Offline Reinforcement Learning: Fundamental Barriers for Value Function Approximation

Dylan J. Foster Akshay Krishnamurthy David Simchi-Levi Yunzong Xu*
Microsoft Research Massachusetts Institute of Technology
{dylanfoster,akshaykr}@microsoft.com {dslevi,yxu}@mit.edu

Abstract

We consider the offline reinforcement learning problem, where the aim is to learn a decision making policy from logged data. Offline RL—particularly when coupled with (value) function approximation to allow for generalization in large or continuous state spaces—is becoming increasingly relevant in practice, because it avoids costly and time-consuming online data collection and is well suited to safety-critical domains. Existing sample complexity guarantees for offline value function approximation methods typically require both (1) distributional assumptions (i.e., good coverage) and (2) representational assumptions (i.e., ability to represent some or all Q-value functions) stronger than what is required for supervised learning. However, the necessity of these conditions and the fundamental limits of offline RL are not well understood in spite of decades of research. This led Chen and Jiang (2019) to conjecture that *concentrability* (the most standard notion of coverage) and *realizability* (the weakest representation condition) alone are not sufficient for sample-efficient offline RL. We resolve this conjecture in the positive by proving that in general, even if both concentrability and realizability are satisfied, any algorithm requires sample complexity polynomial in the size of the state space to learn a non-trivial policy.

Our results show that sample-efficient offline reinforcement learning requires either restrictive coverage conditions or representation conditions that go beyond supervised learning, and highlight a phenomenon called *over-coverage* which serves as a fundamental barrier for offline value function approximation methods. A consequence of our results for reinforcement learning with linear function approximation is that the separation between online and offline RL can be *arbitrarily large*, even in constant dimension.

1 Introduction

In offline reinforcement learning, we aim to evaluate or optimize decision making policies using logged transitions and rewards from historical experiments or expert demonstrations. Offline RL has great promise for decision making applications where actively acquiring data is expensive or cumbersome (e.g., robotics (Pinto and Gupta, 2016; Levine et al., 2018; Kalashnikov et al., 2018)), or where safety is critical (e.g., autonomous driving (Sallab et al., 2017; Kendall et al., 2019) and healthcare (Gottesman et al., 2018, 2019; Wang et al., 2018; Yu et al., 2019; Nie et al., 2021)). In particular, there is substantial interest in combining offline reinforcement learning with function approximation (e.g., deep neural networks) in order to encode inductive biases and enable generalization across large, potentially continuous state spaces, with recent progress on both model-free and model-based approaches (Ross and Bagnell, 2012; Laroche et al., 2019; Fujimoto et al., 2019; Kumar et al., 2019; Agarwal et al., 2020). However, existing algorithms are extremely data-intensive, and offline RL methods—to date—have seen limited deployment in the aforementioned applications. To enable practical deployment going forward, it is paramount that we develop a strong understanding of the statistical foundations for reliable, sample-efficient offline reinforcement learning with function approximation, as well as an understanding of when and why existing methods succeed and how to effectively collect data.

Compared to the basic supervised learning problem, offline reinforcement learning with function approximation poses substantial algorithmic challenges due to two issues: *distribution shift* and *credit assignment*. Within

*Part of this work was completed while Y. Xu was an intern at Microsoft Research.

the literature on *value* function approximation (or, approximate dynamic programming), all existing methods require both (1) distributional conditions, which assert that the logged data has good coverage (addressing distribution shift), and (2) representational conditions, which assert that the function approximator is flexible enough to represent value functions induced by certain policies (addressing credit assignment). Notably, sample complexity analyses for standard offline RL methods (e.g., fitted Q-iteration) require representation conditions considerably more restrictive than what is required for classical supervised learning (Munos, 2003, 2007; Munos and Szepesvári, 2008; Antos et al., 2008), and these methods are known to diverge when these conditions do not hold (Gordon, 1995; Tsitsiklis and Van Roy, 1996, 1997; Wang et al., 2021a). Despite substantial research effort, it is not known whether these conditions constitute fundamental limits, or whether the algorithms can be improved. Resolving this issue would serve as a stepping stone toward developing a theory for offline reinforcement learning that parallels our understanding of supervised (statistical) learning.

The lack of understanding of fundamental limits in offline reinforcement learning was highlighted by Chen and Jiang (2019), who observed that all existing finite-sample analyses for offline RL algorithms based on *concentrability* (Munos, 2003)—the most ubiquitous notion of data coverage—require representation conditions significantly stronger than *realizability*, a standard condition from supervised learning which asserts that the function approximator can represent optimal value functions. Chen and Jiang (2019) conjectured that realizability and concentrability alone do not suffice for sample-efficient offline RL, and noted that proving such a result seemed to be out of reach for existing lower bound techniques. Subsequent progress led to positive results for sample-efficient offline RL under coverage conditions stronger than concentrability (Xie and Jiang, 2021) and impossibility results under weaker coverage conditions (Wang et al., 2020; Zanette, 2021), but the original conjecture remained open.

Contributions. We resolve the conjecture of Chen and Jiang (2019) in the positive. We provide an information-theoretic lower bound which shows that in general, even if both concentrability and realizability are satisfied, any algorithm requires sample complexity polynomial in the size of the state space to learn a near optimal policy. This establishes that sample-efficient offline reinforcement learning in large state spaces is not possible unless more stringent conditions—either distributional or representational—hold.

Our lower bound construction is qualitatively different from previous approaches and holds even when the effective horizon and number of actions are constant and the value function class has constant size. The lower bound highlights the role of a phenomenon we call *over-coverage* (first documented by Xie and Jiang (2021)), wherein the data collection distribution is supported over spurious states not reachable by any policy which—despite being irrelevant for learning in the online setting—create significant uncertainty. Our work shows that the over-coverage phenomenon is a fundamental, information-theoretic barrier for design of offline reinforcement learning algorithms.

1.1 Offline Reinforcement Learning Setting

Markov decision processes. We consider the infinite-horizon discounted reinforcement learning setting. Formally, a Markov decision process $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ consists of a (potentially large/continuous) state space $\mathcal{S}$, action space $\mathcal{A}$, probability transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$, discount factor $\gamma \in [0, 1)$, and initial state distribution d_0. Each (randomized) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ induces a distribution over trajectories $(s_0, a_0, r_0), (s_1, a_1, r_1), \ldots$ via the following process. For $h = 0, 1, \ldots : a_h \sim \pi(s_h)$, $r_h = R(s_h, a_h)$, and $s_{h+1} \sim P(s_h, a_h)$, with $s_0 \sim d_0$. We let $\mathbb{E}^{M,\pi}[\cdot]$ and $\mathbb{P}^{M,\pi}(\cdot)$ denote expectation and probability under this process, respectively.

The expected return for policy π is defined as $J_M(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$, and the value function and Q-function for π are given by

$$V_M^\pi(s) := \mathbb{E}^{M,\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_h \mid s_0 = s\right], \quad \text{and} \quad Q_M^\pi(s,a) := \mathbb{E}^{M,\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_h \mid s_0 = s, a_0 = a\right].$$

It is well-known that there exists a deterministic policy $\pi_M^\star : \mathcal{S} \to \mathcal{A}$ that maximizes $V_M^\pi(s)$ for all $s \in \mathcal{S}$ simultaneously, and thus also maximizes $J_M(\pi)$; Letting $V_M^\star := V_M^{\pi_M^\star}$ and $Q_M^\star := Q_M^{\pi_M^\star}$, we have $\pi_M^*(s) = \arg\max_{a \in \mathcal{A}} Q_M^\star(s, a)$ for all $s \in \mathcal{S}$. Finally, we define the occupancy measure for policy π via $d_M^\pi(s, a) := (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}^{M,\pi}(s_h = s, a_h = a)$. We drop the dependence on the model M when it is clear from context.

Offline policy learning. In the offline policy learning (or, optimization) problem, we do not have direct access to the underlying MDP and instead receive a dataset D_n of tuples (s, a, r, s') with $r = R(s, a)$, $s' \sim P(s, a)$, and $(s, a) \sim \mu$ i.i.d., where $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ is the *data collection distribution*. The goal of the learner is to use the dataset D_n to learn an ε-optimal policy $\widehat{\pi}$, that is:

$$J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \leq \varepsilon,$$

where the expectation $\mathbb{E}[\cdot]$ is over the draw of D_n and any randomness used by the algorithm.

In order to provide sample-efficient learning guarantees that do not depend on the size of the state space, value function approximation methods take advantage of the following conditions.

- **Realizability.** This condition asserts that we have access to a class of candidate value functions $\mathcal{F} \subseteq (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$ (e.g., linear models or neural networks) such that $Q^\star \in \mathcal{F}$. Realizability (that is, a well-specified model) is the most common representation condition in supervised learning and statistical estimation (e.g., Bousquet et al. (2003); Wainwright (2019) and is also widely used in contextual bandits (Agarwal et al., 2012; Foster et al., 2018).
- **Concentrability.** Call a distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$ *admissible* for the MDP M if there exists a (potentially stochastic and non-stationary) policy π and index h such that $\nu(s, a) = \mathbb{P}^\pi[s_h = s, a_h = a]$. This condition asserts that there exists a constant $C_{\mathrm{conc}} < \infty$ such that for all admissible ν,

$$\left\| \frac{\nu}{\mu} \right\|_\infty := \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left\{ \frac{\nu(s, a)}{\mu(s, a)} \right\} \leq C_{\mathrm{conc}}. \tag{1}$$

  Concentrability is a simple but fairly strong notion of coverage which demands that the data distribution uniformly covers all states.

Under these conditions, an offline RL algorithm is said to be sample-efficient if it learns an ε-optimal policy with $\mathrm{poly}(\varepsilon^{-1}, (1-\gamma)^{-1}, C_{\mathrm{conc}}, \log|\mathcal{F}|)$ samples. Notably, such a guarantee depends only on the complexity $\log|\mathcal{F}|$ for the value function class, not on the size of the state space.[1]

Are realizability and concentrability sufficient? While realizability and concentrability are appealing in their simplicity, these assumptions alone are not known to suffice for sample-efficient offline RL. The most well-known line of research (Munos, 2003, 2007; Munos and Szepesvári, 2008; Antos et al., 2008; Chen and Jiang, 2019) analyzes offline RL methods such as fitted Q-iteration under the stronger representation condition that $\mathcal{F}$ is closed under Bellman updates ("completeness"),[2] and obtains $\mathrm{poly}(\varepsilon^{-1}, (1-\gamma)^{-1}, C_{\mathrm{conc}}, \log|\mathcal{F}|)$ sample complexity. Completeness is a widely used assumption, but is substantially more restrictive than realizability and can be violated by adding a single function to $\mathcal{F}$. Subsequent years have seen extensive research into algorithmic improvements and alternative representation and coverage conditions, but the question of whether realizability and concentrability alone are sufficient remained open.

1.2 Main Result

Our main result is an information-theoretic lower bound which shows that realizability and concentrability are not sufficient for sample-efficient offline RL.

Theorem 1 (Main theorem). *For any $S \geq 9$ and $\gamma \in (1/2, 1)$, there exists a family of MDPs $\mathcal{M}$ with $|\mathcal{S}| = S$ and $|\mathcal{A}| = 2$, a value function class $\mathcal{F}$ with $|\mathcal{F}| = 2$, and a data distribution μ such that:*

1. *We have $Q^\star \in \mathcal{F}$ (realizability) and $C_{\mathrm{conc}} \leq 16$ (concentrability) for all models in $\mathcal{M}$.*
2. *Any algorithm using less than $c \cdot S^{1/3}$ samples must have $J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \geq c'/(1-\gamma)$ for some instance in $\mathcal{M}$, where c and c' are absolute numerical constants.*

[1] For infinite function classes ($|\mathcal{F}| = \infty$), one can replace $\log|\mathcal{F}|$ with other standard measures of statistical capacity, such as Rademacher complexity or metric entropy. For example, when $\mathcal{F}$ is a class of d-dimensional linear functions, $\log|\mathcal{F}|$ can be replaced by the dimension d, which is an upper bound on the metric entropy.

[2] Precisely, $\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$, where $\mathcal{T}$ is the *Bellman operator* defined by $[\mathcal{T}f](s, a) = R(s, a) + \mathbb{E}_{s' \sim P(s,a)}[\max_{a'} f(s', a')]$.

Theorem 1 shows that even though realizability and concentrability are satisfied, any algorithm requires at least $\Omega(S^{1/3})$ samples to learn a near-optimal policy. Since S can be arbitrarily large, this shows that sample-efficient offline RL in large state spaces is impossible without stronger representation or coverage conditions.

We next state a variant of Theorem 1 which goes further and shows that even if one replaces realizability ($Q^\star \in \mathcal{F}$) by a substantially stronger representation condition—*all-policy realizability*—which requires that $Q^\pi \in \mathcal{F}$ for *every* policy π rather than just for $\pi^\star$, sample-efficient offline RL is still impossible. This result is established using the same family of MDPs and data distribution as in Theorem 1. The only tradeoff is that Theorem 1′ requires a slightly larger value function class $\mathcal{F}$.

Theorem 1′ (Variant of Theorem 1)**.** *Let $S \geq 9$ and $\gamma \in (1/2, 1)$ be given. For the family of MDPs $\mathcal{M}$ and the data distribution μ described in Theorem 1, there exists a three-dimensional linear function class*

$$\mathcal{F} = \left\{(s,a) \mapsto \langle \phi(s,a), \theta \rangle : \theta \in \mathbb{R}^3, \|\theta\|_\infty \leq (1-\gamma)^{-1}\right\},$$

where ϕ is a known feature map with $\|\phi(s,a)\|_\infty \leq (1-\gamma)^{-1}$, such that:

1. *We have $Q^\pi \in \mathcal{F}$ for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (all-policy realizability) and $C_{\mathrm{conc}} \leq 16$ (concentrability) for all models in $\mathcal{M}$.*
2. *Any algorithm using less than $c \cdot S^{1/3}$ samples must have $J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \geq c'/(1-\gamma)$ for some instance in $\mathcal{M}$, where c and c' are absolute numerical constants.*

We now discuss some consequences and features of our results.

- Our lower bound critically relies on the notion of *over-coverage*, where the data distribution contains states not visited by any admissible policy.[3] The issue of over-coverage was first noted by Xie and Jiang (2021), who observed that it can lead to pathological behavior in certain offline RL algorithms. Our result shows—somewhat surprisingly—that this phenomenon is a fundamental barrier which applies to *any* value approximation method. In particular, we show that over-coverage causes spurious correlations across reachable and unreachable states which leads to significant uncertainty in the dynamics when the number of states is large.

