On The Statistical Complexity of Offline Decision-Making

Thanh Nguyen-Tang [1] **Raman Arora** [1]

Abstract

We study the statistical complexity of offline decision-making with function approximation, establishing (near) minimax-optimal rates for stochastic contextual bandits and Markov decision processes. The performance limits are captured by the pseudo-dimension of the (value) function class and a new characterization of the behavior policy that *strictly* subsumes all the previous notions of data coverage in the offline decision-making literature. In addition, we seek to understand the benefits of using offline data in online decision-making and show nearly minimax-optimal rates in a wide range of regimes.

1. Introduction

Reinforcement learning (RL) has achieved remarkable empirical success in a wide range of challenging tasks, from playing video games at the same level as humans (Mnih et al., 2015), surpassing champions at the game of Go (Silver et al., 2018), to defeating top-ranked professional players in StarCraft (Vinyals et al., 2019). However, many of these systems require extensive online interaction in gameplay with other players who are experts at the task or some form of self-play (Li et al., 2016; Ouyang et al., 2022). Such online interaction may not be affordable in many real-world scenarios due to concerns about cost, safety, and ethics (e.g., healthcare and autonomous driving). Even in domains where online interaction is possible (e.g., dialogue systems), we would still prefer to utilize available historical, pre-collected datasets to learn useful decision-making policies efficiently. Such an approach would allow leveraging plentiful data, possibly replicating the success that supervised learning has had recently (LeCun et al., 2015). Offline RL has emerged as an alternative to allow learning from existing datasets and is particularly attractive when online interaction is prohibitive (Ernst et al., 2005; Lange et al., 2012; Levine et al., 2020).

Nevertheless, learning good policies from offline data presents a unique challenge not present in online decision-making: distributional shift. In essence, the policy that interacts with the environment and collects data differs from the target policy we aim to learn. This challenge becomes more pronounced in real-world problems with large state spaces, where it necessitates function approximation to generalize from observed states to unseen ones.

Representation learning is a basic challenge in machine learning. It is not surprising, then, that function approximation plays a pivotal role in reinforcement learning (RL) problems with large state spaces, mirroring its significance in statistical learning theory (Vapnik, 2013). Empirically, deep RL, which employs neural networks for function approximation, has achieved remarkable success across diverse tasks (Mnih et al., 2015; Schulman et al., 2017). The choice of function approximation class determines the inductive bias we inject into learning, e.g., our belief that the learner's environment is relatively simple even though the state space may be large.

It is natural, then, to understand different function approximation classes in terms of a tight characterization of their complexity and learnability. In statistical supervised learning, specific combinatorial properties of the function class are known to completely characterize sample-efficient supervised learning in both realizable and agnostic settings (Vapnik & Chervonenkis, 1971; Alon et al., 1997; Attias et al., 2023). For offline RL, a similar characterization is not known. With that as our motivation, we pose the following fundamental question that has largely remained unanswered: *What is a sufficient and necessary condition for learnability in offline RL with function approximation?*

We note that given the additional challenge of distribution shift in offline RL, such a characterization would depend not only on the properties of the function class but, more importantly, on the quality of the offline dataset. The existing literature on offline RL provides theoretical understanding only for limited scenarios of distributional shifts. These works capture the quality of offline data via a notion of data coverage. The strongest of these notions is that of uniform coverage, which requires that the behavior policy

[*]Equal contribution [1]Department of Computer Science, Johns Hopkins University, Baltimore 21218, USA. Correspondence to: Thanh Nguyen-Tang <nguyent@cs.jhu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

(aka, the policy used for data collection) covers the space of all feasible trajectories with sufficient probability (Munos & Szepesvári, 2008; Chen & Jiang, 2019). Recent works consider weaker assumptions, including single-policy concentrability, which only requires coverage for the trajectories of the target policy that we want to compete against (Liu et al., 2019; Rashidinejad et al., 2021; Yin & Wang, 2021) and relative condition numbers (Agarwal et al., 2021; Uehara & Sun, 2021). While the works above do not take into account function approximation in their notion of data coverage, Bellman residual ratio (Xie et al., 2021a), Bellman error transfer coefficient (Song et al., 2022) and data diversity (Nguyen-Tang & Arora, 2023) directly incorporate function approximation into the notion of offline data quality. Many of these notions are incompatible in that they give differing views of the landscape of offline learning. Further, there are no known lower bounds establishing the necessity of any of these assumptions. This suggests that there are gaps in our understanding of when offline learning is feasible.

In this paper, we work towards giving a more comprehensive and tighter characterization of problems that are learnable with offline data. Our key contributions are as follows.

- We introduce the notion of policy transfer coefficients, which subsumes other notions of data coverage.
- In conjunction with the pseudo-dimension of the function class, policy transfer coefficients give a tight characterization of offline learnability. Specifically, the class of offline learning problems characterized as learnable by policy transfer coefficients subsumes the problems characterized as learnable in prior literature. We provide (nearly) matching minimax lower and upper bounds for offline learning.
- Our results encompass offline learning in the setting of multi-armed bandits, contextual bandits, and Markov decision processes. We consider a variety of function approximation classes, including linear, neural-network-based, and, more generally, any function class with bounded L_1 covering numbers.
- We extend our results to the setting of hybrid offline-online learning and formally characterize the value of offline data in online learning problems.
- We overcome various technical challenges such as giving the uniform Bernstein's inequality for Bellman-like loss using empirical L_1 covering numbers (see Appendix A and Remark A.4) and removing the blowup of the number of iterations of the Hedge algorithm (see Remark 5.3). These may be of independent interest in themselves.

The rest of the paper is organized as follows. In Section 2, we introduce a formal setup for offline decision-making problems, where we focus mainly on the contextual bandit model to avoid deviating from the main points. In Section 3, we introduce policy transfer coefficients, a new notion of data coverage. In Section 4 and Section 5, we provide lower bounds and upper bounds, respectively, for offline decision-making (in the contextual bandit model). In Section 6, we consider a hybrid offline-online setting. We extend our results to Markov decision processes (MDPs) in Section 7. We conclude with a discussion in Section 8.

2. Background and Problem Formulation

2.1. Stochastic contextual bandits

We represent a stochastic contextual bandit environment with a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{D})$, where $\mathcal{X}$ denotes the set of contexts, $\mathcal{A}$ denotes the space of actions and $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ denotes an unknown joint distribution over the contexts and rewards. Without loss of generality, we take $\mathcal{Y} := [0,1]^{\mathcal{A}}$. A learner interacts with the environment as follows. At each time step, the environment samples $(X, Y) \sim \mathcal{D}$, the learner is presented with the context X, she commits to an action $a \in \mathcal{A}$, and observes reward $Y(a)$.

We model the learner as stochastic. The learner maintains a stochastic policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, i.e., a map from the context space to a distribution over the action space. The value, V^π, of a policy π is defined to be its expected reward,

$$V_{\mathcal{D}}^{\pi} := \mathbb{E}_{(X,Y)\sim\mathcal{D}, A\sim\pi(\cdot|X)}[Y(A)].$$

The sub-optimality of $\hat{\pi}$ w.r.t. any policy π is defined as:

$$\mathrm{SubOpt}_{\mathcal{D}}^{\pi}(\hat{\pi}) = V_{\mathcal{D}}^{\pi} - V_{\mathcal{D}}^{\hat{\pi}}. \tag{1}$$

We often suppress the subscript in $V_{\mathcal{D}}^{\pi}$ and $\mathrm{SubOpt}_{\mathcal{D}}^{\pi}(\hat{\pi})$.

2.2. Offline data

Let $S = \{(x_i, a_i, r_i)\}_{i\in[n]}$ be a dataset collected by a (fixed, but unknown) "behavior" policy μ, i.e., $(x_i, y_i) \overset{i.i.d.}{\sim} \mathcal{D}$, $a_i \sim \mu(\cdot|x_i)$, and $r_i = y_i(a_i)$ for all $i \in [n]$. The goal of offline learning is to learn a policy $\hat{\pi}$ from the offline data such that it has small sub-optimality $\mathrm{SubOpt}_{\mathcal{D}}^{\pi}(\hat{\pi})$ for as wide as possible a range of comparator policies π (possibly including an optimal policy $\pi_{\mathcal{D}}^* \in \arg\max_\pi V_{\mathcal{D}}^{\pi}$).

2.3. Function approximation

A central aspect of any value-based method for a sequential decision-making problem is to employ a certain function class $\mathcal{F} \subset [0,1]^{\mathcal{X}\times\mathcal{A}}$ for modeling rewards in terms of contexts and actions; it is typical to solve a regression problem using squared loss. The choice of the function class reflects learner's inductive bias or prior knowledge about the task

at hand. In particular, we often make the following realizability assumption (Chu et al., 2011; Agarwal et al., 2012; Foster & Rakhlin, 2020).

Assumption 2.1 (Realizability)**.** There exists an $f^* \in \mathcal{F}$ such that $f^*(x,a) = \mathbb{E}[Y(a)|X=x], \forall (x,a) \in (\mathcal{X} \times \mathcal{A})$.

We will utilize a well-understood characterization of complexity of real-valued function classes from statistical learning, that of pseudo-dimension (Pollard, 1984).

Definition 2.2 (Pseudo-dimension)**.** A set $\{z_1, \ldots, z_d\} \subset \mathcal{Z}$ is said to be shattered by $\mathcal{H} \subset \mathbb{R}^{\mathcal{Z}}$ if there exists a vector $r \in \mathbb{R}^d$ such that for all $\epsilon \in \{\pm 1\}^m$, there exists $h \in \mathcal{H}$ such that $\text{sign}(h(z_i) - r_i) = \epsilon_i$ for all $i \in [n]$. The pseudo-dimension $\text{Pdim}(\mathcal{H})$ is the cardinality of the largest set shattered by $\mathcal{H}$.

For example, neural networks of depth L, with ReLU activation, and total number of parameters (weights and biases) equal to W have pseudo-dimension of $\mathcal{O}(WL\log(W))$ (Bartlett et al., 2019).

Assumption 2.3. Define the function class associated with each fixed action, $\mathcal{F}(\cdot, a) := \{f(\cdot, a) : f \in \mathcal{F}\}$. We assume that $\sup_{a \in \mathcal{A}} \text{Pdim}\mathcal{F}(\cdot, a) \leq d$.

The reason we assume that the function class associated with each action has a bounded pseudo-dimension is that we can provide (nearly) matching lower and upper bounds for such a function class. We also provide upper bounds in terms of the covering number of a function class.

Definition 2.4 (Covering number)**.** Let $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$, and $\hat{P}_S(\cdot) = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}(\cdot)$ be the empirical distribution, where δ_z is the Dirac function at z. For any $p > 0$ and any $\mathcal{H} \subset \mathbb{R}^{\mathcal{Z}}$, let $N_p(\mathcal{H}, \epsilon, L_p(\hat{P}_S))$ denote the size of the smallest $\mathcal{H}'$ such that:

$$\forall h \in \mathcal{H}, \exists h' \in \mathcal{H}' : \|h - h'\|_{L_p(\hat{P}_S)} \leq \epsilon,$$

where $\|h - h'\|_{L_p(\hat{P}_S)} := \left(\frac{1}{n}\sum_{i=1}^n |h(z_i) - h'(z_i)|^p\right)^{1/p}$ is a pseudo-metric on $\mathcal{H}$. We define the (worst-case) L_p covering number $N_p(\mathcal{H}, \epsilon, n)$ as:

$$N_p(\mathcal{H}, \epsilon, n) = \sup_{S:|S|=n} N_p(\mathcal{H}, \epsilon, L_p(\hat{P}_S)).$$

Notation. Let $f(x, \pi) := \mathbb{E}_{a \sim \pi(\cdot|x)} f(x,a), \forall x$, and $\mathcal{F}(\cdot, \Pi) := \{f(\cdot, \pi) : f \in \mathcal{F}, \pi \in \Pi\}$, where $\Pi := \Delta(\mathcal{A})^{\mathcal{X}}$ is the set of all possible (Markovian) policies. Let $\mathcal{D} \otimes \pi$ denote the distribution of the random variable (x, a, r), where $(x, y) \sim \mathcal{D}, a \sim \pi(\cdot|x), r = y(a)$. Let l_f denote the random variable $(f(x,a) - r)^2$ where $(x, a, r) \sim \mathcal{D} \otimes \mu$. The empirical and the population means of l_f are represented as $\hat{P}l_f = \frac{1}{n}\sum_{i=1}^n l_f(x_i, a_i, r_i)$ and $Pl_f = \mathbb{E}_{(x,a,r)\sim\mathcal{D}\otimes\mu}[l_f(x,a,r)]$, respectively. The empirical and the population means of f under policy π are denoted as $\hat{P}f(\cdot, \pi) = \frac{1}{n}\sum_{i=1}^n f(x_i, \pi)$, and $Pf(\cdot, \pi) := \mathbb{E}_{x \sim \mathcal{D}}[f(x, \pi)]$, respectively. We write $a \lesssim b$ to mean $a = \mathcal{O}(b)$, suppressing only absolute constants, and $a \widetilde{<} b$ to mean $a = \widetilde{\mathcal{O}}(b)$, further suppressing log factors. Define $[x]_1 := \max\{\sqrt{x}, x\}$.

3. Offline Decision-Making as Transfer Learning

We view offline decision-making as transfer learning where the goal is to utilize pre-collected experiences for learning new tasks. A key observation we leverage is that there are parallels in how the two areas capture distribution shift – a common challenge in both settings. Transfer learning uses various notions of distributional discrepancies (Ben-David et al., 2010; David et al., 2010; Germain et al., 2013; Sugiyama et al., 2012; Mansour et al., 2012; Tripuraneni et al., 2020; Watkins et al., 2023) to capture distribution shift between the source tasks and the target task, much like how offline decision-making uses various notions of data coverage to measure the distributional mismatch due to offline data. We consider a new notion of data coverage inspired by transfer learning, which will be shown shortly to tightly capture the statistical complexity of offline decision-making from a behavior policy.

Definition 3.1 (Policy transfer coefficients)**.** Given any policy π, $\rho \geq 0$ is said to be a policy transfer exponent from μ to π w.r.t. $(\mathcal{D}, \mathcal{F})$ if there exists a finite constant C, called policy transfer factor, such that:

$$\forall f \in \mathcal{F} : \left(\mathbb{E}_{\mathcal{D}\otimes\pi}[f^* - f]\right)^{2\rho} \leq C\mathbb{E}_{\mathcal{D}\otimes\mu}[(f^* - f)^2]. \quad (2)$$

Any such pair (ρ, C) is said to be a policy transfer coefficient from μ to π w.r.t. $(\mathcal{D}, \mathcal{F})$. We denote the *minimal* policy transfer exponent by ρ_π.[1] The *minimal* policy transfer factor corresponding to ρ_π is denoted as C_π.

Remark 3.2. Our definition of policy transfer resembles and is directly inspired by the notion of transfer exponent by (Hanneke & Kpotufe, 2019), which we refer to as Hanneke-Kpotufe (HK) transfer exponent for distinction. A direct adaptation of the HK transfer exponent would result in:

$$\forall f \in \mathcal{F} : \mathbb{E}_{\mathcal{D}\otimes\pi}[(f^* - f)^2]^\rho \leq C\mathbb{E}_{\mathcal{D}\otimes\mu}[(f^* - f)^2]. \quad (3)$$

Note the difference in the LHS of Equation (2) and that of Equation (3). When defining policy transfer coefficients, we use ℓ_2-distance[2] $\mathbb{E}_{\mathcal{D}\otimes\mu}[(f^* - f)^2]$ w.r.t. the behavior policy to control the *expected value gap* $\mathbb{E}_{\mathcal{D}\otimes\pi}[f^* - f]$, whereas the HK transfer exponent directly requires a bound on squared distance $\mathbb{E}_{\mathcal{D}\otimes\pi}[(f^* - f)^2]$ w.r.t. to the target policy. While this appears to be a small change, our notion bears a deeper connection with offline decision-making

[1] If ρ is a policy transfer exponent, so is any $\rho' \geq \rho$.

