Proceedings of Machine Learning Research vol 99:1–36, 2019 32nd Annual Conference on Learning Theory

Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches

Wen Sun WENSUN@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA

Nan Jiang NANJIANG@ILLINOIS.EDU
University of Illinois at Urbana-Champaign, Urbana, IL

Akshay Krishnamurthy AKSHAY@CS.UMASS.EDU
Microsoft Research, New York, NY

Alekh Agarwal ALEKHA@MICROSOFT.COM
Microsoft Research, Redmond, WA

John Langford JCL@MICROSOFT.COM
Microsoft Research, New York, NY

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study the sample complexity of model-based reinforcement learning (henceforth RL) in general contextual decision processes that require strategic exploration to find a near-optimal policy. We design new algorithms for RL with a generic model class and analyze their statistical properties. Our algorithms have sample complexity governed by a new structural parameter called the *witness rank*, which we show to be small in several settings of interest, including factored MDPs. We also show that the witness rank is never larger than the recently proposed *Bellman rank* parameter governing the sample complexity of the model-free algorithm OLIVE (Jiang et al., 2017), the only other provably sample-efficient algorithm for global exploration at this level of generality. Focusing on the special case of factored MDPs, we prove an exponential lower bound for a general class of model-free approaches, including OLIVE, which, when combined with our algorithmic results, demonstrates exponential separation between model-based and model-free RL in some rich-observation settings.

Keywords: Reinforcement Learning, exploration

1. Introduction

Reinforcement learning algorithms can be broadly categorized as model-based or model-free methods. Methods in the former family explicitly model the environment dynamics and then use planning techniques to find a near-optimal policy. In contrast, the latter family models much less, typically only an optimal policy and its value. Algorithms from both families have seen substantial empirical success, but we lack a rigorous understanding of the tradeoffs between them, making algorithm selection difficult for practitioners. This paper provides a new understanding of these tradeoffs, via a comparative analysis between model-based and model-free methods in general RL settings.

Conventional wisdom and intuition suggests that model-based methods are more sample-efficient than model-free methods, since they leverage more supervision. This argument is supported by classical control-theoretic settings like the linear quadratic regulator, where state-of-the-art model-based methods have better dimension dependence than contemporary model-free ones (Tu and Recht, 2018). On the other hand, since models typically have more degrees of freedom (e.g., parameters) and can waste effort on unimportant elements of the environment, one might suspect that model-free methods have better statistical properties. Indeed, recent work in tabular Markov Decision Processes (MDPs) suggest that there is almost no sample-efficiency gap between the two families (Jin et al., 2018). Even worse, in complex environments where function approximation and global exploration are essential, the only algorithms with sample complexity guarantees are model-free (Jiang et al., 2017). In such environments, which of these competing perspectives applies?

To answer this question, we study model-based RL in episodic *contextual decision processes* (CDPs) where high-dimensional observations are used for decision making and the learner needs to perform strategic exploration to find a near-optimal policy. For model-based algorithms, we assume access to a class $\mathcal{M}$ of models and that the true environment is representable by the class, while for model-free algorithms, we assume access to a class of value functions that realizes the optimal value function (with analogous assumptions for policy-based methods). Under such assumptions, we posit:

> *Model-based methods rely on stronger function-approximation capabilities but can be exponentially more sample efficient than their model-free counterparts.*

Our contributions provide evidence for this thesis and can be summarized as follows:

1. We show that there exist MDPs where (1) all model-free methods, given a value function class satisfying the above realizability condition, incur exponential sample complexity (in horizon); and (2) there exist model-based methods that, given a model class containing the true model, obtain polynomial sample complexity. In fact, these MDPs belong to the well-studied *factored MDPs* (Kearns and Koller, 1999), which we use as a running example throughout the paper.

2. We design a new model-based algorithm for general CDPs and show that it has sample complexity governed by a new structural parameter, the *witness rank*. We further show that many concrete settings, including tabular and low rank MDPs, reactive POMDPs, and reactive PSRs have a small witness rank. This algorithm is the first provably-efficient model-based algorithm that does not rely on tabular representations or highly structured control-theoretic settings.

3. We compare our algorithm and the witness rank with the model-free algorithm OLIVE (Jiang et al., 2017) and the Bellman rank, the only other algorithm and structural parameter at this level of generality. We show that the witness rank is never larger, and can be exponentially smaller than the Bellman rank. In particular, our algorithm has polynomial sample complexity in factored MDPs, an exponential gain over OLIVE and any other realizability-based model-free algorithm.

The caveat in our thesis is that model-based methods rely on strong realizability assumptions. In the rich environments we study, where function approximation is essential, some form of realizability is necessary (see Proposition 1 in Krishnamurthy et al. (2016)), but our model-based assumption (See Assumption 1) is strictly stronger than prior value-based analogs (Antos et al., 2008; Krishnamurthy et al., 2016). On the other hand, our results precisely quantify the tradeoffs between model-based and model-free approaches, which may guide future empirical efforts.

2. Preliminaries

We study *Contextual Decision Processes* (CDPs), a general sequential decision making setting where the agent optimizes long-term reward by learning a policy that maps from rich observations (e.g., raw-pixel images) to actions. The term CDP was proposed by Krishnamurthy et al. (2016) and extended by Jiang et al. (2017), with CDPs capturing broad classes of RL problems allowing rich observation spaces including (Partially Observable) MDPs and Predictive State Representations. Please see the above references for further background.

Notation. We use $[N] \triangleq \{1, \ldots, N\}$. For a finite set S, $\Delta(S)$ is the set of distributions over S, and $U(S)$ is the uniform distribution over S. For a function $f : S \to \mathbb{R}$, $\|f\|_\infty$ denotes $\sup_{s\in S} |f(s)|$.

2.1. Basic Definitions

Let $H \in \mathbb{N}$ denote a time horizon and let $\mathcal{X}$ be a large context space of unbounded size, partitioned into subsets $\mathcal{X}_1, \ldots, \mathcal{X}_{H+1}$. A finite horizon episodic CDP is a tuple $(\mathcal{X}, \mathcal{A}, R, P)$ consisting of a (partitioned) context space $\mathcal{X}$, an action space $\mathcal{A}$, a transition operator $P : \{\perp\} \cup (\mathcal{X} \times \mathcal{A}) \to \Delta(\mathcal{X})$, and a reward function $R : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{R})$ with $\mathcal{R} \subseteq [0, 1]$.[1] We assume a *layered Markovian* structure, so that for any $h \in [H]$, $x_h \in \mathcal{X}_h$ and $a \in \mathcal{A}$, the future context and the reward distributions are characterized by x_h, a and moreover $P_{x_h,a} \triangleq P(x_h, a) \in \Delta(\mathcal{X}_{h+1})$. We use $P_0 \triangleq P(\perp) \in \Delta(\mathcal{X}_1)$ to denote the initial context distribution, and we assume $|\mathcal{A}| = K$ throughout.[2] Note that the layering of contexts allows us to implicitly model the level h as part of the context.

A policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ maps each context to a distribution over actions. By executing this policy in the CDP for $h - 1$ steps, we naturally induce a distribution over $\mathcal{X}_h$, and we use $\mathbb{E}_{x_h\sim\pi}[\cdot]$ to denote the expectation with respect to this distribution. A policy π has associated value and action-value functions $V^\pi : \mathcal{X} \to \mathbb{R}^+$ and $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$, defined as

$$\forall h \in [H], x \in \mathcal{X}_h : \; V^\pi(x) \triangleq \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=h}^{H} r_t \,\middle|\, x_h = x\right], \; Q^\pi(x, a) \triangleq \mathop{\mathbb{E}}_{r\sim R(x,a)}[r] + \mathop{\mathbb{E}}_{x'\sim P_{x,a}}\left[V^\pi(x')\right],$$

Here, the expectation is over randomness in the environment and the policy, with actions sampled by π. Note that there is no need to index V and Q by the level h since it is encoded in the context. The value of a policy π is $v^\pi \triangleq \mathbb{E}_{x_1\sim P_0}[V^\pi(x_1)]$ and the goal is to find a policy π that maximizes v^π.

For regularity, we assume that almost surely $\sum_{h=1}^{H} r_h \leq 1$.

Running Example *As a running example, we consider* factored MDPs *(Kearns and Koller, 1999). Let $d \in \mathbb{N}$ and let $\mathcal{O}$ be a small finite set. Define the context space $\mathcal{X} \triangleq [H] \times \mathcal{O}^d$, with the natural partition by time. For a state $x \in \mathcal{X}$ we use $x[i]$ for $i \in [d]$ to denote the value of x on the i^{th} state variable (ignoring the time step h), and similar notation for a subset of state variables. For each state variable $i \in [d]$, the parents of i, $\mathrm{pa}_i \subseteq [d]$ are the subset of state variables that directly influence i. In factored MDPs, the transition dynamics P factorize according to the parent relationships:*

$$\forall h, x \in \mathcal{X}_h, x' \in \mathcal{X}_{h+1}, a \in \mathcal{A}, \;\; P(x' \mid x, a) = \prod_{i=1}^{d} P^{(i)}[x'[i] \mid x[\mathrm{pa}_i], a, h] \tag{1}$$

for conditional distributions $\{P^{(i)}\}_{i=1}^{d}$ of the appropriate dimensions. Note that we always condition on the time point h to allow for non-stationary transitions. This transition operator has $L \triangleq \sum_{i=1}^{d} HK \cdot |\mathcal{O}|^{1+|\mathrm{pa}_i|}$ parameters, which can be much smaller than $dHK|\mathcal{O}|^{1+d}$ for an unfactorized process on $|\mathcal{O}|^d$ states.[3] *When $|\mathrm{pa}_i|$ is small for all i, we can expect algorithms with low sample*

1. We assume Markov transitions w.l.o.g., since context may encode history.
2. Partitioning the context space by time allows us to capture more general time-dependent dynamics, reward, and policy.
3. Actually the full unfactored process has $HK|\mathcal{O}|^{2d}$ parameters. Here we are assuming that the state variables are conditionally independent given the previous state and action.

complexity. Indeed Kearns and Koller (1999) show that factored MDPs can be PAC learned with poly$(H, K, L, \epsilon, \log(1/\delta))$ *samples in the average and discounted reward settings. For more recent development in this line of research, we refer the readers to Diuk et al. (2009); Nguyen et al. (2013); Osband and Van Roy (2014b); Guo and Brunskill (2017) and the references therein.*

2.2. Model Class

Since we are interested in general CDPs with large state spaces (i.e., non-tabular setting), we equip model-based algorithms with a model class $\mathcal{M}$, where all models in $\mathcal{M}$ share the same $\mathcal{X}$ and $\mathcal{A}$ but can differ in reward function R and transition operator P. The environment reward and dynamics are called the *true model* and denoted $M^\star \triangleq (R^\star, P^\star)$. For a model $M \in \mathcal{M}$, π_M, V_M, Q_M, and v_M are the optimal policy, value function, action-value function, and value *in the model* M, respectively. These objects are purely functions of M and do not depend on the environment. For the true model $M^\star$, these quantities are denoted $\pi^\star, V^\star, Q^\star, v^\star$, suppressing subscripts. For $M \triangleq (R, P)$, we denote $(r, x') \sim M_{x,a}$ as sampling a reward and next context from M: $r \sim R(x, a), x' \sim P_{x,a}$. We use $x_h \sim \pi$ to denote a state sampled by executing π in the true environment $M^\star$ for $h - 1$ steps.

We use $\texttt{OP}$ (for Optimal Planning) to represent the operator that maps a model M to its optimal Q function and its optimal policy, that is $\texttt{OP}(M) \triangleq (Q_M, \pi_M)$. We denote $\texttt{OP}(\mathcal{M}) \triangleq \{Q, \pi : \exists M \in \mathcal{M}\ s.t.\ \texttt{OP}(M) = (Q, \pi)\}$ as the set of optimal Q functions and optimal policies derived from the class $\mathcal{M}$.[4] Throughout the paper, when we compare model-based and model-free methods, we use $\mathcal{M}$ as input for the former and $\texttt{OP}(\mathcal{M})$ for the latter.

We assume the model class has *finite (but exponentially large) cardinality and is realizable*.

Assumption 1 (Realizability of $\mathcal{M}$) *We assume the model class $\mathcal{M}$ contains the true model $M^\star$.*

The finiteness assumption is made only to simplify presentation and can be relaxed using standard techniques; see Theorem 8 for a result with infinite model classes. While realizability can also be relaxed (as in Jiang et al. (2017)), it is impossible to avoid it altogether (that is, to compete with $\texttt{OP}(\mathcal{M})$ for arbitrary $\mathcal{M}$) due to exponential lower bounds (Krishnamurthy et al., 2016).

Running Example *For factored MDPs, it is standard to assume the factorization, formally* pa_i *for all* $i \in [d]$*, and the reward function are known (Kearns and Koller, 1999). Thus the natural model class* $\mathcal{M}$ *is just the set of all dynamics of the form (1), which obey the factorization, with shared reward function. While this class is infinite, our techniques apply as shown in the proof of Theorem 8.*

3. Why Model-based RL?

This section contains our first main result, that model-based methods can be *exponentially* more sample-efficient than model-free ones. To our knowledge, this is the first result of this form.

To show such separation, we must prove a lower bound against all model-free methods, and, to do so, we first formally define this class of algorithms. Strehl et al. (2006) define model-free algorithms to be those with $o(|\mathcal{X}|^2|\mathcal{A}|)$ space, but this definition is specialized to the tabular setting and provides little insight for algorithms employing function approximation. In contrast, our definition is information-theoretic: Intuitively, a model-free algorithm does not operate on the context x directly, but rather through the evaluations of a state-action function class $\mathcal{G}$. Formally:

4. As π_M is determined by Q_M, i.e., $\pi_M(x) = \arg\max_a Q_M(x, a)$, we sometimes overload notation and use $\mathcal{Q} = \texttt{OP}(\mathcal{M})$ to represent the set of optimal Q functions derived from $\mathcal{M}$.

Definition 1 (Model-free algorithm) *Given a (finite) function class $\mathcal{G} : (\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$, define the* $\mathcal{G}$-profile $\Phi_{\mathcal{G}} : \mathcal{X} \to \mathbb{R}^{|\mathcal{G}|\times|\mathcal{A}|}$ *by $\Phi_{\mathcal{G}}(x) := [g(x, a)]_{g\in\mathcal{G},a\in\mathcal{A}}$. An algorithm is* model-free *using $\mathcal{G}$ if it accesses x exclusively through $\Phi_{\mathcal{G}}(x)$ for all $x \in \mathcal{X}$ during its entire execution.*

In this definition, $\mathcal{G}$ could be a class of Q functions, a class of policies, or even the union of such classes. As such, it captures both value-function-based algorithms like OLIVE, optimistic Q-learning (Jin et al., 2018), and Delayed Q-learning (Strehl et al., 2006) as well as direct policy search algorithms like policy gradient methods (See Appendix D for a details).[5] In Appendix D, we show that when $\mathcal{G}$ consists of all Q-functions as in the tabular setting, the underlying context/state can be recovered from the $\mathcal{G}$-profile, so Definition 1 introduces no restriction whatsoever. However, beyond tabular settings, the $\mathcal{G}$-profile can obfuscate the context from the agent and may even introduce partial observability. This can lead to a significant loss of information, which can have a dramatic effect on the sample complexity. Such information loss is formalized in the following theorem.

Theorem 2 *Fix $\delta, \epsilon \in (0, 1]$. There exists a family $\mathcal{M}$ of CDPs with horizon H, all with the same reward function, and satisfying $|\mathcal{M}| \le 2^H$,such that*

1. *For any CDP in the family, with probability at least $1 - \delta$, a model based algorithm using $\mathcal{M}$ as the model class (Algorithm 3, Appendix E) outputs $\hat{\pi}$ satisfying $v^{\hat{\pi}} \ge v^\star - \epsilon$ using at most poly$(H, 1/\epsilon, \log(1/\delta))$ trajectories.*
2. *With $\mathcal{G} = \mathtt{OP}(\mathcal{M})$, any model-free algorithm using $o(2^H)$ trajectories outputs a policy $\hat{\pi}$ with $v^{\hat{\pi}} < v^\star - 1/2$ with probability at least $1/3$ on some CDP in the family.*

See Appendix C.2 for the proof. Informally, the result shows that there are CDPs where model-based methods can be exponentially more sample-efficient than any model-free method, when given access to a $\mathcal{G}$ satisfying $Q^\star \in \mathcal{G}, \pi^\star \in \mathcal{G}$. As concrete instances of such methods, the lower bound applies to several recent value-based algorithms for CDPs (Krishnamurthy et al., 2016; Jiang et al., 2017; Dann et al., 2018) as well as any future algorithms developed assuming just realizable optimal value functions or optimal policies. On the other hand, it leaves room for sample efficient model-free techniques that require stronger representation conditions on $\mathcal{G}$. To our knowledge, this is the first information-theoretic separation result for any broad class of model-based/model-free algorithms. Indeed, even the definition of model-free methods is new here.[6]

Given this result, it might seem that model-based methods should always be preferred over model-free ones. However, it is worth also comparing the assumptions required to enable the two paradigms. Since $M^\star \in \mathcal{M}$ for each CDP in the family, we also have $Q^\star \in \mathtt{OP}(\mathcal{M})$. This latter *value-function realizability* assumption is standard in model-free RL with function approximation (Antos et al., 2008; Krishnamurthy et al., 2016), but our model-based analog in Assumption 1 can be substantially stronger. As such, model-based methods operating with realizability typically require more powerful function approximation. Further, while we view setting $\mathcal{G} = \mathtt{OP}(\mathcal{M})$ as the most natural choice for the purposes of comparison, using a more expressive $\mathcal{G}$ may reveal state information and circumvent the lower bound (as we show in Appendix C.3). Thus, while Theorem 2 formalizes an argument in favor of model-based methods, realizability considerations and choice of $\mathcal{G}$ provide important caveats.

