Processes (MDPs) and prove its minimax optimality in the linear MDP setting. By decomposing suboptimality into spurious correlation, intrinsic uncertainty, and optimization error, we demonstrate how our method effectively minimizes suboptimality and achieves optimal policy learning under minimal assumptions on dataset coverage. Our work provides a principled framework that bridges the gap between theory and practice in offline RL.

Keywords: Offline reinforcement learning, pessimism, value iteration, suboptimality decomposition, linear MDPs, minimax optimality.

1 Introduction

Reinforcement learning (RL) has achieved remarkable success in various domains, such as game playing and robotics, where large-scale interactions with environments are feasible. However, in many real-world applications, such as healthcare and autonomous driving, interacting with the environment is costly, risky, or even impossible. These limitations have spurred interest in offline RL, where the goal is to learn effective policies from static datasets collected prior to learning. Offline RL, while promising, poses additional challenges compared to its online counterpart. Without continuous interaction, algorithms must deal with limited exploration, data coverage issues, and potential spurious correlations between actions and rewards.

This paper tackles the fundamental question: ***Can we design a provably efficient algorithm for offline RL under minimal assumptions about the dataset?*** Existing approaches in offline RL often impose strong assumptions on dataset coverage, which are unrealistic in many practical settings. Furthermore, the theoretical understanding of how to mitigate the impact of poor data coverage remains underdeveloped. To address these challenges, we propose a pessimistic value iteration algorithm (PEVI), which leverages a penalty function to combat spurious correlations and address the intrinsic uncertainty in offline data.

The key idea of our approach is to introduce a pessimism-based penalty that adapts the bonus function used in online RL, thereby preventing the algorithm from being misled by poorly covered states or actions. Our contributions are twofold: (i) We introduce a novel decomposition of suboptimality in offline RL into spurious correlation, intrinsic uncertainty, and optimization error. This decomposition allows us to isolate and address the primary sources of inefficiency in offline learning. (ii) We establish theoretical guarantees for PEVI in both general MDPs and the linear MDP setting. In particular, we show that PEVI is minimax optimal for linear MDPs, meaning that it achieves the best possible performance given the inherent uncertainty in the dataset.

Backgrounds The empirical success of online (deep) reinforcement learning (RL) [MKS$^+$15, SHM$^+$16, SSS$^+$17, VEB$^+$17] relies on two ingredients: (i) expressive function approximators, e.g., deep neural networks [LBH15], which approximate policies and values, and (ii) efficient data generators, e.g., game engines [BNVB13] and physics simulators [TET12], which serve as environments. In particular, learning the deep neural network in an online manner often necessitates millions to billions of interactions with the environment. Due to such a barrier of sample complexity, it remains notably more challenging to apply online RL in critical domains, e.g., precision medicine [GJK$^+$19] and autonomous driving [SSSS16], where interactive data collecting processes can be costly and risky. To this end, we study offline RL in this paper, which aims to learn an optimal policy based on a dataset collected a priori without further interactions with the environment. Such datasets are abundantly available in various domains, e.g., electronic health records for precision medicine [CM14] and human driving trajectories for autonomous driving [SKD$^+$20].

In comparison with online RL [LS20, AJK20], offline RL remains even less understood in theory [LGR12, LKTF20], which hinders principled developments of trustworthy algorithms in practice. In particular, as active interactions with the environment are infeasible, it remains unclear how to maximally exploit the dataset without further exploration. Due to such a lack of continuing exploration, which plays a key role in online RL, any algorithm for offline RL possibly suffers from the insufficient coverage of the dataset [WFK20]. Specifically, as illustrated in Section 3, two challenges arise:

(i) the intrinsic uncertainty, that is, the dataset possibly fails to cover the trajectory induced by the optimal policy, which however carries the essential information, and

(ii) the spurious correlation, that is, the dataset possibly happens to cover a trajectory unrelated to the optimal policy, which by chance induces a large cumulative reward and hence misleads the learned policy.

See Figures 1 and 2 for illustrations. As the dataset is collected a priori, which is often beyond the control of the learner, any assumption on the sufficient coverage of the dataset possibly fails to hold in practice [FMP19, ASN20, FKN$^+$20, GWN$^+$20].

In this paper, we aim to answer the following question: ***Is it possible to design a provably efficient algorithm for offline RL under minimal assumptions on the dataset?*** To this end, we propose a pessimistic value iteration algorithm (PEVI), which incorporates a penalty function (pessimism) into the value iteration algorithm [SB18, Sze10]. Here the penalty function simply flips the sign of the bonus function (optimism) for promoting exploration in online RL [JOA10, AOM17], which enables a straightforward implementation of PEVI in practice. Specifically, we study the episodic setting of the Markov decision process (MDP). Our theoretical contribution is fourfold:

(i) We decompose the suboptimality of any algorithm for offline RL into three sources, namely the intrinsic uncertainty, spurious correlation, and optimization error. In particular, we identify the key role of the spurious correlation, even in the multi-armed bandit (MAB), a special case of the MDP.

(ii) For any general MDP, we establish the suboptimality of PEVI under a sufficient condition on the penalty function. In particular, we prove as long as the penalty function is an uncertainty quantifier, which is defined in Section 4.1, pessimism allows PEVI to eliminate the spurious correlation from its suboptimality.

(iii) For the linear MDP [YW19, JYWJ20], we instantiate PEVI by specifying the penalty function. In particular, we prove such a penalty function is an uncertainty quantifier, which verifies the sufficient condition imposed in (ii). Correspondingly, we establish the suboptimality of PEVI for the linear MDP.

(iv) We prove PEVI is minimax optimal for the linear MDP up to multiplicative factors of the dimension and horizon. In particular, we prove the intrinsic uncertainty identified in (i) is impossible to eliminate, as it arises from the information-theoretic lower bound. Moreover, such a fundamental limit certifies an oracle property of PEVI, which is defined in Section 4.2. Specifically, the suboptimality of PEVI only depends on how well the dataset covers the trajectory induced by the optimal policy, which carries the essential information, rather than any trajectory unrelated to the optimal policy, which causes the spurious correlation.

Throughout our theory, we only require an assumption on the compliance of the dataset, that is, the data collecting process is carried out in the underlying MDP of interest. Such an assumption is minimal. In comparison with existing literature, we require no assumptions on the sufficient coverage of the dataset, e.g., finite concentrability coefficients [CJ19] and uniformly lower bounded densities of visitation measures [YBW20], which often fail to hold in practice. Meanwhile, we impose no restrictions on the affinity between the learned policy and behavior policy (for collecting data) [LSAB20], which is often employed as a regularizer (or equivalently, a constraint) in existing literature. See Section 1.1 for a detailed discussion.

1.1 Related Works

Our work adds to the vast body of existing literature on offline RL (also known as batch RL) [LGR12, LKTF20], where a learner only has access to a dataset collected a priori. Existing literature studies two tasks: (i) offline policy evaluation, which estimates the expected cumulative reward or (action- and state-) value functions of a target policy, and (ii) offline policy optimization, which learns an optimal policy that maximizes the expected cumulative reward. Note that (i) is also known as off-policy policy evaluation, which can be adapted to handle the online setting. Also, note that the target policy in (i) is known, while the optimal policy in (ii) is unknown. As (ii) is more challenging than (i), various algorithms for solving (ii), especially the value-based approaches, can be adapted to solve (i). Although we focus on (ii), we discuss the existing works on (i) and (ii) together.

A key challenge of offline RL is the insufficient coverage of the dataset [WFK20], which arises from the lack of continuing exploration [Sze10]. In particular, the trajectories given in the dataset and those induced by the optimal policy (or the target policy) possibly have different distributions, which is also known as distribution shift [LKTF20]. As a result, intertwined with overparameterized function approximators, e.g., deep neural networks, offline RL possibly suffers from the extrapolation error [FMP19], which is large on the states and actions that are less covered by the dataset. Such an extrapolation error further propagates through each iteration of the algorithm for offline RL, as it often relies on bootstrapping [SB18].

To address such a challenge, the recent works [FMP19, LTDC19, JGS$^+$19, WTN19, KFS$^+$19, KZTL20, ASN20, YTY$^+$20, KRNJ20, WNZ$^+$20, SSB$^+$20, NDGL20, LSAB20] demonstrate the empirical success of various algorithms, which fall into two (possibly overlapping) categories: (i) regularized policy-based approaches and (ii) pessimistic value-based approaches. Specifically, (i) regularizes (or equivalently, constrains) the policy to avoid visiting the states and actions that are less covered by the dataset, while (ii) penalizes the (action- or state-) value function on such states and actions.

On the other hand, the empirical success of offline RL mostly eludes existing theory. Specifically, the existing works require various assumptions on the sufficient coverage of the dataset, which is also known as data diversity [LKTF20]. For example, offline policy evaluation often requires the visitation measure of the behavior policy to be lower bounded uniformly over the state-action space. An alternative assumption requires the ratio between the visitation measure of the target policy and that of the behavior policy to be upper bounded uniformly over the state-action space. See, e.g., [JL16, TB16, FCG18, LLTZ18, XMW19, NCDL19, NDK$^+$19, TFL$^+$19, KU19, KU20, JH20, UHJ20, DJW20, YW20, YBW20, ND20, YND$^+$20, ZDLS20] and the references therein. As another example, offline policy optimization often requires the concentrability coefficient to be upper bounded, whose definition mostly involves taking the supremum of a similarly defined ratio over the state-action space. See, e.g., [ASM07, ASM08, MS08, FSM10, FGSM16, SGG$^+$15, CJ19, LCYW19, WCYW19, FYW20, FWXY20, XJ20a, XJ20b, LQM20, ZKB$^+$20] and the references therein.

In practice, such assumptions on the sufficient coverage of the dataset often fail to hold [FMP19, ASN20, FKN$^+$20, GWN$^+$20], which possibly invalidates existing theory. For example, even for the MAB, a special case of the MDP, it remains unclear how to maximally exploit the dataset without such assumptions, e.g., when each action (arm) is taken a different number of times. As illustrated in Section 3, assuming there exists a suboptimal action that is less covered by the dataset, it possibly interferes with the learned policy via the spurious correlation. As a result, it remains unclear how to learn a policy whose suboptimality only depends on how well the dataset covers the optimal action instead of the suboptimal ones. In contrast, our work proves that pessimism resolves such a challenge by eliminating the spurious correlation, which enables exploiting the essential information, e.g., the observations of the optimal action in the dataset, in a minimax optimal manner. Although the optimal action is unknown, our algorithm adapts to identify the essential information in the dataset via the oracle property. See Section 4 for a detailed discussion.

Our work adds to the recent works on pessimism [YTY$^+$20, KRNJ20, KZTL20, LSAB20, BGB20]. Specifically, [YTY$^+$20, KRNJ20] propose a pessimistic model-based approach, while [KZTL20] propose a pessimistic value-based approach, both of which demonstrate empirical successes. From a theoretical perspective, [LSAB20] propose a regularized (and pessimistic) variant of the fitted Q-iteration algorithm [ASM07, ASM08, MS08], which attains the optimal policy within a restricted class of policies without assuming the sufficient coverage of the dataset. In contrast, our work imposes no restrictions on the affinity between the learned policy and behavior policy. In particular, our algorithm attains the information-theoretic lower bound for the linear MDP [YW19, JYWJ20] (up to multiplicative factors of the dimension and horizon), which implies that given the dataset, the learned policy serves as the "best effort" among all policies since no other can do better. From another theoretical perspective, [BGB20] characterize the importance of pessimism, especially when the assumption on the sufficient coverage of the dataset fails to hold. In contrast, we propose a principled framework for achieving pessimism via the notion of uncertainty quantifier, which serves as a sufficient condition for general function approximators. See Section

4 for a detailed discussion. Moreover, we instantiate such a framework for the linear MDP and establish its minimax optimality via the information-theoretic lower bound. In other words, our work complements [BGB20] by proving that pessimism is not only "important" but also optimal in the sense of information theory.

2 Preliminaries

In this section, we first introduce the episodic Markov decision process (MDP) and the corresponding performance metric. Then we introduce the offline setting and the corresponding data collecting process.

2.1 Episodic MDP and Performance Metric

We consider an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ with the state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, transition kernel $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^H$, and reward function $r = \{r_h\}_{h=1}^H$. We assume the reward function is bounded, that is, $r_h \in [0,1]$ for all $h \in [H]$. For any policy $\pi = \{\pi_h\}_{h=1}^H$, we define the (state-)value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ at each step $h \in [H]$ as

$$V_h^\pi(x) = \mathbb{E}_\pi\Big[\sum_{i=h}^H r_i(s_i, a_i) \,\Big|\, s_h = x\Big] \tag{2.1}$$

and the action-value function (Q-function) $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at each step $h \in [H]$ as

$$Q_h^\pi(x, a) = \mathbb{E}_\pi\Big[\sum_{i=h}^H r_i(s_i, a_i) \,\Big|\, s_h = x, a_h = a\Big] \tag{2.2}$$

Here the expectation $\mathbb{E}_\pi$ in Equations (2.1) and (2.2) is taken with respect to the randomness of the trajectory induced by π, which is obtained by taking the action $a_i \sim \pi_i(\cdot \,|\, s_i)$ at the state s_i and observing the next state $s_{i+1} \sim \mathcal{P}_i(\cdot \,|\, s_i, a_i)$ at each step $i \in [H]$. Meanwhile, we fix $s_h = x \in \mathcal{S}$ in Equation (2.1) and $(s_h, a_h) = (x, a) \in \mathcal{S} \times \mathcal{A}$ in Equation (2.2). By the definition in Equations (2.1) and (2.2), we have the Bellman equation

$$V_h^\pi(x) = \langle Q_h^\pi(x, \cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}}, \quad Q_h^\pi(x, a) = \mathbb{E}\big[r_h(s_h, a_h) + V_{h+1}^\pi(s_{h+1}) \,\big|\, s_h = x, a_h = a\big]$$

where $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ is the inner product over $\mathcal{A}$, while $\mathbb{E}$ is taken with respect to the randomness of the immediate reward $r_h(s_h, a_h)$ and next state s_{h+1}. For any function $f : \mathcal{S} \to \mathbb{R}$, we define the transition operator at each step $h \in [H]$ as

$$(\mathbb{P}_h f)(x, a) = \mathbb{E}\big[f(s_{h+1}) \,\big|\, s_h = x, a_h = a\big] \tag{2.3}$$

and the Bellman operator at each step $h \in [H]$ as

$$\begin{aligned}(\mathbb{B}_h f)(x, a) &= \mathbb{E}\big[r_h(s_h, a_h) + f(s_{h+1}) \,\big|\, s_h = x, a_h = a\big] \\ &= \mathbb{E}\big[r_h(s_h, a_h) \,\big|\, s_h = x, a_h = a\big] + (\mathbb{P}_h f)(x, a) \end{aligned} \tag{2.4}$$

For the episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, we use π^*, Q_h^*, and V_h^* to denote the optimal policy, optimal Q-function, and optimal value function, respectively. We have $V_{H+1}^* = 0$ and the Bellman optimality equation

$$V_h^*(x) = \max_{a \in \mathcal{A}} Q_h^*(x, a), \quad Q_h^*(x, a) = (\mathbb{B}_h V_{h+1}^*)(x, a) \tag{2.5}$$

Meanwhile, the optimal policy π^* is specified by

$$\pi_h^*(\cdot \,|\, x) = \mathop{\arg\max}_{\pi_h} \langle Q_h^*(x, \cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}}, \quad V_h^*(x) = \langle Q_h^*(x, \cdot), \pi_h^*(\cdot \,|\, x)\rangle_{\mathcal{A}}$$

where the maximum is taken over all functions mapping from $\mathcal{S}$ to distributions over $\mathcal{A}$. We aim to learn a policy that maximizes the expected cumulative reward. Correspondingly, we define the performance metric as

$$\mathrm{SubOpt}(\pi; x) = V_1^{\pi^*}(x) - V_1^{\pi}(x) \tag{2.6}$$

which is the suboptimality of the policy π given the initial state $s_1 = x$.