[2] The ℓ_2-distance from f to f^* corresponds to the excess risk of f w.r.t. squared loss when assuming realizability.

that the HK transfer exponent cannot capture. Perhaps, the best way to demonstrate this is through a concrete example where the policy transfer coefficient tightly captures the learnability of offline decision-making while the HK transfer exponent fails – we return to this in Example 3.3. For now, it suffices to argue the tightness of our notion more generally. Indeed, our policy transfer exponent notion allows a tighter characterization of the transferability for offline decision-making, as indicated by Jensen's inequality:

$$(\mathbb{E}_{\mathcal{D}\otimes\pi}[f^* - f])^2 \leq \mathbb{E}_{\mathcal{D}\otimes\pi}[(f^* - f)^2].$$

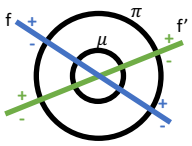
Now, consider a random variable, ζ, that is distributed according to the Bernoulli distribution $\mathrm{Ber}(p)$. Then, we have $\mathbb{E}[\zeta^2]/|\mathbb{E}[\zeta]|^2 = 1/p$. This means that in this example, the transfer factor C_1 (corresponding to $\rho = 1$) in our definition is smaller than the transfer factor implied by the original definition of (Hanneke & Kpotufe, 2019) by a factor of p, which is significant for small values of p.

3.1. Relations with other notions of data coverage

In this section, we highlight the properties of transfer exponents and compare them with other notions of data coverage considered in offline decision-making literature. Perhaps the most common notions of data coverage are that of single-policy concentrability coefficients (Liu et al., 2019; Rashidinejad et al., 2021), relative condition numbers (for linear function classes) (Agarwal et al., 2021; Uehara & Sun, 2021), and data diversity (Nguyen-Tang & Arora, 2023).[3] We demonstrate that policy transfer coefficients strictly generalize all these prior notions, in the sense that bounds on the prior notions of data coverage always imply bounds on transfer exponents but not vice versa. Specifically, there are problem instances for which the existing measures of data coverage tend to infinity, yet these problems are learnable given the characterization in terms of transfer coefficients.

Compared with concentrability coefficients. The concentrability coefficient between π and μ is defined as $\kappa_\pi := \sup_{x,a} \frac{\pi(a|x)}{\mu(a|x)}$. The finiteness of κ_π is widely used as one of the sufficient conditions for sample-efficient offline decision-making. By definition, the policy transfer factor corresponding to the policy transfer exponent of 1, is always upper-bounded by κ_π. The finiteness of κ_π requires the support of μ to contain that of π. However, offline decision-making does not even need overlapping supports between a target policy and the behaviour policy.

Example 3.3. *Consider $|\mathcal{X}| = 1$ and $\mathcal{A} = \mathbb{R}^d$. Let $\mathcal{F}$ be the class of d-dimensional halfspaces that pass through the origin, and Π be the set of d-dimensional spheres centered at the origin, $\Pi = \{\mathit{Uniform}(\{a \in \mathbb{R}^d : \|a\| = r\}) : r \geq 0\}$ (see the figure below).*



Example 3.3, taken from (Hanneke & Kpotufe, 2019), gracefully reflects the "correctness" of policy transfer exponent at characterizing learnability in offline settings, whereas both the HK transfer exponent and the concentrability coefficient fail to do so. In this case, by direct computation we have that $\mathbb{E}_{\mathcal{D}\otimes\pi}[f^* - f] = 0$, thus the minimal policy transfer exponent from any behavior policy $\mu \in \Pi$ to any different policy $\pi \in \Pi$ approaches zero, i.e., $\rho_\pi \to 0$, while the HK transfer exponent, denoted ρ_π^{HK}, is 1 and the concentrability coefficient κ_π is infinity. These different values of data quality measures translate into different predictive bounds of offline learning for Example 3.3. For concreteness, we summarize these bounds in Table 1.

Data quality measure	Predictive bound
Policy transfer exponent $\rho_\pi \to 0$	0
HK transfer exponent $\rho_\pi^{\mathrm{HK}} = 1$	$\sqrt{\frac{C_\pi}{n}}$
Concentrability coefficient $\kappa_\pi = \infty$	∞

Table 1. Predictive bounds for offline learning of Example 3.3 stemming from different measures of data quality. The bounds for the policy transfer exponent and the HK transfer exponent are obtained from Theorem 5.1 (note that our results are applicable to HK transfer exponents as well), while the bound for the concentrability coefficient is obtained from Rashidinejad et al. (2021).

Which of the above data quality measures give a tight characterization of learnability of problem in Example 3.3? With the realizability assumption, the value V^π of any policy in Π is equal to $1/2$, regardless of the policy. Thus, the true sub-optimality is zero. This is tightly bounded by the predictive bound of our policy transfer exponent, while those by HK transfer exponent and concentrability coefficient give vacuous bounds, as shown in Table 1.

Another interesting remark is that for offline decision-making, it is not always necessary to learn the true reward function, a task that requires the sample complexity of $\tilde{\Theta}(\frac{d}{\epsilon})$ – this is captured by the HK transfer exponent at the value of 1; see (Hanneke & Kpotufe, 2019, Example 1).

Compared with the data diversity of Nguyen-Tang & Arora (2023). The notion of data diversity, recently proposed by Nguyen-Tang & Arora (2023) is motivated by the notion of task diversity in transfer learning (Tripuraneni

[3] Data diversity of Nguyen-Tang & Arora (2023), in fact, generalizes Bellman error transfer coefficient of Song et al. (2022) to allow an additive error.

et al., 2020). Nguyen-Tang & Arora (2023) show that data diversity subsumes both the concentrability coefficients and the relative condition numbers and that it can be used to derive strong guarantees for offline decision-making and state-of-the-art bounds when the function class is finite or linear. Their data diversity notion is, in fact, the policy transfer factor corresponding to policy transfer exponent $\rho = 1$. The following example, which is modified from (Hanneke & Kpotufe, 2019, Example 3), shows that data diversity can be infinite while the minimal transfer exponent is finite.

Example 3.4. *Consider $|\mathcal{X}| = 1, \mathcal{A} = [-1, 1], \mathcal{F} = \{f_t : t \in [-1, 1]\}$ where $f_t(a) = \mathbb{1}\{a \geq t\}$ is a 1-dimensional threshold , and let f_0 be the optimal mean reward function. Let $\iota > 0$ be any positive scalar. Consider the following distribution: $\mu(a) \propto a^{2\iota-1}$ for $a \geq 0$ and $\mu(a)$ is uniform for $a \in [-1, 0]$. Let π be the uniform distribution over $\mathcal{A}$. By direct computation, for any $t \geq 0$, $|\mathbb{E}_\pi[f_0 - f_t]| \propto t$ and $\mathbb{E}_\mu[(f_0 - f)^2] \propto t^{2\iota}$. Thus, no $\rho < 2\iota$ can be a policy transfer exponent, as $\lim_{t\to 0} |\mathbb{E}_\pi[f_0 - f_t]|^\rho / \mathbb{E}_\mu[(f_0 - f)^2] \to \infty$. Thus, the bound in Nguyen-Tang & Arora (2023) becomes vacuous. However, as long as the squared Bellman error $\mathbb{E}_\mu[(f_0 - f_t)^2]$ can predict the Bellman error $|\mathbb{E}_\pi[f_0 - f_t]|$, one should expect that offline learning is still possible, albeit at a slower rate.*

As our policy transfer coefficients cover the data diversity of Nguyen-Tang & Arora (2023) as a special case ($\rho = 2$), which, in turn, is a generalization of the relative condition number, we refer to Nguyen-Tang & Arora (2023) for a detailed comparison with the relative condition number.

4. Lower Bounds

Let $\mathcal{B}(\rho, C, d)$ denote the class of offline learning problem instances with any distribution $\mathcal{D}$ over contexts and rewards, any function class $\mathcal{F}$ that satisfies Assumptions 2.1 and 2.3, a behavior policy μ, and all policies $\pi \in \Pi$ such that policy transfer coefficients w.r.t. μ are (ρ, C). For this class, we give a lower bound on the sub-optimality of any offline learning algorithm.

Theorem 4.1. *For any $C > 0, \rho \geq 1, n \geq d \cdot \max\{2^{2\rho-4}C, C^{\frac{1}{\rho-1}}/32\}$, we have*

$$\inf_{\hat{\pi}(\cdot)} \sup_{(\mathcal{D},\mu,\pi,\mathcal{F})\in\mathcal{B}(\rho,C,d)} \mathbb{E}_{\mathcal{D}}\left[\mathrm{SubOpt}^{\pi}_{\mathcal{D}}(\hat{\pi})\right] \gtrsim \left(\frac{Cd}{n}\right)^{\frac{1}{2\rho}},$$

where the infimum is taken over all offline algorithm $\hat{\pi}(\cdot)$ (a randomized mapping from the offline data to a policy).

The lower bound in Theorem 4.1 is information-theoretic, i.e., it applies to *any* algorithm for problem class $\mathcal{B}(\rho, C, d)$. The lower bound is obtained by constructing a set of hard contextual bandit (CB) instances $\{\mathcal{D}_i\}$ that are supported on d data points. Then, for each $\mathcal{D}_i$, we pick the hardest comparator policy $\pi = \pi^*_{\mathcal{D}_i}$ and design a behavior policy that satisfies the policy transfer condition. We pick a simple enough function class that satisfies realizability and ensures that the policy transfer exponents and the pseudo-dimension are bounded. We then proceed to show that given a behavior policy μ, for any two CB instances $\mathcal{D}_i$ and $\mathcal{D}_j$ that are close to each other (i.e., $\mathrm{KL}[(\mathcal{D}_i \otimes \mu)^n \| (\mathcal{D}_j \otimes \mu)^n]$ is small) the corresponding optimal policies disagree. A complete proof is given in Appendix B.

5. Upper Bounds

Next, we show that there exists an offline learning algorithm that is agnostic to the minimal policy transfer coefficient of any policy, yet it can compete uniformly with all comparator policies, as long as their minimal policy transfer exponent is finite. For this algorithm, we give an upper bound for VC-type classes that matches the lower bound in the previous section up to log factors, ignoring the dependence on $K = |\mathcal{A}|$. For more general function classes, we only provide upper bounds.

The general recipe for our algorithm (Algorithm 1) is rather standard. We follow the actor-critic framework for offline RL studied in several prior works (Zanette et al., 2021; Xie et al., 2021a; Nguyen-Tang & Arora, 2023). The algorithm alternates between computing a pessimistic estimate of the actor and improving the actor with the celebrated Hedge algorithm (Freund & Schapire, 1997).

Algorithm 1 Hedge for Offline Decision-Making (OfDM-Hedge)

Input: Offline data S, function class $\mathcal{F}$
Hyperparameters: Confidence parameter β, learning rate η, number of iterations T
Initialize $\pi_1(\cdot|x) = \mathrm{Uniform}(\mathcal{A}), \forall x \in \mathcal{X}$
for $t = 1$ **to** T **do**
  Pessimism: $f_t = \underset{f\in\mathcal{F}:\hat{P}l_f - \hat{P}l_{\hat{f}} \leq \beta}{\arg\min} \hat{P}f(\cdot, \pi_t)$
  Hedge: $\pi_{t+1}(a|x) \propto \pi_t(a|x)e^{\eta f_t(x,a)}, \forall (x, a)$
end for
Output: A randomized policy $\hat{\pi}$ as a uniform distribution over $\{\pi_t\}_{t\in[T]}$.

The following result bounds the suboptimality of OfDM-Hedge up to absolute constants which we ignore for ease of exposition (see Theorem C.1 for exact constants).

Theorem 5.1. *Fix any $\delta \in [0, 1], \epsilon \geq 0$. Assume that $|\mathcal{A}| = K$. Then, for any $(\mathcal{D}, \mathcal{F})$ such that Assumption 2.1 holds, invoking Algorithm 1 with $\beta = \Theta\left(\epsilon + \frac{\log(N_1(\mathcal{F},\epsilon,n)/\delta)}{n}\right)$ and $\eta = \Theta(\sqrt{\frac{\log K}{T}})$ returns a policy $\hat{\pi}$ such that with*

probability at least $1-\delta$, $\forall \pi \in \Pi$, $\forall T \in \mathbb{N}$,

$$\mathbb{E}[\mathrm{SubOpt}^{\pi}_{\mathcal{D}}(\hat{\pi})|S] \lesssim C_{\pi}^{\frac{1}{2\rho_\pi}} \left(\epsilon + \frac{\log(N_1(\mathcal{F},\epsilon,n)/\delta)}{n}\right)^{\frac{1}{2\rho_\pi}}$$
$$+\mathbb{1}_{\{|\mathcal{X}|>1\}} \left\{ \left[\frac{\log(N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n)/\delta)}{n}\right]_1 + \epsilon \right\} + \sqrt{\frac{\log K}{T}}.$$

The upper bound consists of three main terms, which represent the transfer cost for offline decision-making, standard statistical learning error, and the optimization error of the Hedge algorithm, respectively. Note that Theorem 5.1 does not require Assumption 2.3.

Remark 5.2 (**Adaptive to policy transfer coefficients**). OfDM-Hedge does not need to know the policy transfer coefficients of any target policy, yet it is adaptively competing with any target policy π characterized by minimal policy transfer exponent ρ_π and policy transfer factor C_π.

Remark 5.3 (**No cost blowup of the Hedge algorithm**). Note that the first two terms in the upper bound in Theorem 5.1 are completely agnostic to the number of iterations T. Thus, scaling up T to drive the last term to zero does not increase the effective dimension – a desirable property that is absent in the bounds of similar algorithms provided in Zanette et al. (2021); Xie et al. (2021a); Nguyen-Tang & Arora (2023). This difference stems from the fact that prior work uses uniform convergence over all policies that can be generated by the Hedge algorithm over T steps. We, however, recognize that any policy π_t that is generated by a step of the Hedge algorithm is only used in one single form: $f(\cdot,\pi_t)$ for $f \in \mathcal{F}$. Thus, it suffices to control the complexity of $\mathcal{F}(\cdot,\Pi)$ where Π is the set of all possible stationary policies. This elegant trick becomes more clear in terms of benefit in comparison with other works that suffer sub-optimal rates due to this cost blowup of the Hedge algorithm; we refer the reader to Section 7.2 for further discussion.

Remark 5.4 (**Potential barrier for offline decision-making in the fast transfer regime** $\rho_\pi < 1$). The second term in the bound above vanishes when $|\mathcal{X}| = 1$ (i.e., in a multi-armed bandit setting). This yields error rates that are faster than $1/\sqrt{n}$ if the rate of transfer from μ to π is fast (i.e., $\rho_\pi < 1$). One such example is Example 3.3 where $\rho \to 0$, which our upper bounds capture rather precisely. In the general case with $|\mathcal{X}| > 1$, OfDM-Hedge is unable to take advantage of fast policy transfer exponent $\rho_\pi < 1$, as the bound also involves the second term stemming from standard statistical learning over the context domain $\mathcal{X}$. This is quite different from the supervised learning setting of Hanneke & Kpotufe (2019) where we can always ensure a fast transfer rate when $\rho_\pi < 1$. Further, note that our lower bound (Theorem 4.1) excludes the fast transfer regime $\rho_\pi < 1$. So, it remains unclear if, any algorithm in the offline setting can leverage fast transfer.

Next, we instantiate the upper bound in Theorem 5.1 for VC-type function classes.

Example 5.5 (VC-type / parametric classes)**.** *Under the same setting as Theorem 5.1, if Assumption 2.3 holds, the suboptimality of OfDM-Hedge is bounded by (ignoring* $\log(1/\delta)$ *terms):*

$$\max\left\{ \left[\frac{Kd\log(dn)}{n}\right]_1, \left(\frac{C_\pi K d\log(dn)}{n}\right)^{\frac{1}{2\rho_\pi}} \right\}.$$

The VC-type parametric classes include the d-dimensional linear class as a special case. The bound above follows from Theorem 5.1 and using the following inequalities: $\max_{a\in\mathcal{A}} N_1(\mathcal{F}(\cdot,a),\epsilon,n) \le e(d+1)(\frac{2e}{\epsilon})^d$ (Haussler, 1995) and $\max\{N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n), N_1(\mathcal{F},\epsilon,n)\} \le (\max_{a\in\mathcal{A}} N_1(\mathcal{F}(\cdot,a),\epsilon/K,n))^K$ (Lemma F.5).

For the common "large-sample, difficult-transfer" regime, i.e., $n \ge Kd\log(dn)$ and $\rho_\pi \ge 1$, the upper bounds for the parametric classes in Example 5.5 match the lower bound in Theorem 4.1, ignoring constants, log factors and dependence on K.

The upper bound in Theorem 5.1 also applies to nonparametric classes, though we do not provide a (matching) lower bound for this case. We leave that for future work.