  It remains to be seen whether the over-coverage phenomenon represents a serious barrier in practice, or whether it is simply an indication that more care needs to be taken in mathematically formulating the offline RL problem. On one hand, given a simulator, it is certainly possible to sample state-action pairs that are not reachable by any policy. Furthermore, for the family of instances in our construction, any data distribution with non-trivial concentrability must satisfy over-coverage unless the learner has prior knowledge of the specific instance under consideration. On the other hand, in practice one often has additional knowledge about the underlying MDP in the form of policy-induced data, and it is not clear whether our lower bounds can be made to hold in this setting.

- On the technical side, compared to recent lower bounds which establish hardness of offline RL under coverage conditions weaker than concentrability (Wang et al., 2020; Zanette, 2021), our lower bounds have a less geometric, more information-theoretic flavor, and share more in common with lower bounds for sparsity and support testing in statistical estimation (Paninski, 2008; Verzelen and Villers, 2010; Verzelen and Gassiat, 2018; Canonne, 2020). Whereas previous lower bounds show that weak coverage conditions can lead to *under-coverage* and consequently failure of concentrability, our lower bound shows that over-coverage is an issue even when concentrability is satisfied.

  Another interesting feature is that while previous lower bounds (Wang et al., 2020; Zanette, 2021) are based on deterministic MDPs, our construction critically uses stochastic dynamics. This departure is necessary, as the Bellman error minimization algorithm in Chen and Jiang (2019) succeeds under concentrability and realizability if the dynamics are deterministic.[4] Compared to Wang et al. (2020),

[3] Note that while the states may not be reachable for a given MDP in the family $\mathcal{M}$, all states are reachable for *some* MDP in the family.

[4] Deterministic dynamics allow one to avoid the well-known *double sampling* problem and in particular cause the conditional variance in Eq. (3) of Chen and Jiang (2019) to vanish.

who work with a relatively small state space but large horizon and feature dimension, we keep the horizon constant and instead grow the state space, leading to polynomial dependence on S in the lower bound.

- The lower bound has constant suboptimality gap for $Q^\star$, which rules out gap-dependent regret bounds as a path toward sample-efficient offline reinforcement learning.

While our lower bound focuses on policy optimization and infinite-horizon RL for concreteness, it readily extends to the finite-horizon setting (in fact, with $H = 3$), and provides the first impossibility result for offline RL with constant horizon that we are aware of. The lower bound also extends to policy evaluation; see Section 2.4 for discussion.

1.3 Related Work

We close this section with a detailed discussion of some of the most relevant related work.

Lower bounds. While algorithm-specific counterexamples for offline reinforcement learning algorithms have a long history (Gordon, 1995; Tsitsiklis and Van Roy, 1996, 1997; Wang et al., 2021a), information-theoretic lower bounds are a more recent subject of investigation. Wang et al. (2020) (see also Amortila et al. (2020)) consider the setting where $\mathcal{F}$ is linear (i.e., $Q^\star(s,a) = \langle \phi(s,a), \theta \rangle$, where $\phi(s,a) \in \mathbb{R}^d$ is a known feature map). They consider a weaker coverage condition tailored to the linear setting, which asserts that $\lambda_{\min}\left(\mathbb{E}_{(s,a)\sim\mu}\left[\phi(s,a)\phi(s,a)^\top\right]\right) \geq \frac{1}{d}$, and they show that this condition and realizability alone are not strong enough for sample-efficient offline RL. The feature coverage condition is strictly weaker than concentrability, so this does not suffice to resolve the conjecture of Chen and Jiang (2019), but their lower bound complements our own by highlighting the dangers of *under-coverage* and a related error amplification phenomenon.

In more detail, Wang et al. (2020) construct a family of MDPs and a data distribution that satisfy realizability and the feature coverage condition above, but not concentrability, and use this family to prove an exponential sample complexity lower bound. However, for the family of MDPs under consideration, one can obtain polynomial sample complexity for any data distribution satisfying concentrability, simply because all MDPs in the family have deterministic dynamics. Hence, this construction cannot be used to establish impossibility of sample-efficient learning under concentrability and realizability. Instead, the conceptual takeaway is that the feature coverage condition can lead to under-coverage, and may not be the right assumption for offline RL. This point is further highlighted by Amortila et al. (2020) who show that in the infinite-horizon setting, the feature coverage condition can lead to non-identifiability in MDPs with only two states, meaning one cannot learn an optimal policy even with infinitely many samples. Concentrability places stronger restrictions on the data distribution and underlying dynamics, and always implies identifiability when the state and action space are finite.

Further work in this direction includes (1) Zanette (2021), who provides a slightly more general lower bound for linear realizability, and (2) lower bounds for online reinforcement learning with linear realizability (Du et al., 2020; Weisz et al., 2021; Wang et al., 2021b). It is worth noting that Zanette (2021) provides a lower bound with a policy-induced data distribution, where over-coverage cannot occur; however concentrability is not satisfied by his construction.

Upper bounds. Classical analyses for offline reinforcement learning algorithms such as FQI (Munos, 2003, 2007; Munos and Szepesvári, 2008; Antos et al., 2008) provide sample complexity upper bounds in terms of concentrability under the strong representation condition of Bellman completeness. The path-breaking recent work of Xie and Jiang (2021) provides an algorithm which requires only realizability, but uses a stronger coverage condition ("pushforward concentrability") which requires that $P(s' \mid s,a)/\mu(s') \leq C$ for all (s,a,s'). Our results imply that this condition cannot be substantially relaxed.

A complementary line of work (Uehara et al., 2020; Xie and Jiang, 2020; Jiang and Huang, 2020; Uehara et al., 2021) provides upper bounds that require only concentrability and realizability, but assume access to an additional *weight function class* that is flexible enough to represent various occupancy measures for the underlying MDP. These results scale with the complexity of the weight function class. In general, the

complexity of this class may be prohibitively large without prior knowledge; this is witnessed by our lower bound construction.

1.4 Preliminaries

For any $x \in \mathbb{R}$, let $(x)_+ := \max\{x, 0\}$. For an integer $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, \dots, n\}$. For a finite set $\mathcal{X}$, $\mathrm{unif}(\mathcal{X})$ denotes the uniform distribution over $\mathcal{X}$, and $\Delta(\mathcal{X})$ denotes the set of all probability distributions over $\mathcal{X}$. For probability distributions $\mathbb{P}$ and $\mathbb{Q}$ over a measurable space $(\Omega, \mathscr{F})$ with a common dominating measure, we define the total variation distance as $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int |d\mathbb{P} - d\mathbb{Q}|$ and define the χ^2-divergence as $D_{\chi^2}(\mathbb{P} \,\|\, \mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[\left(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right)^2\right] = \int \frac{d\mathbb{P}^2}{d\mathbb{Q}} - 1$ when $\mathbb{P} \ll \mathbb{Q}$ and $+\infty$ otherwise.

2 Fundamental Barriers for Offline Reinforcement Learning

In this section we present our lower bound construction, prove Theorem 1, and discuss consequences and extensions, including Theorem 1′.

2.1 Construction: MDP Family, Value Functions, and Data Distribution

We first provide our lower bound construction, which entails specifying the MDP family $\mathcal{M}$, the value function class $\mathcal{F}$, and the data distribution μ; recall that $S = |\mathcal{S}|$ and $\gamma \in (0, 1)$ are parameters for the construction.

All MDPs in $\mathcal{M}$ belong to a parameterized MDP family with shared transition and reward structure. In what follows, we first describe the structure of the parameterized family (Section 2.1.1) and give intuition behind why this structure leads to statistical hardness (Section 2.1.2). We then provide a specific collection of parameters that gives rise to the hard family $\mathcal{M}$ (Section 2.1.3) and finally complete the construction by specifying the value function class $\mathcal{F}$ and data distribution μ (Section 2.1.4).

2.1.1 MDP Parameterization

Let the discount factor $\gamma \in (0, 1)$ be fixed, and let $S \in \mathbb{N}$ be given. Assume without loss of generality that $S > 5$ and that $(S-5)/4$ is an integer. We consider the parameterized MDP family illustrated in Figure 1a. Each MDP takes the form $M_{I,\alpha,\beta,\mathbf{r}} = \{\mathcal{S}, \mathcal{A}, P_{I,\alpha,\beta}, R_{\mathbf{r}}, \gamma, d_0\}$, and is parametrized by a subset of states I, two probability parameters $\alpha, \beta \in (0, 1)$, and a reward vector $\mathbf{r} = (r_W, r_X, r_Y, r_Z) \in [0, 1]^4$. All MDPs in the family $\{M_{I,\alpha,\beta,\mathbf{r}}\}$ share the same state space $\mathcal{S}$, action space $\mathcal{A}$, discount factor γ, and initial state distribution d_0, and differ only in terms of the transition function $P_{I,\alpha,\beta}$ and the reward function $R_{\mathbf{r}}$.

State space. We consider a layered state space $\mathcal{S} := \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2$ with $|\mathcal{S}| = S$, where $\mathcal{S}_0 := \{\mathfrak{s}\}$ contains a single *initial state* $\mathfrak{s}$ which occurs at $h = 0$, $\mathcal{S}_1 = [S_1]$ is a collection of *first-layer states* which occur at $h = 1$, where $S_1 := S - 5$, and $\mathcal{S}_2 := \{W, X, Y, Z\}$ contains four self-looping *terminal states*, which appear for all $h \geq 2$.

Action space. Our action space is given by $\mathcal{A} = \{1, 2\}$. For states in $\mathcal{S}_1$, the two actions have distinct effects, while for states in $\mathcal{S}_0 \cup \mathcal{S}_2$ both actions have identical effects. As a result, the value of a given policy only depends on the actions it selects at layer 1. For the sake of compactness, we use the symbol $\mathfrak{a}$ as a placeholder to denote either action when taken in $s \in \mathcal{S}_0 \cup \mathcal{S}_2$, since the choice is immaterial.[5]

Transition operator. For an MDP $M_{I,\alpha,\beta,\mathbf{r}}$, we let $I \subseteq \mathcal{S}_1$ parameterize a subset of the first-layer states. We call each $s \in I$ a *planted state* and $s \in \mathcal{S}_1 \setminus I$ an *unplanted state*. The dynamics $P_{I,\alpha,\beta}$ for $M_{I,\alpha,\beta,\mathbf{r}}$ are determined by I and the parameters $\alpha, \beta \in (0, 1)$ as follows (cf. Figure 1a):

- *Layer 0 to 1* ($h = 0$). We define $P_{I,\alpha,\beta}(\cdot \mid \mathfrak{s}, \mathfrak{a}) = \mathrm{Unif}(I)$. That is, from the initial state $\mathfrak{s}$ (and taking either action), the MDP transitions to each planted state with equal probability; unplanted states are not reachable.

[5] It is conceptually simpler to consider a construction where only a single action is available in $\mathcal{S}_0$ and $\mathcal{S}_1$, but this is notationally more cumbersome.

- *Layer 1 to 2* ($h = 1$).
  - Choosing action 1 in layer 1 always leads to state W in layer 2, i.e., $P_{I,\alpha,\beta}(W \mid s, 1) = 1 \; \forall s \in \mathcal{S}_1$. See the blue arrows in Figure 1a.
  - Choosing action 2 in layer 1 transitions stochastically to either $\{X, Y\}$ or $\{X, Z\}$, depending on whether the source state $s \in \mathcal{S}_1$ is planted or unplanted; see the red arrows in Figure 1a. Formally:

$$\begin{aligned} \forall s \in I: \quad & P_{I,\alpha,\beta}(X \mid s, 2) = \alpha, \quad P_{I,\alpha,\beta}(Y \mid s, 2) = 1 - \alpha; \\ \forall s \in \mathcal{S}_1 \setminus I: \quad & P_{I,\alpha,\beta}(X \mid s, 2) = \beta, \quad P_{I,\alpha,\beta}(Z \mid s, 2) = 1 - \beta. \end{aligned}$$

- *Layer 2 onwards* ($h \geq 2$). All states in $\mathcal{S}_2$ self-loop indefinitely. That is $P_{I,\alpha,\beta}(s \mid s, \mathfrak{a}) = 1 \; \forall s \in \mathcal{S}_2$.

Reward function. States in layer 0 and layer 1 have no reward, i.e., $R_{\mathbf{r}}(s, a) = 0, \forall s \in \mathcal{S}_0 \cup \mathcal{S}_1, \forall a \in \mathcal{A}$. Starting from layer 2 (i.e. $\mathcal{S}_2 = \{W, X, Y, Z\}$), each of the self-looping terminal states $\{W, X, Y, Z\}$ has a fixed reward determined by the parameter $\mathbf{r} = (r_W, r_X, r_Y, r_Z)$. In particular, we define $R_{\mathbf{r}}(W, \mathfrak{a}) = r_W$, $R_{\mathbf{r}}(X, \mathfrak{a}) = r_X$, $R_{\mathbf{r}}(Y, \mathfrak{a}) = r_Y$, and $R_{\mathbf{r}}(Z, \mathfrak{a}) = r_Z$ (recall that $\mathfrak{a}$ describes either action in these states).

Initial state distribution. All MDPs in $M_{I,\alpha,\beta,\mathbf{r}}$ start at $\mathfrak{s}$ deterministically (that is, the initial state distribution d_0 puts all the probability mass on $\mathfrak{s}$). Note that since d_0 does not vary between instances, it may be thought of as *known* to the learning algorithm.

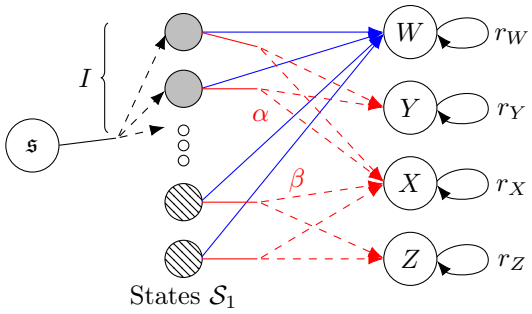
2.1.2 Intuition Behind the Construction

The family of MDPs $\mathcal{M}$ that witnesses our lower bound is a subset of the collection $\{M_{I,\alpha,\beta,\mathbf{r}}\}$. Before specifying the family precisely, we give intuition as to why this MDP structure leads to statistical hardness for offline reinforcement learning. The hardness arises as a result of two general principles, *planted subset structure* and *over-coverage*

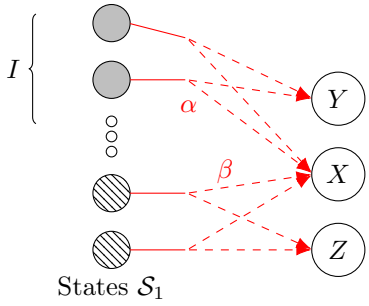
Planted Subset Structure. We use the planted subset structure specified by $I \subseteq \mathcal{S}_1$ as a mechanism to make the state space size appear in the sample complexity lower bound. For exposition, Figure 1b highlights only the parts of the MDP that are relevant to understanding the role of the planted subset structure.

