# Exploration and Apprenticeship Learning in Reinforcement Learning

**Pieter Abbeel** PABBEEL@CS.STANFORD.EDU
**Andrew Y. Ng** ANG@CS.STANFORD.EDU
Computer Science Department, Stanford University Stanford, CA 94305, USA

## Abstract

We consider reinforcement learning in systems with unknown dynamics. Algorithms such as $E^3$ (Kearns and Singh, 2002) learn near-optimal policies by using "exploration policies" to drive the system towards poorly modeled states, so as to encourage exploration. But this makes these algorithms impractical for many systems; for example, on an autonomous helicopter, overly aggressive exploration may well result in a crash. In this paper, we consider the apprenticeship learning setting in which a teacher demonstration of the task is available. We show that, given the initial demonstration, no explicit exploration is necessary, and we can attain near-optimal performance (compared to the teacher) simply by repeatedly executing "exploitation policies" that try to maximize rewards. In finite-state MDPs, our algorithm scales polynomially in the number of states; in continuous-state linear dynamical systems, it scales polynomially in the dimension of the state. These results are proved using a martingale construction over relative losses.

## 1. Introduction

The Markov Decision Processes (MDPs) formalism provides a powerful set of tools for modeling and solving control problems, and many algorithms exist for finding (near) optimal solutions for a given MDP (see, e.g., Bertsekas & Ttsitsiklis, 1996; Sutton & Barto, 1998). To apply these algorithms to control problems in which the dynamics are not known in advance, the parameters of the MDP typically need to be learned from observations of the system.

A key problem in learning an MDP's parameters is that of *exploration*: How can we ensure that all relevant parts of the MDP are visited sufficiently often that we manage to collect accurate statistics for their state transition probabilities? The state-of-the-art answer to this problem is the $E^3$-algorithm (Kearns & Singh, 2002) (and variants/extensions: Kearns & Koller, 1999; Kakade, Kearns & Langford, 2003; Brafman & Tennenholtz, 2002). These algorithms guarantee that near-optimal performance will be obtained in time polynomial in the number of states of the system. The basic idea of $E^3$ is that it will repeatedly apply an "exploration policy," i.e., one that tries to visit state-action pairs whose transition dynamics are still inaccurately modeled. After a polynomial number of iterations, it will deem itself to have modeled enough of the MDP accurately. Then, it will apply an "exploitation policy," which (given the current MDP model) tries to maximize the sum of rewards obtained over time. In the original $E^3$ work (Kearns & Singh, 2002), the algorithm would explicitly use an exploration policy until the model was considered accurate enough, after which it switched to an exploitation policy. In later variants such as (Brafman & Tennenholtz, 2002) this choice of exploration vs. exploitation policy was made less explicitly, but through a reward scheme reminiscent of "optimism in the face of uncertainty," (e.g., Kaelbling, Littman & Moore, 1996). However, the algorithm still tends to end up generating (and using) exploration policies in its initial stage.

To achieve its performance guarantees, the $E^3$-family of algorithms demand that we run exploration policies on the unknown system until we have an accurate model for the entire MDP (or at least for the "reachable" parts of it). The strong bias towards exploration makes the policies generated by the $E^3$-family often unacceptable for running on a real system. Consider for example running $E^3$ on an autonomous helicopter. This would require executing policies that aggressively explore different parts of the state-space, including parts of it that would lead to crashing the helicopter.[1] As a second example, if the system to be controlled is a chemical plant, $E^3$-generated policies may well cause an explosion in the plant through its aggressive exploration of the entire state space. Despite the strong theoretical results, for many robotics and other applications, we do not believe that $E^3$ is a practical algorithm.

In this paper, we consider the apprenticeship learning setting, in which we have available an initial teacher demonstration of the task to be learned. For example, we may

This is the long version of Abbeel and Ng (2005). The main bodies of both papers are identical. This version includes an appendix with complete proofs of all theorems, propositions and lemmas.