5. There are model-free algorithms that elude our definition (for example, ones that approximate the state-action distributions (Chen et al., 2018; Liu et al., 2018)), although these algorithms do not address the exploration setting.
6. Sutton and Barto (2018, Section 11.6) have a closely related definition (where the learner can only observe state features), but the definition is specialized to linear function approximation and is subsumed by ours.

Running Example *The construction in the proof of Theorem 2 is a simple factored MDP with* $d = H$, $|\mathcal{O}| = 4$, $|pa_i| = 1$ *for all* i*, and deterministic dynamics. As we see, our algorithm has polynomial sample complexity in all factored MDPs (and a broad class of other environments).*

The construction implies that model-free methods cannot succeed in factored MDPs. To our knowledge, no information theoretic lower bounds for factored MDPs exist, but the result does agree with known computational and representational barriers, namely (a) that planning is NP-hard (Mundhenk et al., 2000), (b) that $Q^\star$ *and* $\pi^\star$ *may not factorize (Guestrin et al., 2003), and (c) that* $\pi^\star$ *cannot be represented by a polynomially sized circuit (Allender et al., 2003). Our result provides a new form of hardness, namely statistical complexity, for model-free RL in factored MDPs.*

4. Witnessed Model Misfit

In this section we introduce *witnessed model misfit*, a measure of model error, which we later use to eliminate candidate models in our algorithm.

To verify the validity of a candidate model, a natural idea is to compare the samples from the environment with synthetic samples generated from a model M. To formalize this comparison approach, we use Integral Probability Metrics (IPM) (Müller, 1997): for two probability distributions $P_1, P_2 \in \Delta(\mathcal{Z})$ over $z \in \mathcal{Z}$ and a function class $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$ that is symmetric (i.e. if $f \in \mathcal{F}$ then $-f \in \mathcal{F}$ also holds), the IPM with respect $\mathcal{F}$ is: $\sup_{f\in\mathcal{F}} \mathbb{E}_{z\sim P_1}[f(z)] - \mathbb{E}_{z\sim P_2}[f(z)]$. We use IPMs to define witnessed model misfit.

Definition 3 (Witnessed Model Misfit) *For a class* $\mathcal{F} : \mathcal{X}\times\mathcal{A}\times\mathcal{R}\times\mathcal{X} \to \mathbb{R}$*, models* $M, M' \in \mathcal{M}$ *and a time step* $h \in [H]$*, the Witnessed Model Misfit of* M' *witnessed by* M *at level* h *is:*

$$\mathcal{W}(M, M', h; \mathcal{F}) \triangleq \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{\substack{x_h\sim\pi_M \\ a_h\sim\pi_{M'}}} \left[\mathop{\mathbb{E}}_{(r,x')\sim M'_h} [f(x_h, a_h, r, x')] - \mathop{\mathbb{E}}_{(r,x')\sim M^\star_h} [f(x_h, a_h, r, x')] \right], \quad (2)$$

where for a model $M = (R, P)$*,* $(r, x') \sim M_h$ *is shorthand for* $r \sim R_{x_h,a_h}, x' \sim P_{x_h,a_h}$*.*

$\mathcal{W}(M, M', h; \mathcal{F})$ is the IPM between two distributions over $\mathcal{X}\times\mathcal{A}\times\mathcal{R}\times\mathcal{X}$ with the same marginal over $\mathcal{X}\times\mathcal{A}$ but two different conditionals over (r, x'), according to M' and the true model $M^\star$, respectively. The marginal over $\mathcal{X}\times\mathcal{A}$ is the distribution over context-action pairs when π_M, the optimal policy of another candidate model M, is executed in the true environment. We call this witnessed model misfit since M' might successfully masquerade as $M^\star$ unless we find the right context distribution to witness its discrepancy. Below we illustrate the definition with some examples.

Example *[Total Variation] When* $\mathcal{F} = \{f : \|f\|_\infty \le 1\}$*, the witnessed model misfit becomes*

$$\mathcal{W}(M, M', h; \mathcal{F}) = \mathbb{E}\left[\left\| R'_{x_h,a_h} \circ P'_{x_h,a_h} - R^\star_{x_h,a_h} \circ P^\star_{x_h,a_h} \right\|_{TV} | \, x_h \sim \pi_M, a_h \sim \pi_{M'} \right], \quad (3)$$

where $R_{x,a} \circ P_{x,a}$ *is the distribution over* $\mathcal{R}\times\mathcal{X}$ *with* $r \sim R_{x,a}, x' \sim P_{x,a}$ *independently. This is just the total variation distance*[7] *between* $R' \circ P'$ *and* $R^\star \circ P^\star$*, averaged over context-action pairs* $x \sim \pi_M, a \sim \pi_{M'}(\cdot|x)$ *sampled from the true environment.*

Example *[Exponential Family] Suppose the models* $M \triangleq (R, P)$ *are from a conditional exponential family: conditioned on* $(x, a) \in \mathcal{X}\times\mathcal{A}$*, we have* $R_{x,a} \circ P_{x,a} \triangleq \exp(\langle\theta_{x,a}, \mathrm{T}(r, x')\rangle)/Z(\theta_{x,a})$ *for parameters* $\theta_{x,a} \in \Theta \triangleq \{\theta : \|\theta\| \le 1\} \subset \mathbb{R}^m$ *with partition function* $Z(\theta_{x,a})$ *and sufficient statistics*

7. We use $\|P_1 - P_2\|_{\mathrm{TV}} = \sum_{x\in\mathcal{X}} |P_1(x) - P_2(x)|$, differing from the standard definition of $\|\cdot\|_{\mathrm{TV}}$ by a factor of 2.

$\mathrm{T}: \mathcal{R} \times \mathcal{X} \to \mathbb{R}^m$. *With* $\mathcal{V} = \{\mathcal{X} \times \mathcal{A} \to \Theta\}$, *we design* $\mathcal{F} = \{(x, a, r, x') \mapsto \langle v(x, a), \mathrm{T}(r, x')\rangle : v \in \mathcal{V}\}$. *In this setting, witnessed model misfit is*

$$\mathcal{W}(M, M', h; \mathcal{F}) = \mathop{\mathbb{E}}_{x_h \sim \pi_M, a_h \sim \pi_{M'}} \left[\left\| \mathop{\mathbb{E}}_{(r,x') \sim M'_h} [\mathrm{T}(r, x')] - \mathop{\mathbb{E}}_{(r,x') \sim M^\star_h} [\mathrm{T}(r, x')] \right\|_\star \right],$$

with $\|x\|_\star \triangleq \sup\{\langle x, \theta\rangle | \|\theta\| \le 1\}$. *Here, we measure distance, in the dual norm, between the expected sufficient statistics of* (r, x') *sampled from* M' *and the true model* $M^\star$. *See Appendix G.*
Example *[MMD] When* $\mathcal{F}$ *is a unit ball in an RKHS, we obtain MMD (Gretton et al., 2012).*

Witnessed model misfit is also closely related to the *average Bellman error*, introduced by Jiang et al. (2017). Given Q functions, Q and Q', the average Bellman error at time step h is:

$$\mathcal{E}_B(Q, Q', h) \triangleq \mathbb{E}\left[Q'(x_h, a_h) - r_h - Q'(x_{h+1}, a_{h+1}) \,\middle|\, x_h \sim \pi_Q, a_{h:h+1} \sim \pi_{Q'}\right] \quad (4)$$

where π_Q is the greedy policy associated with Q, i.e., $\pi_Q(a|x) \triangleq \mathbf{1}\{a = \mathrm{argmax}_{a'} Q(x, a')\}$, and the random trajectories (w.r.t. which we take expectation) are generated in the true environment $M^\star$.

When the Q functions are derived from a model class, meaning that $\mathcal{Q} = \mathtt{OP}(\mathcal{M})$, we can extend the definition to any pair of models $M, M' \in \mathcal{M}$, using Q_M and $Q_{M'}$. In such cases, the average Bellman error for M, M' is just the model misfit witnessed by the function $f_{M'}(x, a, r, x') = r + V_{M'}(x')$. We conclude this section with an assumption about the class $\mathcal{F}$.

Assumption 2 (Bellman domination using $\mathcal{F}$) $\mathcal{F}$ *is symmetric, finite in size,*[8] $\forall f \in \mathcal{F} : \|f\|_\infty \le 2$, *and the witnessed model misfit (2) satisfies* $\forall M, M' \in \mathcal{M} :\ \mathcal{W}(M, M', h; \mathcal{F}) \ge \mathcal{E}_B(Q_M, Q_{M'}, h)$.

As discussed above, one easy way to satisfy this assumption is to ensure that the special functions $r + V_M(x')$ are contained in $\mathcal{F}$ for all $M \in \mathcal{M}$, but this is not the only way as we will see in Section 6.[9] The Bellman domination condition in Assumption 2 plays an important role in the algorithm we present next, as it allows us to detect the suboptimality of a model in terms of the value attained by its optimal policy in the actual MDP.

5. A Model-based Algorithm

In this section, we present our main algorithm and sample complexity results. We start by describing the algorithm. Then, working towards a statistical analysis, we introduce the *witness rank*, a new structural complexity measure. We end this section with the main sample complexity bounds.

5.1. Algorithm

Since we do not have access to $M^\star$, we must estimate the witnessed model misfit from samples. Since $\mathcal{F}$ will always be clear from the context, we drop it from the arguments to the model misfit for succinctness. Given a dataset $\mathcal{D} \triangleq \{(x_h^{(n)}, a_h^{(n)}, r_h^{(n)}, x_{h+1}^{(n)})\}_{n=1}^N$ with

$$x_h^{(n)} \sim \pi_M,\ a_h^{(n)} \sim U(\mathcal{A}),\ (r_h^{(n)}, x_{h+1}^{(n)}) \sim M^\star_h,$$

denote the importance weight $\rho^{(n)} \triangleq K\pi_{M'}(a_h^{(n)}|x_h^{(n)})$. We simply use the empirical model misfit:

$$\widehat{\mathcal{W}}(M, M', h) \triangleq \max_{f \in \mathcal{F}} \sum_{n=1}^N \frac{\rho^{(n)}}{N} \mathop{\mathbb{E}}_{(r,x') \sim M'_h} \left[f(x_h^{(n)}, a_h^{(n)}, r, x') - f(x_h^{(n)}, a_h^{(n)}, r_h^{(n)}, x_{h+1}^{(n)}) \right]. \quad (5)$$

8. As before, our results apply whenever $\mathcal{F}$ has bounded statistical complexity. We describe a more complicated algorithm with no dependence on the complexity of $\mathcal{F}$ in the appendix.

9. We allow the ℓ_∞ bound of 2 to accommodate these functions whose range can be 2 under our assumptions.

Algorithm 1 Inputs: $(\mathcal{M}, \mathcal{F}, n, n_e, \epsilon, \delta, \phi)$

1: $\mathcal{M}_0 = \mathcal{M}$
2: **for** $t = 1, 2, \ldots$ **do**
3: Choose model optimistically: $M^t = \mathrm{argmax}_{M \in \mathcal{M}_{t-1}} v_M$, set $\pi^t = \pi_{M^t}$
4: Execute π^t to collect n_e trajectories $\{(x_h^i, a_h^i, r_h^i)_{h=1}^H\}_{i=1}^{n_e}$ and set $\hat{v}^{\pi^t} = \frac{1}{n_e}\sum_{i=1}^{n_e}\left(\sum_{h=1}^H r_h^i\right)$
5: **if** $|\hat{v}^{\pi^t} - v_{M^t}| \leq \epsilon/2$ **then** Terminate and output π^t **end if**
6: Find h_t such that $\widehat{\mathcal{E}}_B(M^t, M^t, h_t) \geq \frac{\epsilon}{4H}$ (See (6))
7: Collect trajectories $\{(x_h^{(i)}, a_h^{(i)}, r_h^{(i)})_{h=1}^H\}_{i=1}^n$ where $a_h^{(i)} \sim \pi^t$ for $h \neq h_t$ and $a_{h_t}^{(i)} \sim U(\mathcal{A})$
8: **for** $M' \in \mathcal{M}_{t-1}$ **do** Compute $\widehat{\mathcal{W}}(M^t, M', h_t)$ (See (5)) **end for**
9: Set $\mathcal{M}_t = \{M \in \mathcal{M}_{t-1} : \widehat{\mathcal{W}}(M^t, M, h_t) \leq \phi\}$
10: **end for**

Here the importance weight $\rho^{(n)}$ accounts for distribution mismatch, since we are sampling from $U(\mathcal{A})$ instead of $\pi_{M'}$. Via standard uniform convergence arguments (in Appendix A) we show that $\widehat{\mathcal{W}}(M, M', h)$ provides a high-quality estimate of $\mathcal{W}(M, M', h)$ under Assumption 2.

We also require an estimator for the average Bellman error $\mathcal{E}_B(M, M, h)$. Given a data set $\{(x_n^{(h)}, a_h^{(n)}, r_h^{(i)}, x_{h+1}^{(n)})\}_{n=1}^N$ where $x_h^{(n)} \sim \pi_M, a_h^{(n)} \sim \pi_M$, and $(r_h^{(n)}, x_{h+1}^{(n)}) \sim M_h^\star$, we form an unbiased estimate of $\mathcal{E}_B(M, M, h)$ as

$$\widehat{\mathcal{E}}_B(M, M, h) \triangleq \frac{1}{N}\sum_{n=1}^N \left[Q_M(x_h^{(n)}, a_h^{(n)}) - \left[r_h^{(n)} + V_M(x_{h+1}^{(n)})\right]\right]. \tag{6}$$

The pseudocode is displayed in Algorithm 1. The algorithm is round-based, maintaining a version space of models and eliminating a model from the version space when the discrepancy between the model and the ground truth $M^\star$ is witnessed. The witness distributions are selected using a form of *optimism*: at each round, we select, from all surviving models, the one with the highest predicted value, and use the associated policy for data collection. If the policy achieves a high value in the environment, we simply return it. Otherwise we estimate the witnessed model misfit on the context distributions induced by the policy, and we shrink the version space by eliminating all incorrect models. Then we proceed to the next iteration.

Intuitively, using a simulation lemma analogous to Lemma 1 of Jiang et al. (2017), if M^t is the optimistic model at round t and we do not terminate, then there must exist a time step h_t (line 6) where the average Bellman error is large. Using Assumption 2, this also implies that the witness model misfit for M^t witnessed by M^t itself must be large. Thus, if t is a non-terminal round, we ensure that M^t and potentially many other models are eliminated.

The algorithm is similar to OLIVE, which uses average Bellman error instead of witnessed model misfit to shrink the version space. However, by appealing to Assumption 2, witness model misfit provides a more aggressive elimination criterion, since a large average Bellman error on a distribution immediately implies a large witnessed model misfit on the same distribution, but the converse does not necessarily hold. Since the algorithm uses an aggressive elimination rule, it often requires fewer iterations than OLIVE, as discussed below.

Computational considerations. In this work, we focus on the sample complexity of model-based RL, and Algorithm 1, as stated, admits no obvious efficient implementation. The main bottleneck, for efficiency, is the optimistic computation of the next model in line 3 where we perform a constrained optimization, restricted to the class of all models not eliminated so far. The objective in this problem encapsulates a planning oracle to map models from our class to their values, and the constraints involve enforcing small values of witness model misfit on the prior context distributions. While the witness model misfit is linear in the transition dynamics, finding an optimistic value function induces bilinear, non-convex constraints even in a tabular setting. This resembles known computational difficulties with OLIVE, but we note that the recent hardness result of Dann et al. (2018) for OLIVE does not apply to Algorithm 1, leaving the possibility of an efficient implementation open.

5.2. A structural complexity measure

So far, we have imposed realizability and expressivity assumptions (Assumption 1 and Assumption 2) on $\mathcal{M}$ and $\mathcal{F}$. Unfortunately, these alone do not enable tractable reinforcement learning with polynomial sample complexity, as verified by the following simple lower bound.

Proposition 4 *Fix $H, K \in \mathbb{N}^+$ with $K \geq 2$ and $\epsilon \in (0, \sqrt{1/8})$. There exists a family of MDPs, classes $\mathcal{M}, \mathcal{F}$ satisfying Assumption 1 and Assumption 2 for all MDPs in the family with $|\mathcal{M}| = |\mathcal{F}| = K^{H-1}$, and a constant $c > 0$, such that the following holds: For any algorithm that takes $\mathcal{M}$, $\mathcal{F}$ as inputs and uses $T \leq cK^{H-1}/\epsilon^2$ episodes, the algorithm outputs a policy $\hat{\pi}$ with $v^{\hat{\pi}} < v^\star - \epsilon$ with probability at least $1/3$ for some MDP in the family.*

The proof, provided in Appendix C.1, adapts a construction from Krishnamurthy et al. (2016) for showing that value-based realizability is insufficient for model-free algorithms. The result suggests that we must introduce further structure to obtain polynomial sample complexity guarantees. We do so with a new structural complexity measure, the *witness rank*.

For any matrix $B \in \mathbb{R}^{n \times n}$, define $\mathrm{rank}(B, \beta)$ to be the smallest integer k such that $B = UV^\top$ with $U, V \in \mathbb{R}^{n \times k}$ and for every pair of rows u_i, v_j, we have $\|u_i\|_2 \cdot \|v_j\|_2 \leq \beta$. This generalizes the standard definition of matrix rank, with a condition on the row norms of the factorization.