Recall that for the first-layer states $\mathcal{S}_1 = [S_1]$, the planted subset $I \subset \mathcal{S}_1$ has the property that all states in I (planted states) share the same transition rule, and all states in $\mathcal{S}_1 \setminus I$ (unplanted states) share a different transition rule. The choice of the planted subset $I \subset \mathcal{S}_1$ is combinatorial in nature (for example, the number of all planted subsets of size $S_1/2$ is $\binom{S_1}{S_1/2}$, which is exponential in S_1), which is the basis for the statistical hardness of our construction. Consider Figure 1b, which has transitions from $\mathcal{S}_1$ for action 2 illustrated in red. Each planted state in I (grey) transitions to X and Y with probability α and $1 - \alpha$ respectively, while each unplanted state in $\mathcal{S}_1 \setminus I$ (striped) transitions to X and Z with probability β and $1 - \beta$ respectively. Suppose we are given a batch dataset of independent examples in which a state in $\mathcal{S}_1$ is selected uniformly at random, and we observe a sample from the next state distribution when action 2 is taken; we focus on action 2 because action 1 reveals no information about the underlying MDP. One can show that basic statistical inference tasks such as estimating the size $|I|$ of the planted subset require poly(S_1) sample complexity, as this entails detecting the subset based on data generated from a mixture of planted and unplanted states. For example, it is well known that testing if a distribution is uniform on a set $I \subset [N]$ with $|I| = \Theta(N)$ versus uniform on all of $[N]$ requires poly(N) samples (see e.g., Paninski, 2008; Ingster and Suslina, 2012; Canonne, 2020, Section 5.1).

Building on this hardness, one can also show any algorithm requires at least poly(S_1) samples to reliably estimate the transition probability α if β and $|I|$ are unknown. Intuitively, this arises because the only way to escape from the sample complexity of estimating $|I|$ as a means to estimate α is to directly look at the marginal distribution over $\{X, Y, Z\}$. However, the marginal distribution is uninformative for estimating α when there is uncertainty about β and $|I|$. For example, the marginal probability of transitioning to Y is $(1 - \alpha)|I|/S_1$, from which α cannot be directly recovered if $|I|$ is unknown.



(a) Structure of MDPs in $\mathcal{M}$.



(b) Planted subset structure.

Figure 1: The MDPs in $\mathcal{M}$ are parametrized by a subset I, probabilities α and β, and last-state rewards r_W, r_X, r_Y, r_Z. Layer 1 has states $\mathcal{S}_1$, where $I \subset \mathcal{S}_1$ are the gray *planted states* and the remaining $\mathcal{S}_1 \setminus I$ are the striped *unplanted states*; there are combinatorially many choices for I. From states at layer 1, action 1 (in blue) transitions to state W at layer 2. From planted states, action 2 (in red) transitions with probability α to state X and $1-\alpha$ to state Y, while from unplanted states, action 2 transitions with probability β to X and $(1-\beta)$ to Z. In the starting state $\mathfrak{s}$ and the states at layer 2, both actions have the same effect, with transitions denoted in black. At layer 2, these transitions are decorated with the corresponding reward.

The takeaway is that while estimating α would be trivial if the dataset only consisted of transitions generated from planted states, estimating this parameter when states are drawn uniformly from $\mathcal{S}_1$ is very difficult because an unknown subset comes from unplanted states. This is relevant because—as we will show—one can choose a collection of values for the parameters α, β, $|I|$ such that any near-optimal policy learning algorithm can be used to recover the parameter value.

Another important feature of the planted subset structure is that different choices of $I \subset \mathcal{S}_1$ lead to the same $Q^\star$ function, as long as we ensure that all states in $\mathcal{S}_1$ have the same value. This allows us to construct a large number of MDPs (exponential in $|\mathcal{S}|$) for which the $Q^\star$ function is one of two candidates (i.e., $|\mathcal{F}| = 2$). However, it remains to show that the hardness described above can be embedded in the offline RL setting, since (i) we must ensure concentrability is satisfied, and (ii) the learner observes rewards, not just transitions.

Over-coverage. Returning to Figure 1a, we observe that the transitions from the initial state $\mathfrak{s}$ are such that all planted states in I are *reachable*, but the unplanted states in $\mathcal{S}_1 \setminus I$ are *not reachable by any policy*. In particular, since all unplanted states are unreachable, any state that can only be reached from unplanted states is also unreachable, and hence we can achieve concentrability (1) without covering such states. This allows us to choose the data distribution μ to be (roughly) uniform over all states except for the unreachable state Z. This choice satisfies concentrability, but renders all reward observations uninformative (cf. Section 2.1.1). As a consequence, we show that the task described in Section 2.1.2, i.e., detecting α based on transition data unavoidable for any algorithm with non-trivial offline RL performance.

The key principle at play here is over-coverage. Per the discussion above, we know that if the data distribution μ happened to only be supported over reachable states for a given MDP, all layer-1 examples (s, a, r, s') in D_n would have $s \in I$, which would make estimating α trivial. Our construction for μ is uniform over all states in $\mathcal{S}_1$, and hence satisfies over-coverage, since it is supported over a mix of planted states and spurious (unplanted) states not reachable by any policy. This makes estimating α challenging because—due to correlations between planted and unplanted states—no algorithm can accurately estimate transitions or recover the planted states until the number of samples scales with the number of states.

2.1.3 Specifying the MDP Family

Using the parameterized MDP family $\{M_{I,\alpha,\beta,\mathbf{r}}\}$, we construct the hard family $\mathcal{M}$ for our lower bound by selecting a specific collection of values for the parameters $(I, \alpha, \beta, \mathbf{r})$.

Define $\mathcal{I}_\theta := \{I : |I| = \theta S_1\}$ for all $\theta \in (0,1)$ such that θS_1 is an integer. We define two sub-families of MDPs,

$$\mathcal{M}_1 := \bigcup_{I \in \mathcal{I}_{\theta_1}} \{M_{I,\alpha_1,\beta_1,\mathbf{r}_1}\}, \quad \text{and} \quad \mathcal{M}_2 := \bigcup_{I \in \mathcal{I}_{\theta_2}} \{M_{I,\alpha_2,\beta_2,\mathbf{r}_2}\},$$

where $\mathcal{M}_1$ is specified by $(\theta_1, \alpha_1, \beta_1) = (1/2, 1/4, 3/4)$ and $\mathbf{r}_1 = (4/5, 1, 2/3, 0)$, and $\mathcal{M}_2$ is specified by $(\theta_2, \alpha_2, \beta_2) = (3/4, 1/2, 1/2)$ and $\mathbf{r}_2 = (4/5, 1, 2/3, 2/3)$.[6] Finally, we define the hard family $\mathcal{M}$ via

$$\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2.$$

Let us discuss some basic properties of the construction that will be used to prove the lower bound.

- For all MDPs in $\mathcal{M}$, we have $r_W = 4/5$, $r_X = 1$ and $r_Y = 2/3$. This means there is no uncertainty in the reward function outside of state Z, which has $r_Z = 0$ when $M \in \mathcal{M}_1$ and $r_Z = 2/3$ when $M \in \mathcal{M}_2$. The values for r_Z are chosen to ensure that all states in $\mathcal{S}_1$ have the same value under action 2, given the choices for the other parameters above.
- All MDPs in $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) differ only in the choice of $I \subset \mathcal{S}_1$. This property, along with the fact that all states in $\mathcal{S}_1$ have the same value (guaranteed by our choice of r_Z), ensures that $Q^\star_M$ is the same for all $M \in \mathcal{M}_1$ (resp. $M \in \mathcal{M}_2$). Furthermore, our choice for $(\alpha_1, \beta_1, \mathbf{r}_1)$ (resp. $(\alpha_2, \beta_2, \mathbf{r}_2)$) ensures that the optimal action for all states in $\mathcal{S}_1$ is action 1 (resp. action 2).
- Our choice for $(\theta_1, \alpha_1, \beta_1)$ and $(\theta_2, \alpha_2, \beta_2)$ ensures that the marginal distribution of s' under the process $s \sim \mathrm{Unif}(\mathcal{S}_1), s' \sim P(\cdot \mid s, 2)$ is the same for all $M \in \mathcal{M}$. This property is motivated by the hard inference task described in Section 2.1.2, which requires an uninformative marginal distribution.

The exact numerical values for the MDP parameters chosen above are not essential to the hardness result. Any tuple $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; \mathbf{r}_1 = (r_W, r_X, r_Y, r_{1Z}); \mathbf{r}_2 = (r_W, r_X, r_Y, r_{2Z}))$ can be used to establish a result similar to Theorem 1, as long as it satisfies five general properties described in Appendix A.

2.1.4 Finishing the Construction: Value Functions and Data Distribution

We complete our construction by specifying a value function class $\mathcal{F}$ that satisfies realizability and a data distribution μ that satisfies concentrability (1).

Value function class. Define functions $f_1, f_2 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows; differences are highlighted in blue:

$$f_1(s,a) := \frac{1}{1-\gamma} \cdot \begin{cases} \frac{4}{5}\gamma^2, & s = \mathfrak{s} \\ \frac{4}{5}\gamma, & s \in \mathcal{S}_1,\ a = 1 \\ \frac{3}{4}\gamma, & s \in \mathcal{S}_1,\ a = 2 \\ \frac{4}{5}, & s = W \\ 1, & s = X \\ \frac{2}{3}, & s = Y \\ 0, & s = Z \end{cases} \quad \text{and} \quad f_2(s,a) := \frac{1}{1-\gamma} \cdot \begin{cases} \frac{5}{6}\gamma^2, & s = \mathfrak{s} \\ \frac{4}{5}\gamma, & s \in \mathcal{S}_1,\ a = 1 \\ \frac{5}{6}\gamma, & s \in \mathcal{S}_1,\ a = 2 \\ \frac{4}{5}, & s = W \\ 1, & s = X \\ \frac{2}{3}, & s = Y \\ \frac{2}{3}, & s = Z \end{cases}. \tag{2}$$

The following result is elementary; see Appendix B for a detailed calculation.

Proposition 1. *We have $Q^\star_M = f_1$ for all $M \in \mathcal{M}_1$ and $Q^\star_M = f_2$ for all $M \in \mathcal{M}_2$.*

It follows that by choosing $\mathcal{F} = \{f_1, f_2\}$, realizability holds for all $M \in \mathcal{M}$. Note that this choice of $\mathcal{F}$ satisfies the standard realizability condition that $Q^\star_M \in \mathcal{F}$ for all $M \in \mathcal{M}$ (as in the conjecture of Chen and Jiang (2019)), but—as is—does not satisfy the stronger all-policy realizability condition that $Q^\pi_M \in \mathcal{F}$ for *all* policies π (as in Theorem 1′). We handle this setting in Section 2.4.

[6] Recall that we assume without loss of generality that $S_1/4$ is an integer.

Data distribution. Recall that in the offline RL setting, the learner is provided with an i.i.d. dataset $D_n = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ where $(s_i, a_i) \sim \mu$, $s_i' \sim P(\cdot \mid s_i, a_i)$, and $r_i = R(s_i, a_i)$ (here P and R are the transition and reward functions for the underlying MDP). We define the data collection distribution via:

$$\mu = \frac{1}{8}\mathrm{Unif}(\{\mathfrak{s}\} \times \{1,2\}) + \frac{1}{2}\mathrm{Unif}(\mathcal{S}_1 \times \{1,2\}) + \frac{3}{8}\mathrm{Unif}(\{W, X, Y\} \times \{1,2\}).$$

This choice for μ forces the learner to suffer from the hardness described in Section 2.1.2. Salient properties include: (i) both planted and unplanted states in $\mathcal{S}_1$ are covered, and (ii) the state Z is not covered. Property (i) results in over-coverage, which makes estimating the parameters of the underlying MDP from transitions statistically hard, while property (ii) hides the difference between $\mathbf{r}_1$ and $\mathbf{r}_2$ and hence makes all reward observations uninformative.

We now verify the concentrability condition (1).

- For layer $h = 0$, for all $\pi : \mathcal{S} \to \mathcal{A}$, the distribution of (s_0, a_0) is $d_0 \times \pi$. It follows that

$$\left\| \frac{d_0 \times \pi}{\mu} \right\|_\infty \le \frac{1}{\frac{1}{8} \cdot \frac{1}{2}} = 16.$$

- For layer $h = 1$, for any $\pi : \mathcal{S} \to \mathcal{A}$, the distribution of (s_1, a_1) is $\mathrm{Unif}(I) \times \pi$. We conclude that

$$\left\| \frac{\mathrm{Unif}(I) \times \pi}{\mu} \right\|_\infty \le \frac{\frac{1}{S_1/2}}{\frac{1}{2} \cdot \frac{1}{S_1} \cdot \frac{1}{2}} = 8,$$

  where we have used that $|I| \ge S_1/2$.

- For layer $h \ge 2$, for any $\pi : \mathcal{S} \to \mathcal{A}$, the distribution of (s_h, a_h) (denoted by d_h^π) is supported on $\{W, X, Y\} \times \{1,2\}$. Therefore, we have

$$\left\| \frac{d_h^\pi}{\mu} \right\|_\infty \le \frac{1}{\frac{3}{8} \cdot \frac{1}{3} \cdot \frac{1}{2}} = 16.$$

We conclude that the construction satisfies concentrability with $C_{\mathrm{conc}} = 16$.