[1] Indeed, in our work on an autonomous helicopter flight, our first crash occurred during (manual flight) exploration, when a human pilot was over-aggressive in exploring the boundaries of the flight envelope (moving the control sticks through their extreme ranges), which placed excessive strain on the rotor head assembly and caused it to disintegrate in mid-air.

have a human pilot give us an initial demonstration of helicopter flight. Given this initial training data with which to learn the dynamics, we show that it suffices to only execute exploitation policies (ones that try to do as well as possible, given the current model of the MDP). More specifically, we propose the following algorithm:

1. Have a teacher demonstrate the task to be learned, and record the state-action trajectories of the teacher's demonstration.
2. Use all state-action trajectories seen so far to learn a dynamics model for the system. For this model, find a (near) optimal policy using any reinforcement learning (RL) algorithm.
3. Test that policy by running it on the real system. If the performance is as good as the teacher's performance, stop. Otherwise, *add the state-action trajectories from the (unsuccessful) test to the training set*, and go back to step 2.

Note that the algorithm we described uses a greedy policy with respect to the current estimated model at every iteration. So there is never an explicit exploration step. In practice, exploitation policies tend to be more benign, and thus we believe this is a significantly more palatable algorithm for many applications. For example, unlike $E^3$, this is a procedure that can much more safely and confidently be tried on an autonomous helicopter.[2] Further, if we are designing a controller for a client and each experiment consumes a non-trivial amount of time/resources, we believe it is much more palatable to tell them that the next policy we try will represent our best attempt at solving their problem—i.e., an exploitation policy that represents our current best attempt at controlling the system—rather than that we will be repeatedly running expensive experiments to slowly gather more and more data about the MDP.

We note that the algorithm proposed above also parallels a reasonably common practice in applied control, in which some initial policy is used to collect data and build a model for a simulator. Then, if subsequently a controller is found that works in simulation but not in real-life, the designer tries (usually manually) to adjust the simulator to make it correctly predict the failure of this policy. If machine learning is used to build the simulator, then a natural way to modify the simulator after observing an unsuccessful policy is to add the data obtained from the unsuccessful policy to the training set. Thus, our work can also be viewed as formally analyzing, and thereby attempting to cast light on, the conditions under which a procedure like this can be expected to lead to a good policy.

Previous work has shown the effectiveness of using teacher or expert demonstrations (called apprenticeship learning, also imitation learning, and learning by watching) in various ways for control. Schaal and Atkeson (1994) and Smart and Kaelbling (2000) both give examples where learning is significantly faster when bootstrapping from a teacher. Their methods are somewhat related in spirit, but different in detail from ours (e.g., Smart and Kaelbling, 2000, uses model-free Q-learning, and does not learn the MDP parameters), and had no formal guarantees.

Other examples include Sammut et al. (1992); Kuniyoshi, Inaba & Inoue (1994); Demiris & Hayes (1994); Amit & Mataric (2002); and Pomerleau (1989), which apply supervised learning to learn a parameterized policy from the demonstrations. In these examples, neither the reward function nor the system dynamics need to be specified since a policy is learned directly as a mapping from the states to the actions. This approach has been applied successfully in a variety of applications, but may require careful selection of an appropriate policy class parameterization, and generally lacks strong performance guarantees. Abbeel and Ng (2004) uses the demonstrations to remove the need for explicitly specifying a reward function; there, the system dynamics were assumed to be known.

In what follows, we prove that, with high probability, our algorithm given above terminates with a policy whose performance is comparable to (or better than) the teacher. In the case of discrete state MDPs, the algorithm scales at most polynomially in the number of states. In the case of linearly parameterized dynamical systems, we use a martingale over relative losses to show that the algorithm scales at most polynomially in the dimension of the state space.

This paper is the long version of Abbeel and Ng (2005). The main bodies of both papers are identical. This version includes an appendix with complete proofs of all theorems, propositions and lemmas.

## 2. Preliminaries

A Markov decision process (MDP) is a tuple $(S, \mathcal{A}, T, H, D, R)$, where S is a set of states; $\mathcal{A}$ is a set of actions/inputs; $T = \{P(\cdot|s,a)\}_{s,a}$ is a set of state transition probabilities (here, $P(\cdot|s,a)$ is the state transition distribution upon taking action $a$ in state $s$); $H$ is the horizon time of the MDP, so that the MDP terminates after $H$ steps;[3] $D$ is a distribution over states from which the initial state $s_0$ is drawn; and $R : S \mapsto \mathbb{R}$ is the reward function, which we assume to be non-negative and bounded by $R_{\max}$. A policy $\pi$ is a mapping from states $S$ to a probability distribution over the set of actions $\mathcal{A}$. The utility of a policy $\pi$ in an MDP $M$ is given by $U_M(\pi) = \mathrm{E}[\sum_{t=0}^{H} R(s_t)|\pi, M]$. Here the expectation is over all possible state trajectories in the MDP $M$.