Definition 5 (Witness Rank) *Given a model class $\mathcal{M}$, test functions $\mathcal{F}$, and $\kappa \in (0, 1]$, for $h \in [H]$, define the set of matrices $\mathcal{N}_{\kappa,h}$ such that any matrix $A \in \mathcal{N}_{\kappa,h}$ satisfies:*

$$A \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{M}|}, \quad \kappa \mathcal{E}_B(M, M', h) \leq A(M, M') \leq \mathcal{W}(M, M', h), \forall M, M' \in \mathcal{M},$$

We define the witness rank as

$$\mathrm{W}(\kappa, \beta, \mathcal{M}, \mathcal{F}, h) \triangleq \min_{A \in \mathcal{N}_{\kappa,h}} \mathit{rank}(A, \beta).$$

We typically suppress the dependence on β because it appears only logarithmically in our sample complexity bounds. Any β that is polynomial in other parameters (K, H, and the rank itself) suffices.

To build intuition for the definition, first consider the extreme where $A(M, M') = \mathcal{W}(M, M', h)$. The rank of this matrix corresponds to the number of context distributions required to verify non-zero witnessed model misfit for all incorrect models. This follows from the fact that there are at most $\mathrm{rank}(\mathcal{W})$ linearly independent *rows* (context distributions), so any non-zero column (an incorrect model) must have a non-zero in at least one of these rows. Algorithmically, if we can find the policies π_M corresponding to these rows, we can eliminate *all* incorrect models to find $M^\star$ and hence $\pi^\star$.

At the other extreme, we have $A(M, M') = \kappa \mathcal{E}_B(M, M', h)$, the *Bellman error matrix* . The rank of this matrix, called *Bellman rank*, provides an upper bound on the witness rank by construction, and is known to be small for many natural RL settings, including tabular and low-rank MDPs, reactive POMDPs, and reactive PSRs (see Section 2 of Jiang et al. (2017) for details). The minimum over all sandwiched A matrices in the definition of the witness rank allows a smooth interpolation between these extremes in general. We further note that the choice of the class $\mathcal{F}$ defining the IPM also affects the witness model misfit and hence the witness rank. Adapting this class to the problem structure yields another useful knob to control the witness rank, as we show for the running example of factored MDPs in Section 6.

5.3. Sample complexity results

We now present a sample complexity guarantee for Algorithm 1 using the witness rank. Denote $\mathrm{W}_\kappa \triangleq \max_{h\in[H]} \mathrm{W}(\kappa, \beta, \mathcal{M}, \mathcal{F}, h)$. The main guarantee is the following theorem.

Theorem 6 *Under Assumption 1 and Assumption 2, for any* $\epsilon, \delta, \kappa \in (0, 1]$*, set* $\phi = \frac{\kappa\epsilon}{48H\sqrt{\mathrm{W}_\kappa}}$*, and denote* $T = H\mathrm{W}_\kappa \log(\beta/2\phi)/\log(5/3)$*. Run Algorithm 1 with inputs* $(\mathcal{M}, \mathcal{F}, n_e, n, \epsilon, \delta, \phi)$*, where* $n_e = \Theta\big(H^2 \log(HT/\delta)/\epsilon^2\big)$ *and* $n = \Theta\big(H^2 K\mathrm{W}_\kappa \log(T|\mathcal{M}||\mathcal{F}|/\delta)/(\kappa\epsilon)^2\big)$*. Then with probability at least* $1 - \delta$*, Algorithm 1 outputs a policy* π *such that* $v^\pi \geq v^\star - \epsilon$*. The number of trajectories collected is at most* $\tilde{O}\left(\frac{H^3\mathrm{W}_\kappa^2 K}{\kappa^2\epsilon^2}\log\left(\frac{T|\mathcal{F}||\mathcal{M}|}{\delta}\right)\right)$.

The proof is included in Appendix A. Since, as we have discussed, many popular RL models admit low Bellman rank and hence low witness rank, Theorem 6 verifies that Algorithm 1 has polynomial sample complexity in all of these settings. A noteworthy case that does not have small Bellman rank but does have small witness rank is the factored MDP, which we discuss further in Section 6.

Comparison with OLIVE. The minimum sample complexity is achieved at $\inf_\kappa \mathrm{W}_\kappa/\kappa$, which is never larger than the Bellman rank. In fact when $\kappa = 1$, the sample complexity bounds match in all terms except (a) we replace Bellman rank with witness rank, and (b) we have a dependence on model and test-function complexity $\log(|\mathcal{M}||\mathcal{F}|)$ instead of Q-function complexity $\log|\mathtt{OP}(\mathcal{M})|$. The witness rank is never larger than the Bellman rank and it can be substantially smaller, which is favorable for Algorithm 1. However, we always have $\log|\mathcal{M}| \geq \log|\mathtt{OP}(\mathcal{M})|$ and since we require realizability, the model class can be much larger than the induced Q-function class. Thus the two results are in general incomparable, but for problems where modeling the environment is not much harder than modeling the optimal Q-function (in other words $\log(|\mathcal{M}||\mathcal{F}|) \approx \log|\mathtt{OP}(\mathcal{M})|$), Algorithm 1 can be substantially more sample-efficient than OLIVE.

Adapting to unknown witness rank. In Theorem 6, the algorithm needs to know the value of κ and W_κ, as they are used to determine ϕ and n. In Appendix F, we show that a standard doubling trick can adapt to unknown κ and W_κ. The sample complexity for this adaptation is given by $\tilde{O}(H^3\mathrm{W}_{\kappa^\star}^2 K/((\kappa^\star\epsilon)^2)\log(|\mathcal{M}||\mathcal{F}|/\delta))$, where $\kappa^\star \triangleq \arg\min_{\kappa\in(0,1]} \mathrm{W}_\kappa/\kappa$ minimizes the bound in Theorem 6. A similar technique was used to adapt OLIVE to handle unknown Bellman rank.

Extension to infinite $\mathcal{M}$. Theorem 6 as stated assumes that $\mathcal{M}$ and $\mathcal{F}$ are finite classes. It is desirable to allow rich classes $\mathcal{M}$ to have a better chance of satisfying realizability of $M^\star$ in Assumption 1. Indeed, it is possible to use standard covering arguments to handle the case of infinite $\mathcal{M}$, and we demonstrate this in the context of factored MDPs in Theorem 8.

Extension to infinite $\mathcal{F}$. While our result also extends to infinite $\mathcal{F}$ with bounded statistical complexity, it is desirable to handle even richer classes, for example, $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$ for the total variation distance, which does not admit uniform convergence. To handle such rich classes, we borrow ideas from the Scheffé tournament of Devroye and Lugosi (2012),[10] and extend the method to handle conditional distributions and IPMs induced by an arbitrary class. The analysis here covers the total-variation based witnessed model misfit defined in (3) as a special case.

Theorem 7 *Under Assumption 1 and Assumption 2, but with no restriction on size of $\mathcal{F}$,[11] there exists an algorithm such that: For any $\epsilon, \delta \in (0,1]$, with probability at least $1-\delta$ the algorithm outputs a policy π such that $v^\pi \geq v^\star - \epsilon$ with at most $\tilde{O}\left(\frac{H^3 \mathrm{W}_\kappa^2 K}{\kappa^2\epsilon^2}\log\left(\frac{T|\mathcal{M}|}{\delta}\right)\right)$ trajectories collected, where $T = H\mathrm{W}_\kappa \log(\beta/2\phi)/\log(5/3)$.*

The algorithm modifies Algorithm 1 to incorporate the Scheffé estimator instead of the direct empirical estimate for the witnessed model misfit (5). We defer the details of the algorithm and analysis to Appendix B. The main improvement over Theorem 6 is that the sample complexity here has no dependence on $\mathcal{F}$, so we may use test function classes with unbounded statistical complexity.

6. Case Study on MDPs with Factored Transitions

In this section, we study the running example of factored MDPs in detail. Recall the definition of factored transition dynamics in (1). Following Kearns and Koller (1999), we assume $R^\star$ and $\{\mathrm{pa}_i\}$ are known, and $\mathcal{M}$ is the continuous space of *all* models obeying the factored transition structure and with $R^\star$ as the reward function. For this setting, we have the following guarantee.

Theorem 8 *For MDPs with factored transitions and for any $\epsilon, \delta \in (0,1]$, with probability at least $1-\delta$ a modification of Algorithm 1 (Algorithm 3 in Appendix E) outputs a policy π with $v^\pi \geq v^\star - \epsilon$ using at most $\tilde{O}(d^2 L^3 H K^2 \log(1/\delta)/\epsilon^2)$ trajectories.*

This result should be contrasted with the $\Omega(2^H)$ lower bound from Theorem 2 that actually applies precisely to this setting, where the lower bound construction has description length L polynomial in H (see Appendix C.2 for details). Combining the two results we have demonstrated exponential separation between model-based and model-free algorithms for MDPs with factored transitions.

Comparing with Theorem 6, the main improvement is that we are working with an infinite model class of all possible factored transition operators. The linear scaling with H, which seems to be an improvement, is purely cosmetic as we have $L = \Omega(H)$ here. Theorem 8 involves a slight modification to Algorithm 1, in that it uses a slightly different notion of witnessed model misfit,

$$\mathcal{W}_F(M, M', h) = \max_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{\substack{x_h\sim\pi_M \\ a_h\sim U(\mathcal{A})}}\left[\mathop{\mathbb{E}}_{(r,x')\sim M'_h}[f(x_h,a_h,r,x')] - \mathop{\mathbb{E}}_{(r,x')\sim M^\star_h}[f(x_h,a_h,r,x')]\right]. \quad (7)$$

together with an $\mathcal{F}$ specially designed for factored MDPs (subscript of $\mathcal{W}_F$ indicates adaptation to factored MDPs). The main difference with (3) is that a_h is sampled from $U(\mathcal{A})$ rather than $\pi_{M'}$. This modification is crucial to obtain a low witness rank, since $\pi_{M'}$ is in general not guaranteed to be factored (recall the representation hardness discussed at the end of Section 3). Thanks to uniformly random actions and our choice of $\mathcal{F}$, $\mathcal{W}_F$ essentially computes the sum of the TV-distances across all factors, and the corresponding matrix naturally factorizes and yields low witness rank. On the other hand, the choice of $\pi_{M'}$ for the general case allows a direct comparison with Bellman rank and

10. The classical Scheffé tournament targets the following problem: given a set of distributions $\{P_i\}_{i=1}^K$ over $\mathcal{X}$, and a set of i.i.d samples $\{x_i\}_{i=1}^N$ from $P^\star \in \Delta(\mathcal{X})$, approximate the minimizer $\mathrm{argmin}_{i\in[K]} \|P_i - P^\star\|_{\mathrm{TV}}$.

11. In fact, Assumption 2 holds automatically if we choose $\mathcal{F} = \{f : \|f\|_\infty \leq 2\}$.

leads to better guarantees in general, so we do not use the definition (7) more generally. We defer the details of the algorithm and its analysis to Appendix E.

7. Related Work

For tabular MDPs, a number of sample-efficient RL approaches exist, mostly model-based (Kearns and Singh, 2002; Jaksch et al., 2010; Dann and Brunskill, 2015; Szita and Szepesvári, 2010; Azar et al., 2017), but some are model-free (Strehl et al., 2006; Jin et al., 2018). In contrast, our work focuses on more realistic rich-observation settings.[12] For factored MDPs, all prior sample-efficient algorithms are model-based (Kearns and Koller, 1999; Osband and Van Roy, 2014b). With rich observations, many prior works either focus on structured control settings like LQRs (Abbasi-Yadkori and Szepesvári, 2011; Dean et al., 2018) or Lipschitz-continuous MDPs (Kakade et al., 2003; Ortner and Ryabko, 2012; Pazis and Parr, 2013; Lakshmanan et al., 2015). In LQRs, Tu and Recht (2018) show a gap between model-based and a particular model-free algorithm, but not an algorithm agnostic lower bound, as we show here for factored MDPs. We expect that our algorithm or natural variants are sample-efficient in many of these specific settings.

In more abstract settings, most sample-efficient algorithms are model-free (Wen and Van Roy, 2013; Krishnamurthy et al., 2016; Jiang et al., 2017; Dann et al., 2018). Our work can be seen as a model-based analog to Jiang et al. (2017), which among the above references, studies the most general class of environments.

On the model-based side, Lattimore et al. (2013) and Osband and Van Roy (2014a) obtain sample complexity guarantees; the former makes no assumptions but the guarantee scales linearly with the model class size, and the latter makes continuity assumptions, so both results have more limited scope than ours. Ok et al. (2018) propose a complexity measure for structured RL problems, but their results are for asymptotic regret in tabular or Lipschitz MDPs.

On the empirical side, models are often used to speed up learning (see e.g., Aboaf et al., 1989; Deisenroth et al., 2011, for classical references in robotics). Such results provide empirical evidence that models can be statistically valuable, which complement our theoretical results.

Finally, two recent papers share some technical similarities to our work. Farahmand et al. (2017) also use IPMs to detect model error but their analysis is restricted to test functions that form a ball in an RKHS, and they do not address exploration issues. Xu et al. (2018) devise a model-based algorithm with function approximation, but their algorithm performs local policy improvement and cannot find a globally optimal policy in a sample-efficient manner.

8. Discussion

We study model-based RL in general contextual decision processes. We derive an algorithm for general CDPs and prove that it has sample complexity upper-bounded by a new structural notion called the witness rank, which is small in many settings of interest. Comparing model-based and model-free methods, we show that the former can be exponentially more sample efficient in some settings, but they also require stronger function-approximation capabilities, which can result in worse sample complexity in other cases. Comparing the guarantees here with those derived by Jiang et al. (2017) precisely quantifies these tradeoffs, which we hope guides future design of RL algorithms.

12. In fact, our information-theoretic definition of model-free methods (Definition 1) is uninteresting in the tabular setting.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Conference on Learning Theory*, 2011.

Eric W Aboaf, Steven M Drucker, and Christopher G Atkeson. Task-level robot learning: Juggling a tennis ball more accurately. In *IEEE International Conference on Robotics and Automation*, 1989.

Eric Allender, Sanjeev Arora, Michael Kearns, Cristopher Moore, and Alexander Russell. A note on the representational incompatibility of function approximation and factored dynamics. In *Advances in Neural Information Processing Systems*, 2003.

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.

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

Yichen Chen, Lihong Li, and Mengdi Wang. Scalable bilinear π learning using state and action features. In *International Conference on Machine Learning*, pages 833–842, 2018.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, 2018.

Marc P Deisenroth, Carl E Rasmussen, and Dieter Fox. Learning to control a low-cost manipulator using data-efficient reinforcement learning. In *Robotics: Science and Systems*, 2011.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Carlos Diuk, Lihong Li, and Bethany R Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 249–256. ACM, 2009.

Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, 2017.

Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial nets. *arXiv:1810.02030*, 2018.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Aprtificial Intelligence Research*, 2003.

Zhaohan Daniel Guo and Emma Brunskill. Sample efficient feature selection for factored MDPs. *arXiv preprint arXiv:1703.03454*, 2017.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 2017.

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

Sham Kakade, Michael J Kearns, and John Langford. Exploration in metric state spaces. In *International Conference on Machine Learning*, 2003.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In *International Joint Conference on Artificial Intelligence*, 1999.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 2002.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 2016.

Kailasam Lakshmanan, Ronald Ortner, and Daniil Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *International Conference on Machine Learning*, 2015.

Tor Lattimore, Marcus Hutter, and Peter Sunehag. The sample-complexity of general reinforcement learning. In *International Conference on Machine Learning*, 2013.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1997.

Martin Mundhenk, Judy Goldsmith, Christopher Lusena, and Eric Allender. Complexity of finite-horizon Markov decision process problems. *Journal of the ACM*, 2000.

Trung Nguyen, Zhuoru Li, Tomi Silander, and Tze Yun Leong. Online feature selection for model-based reinforcement learning. In *International Conference on Machine Learning*, pages 498–506, 2013.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems 31*, pages 8888–8896, 2018.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems*, 2012.

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

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014b.

Jason Pazis and Ronald Parr. PAC Optimal Exploration in Continuous Space Markov Decision Processes. In *AAAI*, 2013.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, 2006.

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

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *International Conference on Machine Learning*, 2010.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565*, 2018.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 1992.

Zheng Wen and Benjamin Van Roy. Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems*, 2013.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma. Algorithmic framework for model-based reinforcement learning with theoretical guarantees. *arXiv:1807.03858*, 2018.

Appendix A. Proof of Theorem 6

We first present several lemmas that are useful for proving Theorem 6.

Fact 1 *For any two models M, M', the corresponding average Bellman error can be written as*

$$\begin{aligned}&\mathcal{E}_B(M, M', h) \triangleq \mathcal{E}_B(Q_M, Q_{M'}, h)\\ =&\mathbb{E}_{x_h \sim \pi_M, a_h \sim \pi_{M'}} \left[\mathbb{E}_{(r,x') \sim M'_h} \left[r + V_{M'}(x') \right] - \mathbb{E}_{(r,x') \sim M^\star_h} \left[r + V_{M'}(x') \right] \right]. \end{aligned} \tag{8}$$

Lemma 9 (Lemma 11 of Jiang et al. (2017)) *Consider a closed and bounded set $V \subset \mathbb{R}^d$ and a vector $p \in \mathbb{R}^d$. Let B be any origin-centered enclosing ellipsoid of V. Suppose there exists $v \in V$ such that $p^\top v \geq \kappa$ and define B_+ as the minimum volume enclosing ellipsoid of $\{v \in B : |p^\top v| \leq \frac{\kappa}{3\sqrt{d}}\}$. With vol$(\cdot)$ denoting the (Lebesgue) volume, we have:*

$$\frac{vol(B_+)}{vol(B)} \leq \frac{3}{5}.$$

Recall that V_M, π_M denote the optimal value function and policy derived from model M, and that v_M denotes π_M's value in M. For any policy π, v^π denotes the policy π's value in the true environment.