2.2 Proof of Theorem 1

Having specified the lower bound construction, we proceed to prove Theorem 1. For any MDP $M \in \mathcal{M}$, we know from (2) that the optimal policy $\pi_M^\star$ has

$$\pi_M^\star(s) = \begin{cases} 1, & \text{if } M \in \mathcal{M}_1, \\ 2, & \text{if } M \in \mathcal{M}_2, \end{cases} \quad \forall s \in \mathcal{S}_1,$$

and that $Q_M^\star$ has a constant gap in value between the optimal and suboptimal actions for states in $\mathcal{S}_1$:

$$Q_M^\star(s, \pi_M^\star(s)) - Q_M^\star(s, a) \ge \frac{1}{30}\frac{\gamma}{1-\gamma}, \quad \forall s \in \mathcal{S}_1, \forall a \ne \pi_M^\star(s). \tag{3}$$

Meanwhile, the choice of actions at states $s \in \mathcal{S}_0 \cup \mathcal{S}_2$ does not affect the policy value. This implies that any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ with high reward must choose action 1 for almost all reachable states in layer 1 when $M \in \mathcal{M}_1$, and must choose action 2 for almost all such states when $M \in \mathcal{M}_2$. As a result, any offline RL algorithm with non-trivial performance must reliably distinguish between $M \in \mathcal{M}_1$ and $M \in \mathcal{M}_2$ using the offline dataset D_n. In what follows we make this intuition precise.

For each $M \in \mathcal{M}$, let $\mathbb{P}_n^M$ denote the law of the offline dataset D_n when the underlying MDP is M, and let $\mathbb{E}_n^M$ be the associated expectation operator. We formalize the idea of distinguishing between $M \in \mathcal{M}_1$ and $M \in \mathcal{M}_2$ using Lemma 1, which reduces the task of proving a policy learning lower bound to the task of upper bounding the total variation distance between two *mixture distributions* $\mathbb{P}_n^1 := \frac{1}{|\mathcal{M}_1|}\sum_{M \in \mathcal{M}_1} \mathbb{P}_n^M$ and $\mathbb{P}_n^2 := \frac{1}{|\mathcal{M}_2|}\sum_{M \in \mathcal{M}_2} \mathbb{P}_n^M$.

Lemma 1. *Let $\gamma \in (0,1)$ be fixed. For any offline RL algorithm which takes $D_n = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ as input and returns a stochastic policy $\widehat{\pi}_{D_n} : \mathcal{S} \to \Delta(\mathcal{A})$, we have*

$$\sup_{M \in \mathcal{M}} \left\{ J_M(\pi_M^\star) - \mathbb{E}_n^M[J_M(\widehat{\pi}_{D_n})] \right\} \geq \frac{\gamma^2}{360(1-\gamma)} \left(1 - D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^2\right)\right).$$

Lemma 1 implies that if the difference between the average dataset generated by all $M \in \mathcal{M}_1$ and the average dataset generated by all $M \in \mathcal{M}_2$ is sufficiently small, no algorithm can reliably distinguish $M \in \mathcal{M}_1$ and $M \in \mathcal{M}_2$ based D_n, and hence must have poor performance on some instance. See Appendix C for a proof.

We conclude the proof by bounding $D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^2\right)$. Since directly calculating the total variation distance is difficult, we proceed in two steps. We first design an auxiliary *reference measure* $\mathbb{P}_n^0$, and then bound $D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^0\right)$ and $D_{\mathrm{TV}}\left(\mathbb{P}_n^2, \mathbb{P}_n^0\right)$ separately. For the latter step, we move from total variation distance to χ^2*-divergence* and derive upper bounds on $D_{\chi^2}\left(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0\right)$ (resp. $D_{\chi^2}\left(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0\right)$) using a mix of combinatorial arguments and concentration inequalities. This constitutes the most technical portion of the proof, and formalizes the intuition about hardness of estimation under planted subset structure described in Section 2.1.2.

Our final bound on the total variation distance (proven in Appendix D) is as follows.

Lemma 2. *For all $n \leq \sqrt[3]{(S-5)/400}$, we have*

$$D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^2\right) \leq 3/4.$$

Theorem 1 immediately follows by combining Lemma 1 and Lemma 2. □

2.3 Discussion

Having proven Theorem 1, we briefly interpret the result and discuss some additional consequences.

Separation between online and offline reinforcement learning. In the online reinforcement learning setting, the learner can execute any policy in the underlying MDP and observe the resulting trajectory. Our results show that in general, the separation between the sample complexity of online RL and offline RL can be arbitrarily large, even when concentrability is satisfied. To see this, recall that in the online RL setting, we can evaluate any fixed policy to precision ε using $\mathrm{poly}((1-\gamma)^{-1}) \cdot \varepsilon^{-2}$ trajectories via Monte-Carlo rollouts. Since the class $\mathcal{M}$ we construct only has two possible choices for the optimal policy and has suboptimality gap $\frac{\gamma^2}{1-\gamma}$, we can learn the optimal policy in the online setting using $\mathrm{poly}((1-\gamma)^{-1})$ trajectories, with no dependence on the number of states. On the other hand, Theorem 1 shows that the sample complexity of offline RL for this family can be made arbitrarily large.

Linear function approximation. The observation above is particularly salient in the context of linear function approximation, where $\mathcal{F} = \left\{(s,a) \mapsto \langle \phi(s,a), \theta \rangle : \theta \in \mathbb{R}^d\right\}$ for a known feature map $\phi(s,a)$. Our lower bound construction for Theorem 1 can be viewed as a special case of the linear function approximation setup with $d = 2$ by choosing $\phi(s,a) = (f_1(s,a), f_2(s,a))$. Consequently, our results show that the separation between the complexity of offline RL and online RL with linearly realizable function approximation can be arbitrarily large, even when the dimension is constant. This strengthens one of the results of Zanette (2021), which provides a linearly realizable construction in which the separation between online and offline RL is exponential with respect to dimension.

Why aren't stronger coverage or representation conditions satisfied? While our construction satisfies concentrability and realizability, it fails to satisfy stronger coverage and representation conditions for which sample-efficient upper bounds are known. This is to be expected (or else we would have a contradiction!) but understanding why is instructive. Here we discuss connections to some notable conditions.

Pushforward concentrability. The stronger notion of concentrability that $P(s' \mid s,a)/\mu(s') \leq C$ for all (s,a,s'), which is used in Xie and Jiang (2021), fails to hold because the state Z is not covered by μ. This presents no issue for standard concentrability because Z is not reachable starting from $\mathfrak{s}$.

Completeness. Bellman completeness requires that the value function class $\mathcal{F}$ has $\mathcal{T}_M\mathcal{F} \subseteq \mathcal{F}$ for all $M \in \mathcal{M}$, where $\mathcal{T}_M$ is the Bellman operator for M. We show in (2) that the set of optimal Q-value functions $\{Q^\star_M\}_{M\in\mathcal{M}}$ is small, but completness requires that the class remains closed even when we mix and match value functions and Bellman operators from $\mathcal{M}_1$ and $\mathcal{M}_2$, which results in an exponentially large class in our construction. To see why, first note that by Bellman optimality, we must have $\{Q^\star_M\}_{M\in\mathcal{M}} \subseteq \mathcal{F}$ if $\mathcal{F}$ is complete. We therefore also require $\mathcal{T}_{M'}Q^\star_M \in \mathcal{F}$ for $M \in \mathcal{M}_1$ and $M' \in \mathcal{M}_2$. Unlike the optimal Q-functions, which are constant across $\mathcal{S}_1$, the value of $[\mathcal{T}_{M'}Q^\star_M](s, 2)$ for $s \in \mathcal{S}_1$ depends on whether $s \in I$ or $s \in \mathcal{S}_1 \setminus I$, where I is the collection of planted states for M'.[7] As a result, there are $\binom{S_1}{|I|}$ possible values for the Bellman backup, which means that the cardinality of $\mathcal{F}$ must be exponential in S.

2.4 Extensions

Theorem 1 presents the simplest variant of our lower bound for clarity of exposition. In what follows we sketch some straightforward extensions.

- *Realizability for all policies (Theorem 1′).* The value function class $\mathcal{F} = \{f_1, f_2\}$ in the prequel satisfies the standard realizability condition that $Q^\star_M \in \mathcal{F}$ for all $M \in \mathcal{M}$ as in Chen and Jiang (2019), but does not satisfy the stronger *all-policy realizability* condition that $Q^\pi_M \in \mathcal{F}$ for all $\pi : \mathcal{X} \to \Delta(\mathcal{A})$. This is a simple issue to fix: General policies π will have different initial-state values $Q^\pi_M(\mathfrak{s}, \mathfrak{a})$, but otherwise agree with f_1 or f_2 at every other state.

  In more detail, a general policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ may select different actions on different states $s \in \mathcal{S}_1$, and the value $Q^\pi_M(\mathfrak{s}, \mathfrak{a})$ depends on the probability that π selects action 1 or 2 in $\mathcal{S}_1$. In particular, for any $M \in \mathcal{M}_1$, while we have $Q^\pi_M(s, a) = Q^\star_M(s, a) = f_1(s, a)$ for all $(s, a) \neq (\mathfrak{s}, \mathfrak{a})$, we have

$$Q^\pi_M(\mathfrak{s}, \mathfrak{a}) = \frac{\gamma^2}{1-\gamma} \cdot \left(\frac{4}{5}\mathbb{P}_{s\sim P(\mathfrak{s},\mathfrak{a})}(\pi(s) = 1) + \frac{3}{4}\mathbb{P}_{s\sim P(\mathfrak{s},\mathfrak{a})}(\pi(s) = 2)\right),$$

  which may not belong to $\{f_1(\mathfrak{s}, \mathfrak{a}), f_2(\mathfrak{s}, \mathfrak{a})\}$; MDPs in $\mathcal{M}_2$ satisfy a similar expression for $Q^\pi_M(\mathfrak{s}, \mathfrak{a})$.

  We address this issue by enlarging $\mathcal{F}$ using linear function approximation. Define a feature map

$$\phi(s, a) = \begin{cases} (1, 0, 0), & \text{if } (s, a) = (\mathfrak{s}, \mathfrak{a}), \\ (0, f_1(s, a), f_2(s, a)), & \text{otherwise.} \end{cases}$$

  Then for any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we have

$$Q^\pi_M(s, a) = \begin{cases} \langle\phi(s, a), (Q^\pi_M(\mathfrak{s}, \mathfrak{a}), 1, 0)\rangle, & \text{if } M \in \mathcal{M}_1, \\ \langle\phi(s, a), (Q^\pi_M(\mathfrak{s}, \mathfrak{a}), 0, 1)\rangle, & \text{if } M \in \mathcal{M}_2. \end{cases}$$

  for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. As a result, defining $\mathcal{F} = \left\{(s, a) \mapsto \langle\phi(s, a), \theta\rangle : \theta \in \mathbb{R}^3\right\}$, we have $Q^\pi_M \in \mathcal{F}$ for all π. One can verify that this construction has $\|\phi(s, a)\|_\infty, \|\theta\|_\infty \leq (1 - \gamma)^{-1}$, which proves Theorem 1′.

  Beyond linear function approximation, there are many other ways to enlarge $\mathcal{F}$ (from $\{f_1, f_2\}$) to obtain interesting consequences. For example, observing that $Q^\pi_M(\mathfrak{s}, \mathfrak{a})$ takes on at most $\mathcal{O}(S)$ different values across all $M \in \mathcal{M}$ and all *deterministic* policies π, we can simply define $\mathcal{F} = \{Q^\pi_M\}_{M\in\mathcal{M},\pi:\mathcal{S}\to\mathcal{A}}$, which is a finite class and has $Q^\pi_M \in \mathcal{F}$ for all deterministic π. For this new $\mathcal{F}$, since $|\mathcal{S}|, |\mathcal{F}| \propto S$, the interpretation of the lower bound should be that $n \gtrsim \min\left\{|\mathcal{S}|^{1/3}, |\mathcal{F}|^{1/3}\right\}$ samples are required, which is slightly weaker than the condition that $n \gtrsim |\mathcal{S}|^{1/3}$ in Theorem 1, but nonetheless rules out sample-efficient learning.

- *Policy evaluation.* Our lower bound immediately extends from policy optimization to policy evaluation. Indeed, letting $\pi^\star_1$ and $\pi^\star_2$ denote the optimal policies for $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, we have $|J_M(\pi^\star_1) - J_M(\pi^\star_2)| \propto \frac{\gamma^2}{1-\gamma}$ for all $M \in \mathcal{M}$. It follows that any algorithm which evaluates policies to precision

[7] Recall that f_1 is the optimal Q-function for any $M \in \mathcal{M}_1$ and consider $\mathcal{T}_{M'}f_1$ where $M' \in \mathcal{M}_2$ has planted set I. For $s \in I$, we have $[\mathcal{T}_{M'}f_1](s, 2) = (1/2 \cdot 1 + 1/2 \cdot 2/3)\gamma = 5/6\gamma$ while for $s \in \mathcal{S}_1 \setminus I$, we have $[\mathcal{T}_{M'}f_1](s, 2) = (1/2 \cdot 1 + 1/2 \cdot 0)\gamma = 1/2\gamma$.

$c \cdot \frac{\gamma^2}{1-\gamma}$ with probability at least $1-c'$ for small numerical constants $c, c' > 0$ can be used to select the optimal policy with constant probability, and hence must have $n = \Omega(|\mathcal{S}|^{1/3})$ by our lower bound.

To formally cast this setup in the policy evaluation setting, we take $\Pi = \{\pi_2^\star\}$ as the class of policies to be evaluated (recall that $J_M(\pi_1^\star)$ is constant across all $M \in \mathcal{M}$), and we require a value function class $\mathcal{F}$ such that $Q_M^\pi \in \mathcal{M}$ for all $\pi \in \Pi$, $M \in \mathcal{M}$. It suffices to select $\mathcal{F} = \{Q_M^{\pi_2^\star}, Q_{M'}^{\pi_2^\star}\}$ for an arbitrary choice of $M \in \mathcal{M}_1$ and $M' \in \mathcal{M}_2$, which has $|\mathcal{F}| \leq 2$. This choice for $\mathcal{F}$ is admissible because all $M \in \mathcal{M}_1$ lead to the same Q-value functions for $\pi_2^\star$, and likewise for $\mathcal{M}_2$.

