Is Pessimism Provably Efficient for Offline RL?

Ying Jin [*1] **Zhuoran Yang** [*2] **Zhaoran Wang** [*3]

Abstract

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, 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 emerges from the "irrelevant" trajectories that are less covered by the dataset and not informative for the optimal policy.

1. Introduction

The empirical success of online (deep) reinforcement learning (RL) (Mnih et al., 2015; Silver et al., 2016; 2017; Vinyals et al., 2017) relies on two ingredients: (i) expressive function approximators, e.g., deep neural networks (LeCun et al., 2015), which approximate policies and values, and (ii) efficient data generators, e.g., game engines (Bellemare et al., 2013) and physics simulators (Todorov et al., 2012), 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 (Gottesman et al., 2019) and autonomous driving (Shalev-Shwartz et al., 2016), 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 (Chakraborty and Murphy, 2014) and human driving trajectories for autonomous driving (Sun et al., 2020).

In comparison with online RL (Lattimore and Szepesvári, 2020; Agarwal et al., 2020a), offline RL remains even less understood in theory (Lange et al., 2012; Levine et al., 2020), 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 (Wang et al., 2020a). 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;

(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

[*]Equal contribution [1]Department of Statistics, Stanford University [2]Department of Operations Research and Financial Engineering, Princeton University [3]Department of Industrial Engineering and Management Sciences, Northwestern University . Correspondence to: Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

dataset possibly fails to hold in practice (Fujimoto et al., 2019; Agarwal et al., 2020b; Fu et al., 2020a; Gulcehre et al., 2020).

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 (Sutton and Barto, 2018; Szepesvári, 2010). Here the penalty function simply flips the sign of the bonus function (optimism) for promoting exploration in online RL (Jaksch et al., 2010; Azar et al., 2017), 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 in Section 3, even in the multi-armed bandit (MAB), a special case of the MDP, illustrated in Appendix A.

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

(iii) For the linear MDP (Yang and Wang, 2019; Jin et al., 2020), we instantiate PEVI by specifying the penalty function in Section 4.2. 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 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. See Section 4.3 for discussions.

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 (Chen and Jiang, 2019) and uniformly lower bounded densities of visitation measures (Yin et al., 2020), 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) (Liu et al., 2020), 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) (Lange et al., 2012; Levine et al., 2020), 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 (Wang et al., 2020a), which arises from the lack of continuing exploration (Szepesvári, 2010). 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 (Levine et al., 2020). As a result, intertwined with over-parameterized function approximators, e.g., deep neural networks, offline RL possibly suffers from the extrapolation error (Fujimoto et al., 2019), 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 (Sutton and Barto, 2018).

To address such a challenge, the recent works (Fujimoto et al., 2019; Laroche et al., 2019; Jaques et al., 2019; Wu et al., 2019; Kumar et al., 2019; 2020; Agarwal et al., 2020b; Yu et al., 2020; Kidambi et al., 2020; Wang et al., 2020c; Siegel et al., 2020; Nair et al., 2020; Liu et al., 2020) 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 (Levine et al., 2020). 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., (Jiang and Li, 2016; Thomas and Brunskill, 2016; Farajtabar et al., 2018; Liu et al., 2018; Xie et al., 2019; Nachum et al., 2019a;b; Tang et al., 2019; Kallus and Uehara, 2019; 2020; Jiang and Huang, 2020; Uehara et al., 2020; Duan et al., 2020; Yin and Wang, 2020; Yin et al., 2020; Nachum and Dai, 2020; Yang et al., 2020a; Zhang et al., 2020b) 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., (Antos et al., 2007; 2008; Munos and Szepesvári, 2008; Farahmand et al., 2010; 2016; Scherrer et al., 2015; Chen and Jiang, 2019; Liu et al., 2019; Wang et al., 2019; Fu et al., 2020b; Fan et al., 2020; Xie and Jiang, 2020a;b; Liao et al., 2020; Zhang et al., 2020a) and the references therein.