Lemma 10 (Simulation Lemma) *Fix a model M. Under Assumption 2, we have*

$$v_M - v^{\pi_M} = \sum_{h=1}^{H} \mathcal{E}_B(M, M, h), \qquad \textit{and} \qquad v_M - v^{\pi_M} \le \sum_{h=1}^{H} \mathcal{W}(M, M, h).$$

Proof Start at time step $h = 1$,

$$\begin{aligned}
&\mathbb{E}_{x_1 \sim P_0}[V_M(x_1) - V^{\pi_M}(x_1)] \\
&= \mathbb{E}_{x_1 \sim P_0, a_1 \sim \pi_M}\left[\mathbb{E}_{(r,x_2)\sim M_{x_1,a_1}}\left[r + V_M(x_2)\right] - \mathbb{E}_{(r,x_2)\sim M^\star_{x_1,a_1}}\left[r + V^{\pi_M}(x_2)\right]\right] \\
&= \mathbb{E}_{x_1 \sim P_0, a_1 \sim \pi_M}\left[\mathbb{E}_{(r,x_2)\sim M_{x_1,a_1}}\left[r + V_M(x_2)\right] - \mathbb{E}_{(r,x_2)\sim M^\star_{x_1,a_1}}\left[r + V_M(x_2)\right]\right] \\
&\quad + \mathbb{E}_{x_1 \sim P_0, a_1 \sim \pi_M}\left[\mathbb{E}_{(r,x_2)\sim M^\star_{x_1,a_1}}\left[V_M(x_2)\right] - \mathbb{E}_{(r,x_2)\sim M^\star_{x_1,a_1}}\left[V^{\pi_M}(x_2)\right]\right],
\end{aligned}$$

where the first equality is based on applying Bellman's equation to V_M in M and V^{π_M} in $M^\star$. Now, by Fact 1, the first term above is exactly $\mathcal{E}_B(M, M, 1)$. The second term can be expressed as,

$$\mathbb{E}\left[V_M(x_2) - V^{\pi_M}(x_2) | x_2 \sim \pi_M\right],$$

which we can further expand by applying the same argument recursively to obtain the identity involving the average Bellman errors. For the bound involving the witness model misfit, since $V_M \in \mathcal{F}$, we simply observe that $\mathcal{E}_B(M, M, h) \le \mathcal{W}(M, M, h)$. ∎

Next, we present several concentration results.

Lemma 11 *Fix a policy π, and fix $\epsilon, \delta \in (0, 1)$. Sample $n_e = \frac{\log(2/\delta)}{(2\epsilon)^2}$ trajectories $\{(x_h^{(i)}, a_h^{(i)}, r_h^{(i)})_{h=1}^H\}_{i=1}^{n_e}$ by executing π and set $\hat{v}^\pi = \frac{1}{n_e}\sum_{i=1}^{n_e}\sum_{h=1}^{H} r_h^{(i)}$. With probability at least $1 - \delta$, we have $|\hat{v}^\pi - v^\pi| \le \epsilon$.*

The proof is a direct application of Hoeffding's inequality on the random variables $\sum_{h=1}^{H} r_h^{(i)}$.

Recall the definitions of $\widehat{\mathcal{W}}$ and $\widehat{\mathcal{E}}_B$ from (5) and (6), and the shorthand notation $(r, x') \sim M_h$, which stands for $r \sim R_{x_h,a_h}$ and $x' \sim P_{x_h,a_h}$ (with $(R, P) = M$) whenever the identities of x_h and a_h are clear from context.

Lemma 12 (Deviation Bound for $\widehat{\mathcal{E}}_M$) *Fix h and model $M \in \mathcal{M}$. Sample a dataset $\mathcal{D} = \left\{(x_h^{(i)}, a_h^{(i)}, r_h^{(i)}, x_{h+1}^{(i)})\right\}_{i=1}^{N}$ with $x_h^{(i)} \sim \pi_M, a_h^{(i)} \sim U(\mathcal{A}), (r_h^{(i)}, x_{h+1}^{(i)}) \sim M_h^\star$ of size N. Then with probability at least $1 - \delta$, we have for all $M' \in \mathcal{M}$:*

$$\left|\widehat{\mathcal{W}}(M, M', h) - \mathcal{W}(M, M', h)\right| \le \sqrt{\frac{2K\log(2|\mathcal{M}||\mathcal{F}|/\delta)}{N}} + \frac{2K\log(2|\mathcal{M}||\mathcal{F}|/\delta)}{3N}.$$

Proof

Fix $M' \in \mathcal{M}$ and $f \in \mathcal{F}$, define the random variable $z_i(M', f)$ as:

$$z_i(M', f) \triangleq K\pi_{M'}(a_h^{(i)}|x_h^{(i)}) \left(\mathop{\mathbb{E}}_{(r,x')\sim M'_h} f(x_h^{(i)}, a_h^{(i)}, r, x') - f(x_h^{(i)}, a_h^{(i)}, r_h^{(i)}, x_{h+1}^{(i)}) \right).$$

The expectation of $z_i(M', f)$ is

$$\mathbb{E}[z_i(M', f)] = \underbrace{\mathop{\mathbb{E}}_{x_h\sim\pi_M, a_h\sim\pi_{M'}} [\mathop{\mathbb{E}}_{(r,x')\sim M'_h} [f(x_h, a_h, r, x')] - \mathop{\mathbb{E}}_{(r,x')\sim M^\star_h} [f(x_h, a_h, r, x')]]}_{\triangleq d(M', M^\star, f)},$$

and it is easy to verify that $\mathrm{Var}(z_i(M', f)) \leq 4K$. Hence, we can apply Bernstein's inequality, so that with probability at least $1 - \delta$, we have

$$\left| \frac{1}{N} \sum_{i=1}^{N} z_i(M', f) - d(M', M^\star, f) \right| \leq \sqrt{\frac{2K\log(2/\delta)}{N}} + \frac{2K\log(2/\delta)}{3N}.$$

Via a union bound over $\mathcal{M}$ and $\mathcal{F}$, we have that for all pairs $M' \in \mathcal{M}$, $f \in \mathcal{F}$, with probability at least $1 - \delta$:

$$\left| \frac{1}{N} \sum_{i=1}^{N} z_i(M', f) - d(M', M^\star, f) \right| \leq \sqrt{\frac{2K\log(2|\mathcal{M}||\mathcal{F}|/\delta)}{N}} + \frac{2K\log(2|\mathcal{M}||\mathcal{F}|/\delta)}{3N}. \quad (9)$$

For fixed M', we have shown uniform convergence over $\mathcal{F}$, and this implies that the empirical and the population maxima must be similarly close, which yields the result. ■

Lemma 13 (Deviation Bound on $\widehat{\mathcal{E}}_B$) *Fix model $M \in \mathcal{M}$. Sample a dataset $\mathcal{D} = \left\{ (x_h^{(i)}, a_h^{(i)}, r_h^{(i)}, x_{h+1}^{(i)}) \right\}_{i=1}^{N}$ with $x_h^{(i)} \sim \pi_M, a_h^{(i)} \sim \pi_M, (r_h^{(i)}, x_{h+1}^{(i)}) \sim M^\star_h$ of size N. Then with probability at least $1-\delta$, for any $h \in [H]$, with probability at least $1 - \delta$, we have:*

$$\left| \mathcal{E}_B(M, M, h) - \widehat{\mathcal{E}}_B(M, M, h) \right| \leq \sqrt{\frac{\log(2H/\delta)}{2N}}.$$

The result involves a standard application of Hoeffding's inequality with a union bound over $h \in [H]$, which can also be found in Jiang et al. (2017).

Lemma 14 (Terminate or Explore) *Suppose that for any round t, $\hat{v}^{\pi^t}$ satisfies $\left| v^{\pi^t} - \hat{v}^{\pi^t} \right| \leq \epsilon/8$ and $M^\star$ is never eliminated. Then in any round t, one of the following two statements must hold:*
1. The algorithm does not terminate and there exists a $h \in [H]$ such that $\mathcal{E}_B(M^t, M^t, h) \geq \frac{3\epsilon}{8H}$;
2. The algorithm terminates and outputs a policy π^t which satisfies $v^{\pi^t} \geq v^\star - \epsilon$.

Proof Let us first consider the situation where the algorithm does not terminate, i.e., $|\hat{v}^{\pi^t} - v_{M^t}| \geq \epsilon/2$. Via Lemma 10, we must have

$$\sum_{h=1}^{H} \mathcal{E}_B(M^t, M^t, h) \geq \left| v^{\pi^t} - v_{M^t} \right| = \left| v^{\pi^t} - \hat{v}^{\pi^t} + \hat{v}^{\pi^t} - v_{M^t} \right| \geq \left| \hat{v}^{\pi^t} - v_{M^t} \right| - \left| v^{\pi^t} - \hat{v}^{\pi^t} \right| \geq 3\epsilon/8.$$

By the pigeonhole principle, there must exist $h \in [H]$, such that

$$\mathcal{E}_B(M^t, M^t, h) \geq \frac{3\epsilon}{8H},$$

so we obtain the first claim. For the second claim, if the algorithm terminates at round t, we must have $|\hat{v}^{\pi^t} - v_{M^t}| \leq \epsilon/2$. Based on the assumption that $M^\star$ is never eliminated, and M^t is the optimistic model, we may deduce

$$v^{\pi^t} \geq \hat{v}^{\pi^t} - \frac{\epsilon}{8} \geq v_{M^t} - \frac{5\epsilon}{8} \geq v^\star - \frac{5\epsilon}{8} \geq v^\star - \epsilon. \tag{10}$$

■

Recall the definition of the *witness rank* (Definition 5):

$$\mathrm{W}(\kappa, \beta, \mathcal{M}, \mathcal{F}, h) = \inf\left\{\mathrm{rank}(A) : \kappa\mathcal{E}_B\left(M, M', h\right) \leq A(M, M') \leq \mathcal{W}(M, M', h), \forall M, M' \in \mathcal{M}\right\}.$$

Let us denote $A^\star_{\kappa,h}$ as the matrix that achieves the witness rank $\mathrm{W}(\kappa, \beta, \mathcal{M}, \mathcal{F}, h)$ at time step h. Denote the factorization by $A^\star_{\kappa,h}(M, M') = \langle \zeta_h(M), \chi_h(M')\rangle$ with $\zeta_h, \chi_h \in \mathbb{R}^{\mathrm{W}(\kappa,\beta,\mathcal{M},\mathcal{F},h)}$. Finally, recall that $\beta \geq \max_{M,M',h} \|\zeta_h(M)\|_2 \|\chi_h(M')\|_2$.

Lemma 15 *Fix round t and assume that $\left|\widehat{\mathcal{E}}_B(M^t, M^t, h) - \mathcal{E}_B(M^t, M^t, h)\right| \leq \frac{\epsilon}{8H}$ for all $h \in [H]$ and $\left|v^{\pi^t} - \hat{v}^{\pi^t}\right| \leq \epsilon/8$ hold. If Algorithm 1 does not terminate, then we must have $A^\star_{\kappa,h_t}(M^t, M^t) \geq \frac{\kappa\epsilon}{8H}$.*

Proof We first verify the existence of h_t in the selection rule line 6 in Algorithm 1. From Lemma 14, we know that there exists $h \in [H]$ such that $\mathcal{E}_B(M^t, M^t, h) \geq \frac{3\epsilon}{8H}$, and for this h, we have

$$\widehat{\mathcal{E}}_B(M^t, M^t, h) \geq \frac{3\epsilon}{8H} - \frac{\epsilon}{8H} = \frac{\epsilon}{4H}. \tag{11}$$

While this h may not be the one selected in line 6, it verifies that h_t exists, and further we do know that for h_t

$$\mathcal{E}_B(M^t, M^t, h_t) \geq \frac{2\epsilon}{8H} - \frac{\epsilon}{8H} = \frac{\epsilon}{8H},$$

Now the constraints defining $A^\star_{\kappa,h_t}$ give $A^\star_{\kappa,h_t}(M^t, M^t) \geq \kappa\mathcal{E}_B(M^t, M^t, h_t)$, which proves the lemma. ■

Recall the model elimination criteria at round t: $\mathcal{M}_t = \{M \in \mathcal{M}_{t-1} : \widehat{\mathcal{W}}(M^t, M, h_t) \leq \phi\}$.

Lemma 16 *Suppose that $\left|\widehat{\mathcal{W}}(M^t, M, h_t) - \mathcal{W}(M^t, M, h_t)\right| \leq \phi$ holds for all t, h_t, and $M \in \mathcal{M}$. Then*

1. *$M^\star \in \mathcal{M}_t$, for all t.*
2. *Denote $\widetilde{\mathcal{M}}_t = \{M \in \widetilde{\mathcal{M}}_{t-1} : A^\star_{\kappa,h_t}(M^t, M) \leq 2\phi\}$ with $\widetilde{\mathcal{M}}_0 = \mathcal{M}$. We have $\mathcal{M}_t \subseteq \widetilde{\mathcal{M}}_t$ for all t.*

Observe $\widetilde{\mathcal{M}}_t$ is defined via the matrix $A^{\star}_{\kappa,h}$.

Proof Recall that we have $\mathcal{W}(M^t, M^{\star}, h_t) = 0$. Assuming $M^{\star} \in \mathcal{M}_{t-1}$ and via the assumption in the statement, for every t, we have

$$\widehat{\mathcal{W}}(M^t, M^{\star}, h_t) \leq \mathcal{W}(M^t, M^{\star}, h_t) + \phi = \phi,$$

so $M^{\star}$ will not be eliminated at round t.

For the second result, we know that $\widetilde{\mathcal{M}}_0 = \mathcal{M}$. Assume inductively that, we have $\mathcal{M}_{t-1} \subset \widetilde{\mathcal{M}}_{t-1}$, and let us prove that $\mathcal{M}_t \subset \widetilde{\mathcal{M}}_t$. Towards a contradiction, let us assume that there exists $M \in \mathcal{M}_t$ such that $M \notin \widetilde{\mathcal{M}}_t$. Since $M \in \mathcal{M}_t \subset \mathcal{M}_{t-1} \subset \widetilde{\mathcal{M}}_{t-1}$, the update rule for $\widetilde{\mathcal{M}}_t$ implies that

$$A^{\star}_{\kappa,h_t}(M^t, M) > 2\phi.$$

But, using the deviation bound and the definition of $A_{\kappa^{\star},h}$, we get

$$\widehat{\mathcal{W}}(M^t, M, h_t) \geq \mathcal{W}(M^t, M, h_t) - \phi \geq A^{\star}_{\kappa,h_t}(M^t, M) - \phi > \phi,$$

which contradicts the fact that $M \in \mathcal{M}_t$. Thus, by induction we obtain the result. ■

With our choice of $\phi = \frac{\kappa\epsilon}{48H\sqrt{\mathrm{W}_\kappa}}$, we may now quantify the number of rounds of Algorithm 1 using $\widetilde{\mathcal{M}}_t$.

Lemma 17 (Iteration complexity) *Suppose that*

$$\left|\widehat{\mathcal{W}}(M^t, M, h_t) - \mathcal{W}(M^t, M, h_t)\right| \leq \phi, \qquad \left|\widehat{\mathcal{E}}_B(M^t, M^t, h) - \mathcal{E}_B(M^t, M^t, h)\right| \leq \frac{\epsilon}{8H},$$

hold for all t, h_t, $h \in [H]$, and $M \in \mathcal{M}$, then the number of rounds of Algorithm 1 is at most $H\mathrm{W}_\kappa \log(\frac{\beta}{2\phi})/\log(5/3)$.

Proof From Lemma 15, if the algorithm does not terminate at round t, we find M^t and h_t such that

$$A^{\star}_{\kappa,h_t}(M^t, M^t) = \left\langle \zeta_{h_t}(M^t), \chi_{h_t}(M^t) \right\rangle \geq \frac{\kappa\epsilon}{8H} = 6\sqrt{\mathrm{W}_\kappa}\phi,$$

which uses the value of $\phi = \frac{\kappa\epsilon}{48H\sqrt{\mathrm{W}_\kappa}}$.

Recall the recursive definition of $\widetilde{\mathcal{M}}_t = \{M \in \widetilde{\mathcal{M}}_{t-1} : A^{\star}_{\kappa,h_t}(M^t, M) \leq 2\phi\}$ from Lemma 16. For the analysis, we maintain and update H origin-centered ellipsoids where the h^{th} ellipsoid contains the set $\{\chi_h(M) : M \in \widetilde{\mathcal{M}}_t\}$. Denote O^h_t as the origin-centered minimum volume enclosing ellipsoid (MVEE) of $\{\chi_h(M) : M \in \widetilde{\mathcal{M}}_t\}$. At round t, for $\zeta_{h_t}(M^t)$, we just proved that there exists a vector $\chi_{h_t}(M^t) \in O^{h_t}_{t-1}$ such that $\left\langle \zeta_{h_t}(M^t), \chi_{h_t}(M^t) \right\rangle \geq 6\sqrt{\mathrm{W}_\kappa}\phi$. Denote $O^{h_t}_{t-1,+}$ as the origin-centered MVEE of $\{v \in O^{h_t}_{t-1} : \left\langle \zeta_{h_t}(M^t), v \right\rangle \leq 2\phi\}$. Based on Lemma 9, and the fact that $O^{h_t}_t \subset O^{h_t}_{t-1,+}$, by the definition of $\widetilde{\mathcal{M}}_t$, we have:

$$\frac{\mathrm{vol}(O^{h_t}_t)}{\mathrm{vol}(O^{h_t}_{t-1})} \leq \frac{\mathrm{vol}(O^{h_t}_{t-1,+})}{\mathrm{vol}(O^{h_t}_{t-1})} \leq 3/5,$$

which shows that if the algorithm does not terminate, then we shrink the volume of $O_t^{h_t}$ by a constant factor.