- *Learning an ε-suboptimal policy.* Theorem 1 shows that for any $\gamma \in (1/2, 1)$, $n \gtrsim S^{1/3}$ samples are required to learn a $(1-\gamma)^{-1}$-optimal policy. We can extend the construction to show that more generally, for any $\varepsilon \in (0,1)$, $n \gtrsim \frac{S^{1/3}}{\varepsilon}$ samples are required to learn an $\varepsilon \cdot (1-\gamma)^{-1}$-optimal policy. We modify the MDP family $M_{I,\alpha,\beta,\mathbf{r}}$ by adding a single dummy state $\mathfrak{t}$ with a self-loop and zero reward. The initial state distribution is changed so that $d_0(\mathfrak{t}) = 1-\varepsilon$ and $d_1(\mathfrak{s}) = \varepsilon$. That is, with probability $1-\varepsilon$, the agent begins in $\mathfrak{t}$ and stays there forever, collecting no reward, and otherwise the agent begins at $\mathfrak{s}$ and proceeds as in the original construction. Analogously, we replace the original data distribution μ with $\mu' := (1-\varepsilon)\delta_{\mathfrak{t}} + \varepsilon\mu$, where $\delta_{\mathfrak{t}}$ is a point mass on $\mathfrak{t}$. This preserves the concentrability bound $C_{\mathrm{conc}} \leq 16$. This modification rescales the optimal value functions, and the conclusion of Lemma 1 is replaced by
$$\sup_{M \in \mathcal{M}} \left\{ J_M(\pi_M^\star) - \mathbb{E}_n^M[J_M(\widehat{\pi}_{D_n})] \right\} \geq \varepsilon \cdot \frac{\gamma^2}{360(1-\gamma)} \left(1 - D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^2\right)\right).$$
On the other hand, since samples from the state $\mathfrak{t}$ provide no information about the underlying instance, the effective number of samples is reduced to εn. One can make this intuition precise and prove that $D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^2\right) \leq 3/4$ whenever $\varepsilon n \leq c \cdot S^{1/3}$ for a numerical constant c. Combining this with the previous bound yields the result.

Lastly, it should be clear at this point that our lower bound construction extends to the finite-horizon setting with $H = 3$ by simply removing the self-loops from the terminal states. The only difference is that the optimal Q-value functions require a new calculation since rewards are no longer discounted.

3 Conclusion

We have proven that concentrability and realizability alone are not sufficient for sample-efficient offline reinforcement learning, resolving the conjecture of Chen and Jiang (2019). Our results establish that sample-efficient offline RL requires coverage or representation conditions than go beyond what is required for supervised learning, and show that over-coverage is a fundamental barrier within the standard formulation the offline RL problem. We close by discussing some open questions and directions for future research.

An immediate question is whether it is possible to circumvent our lower bound by restricting to data distributions induced by admissible policies, especially if one is allowed access to full trajectories rather than (s, a, r, s') tuples. However, there is a tradeoff here. While collecting data using policies is reasonable in many settings, in general this requires prior knowledge of the underlying MDP, and obliviously selected data gathering policies can have arbitrarily poor coverage (Xiao et al., 2021). Working with general data distributions $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ can allow for good coverage without prior knowledge, but subjects the learner to the limitations imposed by our lower bound.

More broadly, while our lower bound elucidates the role of concentrability and realizability, it remains to obtain a sharp, distribution-dependent characterization for the sample complexity of offline RL with general function approximation. Such a characterization would need to recover our result and previous results—both positive and negative—as special cases.

Acknowledgements

We thank Nan Jiang for insightful discussions and helpful feedback on a draft of the manuscript.

References

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E Schapire. Contextual bandit learning with predictable rewards. In *International Conference on Artificial Intelligence and Statistics*, 2012.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, 2020.

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the Wang-Foster-Kakade lower bound for the discounted setting. *arXiv:2011.01075*, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, 2003.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, 2020.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *International Conference on Learning Representations*, 2020.

Dylan J Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E Schapire. Practical contextual bandits with regression oracles. *International Conference on Machine Learning*, 2018.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2019.

Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine Learning Proceedings*, 1995.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv:1805.12298*, 2018.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963.

Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer Science & Business Media, 2012.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Neural Information Processing Systems*, 2020.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, 2018.

Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *International Conference on Robotics and Automation*, 2019.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 2019.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, 2019.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 2018.

Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, 2003.

Rémi Munos. Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM Journal on Control and Optimization*, 2007.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 2008.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 2021.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 2008.

Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours. In *International Conference on Robotics and Automation*, 2016.

Yury Polyanskiy. Lecture 1 of Information Theoretic Methods in Statistics and Computer Science. `http://people.lids.mit.edu/yp/homepage/data/LN_fdiv.pdf`, 2020. [Online; accessed Nov/6/2021].

John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

Stéphane Ross and J Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *International Conference on Machine Learning*, 2012.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 1996.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 1997.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, 2020.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv:2102.02981*, 2021.

Nicolas Verzelen and Elisabeth Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 2018.

Nicolas Verzelen and Fanny Villers. Goodness-of-fit tests for high-dimensional Gaussian linear models. *The Annals of Statistics*, 2010.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Ruosong Wang, Dean Foster, and Sham M Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2020.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham M Kakade. Instabilities of offline RL with pre-trained neural representation. *International Conference on Machine Learning*, 2021a.

Yuanhao Wang, Ruosong Wang, and Sham M Kakade. An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. *arXiv:2103.12690*, 2021b.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, 2021.

Chenjun Xiao, Ilbin Lee, Bo Dai, Dale Schuurmans, and Csaba Szepesvari. On the sample complexity of batch reinforcement learning with policy-induced data. *arXiv:2106.09973*, 2021.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, 2021.

Chao Yu, Guoqi Ren, and Jiming Liu. Deep inverse reinforcement learning for sepsis treatment. In *International Conference on Healthcare Informatics*, 2019.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *International Conference on Machine Learning*, 2021.

A General Scheme to Construct Hard Families of Instances

Recall that Section 2.1.3 gives specific numerical values for the parameters that define the model class $\mathcal{M}$ used in our lower bound construction. The precise values are not critical for our proof, and in this section we give general conditions on the parameters under which one can derive a similar lower bound. In doing so, we also provide some intuition behind the specific choice of parameters used for Theorem 1.

In more detail, for any tuple of parameters

$$(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; \mathbf{r}_1 = (r_W, r_X, r_Y, r_{1Z}); \mathbf{r}_2 = (r_W, r_X, r_Y, r_{2Z})),$$

we consider the family of MDPs $\mathcal{M}$ given by

$$\mathcal{M}_1 := \bigcup_{I \in \mathcal{I}_{\theta_1}} M_{I,\alpha_1,\beta_1,\mathbf{r}_1}, \quad \mathcal{M}_2 := \bigcup_{I \in \mathcal{I}_{\theta_2}} M_{I,\alpha_2,\beta_2,\mathbf{r}_2}, \quad \mathcal{M} := \mathcal{M}_1 \cup \mathcal{M}_2.$$

There are 11 independent scalars in the tuple, all of which lie in $[0, 1]$: $\theta_1, \alpha_1, \beta_1, \theta_2, \alpha_2, \beta_2, r_W, r_X, r_Y, r_{1Z}, r_{2Z}$; note that the parameters r_W, r_X, r_Y are shared between $\mathcal{M}_1$ and $\mathcal{M}_2$. The family $\mathcal{M}$ above can be used to derive a hardness result similar to Theorem 1 as long as the following five general equality and inequality constraints are satisfied.

- All $M \in \mathcal{M}$ should have the same marginal distribution for s' under the process $s \sim \mathrm{Unif}(\mathcal{S}_1)$, $s' \sim P(\cdot \mid s, 2)$:

  $$\theta_1(1-\alpha_1) = \theta_2(1-\alpha_2), \quad \text{and} \quad (1-\theta_1)(1-\beta_1) = (1-\theta_2)(1-\beta_2). \tag{4}$$

  This ensures that the learner cannot trivially test whether $M \in \mathcal{M}_1$ or $\mathcal{M}_2$ using marginals, which is tacitly used in the proof of Lemma 2.

- All $M \in \mathcal{M}_1$ (resp. $M \in \mathcal{M}_2$) should have the same value for action 2 across all states in $\mathcal{S}_1$:

  $$\alpha_1 r_X + (1-\alpha_1) r_Y = \beta_1 r_X + (1-\beta_1) r_{1Z}, \quad \text{and} \quad \alpha_2 r_X + (1-\alpha_2) r_Y = \beta_2 r_X + (1-\beta_2) r_{2Z}. \tag{5}$$

  This ensures that the family $\mathcal{M}$ is realizable with $|\mathcal{F}| = 2$.

- The parameters $\theta_1, \alpha_1, \beta_1, \theta_2, \alpha_2, \beta_2$ should be bounded away from 0 and 1:

  $$\theta_1, \alpha_1, \beta_1, \theta_2, \alpha_2, \beta_2 \in (0, 1). \tag{6}$$

  In particular, the distance from the boundary should be a constant independent of $\frac{1}{|\mathcal{S}|}$ and γ.

- Action 1 should be strictly better (resp. worse) than action 2 if $M \in \mathcal{M}_1$ (resp. $M \in \mathcal{M}_2$):

  $$\alpha_1 r_X + (1-\alpha_1) r_Y < r_W < \alpha_2 r_X + (1-\alpha_2) r_Y. \tag{7}$$

  The final lower bound depends on this separation quantitatively.

- For any $M \in \mathcal{M}_1$ and $M' \in \mathcal{M}_2$, the respective state distributions for $\mathcal{S}_1$ should have overlapping support on a constant fraction of states.

  $$\theta_1 + \theta_2 > 1. \tag{8}$$

  This, together with the previous bullet, implies that any algorithm that returns a near-optimal policy can be used to test whether $M \in \mathcal{M}_1$ or $\mathcal{M}_2$.

Any tuple simultaneously satisfying Eqs. (4) to (8) is sufficient for our proof (modulo numerical differences). Naturally, the numerical values for the function class $\mathcal{F}$ defined in (2) must be changed accordingly so that the class contains $Q^\star$ for both $\mathcal{M}_1$ and $\mathcal{M}_2$.

B Computation of Value Functions (Proposition 1)

In this section, we verify Proposition 1, which asserts that $Q^\star_M = f_1$ for all $M \in \mathcal{M}_1$ and $Q^\star_M = f_2$ for all $M \in \mathcal{M}_2$, where f_1 and f_2 are defined in (2). Note that the calculation we present here is based on the precise values for the parameters $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; \mathbf{r}_1, \mathbf{r}_2)$ given in Section 2.1.3, not the general scheme given in Appendix A.

Proof of Proposition 1. Suppose $M \in \mathcal{M}_1$. Let I_M denote the planted subset associated with M. First, for any self-looping terminal state $s \in \{W, X, Y, Z\}$, for any action $a \in \mathcal{A}$, we have

$$V^\star_M(s) = Q^\star_M(s,a) = \sum_{h=0}^{\infty} \gamma^h R_{\mathbf{r}_1}(s,a) = \frac{1}{1-\gamma} \cdot \begin{cases} \frac{4}{5}, & s = W, \\ 1, & s = X, \\ \frac{2}{3}, & s = Y, \\ 0, & s = Z. \end{cases}$$

Next, for any first-layer state $s \in \mathcal{S}_1$, by the Bellman optimality equation, we have

$$\begin{aligned} Q^\star_M(s,1) &= R_{\mathbf{r}_1}(s,1) + \gamma \mathbb{E}_{s' \sim P_{I_M, \alpha_1, \beta_1, \mathbf{r}_1}(s,1)}[V^\star_M(s')] \\ &= 0 + \gamma V^\star_M(W) \\ &= \frac{\gamma}{1-\gamma}\frac{4}{5}, \end{aligned}$$

and likewise

$$\begin{aligned} Q^\star_M(s,2) &= R_{\mathbf{r}_1}(s,2) + \gamma \mathbb{E}_{s' \sim P_{I_M, \alpha_1, \beta_1, \mathbf{r}_1}(s,2)}[V^\star_M(s')] \\ &= \begin{cases} 0 + \gamma[\alpha_1 V^\star_M(X) + (1-\alpha_1) V^\star_M(Y)], & s \in I_M \\ 0 + \gamma[\beta_1 V^\star_M(X) + (1-\beta_1) V^\star_M(Z)], & s \in \mathcal{S}_1 \setminus I_M \end{cases} \\ &= \begin{cases} \frac{\gamma}{1-\gamma}(\frac{1}{4} \times 1 + \frac{3}{4} \times \frac{2}{3}), & s \in I_M \\ \frac{\gamma}{1-\gamma}(\frac{3}{4} \times 1 + \frac{1}{4} \times 0), & s \in \mathcal{S}_1 \setminus I_M \end{cases} \\ &= \frac{\gamma}{1-\gamma}\frac{3}{4}. \end{aligned}$$

Since $Q^\star_M(s,1) = \frac{\gamma}{1-\gamma}\frac{4}{5} > \frac{\gamma}{1-\gamma}\frac{3}{4} = Q^\star_M(s,2)$ for all $s \in \mathcal{S}_1$, we have $\pi^\star_M(s) = 1$ and $V^\star_M(s) = Q^\star_M(s,1) = \frac{\gamma}{1-\gamma}\frac{4}{5}$ for all $s \in \mathcal{S}_1$. Thus, for the initial state $\mathfrak{s}$, for any action $a \in \mathcal{A}$, we have

$$Q^\star_M(\mathfrak{s},a) = R_{\mathbf{r}_1}(\mathfrak{s},a) + \gamma \mathbb{E}_{s' \sim P_{I_M, \alpha_1, \beta_1, \mathbf{r}_1}(\mathfrak{s},a)}[V^\star_M(s')] = 0 + \gamma \mathbb{E}_{s' \sim \mathrm{Unif}(I_M)} V^\star_M(s') = \frac{\gamma^2}{1-\gamma}\frac{4}{5}.$$

Therefore, $Q^\star_M(s,a) = f_1(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Now suppose $M \in \mathcal{M}_2$. Let I_M denote the planted subset associated with M. For any self-looping terminal state $s \in \{W, X, Y, Z\}$ and any action $a \in \mathcal{A}$, we have

$$V^\star_M(s) = Q^\star_M(s,a) = \sum_{h=0}^{\infty} \gamma^h R_{\mathbf{r}_2}(s,a) = \frac{1}{1-\gamma} \cdot \begin{cases} \frac{4}{5}, & s = W, \\ 1, & s = X, \\ \frac{2}{3}, & s = Y, \\ \frac{2}{3}, & s = Z. \end{cases}$$