In practice, such assumptions on the sufficient coverage of the dataset often fail to hold (Fujimoto et al., 2019; Agarwal et al., 2020b; Fu et al., 2020a; Gulcehre et al., 2020), 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 (Yu et al., 2020; Kidambi et al., 2020; Kumar et al., 2020; Liu et al., 2020; Buckman et al., 2020). Specifically, (Yu et al., 2020; Kidambi et al., 2020) propose a pessimistic model-based approach, while (Kumar et al., 2020) propose a pessimistic value-based approach, both of which demonstrate empirical successes. From a theoretical perspective, (Liu et al., 2020) propose a regularized (and pessimistic) variant of the fitted Q-iteration algorithm (Antos et al., 2007; 2008; Munos and Szepesvári, 2008), 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 (Yang and Wang, 2019; Jin et al., 2020) (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, (Buckman et al., 2020) 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 (Buckman et al., 2020) 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

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

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

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

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

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

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

$$(\mathbb{B}_h f)(x,a) = \mathbb{E}\big[r_h(s_h,a_h) + f(s_{h+1}) \,\big|\, s_h = x, a_h = a\big]. \quad (2.4)$$

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

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

Meanwhile, the optimal policy π^* is specified by

$$\pi_h^*(\cdot \,|\, x) = \operatorname*{argmax}_{\pi_h} \langle Q_h^*(x,\cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}},$$
$$V_h^*(x) = \langle Q_h^*(x,\cdot), \pi_h^*(\cdot \,|\, x)\rangle_{\mathcal{A}},$$

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

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

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 $x_h^\tau \sim \mathcal{P}_h(\cdot \,|\, s_h = x_h^\tau, a_h = a_h^\tau)$. 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 2.1 (Compliance of Dataset)**.** 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

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

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 $h \in [H]$, $\{(x_h^\tau, a_h^\tau, r_h^\tau, x_{h+1}^\tau)\}_{\tau\in[K]}$ 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 2.2 (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 2.2 holds if the experimenter follows a fixed behavior policy. More generally, Assumption 2.2 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 2.2 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, the latter deferred to Appendix A.

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 3.1 (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 3.1. See Appendix D. □

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 A, 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, $\widehat{\pi}_h(\cdot \,|\, x) = \mathrm{argmax}_{\pi_h} \langle \widehat{Q}_h(x, \cdot), \pi_h(\cdot \,|\, x) \rangle_{\mathcal{A}}$ (although Equation (3.2) holds for any $\widehat{\pi}$ such that $\widehat{V}_h(x) = \langle \widehat{Q}_h(x, \cdot), \widehat{\pi}_h(\cdot \,|\, x) \rangle_{\mathcal{A}}$).

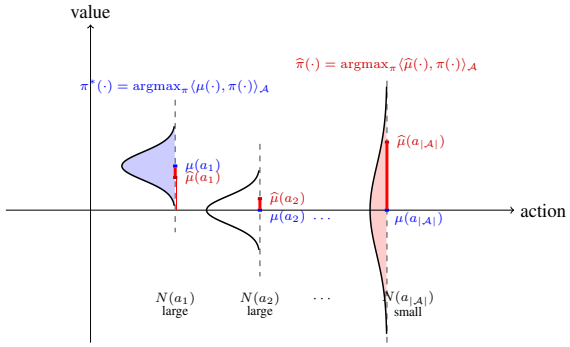

Figure 1. An illustration of the spurious correlation in the MAB, a special case of the MDP, where $\mathcal{S}$ is a singleton, $\mathcal{A}$ is discrete, and $H = 1$. Here $\mu(a)$ is the expected reward of each action $a \in \mathcal{A}$ and $\widehat{\mu}(a)$ is its sample average estimator, which follows the Gaussian distribution in Equation (A.1). 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 $a_{|\mathcal{A}|} = \mathrm{argmax}_{a \in \mathcal{A}} \widehat{\mu}(a)$ 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 $a_1 = \mathrm{argmax}_{a \in \mathcal{A}} \mu(a)$ with probability one.

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 (Jaksch et al., 2010; Abbasi-Yadkori et al., 2011; Russo and Van Roy, 2013; Osband and Van Roy, 2014; Chowdhury and Gopalan, 2017; Azar et al., 2017; Jin et al., 2018; 2020; Cai et al., 2020; Yang et al., 2020b; Ayoub et al., 2020; Wang et al., 2020b). 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 (Yang and Wang, 2019; Jin et al., 2020) 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 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 4.1 (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

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

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

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).