Example 5.6 (Nonparametric classes)**.** *Under the same setting as Theorem 5.1, assume that for all a,* $\mathcal{F}(\cdot,a)$ *scales polynomially with the inverse of the scale, i.e., for some* $p > 0$

$$\log N_1(\epsilon,\mathcal{F}(\cdot,a),n) \le \left(\frac{1}{\epsilon}\right)^p, \forall a \in \mathcal{A}, \forall \epsilon \ge 0.$$

Then, the upper bound in Theorem 5.1 is of the order (ignoring $\log(1/\delta)$ *terms):*

$$\max\left\{ C_\pi^{\frac{1}{2\rho_\pi}} \left(\frac{K}{n}\right)^{\frac{1}{2\rho_\pi(p+1)}}, \left(\frac{K}{n}\right)^{\frac{1}{1+p}}, \left(\frac{K}{n}\right)^{\frac{1}{2(1+p)}} \right\}.$$

In the typical "large-sample difficult-transfer" regime, the above yields an error rate of $C_\pi^{\frac{1}{2\rho_\pi}} \left(\frac{K}{n}\right)^{\frac{1}{2\rho_\pi(p+1)}}$ which vanishes with n. Typical examples of nonparametric classes include infinite-dimension linear classes with $p = 2$ (Zhang, 2002), the class of 1-Lipschitz functions over $[0,1]^d$, with $p = d$ (Slivkins, 2011), the class of Holder-smooth functions of order β over $[0,1]^d$, with $p = d/\beta$ (Rigollet & Zeevi, 2010), and neural networks with spectrally bounded norms, with $p = 2$ (Bartlett et al., 2017).

6. Offline Data-assisted Online Decision-Making

In this section, we consider a hybrid setting, where in addition to the offline data $S = \{(x_i, a_i, r_i)\}_{i\in[n]}$, the learner is allowed to interact with the environment for m rounds. The goal is to output a policy $\hat{\pi}_{\text{hyb}}$ with small $\mathrm{SubOpt}^{\pi}_{\mathcal{D}}(\hat{\pi}_{\text{hyb}})$ w.r.t. π with a high value $V^{\pi}_{\mathcal{D}}$. We assume realizability (Assumption 2.1) and for simplicity, we focus only on VC-type function classes with pseudo-dimension at most d. Let $\rho^* = \rho_{\pi^*}$ and $C^* = C_{\pi^*}$ denote the policy transfer coefficients for an optimal policy π^*.

To avoid deviating from the main point, in this section, we focus on the "large sample, difficult transfer" regime, where $d \geq Kd\log(dn)$ and $\rho^* \geq 1$. The key question we ask is whether a learner can perform better in a hybrid setting than in a purely online or offline setting.

6.1. Lower bounds

We start with a lower bound for any hybrid learner for the class of problems $\mathcal{B}(\rho, C, d)$ as in Section 4.

Theorem 6.1. *For any $C > 0, \rho \geq 1$, and sample size $n \geq d\max\{2^{2\rho-4}C, \frac{C^{\frac{1}{\rho-1}}}{32}\}$, we have*

$$\inf_{\hat{\pi}_{hyb}(\cdot)} \sup_{(\mathcal{D},\mu,\pi,\mathcal{F})\in\mathcal{B}(\rho,C,d)} \mathbb{E}_{\mathcal{D}}\left[\mathrm{SubOpt}^{\pi}_{\mathcal{D}}(\hat{\pi}_{hyb})\right] \gtrsim \min\left\{\left(\frac{Cd}{n}\right)^{\frac{1}{2\rho}}, \sqrt{\frac{d}{m}}\right\},$$

where the infimum is taken over all possible hybrid algorithm $\hat{\pi}_{hyb}(\cdot)$.

Ignoring log factors and dependence on K, the first term in the lower bound matches the upper bound for OfDM-Hedge. Therefore, if that is the dominating term, the learner does not benefit from online interaction. The second term, $\sqrt{d/m}$, matches the upper bound of the state-of-the-art online learner (Simchi-Levi & Xu, 2022) with an online interaction budget of m. In the regime where the latter term is dominating there is no advantage to having offline data.

To summarize, the lower bound suggests that if the policy transfer coefficient is known a priori to the hybrid learner, there is no benefit of mixing the offline data with the online data at least in the worst case. That is, obtaining the nearly minimax-optimal rates for hybrid learning is akin to either discarding the online data and running the best offline learner or ignoring the offline data and running the best online learner. Which algorithm to run depends on the transfer coefficient. That there is no benefit to mixing online and offline data is a phenomenon that has also been discovered in a related setting of policy finetuning (Xie et al., 2021b).

6.2. Upper bounds

The discussion following the lower bound suggests different algorithmic approaches (purely offline vs. purely online) for different regimes defined in terms of the policy transfer coefficient. However, it is unrealistic to assume that the learner has prior knowledge of the transfer coefficient. We present a hybrid learning algorithm (Algorithm 2) that offers the best of both worlds. Without requiring the knowledge of the policy transfer coefficient, it produces a policy with nearly optimal minimax rates.

The key algorithmic idea is rather simple and natural. We invoke both an offline policy optimization algorithm and an online learner, resulting in policies $\hat{\pi}_{\text{off}}$ and $\hat{\pi}_{\text{on}}$, respectively. Half of the interaction budget, i.e., $m/2$ rounds, is utilized for learning $\hat{\pi}_{\text{on}}$. For the remaining $m/2$ rounds, we run the EXP4 algorithm (Auer et al., 2002) with $\hat{\pi}_{\text{off}}$ and $\hat{\pi}_{\text{on}}$ as expert policies. The output of EXP4 is a uniform distribution over all of the iterates. We denote this randomized policy as $\hat{\pi}_{\text{hyb}}$.We can equivalently represent $\hat{\pi}_{\text{hyb}}$ as a distribution over $\{\hat{\pi}_{\text{off}}, \hat{\pi}_{\text{on}}\}$.

Algorithm 2 Hybrid Learning Algorithm

1: **Input:** Offline data S, function class $\mathcal{F}$, m online interactions
2: $\hat{\pi}_{\text{off}} \leftarrow \text{OfflineLearner}(S, \mathcal{F})$
3: $\hat{\pi}_{\text{on}} \leftarrow \text{OnlineLearner}(\frac{m}{2}, \mathcal{F})$
4: $\hat{\pi}_{\text{hyb}} \leftarrow \text{EXP4}(\frac{m}{2}, \{\hat{\pi}_{\text{off}}, \hat{\pi}_{\text{on}}\})$
5: **Output:** $\hat{\pi}_{\text{hyb}}$

Proposition 6.2. *Algorithm 2 return a randomized policy $\hat{\pi}_{hyb}$ such that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\max_{\pi\in\{\hat{\pi}_{off},\hat{\pi}_{on}\}} \mathbb{E}_{\mathcal{D}}[\mathrm{SubOpt}^{\pi}_{\mathcal{D}}(\hat{\pi}_{hyb})|S, S_{on}] \leq \frac{4\sqrt{2\log 2}}{\sqrt{m}} + \frac{32\log(\log(m/2)/\delta)}{3m} + \frac{4}{m},$$

where S is the offline data and S_{on} is the online data collected by $\hat{\pi}_{on}$ in Algorithm 2.

The first term in the bound above comes from the guarantees on EXP4 with two experts (Lattimore & Szepesvári, 2020, Theorem 18.3) and the remaining terms result from an (improved) online-to-batch conversion (Nguyen-Tang et al., 2023, Lemma A.5) and the assumption that contexts are sampled i.i.d. Note that the result above does not require any assumption on $\mathcal{F}$ and $\mathcal{D}$.

Next, we implement Algorithm 2 using FALCON+ (Simchi-Levi & Xu, 2022, Algorithm 2) for online learning and OfDM-Hedge for offline learning. Then, we have the following bound on the output of the hybrid learning algorithm.

Theorem 6.3. *Assume that the closure of $\mathcal{F}$ is convex. Then, given that Assumptions 2.1 and 2.3 hold, we have*

$$\mathbb{E}_{\mathcal{D}}\Big[\mathrm{SubOpt}_{\mathcal{D}}^{\pi^*_{\mathcal{D}}}(\hat{\pi}_{hyb})\Big] \lesssim \frac{1}{\sqrt{m}} + \min\left\{\left(\frac{C^* K d}{n}\right)^{\frac{1}{2\rho^*}}, K\sqrt{\frac{d}{m}}\right\}.$$

In the bound above, the first term is the standard error rate of EXP4, the term $\left(\frac{C^* K d}{n}\right)^{\frac{1}{2\rho^*}}$ is the error rate of OfDM-Hedge (see Theorem 5.1), and $K\sqrt{\frac{d}{m}}$ is the error rate of FALCON+ (which we tailor to our setting in Appendix D.1). Applying Proposition 6.2 yields the minimum over the error rates of the offline and online learning.

In the regime that the EXP4 cost is dominated by the second term, i.e. , when $m \geq \left(\frac{n}{C^* K d}\right)^{1/\rho^*}$, the upper bound on the suboptimality of hybrid learning nearly matches the lower bound, modulo log factors and pesky dependence on K.

6.3. Related works for the hybrid setting

Xie et al. (2021b) consider a related setting of policy fine-tuning, where the online learner is given an additional reference policy μ that is close to the optimal policy and the learner can collect data using μ at any point. They do not consider function approximation and utilize single-policy concentrability coefficients. Song et al. (2022) consider the same hybrid setting as ours (albeit for RL) and show that offline data of good quality can help avoid a need to explore, thereby resulting in an oracle-efficient algorithm which in general is not possible. The statistical complexity they establish for their algorithm is not minimax-optimal, and can get arbitrarily worse with the quality of the offline data. Our guarantees, on the other hand, are adaptive as they revert to online learning guarantees when the quality of the offline data is low. Wagenmaker & Pacchiano (2023) give instance-dependent bounds for hybrid RL with linear function approximation under a uniform coverage condition. Other related works include hybrid RL in tabular MDPs (Li et al., 2023), a Bayesian framework for incorporating offline data into the prior distribution (Tang et al., 2023), and one-shot online learning using offline data (Zhang & Zanette, 2023).

7. Offline Decision-Making in MDPs

In this section, we extend our results to offline decision-making in Markov decision processes (MDPs). We show that the key insights developed for the contextual bandit model extend naturally to offline learning of MDPs as we establish nearly matching upper and lower bounds.

Setup. Let $M = \mathrm{MDP}(\mathcal{X}, \mathcal{A}, [H], \{P_h\}_{h\in[H]}, \{r_h\}_{h\in[H]})$ denote an episodic Markov decision process with state space $\mathcal{X}$, action space $\mathcal{A}$, horizon length H, transition kernel $P_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$, and mean reward functions $r_h : \mathcal{X} \times \mathcal{A} \to [0, 1]$. For any policy $\pi = (\pi_1, \ldots, \pi_H)$ where $\pi_h : \mathcal{X} \to \Delta(\mathcal{A})$, let $\{Q_h^\pi\}_{h\in[H]}$ and $\{V_h^\pi\}_{h\in[H]}$ denote the action-value functions and the state-value functions, respectively. For any policy π, define the Bellman operator $[T_h^\pi g](x, a) := \mathbb{E}_{x'\sim P_h(\cdot|x,a)}[r_h(x, a) + g(x')]$. For simplicity, we assume that the MDP starts in the same initial state at every episode. Here $\mathbb{E}_\pi[\cdot]$ denotes the expectation with respect to the randomness of the trajectory $(x_h, a_h, \ldots, x_H, a_H)$, with $a_i \sim \pi_i(\cdot|x_i)$ and $x_{i+1} \sim P_i(\cdot|x_i, a_i)$ for all i. We assume that $|r_h| \leq 1, \forall h$.

The learner has access to a dataset $S = \{(x_h^{(t)}, a_h^{(t)}, r_h^{(t)})\}_{h\in[H],t\in[n]}$ collected using a behavior policy μ. Define the (value) sub-optimality as $\mathrm{SubOpt}_M^\pi(\hat{\pi}) := V_1^\pi(s_1) - V_1^{\hat{\pi}}(s_1)$. Wherever clear, we drop the subscript M in Q_M^π, V_M^π, d_M^π, and $\mathrm{SubOpt}_M^\pi(\hat{\pi})$.

Function approximation. We consider a function approximation class $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_H)$ where $\mathcal{F}_h \subseteq [0, H - h + 1]^{\mathcal{X}\times\mathcal{A}}$.

We make the following standard assumptions regarding how the function class $\mathcal{F}$ interacts with the underlying MDP M.

Assumption 7.1 (Realizability)**.** $\forall \pi \in \Pi, h \in [H], Q_h^\pi \in \mathcal{F}_h$.

Assumption 7.2 (Bellman completeness)**.** $\forall \pi \in \Pi, \forall h \in [H]$, if $f_{h+1} \in \mathcal{F}_{h+1}$, then $T_h^\pi f_{h+1} \in \mathcal{F}_h$.

For simplicity, we focus on VC-type/parametric $\mathcal{F}$, though, again, our upper bounds can yield a vanishing error rate for non-parametric classes.

Assumption 7.3. $\sup_{h\in[H],a\in\mathcal{A}} \mathrm{Pdim}(\mathcal{F}_h(\cdot, a)) \leq d$.

Definition 7.4 (Policy transfer coefficients)**.** Given any policy π, the minimal policy transfer exponent ρ_π w.r.t. $(M, \mathcal{F}, \mu)$ is the smallest $\rho \geq 0$ such that there exists a finite constant C such that for every $h \in [H]$

$$\forall f_h, g_h \in \mathcal{F}_h : |\mathbb{E}_\pi[f_h - g_h]|^{2\rho} \leq C\mathbb{E}_\mu[(f_h - g_h)^2]. \quad (4)$$

The smallest such C w.r.t ρ_π, denoted C_π, is called the policy transfer factor. The pair (ρ_π, C_π) is said to be the policy transfer coefficient of π.

7.1. Lower bounds

Let $\mathcal{M}(\rho, C, d)$ denote the class of offline learning problem instances with any MDP M, any function class $\mathcal{F}$ that satisfies Assumptions 7.1, 7.2, and 7.3, a behavior policy μ, and all policies $\pi \in \Pi$ such that policy transfer coefficients w.r.t. μ are (ρ, C). For this class, we give a lower bound on the sub-optimality of any offline learning algorithm.

Theorem 7.5. *For $\rho \geq 1$ and $n \geq \frac{CdH^{2\rho}}{32}$, we have*

$$\inf_{\hat{\pi}} \sup_{(M,\mathcal{F},\pi,\mu)\in\mathcal{M}(\rho,C,d)} \mathbb{E}_M[\mathrm{SubOpt}_M^{\pi}(\hat{\pi})] \gtrsim \left(\frac{H^2Cd}{n}\right)^{\frac{1}{2\rho}}.$$

7.2. Upper bounds

We now establish upper bounds for this setting. Due to space limitations, we only state the main result here and defer the details (including concrete algorithms) to Appendix E.2.

Theorem 7.6. *Let $\delta > 0$. Assume that $|\mathcal{A}| = K$. Then, there exists a learning algorithm that for any problem instance in the set $\mathcal{M}(\rho, C, d)$, given offline data S of size n, returns a policy $\hat{\pi}$ such that with probability at least $1-\delta$ over the randomness of generating S, we have*

$$\mathbb{E}\left[\mathrm{SubOpt}_M^{\pi}(\hat{\pi})|S\right] \widetilde{<} H\left(\frac{H^2C_{\pi}(Kd+\log(1/\delta))}{n}\right)^{\frac{1}{2\rho_{\pi}}},$$

relative to any comparator $\pi \in \Pi$.

There is a gap of $\mathcal{O}(H)$ between our upper and lower bounds, something that can potentially be improved using variance-aware algorithms. Nonetheless, Theorem 7.6 improves and generalizes the results of Xie et al. (2021a) and Nguyen-Tang & Arora (2023) on several fronts. First, since policy transfer coefficients subsume other notions of data coverage, our results hold for a larger class of problem instances. Second, our results hold for any function class with finite L_1 covering number. In contrast, the guarantees in Xie et al. (2021a) hold only for finite function classes and in Nguyen-Tang & Arora (2023) for function classes with finite domain-wide L_∞ covering numbers. Third, even when specializing to the setting of Xie et al. (2021a); Nguyen-Tang & Arora (2023) (i.e., $\rho_\pi = 1$ and finite function class), our bounds decay as $n^{-1/2}$ which is optimal whereas their bounds scale as $n^{-1/4}$ in the worst-case. Finally, we provide a lower bound for general function approximation.

8. Discussion

We study the statistical complexity of offline decision-making with value function approximation. We identify a large class of offline learning problems, characterized by the pseudo-dimension of the value function class and a new characterization of the offline data, for which we provide tight minimax lower and upper bounds. We also provide insights into the role of offline data for online decision-making from a minimax perspective.