2.2 Offline Data Collecting Process

We consider the offline setting, that is, a learner only has access to a dataset $\mathcal{D}$ consisting of K trajectories $\{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$, which is collected a priori by an experimenter. In other words, at each step $h \in [H]$ of each trajectory $\tau \in [K]$, the experimenter takes the action a_h^τ at the state x_h^τ, receives the reward $r_h^\tau = r_h(x_h^\tau, a_h^\tau)$, and observes the next state $x_{h+1}^\tau \sim \mathcal{P}_h(\cdot \,|\, s_h = x_h^\tau, a_h = a_h^\tau)$. Here a_h^τ can be arbitrarily chosen, while r_h and $\mathcal{P}_h$ are the reward function and transition kernel of an underlying MDP. We define the compliance of such a dataset with the underlying MDP as follows.

Definition 1 (Compliance)**.** *For a dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$, let $\mathbb{P}_{\mathcal{D}}$ be the joint distribution of the data collecting process. We say $\mathcal{D}$ is compliant with an underlying MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ if*

$$\begin{aligned} &\mathbb{P}_{\mathcal{D}}\big(r_h^\tau = r', x_{h+1}^\tau = x' \,\big|\, \{(x_h^j, a_h^j)\}_{j=1}^{\tau}, \{(r_h^j, x_{h+1}^j)\}_{j=1}^{\tau-1}\big) \\ &\quad = \mathbb{P}\big(r_h(s_h, a_h) = r', s_{h+1} = x' \,\big|\, s_h = x_h^\tau, a_h = a_h^\tau\big) \end{aligned} \tag{2.7}$$

for all $r' \in [0,1]$ and $x' \in \mathcal{S}$ at each step $h \in [H]$ of each trajectory $\tau \in [K]$. Here $\mathbb{P}$ on the right-hand side of Equation (2.7) *is taken with respect to the underlying MDP*

Equation (2.7) implies the following two conditions on $\mathbb{P}_{\mathcal{D}}$ hold simultaneously: (i) at each step $h \in [H]$ of each trajectory $\tau \in [K]$, (r_h^τ, x_{h+1}^τ) only depends on $\{(x_h^j, a_h^j)\}_{j=1}^{\tau} \cup \{(r_h^j, x_{h+1}^j)\}_{j=1}^{\tau-1}$ via (x_h^τ, a_h^τ), and (ii) conditioning on (x_h^τ, a_h^τ), (r_h^τ, x_{h+1}^τ) is generated by the reward function and transition kernel of the underlying MDP. Intuitively, (i) ensures $\mathcal{D}$ possesses the Markov property. Specifically, (i) allows the K trajectories to be interdependent, that is, at each step $h \in [H]$, $\{(x_h^\tau, a_h^\tau, r_h^\tau, x_{h+1}^\tau)\}_{\tau=1}^{K}$ are interdependent across each trajectory $\tau \in [K]$. Meanwhile, (i) requires the randomness of $\{(x_h^j, a_h^j, r_h^j, x_{h+1}^j)\}_{j=1}^{\tau-1}$ to be fully captured by (x_h^τ, a_h^τ) when we examine the randomness of (r_h^τ, x_{h+1}^τ).

Assumption 1 (Data Collecting Process)**.** *The dataset $\mathcal{D}$ that the learner has access to is compliant with the underlying MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$*

As a special case, Assumption 1 holds if the experimenter follows a fixed behavior policy. More generally, Assumption 1 allows a_h^τ to be arbitrarily chosen, even in an adaptive or adversarial manner, in the sense that the experimenter does not necessarily follow a fixed behavior policy. In particular, a_h^τ can be interdependent across each trajectory $\tau \in [K]$. For example, the experimenter can sequentially improve the behavior policy using any algorithm for online RL. Furthermore, Assumption 1 does not require the data collecting process to well explore the state space and action space.

3 *What Causes Suboptimality?*

In this section, we decompose the suboptimality of any policy into three sources, namely the spurious correlation, intrinsic uncertainty, and optimization error. We first analyze the MDP and then specialize the general analysis to the multi-armed bandit (MAB) for illustration.

3.1 Spurious Correlation Versus Intrinsic Uncertainty

We consider a meta-algorithm, which constructs an estimated Q-function $\widehat{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and an estimated value function $\widehat{V}_h : \mathcal{S} \to \mathbb{R}$ based on the dataset $\mathcal{D}$. We define the model evaluation error at each step $h \in [H]$ as

$$\iota_h(x,a) = (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - \widehat{Q}_h(x,a) \tag{3.1}$$

In other words, ι_h is the error that arises from estimating the Bellman operator $\mathbb{B}_h$ defined in Equation (2.4), especially the transition operator $\mathbb{P}_h$ therein, based on $\mathcal{D}$. Note that ι_h in Equation (3.1) is defined in a pointwise manner for all $(x,a) \in \mathcal{S} \times \mathcal{A}$, where $\widehat{V}_{h+1}$ and $\widehat{Q}_h$ depend on $\mathcal{D}$. The suboptimality of the policy $\widehat{\pi}$ corresponding to $\widehat{V}_h$ and $\widehat{Q}_h$ (in the sense that $\widehat{V}_h(x) = \langle \widehat{Q}_h(x,\cdot), \widehat{\pi}_h(\cdot \,|\, x)\rangle_{\mathcal{A}}$), which is defined in Equation (2.6), admits the following decomposition.

Lemma 1 (Decomposition of Suboptimality)**.** *Let $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ be the policy such that $\widehat{V}_h(x) = \langle \widehat{Q}_h(x,\cdot), \widehat{\pi}_h(\cdot \,|\, x)\rangle_{\mathcal{A}}$. For any $\widehat{\pi}$ and $x \in \mathcal{S}$, we have*

$$\begin{aligned}\mathrm{SubOpt}(\widehat{\pi}; x) = & \underbrace{-\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\big[\iota_h(s_h,a_h) \,\big|\, s_1 = x\big]}_{\text{(i): Spurious Correlation}} + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\iota_h(s_h,a_h) \,\big|\, s_1 = x\big]}_{\text{(ii): Intrinsic Uncertainty}} \\ & + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle \widehat{Q}_h(s_h,\cdot), \pi_h^*(\cdot \,|\, s_h) - \widehat{\pi}_h(\cdot \,|\, s_h)\rangle_{\mathcal{A}} \,\big|\, s_1 = x\big]}_{\text{(iii): Optimization Error}}\end{aligned} \tag{3.2}$$

Here $\mathbb{E}_{\widehat{\pi}}$ and $\mathbb{E}_{\pi^*}$ are taken with respect to the trajectories induced by $\widehat{\pi}$ and π^* in the underlying MDP given the fixed functions $\widehat{V}_{h+1}$ and $\widehat{Q}_h$, which determine ι_h

Proof of Lemma 1. See Section A for a detailed proof. □

In Equation (3.2), term (i) is more challenging to control, as $\widehat{\pi}$ and ι_h simultaneously depend on $\mathcal{D}$ and hence spuriously correlate with each other. In Section 3.2, we show such a spurious correlation can "mislead" $\widehat{\pi}$, which incurs a significant suboptimality, even in the MAB. Specifically, assuming hypothetically $\widehat{\pi}$ and ι_h are independent, term (i) is mean zero with respect to $\mathbb{P}_{\mathcal{D}}$ as long as ι_h is mean zero for all $(x,a) \in \mathcal{S} \times \mathcal{A}$, which only necessitates an unbiased estimator of $\mathbb{B}_h$ in Equation (3.1), e.g., the sample average estimator in the MAB. However, as $\widehat{\pi}$ and ι_h are spuriously correlated, term (i) can be rather large in expectation.

In contrast, term (ii) is less challenging to control, as π^* is intrinsic to the underlying MDP and hence does not depend on $\mathcal{D}$, especially the corresponding ι_h, which quantifies the uncertainty that arises from approximating $\mathbb{B}_h \widehat{V}_{h+1}$. In Section 4.3, we show such an intrinsic uncertainty is impossible to eliminate, as it arises from the information-theoretic lower bound. In addition, as the optimization error, term (iii) is nonpositive as long as $\widehat{\pi}$ is greedy with respect to $\widehat{Q}_h$, that is, $\widehat{\pi}_h(\cdot \,|\, x) = \arg\max_{\pi_h} \langle \widehat{Q}_h(x,\cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}}$ (although Equation (3.2) holds for any $\widehat{\pi}$ such that $\widehat{V}_h(x) = \langle \widehat{Q}_h(x,\cdot), \widehat{\pi}_h(\cdot \,|\, x)\rangle_{\mathcal{A}}$).

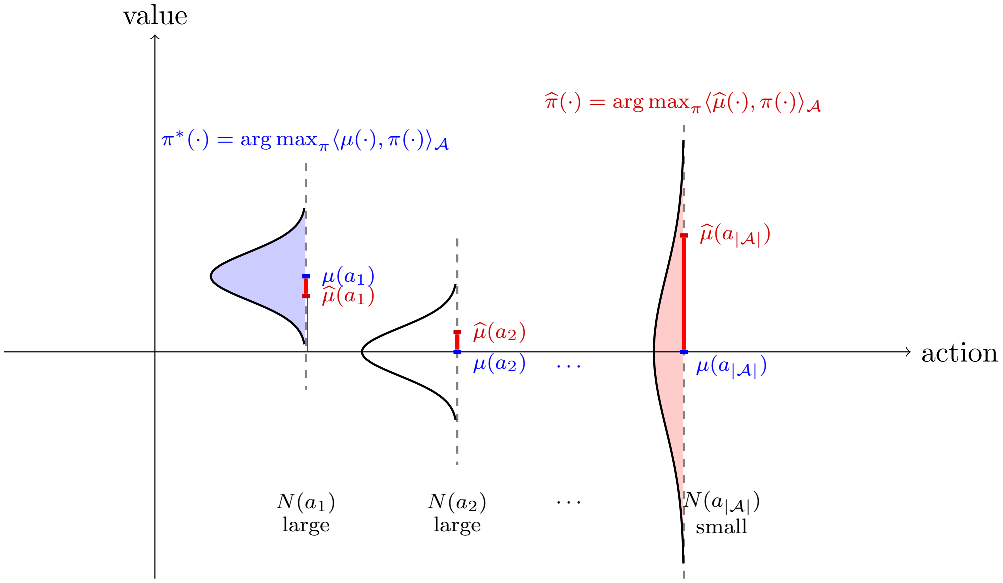
value
$\widehat{\pi}(\cdot) = \arg\max_\pi \langle \widehat{\mu}(\cdot), \pi(\cdot) \rangle_{\mathcal{A}}$
$\pi^*(\cdot) = \arg\max_\pi \langle \mu(\cdot), \pi(\cdot) \rangle_{\mathcal{A}}$
$\widehat{\mu}(a_{|\mathcal{A}|})$
$\mu(a_1)$
$\widehat{\mu}(a_1)$
$\widehat{\mu}(a_2)$
$\mu(a_2)$
$\cdots$
$\mu(a_{|\mathcal{A}|})$
action
$N(a_1)$ large
$N(a_2)$ large
$\cdots$
$N(a_{|\mathcal{A}|})$ small

Figure 1. An illustration of the spurious correlation in the MAB, a special case of the MDP, where $\mathcal{S}$ is a singleton, $\mathcal{A}$ is discrete, and $H = 1$. Here $\mu(a)$ is the expected reward of each action $a \in \mathcal{A}$ and $\widehat{\mu}(a)$ is its sample average estimator, which follows the Gaussian distribution in Equation (3.3). Correspondingly, $\iota(a) = \mu(a) - \widehat{\mu}(a)$ is the model evaluation error. As the greedy policy with respect to $\widehat{\mu}$, $\widehat{\pi}$ wrongly takes the action $a_{|\mathcal{A}|} = \arg\max_{a\in\mathcal{A}} \widehat{\mu}(a)$ with probability one only because $N(a_{|\mathcal{A}|})$ is relatively small, which allows $\widehat{\mu}(a_{|\mathcal{A}|})$ to be rather large, even though $\mu(a_{|\mathcal{A}|}) = 0$. Due to such a spurious correlation, $\widehat{\pi}$ incurs a significant suboptimality in comparison with π^*, which takes the action $a_1 = \arg\max_{a\in\mathcal{A}} \mu(a)$ with probability one.

3.2 Illustration via a Special Case: MAB

We consider the MAB, a special case of the MDP, where $\mathcal{S}$ is a singleton, $\mathcal{A}$ is discrete, and $H = 1$. To simplify the subsequent discussion, we assume without loss of generality

$$r(a) = \mu(a) + \epsilon, \quad \text{where} \ \ \epsilon \sim \mathrm{N}(0, 1)$$

Here $\mu(a)$ is the expected reward of each action $a \in \mathcal{A}$ and ϵ is independently drawn. For notational simplicity, we omit the dependency on $h \in [H]$ and $x \in \mathcal{S}$, as $H = 1$ and $\mathcal{S}$ is a singleton. Based on the dataset $\mathcal{D} = \{(a^\tau, r^\tau)\}_{\tau=1}^K$, where $r^\tau = r(a^\tau)$, we consider the sample average estimator

$$\widehat{\mu}(a) = \frac{1}{N(a)} \sum_{\tau=1}^K r^\tau \cdot \mathbb{1}\{a^\tau = a\}, \quad \text{where} \ \ N(a) = \sum_{\tau=1}^K \mathbb{1}\{a^\tau = a\}$$

Note that $\widehat{\mu}$ serves as the estimated Q-function. Under Assumption 1, we have

$$\widehat{\mu}(a) \sim \mathrm{N}\big(\mu(a), 1/N(a)\big) \tag{3.3}$$

In particular, $\{\widehat{\mu}(a)\}_{a\in\mathcal{A}}$ are independent across each action $a \in \mathcal{A}$ conditioning on $\{a^\tau\}_{\tau=1}^K$. We consider the policy

$$\widehat{\pi}(\cdot) = \arg\max_\pi \langle \widehat{\mu}(\cdot), \pi(\cdot) \rangle_{\mathcal{A}} \tag{3.4}$$

which is greedy with respect to $\widehat{\mu}$, as it takes the action $\arg\max_{a\in\mathcal{A}} \widehat{\mu}(a)$ with probability one.