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 $\mathcal{D}$.
  Set $\overline{Q}_h(\cdot,\cdot)\leftarrow(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\cdot,\cdot)-\Gamma_h(\cdot,\cdot)$.
  Set $\widehat{Q}_h(\cdot,\cdot)\leftarrow\min\{\overline{Q}_h(\cdot,\cdot), H-h+1\}^+$.
  Set $\widehat{\pi}_h(\cdot\,|\,\cdot)\leftarrow\operatorname{argmax}_{\pi_h}\langle\widehat{Q}_h(\cdot,\cdot),\pi_h(\cdot\,|\,\cdot)\rangle_{\mathcal{A}}$.
  Set $\widehat{V}_h(\cdot)\leftarrow\langle\widehat{Q}_h(\cdot,\cdot),\widehat{\pi}_h(\cdot\,|\,\cdot)\rangle_{\mathcal{A}}$.
end for
Output: $\texttt{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=1}^H$.

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

Theorem 4.2 (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 $\mathbb{P}_{\mathcal{D}}(\mathcal{E})\ge 1-\xi$, for any $x\in\mathcal{S}$, $\texttt{Pess}(\mathcal{D})$ in Algorithm 1 satisfies

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

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

Proof of Theorem 4.2. See Section C.1 for a sketch. □

Theorem 4.2 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,

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

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 2.2. 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 (Yang and Wang, 2019; Jin et al., 2020) as follows, where the transition kernel and expected reward function are linear in a feature map.

Definition 4.3 (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

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

for all $(x,a,x')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ at each step $h\in[H]$. Here we assume $\|\phi(x,a)\|\le 1$ for all $(x,a)\in\mathcal{S}\times\mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|,\|\theta_h\|\}\le\sqrt{d}$ 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 empiri-

cal 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

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

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

$$\widehat{w}_h = \Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)\big)\Big),$$
$$\text{where } \Lambda_h = \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I. \tag{4.6}$$

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

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

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

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

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 $\widehat{w}_h \leftarrow \Lambda_h^{-1}(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)))$.
  Set $\Gamma_h(\cdot,\cdot) \leftarrow \beta \cdot (\phi(\cdot,\cdot)^\top \Lambda_h^{-1} \phi(\cdot,\cdot))^{1/2}$.
  Set $\overline{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{w}_h - \Gamma_h(\cdot,\cdot)$.
  Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\{\overline{Q}_h(\cdot,\cdot), H - h + 1\}^+$.
  Set $\widehat{\pi}_h(\cdot \,|\, \cdot) \leftarrow \operatorname{argmax}_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \,|\, \cdot)\rangle_{\mathcal{A}}$.
  Set $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \,|\, \cdot)\rangle_{\mathcal{A}}$.
end for
Output: $\texttt{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=1}^{H}$.

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

Theorem 4.4 (Suboptimality for Linear MDP)**.** Suppose Assumption 2.2 holds and the underlying MDP is a linear MDP. In Algorithm 2, we set

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

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

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

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

Proof of Theorem 4.4. See Section C.2. □

We highlight the following aspects of Theorem 4.4:

"Assumption-Free" Guarantee: Theorem 4.4 only relies on the compliance of $\mathcal{D}$ with the linear MDP. In comparison with existing literature (Antos et al., 2007; 2008; Munos and Szepesvári, 2008; Farahmand et al., 2010; 2016; Scherrer et al., 2015; Liu et al., 2018; Nachum et al., 2019a;b; Chen and Jiang, 2019; Tang et al., 2019; Kallus and Uehara, 2019; 2020; Fan et al., 2020; Xie and Jiang, 2020a;b; Jiang and Huang, 2020; Uehara et al., 2020; Duan et al., 2020; Yin et al., 2020; Qu and Wierman, 2020; Li et al., 2020; Liao et al., 2020; Nachum and Dai, 2020; Yang et al., 2020a; Zhang et al., 2020a;b), 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 (Fujimoto et al., 2019; Laroche et al., 2019; Jaques et al., 2019; Wu et al., 2019; Kumar et al., 2019; Wang et al., 2020c; Siegel et al., 2020; Nair et al., 2020; Liu et al., 2020).

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

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

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

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

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 $\text{rank}(\Sigma_h(x)) \leq r$ 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

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

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 4.5. See Appendix E.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. Note that the estimator $\widehat{w}_h$ in Equation (4.6) is the Bayesian estimator of w_h, under prior $w_h \sim \text{N}(0, \lambda \cdot I)$ and Gaussian response with variance one. Within this equivalent Bayesian framework, the posterior has the closed form

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

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

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

where $\Lambda_h^\dagger = \Lambda_h + \phi(s_h, a_h)\phi(s_h, a_h)^\top$. Here I is the (conditional) mutual information and H is the (conditional) differential entropy. Meanwhile, we have