We remark that our results do not imply that pseudo-dimension and policy transfer coefficients are necessary conditions for learnability in offline decision-making with function approximation. Consequently, there are several notable gaps in our current understanding. First, our lower bounds apply only to parametric function classes, i.e., for function classes with finite pseudo-dimension. Whereas our upper bounds hold also for non-parametric function classes. Can we provide a lower bound for general function classes, including non-parametric ones? Second, in the hybrid setting with a parametric function class, the upper bound of our adaptive algorithm matches the lower bound only in the regime that $m \geq \left(\frac{n}{C^*Kd}\right)^{1/\rho^*}$. Is it possible to design a fully adaptive algorithm (i.e., without the knowledge of policy transfer coefficients) whose upper bound matches the lower bound in any regime of m, including the practical scenarios where the online exploration budget m is small? Third, what are the necessary conditions for the value function class and the data quality that *fully* characterize the statistical complexity of offline decision-making with value function approximation?

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This research was supported, in part, by DARPA GARD award HR00112020004 and NSF CAREER award IIS-1943251.

References

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pp. 19–26. PMLR, 2012.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Attias, I., Hanneke, S., Kalavasis, A., Karbasi, A., and Velegkas, G. Optimal learners for realizable regression: PAC learning and online learning. *arXiv preprint arXiv:2307.03848*, 2023.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

David, S. B., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136. JMLR Workshop and Conference Proceedings, 2010.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Foster, D. and Rakhlin, A. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Foster, D., Agarwal, A., Dudík, M., Luo, H., and Schapire, R. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 1539–1548. PMLR, 2018.

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pp. 738–746. PMLR, 2013.

Hanneke, S. and Kpotufe, S. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Haussler, D. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69 (2):217–232, 1995.

Krishnamurthy, A., Agarwal, A., Huang, T.-K., Daumé III, H., and Langford, J. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 20 (65):1–50, 2019.

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

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. The importance of convexity in learning with squared loss. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pp. 140–146, 1996.

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

Li, G., Zhan, W., Lee, J. D., Chi, Y., and Chen, Y. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. *arXiv preprint arXiv:2305.10282*, 2023.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with stationary distribution correction. In Globerson, A. and Silva, R. (eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1180–1190. AUAI Press, 2019.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Multiple source adaptation and the Rényi divergence. *arXiv preprint arXiv:1205.2628*, 2012.

Mehta, N. A. and Williamson, R. C. From stochastic mixability to fast rates. *Advances in Neural Information Processing Systems*, 27, 2014.

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. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.*, 9:815–857, 2008.

Nguyen-Tang, T. and Arora, R. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Nguyen-Tang, T., Yin, M., Gupta, S., Venkatesh, S., and Arora, R. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9310–9318, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pollard, D. *Convergence of stochastic processes*. David Pollard, 1984.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Rigollet, P. and Zeevi, A. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

Slivkins, A. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pp. 679–702. JMLR Workshop and Conference Proceedings, 2011.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. Hybrid RL: Using both offline and online data can make RL efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Srebro, N., Sridharan, K., and Tewari, A. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Tang, D., Jain, R., Hao, B., and Wen, Z. Efficient online learning with offline datasets for infinite horizon mdps: A bayesian approach. *arXiv preprint arXiv:2310.11531*, 2023.

Tripuraneni, N., Jordan, M., and Jin, C. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33: 7852–7862, 2020.

Tsybakov, A. B. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Van Erven, T., Grunwald, P., Mehta, N. A., Reid, M., Williamson, R., et al. Fast rates in statistical and online learning. 2015.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

Vapnik, V. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280, 1971.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wagenmaker, A. and Pacchiano, A. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning*, pp. 35300–35338. PMLR, 2023.

Watkins, A., Ullah, E., Nguyen-Tang, T., and Arora, R. Optimistic rates for multi-task representation learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Yin, M. and Wang, Y.-X. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Zhang, R. and Zanette, A. Policy finetuning in reinforcement learning via design of experiments using offline data. *arXiv preprint arXiv:2307.04354*, 2023.

Zhang, T. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Zhang, T. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

A. Uniform Bernstein's Inequality

A central tool for our analysis is the uniform concentrability of a subset of functions near the ERM (empirical risk minimizer) predictor. In order to be able to match the lower bounds of offline decision-making (at least) for parametric classes, we require that every ball centered around the ERM predictor within a sufficiently small radius (in fact, of the order of $\mathcal{O}(\frac{1}{n})$) has an $\mathcal{O}(\frac{1}{n})$ order in the excess risk. While fast rates are in general not possible, they are so under certain conditions (Van Erven et al., 2015), including the bounded, squared loss in parametric classes that we consider. A uniform Bernstein's inequality is also presented in (Cucker & Smale, 2002, Theorem B), but using a strong notion of domain-wide L_∞ covering numbers. (Zhang, 2023, Theorem 3.21) presents a version of uniform Bernstein's inequality using the population L_1 covering numbers. The population L_1 however requires the knowledge of the data distribution. Here, we present a more practical version of uniform Bernstein's inequality using the empirical L_1 covering numbers, which is, at least in principle, computable given the empirical data. More importantly, and also as a key technical result in this section, we prove in Proposition A.3 a uniform Bernstein's inequality for Bellman-like loss functions using the empirical L_1 covering numbers. Proposition A.3 applies to handle the data structure generated by RL and, as a special case, naturally applies to contextual bandits. These results are central to our analysis tool and might be of independent interest. Notably, our proof for Proposition A.3 relies on an elegant argument of localizing a function class into a set of balls centered around the covering functions, and then performing uniform convergence in each of such local balls before combining them via a union bound. This localization argument is inspired by a similar argument by (Mehta & Williamson, 2014).

Before stating the uniform Bernstein's inequality for Bellman-like loss functions in Proposition A.3, we start with the uniform Bernstein's inequality for a simpler case, which we also use in our analysis and also serves a good point for demonstrating the localization argument.

A.1. Uniform Bernstein's inequality for generic case

Proposition A.1 (Uniform Bernstein's inequality for generic case)**.** *Let $\mathcal{G}$ be a set of functions $g : \mathcal{Z} \to [-b, b]$. Fix $n \in \mathbb{N}$. Denote $\hat{\mathbb{E}}g := \frac{1}{n}\sum_{i=1}^n g(z_i)$ where $\{z_i\}_i \overset{i.i.d.}{\sim} P$ and $\mathbb{E}g = \mathbb{E}_{z\sim P}[g(z)]$, and $\mathbb{V}[g]$ is the variance of $g(z)$. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\forall g \in \mathcal{G} : \mathbb{E}g - \hat{\mathbb{E}}g \le \inf_{\epsilon>0}\left\{\sqrt{\frac{2\mathbb{V}[g]\log(2N_1(\mathcal{G},\epsilon,n)/\delta)}{n}} + \frac{62b\log(6N_1(\mathcal{G},\epsilon,n)/\delta)}{n} + 61\epsilon\right\}.$$

Fix $\epsilon > 0$ and let $N = N_1(\mathcal{G}, \epsilon, n)$. Let $\{g_i\}_{i\in[N]}$ be an ϵ-cover of $\mathcal{G}$ w.r.t. $L_1(\hat{P}_n)$. For any $i \in [N]$, denote $\mathcal{G}_i = \{g \in \mathcal{G} : \hat{\mathbb{E}}|g - g_i| \le \epsilon\}$. We have $\mathcal{G} \subseteq \cup_{i\in[N]}\mathcal{G}_i$. The proof of Proposition A.1 relies on the following lemma.

Lemma A.2. *Fix any $i \in [N]$. With probability at least $1-\delta$,*

$$\sup_{g\in\mathcal{G}_i}\mathbb{E}|g - g_i| \le 12\epsilon + \frac{12b\log(3/\delta)}{n}.$$

Proof of Lemma A.2. The key idea to obtain fast rates in the above bounds is leveraging the *non-negativity* of the random objects of interest by employing Lemma F.2. With probability at least $1-\delta$, for any $g \in \mathcal{G}_i$, we have

$$\begin{aligned}\mathbb{E}|g - g_i| &\le 4\hat{\mathbb{E}}|g - g_i| + \inf_{\epsilon'}\left\{8\epsilon' + \frac{12b\log(3N_1(\mathcal{G}_i,\epsilon',n)/\delta)}{n}\right\}\\ &\le 12\epsilon + \frac{12b\log(3/\delta)}{n},\end{aligned}$$

where the second inequality follows from that $\hat{\mathbb{E}}|g - g_i| \le \epsilon$, that we choose $\epsilon' = \epsilon$, and that the ϵ-ball $\mathcal{G}_i$ can be ϵ-covered by one point, thus $N_1(\mathcal{G}_i, \epsilon_i, n) = 1$. □

Proof of Proposition A.1. Denote the event $E = E_1 \cap E_2$, where

$$E_1 = \left\{ \forall i \in [N], \mathbb{E}g_i - \hat{\mathbb{E}}g_i \le \frac{2b\log(N/\delta)}{3n} + \sqrt{\frac{2\mathbb{V}[g_i]\log(N/\delta)}{n}} \right\}$$
$$E_2 = \left\{ \sup_{i\in[N]} \sup_{g\in\mathcal{G}_i} \mathbb{E}|g - g_i| \le 12\epsilon + \frac{12b\log(3N/\delta)}{n} \right\}.$$

Now consider under the event E. Since $\mathcal{G} \subseteq \cup_{i\in[N]}\mathcal{G}_i$, for any $g \in \mathcal{G}$, there must exist $i \in [N]$ such that $g \in \mathcal{G}_i$ (consequently, $\hat{\mathbb{E}}|g - g_i| \le \epsilon$). Also notice that

$$\mathbb{V}[g] - \mathbb{V}[g_i] = \mathbb{E}[g^2] - \mathbb{E}[g_i^2] + \mathbb{E}[g_i]^2 - \mathbb{E}[g]^2 \le 2b\mathbb{E}|g - g_i| + 2b|\mathbb{E}[g_i] - \mathbb{E}[g]| \le 4b\mathbb{E}|g - g_i|.$$

Thus, we have

$$\begin{aligned}
\mathbb{E}g - \hat{\mathbb{E}}g &= \mathbb{E}g_i - \hat{\mathbb{E}}g_i + \mathbb{E}(g - g_i) + \hat{\mathbb{E}}(g_i - g) \\
&\le \mathbb{E}g_i - \hat{\mathbb{E}}g_i + \mathbb{E}|g - g_i| + \hat{\mathbb{E}}|g - g_i| \\
&\le \frac{2b\log(N/\delta)}{3n} + \sqrt{\frac{2\mathbb{V}[g_i]\log(N/\delta)}{n}} + \mathbb{E}|g - g_i| + \hat{\mathbb{E}}|g - g_i| \\
&\le \frac{2b\log(N/\delta)}{3n} + \sqrt{\frac{2\mathbb{V}[g]\log(N/\delta)}{n}} + \sqrt{\frac{2|\mathbb{V}[g] - \mathbb{V}[g_i]|\log(N/\delta)}{n}} + \mathbb{E}|g - g_i| + \hat{\mathbb{E}}|g - g_i| \\
&\le \frac{2b\log(N/\delta)}{3n} + \sqrt{\frac{2\mathbb{V}[g]\log(N/\delta)}{n}} + \frac{b\log(N/\delta)}{n} + \frac{1}{b}|\mathbb{V}[g] - \mathbb{V}[g_i]| + \mathbb{E}|g - g_i| + \hat{\mathbb{E}}|g - g_i| \\
&\le \frac{5b\log(N/\delta)}{3n} + \sqrt{\frac{2\mathbb{V}[g]\log(N/\delta)}{n}} + 5\mathbb{E}|g - g_i| + \hat{\mathbb{E}}|g - g_i| \\
&\le \frac{5b\log(N/\delta)}{3n} + \sqrt{\frac{2\mathbb{V}[g]\log(N/\delta)}{n}} + 5(12\epsilon + \frac{12b\log(3N/\delta)}{n}) + \epsilon \\
&\le \frac{62b\log(3N/\delta)}{n} + \sqrt{\frac{2\mathbb{V}[g]\log(N/\delta)}{n}} + 61\epsilon,
\end{aligned}$$

where the fourth inequality follows AM-GM inequality . Finally, note that, by Bernstein's inequality and the union bound, $\Pr(E_1) \ge 1-\delta$; by Lemma A.2 and the union bound, $\Pr(E_2) \ge 1-\delta$. Thus, by the union bound again, $\Pr(E) \ge 1-2\delta$. □

A.2. Uniform Bernstein's inequality for Bellman-like loss classes

We now establish the uniform Bernstein's inequality for the Bellman-like loss functions. The nature of the result and the proof is similar to those for Proposition A.1, but only more involved as we deal with a more structural random tuple.

Consider a tuple of random variables $(x, a, r, x') \in \mathcal{X} \times \mathcal{A} \times [0, 1] \times \mathcal{X}$ distributed according to distribution P. For any $u : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $g : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we define the random variable

$$M(u, g) = (u(x, a) - r - g(x'))^2 - (g^*(x, a) - r - g(x'))^2,$$

where $g^*(x, a) := \mathbb{E}_{(r,x')\sim P(\cdot|x,a)}[r + g(x')]$. Let $\{(x_t, a_t, r_t, x'_t)\}_{t\in[n]}$ be an i.i.d. sample from P. We write M_t in replace of M when we replace (x, a, r, x') in M by (x_t, a_t, r_t, x'_t). We consider function classes $\mathcal{U} \subseteq \{u : \mathcal{X} \times \mathcal{A} \to [0, b]\}$ and $\mathcal{G} \subseteq \{g : \mathcal{X} \to [0, b]\}$. We assume, for simplicity, that $r + g(x') \in [0, b]$ almost surely.

Proposition A.3 (Uniform Bernstein's inequality for Bellman-like loss functions)**.** *Fix any $\epsilon > 0$. With probability at least $1 - \delta$, for any $u \in \mathcal{U}, g \in \mathcal{G}$,*

$$\mathbb{E}[(u(x, a) - g^*(x, a))^2] \le \frac{2}{n}\sum_{t=1}^{n} M_t(u, g) + \inf_{\epsilon>0}\left\{108b\epsilon + b^2\frac{36\log N_1(\mathcal{U}, \epsilon, n) + 83\log N_1(\mathcal{G}, \epsilon, n) + 108\log(12/\delta)}{n}\right\}.$$

In addition, with probability at least $1 - \delta$, for any $u \in \mathcal{U}, g \in \mathcal{G}$,

$$-\frac{1}{n}\sum_{t=1}^{n} M_t(u, g) \le \inf_{\epsilon>0}\left\{30b\epsilon + b^2\frac{4\log N_1(\mathcal{U}, \epsilon, n) + 28\log N_1(\mathcal{G}, \epsilon, n) + 28\log(6/\delta)}{n}\right\}.$$

Remark A.4. Krishnamurthy et al. (2019) develop a uniform Freedman-type inequality for an indicator-type martingale difference sequence, which does not require the data-generating policy to be non-adaptive as in our case. Applying the uniform Freedman-type inequality of Krishnamurthy et al. (2019) "as-is" to the stochastic contextual bandit with a non-adaptive data-generating policy results in larger constants and an additional factor of $\log(K)$ (see (Foster et al., 2018) for why this additional $\log K$ is the case). However, there are deeper reasons why our uniform Bernstein's inequality is preferred over the uniform Freedman-type inequality of Krishnamurthy et al. (2019). First, it seems impossible even if we want to apply the uniform Freedman-type inequality of Krishnamurthy et al. (2019) "as-is" to the RL setting. The reason is that the martingale structure in the uniform Freedman-type inequality of Krishnamurthy et al. (2019) is formed only through the adaptive policy that selects actions based on the historical data, and they require i.i.d. assumption on $\{x_t\}_{t\geq 1}$. In contrast, in RL, even when we assume that the initial state is i.i.d. from a fixed distribution, the states from $h \geq 2$ are not independent once the policy that generates the offline data is adaptive. Second, the uniform Freedman-type inequality of Krishnamurthy et al. (2019) cannot seem to easily apply to the RL setting considered in Proposition A.1, where we consider multiple target regression functions g^*, whereas they require one fixed target function.

Fix $\epsilon > 0$. For simplicity, we write $N_g = N_1(\mathcal{G}, \epsilon, n)$ and $N_u = N_1(\mathcal{U}, \epsilon, n)$.