By Equation (3.1), Lemma 1, and Equation (3.4), we have

$$\mathrm{SubOpt}(\widehat{\pi}; x) \le \underbrace{-\mathbb{E}_{\widehat{\pi}}\big[\iota(a)\big]}_{\text{(i)}} + \underbrace{\mathbb{E}_{\pi^*}\big[\iota(a)\big]}_{\text{(ii)}}, \quad \text{where} \;\; \iota(a) = \mu(a) - \widehat{\mu}(a)$$

Note that $\iota(a)$ is mean zero with respect to $\mathbb{P}_{\mathcal{D}}$ for each action $a \in \mathcal{A}$. Therefore, assuming hypothetically $\widehat{\pi}$ and ι are independent, term (i) is mean zero with respect to $\mathbb{P}_{\mathcal{D}}$. Meanwhile, as π^* and ι are independent, term (ii) is also mean zero with respect to $\mathbb{P}_{\mathcal{D}}$. However, as $\widehat{\pi}$ and ι are spuriously correlated due to their dependency on $\mathcal{D}$, term (i) can be rather large in expectation. See Figure 1 for an illustration. Specifically, we have

$$-\mathbb{E}_{\widehat{\pi}}\big[\iota(a)\big] = \langle \widehat{\mu}(\cdot) - \mu(\cdot), \widehat{\pi}(\cdot) \rangle_{\mathcal{A}} = \big\langle \widehat{\mu}(\cdot) - \mu(\cdot), \mathop{\arg\max}_{\pi} \langle \widehat{\mu}(\cdot), \pi(\cdot) \rangle_{\mathcal{A}} \big\rangle_{\mathcal{A}} \tag{3.5}$$

For example, assuming $\mu(a) = 0$ for each action $a \in \mathcal{A}$, term (i) is the maximum of $|\mathcal{A}|$ Gaussians $\{\mathrm{N}(0, 1/N(a))\}_{a \in \mathcal{A}}$, which can be rather large in expectation, especially when $N(a^\sharp)$ is relatively small for a certain action $a^\sharp \in \mathcal{A}$, e.g., $N(a^\sharp) = 1$. More generally, it is quite possible that $\widehat{\pi}$ takes a certain action $a^\natural \in \mathcal{A}$ with probability one only because $N(a^\natural)$ is relatively small, which allows $\widehat{\mu}(a^\natural)$ to be rather large, even when $\mu(a^\natural)$ is relatively small. Due to such a spurious correlation, $\langle \widehat{\mu}(\cdot) - \mu(\cdot), \widehat{\pi}(\cdot) \rangle_{\mathcal{A}} = \widehat{\mu}(a^\natural) - \mu(a^\natural)$ in Equation (3.5) can be rather large in expectation, which incurs a significant suboptimality. More importantly, such an undesired situation can be quite common in practice, as $\mathcal{D}$ does not necessarily have a "uniform coverage" over each action $a \in \mathcal{A}$. In other words, $N(a^\sharp)$ is often relatively small for at least a certain action $a^\sharp \in \mathcal{A}$.

Going beyond the MAB, that is, $H \ge 1$, such a spurious correlation is further exacerbated, as it is more challenging to ensure each state $x \in \mathcal{S}$ and each action $a \in \mathcal{A}$ are visited sufficiently many times in $\mathcal{D}$. To this end, existing literature [ASM07, ASM08, MS08, FSM10, FGSM16, SGG+15, LLTZ18, NCDL19, NDK+19, CJ19, TFL+19, KU19, KU20, FWXY20, XJ20a, XJ20b, JH20, UHJ20, DJW20, YBW20, QW20, LWC+20, LQM20, ND20, YND+20, ZKB+20, ZDLS20] relies on various assumptions on the "uniform coverage" of $\mathcal{D}$, e.g., finite concentrability coefficients and uniformly lower bounded densities of visitation measures, which however often fail to hold in practice.

4 Pessimism is Provably Efficient

In this section, we present the algorithm and theory. Specifically, we introduce a penalty function to develop a pessimistic value iteration algorithm (PEVI), which simply flips the sign of the bonus function for promoting exploration in online RL [JOA10, AYPS11, RVR13, OVR14, CG17, AOM17, JAZBJ18, JYWJ20, CYJW20, YJW+20a, AJS+20, WSY20]. In Section 4.1, we provide a sufficient condition for eliminating the spurious correlation from the suboptimality for any general MDP. In Section 4.2, we characterize the suboptimality for the linear MDP [YW19, JYWJ20] by verifying the sufficient condition in Section 4.1. In Section 4.3, we establish the minimax optimality of PEVI via the information-theoretic lower bound.

4.1 Pessimistic Value Iteration: General MDP

We consider a meta-algorithm, namely PEVI, which constructs an estimated Bellman operator $\widehat{\mathbb{B}}_h$ based on the dataset $\mathcal{D}$ so that $\widehat{\mathbb{B}}_h \widehat{V}_{h+1} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ approximates $\mathbb{B}_h \widehat{V}_{h+1} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Here $\widehat{V}_{h+1} : \mathcal{S} \to \mathbb{R}$ is an estimated value function constructed by the meta-algorithm based on $\mathcal{D}$. Note

Algorithm 1 Pessimistic Value Iteration (PEVI): General MDP

```
Input: Dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$.
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$.
for step $h = H, H-1, \ldots, 1$ do
    Construct $(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\cdot,\cdot)$ and $\Gamma_h(\cdot,\cdot)$ based on $\mathcal{D}$.    //Estimation & Uncertainty
    Set $\overline{Q}_h(\cdot,\cdot) \leftarrow (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\cdot,\cdot) - \Gamma_h(\cdot,\cdot)$.    //Pessimism
    Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\{\overline{Q}_h(\cdot,\cdot), H-h+1\}^+$.    //Truncation
    Set $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$.    //Optimization
    Set $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$.    //Evaluation
end for
Output: $\texttt{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=1}^H$.
```

that such a construction of $\widehat{\mathbb{B}}_h$ can be implicit in the sense that the meta-algorithm only relies on $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ instead of $\widehat{\mathbb{B}}_h$ itself. We define an uncertainty quantifier with the confidence parameter $\xi \in (0,1)$ as follows. Recall that $\mathbb{P}_\mathcal{D}$ is the joint distribution of the data collecting process.

Definition 2 (ξ-Uncertainty Quantifier)**.** *We say $\{\Gamma_h\}_{h=1}^H$ ($\Gamma_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$) is a ξ-uncertainty quantifier with respect to $\mathbb{P}_\mathcal{D}$ if the event*

$$\mathcal{E} = \Big\{ \big| (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a) - (\mathbb{B}_h \widehat{V}_{h+1})(x,a) \big| \leq \Gamma_h(x,a) \text{ for all } (x,a) \in \mathcal{S} \times \mathcal{A}, h \in [H] \Big\} \tag{4.1}$$

satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}) \geq 1 - \xi$

By Equation (4.1), Γ_h quantifies the uncertainty that arises from approximating $\mathbb{B}_h \widehat{V}_{h+1}$ using $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$, which allows us to develop the meta-algorithm (Algorithm 1).

The following theorem characterizes the suboptimality of Algorithm 1, which is defined in Equation (2.6).

Theorem 4.1 (Suboptimality for General MDP)**.** *Suppose $\{\Gamma_h\}_{h=1}^H$ in Algorithm 1 is a ξ-uncertainty quantifier. Under $\mathcal{E}$ defined in Equation (4.1), which satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}) \geq 1-\xi$, for any $x \in \mathcal{S}$, $\texttt{Pess}(\mathcal{D})$ in Algorithm 1 satisfies*

$$\mathrm{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq 2 \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\Gamma_h(s_h, a_h) \,\big|\, s_1 = x\big] \tag{4.2}$$

Here $\mathbb{E}_{\pi^*}$ is with respect to the trajectory induced by π^* in the underlying MDP given the fixed function Γ_h

Proof of Theorem 4.1. See Section 5.1 for a proof sketch. □

Theorem 4.1 establishes a sufficient condition for eliminating the spurious correlation, which corresponds to term (i) in Equation (3.2), from the suboptimality for any general MDP. Specifically, $-\Gamma_h$ in Algorithm 1 serves as the penalty function, which ensures $-\iota_h$ in Equation (3.2) is nonpositive under $\mathcal{E}$ defined in Equation (4.1), that is,

$$\begin{aligned} -\iota_h(x,a) &= \widehat{Q}_h(x,a) - (\mathbb{B}_h \widehat{V}_{h+1})(x,a) \leq \overline{Q}_h(x,a) - (\mathbb{B}_h \widehat{V}_{h+1})(x,a) \\ &= (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a) - (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - \Gamma_h(x,a) \leq 0 \end{aligned} \tag{4.3}$$

Note that Equation (4.3) holds in a pointwise manner for all $(x, a) \in \mathcal{S} \times \mathcal{A}$. In other words, as long as Γ_h is a ξ-uncertainty quantifier, the suboptimality in Equation (4.2) only corresponds to term (ii) in Equation (3.2), which characterizes the intrinsic uncertainty. In any concrete setting, e.g., the linear MDP, it only remains to specify Γ_h and prove it is a ξ-uncertainty quantifier under Assumption 1. In particular, we aim to find a ξ-uncertainty quantifier that is sufficiently small to establish an adequately tight upper bound of the suboptimality in Equation (4.2). In the sequel, we show it suffices to employ the bonus function for promoting exploration in online RL.

4.2 Pessimistic Value Iteration: Linear MDP

As a concrete setting, we study the instantiation of PEVI for the linear MDP. We define the linear MDP [YW19, JYWJ20] as follows, where the transition kernel and expected reward function are linear in a feature map.

Definition 3 (Linear MDP)**.** *We say an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ is a linear MDP with a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d unknown (signed) measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that*

$$\mathcal{P}_h(x' \,|\, x, a) = \langle \phi(x, a), \mu_h(x') \rangle, \quad \mathbb{E}\big[r_h(s_h, a_h) \,\big|\, s_h = x, a_h = a\big] = \langle \phi(x, a), \theta_h \rangle \tag{4.4}$$

for all $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ at each step $h \in [H]$. Here we assume $\|\phi(x, a)\| \leq 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \leq \sqrt{d}$ at each step $h \in [H]$, where with an abuse of notation, we define $\|\mu_h(\mathcal{S})\| = \int_{\mathcal{S}} \|\mu_h(x)\| \,\mathrm{d}x$

We specialize the meta-algorithm (Algorithm 1) by constructing $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$, Γ_h, and $\widehat{V}_h$ based on $\mathcal{D}$, which leads to the algorithm for the linear MDP (Algorithm 2). Specifically, we construct $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ based on $\mathcal{D}$ as follows. Recall that $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ approximates $\mathbb{B}_h \widehat{V}_{h+1}$, where $\mathbb{B}_h$ is the Bellman operator defined in Equation (2.4), and $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau, h=1}^{K, H}$ is the dataset. We define the empirical mean squared Bellman error (MSBE) as

$$M_h(w) = \sum_{\tau=1}^{K} \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau) - \phi(x_h^\tau, a_h^\tau)^\top w\big)^2$$

at each step $h \in [H]$. Correspondingly, we set

$$(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x, a) = \phi(x, a)^\top \widehat{w}_h, \quad \text{where} \quad \widehat{w}_h = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} M_h(w) + \lambda \cdot \|w\|_2^2 \tag{4.5}$$

at each step $h \in [H]$. Here $\lambda > 0$ is the regularization parameter. Note that $\widehat{w}_h$ has the closed form

$$\widehat{w}_h = \Lambda_h^{-1} \Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)\big) \Big)$$

$$\text{where} \quad \Lambda_h = \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I \tag{4.6}$$

Meanwhile, we construct Γ_h based on $\mathcal{D}$ as

$$\Gamma_h(x, a) = \beta \cdot \big(\phi(x, a)^\top \Lambda_h^{-1} \phi(x, a)\big)^{1/2} \tag{4.7}$$

Algorithm 2 Pessimistic Value Iteration (PEVI): Linear MDP

```
Input: Dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$.
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$.
for step $h = H, H-1, \ldots, 1$ do
    Set $\Lambda_h \leftarrow \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I$.
    Set $\widehat{w}_h \leftarrow \Lambda_h^{-1}(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)))$.      //Estimation
    Set $\Gamma_h(\cdot,\cdot) \leftarrow \beta \cdot (\phi(\cdot,\cdot)^\top \Lambda_h^{-1} \phi(\cdot,\cdot))^{1/2}$.      //Uncertainty
    Set $\overline{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{w}_h - \Gamma_h(\cdot,\cdot)$.      //Pessimism
    Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\{\overline{Q}_h(\cdot,\cdot), H-h+1\}^+$.      //Truncation
    Set $\widehat{\pi}_h(\cdot \,|\, \cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \,|\, \cdot)\rangle_{\mathcal{A}}$.      //Optimization
    Set $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \,|\, \cdot)\rangle_{\mathcal{A}}$.      //Evaluation
end for
Output: $\texttt{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=1}^{H}$.
```

at each step $h \in [H]$. Here $\beta > 0$ is the scaling parameter. In addition, we construct $\widehat{V}_h$ based on $\mathcal{D}$ as

$$\widehat{Q}_h(x,a) = \min\{\overline{Q}_h(x,a), H-h+1\}^+, \quad \text{where } \overline{Q}_h(x,a) = (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a) - \Gamma_h(x,a),$$
$$\widehat{V}_h(x) = \langle \widehat{Q}_h(x,\cdot), \widehat{\pi}_h(\cdot \,|\, x)\rangle_{\mathcal{A}}, \quad \text{where } \widehat{\pi}_h(\cdot \,|\, x) = \arg\max_{\pi_h} \langle \widehat{Q}_h(x,\cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}}$$

The following theorem characterizes the suboptimality of Algorithm 2, which is defined in Equation (2.6).

Theorem 4.2 (Suboptimality for Linear MDP)**.** *Suppose Assumption 1 holds and the underlying MDP is a linear MDP. In Algorithm 2, we set*[1]

$$\lambda = 1, \quad \beta = c \cdot dH\sqrt{\zeta}, \quad \textit{where } \zeta = \log(2dHK/\xi)$$

Here $c > 0$ is an absolute constant and $\xi \in (0,1)$ is the confidence parameter. The following statements hold: (i) $\{\Gamma_h\}_{h=1}^{H}$ in Algorithm 2, which is specified in Equation (4.7), is a ξ-uncertainty quantifier, and hence (ii) under $\mathcal{E}$ defined in Equation (4.1), which satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1-\xi$, for any $x \in \mathcal{S}$, $\texttt{Pess}(\mathcal{D})$ in Algorithm 2 satisfies

$$\mathrm{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq 2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big] \tag{4.8}$$

Here $\mathbb{E}_{\pi^*}$ is with respect to the trajectory induced by π^* in the underlying MDP given the fixed matrix Λ_h

Proof of Theorem 4.2. See Section 5.2 for a proof sketch. □

We highlight the following aspects of Theorem 4.2:

"Assumption-Free" Guarantee: Theorem 4.2 only relies on the compliance of $\mathcal{D}$ with the linear MDP. In comparison with existing literature [ASM07, ASM08, MS08, FSM10, FGSM16,

[1] As a side note, the factor d in B can be improved with a sample splitting trick; we apply this trick and defer the corresponding discussion to the kernel setting in Section 4.4.