Denote $\Phi \triangleq \sup_{M\in\mathcal{M},h} \|\zeta_h(M)\|_2$ and $\Psi \triangleq \sup_{M\in\mathcal{M},h} \|\chi_h(M)\|_2$. For O_0^h, we have $\mathrm{vol}(O_0^h) \leq c_{\mathrm{W}_\kappa} \Psi^{\mathrm{W}_\kappa}$ where c_{W_κ} is the volume of the unit Euclidean ball in W_κ-dimensions. For any t, we have

$$O_t^h \supseteq \{q \in \mathbb{R}^{\mathrm{W}_\kappa} : \max_{p:\|p\|_2\leq\Phi} \langle q, p\rangle \leq 2\phi\} = \{q \in \mathbb{R}^{\mathrm{W}_\kappa} : \|q\|_2 \leq 2\phi/\Phi\}$$

Hence, we must have that at termination, $\mathrm{vol}(O_T^h) \geq c_{\mathrm{W}_\kappa}(2\phi/\Phi)^{\mathrm{W}_\kappa}$. Using the volume of O_0^h and the lower bound of the volume of O_T^h and the fact that every round we shrink the volume of $O_t^{h_t}$ by a constant factor, we must have that for any $h \in [H]$, the number of rounds for which $h_t = h$ is at most:

$$\mathrm{W}_\kappa \log(\frac{\Phi\Psi}{2\phi})/\log(5/3). \tag{12}$$

Using the definition $\beta \geq \Phi\Psi$, this gives an iteration complexity of $H\mathrm{W}_\kappa \log\left(\frac{\beta}{2\phi}\right)/\log(5/3)$. ∎

We are now ready to prove Theorem 6. Note that we are using $A_\kappa^\star$, rather than relying on $\mathcal{E}_B$ or $\mathcal{W}$.

Proof [Proof of Theorem 6]

Below we condition on three events: (1) $\left|\widehat{\mathcal{W}}(M^t, M, h_t) - \mathcal{W}(M^t, M, h_t)\right| \leq \phi$ for all t and $M \in \mathcal{M}$, (2) $\left|\widehat{\mathcal{E}}_B(M^t, M^t, h) - \mathcal{E}_B(M^t, M^t, h)\right| \leq \frac{\epsilon}{8H}$ for all t and $h \in [H]$, and (3) $\left|v^{\pi^t} - \hat{v}^{\pi^t}\right| \leq \epsilon/8$ for all t.

Under the first and second condition, from the lemma above, we know that the algorithm must terminate in at most $T = \mathrm{W}_\kappa H \log(\beta/(2\phi))/\log(5/3)$ rounds. Once the algorithm terminates, based on Lemma 14, we know that we must have found a policy that is ϵ-optimal.

Now, we show that with our choices for n, n_e, and ϕ, the above conditions hold with probability at least $1-\delta$. Based on value of $n_e = 32\frac{H^2\log(6HT/\delta)}{\epsilon^2}$, and Lemma 11, we can verify that the third condition $|v^{\pi^t} - \hat{v}^{\pi^t}| \leq \epsilon/8$ for all $t \in [T]$ with probability $1-\delta/3$, and the condition $\left|\hat{\mathcal{E}}_B(M^t, M^t, h) - \mathcal{E}_B(M^t, M^t, h)\right| \leq \epsilon/(8H)$ holds for all $t \in [T]$ and $h \in [H]$ with probability at least $1-\delta/3$. Based on the value of $n = 18432H^2K\mathrm{W}_\kappa \log(12T|\mathcal{M}||\mathcal{F}|/\delta)/(\kappa\epsilon)^2$, the value of ϕ, and the deviation bound from Lemma 12, we can verify that the condition $\left|\widehat{\mathcal{W}}(M^t, M, h_t) - \mathcal{W}(M^t, M, h_t)\right| \leq \phi$ holds for all $t \in [T]$, $M \in \mathcal{M}$ with probability at least $1-\delta/3$. Together these ensure the algorithm terminates in T iterations. The number trajectories is at most $(n_e + n)\cdot T$, and the result follows by substitute the value of n_e, n, and T. ∎

Appendix B. Proof of Theorem 7

We are interested in generalizing Theorem 6 to accommodate a broader class of test functions $\mathcal{F}$, for example $\{f : \|f\|_\infty \leq 1\}$ that induces the total-variation distance. This class is not a Glivenko-Cantelli class, so it does not enable uniform convergence, and we cannot simply use empirical mean estimator as in (5).

Algorithm 2 Extension to $\mathcal{F}$ with Unbounded Complexity. Arguments: $(\mathcal{M}, \mathcal{F}, \epsilon, \delta, \epsilon)$

1: Compute $\tilde{\mathcal{F}}$ from $\mathcal{F}$ and $\mathcal{M}$ via (13)
2: Set $\phi = \kappa\epsilon/(48H\sqrt{\mathrm{W}_\kappa})$ and $T = H\mathrm{W}_\kappa \log(\beta/2\phi)/\log(5/3)$
3: Set $n_e = \Theta(H^2 \log(6HT/\delta)/\epsilon^2)$ and $n = \Theta(H^2 K \mathrm{W}_\kappa \log(12T|\mathcal{M}||\tilde{\mathcal{F}}|/\delta)/(\kappa^2\epsilon^2))$
4: Run Algorithm 1 with inputs $(\mathcal{M}, \tilde{\mathcal{F}}, n_e, n, \epsilon, \delta, \phi)$ and return the found policy.

The key is to define a much smaller function class $\widetilde{\mathcal{F}} \subset \mathcal{F}$ that does enjoy uniform convergence, and at the same time is expressive enough such that the witnessed model misfit w.r.t. $\widetilde{\mathcal{F}}$ is the same as that w.r.t. $\mathcal{F}$. To define $\widetilde{\mathcal{F}}$, we need one new definition. For a model M and a policy π, we use $x_h \sim (\pi, M)$ to denote that x_h is sampled by executing π *in the model* M, instead of the true environment, for h steps. With this notation, define $f_{\pi,M_1,M_2,h}$ as:

$$\operatorname*{argmax}_{f\in\mathcal{F}} \mathbb{E}\left[\mathop{\mathbb{E}}_{(r,x_{h+1})\sim M_2}[f(x_h,a_h,r,x_{h+1})] - \mathop{\mathbb{E}}_{(r,x_{h+1})\sim M_1}[f(x_h,a_h,r,x_{h+1})] \mid x_h \sim (\pi, M_1), a_h \sim \pi_{M_2}\right].$$

Note that the maximum over $\mathcal{F}$ is always attained due to the boundedness assumption on $f \in \mathcal{F}$, and hence this definition is without loss of generality. Now we define

$$\widetilde{\mathcal{F}} \triangleq \left\{\pm f_{\pi_{M_3},M_1,M_2,h} : M_1, M_2, M_3 \in \mathcal{M}, h \in [H]\right\}. \tag{13}$$

This construction is based on the Scheffé estimator, which was originally developed for density estimation in total variation (Devroye and Lugosi, 2012). As we have done here, the idea is to define a smaller function class containing just the potential maximizers. Importantly, this smaller function class is computed *independently* of the data, so there is no risk of overfitting. The main innovation here is that we extend the Scheffé estimator to conditional distributions, and also to handle arbitrary classes $\mathcal{F}$.

Lemma 18 *For any true model $M^\star \in \mathcal{M}$, policy π_M, $h \in [H]$, and target model M', we have*

$$\mathcal{W}(M, M', h; \mathcal{F}) = \mathcal{W}(M, M', h; \widetilde{\mathcal{F}}).$$

Moreover $|\widetilde{\mathcal{F}}| \le 2|\mathcal{M}|^3 H$.

Proof The bound on $|\widetilde{\mathcal{F}}|$ is immediate. For the other claim, by the realizability assumption for $\mathcal{M}$, $\widetilde{\mathcal{F}}$ contains the functions $f_{\pi_M,M^\star,M',h}$ for each (M, M', h) pair. These are precisely the test functions that maximize the witness model misfit for $\mathcal{F}$, and so the IPM induced by $\widetilde{\mathcal{F}}$ achieves exactly the same values. ■

Replacing $\widehat{\mathcal{W}}(M, M', h)$ in (5), which uses $\mathcal{F}$, to instead use $\widetilde{\mathcal{F}}$, we obtain Algorithm 2 and Theorem 7 as a corollary to Theorem 6. The key is that we have eliminated the dependence on $|\mathcal{F}|$ in the bound.

Appendix C. Lower Bounds and the Separation Result

C.1. Proof of Proposition 4

To prove Proposition 4, we need the following lower bound for best-arm identification in stochastic multi-armed bandits.

Lemma 19 (Theorem 2 from Krishnamurthy et al. (2016)) *For $K \geq 2$, $\epsilon < \sqrt{1/8}$, and any best-arm identification algorithm, there exists a multi-armed bandit problem for which the best arm $i^\star$ is ϵ better than all others, but for which the estimate $\hat{i}$ of the best arm must have $\mathbb{P}[\hat{i} \neq i^\star] \geq 1/3$ unless the number of samples collected is at least $K/(72\epsilon^2)$.*

Proof [Proof of Proposition 4] Below we explicitly give the construction of $\mathcal{M}$. Every MDP in this family shares the same reward function, and actually also shares the same transition structure for all levels $h \in [H-1]$. The models only differ in their transition at the last time step.

Fix H and $K \geq 2$. Each MDP $M^{\mathbf{a}^\star} \in \mathcal{M}$ corresponds to an action sequence $\mathbf{a}^\star = \{a_1^\star, a_2^\star, \ldots, a_{H-1}^\star\}$ where $a_i^\star \in [K]$. Thus there are K^{H-1} models. The reward function, which is shared by all models, is

$$R(x) \triangleq \mathbf{1}\{x = x^\star\} \tag{14}$$

where $x^\star$ is a special state that only appears at level H. Let x' denote another special state at level H.

For any model $M^{\mathbf{a}^\star}$, at any level $h < H-1$, the state x_h is simply the history of actions $x_h \triangleq \{a_1, a_2, \ldots a_{h-1}\}$ applied so far, and taking $a \in \mathcal{A}$ at state x_h deterministically transitions to $x_h \circ a \triangleq \{a_1, a_2, \ldots, a_{h-1}, a\}$. The transition at level $h = H-1$ is defined as follows:

$$P^{\mathbf{a}^\star}(x_H | x_{H-1}, a_{H-1}) \triangleq \begin{cases} 0.5 + \epsilon \mathbf{1}\{x_{H-1} \circ a_{H-1} = \mathbf{a}^\star\}, & x_H = x^\star \\ 0.5 - \epsilon \mathbf{1}\{x_{H-1} \circ a_{H-1} = \mathbf{a}^\star\}, & x_H = x'. \end{cases} \tag{15}$$

Thus, in each model $M^{\mathbf{a}^\star}$, each action sequence $\{a_1, a_2, \ldots, a_{H-1}\}$ can be regarded as an arm in MAB problem with K^{H-1} arms, where all the arms yield Ber(0.5) reward except for the optimal arm $\mathbf{a}^\star$ which yields Ber$(0.5+\epsilon)$ reward. In fact, this construction is information-theoretically equivalent to the construction used in the standard MAB lower bound, which appears in the proof of Lemma 19. That lower bound directly applies and since we have K^{H-1} arms here, the result follows. ■

C.2. Proof of Theorem 2

Theorem 2 has two claims: (1) There exists a family of MDPs in which Algorithm 3 achieves polynomial sample complexity, and (2) *Any* model-free algorithm will incur exponential sample complexity in this family. As we have discussed, the actual result is stronger in that the model class consists of factored MDPs under a particular structure, and our algorithm can handle any class of factored MDPs with an arbitrary (but known) structure.

The rest of this subsection is organized as follows: Appendix C.2.1 describes the family of MDPs we construct. Since the MDPs obey a factored structure, we can learn this family using our Algorithm 3 and its guarantees in Theorem 8 immediately applies, which proves the second claim. Then, the first claim is proved in Appendix C.2.2, where we leverage the definition of model-free algorithm (Definition 1) to induce information-theoretic hardness.

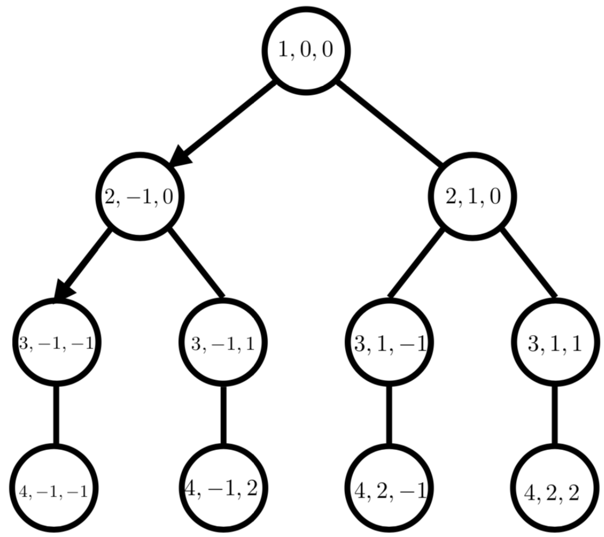

Figure 1: An example of the factored MDP construction in the proof of Theorem 2, with $d = 2$ and $H = 4$. All models are deterministic, and each model is uniquely indexed by a sequence of actions $\mathbf{p}$. (Here $\mathbf{p} = \{-1, -1\}$, as indicated by the black arrows.) The first coordinate in each state encodes the level h. Each state at level $h \leq H - 1$ encodes the sequence of actions leading to it using bits from the second to the last (padded with 0's). The last transition is designed such that the agent always lands in a state that contains "2" unless it follows path $\mathbf{p}$.

C.2.1. MODEL CLASS CONSTRUCTION AND SAMPLE EFFICIENCY OF ALGORITHM 3

Model Class Construction. We prove the claim by constructing a family of factored MDPs (recall (1)) that share the same reward function R but differ in their transition operators. The set of such transition operators is denoted as $\mathcal{P}$, and we use $P \in \mathcal{P}$ to refer to an MDP instance.

Fix $d > 2$ and set $H \triangleq d + 2$. The state variables take values in $\mathcal{O} = \{-1, 0, 1, 2\}$. The state space is $\mathcal{X} = [H] \times \mathcal{O}^d$ with the natural partition across time steps and the action space is $\mathcal{A} = \{-1, +1\}$. The initial state is fixed as $x = 1 \circ [0]^d$, where $[a]^d$ stands for a d-dimensional vector where every coordinate is a and $\circ$ denotes concatenation. Our model class contains 2^d models, each of which is uniquely indexed by an action sequence (or a *path*) of length d, $\mathbf{p} = \{p_1, \dots, p_d\}$ with $p_i \in \{-1, 1\}$. Fixing $\mathbf{p}$, we describe the transition dynamics for $P^{\mathbf{p}}$ below. All models share the same reward function, which will be described afterwards.

In $P^{\mathbf{p}}$, the parent of the i^{th} factor is itself so that each factor evolves independently. Furthermore, all transitions are deterministic, so we abuse notation and let $P_h^{\mathbf{p},i}(\cdot, \cdot)$ denote the deterministic value of the i^{th} factor at time step $h + 1$, as a function of its value at step h and action a. That is, if at time step h we are in state $(h, x_1, \dots, x_d)$, upon taking action a we will transition deterministically to $(h + 1, P_h^{\mathbf{p},1}(x_1, a), \dots, P_h^{\mathbf{p},d}(x_d, a))$.

Levels 1 to $H-1$ form a complete binary tree; see Figure 1 for an illustration. For any layer $h \le H-2$,

$$P_h^{\mathbf{p},i}(v,a) = v, \quad \forall v \in \mathcal{O}, a \in \mathcal{A},\, i \neq h;$$
$$P_h^{\mathbf{p},i}(v,a) = a, \quad \forall v \in \mathcal{O}, a \in \mathcal{A},\, i = h.$$

In words, any internal state at level $h \le H-1$ simply encodes the sequence of actions that leads to it. These transitions *do not* depend on the planted path $\mathbf{p}$ and are identical across all models. Note that it is not possible to have $x_i = 2$ for any $i \in [d], h \le H-1$.

Now we define the transition from level $H-1$ to H, where each state only has 1 action, say $+1$:

$$P_{H-1}^{\mathbf{p},i}(p_i, +1) = p_i, \quad \forall i \in [d], \qquad \text{and} \qquad P_{H-1}^{\mathbf{p},i}(\bar{p}_i, +1) = 2, \quad i \in [d].$$

Here $\bar{p}_i$ is the negation of p_i. In words, the state at level H simply copies the state at level $H-1$, except that the i^{th} factor will take value 2 if it disagrees with p_i (see Figure 1). Thus, the agent arrives at a state without the symbol "2" at level H only if it follows the action sequence $\mathbf{p}$.

The reward function is shared across all models. Non-zero rewards are only available at level H, where each state only has 1 action. The reward is 1 if x does not contain the symbol "2" and the reward is 0 otherwise. Formally

$$R((h, x_1, \ldots, x_d)) \triangleq \mathbf{1}\{h = H\} \prod_{i=1}^{d} \mathbf{1}\{x_i \neq 2\}. \tag{16}$$

Sample Efficiency of Algorithm 3 For this family of factored MDPs, we have $K = 2$ and $d = H-2$. The remaining parameter of interest is L, on which we provide a coarse upper bound: $L \le dH|\mathcal{A}||\mathcal{O}|^2 = O(H^2)$ since $|\mathrm{pa}_i| = 1$ for all i and $|\mathcal{O}| = 4$. Given that our Algorithm 3 works for factored MDPs of any structure, the guarantees in Theorem 8 immediately applies and we obtain a sample complexity that is polynomial in H and $\log(1/\delta)$. This proves the first claim of Theorem 2.