Lemma A.5. *For any $u \in \mathcal{U}$ and any $g \in \mathcal{G}$, we have*

$$\mathbb{E}[M(u,g)] = \mathbb{E}[(u(x,a) - g^*(x,a))^2],$$
$$\mathbb{V}[M(u,g)] \leq 4b^2\mathbb{E}[M(u,g)].$$

Let $\{u_i\}_{i\in N_u(\epsilon)}$ and $\{g_j\}_{j\in N_g(\epsilon)}$ be the corresponding ϵ-cover of $\mathcal{U}$ and $\mathcal{G}$. Let $\mathcal{U}_i$ be the set of functions $u \in \mathcal{U}$ such that $\|u - u_i\|_{L_1(\{(x_t,a_t)\}_{t\in[n]})} \leq \epsilon$. Similarly, we define $\mathcal{G}_j$.

The following lemma establishes a finite function class variant of Proposition A.3, over (a finite number of) the covering functions of $\mathcal{U}$ and $\mathcal{G}$.

Lemma A.6. *With probability $1-\delta$, for any $(i,j) \in [N_u] \times [N_g]$, we have*

$$\mathbb{E}[(u_i(x,a) - g_j^*(x,a))^2] \leq \frac{2}{n}\sum_{t=1}^{n} M_t(u_i, g_j) + \frac{12b^2\log(N_u N_g/\delta)}{n}.$$

In addition, with probability $1-\delta$, for any $(i,j) \in [N_u] \times [N_g]$, we have

$$-\frac{1}{n}\sum_{t=1}^{n} M_t(u_i, g_j) \leq \frac{4b^2\log(N_u N_g/\delta)}{n}.$$

Proof of Lemma A.6. For any $(i,j) \in [N_u] \times [N_g]$, by the Freedman's inequality, with probability at least $1-\delta$, for any $t \in [0, \frac{1}{b^2}]$

$$\begin{aligned}\mathbb{E}[(u_i(x,a) - g_j^*(x,a))^2] &\leq \frac{1}{n}\sum_{t=1}^{n} M_t(u_i, g_j) + (e-2)\lambda\mathbb{V}[M(u_i,g_j)] + \frac{\log(1/\delta)}{\lambda n} \\ &\leq \frac{1}{n}\sum_{t=1}^{n} M_t(u_i, g_j) + 4b^2(e-2)\lambda\mathbb{E}M(u_i,g_j) + \frac{\log(1/\delta)}{\lambda n},\end{aligned}$$

where the second inequality follows from Lemma A.5 (the second inequality). By choosing $\lambda = \frac{1}{8(e-2)b^2}$, the above inequality becomes:

$$\mathbb{E}[(u_i(x,a) - g_j^*(x,a))^2] \leq \frac{2}{n}\sum_{t=1}^{n} M_t(u_i, g_j) + \frac{12\log(1/\delta)}{n}.$$

Taking the union bound over $(i,j) \in [N_u] \times [N_g]$ completes the proof for the first part.

The second part follows similarly except the only difference is that we choose $\lambda = \frac{1}{4(e-2)b^2}$ in the Freedman's inequality. □

The following lemma shows that a ball in the L_1 distance is provably contained within a ball with an empirical L_1 distance with a slightly larger radius by a margin that scales at "fast rates" $\tilde{O}(\frac{1}{n})$.

Lemma A.7. *1. Fix $i \in [N_u]$. With probability at least $1-\delta$,*

$$\sup_{u \in \mathcal{U}_i} \mathbb{E}|u - u_i| \le 12\epsilon + \frac{12b\log(3/\delta)}{n}.$$

2. Fix $j \in [N_g]$. With probability at least $1-\delta$,

$$\sup_{g \in \mathcal{G}_j} \mathbb{E}|g - g_j| \le 12\epsilon + \frac{12b\log(3/\delta)}{n}.$$

3. Fix $j \in [N_g]$. With probability at least $1-\delta$, we have

$$\sup_{g \in \mathcal{G}_j} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{x' \sim P(\cdot|x_t,a_t)}|g - g_j|(x') \le 12\epsilon + \frac{12b\log(3/\delta)}{n}.$$

Proof of Lemma A.7. The nature of these results are similar to Lemma A.2 in the generic case, where the key idea to obtain fast rates in the estimation errors is leveraging the *non-negativity* of the random objects of interest by employing Lemma F.2. We start with Part 1. With probability at least $1-\delta$, for any $u \in \mathcal{U}_i$, we have

$$\mathbb{E}|u - u_i| \le 4\hat{\mathbb{E}}|u - u_i| + \inf_{\epsilon'}\left\{8\epsilon' + \frac{12b\log(3N_1(\epsilon', \mathcal{U}_i, n)/\delta)}{n}\right\} \le 12\epsilon + \frac{12b\log(3/\delta)}{n}.$$

where the second inequality follows from that $\hat{\mathbb{E}}|u - u_i| \le \epsilon$, that we choose $\epsilon' = \epsilon$, and that the ϵ-ball $\mathcal{U}_i$ can be ϵ-covered by one point, thus $N_1(\epsilon, \mathcal{U}_i, n) = 1$.

Part 2 follows exactly as Part 1. For Part 3, we notice

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}|g - g_j|(x'_t)|\{(x_t, a_t)\}_{t\in[n]}\right] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{x' \sim P(\cdot|x_t,a_t)}|g - g_j|(x').$$

Thus, the same arguments in Part 1 and Part 2 apply to Part 3. □

We are now ready to prove Proposition A.3.

Proof of Proposition A.3. We start with the following simple discretization: For any u, u', g, g', we have

$$M(u, g) - M(u', g') \le 2b|u - u'|(x, a) + 4b|g - g'|(x') + 2b\mathbb{E}_{x' \sim P(\cdot|x,a)}|g - g'|(x'), \tag{5}$$

$$(u - g_*)^2(x, a) - (u' - g'_*)^2(x, a) \le 2b|u - u'|(x, a) + 2b\mathbb{E}_{x' \sim P(\cdot|x,a)}|g - g'|(x'). \tag{6}$$

Consider the following event:

$$\begin{aligned} E_1 = &\left\{\sup_{i\in[N_u]}\sup_{u\in\mathcal{U}_i}\mathbb{E}|u - u_i| \le 12\epsilon + \frac{12b\log(3N_u/\delta)}{n}\right\} \cap \\ &\left\{\sup_{j\in[N_g]}\sup_{g\in\mathcal{G}_j}\mathbb{E}|g - g_j| \le 12\epsilon + \frac{12b\log(3N_g/\delta)}{n}\right\} \cap \\ &\left\{\sup_{j\in[N_g]}\sup_{g\in\mathcal{G}_j}\hat{\mathbb{E}}\mathbb{E}_{x' \sim P(\cdot|x,a)}|g - g_j|(x') \le 12\epsilon + \frac{12b\log(3N_g/\delta)}{n}\right\} \cap \\ &\left\{\mathbb{E}[(u_i(x, a) - g_j^*(x, a))^2] \le \frac{2}{n}\sum_{t=1}^{n} M_t(u_i, g_j) + \frac{12b^2\log(N_u N_g/\delta)}{n}\right\}. \end{aligned}$$

Note that for any $u \in \mathcal{U}$ and $g \in \mathcal{G}$, there exist $i \in [N_u]$ and $j \in N_g$ such that $u \in \mathcal{U}_i$ and $g \in \mathcal{G}_j$. Thus, under event E_1, for any $u \in \mathcal{U}, g \in \mathcal{G}$, we have

$$\begin{aligned}
\mathbb{E}[(u(x,a) - g_*(x,a))^2] &\le \mathbb{E}[(u_i(x,a) - g_j^*(x,a))^2] + 2b\mathbb{E}|u - u_i| + 2b\mathbb{E}|g - g_j| \\
&\le \frac{2}{n}\sum_{t=1}^{n} M_t(u_i, g_j) + \frac{12b^2\log(N_u N_g/\delta)}{n} + 2b\mathbb{E}|u - u_i| + 2b\mathbb{E}|g - g_j| \\
&\le \frac{2}{n}\sum_{t=1}^{n} M_t(u, g) + 4b\hat{\mathbb{E}}|u - u'| + 8b\hat{\mathbb{E}}|g - g'| + 4b\hat{\mathbb{E}}\mathbb{E}_{x'\sim P(\cdot|x,a)}|g - g'|(x') \\
&+ \frac{12b^2\log(N_u N_g/\delta)}{n} + 2b\mathbb{E}|u - u_i| + 2b\mathbb{E}|g - g_j| \\
&\le \frac{2}{n}\sum_{t=1}^{n} M_t(u, g) + 4b\epsilon + 8b\epsilon + 4b(12\epsilon + \frac{12b\log(3N_g/\delta)}{n}) \\
&+ \frac{12b^2\log(N_u N_g/\delta)}{n} + 2b(12\epsilon + \frac{12b\log(3N_u/\delta)}{n}) + 2b(12\epsilon + \frac{12b\log(3N_g/\delta)}{n}) \\
&\le \frac{2}{n}\sum_{t=1}^{n} M_t(u, g) + 108b\epsilon + b^2\frac{36\log(N_u) + 83\log(N_g) + 108\log(3/\delta)}{n},
\end{aligned}$$

where the first inequality follows from Equation (6), the second inequality follows from E_1, the third inequality follows from Equation (5), and the fourth inequality follows from E_1. By Lemma A.7, Lemma A.6 and the union bound, we have $\Pr(E_1) \ge 1 - 4\delta$. Thus, we complete the first part of the theorem.

For the second part, the proof follows similarly. In particular, we consider the following event:

$$\begin{aligned}
E_2 = &\left\{\sup_{j\in[N_g]}\sup_{g\in\mathcal{G}_j} \hat{\mathbb{E}}\mathbb{E}_{x'\sim P(\cdot|x,a)}|g - g_j|(x') \le 12\epsilon + \frac{12b\log(3N_g/\delta)}{n}\right\} \cap \\
&\left\{-\frac{1}{n}\sum_{t=1}^{n} M_t(u_i, g_j) \le \frac{4b^2\log(N_u N_g/\delta)}{n}\right\}.
\end{aligned}$$

It follows from Lemma A.6 (second part), Lemma A.7, and a union bound, that $\Pr(E_2) \ge 1 - 2\delta$.

Finally, note that under event E_2, for any $u \in \mathcal{U}, g \in \mathcal{G}$, we have

$$\begin{aligned}
-\frac{1}{n}\sum_{t=1}^{n} M_t(u, g) &\le -\frac{1}{n}\sum_{t=1}^{n} M_t(u_i, g_j) + 2b\hat{\mathbb{E}}|u - u_i| + 4b\hat{\mathbb{E}}|g - g_j| + 2b\hat{\mathbb{E}}\mathbb{E}_{x'\sim P(\cdot|x,a)}|g - g_j|(x') \\
&\le \frac{4b^2\log(N_u N_g/\delta)}{n} + 2b\epsilon + 4b\epsilon + 2b(12\epsilon + \frac{12b\log(3N_g/\delta)}{n}) \\
&\le 30b\epsilon + b^2\frac{4\log(N_u) + 28\log(N_g) + 28\log(3/\delta)}{n},
\end{aligned}$$

where the first inequality follows from Eq. (5), the second inequality follows from the conditions in event E_2.

□

B. Proofs of Section 4

Proof of Theorem 4.1. Fix any (C, ρ, d) as stated in the theorem. Let $\epsilon = \left(\frac{Cd}{32n}\right)^{1/(2\rho)}$. We have $0 < \epsilon < 1/2$ and $\frac{\epsilon^{2\rho-2}}{C} \le 1$.

Construction of hard instances. Let $\mathcal{A} = \{a_1, a_2\}$. Pick any d mutually distinct points $x_1, \ldots, x_d$. We construct a family of the context-reward distributions $\mathcal{D}_\sigma$ indexed by $\sigma \in \{-1, 1\}^d$ where $\mathcal{D}_\sigma(x, y) = P_X(x) \times P^\sigma_{Y|X}(y)$, $P_X(x_i) = \frac{1}{d}$, $P^\sigma_{Y|X}(y(a_1)|x_i) = \text{Ber}(\frac{1}{2})$, and $P^\sigma_{Y|X}(y(a_2)|x_i) = \text{Ber}(\frac{1}{2} + \sigma_i\frac{\epsilon}{2})$. Choose any μ such that

$\mu(a_2|x_i) = \frac{\epsilon^{2\rho-2}}{C} \leq 1, \forall i \in [d]$. Now we choose the function class $\mathcal{F}$ as follows: $\mathcal{F}_{a_1} = \{f(x) \equiv \frac{1}{2}\}$ and $\mathcal{F}_{a_2} = \{f_\alpha(x_i) := \frac{1}{2} + \alpha_i \frac{\epsilon}{2}, \forall i \in [d] | \alpha \in \{-1,1\}^d\}$.

Verification. We now verify that for any σ, $(\mathcal{D}_\sigma, \mu, \pi^*_{\mathcal{D}_\sigma}, \mathcal{F}) \in \mathcal{B}(\rho, C, d)$. First, it is easy to see that $\text{Pdim}(\mathcal{F}_{a_2}) = d$ and $\text{Pdim}(\mathcal{F}_{a_1}) = 0$. Second, we have $f^*_\sigma(x_i, a_1) = \frac{1}{2}$ and $f^*_\sigma(x_i, a_2) = \frac{1}{2} + \sigma_i \frac{\epsilon}{2}$, where $f^*_\sigma(x, a) := \mathbb{E}_{P_\sigma}[Y(a)|X = x]$, thus $f^*_\sigma \in \mathcal{F}, \forall \sigma$. Third, for any $f \in \mathcal{F}$, we have $f(\cdot, a_2) = f_\alpha$ for some $\alpha \in \{-1,1\}^d$ and for any policy π, we have

$$\mathcal{E}_{\sigma,\pi}(f) = \epsilon^2 \sum_{i=1}^{d} \pi(a_2|x_i) \frac{\mathbb{1}\{\sigma_i \neq \alpha_i\}}{d},$$

where $\mathcal{E}_{\sigma,\pi}(f)$ denotes the excess risk of f under $\mathcal{D}_\sigma \otimes \pi$, i.e., under the squared loss and realizability, $\mathcal{E}_{\sigma,\pi}(f) = \mathbb{E}_{\mathcal{D}_\sigma \otimes \pi}[(f - f^*)^2]$.

Since we choose $\mu(a_2|x_i) = \frac{\epsilon^{2\rho-2}}{C} \leq 1, \forall i \in [d]$, we have

$$\mathcal{E}_{\sigma,\mu}(f) = \frac{\epsilon^{2\rho}}{C} \frac{\text{dist}(\sigma, \alpha)}{d} \geq \frac{1}{C} \left(\epsilon^2 \frac{\text{dist}(\sigma, \alpha)}{d} \right)^\rho \geq \frac{1}{C} \mathcal{E}^\rho_{\sigma,\pi^*}(f),$$

since $\text{dist}(\sigma, \alpha) \leq d$ and $\rho \geq 1$. Thus, we have $(\mathcal{D}_\sigma, \mu, \pi^*_{P_\sigma}, \mathcal{F}) \in \mathcal{B}(\rho, C, d)$ for any $\sigma \in \{-1,1\}^d$.

Reduction to testing. For any policy π, we have $V^\pi_\sigma := V^\pi_{\mathcal{D}_\sigma} = \sum_{i=1}^d \frac{1}{d} \left(\pi(a_1|x_i)(\frac{1}{2}) + \pi(a_2|x_i)(\frac{1}{2} + \sigma_i \frac{\epsilon}{2}) \right)$. Thus, we can compute the sub-optimality of the output policy $\hat{\pi}$ of any offline learner by

$$V^*_\sigma - V^{\hat{\pi}}_\sigma = \sum_{i=1}^{d} \frac{\epsilon}{2d} \left(\mathbb{1}\{\sigma_i\} - \hat{\pi}(a_2|x_i)\sigma_i \right).$$

Let $\hat{\sigma}_i = \mathbb{1}\{\hat{\pi}(a_2|x_i) \geq \frac{1}{2}\}$. We have $\mathbb{1}\{\sigma_i\} - \hat{\pi}(a_2|x_i)\sigma_i \geq \frac{|\sigma_i - \hat{\sigma}_i|}{4}$, and thus, we have

$$\mathbb{E}_\sigma \left[V^*_\sigma - V^\pi_\sigma \right] \geq \frac{\epsilon}{4d} \mathbb{E}_\sigma \left[\text{dist}(\sigma, \hat{\sigma}) \right], \tag{7}$$

where $\text{dist}(\sigma, \hat{\sigma}) := \sum_{i=1}^d \mathbb{1}\{\sigma_i \neq \hat{\sigma}_i\}$ denotes the Hamming distance between two binary vectors σ and $\hat{\sigma}$, and $\mathbb{E}_\sigma$ denotes the expectation over the randomness of $\hat{\sigma}$ with respect to P_σ.