SGG$^+$15, LLTZ18, NCDL19, NDK$^+$19, CJ19, TFL$^+$19, KU19, KU20, FWXY20, XJ20a, XJ20b, JH20, UHJ20, DJW20, YBW20, QW20, LWC$^+$20, LQM20, ND20, YND$^+$20, ZKB$^+$20, ZDLS20], we require no assumptions on the "uniform coverage" of $\mathcal{D}$, e.g., finite concentrability coefficients and uniformly lower bounded densities of visitation measures, which often fail to hold in practice. Meanwhile, we impose no restrictions on the affinity between $\texttt{Pess}(\mathcal{D})$ and a fixed behavior policy that induces $\mathcal{D}$, which is often employed as a regularizer (or equivalently, a constraint) in existing literature [FMP19, LTDC19, JGS$^+$19, WTN19, KFS$^+$19, WNZ$^+$20, SSB$^+$20, NDGL20, LSAB20].

Intrinsic Uncertainty Versus Spurious Correlation: The suboptimality in Equation (4.8) only corresponds to term (ii) in Equation (3.2), which characterizes the intrinsic uncertainty. Note that Λ_h depends on $\mathcal{D}$ but acts as a fixed matrix in the expectation, that is, $\mathbb{E}_{\pi^*}$ is only taken with respect to (s_h, a_h), which lies on the trajectory induced by π^*. In other words, as π^* is intrinsic to the underlying MDP and hence does not depend on $\mathcal{D}$, the suboptimality in Equation (4.8) does not suffer from the spurious correlation, that is, term (i) in Equation (3.2), which arises from the dependency of $\widehat{\pi} = \texttt{Pess}(\mathcal{D})$ on $\mathcal{D}$.

The following corollary proves as long as the trajectory induced by π^* is "covered" by $\mathcal{D}$ sufficiently well, the suboptimality of Algorithm 2 decays at a $K^{-1/2}$ rate.

Corollary 1 (Sufficient "Coverage"). *Suppose there exists an absolute constant $c^\dagger > 0$ such that the event*

$$\mathcal{E}^\dagger = \Big\{\Lambda_h \geq I + c^\dagger \cdot K \cdot \mathbb{E}_{\pi^*}\big[\phi(s_h, a_h)\phi(s_h, a_h)^\top \,\big|\, s_1 = x\big] \text{ for all } x \in \mathcal{S}, h \in [H]\Big\} \tag{4.9}$$

satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}^\dagger) \geq 1 - \xi/2$. Here Λ_h is defined in Equation (4.6) and $\mathbb{E}_{\pi^*}$ is taken with respect to the trajectory induced by π^* in the underlying MDP. In Algorithm 2, we set

$$\lambda = 1, \quad \beta = c \cdot dH\sqrt{\zeta}, \quad \text{where } \zeta = \log(4dHK/\xi)$$

Here $c > 0$ is an absolute constant and $\xi \in (0,1)$ is the confidence parameter. For $\texttt{Pess}(\mathcal{D})$ in Algorithm 2, the event

$$\mathcal{E}' = \Big\{\mathrm{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq c' \cdot d^{3/2}H^2K^{-1/2}\sqrt{\zeta} \text{ for all } x \in \mathcal{S}\Big\} \tag{4.10}$$

satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}') \geq 1 - \xi$, where $c' > 0$ is an absolute constant that only depends on $c^\dagger$ and c. In particular, if $\mathrm{rank}(\Sigma_h(x)) \leq r$ *for all $x \in \mathcal{S}$ at each step $h \in [H]$, where*

$$\Sigma_h(x) = \mathbb{E}_{\pi^*}\big[\phi(s_h, a_h)\phi(s_h, a_h)^\top \,\big|\, s_1 = x\big]$$

for $\texttt{Pess}(\mathcal{D})$ in Algorithm 2, the event

$$\mathcal{E}'' = \Big\{\mathrm{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq c'' \cdot dH^2K^{-1/2}\sqrt{\zeta} \text{ for all } x \in \mathcal{S}\Big\} \tag{4.11}$$

satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}'') \geq 1 - \xi$, where $c'' > 0$ is an absolute constant that only depends on $c^\dagger$, c, and r.

Proof of Corollary 1. See Appendix B.3 for a detailed proof. □

Intrinsic Uncertainty as Information Gain: To understand Equation (4.8), we interpret the intrinsic uncertainty in the suboptimality, which corresponds to term (ii) in Equation (3.2), from a Bayesian perspective. Recall that constructing $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}$ based on $\mathcal{D}$ at each step $h \in [H]$ involves solving the linear regression problem in Equation (4.5), where $\phi(x_h^\tau, a_h^\tau)$ is the covariate, $r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)$ is the response, and $\widehat{w}_h$ is the estimated regression coefficient. Here $\widehat{V}_{h+1}$ acts as a fixed function. We consider the Bayesian counterpart of such a linear regression problem. Note that the estimator $\widehat{w}_h$ specified in Equation (4.6) is the Bayesian estimator of w_h under the prior $w_h \sim \mathrm{N}(0, \lambda \cdot I)$ and Gaussian distribution of the response with variance one (conditioning on the covariate). Under the Bayesian framework, the posterior has the closed form

$$w_h \,|\, \mathcal{D} \sim \mathrm{N}(\widehat{w}_h, \Lambda_h^{-1}) \tag{4.12}$$

where $\widehat{w}_h$ and Λ_h are defined in Equation (4.6). Correspondingly, we have

$$\begin{aligned}\mathrm{I}\big(w_h; \phi(s_h, a_h) \,\big|\, \mathcal{D}\big) &= \mathrm{H}(w_h \,|\, \mathcal{D}) - \mathrm{H}\big(w_h \,\big|\, \mathcal{D}, \phi(s_h, a_h)\big) \\ &= 1/2 \cdot \log \frac{\det(\Lambda_h^\dagger)}{\det(\Lambda_h)}, \quad \text{where} \quad \Lambda_h^\dagger = \Lambda_h + \phi(s_h, a_h)\phi(s_h, a_h)^\top\end{aligned}$$

Here I is the (conditional) mutual information and H is the (conditional) differential entropy. Meanwhile, we have

$$\begin{aligned}\log \frac{\det(\Lambda_h^\dagger)}{\det(\Lambda_h)} &= \log\det\big(I + \Lambda_h^{-1/2}\phi(s_h, a_h)\phi(s_h, a_h)^\top \Lambda_h^{-1/2}\big) \\ &= \log\big(1 + \phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big) \approx \phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\end{aligned}$$

where the second equality follows from the matrix determinant lemma and the last equality holds when $\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)$ is close to zero. Therefore, in Equation (4.8), we have

$$\mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big] \approx \sqrt{2} \cdot \mathbb{E}_{\pi^*}\Big[\mathrm{I}\big(w_h; \phi(s_h, a_h) \,\big|\, \mathcal{D}\big)^{1/2} \,\Big|\, s_1 = x\Big]$$

In other words, the suboptimality in Equation (4.8), which corresponds to the intrinsic uncertainty, can be cast as the mutual information between $w_h \,|\, \mathcal{D}$ in Equation (4.12) and $\phi(s_h, a_h)$ on the trajectory induced by π^* in the underlying MDP. In particular, such a mutual information can be cast as the information gain [Sch91, Sch10, SGS11, SP12, HCD$^+$16, RVR16, RVR18] for estimating w_h, which is induced by observing $\phi(s_h, a_h)$ in addition to $\mathcal{D}$. In other words, such a mutual information characterizes how much uncertainty in $w_h \,|\, \mathcal{D}$ can be eliminated when we additionally condition on $\phi(s_h, a_h)$.

Illustration via a Special Case: Tabular MDP: To understand Equation (4.8), we consider the tabular MDP, a special case of the linear MDP, where $\mathcal{S}$ and $\mathcal{A}$ are discrete. Correspondingly, we set ϕ in Equation (4.4) as the canonical basis of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. When $\mathcal{S}$ is a singleton and $H = 1$, the tabular MDP reduces to the MAB, which is discussed in Section 3.2. Specifically, in the tabular MDP, we have

$$\begin{aligned}\Lambda_h &= \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I \\ &= \mathrm{diag}\big(\{N_h(x, a) + \lambda\}_{(x,a)\in\mathcal{S}\times\mathcal{A}}\big) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}, \quad \text{where} \quad N_h(x, a) = \sum_{\tau=1}^K \mathbb{1}\left\{(x_h^\tau, a_h^\tau) = (x, a)\right\}\end{aligned}$$

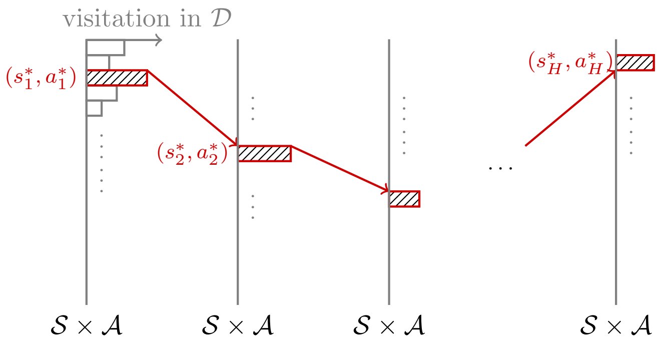
visitation in $\mathcal{D}$
(s_1^*, a_1^*)
(s_2^*, a_2^*)
(s_H^*, a_H^*)
$\mathcal{S} \times \mathcal{A}$
$\mathcal{S} \times \mathcal{A}$
$\mathcal{S} \times \mathcal{A}$
$\mathcal{S} \times \mathcal{A}$

Figure 2. An illustration of the oracle property in the tabular MDP, where the transition kernel is deterministic. The histogram depicts the number of times (s_h, a_h) is visited in $\mathcal{D}$. The suboptimality in Equation (4.8) only depends on the number of times (s_h^*, a_h^*), which lies on the trajectory induced by π^*, is visited in $\mathcal{D}$, even though π^* is unknown a priori.

To simplify the subsequent discussion, we assume $\mathcal{P}_h$ is deterministic at each step $h \in [H]$. Let $\{(s_h^*, a_h^*)\}_{h=1}^H$ be the trajectory induced by π^*, which is also deterministic. In Equation (4.8), we have

$$\mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big] = \big(N_h(s_h^*, a_h^*) + \lambda\big)^{-1/2}$$

In other words, the suboptimality in Equation (4.8) only depends on how well $\mathcal{D}$ "covers" the trajectory induced by π^* instead of its "uniform coverage" over $\mathcal{S}$ and $\mathcal{A}$. In particular, as long as $(s_h^\diamond, a_h^\diamond)$ lies off the trajectory induced by π^*, how well $\mathcal{D}$ "covers" $(s_h^\diamond, a_h^\diamond)$, that is, $N_h(s_h^\diamond, a_h^\diamond)$, does not affect the suboptimality in Equation (4.8). See Figure 2 for an illustration.

Oracle Property: Following existing literature [DJ94, FL01, Zou06], we refer to such a phenomenon as the oracle property, that is, the algorithm incurs an "oracle" suboptimality that automatically "adapts" to the support of the trajectory induced by π^*, even though π^* is unknown a priori. From another perspective, assuming hypothetically π^* is known a priori, the error that arises from estimating the transition kernel and expected reward function at (s_h^*, a_h^*) scales as $N_h(s_h^*, a_h^*)^{-1/2}$, which can not be improved due to the information-theoretic lower bound.

Outperforming Demonstration: Assuming hypothetically $\mathcal{D}$ is induced by a fixed behavior policy $\overline{\pi}$ (namely the demonstration), such an oracle property allows $\texttt{Pess}(\mathcal{D})$ to outperform $\overline{\pi}$ in terms of the suboptimality, which is defined in Equation (2.6). Specifically, it is quite possible that $r_h(s_h^\diamond, a_h^\diamond)$ is relatively small and $N_h(s_h^\diamond, a_h^\diamond)$ is rather large for a certain $(s_h^\diamond, a_h^\diamond)$, which is "covered" by $\mathcal{D}$ but lies off the trajectory induced by π^*. Correspondingly, the suboptimality of $\overline{\pi}$ can be rather large. On the other hand, as discussed above, $r_h(s_h^\diamond, a_h^\diamond)$ and $N_h(s_h^\diamond, a_h^\diamond)$ do not affect the suboptimality of $\texttt{Pess}(\mathcal{D})$, which can be relatively small as long as $N_h(s_h^*, a_h^*)$ is sufficiently large. Here (s_h^*, a_h^*) is "covered" by $\mathcal{D}$ and lies on the trajectory induced by π^*.

Well-Explored Dataset: To connect existing literature [DJW20], the following corollary specializes Theorem 4.2 under the additional assumption that the data collecting process well explores $\mathcal{S}$ and $\mathcal{A}$.

Corollary 2 (Well-Explored Dataset). *Suppose $\mathcal{D}$ consists of K trajectories $\{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ independently and identically induced by a fixed behavior policy $\overline{\pi}$ in the linear MDP. Meanwhile, suppose there exists an absolute constant $\underline{c} > 0$ such that*

$$\lambda_{\min}(\Sigma_h) \geq \underline{c}/d, \quad \textit{where} \;\; \Sigma_h = \mathbb{E}_{\overline{\pi}}\big[\phi(s_h, a_h)\phi(s_h, a_h)^\top\big]$$

at each step $h \in [H]$. Here $\mathbb{E}_{\overline{\pi}}$ is taken with respect to the trajectory induced by $\overline{\pi}$ in the underlying MDP. In Algorithm 2, we set

$$\lambda = 1, \quad \beta = c \cdot dH\sqrt{\zeta}, \quad \textit{where} \;\; \zeta = \log(4dHK/\xi)$$

Here $c > 0$ is an absolute constant and $\xi \in (0,1)$ is the confidence parameter. Suppose we have $K \geq C \cdot d \log(4dH/\xi)$, where $C > 0$ is a sufficiently large absolute constant that depends on $\underline{c}$. For $\texttt{Pess}(\mathcal{D})$ in Algorithm 2, the event

$$\mathcal{E}^* = \Big\{ \text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq c' \cdot d^{3/2} H^2 K^{-1/2} \sqrt{\zeta} \text{ for all } x \in \mathcal{S} \Big\} \tag{4.13}$$

satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}^*) \geq 1 - \xi$. Here $c' > 0$ is an absolute constant that only depends on $\underline{c}$ and c.

Proof of Corollary 2. See Appendix B.4 for a detailed proof. □

The suboptimality in Equation (4.13) parallels the policy evaluation error established in [DJW20], which also scales as $H^2 K^{-1/2}$ and attains the information-theoretic lower bound for offline policy evaluation. In contrast, we focus on offline policy optimization, which is more challenging. As $K \to \infty$, the suboptimality in Equation (4.13) goes to zero.