For any first-layer state $s \in \mathcal{S}_1$, by the Bellman optimality equation, we have

$$\begin{aligned} Q^\star_M(s,1) &= R_{\mathbf{r}_2}(s,1) + \gamma \mathbb{E}_{s' \sim P_{I_M, \alpha_2, \beta_2, \mathbf{r}_2}(s,1)}[V^\star_M(s')] \\ &= 0 + \gamma V^\star_M(W) \\ &= \frac{\gamma}{1-\gamma}\frac{4}{5}, \end{aligned}$$

and likewise

$$\begin{aligned}
Q^\star_M(s,2) &= R_{\mathbf{r}_2}(s,2) + \gamma \mathbb{E}_{s' \sim P_{I_M,\alpha_2,\beta_2,\mathbf{r}_2}(s,2)}[V^\star_M(s')] \\
&= \begin{cases} 0 + \gamma[\alpha_2 V^\star_M(X) + (1-\alpha_2) V^\star_M(Y)], & s \in I_M \\ 0 + \gamma[\beta_2 V^\star_M(X) + (1-\beta_2) V^\star_M(Z)], & s \in \mathcal{S}_1 \setminus I_M \end{cases} \\
&= \begin{cases} \frac{\gamma}{1-\gamma}(\frac{1}{2} \times 1 + \frac{1}{2} \times \frac{2}{3}), & s \in I_M \\ \frac{\gamma}{1-\gamma}(\frac{1}{2} \times 1 + \frac{1}{2} \times \frac{2}{3}), & s \in \mathcal{S}_1 \setminus I_M \end{cases} \\
&= \frac{\gamma}{1-\gamma}\frac{5}{6}.
\end{aligned}$$

Since $Q^\star_M(s,1) = \frac{\gamma}{1-\gamma}\frac{4}{5} < \frac{\gamma}{1-\gamma}\frac{5}{6} = Q^\star_M(s,2)$ for all $s \in \mathcal{S}_1$, we have $\pi^\star_M(s) = 2$ and $V^\star_M(s) = Q^\star_M(s,2) = \frac{\gamma}{1-\gamma}\frac{5}{6}$ for all $s \in \mathcal{S}_1$. Thus, for the initial state $\mathfrak{s}$, for any action $a \in \mathcal{A}$, we have

$$Q^\star_M(\mathfrak{s},a) = R_{\mathbf{r}_2}(\mathfrak{s},a) + \gamma \mathbb{E}_{s' \sim P_{I_M,\alpha_2,\beta_2,\mathbf{r}_2}(\mathfrak{s},a)}[V^\star_M(s')] = 0 + \gamma \mathbb{E}_{s' \sim \mathrm{Unif}(I_M)} V^\star_M(s') = \frac{\gamma^2}{1-\gamma}\frac{5}{6}.$$

It follows that $Q^\star_M(s,a) = f_2(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. □

C Proof of Lemma 1

We now prove Lemma 1. Before proceeding, let us note that this lemma is proven only for the precise values for the parameters $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; \mathbf{r}_1, \mathbf{r}_2)$ given in Section 2.1.3. One could establish a more general lemma using the generic parameters introduced in Appendix A, but this would require changing the numerical constants appearing in the statement.

We begin the proof by lower bounding the regret for any MDP in the family $\mathcal{M}$. For any $i \in \{1,2\}$, any $I \in \mathcal{I}_{\theta_i}$, and any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, applying the performance difference lemma (Kakade and Langford (2002)) on the MDP $M := M_{I,\alpha_i,\beta_i,\mathbf{r}_i}$ gives

$$\begin{aligned}
J_M(\pi^\star_M) - J_M(\pi) &= \gamma \mathbb{E}_{s \sim \mathrm{Unif}(I)}[Q^\star_M(s, \pi^\star_M(s)) - Q^\star_M(s, \pi(s))] \\
&\geq \gamma \mathbb{P}_{s \sim \mathrm{Unif}(I)}(\pi(s) \neq \pi^\star_M(s)) \cdot \inf_{s \in \mathcal{S}_1} |Q^\star_M(s,1) - Q^\star_M(s,2)| \\
&\overset{\text{(i)}}{\geq} \gamma \mathbb{P}_{s \sim \mathrm{Unif}(I)}(\pi(s) \neq \pi^\star_M(s)) \cdot \frac{1}{30}\frac{\gamma}{1-\gamma} \\
&= \frac{\gamma^2}{30(1-\gamma)} \mathbb{P}_{s \sim \mathrm{Unif}(I)}(\pi(s) \neq \pi^\star_M(s)) \\
&= \frac{\gamma^2}{30(1-\gamma)} \frac{|\{s \in I : \pi(s) \neq \pi^\star_M(s)\}|}{|I|} \\
&\overset{\text{(ii)}}{\geq} \frac{\gamma^2}{30(1-\gamma)} \frac{(|\{s \in \mathcal{S}_1 : \pi(s) \neq \pi^\star_M(s)\}| + |I| - S_1)_+}{|I|} \\
&= \frac{\gamma^2}{30(1-\gamma)} \frac{(\mathbb{P}_{s \sim \mathrm{Unif}(\mathcal{S}_1)}(\pi(s) \neq \pi^\star_M(s)) + \theta_i - 1)_+}{\theta_i} \\
&= \frac{\gamma^2}{30(1-\gamma)} \left(\frac{\mathbb{P}_{s \sim \mathrm{Unif}(\mathcal{S}_1)}(\pi(s) \neq \pi^\star_M(s)) + \theta_i - 1}{\theta_i} \right)_+ ,
\end{aligned} \tag{9}$$

where (i) follows from the calculation for the value function gap in (3), and (ii) follows from the fact that for any $I_1, I_2 \subset \mathcal{S}_1$, $|I_1 \cap I_2| = |I_1| + |I_2| - |I_1 \cup I_2| \geq (|I_1| + |I_2| - S_1)_+$.

Now, consider any fixed offline reinforcement learning algorithm which takes the offline dataset D_n as an input and returns a stochastic policy $\widehat{\pi}_{D_n} : \mathcal{S} \to \Delta(\mathcal{A})$. Recall that $\mathbb{P}^M_n$ denotes the law of the offline dataset $D_n = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ when the data collection distribution is μ and the underlying MDP is M, and

that $\mathbb{E}_n^M$ is the associated expectation operator. Define events $A_1 := \{\mathbb{P}_{s\sim\mathrm{Unif}(\mathcal{S}_1)}(\widehat{\pi}_{D_n}(s) \neq 1) \geq 5/8\}$ and $A_2 := \{\mathbb{P}_{s\sim\mathrm{Unif}(\mathcal{S}_1)}(\widehat{\pi}_{D_n}(s) \neq 2) \geq 3/8\}$, and note that A_1 and A_2 are complementary, i.e., $A_2 = A_1^c$. By (9), for any $M \in \mathcal{M}_i$ ($i \in \{1,2\}$), we have

$$\begin{aligned}
\mathbb{E}_n^M[J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] &\geq \frac{\gamma^2}{30(1-\gamma)}\mathbb{E}_n^M\left[\left(\frac{\mathbb{P}_{s\sim\mathrm{Unif}(\mathcal{S}_1)}(\pi(s) \neq \pi_M^\star(s)) + \theta_i - 1}{\theta_i}\right)_+\right] \\
&\geq \begin{cases} \frac{\gamma^2}{30(1-\gamma)}\mathbb{P}_n^M(A_1)\left(\frac{5/8+1/2-1}{1/2}\right)_+, & \text{if } i=1 \\ \frac{\gamma^2}{30(1-\gamma)}\mathbb{P}_n^M(A_2)\left(\frac{3/8+3/4-1}{3/4}\right)_+, & \text{if } i=2 \end{cases} \qquad (10) \\
&\geq \frac{\gamma^2}{180(1-\gamma)}\mathbb{P}_n^M(A_i), \qquad (11)
\end{aligned}$$

where the second inequality follows from the definition of A_i and the parameter value for θ_i.

For each $i \in \{1,2\}$, we apply (10) to all MDPs in $\mathcal{M}_i$ and average to obtain

$$\frac{1}{|\mathcal{M}_i|}\sum_{M\in\mathcal{M}_i}\mathbb{E}_n^M[J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \geq \frac{\gamma^2}{180(1-\gamma)}\frac{1}{|\mathcal{M}_i|}\sum_{M\in\mathcal{M}_i}\mathbb{P}_n^M(A_i).$$

Combining the inequalities for $i=1$ and $i=2$, we have

$$\begin{aligned}
&\max_{M\in\mathcal{M}}\mathbb{E}_n^M[J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \\
&\geq \frac{1}{2|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1}\mathbb{E}_n^M[J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] + \frac{1}{2|\mathcal{M}_2|}\sum_{M\in\mathcal{M}_2}\mathbb{E}_n^M[J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \\
&\geq \frac{\gamma^2}{360(1-\gamma)}\left\{\frac{1}{|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1}\mathbb{P}_n^M(A_1) + \frac{1}{|\mathcal{M}_2|}\sum_{M\in\mathcal{M}_2}\mathbb{P}_n^M(A_2)\right\} \\
&\geq \frac{\gamma^2}{360(1-\gamma)}\left(1 - D_{\mathrm{TV}}\left(\frac{1}{|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1}\mathbb{P}_n^M, \frac{1}{|\mathcal{M}_2|}\sum_{M\in\mathcal{M}_2}\mathbb{P}_n^M\right)\right),
\end{aligned}$$

where the last inequality follows from the elementary fact that for any pair of probability measures $\mathbb{P}, \mathbb{Q}$ and any event E,

$$\mathbb{P}(E) + \mathbb{Q}(E^c) = 1 - \max\{\mathbb{P}(E^c) - \mathbb{Q}(E^c), \mathbb{Q}(E) - \mathbb{P}(E)\} \geq 1 - D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}).$$

D Proof of Lemma 2

This proof is organized as follows. In Appendix D.1, we introduce a reference measure and move from the total variation distance to the χ^2-divergence. This allows us to reduce the task of upper bounding $D_{\mathrm{TV}}(\mathbb{P}_n^1, \mathbb{P}_n^2)$ to the task of upper bounding two manageable density ratios (Eqs. (13) and (14) in the sequel). We develop several intermediate technical lemmas related to the density ratios in Appendix D.2, and in Appendix D.3 we put everything together to bound the density ratios, thus completing the proof of Lemma 2.

For the statement of Lemma 2 and the main subsections of this appendix (Appendices D.1 and D.3), we only consider the specific values for the parameters

$$(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; \mathbf{r}_1 = (r_W, r_X, r_Y, r_{1Z}); \mathbf{r}_2 = (r_W, r_X, r_Y, r_{2Z}))$$

given in Section 2.1.3. However, in Appendix D.2, which contains intermediate technical lemmas, results are presented under a slightly more general setup (explained at the beginning of the subsection).

D.1 Introducing a Reference Measure and Moving to χ^2-Divergence

Directly calculating the total variation distance $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)$ is challenging, so we design an auxillary *reference measure* $\mathbb{P}_n^0$ which serves as an intermediate quantity to help with the upper bound. The reference measure $\mathbb{P}_n^0$ lies in the same measurable space as $\mathbb{P}_n^1$ and $\mathbb{P}_n^2$, and is defined as follows:

$$\mathbb{P}_n^0(\{(s_i, a_i, r_i, s_i')\}_{i=1}^n) := \prod_{i=1}^n \mu(s_i, a_i) \mathbb{1}_{\{r_i = R_0(s_i, a_i)\}} P_0(s_i' \mid s_i, a_i), \quad \forall\, \{(s_i, a_i, r_i, s_i')\}_{i=1}^n,$$

where

$$R_0(s, \mathfrak{a}) := \begin{cases} 0, & s \in \mathcal{S}_0 \cup \mathcal{S}_1, \\ r_W = 4/5, & s = W, \\ r_X = 1, & s = X, \\ r_Y = 2/3, & s = Y, \\ 0, & s = Z, \end{cases}$$

and

$$\begin{aligned} P_0(\cdot \mid \mathfrak{s}, \mathfrak{a}) &= \mathrm{Unif}(\mathcal{S}_1), \\ P_0(\cdot \mid s, 1) &= W, \text{ w.p. } 1, \qquad \forall s \in \mathcal{S}_1, \\ P_0(\cdot \mid s, 2) &= \begin{cases} X, & \text{w.p. } \theta_1\alpha_1 + (1-\theta_1)\beta_1, \\ Y, & \text{w.p. } \theta_1(1-\alpha_1), \\ Z, & \text{w.p. } (1-\theta_1)(1-\beta_1), \end{cases} \qquad \forall s \in \mathcal{S}_1, \\ P_0(\cdot \mid s, \mathfrak{a}) &= s, \text{ w.p. } 1, \qquad \forall s \in \mathcal{S}_2. \end{aligned}$$

The reference measure $\mathbb{P}_n^0$ can be understood as the law of D_n when the data collection distribution is μ and the underlying MDP is $M_0 := (\mathcal{S}, \mathcal{A}, P_0, R_0, \gamma, d_0)$. The value of $R_0(Z, \mathfrak{a})$ is immaterial, as the data collection distribution μ is not supported on $(Z, \mathfrak{a})$; we choose $R_0(Z, \mathfrak{a}) = 0$ for concreteness.

An important feature of the reference measure is that the families $\mathcal{M}_1$ and $\mathcal{M}_2$ in our construction satisfy the constraint (4), so that $\theta_1(1-\alpha_1) = \theta_2(1-\alpha_2)$ and $(1-\theta_1)(1-\beta_1) = (1-\theta_2)(1-\beta_2)$. As such, even though we define the transition operator P_0 above based on the tuple $(\theta_1, \alpha_1, \beta_1)$, substituting in $(\theta_2, \alpha_2, \beta_2)$ leads to the same operator.

Starting with the triangle inequality for the total variation distance, we have

$$\begin{aligned} D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) &\le D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^0\big) + D_{\mathrm{TV}}\big(\mathbb{P}_n^2, \mathbb{P}_n^0\big) \\ &\le \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0)} + \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0)}, \end{aligned} \tag{12}$$

where the last inequality follows from the fact that $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \le \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P} \,\|\, \mathbb{Q})}$ for any $\mathbb{P}, \mathbb{Q}$ (see Proposition 7.2 or Section 7.6 of Polyanskiy (2020)).

In what follows, we derive simplified expressions for $D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0)$ and $D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0)$. We first expand and simplify $D_{\chi^2}(\mathbb{P}_n^1, \mathbb{P}_n^0)$, then obtain a similar expression for $D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0)$.