The worst-case Hamming distance $\sup_{\sigma \in \{-1,1\}^d} \mathbb{E}_\sigma \left[\text{dist}(\sigma, \hat{\sigma}) \right]$ can be lower-bounded using the standard tools in hypothesis testing:

$$\begin{aligned} \sup_{\sigma \in \{-1,1\}^d} \mathbb{E}_\sigma \left[\text{dist}(\sigma, \hat{\sigma}) \right] &\geq \frac{d}{2} \min_{\sigma, \sigma' : \text{dist}(\sigma, \sigma') = 1} \inf_\psi \left[\mathcal{D}_\sigma(\psi \neq \sigma) + \mathcal{D}_{\sigma'}(\psi \neq \sigma') \right] \\ &\geq \frac{d}{2} \left(1 - \sqrt{\frac{1}{2} \max_{\sigma, \sigma' : \text{dist}(\sigma, \sigma') = 1} \text{KL}\left((\mathcal{D}_\sigma \otimes \mu)^n \| (\mathcal{D}_{\sigma'} \otimes \mu)^n \right)} \right), \end{aligned} \tag{8}$$

where the first inequality follows Assouad's lemma (Tsybakov, 1997, Lemma 2.12) and the second inequality follows from (Tsybakov, 1997, Theorem 2.12).

We now compute the KL distance: For any σ and σ' such that $\text{dist}(\sigma, \sigma') = 1$, let $i^* \in [d]$ be the (only) coordinate that σ

differs from σ', we have

$$\begin{aligned}
&\mathrm{KL}\left((\mathcal{D}_\sigma \otimes \mu)^n \| (\mathcal{D}_{\sigma'} \otimes \mu)^n\right) = n\mathrm{KL}\left(\mathcal{D}_\sigma \otimes \mu \| \mathcal{D}_{\sigma'} \otimes \mu\right) \\
&= n\mathbb{E}_{P_X}\mathrm{KL}\left(\mu \otimes P_\sigma(Y|X) \| \mu \otimes P_{\sigma'}(Y|X)\right) \\
&= \frac{n}{d}\sum_{i=1}^{d}\mathrm{KL}\left(\mu(a_1|x_i)\mathrm{Ber}(\frac{1}{2}) + \mu(a_2|x_i)\mathrm{Ber}(\frac{1}{2}+\sigma_i\frac{\epsilon}{2}) \middle\| \mu(a_1|x_i)\mathrm{Ber}(\frac{1}{2}) + \mu(a_2|x_i)\mathrm{Ber}(\frac{1}{2}+\sigma_i'\frac{\epsilon}{2})\right) \\
&\leq \frac{n}{d}\sum_{i=1}^{d}\mu(a_2|x_i)\mathrm{KL}\left(\mathrm{Ber}(\frac{1}{2}+\sigma_i\frac{\epsilon}{2}) \middle\| \mathrm{Ber}(\frac{1}{2}+\sigma_i'\frac{\epsilon}{2})\right) \\
&= \frac{n}{d}\mu(a_2|x_{i^*})\mathrm{KL}\left(\mathrm{Ber}(\frac{1}{2}+\sigma_{i^*}\frac{\epsilon}{2}) \middle\| \mathrm{Ber}(\frac{1}{2}+\sigma_{i^*}'\frac{\epsilon}{2})\right) \\
&\leq 16\frac{n}{d}\mu(a_2|x_{i^*})\epsilon^2 = \frac{16n\epsilon^{2\rho}}{Cd} = \frac{1}{2},
\end{aligned}$$

where the first inequality uses the convexity of KL divergence, the second inequality uses a basic KL upper bound that $\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2}+z\frac{\epsilon}{2}\right) \middle\| \mathrm{Ber}\left(\frac{1}{2}-z\frac{\epsilon}{2}\right)\right) \leq 16\epsilon^2$ for any $z \in \{-1, 1\}$, and the second-last equality plugs in the choice of μ, and the last equality plugs in the choice of ϵ. Now, plugging the above inequality into Equation (8) and Equation (7), we have

$$\max_{\sigma\in\{-1,1\}^d}\mathbb{E}_\sigma\left[V_\sigma^* - V_\sigma^\pi\right] \geq \frac{\epsilon}{16} = \frac{1}{16}\left(\frac{Cd}{32n}\right)^{1/(2\rho)}.$$

□

C. Proofs of Section 5

We restate Theorem 5.1 in an exact form with disclosed constants.

Theorem C.1. *Fix any $\delta \in [0,1]$. Invoke Algorithm 1 with $\beta = 32\epsilon + \frac{4\log N_1(\mathcal{F},\epsilon,n)+24\log(12/\delta)}{n}$. Then, for any $(\mathcal{D},\mathcal{F})$ such that Assumption 2.1 holds, with probability at least $1-\delta$, for any $\pi \in \Pi$,*

$$\begin{aligned}
\mathbb{E}[\mathrm{SubOpt}_{\mathcal{D}}^\pi(\hat{\pi})|S] \leq &\left(C_{\rho(\pi)}\inf_{\epsilon\geq 0}\left\{172\epsilon + \frac{44\log N_1(\mathcal{F},\epsilon,n) + 156\log(24/\delta)}{n}\right\}\right)^{1/(2\rho(\pi))} + 4\sqrt{\frac{\log K}{T}} \\
&+ \mathbb{1}_{\{|\mathcal{X}|>1\}}\cdot\inf_{\epsilon\geq 0}\left\{\sqrt{\frac{2\log(2N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n)/\delta)}{n}} + \frac{62\log(6N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n)/\delta)}{n} + 61\epsilon\right\}.
\end{aligned}$$

Our proof relies on the following lemmas.

Lemma C.2. *Fix any $\epsilon > 0$, and $\delta \in [0,1]$. Define the version space*

$$\mathcal{F}(\beta) := \{f \in \mathcal{F} : \hat{P}l_f - \hat{P}l_{\hat{f}} \leq \beta\},$$

where $\beta = 32\epsilon + \frac{4\log N_1(\mathcal{F},\epsilon,n)+24\log(12/\delta)}{n}$. Then, with probability at least $1-\delta$, $f^* \in \mathcal{F}(\beta)$, and

$$\sup_{f\in\mathcal{F}(\beta)}\mathbb{E}_{\mathcal{D}\otimes\mu}[(f-f^*)^2] \leq 172\epsilon + \frac{44\log N_1(\mathcal{F},\epsilon,n) + 156\log(24/\delta)}{n}.$$

Proof of Lemma C.2. Lemma C.2 is an instantiation of Proposition A.3 for the contextual bandit case where we have a tuple of random variables (x,a,r) instead of having an additional transition variable x'. In the contextual bandit case, we simply use $\mathcal{G} = \{0\}$ the set of the zero function, and remove all the terms regarding $N_1(\mathcal{G},\epsilon,n)$ (which is 1 in this case) in the RHS of Proposition A.3. □

Lemma C.3. *Choose* $\eta = \sqrt{\frac{\log K}{4(e-2)T}}$ *and* $T \geq \frac{\log K}{e-2}$. *Then we have*

$$\forall \pi, \sum_{t=1}^{T} \hat{P}(f_t(\cdot,\pi) - f_t(\cdot,\pi_t)) \leq 4\sqrt{T \log K}.$$

Proof of Lemma C.3. It suffices to show a stronger result, that any $x \in \mathcal{X}$ and $\pi \in \Pi$,

$$\sum_{t=1}^{T} f_t(x,\pi) - f_t(x,\pi_t) \leq 4\sqrt{T \log K}.$$

This is a standard guarantee of the Hedge algorithm, where a complete proof can be found at (Nguyen-Tang & Arora, 2023, Lemma B.6). □

Proof of Theorem C.1.

$$\begin{aligned}
V^{\pi} - V^{\pi_t} &= Pf^*(\cdot,\pi) - Pf^*(\cdot,\pi_t) \\
&= \mathbb{E}_{\mathcal{D}\otimes\pi}[f^* - f_t] + (P-\hat{P})f_t(\cdot,\pi) + \hat{P}(f_t(\cdot,\pi) - f_t(\cdot,\pi_t)) + (\hat{P}-P)f^*(\cdot,\pi_t) + \hat{P}(f_t(\cdot,\pi_t) - f^*(\cdot,\pi_t)), \\
&\leq \underbrace{\mathbb{E}_{\mathcal{D}\otimes\pi}[f^* - f_t]}_{I_1} + \underbrace{\hat{P}(f_t(\cdot,\pi) - f_t(\cdot,\pi_t))}_{I_2} + 2\underbrace{\sup_{g\in\mathcal{F}(\cdot,\Pi)} |(P-\hat{P})g|}_{I_3} + \underbrace{\hat{P}(f_t(\cdot,\pi_t) - f^*(\cdot,\pi_t))}_{I_4}.
\end{aligned}$$

We bound each term I_1, I_2, I_3, I_4 separately. Regarding term I_2, by Lemma C.3, we have

$$\sum_{t=1}^{T} \hat{P}(f_t(\cdot,\pi) - f_t(\cdot,\pi_t)) \leq 4\sqrt{T \log K}.$$

Consider the event $E = E_1 \cap E_2 \cap E_3$, where

$$\begin{aligned}
E_1 &:= \left\{ \sup_{g\in\mathcal{F}(\cdot,\Pi)} |(P-\hat{P})g| \leq \mathbb{1}_{\{|\mathcal{X}|>1\}} \inf_{\epsilon>0} \left\{ \sqrt{\frac{2\log(2N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n)/\delta)}{n}} + \frac{62\log(6N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n)/\delta)}{n} + 61\epsilon \right\} \right\}, \\
E_2 &:= \{f^* \in \mathcal{F}(\beta)\}, \\
E_3 &:= \left\{ \sup_{f\in\mathcal{F}(\beta)} \mathbb{E}_{\mathcal{D}\otimes\mu}[(f-f^*)^2] \leq 172\epsilon + \frac{44\log N_1(\mathcal{F},\epsilon,n) + 156\log(24/\delta)}{n} \right\}.
\end{aligned}$$

Due to pessimism of Algorithm 1, $f_t = \underset{f\in\mathcal{F}:\hat{P}l_f - \hat{P}l_{\hat{f}} \leq \beta}{\arg\min} \hat{P}f(\cdot,\pi_t)$. Thus, under E_2, $I_4 = \hat{P}f_t(\cdot,\pi_t) - \hat{P}f^*(\cdot,\pi_t) \leq 0$. By the policy transfer definition, we have

$$I_1 = \mathbb{E}_{\mathcal{D}\otimes\pi}[f^* - f_t] \leq \left(C_\pi \mathbb{E}_{\mathcal{D}\otimes\mu}[(f^*-f)^2]\right)^{1/(2\rho_\pi)}.$$

Thus, under event E, we have

$$\begin{aligned}
V^{\pi} - \frac{1}{T}\sum_{t=1}^{T} V^{\pi_t} \leq& \left(C_\pi \left(172\epsilon + \frac{44\log N_1(\mathcal{F},\epsilon,n) + 156\log(24/\delta)}{n} \right) \right)^{1/(2\rho_\pi)} + 4\sqrt{\frac{\log K}{T}} \\
&+ \mathbb{1}\{|\mathcal{X}|>1\} \wedge \inf_{\epsilon>0} \left\{ \sqrt{\frac{2\log(2N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n)/\delta)}{n}} + \frac{62\log(6N_1(\mathcal{F}(\cdot,\Pi),\epsilon,n)/\delta)}{n} + 61\epsilon \right\}.
\end{aligned}$$

Now we bound the probability of each of the events E_1, E_2, E_3 and combine them via the union bound for the probability of E. Note that if $|\mathcal{X}| = 1$, we have a trivial inequality $I_3 = 0$. Combining with Proposition A.1, we have $\Pr(E_1) \geq 1-\delta$. By Lemma C.2, we have $\Pr(E_2 \cap E_3) \geq 1-\delta$. Thus, $\Pr(E) \geq 1-2\delta$.

□

D. Proofs for Section 6

D.1. The expected sub-optimality guarantees for FALCON+ (Simchi-Levi & Xu, 2022, Algorithm 2)

When we employ FALCON+ (Simchi-Levi & Xu, 2022, Algorithm 2) for the online learning step in Algorithm 2, the returned policy $\hat{\pi}_{\text{on}}$ has the expected regret over $m/2$ rounds of order $\sqrt{K\mathcal{E}_{\mathcal{F},\delta/\log(m/2)}(m)m}$, where $\mathcal{E}_{\mathcal{F},\delta}(n)$ is a learning guarantee for the offline regression oracle (see their Assumption 2).

Assumption D.1 ((Simchi-Levi & Xu, 2022, Assumption 2))**.** Let π be an arbitrary policy. Given n training samples of the form $(x_i, a_i, r_i(a_i))$ i.i.d. drawn according to $(x_i, r_i) \sim \mathcal{D}$ and $a_i \sim \pi(\cdot|x_i)$, the offline regression oracle returns a predictor $\hat{f} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. For any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}},a\sim\pi(\cdot|x)}\left[(\hat{f}(x,a) - f^*(x,a))^2\right] \leq \mathcal{E}_{\mathcal{F},\delta}(n).$$

Note that, the output for running FALCON++ for $m/2$ iterates is a sequence of $m/2$ policies. $\hat{\pi}_{\text{on}}$ is an uniform distribution over such $m/2$ iterates.

Since the context x is i.i.d. from $\mathcal{D}_{\mathcal{X}}$, we can transform the expected regret $\sqrt{K\mathcal{E}_{\mathcal{F},\delta/\log(m/2)}(m)m}$ into the expected sub-optimality bound of order $\sqrt{K\mathcal{E}_{\mathcal{F},\delta/\log(m/2)}(m)} + \frac{1}{m}$ using the improved online-to-batch argument (Nguyen-Tang et al., 2023), wherein $\frac{1}{m}$ is the cost of converting a regret bound into a sub-optimality bound. It is known that if the closure of $\mathcal{F}$ is convex (which we assume here for simplicity of comparison), the sample complexity of learning $\mathcal{F}$ using squared loss is $\frac{\text{Pdim}(\mathcal{F})}{\epsilon}(\log(1/\epsilon) + \log(1/\delta))$ (Lee et al., 1996). Thus, $\mathcal{E}_{\mathcal{F},\delta}(n) \leq \frac{\text{Pdim}(\mathcal{F})}{n}\log(n/\delta)$. Finally, note that $\text{Pdim}(\mathcal{F}) \leq Kd$.

D.2. Proof of Theorem 6.1

Proof of Theorem 6.1. Construct the hard problem instances exactly like the ones in the proof of Theorem 4.1, except that we now choose

$$\epsilon = \min\left\{\left(\frac{Cd}{64n}\right)^{1/(2\rho)}, \frac{1}{8}\sqrt{\frac{d}{m}}\right\}.$$

The verification and the reduction to testing follow exactly, except for the step in which we compute the KL divergence of the observations between two different models. Specifically, denote Q_σ as the probability of the pre-collected data and the online data under the model P_σ. Let $S_{\text{on}} = \{\tilde{x}_i, \tilde{a}_i, \tilde{r}_i\}_{i\in[m]}$ be the random online data collected during the online phase by a hybrid algorithm ALG. Note that $\tilde{a}_t$ depends on $\tilde{x}_t, \{\tilde{x}_i, \tilde{a}_i, \tilde{r}_i\}_{i\in[t-1]}$, and S_{off} for any $t \in [m]$. We denote this conditional distribution by $P_{\text{ALG}}(\tilde{a}_t|\tilde{x}_t, \{\tilde{x}_i, \tilde{a}_i, \tilde{r}_i\}_{i\in[t-1]}, S_{\text{off}})$. Note that the conditional distribution P_{ALG} depends only on the algorithm ALG and invariant to any underlying model P_σ, thus we have

$$\frac{Q_\sigma(S_{\text{on}})}{Q_{\sigma'}(S_{\text{on}})} = \prod_{i\in[m]} \frac{P_\sigma(\tilde{r}_i|\tilde{x}_i, \tilde{a}_i)}{P_{\sigma'}(\tilde{r}_i|\tilde{x}_i, \tilde{a}_i)}.$$

Consider any σ and σ' such that $\text{dist}(\sigma, \sigma') = 1$. We thus have

$$\begin{aligned}\sum_{S_{\text{off}}\cup S_{\text{on}}} Q_\sigma(S_{\text{on}}) \log \frac{Q_\sigma(S_{\text{on}})}{Q_{\sigma'}(S_{\text{on}})} &= \sum_{S_{\text{off}}\cup S_{\text{on}}} Q_\sigma(S_{\text{on}}) \prod_{i\in[m]} \frac{P_\sigma(\tilde{r}_i|\tilde{x}_i, \tilde{a}_i)}{P_{\sigma'}(\tilde{r}_i|\tilde{x}_i, \tilde{a}_i)} \\ &\leq \frac{m}{d}\sum_{i=1}^{d} \text{KL}\left(\text{Ber}(\frac{1}{2} + \sigma_i\frac{\epsilon}{2})\middle\|\text{Ber}(\frac{1}{2} + \sigma_i'\frac{\epsilon}{2})\right) \leq \frac{m}{d}16\epsilon^2.\end{aligned}$$

Hence, we have

$$\begin{aligned}\text{KL}(Q_\sigma\|Q_{\sigma'}) &= \sum_{S_{\text{off}}\cup S_{\text{on}}} Q_\sigma(S_{\text{off}} \cup S_{\text{on}}) \log \frac{Q_\sigma(S_{\text{off}} \cup S_{\text{on}})}{Q_{\sigma'}(S_{\text{off}} \cup S_{\text{on}})} \\ &= \sum_{S_{\text{off}}} Q_\sigma(S_{\text{off}}) \log \frac{Q_\sigma(S_{\text{off}})}{Q_{\sigma'}(S_{\text{off}})} + \sum_{S_{\text{off}}\cup S_{\text{on}}} Q_\sigma(S_{\text{on}}) \log \frac{Q_\sigma(S_{\text{on}})}{Q_{\sigma'}(S_{\text{on}})} \\ &\leq \frac{16n\epsilon^{2\rho}}{Cd} + \frac{16m\epsilon^2}{d} \leq 1/4 + 1/4 = 1/2,\end{aligned}$$

where the first term in the first inequality follows from the upper bound of KL of two distributions of the offline data in the proof of Theorem 4.1, the second term of the first inequality follows from a direct calculation, and the last inequality follows from the choice of ϵ presented at the beginning of the proof.