4.3 Minimax Optimality: Information-Theoretic Lower Bound

We establish the minimax optimality of Theorems 4.1 and 4.2 via the following information-theoretic lower bound. Recall that $\mathbb{P}_{\mathcal{D}}$ is the joint distribution of the data collecting process.

Theorem 4.3 (Information-Theoretic Lower Bound)**.** *For the output $\texttt{Algo}(\mathcal{D})$ of any algorithm, there exist a linear MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, an initial state $x \in \mathcal{S}$, and a dataset $\mathcal{D}$, which is compliant with $\mathcal{M}$, such that*

$$\mathbb{E}_{\mathcal{D}}\left[\frac{\text{SubOpt}\big(\texttt{Algo}(\mathcal{D}); x\big)}{\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x \Big]} \right] \geq c, \tag{4.14}$$

where $c > 0$ is an absolute constant. Here $\mathbb{E}_{\pi^*}$ is taken with respect to the trajectory induced by π^* in the underlying MDP given the fixed matrix Λ_h. Meanwhile, $\mathbb{E}_{\mathcal{D}}$ is taken with respect to $\mathbb{P}_{\mathcal{D}}$, where $\texttt{Algo}(\mathcal{D})$ and Λ_h depend on $\mathcal{D}$

Proof of Theorem 4.3. See Section 5.3 for a proof sketch and Appendix C.3 for a detailed proof. □

Theorem 4.3 matches Theorem 4.2 up to β and absolute constants. Although Theorem 4.3 only establishes the minimax optimality, Proposition 1 further certifies the local optimality on the constructed set of worst-case MDPs via a more refined instantiation of the meta-algorithm (Algorithm 1). See Appendix C.4 for a detailed discussion.

4.4 Pessimistic Value Iteration: Reproducing Kernel Hilbert Spaces

In this section, we study the Pessimistic Value Iteration in greater generality with kernel function approximation, covering the linear setting of Algorithm 2 as a special case. The algorithm we develop in this part slightly modifies the generic Algorithm 1 with a data splitting trick: we use distinct (and reverse-ordered) subsets of the offline dataset for the value iteration at each time step. Despite a (limited) reduction in the size of available sample, this modification allows us to remove the dependence on the covering number of the kernel function classes in the analysis of suboptimality upper bounds, thereby being particularly favorable if the covering number is large.

4.4.1 Basics of Reproducing Kernel Hilbert Space

To simplify the notations, we let $z = (s, a)$ denote the state-action pair and denote $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$ for any $h \in [H]$. Without loss of generality, we view $\mathcal{Z}$ as a compact subset of $\mathbb{R}^d$ where the dimension d is fixed. Let $\mathcal{H}$ be an RKHS of functions on $\mathcal{Z}$ with kernel function $K \colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, innder product $\langle \cdot, \cdot \rangle \colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ and RKHS norm $\| \cdot \|_{\mathcal{H}} \colon \mathcal{H} \to \mathbb{R}$. By definition of RKHS, there exists a *feature mapping* $\phi \colon \mathcal{Z} \to \mathcal{H}$ such that $f(z) = \langle f, \phi(z) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and all $z \in \mathcal{Z}$. Also, the kernel function admits the feature representation $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ for any $x, y \in \mathcal{H}$. Throughout the section, we assume that the kernel function is uniformly bounded as $\sup_{z \in \mathcal{Z}} K(z, z) < \infty$. Without loss of generality, we assume $\sup_{z \in \mathcal{Z}} K(z, z) \leq 1$ hence $\|\phi(z)\|_{\mathcal{H}} \leq 1$ for all $z \in \mathcal{Z}$.

Let $\mathcal{L}^2(\mathcal{Z})$ be the space of square-integrable functions on $\mathcal{Z}$ with respect to Lebesgue measure and let $\langle \cdot, \cdot \rangle_{\mathcal{L}^2}$ be the inner product for $\mathcal{L}^2(\mathcal{Z})$. The kernel function K induces an integral operator $T_K \colon \mathcal{L}^2(\mathcal{Z}) \to \mathcal{L}^2(\mathcal{Z})$ defined by

$$T_K f(z) = \int_{\mathcal{Z}} K(z, z') \cdot f(z') \mathrm{d}z', \quad \forall f \in \mathcal{L}^2(\mathcal{Z}) \tag{4.15}$$

Mercer's Theorem [SC08] implies that there exists a countable and non-increasing sequence of non-negative eigenvalues $\{\sigma_i\}_{i \geq 1}$ for the integral operator T_K, and the associated eigenfunctions $\{\psi_i\}_{i \geq 1}$ form an orthogonal basis of $\mathcal{L}^2(\mathcal{Z})$. Moreover, the kernel function admits a spectral representation $K(z, z') = \sum_{i=1}^{\infty} \sigma_i \cdot \psi_i(z) \cdot \psi_i(z')$ for all $z, z' \in \mathcal{Z}$. The eigenfunctions $\{\psi_i\}_{i \geq 1}$ also enables us to write the RKHS $\mathcal{H}$ as a subset of $\mathcal{L}^2(\mathcal{Z})$:

$$\mathcal{H} = \left\{ f \in \mathcal{L}^2(\mathcal{Z}) \colon \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle_{\mathcal{L}^2}^2}{\sigma_i} < \infty \right\}$$

such that the $\mathcal{H}$-inner product of any $f, g \in \mathcal{H}$ can be represented as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sigma_i^{-1} \cdot \langle f, \psi_i \rangle_{\mathcal{L}^2} \cdot \langle g, \psi_i \rangle_{\mathcal{L}}^2$$

Then the scaled eigenfunctions $\{\sqrt{\sigma_i} \psi_i\}_{i \geq 1}$ form an orthogonal basis for $\mathcal{H}$, and the feature mappling ϕ can be written as $\phi(z) = \sum_{i=1}^{\infty} \sigma_i \psi_i(z) \cdot \psi_i$ for any $z \in \mathcal{Z}$. In particular, assuming $\sigma_j = 0$ once $j > \gamma$ for some $\gamma \in \mathbb{N}$, there exists a feature mapping $\phi \colon \mathcal{X} \to \mathbb{R}^{\gamma}$, and $\mathcal{H}$ is a γ-dimensional RKHS. In this case, the kernel function approximation approach we introduce here recovers the linear setting in Section 4.2 with $d = \gamma$.

4.4.2 Pessimistic Value Iteration for Kernel Function Approximation with Data Splitting

We propose a slight modification of Algorithm 1 where we use a distinct fold of trajectories for value iteration at each step $h \in [H]$. The goal of such a data-splitting trick is to remove the dependency across $h \in [H]$ in Algorithm 1. In specific, recall that $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ in Line 4 of Algorithm 1 is obtained via the ridge regression given in Equation (4.5), which uses the whole dataset $\mathcal{D}$. As a result, for all $h \in [H]$, $\{x_h^{\tau}, a_h^{\tau}\}_{\tau \in [K]}$ are $\widehat{V}_{h+1}$ are statistically dependent. When quantifying the uncertainty of $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$, to handle such a statistical dependency, in the proof of Theorem 4.2, we additionally seek uniform concentration over a value function class that contains $\widehat{V}_{h+1}$. The covering number of such a function class partly determines the scaling parameter β in Theorem 4.2. When the function class has a large covering number, the uniform convergence approach for handling the

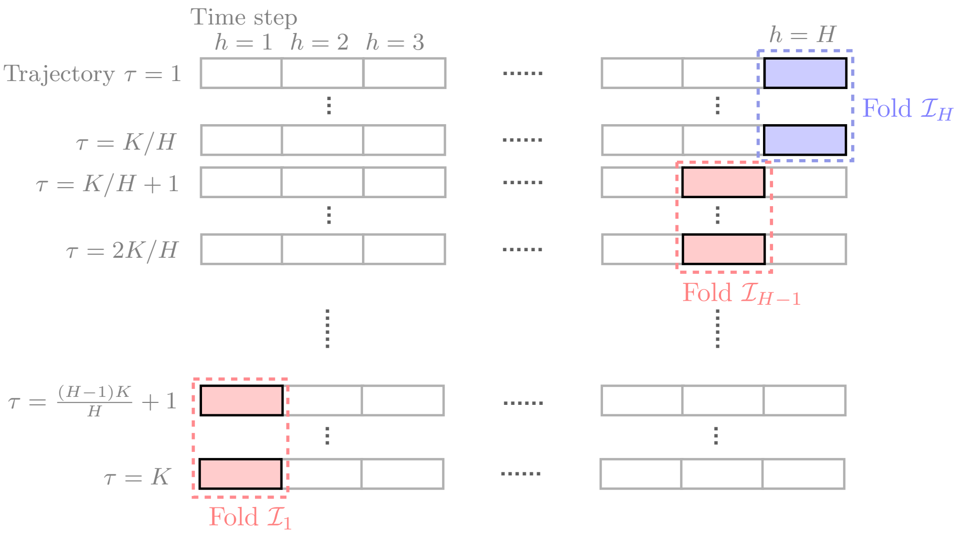

Figure 3. Illustration of the data-splitting step with a reverse ordering for all trajectories in the offline dataset.

dependency between $\{x_h^\tau, a_h^\tau\}_{\tau\in[K]}$ and $\widehat{V}_{h+1}$ would result in an excessively large β such that the uncertainty quantifier is loose. To resolve this issue, in the following, we split the dataset $\mathcal{D}$ into H parts $\{\mathcal{D}_h\}_{h\in[H]}$ and use each $\mathcal{D}_h$ in the construction of $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}$. Such a data splitting mechanism removes the undesirable statistical dependency. As we will show in Proposition 5, for linear function approximation, with data splitting, it suffices to choose $\beta = \widetilde{\mathcal{O}}(\sqrt{d}H)$ instead of $\beta = \widetilde{\mathcal{O}}(dH)$ in Theorem 4.2, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

Specifically, in the sequel, we first split the trajectories $\tau \in [K]$ into H disjoint and equally-sized folds $\mathcal{D}_h$ for all $h\in[H]$. Without loss of generality, we assume that $K/H \in \mathbb{Z}$. In specific, trajectories with $\tau \in \mathcal{I}_H = \{1,\ldots,K_0\}$ are used for the construction of $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}$ at time step $h = H$. Here we define $K_0 = K/H$ for simplicity. For all $h \in [H]$, let $\mathcal{I}_h = \{(H-h)\cdot K_0 + 1, \ldots, (H-h+1)\cdot K_0\}$. Then trajectories with $\tau \in \mathcal{I}_h$ is used to construct $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}$. An illustration of the data-splitting step is in Figure 3.

We then instantiate the pessimistic value iterations by constructing $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}$, Γ_h and $\widehat{V}_h$. To be specific, we define the empirical mean squared error (MSBE) as

$$M_h(f) = \sum_{\tau\in\mathcal{I}_h} \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau) - f(x_h^\tau, a_h^\tau)\big)^2$$

at each step $h\in[H]$ for all $f \in \mathcal{H}$, where only trajectories in fold $\mathcal{D}_h$ are involved. The empirical Bellman update is obtained from a kernel ridge regression so that

$$(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(z) = \widehat{f}_h(z), \quad \text{where} \quad \widehat{f}_h = \operatorname*{arg\,min}_{f\in\mathcal{H}} \; M_h(f) + \lambda\cdot\|f\|_{\mathcal{H}}^2$$

for some regularization parameter $\lambda > 0$. Following the same arguments as in [YJW$^+$20b], we note that $\widehat{f}_h$ admits a closed-form solution

$$\widehat{f}_h(z) = k_h(z)^\top (K_h + \lambda\cdot I)^{-1} y_h \tag{4.16}$$

where we define the Gram matrix $K_h \in \mathbb{R}^{K/H\times K/H}$ and the function $k_h\colon \mathcal{Z}\to\mathbb{R}^{K/H}$ as

$$K_h = \big[K(z_h^\tau, z_h^{\tau'})\big]_{\tau,\tau'\in\mathcal{I}_h} \in \mathbb{R}^{K/H\times K/H}, \quad k_h(z) = \big[K(z_h^\tau, z)\big]_{\tau\in\mathcal{I}_h}^\top \in \mathbb{R}^{K/H} \tag{4.17}$$

Algorithm 3 Pessimistic Value Iteration (PEVI): RKHS Approximation with data splitting

```
Input: Dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$.
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$.
Data splitting: Randomly split $[K]$ into disjoint and equal-sized folds $\{\mathcal{I}_h\}_{h=1}^H$, $|\mathcal{I}_h| = K/H$.
for step $h = H, H-1, \ldots, 1$ do
    Compute the Gram matrix $K_h \in \mathbb{R}^{K/H}$, function $k_h\colon \mathcal{Z} \to \mathbb{R}^{K/H}$ and response $y_h \in \mathbb{R}^{K/H}$
    as Equations (4.17) and (4.18).                                                          //Estimation
    Set $\Gamma_h(\cdot,\cdot) \leftarrow \beta \cdot \lambda^{-1/2} \cdot (K(\cdot,\cdot;\cdot,\cdot) - k_h(\cdot,\cdot)^\top (K_h + \lambda I)^{-1} k_h(\cdot,\cdot))^{1/2}$.   //Uncertainty
    Set $\overline{Q}_h(\cdot,\cdot) \leftarrow k_h(\cdot,\cdot)^\top (K_h + \lambda I)^{-1} y_h$.                  //Pessimism
    Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\{\overline{Q}_h(\cdot,\cdot), H-h+1\}^+$.                        //Truncation
    Set $\widehat{\pi}_h(\cdot\,|\,\cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot\,|\,\cdot)\rangle_{\mathcal{A}}$.            //Optimization
    Set $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot\,|\,\cdot)\rangle_{\mathcal{A}}$.                       //Evaluation
end for
Output: Pess($\mathcal{D}$) = $\{\widehat{\pi}_h\}_{h=1}^H$.
```

and the entry of the response vector $y_h \in \mathbb{R}^{K/H}$ corresponding to $\tau \in \mathcal{I}_h$ is

$$[y_h]_\tau = r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau) \tag{4.18}$$

Moreover, we construct Γ_h via

$$\Gamma_h(z) = \beta \cdot \lambda^{-1/2} \cdot \big(K(z,z) - k_h(z)^\top (K_h + \lambda I)^{-1} k_h(z)\big)^{1/2} \tag{4.19}$$

where $\beta > 0$ is a scaling parameter. Finally, we construct the pessimistic Q-function by

$$\widehat{Q}_h(z) = \min\big\{\overline{Q}_h(z), H-h+1\big\}^+, \quad \text{where } \overline{Q}_h(z) = \widehat{f}_h(z) - \Gamma_h(z),$$
$$\widehat{V}_h(x) = \big\langle \widehat{Q}_h(x,\cdot), \widehat{\pi}_h(\cdot\,|\,x)\big\rangle_{\mathcal{A}}, \quad \text{where } \widehat{\pi}_h(\cdot\,|\,x) = \arg\max_{\pi_h} \big\langle \widehat{Q}_h(x,\cdot), \widehat{\pi}_h(\cdot\,|\,x)\big\rangle_{\mathcal{A}}$$

Recall that we define the Bellman operator $\mathbb{B}_h$ on any value function $V\colon \mathcal{S} \to \mathbb{R}$ in Equation (2.4). We impose the following structural assumption for the kernel setting.