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

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

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

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 (Schmidhuber, 1991; 2010; Sun et al., 2011; Still and Precup, 2012; Houthooft et al., 2016; Russo and Van Roy, 2016; 2018) 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 A. Specifically, in the tabular MDP, we have

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

where $N_h(x, a) = \sum_{\tau=1}^{K} \mathbb{1}\{(x_h^\tau, a_h^\tau) = (x, a)\}$. 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

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

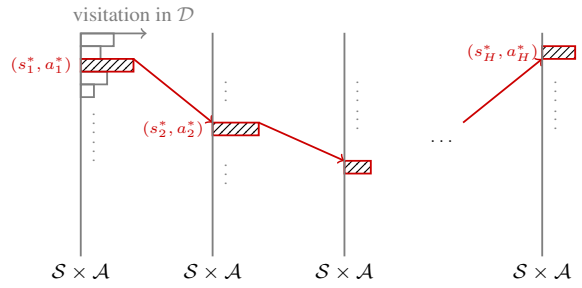

Figure 2. An illustration of the oracle property in the tabular MDP, where the transition kernel is deterministic. The histogram depicts the number of times (s_h, a_h) is visited in $\mathcal{D}$. The suboptimality in Equation (4.8) only depends on the number of times (s_h^*, a_h^*), which lies on the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

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 (Donoho and Johnstone, 1994; Fan and Li, 2001; Zou, 2006), 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 $\bar{\pi}$ (namely the demonstration), such an oracle property allows $\texttt{Pess}(\mathcal{D})$ to outperform $\bar{\pi}$ 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 $\bar{\pi}$ 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 (Duan et al., 2020), Corollary B.1 specializes Theorem 4.4 under the additional assumption that the data collecting process well explores $\mathcal{S}$ and $\mathcal{A}$. The suboptimality in Equation (B.1) parallels the policy evaluation error established in (Duan et al., 2020), 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 (B.1) goes to zero.

4.3. Minimax Optimality: Information-Theoretic Lower Bound

We establish the minimax optimality of Theorems 4.2 and 4.4 via the following information-theoretic lower bound. Recall that $\mathbb{P}_{\mathcal{D}}$ is the joint distribution of the data collecting process.

Theorem 4.6 (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

$$\mathbb{E}_{\mathcal{D}}\left[\frac{\mathrm{SubOpt}\big(\texttt{Algo}(\mathcal{D}); x\big)}{\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big]}\right] \geq c, \tag{4.13}$$

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 4.6. See Section C.3 for a proof sketch and Appendix F.3 for a detailed proof. □

Theorem 4.6 matches Theorem 4.4 up to β and absolute constants. Although Theorem 4.6 only establishes the minimax optimality, Proposition F.2 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 F.4 for a detailed discussion.

Acknowledgement

The authors would like to thank Alekh Agarwal, Lin Yang, Yining Wang, Chi Jin, Mengdi Wang, Yaqi Duan, and Csaba Szepesvári for helpful discussions. Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

Agarwal, A., Jiang, N. and Kakade, S. M. (2020a). *Reinforcement learning: Theory and algorithms*. MIT.

Agarwal, R., Schuurmans, D. and Norouzi, M. (2020b). An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*.

Antos, A., Szepesvári, C. and Munos, R. (2007). Fitted Q-iteration in continuous action-space MDPs. In *Advances in Neural Information Processing Systems*.

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

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M. and Yang, L. F. (2020). Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*.

Azar, M. G., Osband, I. and Munos, R. (2017). Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*.

Bellemare, M. G., Naddaf, Y., Veness, J. and Bowling, M. (2013). The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, **47** 253–279.

Buckman, J., Gelada, C. and Bellemare, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*.

Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, **1** 447–464.

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

Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. *arXiv preprint arXiv:1704.00445*.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81** 425–455.

Duan, Y., Jia, Z. and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96** 1348–1360.

Fan, J., Wang, Z., Xie, Y. and Yang, Z. (2020). A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C. and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, **17** 4809–4874.

Farahmand, A.-m., Szepesvári, C. and Munos, R. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*.

Farajtabar, M., Chow, Y. and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*.

Fu, J., Kumar, A., Nachum, O., Tucker, G. and Levine, S. (2020a). D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

Fu, Z., Yang, Z. and Wang, Z. (2020b). Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*.

Fujimoto, S., Meger, D. and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F. and Celi, L. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, **25** 16–32.