The worst-case Hamming distance $\sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[\text{dist}(\sigma,\hat{\sigma})\right]$ can be lower-bounded using the standard tools in hypothesis testing:

$$\begin{aligned}
\sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[\text{dist}(\sigma,\hat{\sigma})\right] &\geq \frac{d}{2} \min_{\sigma,\sigma':\text{dist}(\sigma,\sigma')=1} \inf_{\psi} \left[Q_\sigma(\psi\neq\sigma) + Q_{\sigma'}(\psi\neq\sigma')\right] \\
&\geq \frac{d}{2}\left(1-\sqrt{\frac{1}{2}\max_{\sigma,\sigma':\text{dist}(\sigma,\sigma')=1} \text{KL}\left(Q_\sigma \| Q_{\sigma'}\right)}\right) \\
&\geq \frac{d}{4},
\end{aligned}$$

where the first inequality follows Assouad's lemma (Tsybakov, 1997, Lemma 2.12) and the second inequality follows from (Tsybakov, 1997, Theorem 2.12).

Let $\hat{\sigma}_i = \mathbb{1}\{\hat{\pi}(a_2|x_i) \geq \frac{1}{2}\}$. We have $\mathbb{1}\{\sigma_i\} - \hat{\pi}(a_2|x_i)\sigma_i \geq \frac{|\sigma_i - \hat{\sigma}_i|}{4}$, and thus, we have

$$\sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[V^*_\sigma - V^\pi_\sigma\right] \geq \frac{\epsilon}{4d} \sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[\text{dist}(\sigma,\hat{\sigma})\right] \geq \frac{\epsilon}{16} = \frac{1}{16}\min\left\{\left(\frac{Cd}{64n}\right)^{1/(2\rho)}, \frac{1}{8}\sqrt{\frac{d}{m}}\right\}.$$

□

E. Proofs for Section 7

E.1. Proof of Theorem 7.5

Proof of Theorem 7.5. The proof structure follows similarly as that of Theorem 4.1.

Construction of a family of hard MDPs. Each MDP M_σ is characterized by $\sigma \in \{-1,1\}^d$. For any σ, M_σ is a deterministic MDP with the state space is $\mathcal{S} = \{x_1, \bar{x}_1, \tilde{x}_1, \ldots, x_d, \bar{x}_d, \tilde{x}_d\}$, the action space $\mathcal{A} = \{a_1, a_2\}$ (see Figure 1). Each MDP M_σ starts uniformly at one of d states $x_1, \ldots, x_d$. Form each state x_i for any $i \in [d]$, one can follow the "blue" path by taking action a_1 or the "red" path by taking action a_2. This always lead to an absorbing state ($\bar{x}_i$ in the blue path and $\tilde{x}_i$ in the red path). Taking any action from an absorbing state leads to the same state. If one starts from x_i for any $i \in [d]$, the reward for every blue arrow in the graph (resp. red arrow) is $\frac{1}{2}$ (resp. $\frac{1}{2} + \sigma_i\frac{\epsilon}{2}$). Also note that all the MDPs in the family share the same (deterministic) dynamics (they are only different by reward labelling).

We can compute exactly Q-functions of each policy under each MDP M_σ. Note that starting from $h \geq 2$, the value function does not depend on the action being taking. The total reward in any trajectory essentially depends on which initial state one starts with and which action one takes from the initial state. Thus under any MDP M_σ, its Q-value functions under a policy π have the property that they are completely agnostic to π – which comes handy in satisfying the value realizability and Bellman completeness. This property is inspired by the construction by (Foster et al., 2021), though our constructions are different and much simpler.

Specifically, for any π, we have

$$\begin{aligned}
Q^\pi_h(\bar{x}_i, a) &= \frac{H-h+1}{2}, \forall h \geq 1, a \in \mathcal{A}, i \in [d] \\
Q^\pi_h(\tilde{x}_i, a) &= (H+1-h)(\frac{1}{2} + \sigma_i\frac{\epsilon}{2}), \forall h \geq 1, a \in \mathcal{A}, i \in [d] \\
Q^\pi_1(x_i, a_1) &= \frac{H}{2}, \forall i \in [d] \\
Q^\pi_1(x_i, a_2) &= H(\frac{1}{2} + \sigma_i\frac{\epsilon}{2}), \forall i \in [d].
\end{aligned}$$

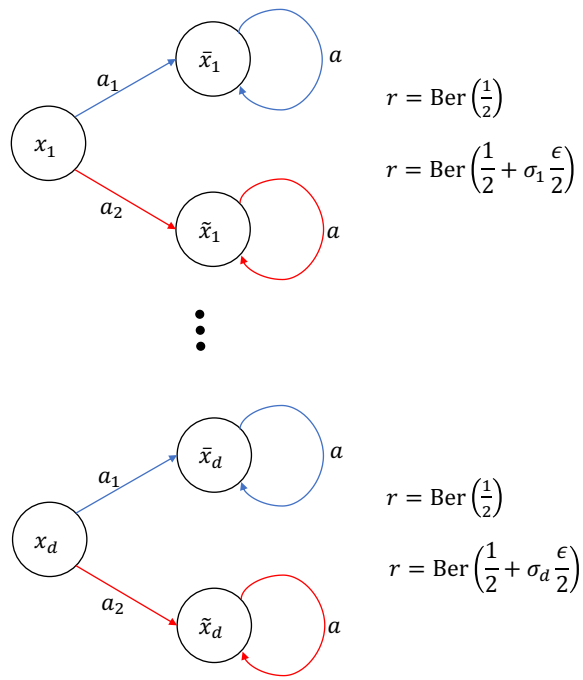


Figure 1. Hard MDPs

We construction the following function class $\mathcal{F} = \{f^\sigma : \sigma \in \{-1,1\}^d\}$ where

$$f_1^\sigma(x_i, a_1) = \frac{H}{2}, \qquad f_1^\sigma(x_i, a_2) = H(\frac{1}{2} + \sigma_i \frac{\epsilon}{2}),$$
$$\forall h \in [2, H], f_h^\sigma(\bar{x}_i, a) = (H+1-h)\frac{1}{2}, \qquad f_h^\sigma(\tilde{x}_i, a) = (H+1-h)(\frac{1}{2} + \sigma_i \frac{\epsilon}{2}).$$

It is easy to see that for any $\sigma \in \{-1,1\}^d$, the pair $(\mathcal{F}, M_\sigma)$ satisfies the value realizability and the Bellman completness. The value realizability follows from that $\mathcal{F}$ is constructed as the bare minimum function class that contains all possible Q-value functions of the MDPs in the class. The Bellman completeness follows from the value realizability and that the Bellman backup under any policy π on a function in $\mathcal{F}$ does not depend on π.

We have

$$V_1^\pi = \frac{H}{d} \sum_{i=1}^{d} \left(\pi_1(a_1|x_i)\frac{1}{2} + \pi_1(a_2|x_i)(\frac{1}{2} + \sigma_i \frac{\epsilon}{2}) \right).$$

Thus, we have

$$\mathbb{E}_\sigma \left[V_\sigma^* - V_\sigma^\pi \right] \geq \frac{H\epsilon}{4d} \mathbb{E}_\sigma \left[\text{dist}(\sigma, \hat{\sigma}) \right], \tag{9}$$

For any $f \in \mathcal{F}$, we have $f = f^\alpha$ for some $\alpha \in \{-1,1\}^d$. For any policy π, we have

$$\forall h \in [H], \mathbb{E}_\pi[(f_h^\sigma - f_h^\alpha)^2] = \epsilon^2 (H-1+1)^2 \sum_{i=1}^{d} \pi_1(a_2|x_i) \frac{\mathbb{1}\{\sigma_i \neq \alpha_i\}}{d}.$$

Choose

$$\mu(a_2|x_i) = \frac{\epsilon^{2\rho-2} H^{2\rho-2}}{C}.$$

This is valid only when $C \geq (\epsilon H)^{2\rho-2}$. Then we have

$$\forall h \in [H], C|\mathbb{E}_\pi[f_h^\sigma - f_h^\alpha]|^2 \leq C\mathbb{E}_\pi[(f_h^\sigma - f_h^\alpha)^2] \leq \mathbb{E}_\mu[(f_h^\sigma - f_h^\alpha)^2]^\rho.$$

The offline data can be equivalently reduced into $S_n = \{(s_1^t, a_1^t, r_1^t)\}_{t\in[n]}$ because the information in the first timestep $h = 1$ fully captures the information in subsequent time steps $h \geq 1$. For any σ and σ' such that $\text{dist}(\sigma, \sigma') = 1$, let $i^* \in [d]$ be the (only) coordinate that σ differs from σ', we have

$$\text{KL}\left(\Pr_\sigma(S_n) \| \Pr_{\sigma'}(S_n)\right) \leq 16\frac{n}{d}\mu(a_2|x_{i^*})\epsilon^2 \leq \frac{16nH^{2\rho-2}\epsilon^{2\rho}}{Cd} = \frac{1}{2},$$

where we choose

$$\epsilon = \left(\frac{Cd}{32nH^{2\rho-2}}\right)^{\frac{1}{2\rho}}.$$

The worst-case Hamming distance $\sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[\text{dist}(\sigma, \hat{\sigma})\right]$ can be lower-bounded using the standard tools in hypothesis testing:

$$\begin{aligned}
\sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[\text{dist}(\sigma, \hat{\sigma})\right] &\geq \frac{d}{2} \min_{\sigma,\sigma':\text{dist}(\sigma,\sigma')=1} \inf_\psi \left[\Pr_\sigma(\psi \neq \sigma) + \Pr_{\sigma'}(\psi \neq \sigma')\right] \\
&\geq \frac{d}{2}\left(1 - \sqrt{\frac{1}{2} \max_{\sigma,\sigma':\text{dist}(\sigma,\sigma')=1} \text{KL}\left(\Pr_\sigma(S_n) \| \Pr_{\sigma'}(S_n)\right)}\right) \\
&\geq \frac{d}{4},
\end{aligned}$$

where the first inequality follows Assouad's lemma (Tsybakov, 1997, Lemma 2.12) and the second inequality follows from (Tsybakov, 1997, Theorem 2.12).

Plugging into Equation (9), we have

$$\begin{aligned}
\sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[V_\sigma^* - V_\sigma^\pi\right] &\geq \frac{H\epsilon}{4d} \sup_{\sigma\in\{-1,1\}^d} \mathbb{E}_\sigma\left[\text{dist}(\sigma, \hat{\sigma})\right] \\
&\geq \frac{H\epsilon}{16} = \frac{1}{16}\left(\frac{H^2Cd}{32n}\right)^{\frac{1}{2\rho}}.
\end{aligned}$$

□

E.2. Proof of Theorem 7.6

We first re-state Theorem 7.6.

Theorem E.1. *Let Assumption 7.1, Assumption 7.2, and Assumption 7.3 hold and assume that $|\mathcal{A}| = K$. There exists a (possibly randomized) learning algorithm $\hat{\pi}$ such that for any MDP M, and any $\delta \in [0, 1]$, with probability at least $1 - \delta$, for any π such that $\rho_\pi \geq 0.5$,*

$$\mathbb{E}\left[\text{SubOpt}_M^\pi(\hat{\pi})|S\right] = C_\pi^{\frac{1}{2\rho_\pi}} H^{1-\frac{1}{2\rho_\pi}} \left(H^2\epsilon + H^3\frac{d(\epsilon, n) + \log(H/\delta)}{n}\right)^{\frac{1}{2\rho_\pi}},$$

where $d(\epsilon, n) := \max_{h\in[H]}\{\log N_1(\mathcal{F}_h, \epsilon, n) \vee \log N_1(\mathcal{F}_h(\cdot, \Pi_h), \epsilon, n)\}$.

We provide a specific algorithm, Algorithm 3, that obtains the bounds in Theorem E.1. Algorithm 3 is essentially the OfDM-Hedge algorithm (Algorithm 1) extended to MDPs. Note that Algorithm 3 also already appears in the prior works of (Xie et al., 2021a; Nguyen-Tang & Arora, 2023).

Algorithm 3 Hedge for Offline Decision-Making in MDP (OfDM-Hedge-MDP)

Input: Offline data S, function class $\mathcal{F}$
Hyperparameters: Confidence parameter β, learning rate η, iteration number T
Initialize $\pi^{(1)} = \{\pi_h^{(1)}\}_{h\in[H]}$, where $\pi_h^{(1)}(\cdot|x) = \text{Uniform}(\mathcal{A}), \forall x \in \mathcal{X}_h$
for $t = 1$ **to** T **do**
  Pessimism: $f^{(t)} = \underset{f\in\mathcal{F}(\beta,\pi^{(t)})}{\arg\min} f_1(s_1, \pi_1^{(t)})$ where

$$\mathcal{F}(\beta, \pi^{(t)}) := \left\{ f \in \mathcal{F} : \sum_{i\in[n]} \sum_{h\in[H]} L_i(f_h, f_{h+1}, \pi^{(t)}) - \inf_{g\in\mathcal{F}} \sum_{i\in[n]} \sum_{h\in[H]} L_i(g_h, f_{h+1}, \pi^{(t)}) \le \beta \right\}$$

  Hedge: $\pi_h^{(t+1)}(a|x) \propto \pi_h^{(t)}(a|x) e^{\eta f_h^{(t)}(x,a)}, \forall (x, a, h)$
end for
Output: A randomized policy $\hat{\pi}$ as a uniform distribution over $\{\pi^{(t)}\}_{t\in[T]}$.

Notations. For convenience, we denote the element-wise functionals indexed by functions and policies:

$$\begin{aligned} L_i(f_h, f_{h+1}, \pi) &:= (f_h(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_{h+1}(x_{h+1}^{(i)}, \pi_{h+1}))^2, \\ Z_i(f_h, f_{h+1}, \pi) &:= L_i(f_h, f_{h+1}, \pi) - L_i(T_h^\pi f_{h+1}, f_{h+1}, \pi), \\ \mathcal{E}_h^\pi(f_h, f_{h+1})(x, a) &:= (T_h^\pi f_h - f_{h+1})(x, a). \end{aligned}$$

A key starting point for the proof of Theorem E.1 of our upper bounds in this section is the error decomposition lemma that relies on a notion of induced MDPs, used originally in (Zanette et al., 2021) and adopted in (Nguyen-Tang & Arora, 2023).