Assumption 2. *Let $R_Q > 0$ be some fixed constant and we define function class $\mathcal{Q}^* = \{f \in \mathcal{H}\colon \|f\|_{\mathcal{H}} \le R_Q H\}$. Then for any $h \in [H]$ and any $Q\colon \mathcal{S} \times \mathcal{A} \to [0,H]$, it holds that $\mathbb{B}_h V \in \mathcal{Q}^*$ for $V(x) = \max_{a \in \mathcal{A}} Q(x,a)$.*

The above assumption states that the Bellman operator maps any bounded function into a bounded RKHS-norm ball. When the RKHS has finite spectrum (i.e., $\sigma_j = 0$ once $j > \gamma$ for some $\gamma \in \mathbb{N}$), a sufficient condition for Assumption 2 to hold is the linear MDP defined in Definition 3.

Besides the closeness assumption on the Bellman operator, we also define the maximal information gain [SKKS09] as a characterization of the complexity of $\mathcal{H}$:

$$G(n,\lambda) = \sup\big\{1/2 \cdot \log\det(I + K_{\mathcal{C}}/\lambda)\colon \mathcal{C} \subset \mathcal{Z},\ |\mathcal{C}| \le n\big\} \tag{4.20}$$

Here $K_{\mathcal{C}}$ is the Gram matrix for the set $\mathcal{C}$, defined similarly as Equation (4.17). In particular, when $\mathcal{H}$ has γ-finite spectrum, $G(n,\lambda) = \mathcal{O}(\gamma \cdot \log n)$ recovers the dimensionality of the linear space up to a logarithmic factor. More importantly, information gain defined in (4.20) offers a characterization of the effective dimension of $\mathcal{H}$ especially when $\mathcal{H}$ is infinite-dimensional.

The suboptimality of the output of Algorithm 3 is characterized by the following theorem.

Theorem 4.4. *Suppose Assumption 2 holds, and there exists some $\lambda \geq 1 + 1/K$ and $B > 0$ satisfying*

$$2\lambda R_Q^2 + 2G(K/H, 1 + 1/K) + 2K/H \cdot \log(1 + 1/K) + 4\log\bigl(H/\xi\bigr) \leq (B/H)^2 \tag{4.21}$$

We set $\beta = B$ in Algorithm 3. Then with probability at least $1 - \xi$ with repsect to $\mathbb{P}_{\mathcal{D}}$, it holds simultaneously for all $x \in \mathcal{S}$ and all $h \in [H]$ that

$$\mathrm{SubOpt}\bigl(\texttt{Pess}(\mathcal{D}); x\bigr) \leq 2\sqrt{2}B \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Bigl[\bigl\{\log\det\bigl(I + K_h(s_h, a_h)/\lambda\bigr) - \log\det(I + K_h/\lambda)\bigr\}^{1/2} \Bigm| s_1 = x\Bigr]$$

Here $\mathbb{E}_{\pi^*}$ is taken with respect to the trajectory induced by π^* in the underlying MDP given the fixed Gram matrix $K_h \in \mathbb{R}^{K\times K}$ and the fixed operator $k_h\colon \mathcal{Z} \to \mathbb{R}^K$ defined in Equation (4.17), and for any $z \in \mathcal{Z}$, we define $K_h(z) \in \mathbb{R}^{(K+1)\times(K_1)}$ as the Gram matrix for $\{z_h^\tau\}_{\tau\in[H]} \cup \{z\}$.

Proof of Theorem 4.4. See Appendix D.1 for a detailed proof. □

Theorem 4.4 expresses the suboptimality upper bound in a generic form consisting of two parts: (i) a parameter $B > 0$ that depends on the kernel function class, as well as (ii) an information quantity

$$I_{\mathcal{D}} = \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Bigl[\bigl\{\log\det\bigl(I + K_h(s_h, a_h)/\lambda\bigr) - \log\det(I + K_h/\lambda)\bigr\}^{1/2} \Bigm| s_1 = x\Bigr] \tag{4.22}$$

that only depends on the optimal policy π^* and the offline dataset. In the same spirit of our preceding results, pessimism eliminates the spurious correlation (c.f. Equation (3.2)) with a properly constructed uncertainty quantifier $\{\Gamma_h\}_{h=1}^H$. When the RKHS has γ-finite spectrum with $\sigma_j = 0$ for all $j > \gamma$, $\mathcal{I}_{\mathcal{D}}$ reduces to the one in Theorem 4.2 for the linear setting (with data splitting).

In what follows, we interpret the generic bound in Theorem 4.4 under specific conditions on $\mathcal{H}$, focusing on the resulting forms of the two components. We would see the effect of sample splitting in our discussion. Firstly, the parameter $B > 0$ depends on the information gain $G(K/H, 1+1/K)$, which can be viewed as a characterization of the complexity of $\mathcal{H}$. We provide explicit choices of B under various eigenvalue decay conditions of $\mathcal{H}$ that decide such complexity.

Assumption 3 (Eigenvalue Decay of $\mathcal{H}$)**.** *Let $\{\sigma_j\}_{j\geq 1}$ be the eigenvalues induced by the integral opretaor T_K defined in Equation (4.15) and $\{\psi_j\}_{j\geq 1}$ be the associated eigenfunctions. We assume that $\{\sigma_j\}_{j\geq 1}$ satisfies one of the following conditions for some constant $\gamma > 0$.*

(i) γ-finite spectrum: $\sigma_j = 0$ for all $j > \gamma$, where γ is a positive integer.

(ii) γ-exponential decay: there exists some constants $C_1, C_2 > 0$, $\tau \in [0, 1/2)$ and $C_\psi > 0$ such that $\sigma_j \leq C_1 \cdot \exp(-C_2 \cdot j^\gamma)$ and $\sup_{z\in\mathcal{Z}} \sigma_j^\tau \cdot |\psi_j(z)| \leq C_\psi$ for all $j \geq 1$.

(iii) γ-polynomial decay: there exists some constants $C_1 > 0$, $\tau \in [0, 1/2)$ and $C_\psi > 0$ such that $\sigma_j \leq C_1 \cdot j^{-\gamma}$ and $\sup_{z\in\mathcal{Z}} \sigma_j^\tau \cdot |\psi_j(z)| \leq C_\psi$ for all $j \geq 1$, where $\gamma > 1$.

The γ-finite spectrum condition is satisfied by the linear MDP [JYWJ20] with feature dimension γ, and Algorithm 3 reduces to the algorithm for linear MDP established in preceding sections (with data splitting). Also, the exponential and polynomial decay are relatively mild conditions compared to those in the literature. We refer the readers to Section 4.1 of [YJW$^+$20b] for a detailed discussion on the eigenvalue decay conditions. Under the conditions in Assumption 3, Proposition 5 establishes the concrete choices of B for Theorem 4.4.

Proposition 5. *Under Assumptions 2 and 3, we set $\lambda \geq 1 + 1/K$ and $\beta = B$ in Algorithm 3, where*

$$B = \begin{cases} C \cdot H \cdot \big\{\sqrt{\lambda} + \sqrt{\gamma \cdot \log(KH/\xi)}\big\} & \gamma\text{-finite spectrum,} \\ C \cdot H \cdot \big\{\sqrt{\lambda} + \sqrt{(\log(KH/\xi))^{1+1/\gamma}}\big\} & \gamma\text{-exponential decay,} \\ C \cdot H \cdot \big\{\sqrt{\lambda} + K^{\frac{d+1}{2(\gamma+d)}} H^{-\frac{d+1}{2(\gamma+d)}} \cdot \sqrt{\log(KH/\xi)}\big\} & \gamma\text{-polynomial decay} \end{cases}$$

Here $C > 0$ is an absolute constant that does not depend on K or H. Then with probability at least $1 - \xi$ with respect to $\mathcal{D}$, it holds that

$$\mathrm{SubOpt}\big(\mathtt{Pess}(\mathcal{D}); x\big) \leq \begin{cases} C \cdot I_{\mathcal{D}} \cdot H \cdot \big\{\sqrt{\lambda} + \sqrt{\gamma \cdot \log(KH/\xi)}\big\} & \gamma\text{-finite spectrum,} \\ C \cdot I_{\mathcal{D}} \cdot H \cdot \big\{\sqrt{\lambda} + \sqrt{(\log KH/\xi)^{1+1/\gamma}}\big\} & \gamma\text{-exponential decay,} \\ C \cdot I_{\mathcal{D}} \cdot H \cdot \big\{\sqrt{\lambda} + K^{\frac{d+1}{2(\gamma+d)}} H^{-\frac{d+1}{2(\gamma+d)}} \cdot \sqrt{\log(KH/\xi)}\big\} & \gamma\text{-polynomial decay} \end{cases}$$

for some absolute constant $C > 0$ that does not depend on K or H.

Proof of Proposition 5. See Appendix D.2 for a detailed proof. □

Under γ-finite spectrum condition, taking $\lambda = 1 + 1/K$ in Algorithm 3 leads to a variant of Algorithm 2 with sample splitting; setting $\lambda = 1 + 1/K$ instead of $\lambda = 1$ does not change the order of upper bounds in Theorem 4.2 for linear MDP. Firstly, comparing $B = \widetilde{\mathcal{O}}(\sqrt{\gamma}H)$ in Proposition 5 to $B = \widetilde{\mathcal{O}}(\gamma H)$ for linear MDP where $d = \gamma$ in Theorem 4.2, data splitting improves the upper bound by a factor of $\sqrt{\gamma}$ since it removes the dependence of B on the covering number of the (linear) function class. Here $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors. On the other hand, in ideal settings such as the well-explored case of Corollary 2, $\mathcal{I}_{\mathcal{D}}$ is of order $\widetilde{\mathcal{O}}(\sqrt{d}H \cdot |\mathcal{I}_h|^{-1/2})$, where the reduction in sample size incurs an additional factor of $\sqrt{H}$. Thus, the data splitting approach is favorable if the horizon H is of a smaller order than $d = \gamma$. In general, the data splitting approach improves sample efficiency if the kernel function class has a covering number that is larger than $\exp(H)$.

To further understand the behavior of $\mathcal{I}_{\mathcal{D}}$ beyond the γ-finite spectrum setting, we now consider a special case where the offline dataset consists of i.i.d. trajectories induced by some behavior policy. This offers a more clear illustration of the learning performance by certain population quantities that characterzie how close the behavior policy is to π^*.

Offline dataset from i.i.d. sampling. We study a special case where the offline dataset consist of i.i.d. trajectories from some fixed behavior policy π^b; this enables us to translate $\mathcal{I}_{\mathcal{D}}$ into population quantities with specific choices of λ. The learning performance would depend on the "coverage" of π^b for the optimal policy π^*, communicated by the following notion of "effective dimension".

Definition 4 (Effective dimension)**.** *Let $K_h \in \mathbb{R}^{K/H \times K/H}$ be defined in (4.17) for all $h \in [H]$. Denote $\Sigma_h = \mathbb{E}_{\pi^b}[\phi(z_h)\phi(z_h)^\top \,|\, s_1 = x]$, $\Sigma_h^* = \mathbb{E}_{\pi^*}[\phi(z_h)\phi(z_h)^\top \,|\, s_1 = x]$, where $\mathbb{E}_{\pi^*}$ is taken with respect to (x_h, a_h) induced by the optimal policy π^*, and $\mathbb{E}_{\pi^b}$ is similarly induced by the behavior policy π^b. We define the (sample) effective dimension as*

$$d_{\mathit{eff}}^{\mathit{sample}} = \sum_{h=1}^{H} \mathit{Tr}\big((K_h + \lambda \mathcal{I}_{\mathcal{H}})^{-1}\Sigma_h^*\big)^{1/2} \tag{4.23}$$

Moreover, we define the population effective dimension under π^b as

$$d_{\mathit{eff}}^{\mathit{pop}} = \sum_{h=1}^{H} \mathit{Tr}\big((K/H \cdot \Sigma_h + \lambda \mathcal{I}_{\mathcal{H}})^{-1}\Sigma_h^*\big)^{1/2} \tag{4.24}$$

Corollary 3. *Suppose $\mathcal{D}$ consists of i.i.d. trajectories sampled from behavior policy π^b, and Assumption 3 holds; in case (iii) γ-polynomial decay, we additionally assume $\gamma(1-2\tau) > 1$. In Algorithm 3, we set $B > 0$ as in Proposition 5 and*

$$\lambda = \begin{cases} C \cdot \gamma \cdot \log(K/\xi) & \gamma\text{-finite spectrum,} \\ C \cdot \big[\log(K/\xi)\big]^{1+1/\gamma} & \gamma\text{-exponential decay,} \\ C \cdot (K/H)^{\frac{2}{\gamma(1-2\tau)-1}} \cdot \log(K/\xi) & \gamma\text{-polynomial decay} \end{cases}$$

where $C > 0$ is a sufficiently large absolute constant that does not depend on K or H. Then with probability at least $1-\xi$ with respect to $\mathcal{D}$, it holds that

$$\mathrm{SubOpt}\big(\mathtt{Pess}(\mathcal{D}); x\big) \le \begin{cases} C' \cdot d_{\mathit{eff}}^{\mathit{pop}} \cdot H \cdot \sqrt{\gamma \cdot \log(KH/\xi)}) & \gamma\text{-finite spectrum,} \\ C \cdot d_{\mathit{eff}}^{\mathit{pop}} \cdot H \cdot \sqrt{(\log(KH/\xi))^{1+1\gamma}} & \gamma\text{-exponential decay,} \\ C' \cdot d_{\mathit{eff}}^{\mathit{pop}} \cdot K^{\kappa^*} H^{\nu^*} \cdot \sqrt{\log(KH/\xi)} & \gamma\text{-polynomial decay} \end{cases} \tag{4.25}$$

where $C' > 0$ is an absolute constant that does not depend on K or H, and

$$\kappa^* = \frac{d+1}{2(\gamma+d)} + \frac{1}{\gamma(1-2\tau)-1}, \quad \nu^* = 1 - \frac{d+1}{2(\gamma+d)} - \frac{1}{\gamma(1-2\tau)-1}$$

The same results also apply to $d_{\mathit{eff}}^{\mathit{sample}}$.

Proof of Corollary 3. See Appendix D.3 for a detailed proof. □

Parallel to the linear setting, Corollary 3 demonstrates the performance of our method in terms of $d_{\text{eff}}^{\text{pop}}$ that depends on the relationship between Σ_h (from π^b) and Σ_h^* (from π^*). When Σ_h and Σ_h^* are close (i.e., π^b covers π^* well), we have $d_{\text{eff}}^{\text{pop}} \approx \widetilde{\mathcal{O}}(H^{3/2}K^{1/2})$; in this case, the suboptimality is of order $K^{-1/2}$ under γ-finite spectrum and γ-exponential decay, while for γ-polynomial decay we obtain a sublinear rate of $K^{\kappa^*-1/2}$.