C.2.2. SAMPLE INEFFICIENCY OF MODEL-FREE ALGORITHMS

We prove the second claim by showing that any model-free algorithm—that is, any algorithm that always accesses state x exclusively through $[f(x,\cdot)]_{f\in\mathcal{G}}$—will incur exponential sample complexity when given $\mathcal{G} = \mathtt{OP}(\mathcal{P})$ as input. To show this, we construct another class of non-factored models, such that (1) learning in this new class is intractable, and (2) the two families are indistinguishable to any model-free algorithm. The new model class is obtained by transforming each $P^{\mathbf{p}} \in \mathcal{P}$ into $\tilde{P}^{\mathbf{p}}$. $\tilde{P}^{\mathbf{p}}$ has the same state space and transitions as $P^{\mathbf{p}}$, except for the transition from level $H-1$ to H. This last transition is:

$$\tilde{P}_h^{\mathbf{p}}((H-1, x_1, \ldots, x_d)) = \begin{cases} (H, x_1, \ldots, x_d) & \text{if } x_i = \mathbf{p}_i \ \forall i \in [d] \\ H \circ [2]^d & \text{otherwise.} \end{cases}$$

The reward function is the same as in the original model class, given in (16). This construction is equivalent to a multi-armed bandit problem with one optimal arm among 2^{H-2} arms, so the sample complexity of *any algorithm* (not necessarily restricted to model-free ones) is $\Omega(2^H)$.[13] In fact this model class is almost identical to the one used in the proof of Proposition 4.

To prove that the two model families are indistinguishable for model-free algorithms (Definition 1), we show that the $\mathcal{G}$-profiles in $P^{\mathbf{p}}$ are identical to those in $\tilde{P}^{\mathbf{p}}$. This implies that the behavior of a model-free algorithm is identical in $P^{\mathbf{p}}$ and $\tilde{P}^{\mathbf{p}}$, so that the sample complexity must be identical, and hence $\Omega(2^H)$.

Let $\mathcal{M} = \{P^{\mathbf{p}}\}_{\mathbf{p}\in\{-1,1\}^d}$ and $\widetilde{\mathcal{M}} = \{\tilde{P}^{\mathbf{p}}\}_{\mathbf{p}\in\{-1,1\}^d}$. Let $\mathcal{Q}, \Pi$ to be the Q class and policy classs from $\mathtt{OP}(\mathcal{M})$, $\widetilde{\mathcal{Q}}$ and $\widetilde{\Pi}$ be the policy class from $\mathtt{OP}(\widetilde{\mathcal{M}})$. Since all MDPs of interest have fully deterministic dynamics, and non-zero rewards only occur at the last step, it suffices to show that for any deterministic sequence of actions, **a**, (1) the final reward has the same distribution for $P^{\mathbf{p}}$ and $\tilde{P}^{\mathbf{p}}$, and (2) the Q-profiles $[Q(x_h,\cdot)]_{Q\in\mathcal{Q}}$ and $[Q(x_h,\cdot)]_{Q\in\tilde{\mathcal{Q}}}$ are equivalent at all states generated by taking **a** in $P^{\mathbf{p}}$ and $\tilde{P}^{\mathbf{p}}$, respectively.[14] The reward equivalence is obvious, so it remains to study the Q-profiles.

In $P^{\mathbf{p}}$ and at level H, since the reward function is shared, the Q-profile is $[1]^{|\mathcal{Q}|}$ for the state without "2" and $[0]^{|\mathcal{Q}|}$ otherwise. Thus, upon taking $\mathbf{a} = \mathbf{p}$ we see the Q-profile $[1]^{|\mathcal{Q}|}$ and otherwise we see $[0]^{|\mathcal{Q}|}$. Similarly, in $\tilde{P}^{\mathbf{p}}$ the Q-profile is $[0]^{|\tilde{\mathcal{Q}}|}$ if the state is $H \circ [2]^d$ and it is $[1]^{|\tilde{\mathcal{Q}}|}$ otherwise. The equivalence here is obvious as $|\mathcal{Q}| = |\widetilde{\mathcal{Q}}| = 2^d$.

For level $H-1$, no matter the true model path $\mathbf{p}$, the $Q^{\mathbf{p}'}$ associated with path $\mathbf{p}'$ has value $Q^{\mathbf{p}'}(\mathbf{a}, +1) = \mathbf{1}\{\mathbf{a} = \mathbf{p}'\}$ at state $\mathbf{a}$. Hence the Q-profile at $\mathbf{a}$ can be represented as $[\mathbf{1}\{\mathbf{a} = \mathbf{p}'\}]_{\mathbf{p}'\in\{-1,1\}^d}$, for both $P^{\mathbf{p}}$ and $\tilde{P}^{\mathbf{p}}$. Note that the Q-profile does not depend on the true model $\mathbf{p}$ because all models agree on the dynamics before the last step. Similarly, for $h < H-1$ where each state has two actions $\{-1, 1\}$, we have:

$$Q^{\mathbf{p}'}(\mathbf{a}_{1:h-1}, -1) = \mathbf{1}\left\{\mathbf{a}_{1:h-1} \circ \text{-}1 = \mathbf{p}'_{1:h}\right\}, Q^{\mathbf{p}'}(\mathbf{a}_{1:h-1}, 1) = \mathbf{1}\left\{\mathbf{a}_{1:h-1} \circ 1 = \mathbf{p}'_{1:h}\right\}.$$

Hence, the Q-profile can be represented as:

$$[(\mathbf{1}\left\{\mathbf{a}_{1:h-1} \circ \text{-}1 = \mathbf{p}'_{1:h}\right\}, \mathbf{1}\left\{\mathbf{a}_{1:h-1} \circ 1 = \mathbf{p}'_{1:h}\right\})]_{\mathbf{p}'\in\{-1,1\}^d},$$

again with no difference between $P^{\mathbf{p}}$ and $\tilde{P}^{\mathbf{p}}$. Thus, the model $P^{\mathbf{p}}$ and $\tilde{P}^{\mathbf{p}}$ induce exactly the same Q-profile for all paths, implying that any model-free algorithm (in the sense of Definition 1), must behave identically on both. Since the family $\widetilde{\mathcal{M}} = \{\tilde{P}^{\mathbf{p}}\}_{\mathbf{p}}$ admits an information-theoretic sample complexity lower bound of $\Omega(2^H)$, this same lower bound applies to $\mathcal{M} = \{P^{\mathbf{p}}\}_{\mathbf{p}}$ for model-free algorithms.

C.3. Circumventing the Lower Bound via Overparameterization

In Theorem 2, the $\mathcal{G}$-profile resulting from the class $\mathcal{M}$ (Appendix C.2.1) obfuscates the true context, a property critical for the separation result. In this section, we show that by increasing the expressiveness of the model class, the induced $\mathcal{G}$-profile could reveal the context and circument the lower bound. More directly, the lower bound is sensitive to the choice of $\mathcal{G}$.

13. Note that the reward function is known and non-random, so we do not have any dependence on an accuracy parameter ϵ.

14. Since each $\pi \in \Pi$ is just derived from some $Q \in \mathcal{Q}$, the equivalence between two Q-profiles implies the equivalence between two Π-profiles, which further implies equivalence in $\mathcal{G}$-profiles.

This sensitivity raises the question: what is the right choice of $\mathcal{G}$ for comparing model-based and model-free methods? Since our construction considers only a small subset of all possible factored MDPs in correspondence with $\mathcal{M}$, the class $\mathcal{G} = \texttt{OP}(\mathcal{M})$ is the smallest class that guarantees $(Q^\star, \pi^\star) \in \mathcal{G}$. Therefore, this choice amounts to *proper learning*, which we argue is a natural choice for the purpose of proving lower bounds. On the other hand, in this section we show that *improper learning* or "overparametrization" can circumvent the lower bound.

Recall that in Appendix C.2.1, every model $M \in \mathcal{M}$ uses the same *true* reward function from (16). Here we create a larger model class $\mathcal{M}'$ and take $\mathcal{G}' = \texttt{OP}(\mathcal{M}')$. First define a set of new reward functions:

$$R_i^{(-1)}(H, \mathbf{x}) \triangleq \mathbf{1}\{x_i \neq -1\}, R_i^{(1)}(H, \mathbf{x}) \triangleq \mathbf{1}\{x_i \neq 1\}, R_i^{(2)}(H, \mathbf{x}) \triangleq \mathbf{1}\{x_i \neq 2\} \;\; \forall i \in [d] \quad (17)$$

We set $\mathcal{M}' = \mathcal{M} \cup \{(P, R_i^{(j)}) : P \in \mathcal{P}, i \in [d], j \in \{-1, 1, 2\}\}$. Namely for every transition structure $P \in \mathcal{P}$, we pair it with each new reward function. Note that $|\mathcal{M}'| = (3d+1)\,|\mathcal{M}|$.

To circumvent the lower bound, we simply show that with $\mathcal{G}' = \texttt{OP}(\mathcal{M}')$, the $\mathcal{G}'$-profile actually reveals the context $\mathbf{x}$. Let us focus on a single coordinate $i \in [d]$, and pick any transition operator $P \in \mathcal{P}$. Observe that at level H the Q function corresponding to transition operator P is just the associated reward, and so the $\mathcal{G}$-profile reveals $R_i^{(-1)}, R_i^{(1)}, R_i^{(2)}$. Since we know that $x_i = \{-1, 1, 2\}$ exactly one of these will evaluate to zero, allowing us to recover the i^{th} bit. In particular this allows us to immediately identify the correct action for time step $h = i$. Since we can do this for every $i \in [d]$, using $\mathcal{G}' = \texttt{OP}(\mathcal{M}')$, we can easily obtain an algorithm with $O(1)$ sample complexity.

Appendix D. $\mathcal{G}$-profiles in tabular settings

Here we show that the $\mathcal{G}$-profile yields no information loss in tabular environments. Thus from the perspective of Definition 1, model-based and model-free algorithms are information-theoretically equivalent.

In tabular settings, the state space $\mathcal{X}$ and action space $\mathcal{A}$ are both finite and discrete. It is also standard to use a fully expressive Q-function class, that is $\mathcal{Q} = \{Q : \mathcal{X} \times \mathcal{A} \to [0,1]\}$, where the range here arises due to the bounded reward. We simply set $\mathcal{G} = \mathcal{Q}$ here. For each state $x \in \mathcal{X}$ define the function Q^x such that for all $a \in \mathcal{A}$, $Q^x(x', a) = \mathbf{1}\{x = x'\}$. Observe that since $\mathcal{Q}$ is fully expressive, we are ensured that $Q^x \in \mathcal{Q}, \forall x \in \mathcal{X}$.

At any state $x' \in \mathcal{X}$, from the Q-profile $\Phi_{\mathcal{Q}}(x')$ we can always extract the values $[Q^x(x', a)]_{x \in \mathcal{X}}$ for some fixed action a. By construction of the Q^x functions, exactly one of these values will be one, while all others will be zero, and thus we can recover the state x' simply by examining a few values in $\Phi_{\mathcal{Q}}(x')$. In other words, the mapping $x \mapsto \Phi_{\mathcal{Q}}(x)$ is invertible in the tabular case, and so there is no information lost through the projection. Hence in tabular setting, one can run classic model-free algorithms such as Q-learning (Watkins and Dayan, 1992) under our definition.

Our definition can also be applied to parameterized Q-function class $\mathcal{Q} \triangleq \{Q(\cdot, \cdot|\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$. To perform gradient-based update on the parameter θ, we can use Q-profile as follows. Given any state-action pair (x, a), we can approximate $\nabla_{\theta_i} Q(x, a|\theta), \forall i \in [d]$, to an arbitrary accuracy, using finite differencing:

$$\nabla_{\theta_i} Q(x, a|\theta) = \lim_{\delta \to 0} \frac{Q(x, a|\theta + \delta e_i) - Q(x, a|\theta - \delta e_i)}{2\delta},$$

Algorithm 3 Variant of Algorithm 1 for factored MDPs. Arguments: $(\mathcal{M}, n, n_e, \epsilon, \delta, \phi)$

1: Run Algorithm 1 with $\mathcal{F}$ in (18), except in line 8, estimate $\widehat{\mathcal{W}}_F(M^t, M', h_t)$ via (19).

where e_i is the vector with zero everywhere except one in the i-th entry, and $Q(x,a|\theta+\delta e_i)$ and $Q(x,a|\theta-\delta e_i)$ can be extracted from the Q-profile $\Phi_{\mathcal{Q}}(x)$.

The Q-profile can also be used to estimate policy gradient on policies induced from the parameterized Q functions. Denote $\Pi_{\mathcal{Q}}$ as the policy class induced from $\mathcal{Q}$, e.g., $\pi(a|x;\theta) \propto \exp(Q(x,a|\theta))$. Policy gradient method often involves computing the gradient of the log likelihood of the policy (e.g., REINFORCE (Williams, 1992)): $\nabla_{\theta_i} \log(\pi(x|a;\theta)), \forall i \in [d]$, which via chain rule, is determined by $\nabla_\theta Q(x,a|\theta)$. Hence, with the finite differencing technique we introduced above for computing $\nabla_\theta Q(x,a|\theta)$, we can use Q-profile to compute $\nabla_\theta \log \pi(x|a;\theta)$.

Appendix E. Proof of Theorem 8

Here we prove Theorem 8, which states that Algorithm 3 can handle factored MDPs, where $\mathcal{M}$ is the infinite class of all possible factored MDPs under the given structure (i.e., $\{\text{pa}_i\}$ are known). Since the only difference between two models is their transitions, we use $\mathcal{P} = \{P : (R^\star, P) \in \mathcal{M}\}$ to represent the model class, and use P and M interchangeably sometimes.

As an input to the algorithm, we supply an $\mathcal{F}$ tailored for factored MDPs that always guarantees Bellman domination (up to a multiplicative constant; see Lemma 26). In particular,

$$\mathcal{F} = \{g_1 + \dots g_d : g_i \in \mathcal{G}_i\}, \tag{18}$$

where each $\mathcal{G}_i = (\mathcal{O}^{|\text{pa}_i|} \times \mathcal{A} \times [H] \times \mathcal{O} \to \{-1,1\})$. Note that functions in $\mathcal{F}$ operate on (x,a,r,x') and here we are using a slightly incorrect but intuitive notation: $g_i \in \mathcal{G}_i$ takes (x,a,r,x') as input, and only looks at $(x[\text{pa}_i], h, a, x'[i])$ to determine a binary output value, and $\mathcal{G}_i$ is the set of all functions of this form. The IPM induced by $\mathcal{F}$ is the sum of total variation for each factor, and

$$|\mathcal{F}| = \prod_{i=1}^{d} 2^{HK|\mathcal{O}|^{1+|\text{pa}_i|}} = 2^L,$$

so its logarithmic size is polynomial in L and allows uniform convergence. One slightly unusual property of $\mathcal{F}$, compared to how it is used in other results in the main text, is that functions in $\mathcal{F}$ has ℓ_∞ norm bounded by d instead of a constant, and this magnitude will be manifested in the sample complexity through concentration bounds.

Besides the specific choice of $\mathcal{F}$, we also need an important change in how we estimate the model misfit $\mathcal{W}_F$ defined in (7). Since $\mathcal{W}_F$ is defined w.r.t. uniformly random actions, we change our estimate accordingly by simply dropping the importance weight in line 8: Given dataset $\{(x_h^{(i)}, a_h^{(i)}, r_h^{(i)}, x_{h+1}^{(i)})\}_{i=1}^n$ generated in line 7 of Algorithm 1 using roll-in policy π_M, the new estimator is

$$\widehat{\mathcal{W}}_F(M, M', h) \triangleq \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathop{\mathbb{E}}_{(r,x') \sim M'_h} [f(x_h^{(i)}, a_h^{(i)}, r, x')] - f(x_h^{(i)}, a_h^{(i)}, r_h^{(i)}, x_{h+1}^{(i)}) \right). \tag{19}$$

We now state a formal version of Theorem 8, which includes the specification of input parameters to Algorithm 3, and prove it in the remainder of this section.