Gulcehre, C., Wang, Z., Novikov, A., Paine, T. L., Colmenarejo, S. G., Zolna, K., Agarwal, R., Merel, J., Mankowitz, D., Paduraru, C. et al. (2020). RL Unplugged: Benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F. and Abbeel, P. (2016). VIME: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*.

Jaksch, T., Ortner, R. and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, **11** 8–36.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S. and Picard, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. In *Advances in Neural Information Processing Systems*.

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

Jin, C., Allen-Zhu, Z., Bubeck, S. and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*.

Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*.

Kallus, N. and Uehara, M. (2019). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*.

Kallus, N. and Uehara, M. (2020). Doubly robust off-policy value and gradient estimation for deterministic policies. *arXiv preprint arXiv:2006.03900*.

Kidambi, R., Rajeswaran, A., Netrapalli, P. and Joachims, T. (2020). MOReL: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*.

Kumar, A., Fu, J., Soh, M., Tucker, G. and Levine, S. (2019). Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*.

Kumar, A., Zhou, A., Tucker, G. and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*.

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

Laroche, R., Trichelair, P. and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge.

Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521** 436–444.

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

Li, G., Wei, Y., Chi, Y., Gu, Y. and Chen, Y. (2020). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*.

Liao, P., Qi, Z. and Murphy, S. (2020). Batch policy learning in average reward Markov decision processes. *arXiv preprint arXiv:2007.11771*.

Liu, B., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*.

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

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al. (2015). Human-level control through deep reinforcement learning. *Nature*, **518** 529–533.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, **9** 815–857.

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

Nachum, O. and Dai, B. (2020). Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L. and Schuurmans, D. (2019b). AlgaeDICE: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*.

Nair, A., Dalal, M., Gupta, A. and Levine, S. (2020). Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*.

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. *arXiv preprint arXiv:2002.00260*.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, **17** 2442–2471.

Russo, D. and Van Roy, B. (2018). Learning to optimize via information-directed sampling. *Operations Research*, **66** 230–252.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B. and Geist, M. (2015). Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, **16** 1629–1676.

Schmidhuber, J. (1991). Curious model-building control systems. In *International Joint Conference on Neural Networks*.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, **2** 230–247.

Shalev-Shwartz, S., Shammah, S. and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Siegel, N. Y., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R. and Riedmiller, M. (2020). Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V. and Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, **529** 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M. and Bolton, A. (2017). Mastering the game of Go without human knowledge. *Nature*, **550** 354.

Still, S. and Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, **131** 139–148.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B. et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Computer Vision and Pattern Recognition*.

Sun, Y., Gomez, F. and Schmidhuber, J. (2011). Planning to be surprised: Optimal Bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence*.

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

Szepesvári, C. (2010). *Algorithms for reinforcement learning*. Morgan & Claypool.

Tang, Z., Feng, Y., Li, L., Zhou, D. and Liu, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*.

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

Todorov, E., Erez, T. and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems*.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, **8** 1–230.

Uehara, M., Huang, J. and Jiang, N. (2020). Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*.

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J. et al. (2017). StarCraft II: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*.

Wang, L., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*.

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

Wang, R., Salakhutdinov, R. R. and Yang, L. (2020b). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*.

Wang, Z., Novikov, A., Zolna, K., Merel, J. S., Springenberg, J. T., Reed, S. E., Shahriari, B., Siegel, N., Gulcehre, C., Heess, N. et al. (2020c). Critic regularized regression. In *Advances in Neural Information Processing Systems*.

Wu, Y., Tucker, G. and Nachum, O. (2019). Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

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

Xie, T. and Jiang, N. (2020b). $Q^\star$-approximation schemes for batch reinforcement learning: A theoretical comparison. *arXiv preprint arXiv:2003.03924*.

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

Yang, L. and Wang, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*.

Yang, M., Nachum, O., Dai, B., Li, L. and Schuurmans, D. (2020a). Off-policy evaluation via the regularized Lagrangian. In *Advances in Neural Information Processing Systems*.

Yang, Z., Jin, C., Wang, Z., Wang, M. and Jordan, M. I. (2020b). Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv preprint arXiv:2011.04622*.

Yin, M., Bai, Y. and Wang, Y.-X. (2020). Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*.

Yin, M. and Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*.

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*. Springer, 423–435.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C. and Ma, T. (2020). MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvári, C. and Wang, M. (2020a). Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*.

Zhang, R., Dai, B., Li, L. and Schuurmans, D. (2020b). GenDICE: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101** 1418–1429.