5 Proof Sketch

In this section, we sketch the proofs of the main results in Section 4. In Section 5.1, we sketch the proof of Theorem 4.1, which handles any general MDP. In Section 5.2, we specialize it to the linear MDP, which is handled by Theorem 4.2. In Section 5.3, we sketch the proof of Theorem 4.3, which establishes the information-theoretic lower bound.

5.1 Suboptimality of PEVI: General MDP

Recall that we define the model evaluation errors $\{\iota_h\}_{h=1}^H$ in Equation (3.1), which are based on the (action- and state-)value functions $\{(\widehat{Q}_h, \widehat{V}_h)\}_{h=1}^H$ constructed by PEVI. Also, recall that we define the ξ-uncertainty quantifiers $\{\Gamma_h\}_{h=1}^H$ in Definition 2. The key to the proof of Theorem 4.1 is to show that for all $h \in [H]$, the constructed Q-function $\widehat{Q}_h$ in Algorithm 1 is a pessimistic estimator of the optimal Q-function Q_h^*. To this end, in the following lemma, we prove that under the event $\mathcal{E}$ defined in Equation (4.1), ι_h lies within $[0, 2\Gamma_h]$ in a pointwise manner for all $h \in [H]$. Recall that $\mathbb{P}_{\mathcal{D}}$ is the joint distribution of the data collecting process.

Lemma 2 (Pessimism for General MDP)**.** *Suppose that $\{\Gamma_h\}_{h=1}^H$ in Algorithm 1 are ξ-uncertainty quantifiers. Under $\mathcal{E}$ defined in Equation* (4.1)*, which satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$, we have*

$$0 \leq \iota_h(x,a) \leq 2\Gamma_h(x,a), \quad \textit{for all} \ \ (x,a) \in \mathcal{S} \times \mathcal{A}, \ h \in [H] \tag{5.1}$$

Proof of Lemma 2. See Appendix B.1 for a detailed proof. □

In Equation (5.1), the nonnegativity of $\{\iota_h\}_{h=1}^H$ implies the pessimism of $\{\widehat{Q}_h\}_{h=1}^H$, that is, $\widehat{Q}_h \leq Q_h^*$ in a pointwise manner for all $h \in [H]$. To see this, note that the definition of $\{\iota_h\}_{h=1}^H$ in Equation (3.1) gives

$$Q_h^*(x,a) - \widehat{Q}_h(x,a) \geq (\mathbb{B}_h V_{h+1}^*)(x,a) - (\mathbb{B}_h \widehat{V}_{h+1})(x,a) = (\mathbb{P}_h V_{h+1}^*)(x,a) - (\mathbb{P}_h \widehat{V}_{h+1})(x,a) \tag{5.2}$$

which together with Equations (2.3) and (2.5) further implies

$$\begin{aligned} Q_h^*(x,a) - \widehat{Q}_h(x,a) &\geq \mathbb{E}\big[\max_{a' \in \mathcal{A}} Q_{h+1}^*(s_{h+1}, a') - \langle \widehat{Q}_{h+1}(s_{h+1}, \cdot), \widehat{\pi}_{h+1}(\cdot \,|\, s_{h+1})\rangle_{\mathcal{A}} \,\big|\, s_h = x, a_h = a\big] \\ &\geq \mathbb{E}\big[\langle Q_{h+1}^*(s_{h+1}, \cdot) - \widehat{Q}_{h+1}(s_{h+1}, \cdot), \widehat{\pi}_{h+1}(\cdot \,|\, s_{h+1})\rangle_{\mathcal{A}} \,\big|\, s_h = x, a_h = a\big] \end{aligned} \tag{5.3}$$

for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$. Also, note that $\widehat{V}_{H+1} = V_{H+1}^* = 0$. Therefore, Equation (5.2) implies $Q_H^* \geq \widehat{Q}_H$ in a pointwise manner. Moreover, by recursively applying Equation (5.3), we have $Q_h^* \geq \widehat{Q}_h$ in a pointwise manner for all $h \in [H]$. In other words, Lemma 2 implies that the pessimism of $\{\widehat{Q}_h\}_{h=1}^H$ holds with probability at least $1 - \xi$ as long as $\{\Gamma_h\}_{h=1}^H$ in Algorithm 1 are ξ-uncertainty quantifiers, which serves as a sufficient condition that can be verified. Meanwhile, the upper bound of $\{\iota_h\}_{h=1}^H$ in Equation (5.1) controls the underestimation bias of $\{\widehat{Q}_h\}_{h=1}^H$, which arises from pessimism.

Based on Lemma 2, we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. We upper bound the three terms on the right-hand side of Equation (3.2) respectively. Specifically, we apply Lemma 1 by setting $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ as the output of Algorithm 1, that is, $\widehat{\pi} = \texttt{Pess}(\mathcal{D})$. As $\widehat{\pi}_h$ is greedy with respect to $\widehat{Q}_h$ for all $h \in [H]$, term (iii) in Equation (3.2) is nonpositive. Therefore, we have

$$\text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq \underbrace{-\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\big[\iota_h(s_h, a_h) \,\big|\, s_1 = x\big]}_{\text{(i)}} + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\iota_h(s_h, a_h) \,\big|\, s_1 = x\big]}_{\text{(ii)}} \tag{5.4}$$

for all $x \in \mathcal{S}$, where terms (i) and (ii) characterize the spurious correlation and intrinsic uncertainty, respectively. To upper bound such two terms, we invoke Lemma 2, which implies

$$\text{(i)} \leq 0, \quad \text{(ii)} \leq 2\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\Gamma_h(s_h, a_h) \,\big|\, s_1 = x\big] \tag{5.5}$$

for all $x \in \mathcal{S}$. Combining Equations (5.4) and (5.5), we obtain Equation (4.2) under $\mathcal{E}$ defined in Equation (4.1). Meanwhile, by Definition 2, we have $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$. Therefore, we conclude the proof of Theorem 4.1. □

5.2 Suboptimality of PEVI: Linear MDP

Based on Theorem 4.1, we are ready to prove Theorem 4.2, which is specialized to the linear MDP defined in Definition 3.

Proof of Theorem 4.2. It suffices to show that $\{\Gamma_h\}_{h=1}^H$ specified in Equation (4.7) are ξ-uncertainty quantifiers, which are defined in Definition 2. In the following lemma, we prove that such a statement holds when the regularization parameter $\lambda > 0$ and scaling parameter $\beta > 0$ in Algorithm 2 are properly chosen.

Lemma 3 (ξ-Uncertainty Quantifier for Linear MDP)**.** *Suppose that Assumption 1 holds and the underlying MDP is a linear MDP. In Algorithm 2, we set*

$$\lambda = 1, \quad \beta = c \cdot dH\sqrt{\zeta}, \quad \textit{where} \;\; \zeta = \log(2dHK/\xi)$$

Here $c > 0$ is an absolute constant and $\xi \in (0, 1)$ is the confidence parameter. It holds that $\{\Gamma_h\}_{h=1}^H$ specified in Equation (4.7) are ξ-uncertainty quantifiers, where $\{\widehat{V}_{h+1}\}_{h=1}^H$ used in Equation (4.1) are obtained by Algorithm 2

Proof of Lemma 3. See Appendix B.2 for a detailed proof. □

As Lemma 3 proves that $\{\Gamma_h\}_{h=1}^H$ specified in Equation (4.7) are ξ-uncertainty quantifiers, $\mathcal{E}$ defined in Equation (4.1) satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$. Recall that $\mathbb{P}_{\mathcal{D}}$ is the joint distribution of the data collecting process. By specializing Theorem 4.1 to the linear MDP, we have

$$\begin{aligned} \text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) &\leq 2\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\Gamma_h(s_h, a_h) \,\big|\, s_1 = x\big] \\ &= 2\beta \sum_{h=1}^H \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big] \end{aligned}$$

for all $x \in \mathcal{S}$ under $\mathcal{E}$ defined in Equation (4.1). Here the last equality follows from Equation (4.7). Therefore, we conclude the proof of Theorem 4.2. □

5.3 Minimax Optimality of PEVI

In this section, we sketch the proof of Theorem 4.3, which establishes the minimax optimality of Theorem 4.2 for the linear MDP. Specifically, in Section 5.3.1, we construct a class $\mathfrak{M}$ of linear MDPs and a worst-case dataset $\mathcal{D}$, while in Section 5.3.2, we prove Theorem 4.3 via the information-theoretic lower bound.

5.3.1 Construction of a Hard Instance

In the sequel, we construct a class $\mathfrak{M}$ of linear MDPs and a worst-case dataset $\mathcal{D}$, which is compliant with the underlying MDP as defined in Definition 1.

Linear MDP: We define the following class of linear MDPs

$$\mathfrak{M} = \big\{M(p_1, p_2, p_3) : p_1, p_2, p_3 \in [1/4, 3/4] \text{ with } p_3 = \min\{p_1, p_2\}\big\} \tag{5.6}$$

where $M(p_1, p_2, p_3)$ is an episodic MDP with the horizon $H \geq 2$, state space $\mathcal{S} = \{x_0, x_1, x_2\}$, and action space $\mathcal{A} = \{b_j\}_{j=1}^A$ with $|\mathcal{A}| = A \geq 3$. In particular, we fix the initial state as $s_1 = x_0$. For the transition kernel, at the first step $h = 1$, we set

$$\begin{aligned}
&\mathcal{P}_1(x_1 \,|\, x_0, b_1) = p_1, \quad \mathcal{P}_1(x_2 \,|\, x_0, b_1) = 1 - p_1,\\
&\mathcal{P}_1(x_1 \,|\, x_0, b_2) = p_2, \quad \mathcal{P}_1(x_2 \,|\, x_0, b_2) = 1 - p_2\\
&\mathcal{P}_1(x_1 \,|\, x_0, b_j) = p_3, \quad \mathcal{P}_1(x_2 \,|\, x_0, b_j) = 1 - p_3, \quad \text{for all} \;\; j \in \{3, \ldots, A\}
\end{aligned} \tag{5.7}$$

Meanwhile, at any subsequent step $h \in \{2, \ldots, H\}$, we set

$$\mathcal{P}_h(x_1 \,|\, x_1, a) = \mathcal{P}_h(x_2 \,|\, x_2, a) = 1, \quad \text{for all} \;\; a \in \mathcal{A}$$

In other words, $x_1, x_2 \in \mathcal{S}$ are the absorbing states. Here $\mathcal{P}_1(x_1 \,|\, x_0, b_1)$ abbreviates $\mathcal{P}_1(s_2 = x_1 \,|\, s_1 = x_0, a_1 = b_1)$. For the reward function, we set

$$\begin{aligned}
&r_1(x_0, a) = 0, \quad \text{for all} \;\; a \in \mathcal{A},\\
&r_h(x_1, a) = 1, \quad r_h(x_2, a) = 0, \quad \text{for all} \;\; a \in \mathcal{A}, \; h \in \{2, \ldots, H\}
\end{aligned} \tag{5.8}$$

See Figure 4 for an illustration. Note that $M(p_1, p_2, p_3)$ is a linear MDP, which is defined in Definition 3 with the dimension $d = A + 2$. To see this, we set the corresponding feature map $\phi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ as

$$\begin{aligned}
&\phi(x_0, b_j) = (\boldsymbol{e}_j, 0, 0) \in \mathbb{R}^{A+2}, \quad \text{for all} \;\; j \in [A]\\
&\phi(x_1, a) = (\mathbf{0}_A, 1, 0) \in \mathbb{R}^{A+2}, \quad \text{for all} \;\; a \in \mathcal{A},\\
&\phi(x_2, a) = (\mathbf{0}_A, 0, 1) \in \mathbb{R}^{A+2}, \quad \text{for all} \;\; a \in \mathcal{A}
\end{aligned} \tag{5.9}$$

where $\{\boldsymbol{e}_j\}_{j=1}^A$ and $\mathbf{0}_A$ are the canonical basis and zero vector in $\mathbb{R}^A$, respectively.

As $x_1, x_2 \in \mathcal{S}$ are the absorbing states, the optimal policy π_1^* at the first step $h = 1$ is a deterministic policy, which by Equation (5.8) selects the action $a \in \mathcal{A}$ that induces the largest transition probability into the desired state x_1. In other words, at the first step $h = 1$, we have

$$\pi_1^*(b_{j^*} \,|\, x_0) = 1, \quad \text{where} \;\; j^* = \mathop{\arg\max}_{j \in \{1,2\}} p_j \tag{5.10}$$

Here we assume without loss of generality $p_1 \neq p_2$ in Equation (5.7). Meanwhile, at any subsequent step $h \in \{2, \ldots, H\}$, an arbitrary policy π_h is optimal, as the action $a \in \mathcal{A}$ selected by π_h does not affect the transition probability. Therefore, for any policy $\pi = \{\pi_h\}_{h=1}^H$, the suboptimality of π for the linear MDP $\mathcal{M} = M(p_1, p_2, p_3)$ takes the form

$$\mathrm{SubOpt}(\mathcal{M}, \pi; x_0) = p_{j^*} \cdot (H - 1) - \sum_{j=1}^{A} p_j \cdot \pi_1(b_j \,|\, x_0) \cdot (H - 1) \tag{5.11}$$

where for notational simplicity, we define $p_j = p_3$ for all $j \in \{3, \ldots, A\}$. Here with an abuse of notation, we incorporate the explicit dependency on the underlying MDP $\mathcal{M} \in \mathfrak{M}$ into the suboptimality $\mathrm{SubOpt}(\pi; x_0)$.

Dataset: We specify the worst-case data collecting process $\mathbb{P}_{\mathcal{D}}$ as follows. Given a linear MDP $\mathcal{M} \in \mathfrak{M}$, the dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ consists of K trajectories starting from the same initial state x_0, that is, $x_1^\tau = x_0$ for all $\tau \in [K]$. The initial actions $\{a_1^\tau\}_{\tau=1}^K$ are predetermined.