Theorem 20 (Formal version of Theorem 8) *Let $M^\star$ be a factored MDP with known structure (1). For any $\epsilon, \delta \in (0,1]$, set $\beta = O(L/K)$, $\kappa = 1/K$, $\mathrm{W}_{\kappa,F} = L_h/|\mathcal{O}|$,*[15] *$\phi = \frac{\kappa\epsilon}{48H\sqrt{\mathrm{W}_{\kappa,F}}}$, and $T = H\mathrm{W}_{\kappa,F}\log(\beta/2\phi)/\log(5/3)$. Run Algorithm 3 with inputs $(\mathcal{M}, n_e, n, \epsilon, \delta, \phi)$, where $\mathcal{M}$ is the infinite class of all possible factored MDPs with the given structure, and*

$$n_e = \Theta\left(\frac{H^2\log(HT/\delta)}{\epsilon^2}\right), n = \Theta\left(\frac{d^2(L\log(\frac{dKL}{\epsilon}) + \log(3T/\delta))}{|\mathcal{O}|\epsilon^2}LHK^2\right),$$

then with probability at least $1-\delta$, the algorithm outputs a policy π such that $v^\pi \geq v^\star - \epsilon$, using at most the following number of sample trajectories:

$$\tilde{O}\left(\frac{d^2L^3HK^2\log(1/\delta)}{\epsilon^2}\right).$$

E.1. Concentration Result

E.1.1. COVER CONSTRUCTION

We prove uniform convergence by discretizing the CPTs in facotred MDPs and constructing a cover of $\mathcal{P}$. Let $\alpha \in (0,1)$ be the discretization resolution, whose precise value will be set later. For convenience we also assume that $2/\alpha$ is an odd integer. Recall that a factored MDP is fully specified by the CPTs:

$$\{P^{(i)}[o \mid x[\mathrm{pa}_i], a, h] : o \in \mathcal{O}, x[\mathrm{pa}_i] \in \mathcal{O}^{|\mathrm{pa}_i|}, a \in \mathcal{A}, h \in [H]\}_{i=1}^d.$$

Since each of these probabilities takes value in $[0,1]$, we start with an improper cover of $\mathcal{P}$ by discretizing this range and considering cover centers $\{\alpha/2, 3\alpha/2, 5\alpha/2, \dots, 1-\alpha/2\}$ for each $o \in \mathcal{O}, x[\mathrm{pa}_i] \in \mathcal{O}^{|\mathrm{pa}_i|}, a \in \mathcal{A}, h \in [H], i \in [d]$. Note that any number in $[0,1]$ will be $(\alpha/2)$-close to one of these $(1/\alpha)$ values. Altogether the discretization yields

$$\prod_{i=1}^d (1/\alpha)^{HK|\mathcal{O}|^{1+|\mathrm{pa}_i|}} = (1/\alpha)^L$$

(possibly unnormalized) CPTs. For the purpose of cover construction, the distance between two CPTs P and P' is defined as

$$\max_{o\in\mathcal{O}, x[\mathrm{pa}_i]\in\mathcal{O}^{|\mathrm{pa}_i|}, a\in\mathcal{A}, h\in[H]} |P^{(i)}[o \mid x[\mathrm{pa}_i], a, h] - P'^{(i)}[o \mid x[\mathrm{pa}_i], a, h]|. \tag{20}$$

Under this distance, any MDP in $\mathcal{P}$ will be $(\alpha/2)$-close to one of the discretized CPTs, hence we say the discretization yields a $(\alpha/2)$-cover of $\mathcal{P}$ with size $(1/\alpha)^L$.

Note that the above cover is improper because many cover centers violate the normalization constraints. We convert this improper cover to a proper one by (1) discarding all cover centers whose $\alpha/2$ radius ball contains no valid models, and (2) replacing every remaining invalid cover center with

15. Here we treat $\mathrm{W}_{\kappa,F}$ as an algorithm parameter, and its value is an upper bound on the actual witness rank (27).

a valid model in its $\alpha/2$ radius ball. This yields an α-cover with size $(1/\alpha)^L$ whose cover centers are all valid models. We denote the set of cover centers as $\mathcal{P}_c$.

E.1.2. UNIFORM CONVERGENCE OF $\widehat{\mathcal{W}}_F$

Recall the definition of $\widehat{\mathcal{W}}_F$ from (19). Our main concentration result is the following lemma.

Lemma 21 (Concentration of $\widehat{\mathcal{W}}_F$ in factored MDPs) *Fix h and model $P \in \mathcal{P}$. Sample a dataset $\mathcal{D} = \left\{(x_h^{(i)}, a_h^{(i)}, r_h^{(i)}, x_{h+1}^{(i)})\right\}_{i=1}^n$ with $x_h^{(i)} \sim \pi_M$, $a_h^{(i)} \sim U(\mathcal{A})$, $(r_h^{(i)}, x_{h+1}^{(i)}) \sim M_h^\star$ of size n. Fix any ϕ and $\delta > 0$. With probability at least $1-\delta$, we have for all $P' \in \mathcal{P}$: $\left|\widehat{\mathcal{W}}_F(M, M', h) - \mathcal{W}_F(M, M', h)\right| \leq \phi$, as long as*

$$n \geq \frac{8d^2(L\log(\frac{8d|\mathcal{O}|}{\phi}) + \log(2/\delta))}{\phi^2}.$$

We first prove a helper lemma, which quantifies the error introduced by approximating $\mathcal{P}$ with $\mathcal{P}_c$.

Lemma 22 *For any $P' \in \mathcal{P}$, let P'_c be its closest model in $\mathcal{P}_c$. For any $f \in \mathcal{F}$ and any x, a,*

$$|\mathop{\mathbb{E}}_{(r,x')\sim M'_{(x,a)}}[f(x,a,r,x')] - \mathop{\mathbb{E}}_{(r,x')\sim (M'_c)_{(x,a)}}[f(x,a,r,x')]| \leq d|\mathcal{O}|\alpha, \tag{21}$$

where $(r, x') \sim M'_{(x,a)}$ is the shorthand for $r \sim R'(x,a), x' \sim P'_{(x,a)}$.

Proof Recall the definition of $f \in \mathcal{F}$ tailored for factored MDPs: $f = g_1 + \cdots + g_d$, with each $\|g_i\|_\infty \leq 1$. By triangle inequality, we have:

$$\begin{aligned}
\text{LHS} &\leq \sum_{i=1}^d \left| \mathop{\mathbb{E}}_{r,x'\sim M'_{(x,a)}}[g_i(x,a,r,x')] - \mathop{\mathbb{E}}_{r,x'\sim (M'_c)_{(x,a)}}[g_i(x,a,r,x')] \right| \\
&\leq \sum_{i=1}^d \left\| (P')_{x,a}^{(i)} - (P'_c)_{x,a}^{(i)} \right\|_{\text{TV}}, && \text{(Hölder)} \\
&= \sum_{i=1}^d \sum_{o\in\mathcal{O}} \left| (P')^{(i)}[o|x[\text{pa}_i], a, h] - (P'_c)^{(i)}[o|x[\text{pa}_i], a, h] \right| \\
&\leq d|\mathcal{O}|\alpha. && (\mathcal{P}_c \text{ yields } \alpha\text{-cover under distance defined in (20)})
\end{aligned}$$

■

Now we are ready to prove the main concentration result for factored MDPs.

Proof [Proof of Lemma 21] To argue uniform convergence for $\mathcal{P}$, we first apply Hoeffding's inequality and union bound to $\mathcal{P}_c$. For any fixed f, $\widehat{\mathcal{W}}_F$ is the average of i.i.d. random variables with range $[-\|f\|_\infty, \|f\|_\infty]$. For the $\mathcal{F}$ that we use for factored MDPs, $\|f\|_\infty \leq d$, so with probability at least $1-\delta$, $\forall P'_c \in \mathcal{P}_c$,

$$\left|\widehat{\mathcal{W}}_F(M, M'_c, h) - \mathcal{W}_F(M, M'_c, h)\right| \leq 2d\sqrt{\frac{\log(2|\mathcal{P}_c||\mathcal{F}|/\delta)}{2n}}.$$

We then follow a standard argument to decompose the estimation error for any $P' \in \mathcal{P}$ into three terms:

$$\left|\widehat{\mathcal{W}}_F(M,M',h) - \mathcal{W}_F(M,M',h)\right| \le \left|\widehat{\mathcal{W}}_F(M,M'_c,h) - \mathcal{W}_F(M,M'_c,h)\right| \\ + \left|\widehat{\mathcal{W}}_F(M,M',h) - \widehat{\mathcal{W}}_F(M,M'_c,h)\right| + \left|\mathcal{W}_F(M,M',h) - \mathcal{W}_F(M,M'_c,h)\right|.$$

We have an upper bound on the first term , so it suffices to upper-bound the other two terms. For the second term,

$$\left|\widehat{\mathcal{W}}_F(M,M',h) - \widehat{\mathcal{W}}_F(M,M'_c,h)\right| \\ \le \frac{1}{n}\max_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\mathop{\mathbb{E}}_{(r,x')\sim M'_h}[f(x_h^{(i)},a_h^{(i)},r,x')] - \sum_{i=1}^{n}\mathop{\mathbb{E}}_{(r,x')\sim(M'_c)_h}[f(x_h^{(i)},a_h^{(i)},r,x')]\right|.$$

where we use the fact that for any functionals μ_1,μ_2, we have $|\max_f \mu_1(f) - \max_f \mu_2(f)| \le \max_f |\mu_1(f) - \mu_2(f)|$. Now using Lemma 22, we can show that:

$$\left|\widehat{\mathcal{W}}_F(M,M',h) - \widehat{\mathcal{W}}_F(M,M'_c,h)\right| \le \frac{1}{n}(nd\,|\mathcal{O}|\,\alpha) = d\,|\mathcal{O}|\,\alpha.$$

$|\mathcal{W}_F(M,M',h) - \mathcal{W}_F(M,M'_c,h)|$ has the same upper bound using exactly the same argument. So finally we conclude that for all $P' \in \mathcal{P}$,

$$\left|\widehat{\mathcal{W}}_F(M,M',h) - \mathcal{W}_F(M,M',h)\right| \le 2d\sqrt{\frac{\log(2|\mathcal{P}_c||\mathcal{F}|/\delta)}{2n}} + 2d|\mathcal{O}|\alpha.$$

To guarantee that the deviation is no more than ϕ, we back up the necessary sample size n from the above expression. Let $\alpha = \frac{\phi}{4d|\mathcal{O}|}$, so $2d|\mathcal{O}|\alpha \le \phi/2$. We then want

$$2d\sqrt{\frac{\log(2(8d|\mathcal{O}|/\phi)^L/\delta)}{2n}} \le \phi/2.$$

It is easy to verify that the sample size given in the lemma statement satisfies this inequality. ■

E.2. Low Witness Rank and Bellman Domination

In this subsection we establish several important properties of $\mathcal{W}_F$ which will be directly useful in proving Theorem 8. To start, we provide a form of $\mathcal{W}_F$ that is equivalent to the definition provided in (7). The proof is elementary and omitted.

Lemma 23 (Alternative definition of $\mathcal{W}_F$)

$$\mathcal{W}_F(M,M',h) = \mathbb{E}\left[\sum_{i=1}^{d}\left\|P'^{(i)}(\cdot|x_h[\mathrm{pa}_i],a_h) - P^{\star(i)}(\cdot|x_h[\mathrm{pa}_i],a_h)\right\|_{TV}|x_h \sim \pi_P, a_h \sim U(\mathcal{A})\right]$$

Using this lemma, we show two important properties of $\mathcal{W}_F$: (1) that the matrix $\mathcal{W}_F$ has rank at most $\sum_{i=1}^{d} K\,|\mathcal{O}|^{|\mathrm{pa}_{(i)}|}$ (Proposition 24), which is less than L, the description length of the factored

MDP, and (2) that we can upper-bound $\mathcal{E}_B$ using $\mathcal{W}_F$ (Lemma 26). For the remainder, it will be convenient to use the notation $L_h = \sum_{i=1}^d K|\mathcal{O}|^{1+|\text{pa}_{(i)}|}$ to be the number of parameters needed to specify the conditional probability table at a single level h.

Proposition 24 *There exists $\zeta_h : \mathcal{P} \to \mathbb{R}^{L_h/|\mathcal{O}|}$ and $\chi_h : \mathcal{P} \to \mathbb{R}^{L_h/|\mathcal{O}|}$, such that for any $P, P' \in \mathcal{P}$, and $h \in [H]$, (recall that $M = (R, P)$ and $M' = (R, P')$)*

$$\mathcal{W}_F(M, M', h) = \left\langle \zeta_h(M), \chi_h(M') \right\rangle ,$$

and $\|\zeta_h(M)\|_2 \cdot \|\chi_h(M')\|_2 \le O(L_h/K)$.

Proof Given any policy π, let us denote $\eta_h^\pi(x) \in \Delta(\mathcal{X}_h)$ as the state distribution resulting from π at time step h. Then we can write $\eta_h^\pi(x) = \eta_h^\pi(x[u])\eta_h^\pi(x[-u]|x[u])$, where for a subset $u \subset [d]$, we write $x[u]$ to denote the corresponding assignment of those state variables in x, and $-u = [d] \setminus u$ is the set of remaining variables. We use η_h^π to denote the probability mass function and we use $\mathbb{P}_h^\pi$ to denote the distribution.

For any $P, P' \in \mathcal{P}$, we can factorize $\mathcal{W}_F(M, M', h)$ as follows:

$$\begin{aligned}
&\mathcal{W}_F(M, M', h) = \mathbb{E}\left[\sum_{i=1}^d \left\|P'^{,(i)}(\cdot|x_h[\text{pa}_i], a_h) - P^{\star,(i)}(\cdot|x_h[\text{pa}_i], a_h)\right\|_{\text{TV}} |x_h \sim \pi_M, a_h \sim U(\mathcal{A})\right] \\
&= \frac{1}{K}\sum_{i=1}^d \sum_{x_h,a} \eta_h^{\pi_M}(x_h) \left\|P^{\star,(i)}(\cdot|x_h[\text{pa}_i], a) - P'^{,(i)}(\cdot|x_h[\text{pa}_i], a)\right\|_{\text{TV}} \\
&= \frac{1}{K}\sum_{i=1}^d \sum_{x_h,a} \eta_h^{\pi_M}(x_h[\text{pa}_i])\eta_h^{\pi_M}(x_h[-\text{pa}_i]|x_h[\text{pa}_i]) \left\|P^{\star,(i)}(\cdot|x_h[\text{pa}_i], a) - P'^{,(i)}(\cdot|x_h[\text{pa}_i], a)\right\|_{\text{TV}} \\
&= \frac{1}{K}\sum_{i=1}^d \sum_{a} \sum_{u \in \mathcal{O}^{|\text{pa}_i|}} \mathbb{P}_h^{\pi_M}[x_h[\text{pa}_i] = u] \left\|P^{\star,(i)}(\cdot|u, a) - P'^{,(i)}(\cdot|u, a)\right\|_{\text{TV}} \\
&= \left\langle \zeta_h(M), \chi_h(M') \right\rangle .
\end{aligned}$$

Here $\zeta_h(M)$ is indexed by $(i, a, u) \in [d] \times \mathcal{A} \times \mathcal{O}^{|\text{pa}_i|}$ with value

$$\zeta_h(i, a, u; M) \triangleq \mathbb{P}_h^{\pi_M}(x_h[\text{pa}_i] = u)/K.$$

$\chi_h(M')$ is also indexed by i, a, u, with value

$$\chi_h(i, a, u; M') \triangleq \left\|P^{\star,(i)}(\cdot|u, a) - P'^{,(i)}(\cdot|u, a)\right\|_{\text{TV}} .$$

Note that ζ_h's value only depends on M, while χ_h's values only depend on M'. Moreover the dimensions of ζ_h and χ_h are $\sum_{i=1}^d K|\mathcal{O}|^{|\text{pa}_i|} = L_h/|\mathcal{O}|$, each entry of ζ_h is bounded by $1/K$, and each entry of χ_h is at most 2. Hence, we must have $\beta = \sup_{\zeta,\chi} \|\zeta\|_2 \cdot \|\chi\|_2 \le O(L/K)$. Note that we omit $|\mathcal{O}|$ from the denominator as L_h has a higher exponent on $|\mathcal{O}|$ and the quantity is being treated as a constant in the big-oh notation. ■

We now proceed to prove Bellman domination (up to a constant), which relies on the following lemma on the tensorization property of total variation:

Lemma 25 *Let $P_1, \dots, P_n$ and $Q_1, \dots, Q_n$ be distributions where $P_i \in \Delta(\mathcal{S}_i)$ for finite sets $\mathcal{S}_i$. Define the product measures $P^{(n)}, Q^{(n)}$ as $P^{(n)}(s_1, \dots, s_n) \triangleq \prod_{i=1}^n P_i(s_i)$. Then*

$$\left\| P^{(n)} - Q^{(n)} \right\|_{TV} \leq \sum_{i=1}^{n} \| P_i - Q_i \|_{TV} .$$

Proof Define $W_i \in \Delta(\mathcal{S}_1 \times \cdots \times \mathcal{S}_n)$ with $W_i(s_{1:n}) = \prod_{j=1}^i P_j(s_j) \prod_{j=i+1}^n Q_j(s_j)$, with $i \in \{0, \dots, n\}$. This gives $W_0 = Q^{(n)}$, and $W_n = P^{(n)}$. Now, by telescoping, we have

$$\left\| P^{(n)} - Q^{(n)} \right\|_{\mathrm{TV}} = \| W_0 - W_n \|_{\mathrm{TV}} \leq \sum_{i=0}^{n-1} \| W_i - W_{i+1} \|_{\mathrm{TV}} .$$

For $\| W_i - W_{i+1} \|_{\mathrm{TV}}$, we have

$$\| W_i - W_{i+1} \|_{\mathrm{TV}} = \left\| \prod_{j=1}^{i} P_j \prod_{j=i+1}^{n} Q_j - \prod_{j=1}^{i+1} P_j \prod_{j=i+2}^{n} Q_j \right\|_{\mathrm{TV}} = \| Q_{i+1} - P_{i+1} \|_{\mathrm{TV}} .$$

■

With this helper lemma, the following lemma shows the Bellman domination.

Lemma 26 $\frac{1}{K} \mathcal{E}_B(Q_M, Q_{M'}, h) \leq \mathcal{W}_F(M, M', h)$.