[2] For example, in our autonomous helicopter work, no exploitation policy that we have ever used—out of many dozens—has ever deliberately jerked the helicopter back-and-forth in the manner described in footnote 1.

[3] Any infinite horizon MDP with discounted rewards can be $\epsilon$-approximated by a finite horizon MDP, using a horizon $H_\epsilon = \lceil \log_\gamma(\epsilon(1-\gamma)/R_{\max}) \rceil$.

Specifying an MDP therefore requires specifying each item of the tuple $(S, \mathcal{A}, T, H, D, R)$. In practice, the state transitions probabilities $T$ are usually the most difficult element of this tuple to specify, and must often be learned from data. More precisely, the state space $S$ and action space $\mathcal{A}$ are physical properties of the system being controlled, and thus easily specified. $R$ (and $H$) is typically given by the task specification (or otherwise can be learned from a teacher demonstration, as in Abbeel & Ng, 2004). Finally, $D$ is usually either known or can straightforwardly be estimated from data. Thus, in the sequel, we will assume that $S, \mathcal{A}, H, D$ and $R$ are given, and focus exclusively on the problem of learning the state transition dynamics $T$ of the MDP.

Consider an MDP $M = (S, \mathcal{A}, T, H, D, R)$, and suppose we have some approximation $\hat{T}$ of the transition probabilities. Thus, $\hat{M} = (S, \mathcal{A}, \hat{T}, H, D, R)$ is our approximation to $M$. The Simulation Lemma (stated below) shows that so long as $\hat{T}$ is close to $T$ on states that are visited with high probability by a policy $\pi$, then the utility of $\pi$ in $\hat{M}$ is close to the utility of $\pi$ in $M$. (Related results are also given in Kearns & Singh, 2002; Kearns & Koller, 1999; Kakade, Kearns & Langford, 2003; Brafman & Tennenholtz, 2002.)

**Lemma 1** (Simulation Lemma)**.** *Let any $\epsilon, \eta \geq 0$ be given. Let an MDP $M = (S, \mathcal{A}, T, H, D, R)$ be given. Let $\hat{M} = (S, \mathcal{A}, \hat{T}, H, D, R)$ be another MDP which only differs from $M$ in its transition probabilities. Let $\pi$ be a policy over the state-action sets $S, \mathcal{A}$, so that $\pi$ can be applied to both $M$ and $\hat{M}$. Assume there exists a set of state-action pairs $\overline{SA}_\eta \subseteq S \times \mathcal{A}$ such that the following holds*

$$\begin{array}{ll} \text{(i)} & \forall (s,a) \in \overline{SA}_\eta,\ \ d_{\text{var}}(P(\cdot|s,a), \hat{P}(\cdot|s,a)) \leq \epsilon, \\ \text{(ii)} & P(\{(s_t,a_t)\}_{t=0}^H \subseteq \overline{SA}_\eta | \pi, M) \geq 1 - \eta. \end{array}$$

*(Above, $d_{\text{var}}$ denotes variational distance.[4]) Then we have*

$$|U_M(\pi) - U_{\hat{M}}(\pi)| \leq H^2 \epsilon R_{\max} + \eta H R_{\max}.$$

Consider the special case where every state-action pair $(s, a) \in S \times \mathcal{A}$ satisfies condition (i), in other words, $\overline{SA}_\eta = S \times \mathcal{A}$ and thus condition (ii) is satisfied for $\eta = 0$. Then the Simulation Lemma tells us that accurate transition probabilities are sufficient for accurate policy evaluation. The Simulation Lemma also shows that not necessarily all state-action pairs' transition probabilities need to be accurately modeled: it is sufficient to accurately model a subset of state-action pairs $\overline{SA}_\eta$ such that the probability of leaving this set $\overline{SA}_\eta$ under the policy $\pi$ is sufficiently small.