For each MDP $M \in \mathcal{M}$, let P_M and R_M denote the associated transition and reward functions. Observe that our construction for P_M, R_M, and μ (see Section 2.1) ensures that for any $(s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times [0, 1] \times \mathcal{S}$ with $\mu(s, a)\mathbb{1}_{\{r = R_0(s,a)\}} P_0(s' \mid s, a) = 0$, we have $\mu(s, a)\mathbb{1}_{\{r = R_M(s,a)\}} P_M(s' \mid s, a) = 0$. As a result, we have

$\mathbb{P}_n^M \ll \mathbb{P}_n^0$ for any $M \in \mathcal{M}$, which implies that $\mathbb{P}_n^1, \mathbb{P}_n^2 \ll \mathbb{P}_n^0$. Hence, we can expand the χ^2-divergence as

$$
\begin{aligned}
&D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0) \\
&= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[\left(\frac{\frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{P}_n^M(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n)}{\mathbb{P}_n^0(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n)} \right)^2 \right] - 1 \\
&= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[\left(\frac{\frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \prod_{i=1}^n \mu(s_i,a_i) \mathbb{1}_{\{r_i = R_M(s_i,a_i)\}} P_M(s_i' \mid s_i,a_i)}{\prod_{i=1}^n \mu(s_i,a_i) \mathbb{1}_{\{r_i = R_0(s_i,a_i)\}} P_0(s_i' \mid s_i,a_i)} \right)^2 \right] - 1 \\
&= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[\left(\frac{\frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \prod_{i=1}^n P_M(s_i' \mid s_i,a_i)}{\prod_{i=1}^n P_0(s_i' \mid s_i,a_i)} \right)^2 \right] - 1 \\
&= \frac{1}{|\mathcal{M}_1|^2} \sum_{M,M' \in \mathcal{M}_1} \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[\frac{\prod_{i=1}^n P_M(s_i' \mid s_i,a_i) P_{M'}(s_i' \mid s_i,a_i)}{\prod_{i=1}^n P_0^2(s_i' \mid s_i,a_i)} \right] - 1 \\
&= \frac{1}{|\mathcal{M}_1|^2} \sum_{M,M' \in \mathcal{M}_1} \left(\mathbb{E}_{(s,a) \sim \mu, s' \sim P_0(\cdot \mid s,a)} \left[\frac{P_M(s' \mid s,a) P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)} \right] \right)^n - 1, \qquad (13)
\end{aligned}
$$

where the third equality follows because (i) $R_M(s,a) = R_0(s,a), \forall M \in \mathcal{M}, \forall a \in \mathcal{A}, \forall s \neq Z$, and (ii) state Z is not covered by μ. Indeed, since the reward function for every MDP in $\mathcal{M}$ is the same as R_0 *for all* (s,a) *covered by* μ, the rewards $r_1, \ldots, r_n$ in D_n are completely uninformative in our construction—they have the same distribution regardless of the underlying MDP. This is why the final expression for $D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0)$ in (13) is completely independent of the reward distribution for both measures.

Using an identical calculation, we also have

$$
D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0) = \frac{1}{|\mathcal{M}_2|^2} \sum_{M,M' \in \mathcal{M}_2} \left(\mathbb{E}_{(s,a) \sim \mu, s' \sim P_0(\cdot \mid s,a)} \left[\frac{P_M(s' \mid s,a) P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)} \right] \right)^n - 1. \qquad (14)
$$

Equipped with these expressions for the χ^2-divergence, the next step in the proof of Lemma 2 is to upper bound the right-hand side for Eqs. (13) and (14). This is done in Appendix D.3, but before proceeding we require several intermediate technical lemmas.

D.2 Technical Lemmas for Density Ratios

In this subsection, we state a number of technical lemmas which can be used to bound the density ratio appearing inside the square in Eqs. (13) and (14) for generic MDPs $M_{I,\alpha,\beta,\mathbf{r}}$ with $I \in \mathcal{I}_\theta$. The lemmas hold for any choice of (θ, α, β), and are independent of the reward parameter $\mathbf{r}$. For this general setup, we work with a variant of the reference operator P_0 defined based on the values (θ, α, β) via

$$
\begin{aligned}
&P_0(\cdot \mid \mathfrak{s}, \mathfrak{a}) = \mathrm{Unif}(\mathcal{S}_1), \\
&P_0(\cdot \mid s, 1) = W, \text{ w.p. } 1, \qquad \forall s \in \mathcal{S}_1, \\
&P_0(\cdot \mid s, 2) = \begin{cases} X, & \text{w.p. } \theta\alpha + (1-\theta)\beta, \\ Y, & \text{w.p. } \theta(1-\alpha), \\ Z, & \text{w.p. } (1-\theta)(1-\beta), \end{cases} \qquad \forall s \in \mathcal{S}_1, \\
&P_0(\cdot \mid s, \mathfrak{a}) = s, \text{ w.p. } 1, \qquad \forall s \in \mathcal{S}_2.
\end{aligned}
$$

In Appendix D.3, we instantiate the results from this subsection with $(\theta_i, \alpha_i, \beta_i)$ for $i \in \{1,2\}$. Recall that per the discussion in Appendix D.1, our specific parameter choices for the families $\mathcal{M}_1$ and $\mathcal{M}_2$ induce the same reference operator P_0.

Lemma 3. *For all* $I, I' \in \mathcal{I}_\theta$, $(2\theta - 1)_+ S_1 \leq |I \cap I'| \leq \theta S_1$.

Proof. Since $|I| = |I'| = \theta S_1$, we have $|I \cap I'| \le |I| = \theta S_1$ and

$$|I \cap I'| = |I| + |I'| - |I \cup I'| \ge |I| + |I'| - S_1 = (2\theta - 1)S_1.$$

Since $|I \cap I'| \ge 0$ trivially, the result follows. □

The next lemma controls the density ratio for states in $\mathcal{S}_1$ when action 2 is taken. To state the result compactly, we define

$$\phi_{\theta,\alpha,\beta} := \theta^2 \left(\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta} + \frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)} \right). \tag{15}$$

Lemma 4. *For all $I, I' \in \mathcal{I}_\theta$, we have*

$$\mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}_1), s'\sim P_0(\cdot|s,2)} \left[\frac{P_I(s' \mid s,2) P_{I'}(s' \mid s,2)}{P_0^2(s' \mid s,2)} \right] = 1 + \phi_{\theta,\alpha,\beta} \cdot \left(\frac{|I \cap I'|}{\theta^2 S_1} - 1 \right).$$

Proof. For any $I, I' \in \mathcal{I}_\theta$, we observe that

$$\mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}_1), s'\sim P_0(\cdot|s,2)} \left[\frac{P_I(s' \mid s,2) P_{I'}(s' \mid s,2)}{P_0^2(s' \mid s,2)} \right] = \mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}_1)} \left[\sum_{s'\in\{X,Y,Z\}} \frac{P_I(s' \mid s,2) P_{I'}(s' \mid s,2)}{P_0(s' \mid s,2)} \right].$$

To proceed, we calculate the value of the ratio $\frac{P_I(s'|s,2)P_{I'}(s'|s,2)}{P_0(s'|s,2)}$ for each possible choice for $s \in \mathcal{S}_1$ and $s' \in \{X, Y, Z\}$ in Table 1 below.

	$s' = X$	$s' = Y$	$s' = Z$
$s \in I \cap I'$	$\alpha^2/(\theta\alpha + (1-\theta)\beta)$	$(1-\alpha)/\theta$	0
$s \in (I \cup I') \setminus (I \cap I')$	$\alpha\beta/(\theta\alpha + (1-\theta)\beta)$	0	0
$s \notin (I \cup I')$	$\beta^2/(\theta\alpha + (1-\theta)\beta)$	0	$(1-\beta)/(1-\theta)$

Table 1: Value of $\frac{P_I(s'|s,2)P_{I'}(s'|s,2)}{P_0(s'|s,2)}$ for all possible pairs (s, s').

Define $t := |I \cap I'|$. From Lemma 3, we must have $t \in [(2\theta-1)_+ S, \theta S]$. We also have $|I \cup I'| = |I| + |I'| - |I \cap I'| = 2\theta S - t$. Hence, the event in the first row of Table 1 occurs with probability $|I \cap I'|/S = t/S$, the event in the second row occurs with probability $|(I \cup I') \setminus (I \cap I')|/S = (2\theta S - 2t)/S$ and the event in the third row occurs with probability $|S \setminus (I \cup I')|/S = ((1-2\theta)S + t)/S$. Using these values, we obtain

$$\begin{aligned}
&\mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}_1)} \left[\sum_{s'\in\{X,Y,Z\}} \frac{P_I(s' \mid s,2) P_{I'}(s' \mid s,2)}{P_0(s' \mid s,2)} \right] \\
&= \frac{t}{S} \cdot \left(\frac{\alpha^2}{\theta\alpha + (1-\theta)\beta} + \frac{1-\alpha}{\theta} \right) + \left(2\theta - \frac{2t}{S} \right) \cdot \frac{\alpha\beta}{\theta\alpha + (1-\theta)\beta} \\
&\quad + \left(1 - 2\theta + \frac{t}{S} \right) \cdot \left(\frac{\beta^2}{\theta\alpha + (1-\theta)\beta} + \frac{1-\beta}{1-\theta} \right) \\
&= \frac{t}{S} \left(\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta} + \frac{1-\theta\beta-\alpha(1-\theta)}{\theta(1-\theta)} \right) + \frac{2\theta(\alpha-\beta)\beta+\beta^2}{\theta(\alpha-\beta)+\beta} + \frac{(1-2\theta)(1-\beta)}{1-\theta} \\
&= \frac{t}{S} \left(\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta} + \frac{1-\theta\beta}{\theta(1-\theta)} - \frac{\alpha}{\theta} \right) + \frac{2\theta(\alpha-\beta)\beta+\beta^2}{\theta(\alpha-\beta)+\beta} + 1 - \beta - \frac{\theta(1-\beta)}{1-\theta} \\
&= \left(\frac{t}{S} - \theta^2 \right) \left(\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta} + \frac{1-\theta\beta}{\theta(1-\theta)} - \frac{\alpha}{\theta} \right) \\
&\quad + \underbrace{\theta^2 \cdot \frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}}_{\text{(i)}} + \underbrace{\theta^2 \cdot \frac{1-\theta\beta}{\theta(1-\theta)}}_{\text{(ii)}} - \underbrace{\theta^2 \cdot \frac{\alpha}{\theta}}_{\text{(iii)}} + \underbrace{\frac{2\theta(\alpha-\beta)\beta+\beta^2}{\theta(\alpha-\beta)+\beta}}_{\text{(i)}} - \underbrace{\frac{\theta(1-\beta)}{1-\theta}}_{\text{(ii)}} + \underbrace{1-\beta}_{\text{(iii)}}.
\end{aligned}$$

Grouping the terms in the second line together, we find that (i) $= \theta(\alpha-\beta)+\beta$, (ii) $= \theta\beta$, and (iii) $= 1-\theta\alpha-\beta$, and by summing,

$$\text{(i)} + \text{(ii)} + \text{(iii)} = 1.$$

Hence, we have the upper bound

$$\begin{aligned}&\left(\frac{t}{S}-\theta^2\right)\left(\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{1-\theta\beta}{\theta(1-\theta)}-\frac{\alpha}{\theta}\right)+1\\&=\left(\frac{t}{S}-\theta^2\right)\left(\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)}\right)+1.\end{aligned}$$

Recalling the definition of $\phi_{\theta,\alpha,\beta}$, this completes the proof. □

The next lemma bounds magnitude of $\phi_{\theta,\alpha,\beta}$ in terms of the parameter θ.

Lemma 5. *For any $\alpha, \beta, \theta \in (0,1)$, we have*

$$\theta^2|\alpha-\beta| \le \phi_{\theta,\alpha,\beta} \le \frac{\theta}{1-\theta}(1-\min\{\alpha,\beta\}) \le \frac{\theta}{1-\theta}.$$

Proof. Recall that $\phi_{\theta,\alpha,\beta} = \theta^2\Big(\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)}\Big)$. We consider two cases.

Case 1: $\alpha \le \beta$**.** Assume $\alpha < \beta$, as the result is immediate if $\alpha = \beta$. We have

$$\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)} \ge 0+\frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)} \ge \frac{\theta(\alpha-\beta)+\beta-\alpha}{\theta(1-\theta)} = \frac{\beta-\alpha}{\theta} > |\alpha-\beta|$$

and

$$\begin{aligned}\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)} &= \frac{(\beta-\alpha)^2}{\beta-\theta(\beta-\alpha)}+\frac{\alpha-\beta}{1-\theta}+\frac{1-\alpha}{\theta(1-\theta)}\\&\le \frac{(\beta-\alpha)^2}{(\beta-\alpha)-\theta(\beta-\alpha)}+\frac{\alpha-\beta}{1-\theta}+\frac{1-\alpha}{\theta(1-\theta)}\\&= \frac{1-\alpha}{\theta(1-\theta)},\end{aligned}$$

where the inequality above uses that $\beta > (\beta-\alpha) > \theta(\beta-\alpha)$.

Case 2: $\alpha > \beta$**.** We have

$$\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)} \ge 0+\frac{\theta(\alpha-\beta)}{\theta(1-\theta)} = \frac{\alpha-\beta}{1-\theta} > |\alpha-\beta|$$

and

$$\begin{aligned}\frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{\theta(\alpha-\beta)+1-\alpha}{\theta(1-\theta)} &= \frac{(\alpha-\beta)^2}{\theta(\alpha-\beta)+\beta}+\frac{\alpha-\beta}{1-\theta}+\frac{1-\alpha}{\theta(1-\theta)}\\&\le \frac{\alpha-\beta}{\theta}+\frac{\alpha-\beta}{1-\theta}+\frac{1-\alpha}{\theta(1-\theta)}\\&= \frac{1-\beta}{\theta(1-\theta)},\end{aligned}$$

where the inequality uses that $\theta(\alpha-\beta)+\beta > \theta(\alpha-\beta) > 0$.

The lemma immediately follows. □

The final lemma in this section controls the density ratio for the initial state $\mathfrak{s}$.