Proof

$$\begin{aligned}
&\mathcal{E}_B(Q_M, Q_{M'}, h) = \mathbb{E}\left[Q'(x_h, a_h) - r_h - Q'(x_{h+1}, a_{h+1}) \,\middle|\, x_h \sim \pi_M, a_{h:h+1} \sim \pi_{M'} \right] \\
&= \mathbb{E}\left[\mathop{\mathbb{E}}_{x_{h+1} \sim P'_{x_h, a_h}} [V_{M'}(x_{h+1})] - \mathop{\mathbb{E}}_{x_{h+1} \sim P^\star_{x_h, a_h}} [V_{M'}(x_{h+1})] \mid x_h \sim \pi_M, a_h \sim \pi_{M'} \right] \\
&\leq \mathbb{E}\left[\sum_{a_h} \pi_{M'}(a_h | x_h) \left| \mathop{\mathbb{E}}_{x_{h+1} \sim P'_{x_h, a_h}} [V_{M'}(x_{h+1})] - \mathop{\mathbb{E}}_{x_{h+1} \sim P^\star_{x_h, a_h}} [V_{M'}(x_{h+1})] \right| \mid x_h \sim \pi_M \right] \\
&\leq \mathbb{E}\left[\sum_{a_h} \pi_{M'}(a_h | x_h) \left\| P'_{x_h, a_h} - P^\star_{x_h, a_h} \right\|_{\mathrm{TV}} \mid x_h \sim \pi_M \right] \qquad \text{(Hölder and boundedness of } V_{M'}) \\
&\leq K \, \mathbb{E}\left[\mathbb{E}_{a_h \sim U(\mathcal{A})} \left[\left\| P'_{x_h, a_h} - P^\star_{x_h, a_h} \right\|_{\mathrm{TV}} \right] \mid x_h \sim \pi_M \right] \leq K \mathcal{W}_F(M, M', h).
\end{aligned}$$

The first step follows as we expand the definition of $Q'(x_h, a_h)$ by Bellman equation in M', and realize that the immediate reward cancels out with r_h in expectation as the reward function is known. The last step follows from Lemma 25 and Lemma 23. ■

Combining Proposition 24 and Lemma 26 we arrive at the following corollary:

Corollary 27 *Recall the definition of witness rank in Definition 5. Set $\kappa = \frac{1}{K}$, we have* $\mathrm{W}_{\kappa, F} \triangleq \mathrm{W}(\frac{1}{K}, \beta, \mathcal{M}, \mathcal{F}, h) \leq L_h / |\mathcal{O}|$,[16] *for $\beta = O(L/K)$.*

16. Note that we abuse the definition of W to use $\mathcal{W}_F$ instead of $\mathcal{W}$ in Definition 5 here.

Algorithm 4 Guessing $\mathrm{W}_{\kappa^\star}/\kappa^\star$, Arguments: $(\mathcal{M}, \mathcal{F}, \epsilon, \delta)$

for epoch $i = 1, 2, ...$ **do**
  Set $N_i = 2^{i-1}$ and $\delta_i = \delta/(i(i+1))$
  for $j = 1, 2, \ldots$ **do**
    Set $\kappa_{i,j} = (1/2)^{j-1}$, $\delta_{i,j} = \delta_i/(j(j+1))$, and $\mathrm{W}_{i,j} = N_i \kappa_{i,j}$
    if $\mathrm{W}_{i,j} < 1$ **then**
      Break
    end if
    Set $T_{i,j} = H\mathrm{W}_{i,j} \log(\beta/(2\phi))/\log(5/3)$ and $\phi_{i,j} = \epsilon\kappa_{i,j}/(48H\sqrt{\mathrm{W}_{i,j}})$
    Set $n_{e_{i,j}} = \Theta\left(\frac{H^2 \log(6HT_{i,j}\delta_{i,j})}{\epsilon^2}\right)$ and $n_{i,j} = \Theta\left(\frac{H^2 K \mathrm{W}_{i,j} \log(12T_{i,j}|\mathcal{M}||\mathcal{F}|\delta_{i,j})}{\kappa_{i,j}^2 \epsilon^2}\right)$
    Run Algorithm 1 with $(\mathcal{M}, \mathcal{F}, n_{i,j}, n_{e_{i,j}}, \epsilon, \delta_{i,j}, \phi)$ for $T_{i,j}$ iterations
    If Algorithm 1 returns a policy, then break and return the policy
  end for
end for

E.3. Proof of Theorem 8

The proof is largely the same as that of Theorem 6, and the only difference is that we use $\widehat{\mathcal{W}}_F$ as the estimator and handle infinite $\mathcal{M}$, whose uniform convergence property is provided in Section E.1.2. Following the proof of Theorem 6, within the high probability events,the algorithm must terminate in $T = \mathrm{W}_{\kappa,F} H \log(\beta/(2\phi))/\log(5/3)$ iterations, where $\kappa = 1/K$ and $\mathrm{W}_{\kappa,F} \le L_h/|\mathcal{O}|$ (Corollary 27) . As in the previous proof we still set $\phi = \frac{\kappa\epsilon}{48H\sqrt{\mathrm{W}_{\kappa,F}}}$. Plugging this value into Lemma 21 and requiring that each of these estimation events succeeds with probability at least $1 - \delta/3T$, we have

$$n = \Theta\left(\frac{d^2(L\log(\frac{dKL}{\epsilon}) + \log(3T/\delta))}{|\mathcal{O}|\epsilon^2} LHK^2\right) = \tilde{O}\left(\frac{d^2 L^2 H K^2 \log(1/\delta)}{\epsilon^2}\right).$$

Here the $|\mathcal{O}|^2$ on the denominator is dropped due to its negligible magnitude compared to L. The rest of the proof is unchanged: Since estimating $\mathcal{W}_F$ requires much more samples than other estimation events, the order of the overall sample complexity is determined by the above expression multiplied by T, which gives the desired sample complexity.

Appendix F. Extension to Unknown Witness Rank

Algorithm 1 and its analysis assumes that we know κ and W_κ (in fact any finite upper bound of W_κ), which could be a strong assumption in some cases. In this section, we show that we can apply a standard doubling trick to handle the situation where κ and W_κ are unknown.

Let us consider the quantity W_κ/κ. Let us denote $\kappa^\star = \arg\min_{\kappa\in(0,1]} \mathrm{W}_\kappa/\kappa$. Note that the sample complexity of Algorithm 1 is minimized at $\kappa^\star$. Algorithm 4 applies the doubling trick to guess $\mathrm{W}_{\kappa^\star}$ and $\kappa^\star$ jointly with Algorithm 1 as a subroutine. In the algorithm, N_i in the outer loop denotes a guess for W/κ as a whole, and in the inner loop we use $\kappa_{i,j}$ to guess κ, while setting $\mathrm{W}_{i,j} = N_i\kappa_{i,j}$, which we use to set the parameter ϕ and n. The following theorem characterizes the its sample complexity.

Theorem 28 *For any $\epsilon, \delta \in (0,1)$, with $\mathcal{M}$ and $\mathcal{F}$ satisfying Assumption 1 and Assumption 2, with probability at least $1-\delta$, Algorithm 4 terminates and outputs a policy π with $v^\pi \geq v^\star - \epsilon$, using at most*

$$\tilde{O}\left(\frac{H^3 K \mathrm{W}_{\kappa^\star}^2 \log(|\mathcal{M}||\mathcal{F}|/\delta)}{(\kappa^\star \epsilon)^2}\right) \text{ trajectories.}$$

Proof Consider the j^{th} iteration in the i^{th} epoch. Based on the value of $\phi_{i,j}$, $n_{e_{i,j}}$, $n_{i,j}$, using Lemma 12 and Lemma 11, with probability at least $1-\delta_{i,j}$, for any $t \in [1, T_{i,j}]$ during the run of Algorithm 1, we have

$$|v^{\pi^t} - \hat{v}^{\pi^t}| \leq \epsilon/8, \tag{22}$$

$$|\hat{\mathcal{E}}_B(M^t, M^t, h) - \mathcal{E}_B(M^t, M^t, h)| \leq \epsilon/(8H), \forall h \in [H], \tag{23}$$

$$|\widehat{\mathcal{W}}(M^t, M', h_t) - \mathcal{W}(M^t, M', h_t)| \leq \phi_{i,j}, \forall M' \in \mathcal{M}. \tag{24}$$

The first condition above ensures that if Algorithm 1 terminates in the j^{th} iteration and the i^{th} epoch and outputs π, then π must be near-optimal, based on Lemma 14. The third inequality above together with the elimination criteria in Algorithm 1 ensures that $M^\star$ is never eliminated.

Denote $i_\star$ as the epoch where $2\mathrm{W}_{\kappa^\star}/\kappa^\star \leq N_{i_\star} \leq 4\mathrm{W}_{\kappa^\star}/\kappa^\star$, and $j_\star$ as the iteration inside the $i_\star^{\text{th}}$ epoch where $\kappa^\star/2 \leq \kappa_{i_\star,j_\star} \leq \kappa^\star$. Since $\mathrm{W}_{i_\star,j_\star} = N_{i_\star}\kappa_{i_\star,j_\star}$, we have:

$$\mathrm{W}_{\kappa^\star} \leq \mathrm{W}_{i_\star,j_\star} \leq 4\mathrm{W}_{\kappa^\star}. \tag{25}$$

Below we condition on the event that $M^\star$ is not eliminated during any epoch before $i_\star$, and any iteration before $j_\star$ in the $i_\star^{\text{th}}$ epoch. We analyze the $j_\star^{\text{th}}$ iteration in the $i_\star^{\text{th}}$ epoch below. Since $N_{i_\star} = 2^{i_\star - 1}$ and $N_{i_\star} \leq 4\mathrm{W}_{\kappa^\star}/\kappa^\star$, we must have $i_\star \leq 1 + \log_2(4\mathrm{W}_{\kappa^\star}/\kappa^\star)$. Also note that we have with these settings that $\mathrm{W}_{i^\star,j^\star}/(\kappa_{i^\star,j^\star})^2 \geq \mathrm{W}_{\kappa^\star}/(\kappa^\star)^2$, so that the number of samples we use at round $(i^\star, j^\star)$is at least as large, and the parameter ϕ is no larger than what we would have if we knew $\kappa^\star$ and used it in Algorithm 1.

Based on the value of $\phi_{i_\star,j_\star}$, $n_{i_\star,j_\star}$, $n_{e_{i_\star,j_\star}}$ and $T_{i_\star,j_\star}$, we know with probability at least $1-\delta_{i_\star,j_\star}$, for any $t \in [1, T_{i_\star,j_\star}]$ in the execution of Algorithm 1, inequalities (22), (23), and (24) hold. Conditioned on this event and since $\phi_{i^\star,j^\star}$ is small enough as observed above, similar to the proof of Lemma 17, we can show that Algorithm 1 must terminate in at most $H\mathrm{W}_{\kappa^\star}\log(\beta/2\phi)/\log(5/3)$ many rounds in this iteration.

From (25), we know that $\mathrm{W}_{i_\star,j_\star} \geq \mathrm{W}_{\kappa^\star}$, which implies that $T_{i_\star,j_\star} \geq H\mathrm{W}_{\kappa^\star}\log(\beta/2\phi)/\log(5/3)$. In other words, in the $j_\star^{\text{th}}$ iteration of the $i_\star^{\text{th}}$ epoch, we run Algorithm 1 long enough to guarantee that it terminates and outputs a policy. We have already ensured that if it terminates, it must output a policy π with $v^\pi \geq v^\star - \epsilon$ (this is true for any (i,j) pair).

Now we calculate the sample complexity. In the i^{th} epoch, since we terminate when $\mathrm{W}_{i,j} < 1$, the number of iterations is at most $\log_2 N_i < i$. Hence the number of trajectories collected in this epoch is at most

$$\sum_{j=1}^{i}(n_{e_{i,j}} + n_{i,j})T_{i,j} = \sum_{j=1}^{i} O\left(H^3 K \mathrm{W}_{i,j}^2 \log(T_{i,j}|\mathcal{M}||\mathcal{F}|/\delta_{i,j})/(\epsilon\kappa_{i,j})^2\right)$$
$$= O\left(iH^3 K N_i^2 \log(T_{i,1}|\mathcal{M}||\mathcal{F}|/\delta_{i,i})/\epsilon^2\right),$$

where we used the fact that $N_i = \mathrm{W}_{i,j}/\kappa_{i,j}$, $T_{i,1} \geq T_{i,j}$, and $\delta_{i,i} \leq \delta_{i,j}$. Note that $\sum_{i=1}^{i_\star-1} iN_i^2 = \sum_{i=1}^{i_\star-1} i(2^{i-1})^2 \leq (i_\star - 1)(2^{i_\star-1})^2/3 = O(i_\star N_{i_\star}^2)$. Hence the sample complexity in the $i_\star^{\text{th}}$ epoch dominates the total sample complexity, which is

$$\tilde{O}\left((1 + \log_2(4\mathrm{W}_{\kappa^\star}/\kappa^\star))H^3 K \mathrm{W}_{\kappa^\star}^2 \log(T_{i_\star,0}|\mathcal{M}||\mathcal{F}|/\delta_{i_\star,i_\star})/(\kappa^\star\epsilon)^2\right),$$

where we used the fact that $i_\star \leq 1+\log_2(2\mathrm{W}_{\kappa^\star}/\kappa^\star)$, and $N_{i_\star} \leq 2\mathrm{W}_{\kappa^\star}/\kappa^\star$. Applying a union bound over (i,j), with $i \leq i_\star$, since we have $\sum_{i=1}^{i_\star}\sum_{j=1}^{i} \delta_{i,j} \leq \sum_{i=1}^{i_\star} \delta_{i_\star,1} = \sum_{i=1}^{i_\star} \delta/(i(i+1)) \leq \delta$, the failure probability is at most δ, which proves the theorem. ■

Appendix G. Details on Exponential Family Model Class

For any model $M \in \mathcal{M}$, conditioned on $(x,a) \in \mathcal{X}\times\mathcal{A}$, we assume $M_{x,a} \triangleq \exp(\langle \theta_{x,a}, \mathrm{T}(r,x')\rangle)/Z(\theta_{x,a})$ with $\theta_{x,a} \in \Theta \subset \mathbb{R}^m$. Without loss of generality, we assume $\|\theta_{x,a}\| \leq 1$, and $\Theta = \{\theta : \|\theta\| \leq 1\}$. We design $\mathcal{V} = \{\mathcal{X}\times\mathcal{A} \to \Theta\}$, i.e., $\mathcal{V}$ contains all mappings from $(\mathcal{X}\times\mathcal{A})$ to Θ. We design $\mathcal{F} = \{(x,a,r,x') \mapsto \langle v(x,a), \mathrm{T}(r,x')\rangle : v \in \mathcal{V}\}$. Using Definition 3, we have:

$$\begin{aligned}
\mathcal{W}(M,M',h) &= \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{x_h\sim\pi_M, a_h\sim\pi_{M'}}\left[\mathop{\mathbb{E}}_{(r,x')\sim M'_h}[f(x_h,a_h,r,x')] - \mathop{\mathbb{E}}_{(r,x')\sim M^\star_h}[f(x_h,a_h,r,x')]\right] \\
&= \mathop{\mathbb{E}}_{x_h\sim\pi_M, a_h\sim\pi_{M'}}\left[\sup_{\theta\in\Theta}\left(\mathop{\mathbb{E}}_{(r,x')\sim M'_h}\left[\langle\theta, \mathrm{T}(r,x')\rangle\right] - \mathop{\mathbb{E}}_{(r,x')\sim M^\star_h}\left[\langle\theta, \mathrm{T}(r,x')\rangle\right]\right)\right] \\
&= \mathop{\mathbb{E}}_{x_h\sim\pi_M, a_h\sim\pi_{M'}}\left[\left\|\mathop{\mathbb{E}}_{(r,x')\sim M'_h}[\mathrm{T}(r,x')] - \mathop{\mathbb{E}}_{(r,x')\sim M^\star_h}[\mathrm{T}(r,x')]\right\|_\star\right],
\end{aligned}$$

where the second equality uses the fact that $\mathcal{V}$ contains all possible mappings from $\mathcal{X}\times\mathcal{A} \to \Theta$.

We assume that for any $\theta \in \Theta$, the hessian of the log partition function $\nabla^2 \log(Z(\theta))$ is positive definite with eigenvalues bounded between $[\gamma, \beta]$ with $0 \leq \gamma \leq \beta$. Below, we show that under the above assumptions, Bellman domination required for Assumption 2 still holds up to a constant.

Claim 29 (Bellman Domination for Exponential Families) *In the exponential family setting, we have*

$$\frac{\gamma}{2\sqrt{2\beta}}\mathcal{E}_B(M,M',h) \leq \mathcal{W}(M,M',h).$$

Proof We leverage Theorem 3.2 from Gao et al. (2018), which implies that

$$\frac{\gamma}{\sqrt{\beta}} \mathop{\mathbb{E}}_{x_h\sim\pi_M, a_h\sim\pi_{M'}} \sqrt{D_{KL}(M'_{x_h,a_h}||M^\star_{x_h,a_h})} \leq \mathcal{W}(M,M',h).$$

By Pinsker's inequality, we have:

$$\frac{\gamma}{\sqrt{2\beta}} \mathop{\mathbb{E}}_{x_h\sim\pi_M, a_h\sim\pi_{M'}} \left\|M'_{x_h,a_h} - M^\star_{x_h,a_h}\right\|_{\mathrm{TV}} \leq \mathcal{W}(M,M',h),$$

where the LHS is the witness model misfit defined using total variation directly (3).

On the other hand, we know that $r + V_M(x)$ for any $M \in \mathcal{M}$ is upper bounded by 2 via our regularity assumption on the reward. Hence, the TV-based witness model misfit upper bounds Bellman error as follows:

$$2 \mathop{\mathbb{E}}_{x_h \sim \pi_M, a_h \sim \pi_{M'}} \left\| M'_{x_h,a_h} - M^\star_{x_h,a_h} \right\|_{\text{TV}} \geq \mathcal{E}_B(M, M', h),$$

which concludes the proof. ■

Note that the constant $\gamma/(2\sqrt{2\beta})$ can be absorbed into κ in the definition of witness rank (Definition 5).