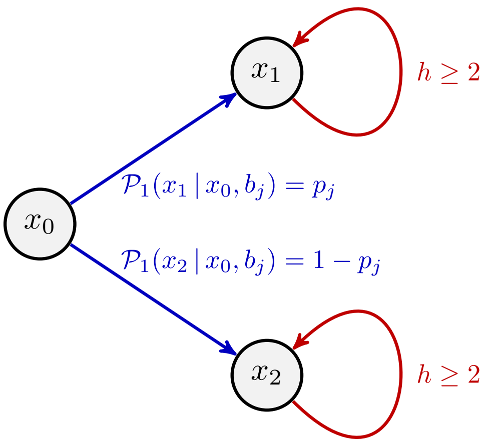
x_1
$h \geq 2$
$\mathcal{P}_1(x_1 \,|\, x_0, b_j) = p_j$
x_0
$\mathcal{P}_1(x_2 \,|\, x_0, b_j) = 1 - p_j$
x_2
$h \geq 2$

Figure 4. An illustration of the episodic MDP $\mathcal{M} = M(p_1, p_2, p_3) \in \mathfrak{M}$ with the state space $\mathcal{S} = \{x_0, x_1, x_2\}$ and action space $\mathcal{A} = \{b_j\}_{j=1}^A$. Here we fix the initial state as $s_1 = x_0$, where the agent takes the action $a \in \mathcal{A}$ and transits into the second state $s_2 \in \{x_1, x_2\}$. At the first step $h = 1$, the transition probability satisfies $\mathcal{P}_1(x_1 \,|\, x_0, b_j) = p_j$ and $\mathcal{P}_1(x_2 \,|\, x_0, b_j) = 1 - p_j$ for all $j \in [A]$, where for notational simplicity, we define $p_j = p_3$ for all $j \in \{3, \ldots, A\}$. Also, $x_1, x_2 \in \mathcal{S}$ are the absorbing states. Meanwhile, the reward function satisfies $r_h(x_0, a) = 0$, $r_h(x_1, a) = 1$, and $r_h(x_2, a) = 0$ for all $a \in \mathcal{A}$ and $h \in [H]$.

The subsequent states $\{x_h^\tau\}_{\tau \in [K], h \geq 2}$ are sampled from the underlying MDP $\mathcal{M} = M(p_1, p_2, p_3)$, while the subsequent actions $\{a_h^\tau\}_{\tau \in [K], h \geq 2}$ are arbitrarily chosen, as they do not affect the state transitions. The state transitions in the different trajectories are independent. The immediate reward r_h^τ satisfies $r_h^\tau = r_h(x_h^\tau, a_h^\tau)$. Note that such a dataset $\mathcal{D}$ satisfies Assumption 1, that is, $\mathcal{D}$ is compliant with the linear MDP $\mathcal{M} \in \mathfrak{M}$.

We define

$$n_j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = b_j\}, \quad \{\kappa_j^i\}_{i=1}^{n_j} = \big\{r_2^\tau \colon a_1^\tau = b_j \text{ with } \tau \in [K]\big\}, \quad \text{for all } \ j \in [A] \tag{5.12}$$

In other words, assuming that $1 \leq \tau_1 < \tau_2 < \cdots < \tau_{n_j} \leq K$ are the episode indices such that $a_1^{\tau_i} = b_j$ for all $i \in [n_j]$, we define $\kappa_j^i = r_2^{\tau_i}$ for all $j \in [A]$. By such a construction, $\{\kappa_j^i\}_{i,j=1}^{n_j, A}$ are the realizations of K independent Bernoulli random variables, which satisfy

$$\mathbb{E}_{\mathcal{D}}[\kappa_j^i] = p_j, \quad \text{for all } \ i \in [n_j], \ j \in [A] \tag{5.13}$$

Note that knowing the value of the immediate reward r_2^τ is sufficient for determining the value of the second state x_2^τ. Meanwhile, recall that $x_1, x_2 \in \mathcal{S}$ are the absorbing states. Therefore, for learning the optimal policy π^*, the original dataset $\mathcal{D}$ contains the same information as the reduced dataset $\mathcal{D}_1 = \{(x_1^\tau, a_1^\tau, x_2^\tau, r_2^\tau)\}_{\tau=1}^K$, where the randomness only comes from the state transition at the first step $h = 1$ of each trajectory $\tau \in [K]$. Correspondingly, the probability of observing the dataset $\mathcal{D}_1$ takes the form

$$\mathbb{P}_{\mathcal{D} \sim \mathcal{M}}(\mathcal{D}_1) = \prod_{\tau=1}^K \mathbb{P}_{\mathcal{M}}\big(r_2(s_2, a_2) = r_2^\tau \,\big|\, s_1 = x_1^\tau, a_1 = a_1^\tau\big)$$

$$= \prod_{j=1}^{A} \Big(\prod_{i=1}^{n_j} p_j^{\kappa_j^i} \cdot (1-p_j)^{1-\kappa_j^i} \Big) = \prod_{j=1}^{A} \Big(p_j^{\sum_{i=1}^{n_j} \kappa_j^i} \cdot (1-p_j)^{n_j - \sum_{i=1}^{n_j} \kappa_j^i} \Big) \tag{5.14}$$

Here $\mathbb{P}_{\mathcal{D}\sim\mathcal{M}}$ denotes the randomness of the dataset $\mathcal{D}$, which is compliant with the underlying MDP $\mathcal{M} = M(p_1, p_2, p_3)$, while $\mathbb{P}_{\mathcal{M}}$ denotes the randomness of the immediate rewards and state transitions. In the second equality, we apply the definition of $\{\kappa_j^i\}_{i=1}^{n_j}$ in Equation (5.12). By such a definition, $\mathbb{P}_{\mathcal{D}\sim\mathcal{M}}(\mathcal{D}_1)$ in Equation (5.14) is the likelihood of the linear MDP $\mathcal{M} \in \mathfrak{M}$ given the reduced dataset $\mathcal{D}_1$ (or equivalently, the original dataset $\mathcal{D}$, assuming that the subsequent actions $\{a_h^\tau\}_{\tau\in[K],h\geq 2}$ are predetermined).

5.3.2 Information-Theoretic Lower Bound

The proof of Theorem 4.3 is based on the Le Cam method [LC12, Yu97]. Specifically, we construct two linear MDPs $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$, where the class $\mathfrak{M}$ of linear MDPs is defined in Equation (5.6). Such a construction ensures that (i) the distribution of the dataset $\mathcal{D}$, which is compliant with the underlying MDP, is similar across $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$, and (ii) the suboptimality of any policy π, which is constructed based on the dataset, is different across $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$. In other words, it is hard to distinguish $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$ based on $\mathcal{D}$, while π obtained from $\mathcal{D}$ can not achieve a desired suboptimality for $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$ simultaneously. Such a construction captures the fundamental hardness of offline RL for the linear MDP.

For any $p, p^* \in [1/4, 3/4]$, where $p < p^*$, we set

$$\mathcal{M}_1 = M(p^*, p, p), \quad \mathcal{M}_2 = M(p, p^*, p) \tag{5.15}$$

Based on $\mathcal{D}$, whose likelihood is specified in Equation (5.14), we aim to test whether the underlying MDP is $\mathcal{M}_1$ or $\mathcal{M}_2$. The following lemma establishes a reduction from learning the optimal policy π^* to testing the underlying MDP $\mathcal{M} \in \mathfrak{M}$. Recall that for any $\ell \in \{1, 2\}$, n_ℓ is defined in Equation (5.12).

Lemma 4 (Reduction to Testing)**.** *For the dataset $\mathcal{D}$ specified in Section 5.3.1, the output $\texttt{Algo}(\mathcal{D})$ of any algorithm satisfies*

$$\begin{aligned} &\max_{\ell\in\{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Big[\mathit{SubOpt}\big(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\big)\Big] \\ &\qquad \geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot (p^* - p) \cdot (H-1) \cdot \Big(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big]\Big) \end{aligned}$$

where $\pi = \{\pi_h\}_{h=1}^H = \texttt{Algo}(\mathcal{D})$. For any $\ell \in \{1, 2\}$, $\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}$ is taken with respect to the randomness of $\mathcal{D}$, which is compliant with the underlying MDP $\mathcal{M}_\ell$

Proof of Lemma 4. See Appendix C.1 for a detailed proof. □

As specified in Equation (5.10), for the underlying MDP $\mathcal{M}_1$, the optimal policy π_1^* takes the initial action b_1 with probability one at the initial state x_0, while for $\mathcal{M}_2$, π_1^* takes b_2 with probability one at x_0. We consider the following hypothesis testing problem

$$H_0\colon \mathcal{M} = \mathcal{M}_1 \quad \text{versus} \quad H_1\colon \mathcal{M} = \mathcal{M}_2 \tag{5.16}$$

based on the dataset $\mathcal{D}$. For such a problem, any test function ψ is a binary map such that $\psi(\mathcal{D}) = 0$ means the null hypothesis H_0 is accepted, while $\psi(\mathcal{D}) = 1$ means H_0 is rejected. For the output $\pi = \{\pi_h\}_{h=1}^H = \texttt{Algo}(\mathcal{D})$ of any algorithm, we define

$$\psi_{\texttt{Algo}}(\mathcal{D}) = \mathbb{1}\{a \neq b_1\}, \quad \text{where} \;\; a \sim \pi_1(\cdot \,|\, x_0) \tag{5.17}$$

Correspondingly, the risk of the (randomized) test function $\psi_{\texttt{Algo}}$ takes the form

$$\begin{aligned} \mathrm{Risk}(\psi_{\texttt{Algo}}) &= \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[\mathbb{1}\{\psi_{\texttt{Algo}}(\mathcal{D}) = 1\}\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\mathbb{1}\{\psi_{\texttt{Algo}}(\mathcal{D}) = 0\}\big] \\ &= \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big] \end{aligned} \tag{5.18}$$

Therefore, Lemma 4 lower bounds the suboptimality of any policy $\pi = \{\pi_h\}_{h=1}^H = \texttt{Algo}(\mathcal{D})$ by the risk of a (randomized) test function, which is induced by π, for the corresponding hypothesis testing problem defined in Equation (5.16). Such an approach mirrors the Le Cam method [LC12, Yu97] for establishing the minimax optimality in statistical estimation. In particular, a careful choice of $p, p^* \in [1/4, 3/4]$ leads to the information-theoretic lower bound established in Theorem 4.3. See Appendix C.3 for a detailed proof.

6 Conclusion

We study offline reinforcement learning (RL), which aims to learn an optimal policy based on a dataset collected a priori. Due to the lack of further interactions with the environment, offline RL suffers from the insufficient coverage of the dataset, which eludes most existing theoretical analysis. In this paper, we propose a pessimistic variant of the value iteration algorithm (PEVI), which incorporates an uncertainty quantifier as the penalty function. Such a penalty function simply flips the sign of the bonus function for promoting exploration in online RL, which makes it easily implementable and compatible with general function approximators.

Without assuming the sufficient coverage of the dataset (e.g., finite concentrability coefficients or uniformly lower bounded densities of visitation measures), we establish a data-dependent upper bound on the suboptimality of PEVI for general Markov decision processes (MDPs). When specialized to linear MDPs, it matches the information-theoretic lower bound up to multiplicative factors of the dimension and horizon. In other words, pessimism is not only provably efficient but also minimax optimal. In particular, given the dataset, the learned policy serves as the "best effort" among all policies, as no other policies can do better. Our theoretical analysis identifies the critical role of pessimism in eliminating a notion of spurious correlation, which arises from the "irrelevant" trajectories that are less covered by the dataset and not informative for the optimal policy.

Our results underscore the importance of addressing uncertainty and spurious correlations in offline RL. PEVI provides a robust framework for learning optimal policies without exploration by carefully incorporating penalties for uncertain states and actions. Its minimax optimality for linear MDPs highlights its effectiveness under practical constraints, making it a valuable tool for offline RL tasks. Future research could extend these principles to more complex environments and function approximation settings.

References

[AJK20] Alekh Agarwal, Nan Jiang, and Sham M Kakade. *Reinforcement learning: Theory and algorithms.* MIT, 2020.

[AJS$^+$20] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

[ASM07] András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In *Advances in Neural Information Processing Systems*, 2007.

[ASM08] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

[ASN20] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114, 2020.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[BGB20] Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

[BNVB13] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[CG17] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. *arXiv preprint arXiv:1704.00445*, 2017.

[CJ19] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

[CM14] Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, 1:447–464, 2014.

[CYJW20] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294, 2020.

[DJ94] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[DJW20] 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, 2020.

[FCG18] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456, 2018.

[FGSM16] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

[FKN$^+$20] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[FL01] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[FMP19] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

[FSM10] Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

[FWXY20] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pages 486–489, 2020.

[FYW20] Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*, 2020.

[GJK+19] O Gottesman, F Johansson, M Komorowski, A Faisal, D Sontag, F Doshi-Velez, and LA Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–32, 2019.

[GWN+20] Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gómez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, et al. RL Unplugged: Benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*, 2020.

[HCD+16] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, 2016.

[JAZBJ18] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[JGS+19] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

[JH20] Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. In *Advances in Neural Information Processing Systems*, 2020.

[JL16] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661, 2016.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4):8–36, 2010.

[JYWJ20] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[KFS+19] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11784–11794, 2019.

[KRNJ20] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

[KU19] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

[KU20] Nathan Kallus and Masatoshi Uehara. Doubly robust off-policy value and gradient estimation for deterministic policies. *arXiv preprint arXiv:2006.03900*, 2020.

[KZTL20] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[LC12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer, 2012.

[LCYW19] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, pages 10565–10576, 2019.

[LGR12] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

[LKTF20] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[LLTZ18] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, 2018.

[LQM20] Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward Markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge, 2020.

[LSAB20] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

[LTDC19] Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661, 2019.

[LWC$^+$20] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*, 2020.

[MKS$^+$15] 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.

[MS08] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

[NCDL19] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019.

[ND20] Ofir Nachum and Bo Dai. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[NDGL20] Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

[NDK$^+$19] 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.

[OVR14] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, 2014.

[QW20] Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. *arXiv preprint arXiv:2002.00260*, 2020.

[RVR13] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

[RVR16] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[RVR18] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT, 2018.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

[Sch91] Jürgen Schmidhuber. Curious model-building control systems. In *International Joint Conference on Neural Networks*, pages 1458–1463, 1991.

[Sch10] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[SGG$^+$15] Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16:1629–1676, 2015.

[SGS11] Yi Sun, Faustino Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal Bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence*, pages 41–51, 2011.

[SHM$^+$16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[SKD$^+$20] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[SKKS09] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[SP12] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

[SSB$^+$20] Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

[SSS^{+}17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

[SSSS16] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[Sze10] Csaba Szepesvári. *Algorithms for reinforcement learning.* Morgan & Claypool, 2010.

[TB16] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

[TET12] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

[TFL^{+}19] Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, May 2015.

[UHJ20] 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, 2020.

[VEB^{+}17] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. StarCraft II: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[WCYW19] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

[WFK20] Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline RL with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

[WNZ^{+}20] Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. In *Advances in Neural Information Processing Systems*, 2020.

[WSY20] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, 2020.

[WTN19] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[XJ20a] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020.

[XJ20b] Tengyang Xie and Nan Jiang. $Q^{\star}$-approximation schemes for batch reinforcement learning: A theoretical comparison. *arXiv preprint arXiv:2003.03924*, 2020.

[XMW19] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pages 9668–9678, 2019.

[YBW20] Ming Yin, Yu Bai, and Yu-Xiang Wang. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*, 2020.

[YJW$^+$20a] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv preprint arXiv:2011.04622*, 2020.

[YJW$^+$20b] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

[YND$^+$20] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized Lagrangian. In *Advances in Neural Information Processing Systems*, 2020.

[YTY$^+$20] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[YW19] Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.

[YW20] 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, 2020.

[ZCA21] Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

[ZDLS20] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDICE: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020.

[ZKB$^+$20] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, 2020.

[Zou06] Hui Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.