Definition E.2 (Induced MDPs)**.** For any policy π and any sequence of functions $Q = \{Q_h\}_{h\in[H]} \in \{\mathcal{X} \times \mathcal{A} \to \mathbb{R}\}^H$, the (Q, π)-induced MDPs, denoted by $M(Q, \pi)$ is the MDP that is identical to the original MDP M except only that the expected reward of $M(Q, \pi)$ is given by $\{r_h^{\pi,Q}\}_{h\in[H]}$, where

$$r_h^{\pi,Q}(x, a) := r_h(x, a) - (T_h^\pi f_h - f_{h+1})(x, a).$$

By definition of $M(\pi, Q)$, Q is the fixed point of the Bellman equation $Q_h = T_{h,M(\pi,Q)}^\pi Q_{h+1}$.

Lemma E.3. *For any policy π and any sequence of functions $Q = \{Q_h\}_{h\in[H]} \in \{\mathcal{X} \times \mathcal{A} \to \mathbb{R}\}^H$, we have*

$$Q_{M(\pi,Q)}^\pi = Q,$$

where $M(\pi, Q)$ is the induced MDP given in Definition E.2.

A key lemma that we use is the following error decomposition.

Lemma E.4 (Error decomposition)**.** *For any action-value functions $Q \in \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}^H$ and any policies $\pi, \tilde{\pi} \in \Pi$, we have*

$$\text{SubOpt}_\pi^M(\tilde{\pi}) = \sum_{h=1}^H \mathbb{E}_\pi[\mathcal{E}_h^{\tilde{\pi}}(Q_h, Q_{h+1})(x_h, a_h)] + Q_1(x_1, \tilde{\pi}_1) - V_{1,M}^{\tilde{\pi}}(x_1) + \text{SubOpt}_\pi^{M(Q,\tilde{\pi})}(\tilde{\pi}).$$

Proof of Lemma E.4. We have

$$\begin{aligned} &\text{SubOpt}_\pi^M(\tilde{\pi}) = V_1^\pi(x_1) - V_1^{\tilde{\pi}}(x_1) \\ &= \left(V_1^\pi(x_1) - V_{1,M(Q,\tilde{\pi})}^\pi(x_1)\right) + \left(V_{1,M(Q,\tilde{\pi})}^{\tilde{\pi}}(x_1) - V_1^{\tilde{\pi}}(x_1)\right) + \left(V_{1,M(Q,\tilde{\pi})}^\pi(x_1) - V_{1,M(Q,\tilde{\pi})}^{\tilde{\pi}}(x_1)\right) \\ &= \sum_{h=1}^H \mathbb{E}_\pi[\mathcal{E}_h^{\tilde{\pi}}(Q_h, Q_{h+1})(x_h, a_h)] + Q_1(x_1, \tilde{\pi}_1) - V_1^{\tilde{\pi}}(x_1) + \text{SubOpt}_\pi^{M(Q,\tilde{\pi})}(\tilde{\pi}), \end{aligned}$$

where in the last equality, for the first term, we use, by Definition E.2, that

$$V_1^\pi(x_1) - V_{1,M(Q,\tilde{\pi})}^\pi(x_1) = \sum_{h=1}^H \mathbb{E}_\pi\left[r_h(x_h,a_h) - r_h^{\tilde{\pi},Q}(x_h,a_h)\right] = \sum_{h=1}^H \mathbb{E}_\pi[\mathcal{E}_h^{\tilde{\pi}}(Q_h,Q_{h+1})(x_h,a_h)],$$

for the second term, we use, by Lemma E.3, that

$$V_{1,M(Q,\tilde{\pi})}^\pi(x_1) = Q_{1,M(Q,\tilde{\pi})}^\pi(x_1,\tilde{\pi}_1) = Q_1(x_1,\tilde{\pi}).$$

□

Lemma E.5 ((Nguyen-Tang & Arora, 2023)). *Consider an arbitrary sequence of value functions $\{Q^t\}_{t\in[T]}$ such that $\max_{h,t}\|Q_h^t\|_\infty \le b$ and define the following sequence of policies $\{\pi^t\}_{t\in[T+1]}$ where*

$$\pi^1(\cdot|s) = \text{Uniform}(\mathcal{A}), \forall s,$$
$$\pi_h^{t+1}(a|s) \propto \pi_h^t(a|s)\exp\left(\eta Q_h^t(s,a)\right), \forall (s,a,h,t).$$

Suppose $\eta = \sqrt{\frac{\ln K}{4(e-2)b^2T}}$ and $T \ge \frac{\ln K}{(e-2)}$. For an arbitrary policy $\pi \in \Pi$, we have

$$\sum_{t=1}^T \left(V_{1,M(\pi^t,Q^t)}^\pi(x_1) - V_{1,M(\pi^t,Q^t)}^{\pi^t}(x_1)\right) \le 4Hb\sqrt{T\log K}.$$

Proof of Theorem E.1. By Lemma E.4, for every π, t, we have

$$\text{SubOpt}_\pi^M(\pi^{(t)}) = \underbrace{\sum_{h=1}^H \mathbb{E}_\pi[(T_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)})(s_h,a_h)]}_{A_t} + \underbrace{f_1^{(t)}(s_1,\pi_1^{(t)}) - V_{1,M}^{\pi^{(t)}}(s_1)}_{B_t} + \underbrace{\text{SubOpt}_\pi^{M(f^{(t)},\pi^{(t)})}(\pi^{(t)})}_{C_t}.$$

Thus, we have

$$\mathbb{E}\left[\text{SubOpt}_M^\pi(\hat{\pi})|S\right] = \frac{1}{T}\sum_{t=1}^T A_t + B_t + C_t.$$

Note that by Lemma E.5, we have

$$\sum_{t=1}^T C_t \le 4H^2\sqrt{T\ln K}.$$

Thus, we now only need to bound $\sum_{t=1}^T A_t$ and $\sum_{t=1}^T B_t$. Note that for every π and $\tilde{\pi}$, we have

$$\begin{aligned}\sum_{h=1}^H \mathbb{E}_\pi\left[(T_h^{\tilde{\pi}} f_{h+1} - f_h)(s_h,a_h)\right] &\le \sum_{h=1}^H C_\pi^{1/(2\rho_\pi)}\left(\mathbb{E}_\mu\left[(T_h^{\tilde{\pi}} f_{h+1} - f_h)(s_h,a_h)^2\right]\right)^{1/(2\rho_\pi)} \\ &\le C_\pi^{1/(2\rho_\pi)} H^{1-1/(2\rho_\pi)}\left(\mathbb{E}_\mu\left[\sum_{h=1}^H (T_h^{\tilde{\pi}} f_{h+1} - f_h)(s_h,a_h)^2\right]\right)^{1/(2\rho_\pi)}, \end{aligned} \tag{10}$$

where the first inequality follows from the Bellman completeness assumption and the transfer exponent definition, the second inequality follows from Jensen's inequality for a concave function $x \mapsto x^{1/(2\rho_\pi)}$ as long as $\rho_\pi \ge 1/2$. Thus, to bound $\sum_{t=1}^T A_t$, it suffices to bound the in-distribution squared Bellman error $\mathbb{E}_\mu\left[\sum_{h=1}^H (T_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)})(s_h,a_h)^2\right]$. This relies on the uniform Bernstein's inequality for Bellman-like loss functions we have developed in Appendix A.

Lemma E.6 (Uniform Bernstein's inequality for Bellman-like loss functions)**.** *Fix any $\epsilon > 0$. With probability at least $1-\delta$, for any $f \in \mathcal{F}, \pi \in \Pi$,*

$$\mathbb{E}[(f_h - T_h^\pi f_{h+1})^2] \leq \frac{2}{n}\sum_{t=1}^n Z_t(f_h, f_{h+1}, \pi) + \inf_{\epsilon>0}\left\{108H\epsilon + H^2\frac{36\log N_1(\mathcal{F}_h,\epsilon,n) + 83\log N_1(\mathcal{F}_{h+1}(\cdot,\Pi_{h+1}),\epsilon,n) + 108\log(12/\delta)}{n}\right\}.$$

In addition, with probability at least $1-\delta$, for any $f \in \mathcal{F}, \pi \in \Pi$,

$$-\frac{1}{n}\sum_{t=1}^n Z_t(f_h, f_{h+1}, \pi) \leq \inf_{\epsilon>0}\left\{32H\epsilon + H^2\frac{4\log N_1(\mathcal{F}_h,\epsilon,n) + 28\log N_1(\mathcal{F}_{h+1}(\cdot,\Pi_{h+1}),\epsilon,n) + 24\log(6/\delta)}{n}\right\}.$$

Proof of Lemma E.6. This is a direct application of Proposition A.3.

□

Now let's fix $\epsilon \geq 0$ and $\delta \in [0,1]$. We use c to denote an absolute constant that can change its value at every of its appearance, as we are not interested in absolute constant factors and would like to simplify the presentation. Set β in Algorithm 3 by

$$\beta = c\left(H^2\epsilon + H^3\frac{d(\epsilon,n) + \log(H/\delta)}{n}\right).$$

By the second part of Lemma E.6, Assumption 7.1, and Assumption 7.2, we have

$$\Pr\left(\forall \pi, Q^\pi \in \mathcal{F}(\beta,\pi)\right) \geq 1-\delta.$$

Thus, we have

$$\Pr\left(B_t \leq 0, \forall t\right) \geq 1-\delta.$$

For bounding $\sum_{t=1}^T A_t$, note that we have

$$\sum_{i=1}^n Z_i(f_h^{(t)}, f_{h+1}^{(t)}, \pi^{(t)}) \leq \sum_{i\in[n]}\sum_{h\in[H]} L_i(f_h^{(t)}, f_{h+1}^{(t)}, \pi^{(t)}) - \inf_{g\in\mathcal{F}}\sum_{i\in[n]}\sum_{h\in[H]} L_i(g_h, f_{h+1}^{(t)}, \pi^{(t)}) \leq \beta,$$

where the first inequality follows from Assumption 7.2 and the second inequality follows from the design of Algorithm 3. Thus, by the first part of Lemma E.6, with probability at least $1-\delta$, we have

$$\forall t, \mathbb{E}\left[\sum_{h=1}^H (T_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)})(s_h,a_h)^2\right] \leq c\left(H^2\epsilon + H^3\frac{d(\epsilon,n)+\log(H/\delta)}{n}\right).$$

Plugging the above inequality into the RHS of Equation (10) leads to an upper bound on $\sum_{t=1}^T A_t$.

□

F. Support Lemmas

Lemma F.1 (Freedman's inequality)**.** *Let $X_1, \ldots, X_T$ be* any *sequence of real-valued random variables. Denote $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|X_1,\ldots,X_{t-1}]$. Assume that $X_t \leq R$ for some $R > 0$ and $\mathbb{E}_t[X_t] = 0$ for all t. Define the random variables*

$$S := \sum_{t=1}^T X_t, \quad V := \sum_{i=1}^T \mathbb{E}_t[X_t^2].$$

Then for any $\delta > 0$, with probability at least $1-\delta$, for any $\lambda \in [0, 1/R]$,

$$S \leq (e-2)\lambda V + \frac{\ln(1/\delta)}{\lambda}.$$

The following lemma exploits the non-negativity of the function class to obtain a fast estimation error rate when relating the population quantity to the empirical one.

Lemma F.2. *Consider any function class $\mathcal{H} \subseteq \{\mathcal{Z} \to [0, b]\}$ for some $b > 0$ and let $S_n = \{z_1, \ldots, z_n\}$ be an i.i.d. sample from a distribution $P \in \Delta(\mathcal{Z})$. For any $\delta \in (0, 1)$, with probability at least $1-\delta$ over the randomness of S_n, we have*

$$\forall h \in \mathcal{H} : Ph \leq 4\hat{P}_n h + \inf_{\epsilon > 0}\left[8\epsilon + \frac{12b\ln(3N_1(\mathcal{H}, \epsilon, S_n)/\delta)}{n}\right].$$

Remark F.3. Lemma F.2 is a generalization of (Zhang, 2023, Theorem 4.12) from a function range $[0, 1]$ to an arbitrary function range $[0, b]$, i.e. setting $b = 1$ in the above lemma reduces to (Zhang, 2023, Theorem 4.12).

Proof of Lemma F.2. We start from (Zhang, 2023, Theorem 4.12, with $\gamma = 0.5$ in their theorem) which corresponds to the case $b = 1$ of the above lemma. Let $\mathcal{H}/b := \{h/b : h \in \mathcal{H}\}$. Since $\|h'\|_\infty \leq 1, \forall h' \in \mathcal{H}/b$, we can apply (Zhang, 2023, Theorem 4.12) to $\mathcal{H}/b$: With probability at least $1-\delta$: $\forall h \in \mathcal{H}$, we have

$$\begin{aligned} P\frac{h}{b} &\leq 4\hat{P}_n\frac{h}{b} + \inf_{\epsilon>0}\left[8\epsilon + \frac{12\ln(3N_1(\mathcal{H}/b, \epsilon, S_n)/\delta)}{n}\right] \\ &\leq 4\hat{P}_n\frac{h}{b} + \inf_{\epsilon>0}\left[8\epsilon + \frac{12\ln(3N_1(\mathcal{H}, \epsilon b, S_n)/\delta)}{n}\right]. \end{aligned}$$

The above inequality implies that

$$\begin{aligned} Ph &\leq 4\hat{P}_n h + \inf_{\epsilon>0}\left[8b\epsilon + \frac{12b\ln(3N_1(\mathcal{H}, \epsilon b, S_n)/\delta)}{n}\right] \\ &\leq 4\hat{P}_n h + \inf_{\epsilon>0}\left[8\epsilon + \frac{12b\ln(3N_1(\mathcal{H}, \epsilon, S_n)/\delta)}{n}\right], \end{aligned}$$

where the last inequality follows from replacing ϵ by ϵ/b in the first inequality. □

Remark F.4. It is possible to obtain a tighter bound specified by the Rademacher complexity of the function class $\mathcal{H}$ in the above lemma, if we are willing to make an additional assumption that each function in $\mathcal{H}$ is smooth (and non-negative, which is already satisfied in the above lemma). The fast rates are achievable via the optimistic rate framework of (Srebro et al., 2010). The smooth and non-negative condition is satisfied anyway in our case as we use squared loss. However, this result comes at the cost of a large absolute constant in the upper bound. Also, this stronger upper bound ultimately does not benefit our case as our bounds still depend on log-covering numbers, instead of entirely depending on Rademacher complexity.

Lemma F.5. *Let $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, let $\Pi = \{\mathcal{X} \to \Delta(\mathcal{A})\}$ be the set of all policies. Suppose that $|\mathcal{A}| = K$. For any $p \geq 1, n \in \mathbb{N}$, we have*

$$\max\{N_p(\mathcal{F}, \epsilon, n), N_p(\mathcal{F}(\cdot, \Pi), \epsilon, n)\} \leq \prod_{a\in\mathcal{A}} N_p(\mathcal{F}(\cdot, a), \frac{\epsilon}{K^{1/p}}, n).$$

Proof of Lemma F.5. We will first prove that:

$$N_p(\mathcal{F}(\cdot, \Pi), \epsilon, n) \leq \prod_{a\in\mathcal{A}} N_p(\mathcal{F}(\cdot, a), \frac{\epsilon}{K^{1/p}}, n).$$

The other inequality can be proved similarly. Fix any $\epsilon > 0, p \geq 1, n \in \mathbb{N}$. Let $N_a = N_p(\mathcal{F}(\cdot, a), \epsilon', n), \forall a \in \mathcal{A}$. For any $g \in \mathcal{F}(\cdot, \Pi)$, $g = f(\cdot, \pi)$ for some $f \in \mathcal{F}, \pi \in \Pi$. Let f' such that $\left(\frac{1}{n}\sum_{i=1}^n |f(x_i, a) - f'(x_i, a)|^p\right)^{1/p} \leq \frac{\epsilon}{K^{1/p}}$ for any a. Define $g' = f'(\cdot, \pi)$. We have

$$
\begin{aligned}
\|g - g'\|_n^p &= \frac{1}{n}\sum_{i=1}^n |f(x_i, \pi) - f'(x_i, \pi)|^p \\
&\leq \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}_{a\sim\pi(\cdot|x_i)}\left[|f(x_i, a) - f'(x_i, a)|\right]\right)^p \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{a\sim\pi(\cdot|x_i)}\left[|f(x_i, a) - f'(x_i, a)|^p\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n \max_{a\in\mathcal{A}} |f(x_i, a) - f'(x_i, a)|^p \\
&\leq \frac{1}{n}\sum_{i=1}^n \sum_{a\in\mathcal{A}} |f(x_i, a) - f'(x_i, a)|^p \\
&\leq \frac{1}{n}\sum_{a\in\mathcal{A}} \sum_{i=1}^n |f(x_i, a) - f'(x_i, a)|^p \leq \epsilon^p.
\end{aligned}
$$

□