Lemma 6. *For any $I, I' \in \mathcal{I}_\theta$, we have*

$$\mathbb{E}_{s' \sim P_0(\cdot \mid \mathfrak{s}, \mathfrak{a})}\left[\frac{P_I(s' \mid \mathfrak{s}, \mathfrak{a}) P_{I'}(s' \mid \mathfrak{s}, \mathfrak{a})}{P_0^2(s' \mid \mathfrak{s}, \mathfrak{a})}\right] = \frac{|I \cap I'|}{\theta^2 S_1}.$$

Proof. Let $I, I' \in \mathcal{I}_\theta$ be given, and observe that

$$\mathbb{E}_{s' \sim P_0(\cdot \mid \mathfrak{s}, \mathfrak{a})}\left[\frac{P_I(s' \mid \mathfrak{s}, \mathfrak{a}) \times P_{I'}(s' \mid \mathfrak{s}, \mathfrak{a})}{P_0^2(s' \mid \mathfrak{s}, \mathfrak{a})}\right] = \mathbb{E}_{s' \sim \mathrm{Unif}(\mathcal{S}_1)}\left[\frac{\mathbb{1}_{\{s' \in I \cap I'\}}}{\theta^2}\right] = \frac{|I \cap I'|}{\theta^2 S_1}.$$

□

D.3 Completing the Proof

To keep notation compact, define

$$g_{\theta,\alpha,\beta}(t; n) := \left(\left(\frac{t}{\theta^2 S} - 1\right)\frac{2\phi_{\theta,\alpha,\beta} + 1}{8} + 1\right)^n.$$

For all $M \in \mathcal{M}$, $P_M(\cdot \mid s, a)$ and $P_0(\cdot \mid s, a)$ differ only when $(s, a) = (\mathfrak{s}, \mathfrak{a})$ or $(s, a) \in \mathcal{S}_1 \times \{2\}$, so—recalling the value of μ—we have

$$\begin{aligned}
&\frac{1}{|\mathcal{M}_1|^2}\sum_{M, M' \in \mathcal{M}_1}\left(\mathbb{E}_{\substack{(s,a)\sim\mu,\\ s' \sim P_0(\cdot \mid s,a)}}\left[\frac{P_M(s' \mid s, a) P_{M'}(s' \mid s, a)}{P_0^2(s' \mid s, a)}\right]\right)^n \\
&= \frac{1}{|\mathcal{M}_1|^2}\sum_{M, M' \in \mathcal{M}_1}\left(\frac{1}{4}\mathbb{E}_{\substack{s\sim\mathrm{Unif}(\mathcal{S}_1),\\ s' \sim P_0(\cdot \mid s,2)}}\left[\frac{P_M(s' \mid s, 2) P_{M'}(s' \mid s, 2)}{P_0^2(s' \mid s, 2)}\right] + \frac{1}{8}\mathbb{E}_{s' \sim P_0(\cdot \mid \mathfrak{s}, \mathfrak{a})}\left[\frac{P_M(s' \mid \mathfrak{s}, \mathfrak{a}) P_{M'}(s' \mid \mathfrak{s}, \mathfrak{a})}{P_0^2(s' \mid \mathfrak{s}, 2)}\right] + \frac{5}{8}\right)^n \\
&= \frac{1}{\binom{S_1}{\theta_1 S_1}^2}\sum_t \sum_{I, I' \in \mathcal{I}_{\theta_1} : |I \cap I'| = t}\left(\frac{1}{4}\left(\left(\frac{t}{\theta_1^2 S_1} - 1\right)\phi_{\theta_1,\alpha_1,\beta_1} + 1\right) + \frac{1}{8}\frac{t}{\theta_1^2 S_1} + \frac{5}{8}\right)^n,
\end{aligned}$$

where we have used the expressions for the density ratio from Lemmas 4 and 6. We further simplify to

$$\begin{aligned}
&= \frac{1}{\binom{S_1}{\theta_1 S_1}^2}\sum_t \sum_{I, I' \in \mathcal{I}_{\theta_1} : |I \cap I'| = t}\left(\left(\frac{t}{\theta_1^2 S_1} - 1\right)\frac{2\phi_{\theta_1,\alpha_1,\beta_1} + 1}{8} + 1\right)^n \\
&= \sum_{t=(2\theta_1 - 1)_+ S_1}^{\theta_1 S_1} \frac{\binom{\theta_1 S_1}{t}\binom{S_1 - \theta_1 S_1}{\theta_1 S_1 - t}}{\binom{S_1}{\theta S_1}}\left(\left(\frac{t}{\theta_1^2 S_1} - 1\right)\frac{2\phi_{\theta_1,\alpha_1,\beta_1} + 1}{8} + 1\right)^n \\
&= \sum_{t=(2\theta_1 - 1)_+ S_1}^{\theta_1 S_1} \frac{\binom{\theta_1 S_1}{t}\binom{S_1 - \theta_1 S_1}{\theta_1 S_1 - t}}{\binom{S_1}{\theta_1 S_1}} g_{\theta_1,\alpha_1,\beta_1}(t; n),
\end{aligned}$$

where the second equality uses Lemma 3. Applying the same calculation for $\mathcal{M}_2$, we also have that

$$\begin{aligned}
&\frac{1}{|\mathcal{M}_2|^2}\sum_{M, M' \in \mathcal{M}_2}\left(\mathbb{E}_{(s,a)\sim\mu, s' \sim P_0(\cdot \mid s,a)}\left[\frac{P_M(s' \mid s, a) P_{M'}(s' \mid s, a)}{P_0^2(s' \mid s, a)}\right]\right)^n \\
&= \sum_{t=(2\theta_2 - 1)_+ S_1}^{\theta_2 S_1} \frac{\binom{\theta_2 S_1}{t}\binom{S_1 - \theta_2 S_1}{\theta_2 S_1 - t}}{\binom{S_1}{\theta_2 S_1}} g_{\theta_2,\alpha_2,\beta_2}(t; n).
\end{aligned}$$

Therefore, to upper bound the right-hand sides of Eqs. (13) and (14), we only need to upper bound the quantity

$$\sum_{t=(2\theta - 1)_+ S_1}^{\theta S_1} \frac{\binom{\theta S_1}{t}\binom{S_1 - \theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} g_{\theta,\alpha,\beta}(t; n), \tag{16}$$

for both $(\theta, \alpha, \beta) = (\theta_1, \alpha_1, \beta_1)$ and $(\theta, \alpha, \beta) = (\theta_2, \alpha_2, \beta_2)$. To upper bound this quantity, we use the following two lemmas.

Lemma 7 (Monotonicity of $g_{\theta,\alpha,\beta}$)**.** *For any $\theta, \alpha, \beta \in (0,1)$ and any $n \in \mathbb{N}$, the function $t \mapsto g_{\theta,\alpha,\beta}(t;n)$ is non-decreasing for $t \in [(2\theta-1)_+ S_1, \theta S_1]$.*

Proof. By Lemma 4, we know that $\left(\frac{t}{\theta^2 S_1} - 1\right)\phi_{\theta,\alpha,\beta} + 1 \geq 0$ for all $t \in [(2\theta-1)_+ S_1, \theta S_1]$, and hence

$$\left(\frac{t}{\theta^2 S_1} - 1\right)\frac{2\phi_{\theta,\alpha,\beta}+1}{8} + 1 = \frac{1}{4}\left(\left(\frac{t}{\theta^2 S_1} - 1\right)\phi_{\theta,\alpha,\beta} + 1\right) + \frac{1}{8}\frac{t}{\theta^2 S_1} + \frac{5}{8} \geq 0$$

for all $t \in [(2\theta-1)_+ S_1, \theta S_1]$. This ensures that we are in the domain where $x \mapsto x^n$ is non-decreasing. Next, by Lemma 5, we know that $\phi_{\theta,\alpha,\beta} \geq 0$, so the coefficient on t is non-negative. It follows that $g_{\theta,\alpha,\beta}(t;n)$ is non-decreasing in $t \in [(2\theta-1)_+ S_1, \theta S_1]$. □

Lemma 8 (Hypergeometric tail bound)**.** *For any $\theta \in \{\theta_1, \theta_2\}$ and $\epsilon \in (0, \theta^2 S_1)$, we have*

$$\sum_{t \geq (\theta+\epsilon)\cdot\theta S_1} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} \leq \exp(-2\epsilon^2 \theta S_1). \tag{17}$$

Proof. Let $\mathrm{Hyper}(t; K, N, N') := \binom{K}{t}\binom{N-K}{N'-t}/\binom{N}{N'}$ denote the hypergeometric probability mass function, which corresponds to the probability that exactly t balls are blue when N' balls are sampled without replacement from a jar containing N total balls, K of which are blue (see, e.g., Chapter 2.1.4 of Rice (2006) for background on the hypergeometric distribution). We observe that the term $\frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}}$ arising in Eqs. (16) and (17) is precisely $\mathrm{Hyper}(t; \theta S_1, S_1, \theta S_1)$, which corresponds to the process in which we sample θS_1 balls without replacement from a jar with S_1 balls, θS_1 of which are blue.

We now apply a classical tail bound for hypergeometric random variables.

Lemma 9 (Hoeffding (1963))**.** *Let $X \sim \mathrm{Hyper}(K, N, N')$ and define $p = K/N$. Then for any $0 < \epsilon < pN'$, we have*

$$\mathbb{P}[X \geq (p+\epsilon)N'] \leq \exp\left(-2\epsilon^2 N'\right).$$

Instantiating this bound with $\mathrm{Hyper}(\theta S_1, S_1, \theta S_1)$ (noting that θS_1 is an integer), we have $p = \theta$ and

$$\sum_{t \geq (\theta+\epsilon)\cdot\theta S_1} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} = \mathbb{P}[X \geq (\theta+\epsilon)\cdot\theta S_1] \leq \exp(-2\epsilon^2\theta S_1).$$

□

Returning to the quantity in (16), for any $(\theta, \alpha, \beta) \in \{(\theta_1, \alpha_1, \beta), (\theta_2, \alpha_2, \beta_2)\}$ and any $\epsilon \in (0, \theta^2 S_1)$ we can split the sum and upper bound as follows:

$$\begin{aligned}
&\sum_{t=(2\theta-1)_+ S_1}^{\theta S_1} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} g_{\theta,\alpha,\beta}(t;n) \\
&\leq \sum_{t=0}^{\lfloor(\theta+\epsilon)\theta S_1\rfloor} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} g_{\theta,\alpha,\beta}(t;n) + \exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n) \\
&\leq \left(\sum_{t=0}^{\lfloor(\theta+\epsilon)\theta S_1\rfloor} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}}\right) g_{\theta,\alpha,\beta}((\theta+\epsilon)\theta S_1; n) + \exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n) \\
&\leq g_{\theta,\alpha,\beta}((\theta+\epsilon)\theta S_1; n) + \exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n),
\end{aligned} \tag{18}$$

where the first two inequalities follow from Lemmas 7 and 8 and the last uses that the sum in the penultimate line is at most 1. We further calculate

$$\begin{aligned} g_{\theta,\alpha,\beta}((\theta+\epsilon)\theta S_1; n) &= \left(\left(\frac{(\theta+\epsilon)\theta S_1}{\theta^2 S_1} - 1\right)\frac{2\phi_{\theta,\alpha,\beta}+1}{8} + 1\right)^n \\ &= \left(\frac{2\epsilon\phi_{\theta,\alpha,\beta}+\epsilon}{8\theta} + 1\right)^n \\ &\le \left(\frac{\epsilon}{4(1-\theta)\theta} + 1\right)^n, \end{aligned} \tag{19}$$

where the inequality follows from Lemma 5. Similarly, we have

$$\begin{aligned} \exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n) &= \exp(-2\epsilon^2\theta S_1)\cdot\left(\left(\frac{\theta S_1}{\theta^2 S_1} - 1\right)\frac{2\phi_{\theta,\alpha,\beta}+1}{8} + 1\right)^n \\ &\le \exp\bigl(-2\epsilon^2\theta S_1\bigr)\cdot\left(\left(\frac{1}{\theta} - 1\right)\frac{2\theta/(1-\theta)+1}{8} + 1\right)^n \\ &\le \exp\bigl(-2\epsilon^2\theta S_1\bigr)\cdot\left(1 + \frac{1}{4\theta}\right)^n \\ &= \exp\bigl(n\ln(1+1/4\theta) - 2\epsilon^2\theta S_1\bigr) \\ &\le \exp\bigl(n/(4\theta) - 2\epsilon^2\theta S_1\bigr), \end{aligned} \tag{20}$$

where the first inequality follows from Lemma 5 and the last inequality uses that $\log(1+x) \le x$.

Combining Eqs. (13), (14), (16) and (18) to (20) and instantiating the bounds for $(\theta_1, \alpha_1, \beta_1)$ and $(\theta_2, \alpha_2, \beta_2)$, we have

$$D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0) \le \inf_{\epsilon\in(0,\theta_1^2 S_1)}\left\{\left(\frac{\epsilon}{4(1-\theta_1)\theta_1} + 1\right)^n + \exp\bigl(n/(4\theta_1) - 2\epsilon^2\theta_1 S_1\bigr)\right\} - 1.$$

$$D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0) \le \inf_{\epsilon\in(0,\theta_2^2 S_1)}\left\{\left(\frac{\epsilon}{4(1-\theta_2)\theta_2} + 1\right)^n + \exp\bigl(n/(4\theta_2) - 2\epsilon^2\theta_2 S_1\bigr)\right\} - 1.$$

Let $c \in (0, 1/4)$ be an arbitrary constant. For each $i \in \{1,2\}$, we set $\epsilon = 4c\cdot\frac{(1-\theta_i)\theta_i}{n}$ (which belongs to $(0, \theta_i^2 S_1)$ because $\epsilon < \theta_i$ since $n \ge 1$ and $\theta_i S_1 \ge 1$ by assumption). Then we have

$$\left(\frac{\epsilon}{4(1-\theta_i)\theta_i} + 1\right)^n \le \left(1 + \frac{c}{n}\right)^n \le e^c \le 1 + 2c, \quad \forall i \in \{1,2\},$$

and

$$D_{\chi^2}(\mathbb{P}_n^i \,\|\, \mathbb{P}_n^0) \le 2c + \exp\left(\frac{n}{4\theta_i} - 32c^2\theta_i\frac{(1-\theta_i)^2\theta_i^2}{n^2}S_1\right), \quad \forall i \in \{1,2\}.$$

In particular, whenever $S_1 \ge \max_{i\in\{1,2\}} \frac{n^3}{64c^2\theta_i^4(1-\theta_i)^2}$, we have

$$D_{\chi^2}(\mathbb{P}_n^i \,\|\, \mathbb{P}_n^0) \le 2c + \exp(-n/(4\theta_i)), \quad \forall i \in \{1,2\}.$$

Plugging in the values $\theta_1 = 1/2$, $\theta_2 = 3/4$ and setting $c = 1/10$, we have that whenever $n \ge 5$ and $S_1 > 400n^3$,

$$D_{\chi^2}(\mathbb{P}_n^i \,\|\, \mathbb{P}_n^0) \le \frac{1}{5} + \exp(-n/3) \le \frac{2}{5}, \quad \forall i \in \{1,2\}.$$

Combining this with (12), we have that $D_{\mathrm{TV}}\bigl(\mathbb{P}_n^1, \mathbb{P}_n^2\bigr) \le \sqrt{2/5} < 3/4$, which proves the lemma.

□