Let there be some event that has probability bounded away from zero. Suppose we would like to observe that event some minimum number of times in a set of IID experiments. The following lemma allows us to prove bounds on how often we need to repeat the experiment to see that event at least the desired number of times (with high probability).

**Lemma 2.** *Let any $\delta > 0$ and $a > 0$ be given. Let $\{X_i\}_{i=1}^m$ be IID Bernoulli($\phi$) random variables. Then for $\sum_{i=1}^m X_i \geq a$ to hold with probability at least $1 - \delta$, it suffices that $m \geq \frac{2}{\phi}(a + \log \frac{1}{\delta})$.*

## 3. Problem description

The problems we are concerned with in this paper are control tasks that can be described by an MDP $M = (S, \mathcal{A}, T, H, D, R)$. However the system dynamics $T$ are unknown. Everything else in the specification of the MDP is assumed to be known. We consider two specific classes of state-action spaces and transition probabilities, which we will refer to as discrete dynamics and linearly parameterized dynamics respectively.

- Discrete dynamics: The sets $S$ and $\mathcal{A}$ are finite sets. The system dynamics $T$ can be described by a set of transition probabilities $P(s'|s,a)$, which denote the probability of the next-state being $s'$ given the current state is $s$ and the current action is $a$. More specifically we have a multinomial distribution $P(\cdot|s,a)$ over the set of states $S$ for all state-action pairs $(s,a) \in S \times \mathcal{A}$.
- Linearly parameterized dynamics: The sets $S = \mathbb{R}^{n_S}$ and $\mathcal{A} = \mathbb{R}^{n_\mathcal{A}}$ are now continuous. We assume the system obeys the following dynamics:[5]

$$x_{t+1} = A\phi(x_t) + Bu_t + w_t, \tag{1}$$

where $\phi(\cdot) : \mathbb{R}^{n_S} \to \mathbb{R}^{n_S}$. Thus, the next-state is a linear function of some (possibly non-linear) features of the current state (plus noise). This generalizes the familiar LQR model from classical control (Anderson & Moore, 1989) to non-linear settings. For example, the (body-coordinates) helicopter model used in (Ng et al., 2004) was of this form, with a particular choice of non-linear $\phi$, and the unknown parameters $A$ and $B$ were estimated from data. The process noise $\{w_t\}_t$ is IID with $w_t \sim \mathcal{N}(0, \sigma^2 I_{n_S})$. Here $\sigma^2$ is a fixed, known, constant. We also assume that $\|\phi(s)\|_2 \leq 1$ for all $s$, and that the inputs $u_t$ satisfy $\|u_t\|_2 \leq 1$.[6]

## 4. Algorithm

Let $\pi_T$ be the policy of a teacher. Although it is natural to think of $\pi_T$ as a good policy for the MDP, we do not assume this to be the case. Let any $\alpha > 0$ be given. Our algorithm (with parameters $N_T$ and $k_1$) is as follows:

[4] Let $P(\cdot), Q(\cdot)$ be two probability distributions over a set $\mathcal{X}$, then the variational distance $d_{\text{var}}(P,Q)$ is defined as follows: $d_{\text{var}}(P,Q) = \frac{1}{2}\int_{x \in \mathcal{X}} |P(x) - Q(x)| dx$.

[5] We chose to adhere to the most commonly used notation for continuous systems. I.e., states are represented by $x$, inputs by $u$ and the system matrices by $A$ and $B$. We use script $\mathcal{A}$ for the set of actions and standard font $A$ for the system matrix.

[6] The generalizations to unknown $\sigma^2$, to non-diagonal noise covariances, and to non-linear features over the inputs ($B\psi(u_t)$ replacing $Bu_t$) offer no special difficulties.

1. Run $N_T$ trials under the teacher's policy $\pi_T$. Save the state-action trajectories encountered during these trials. Compute $\hat{U}_M(\pi_T)$—an estimate of the utility of the teacher's policy $\pi_T$ for the real system $M$—by averaging the sum of rewards accumulated in each of the $N_T$ trials. Initialize $i = 1$.
2. Using all state-action trajectories saved so far, estimate the system dynamics $T$ using maximum likelihood estimation for the discrete dynamics case, and regularized linear regression for the linearly parameterized dynamics case (as described below). Call the estimated dynamics $\hat{T}^{(i)}$.
3. Find a $\alpha/8$ optimal policy[7] for the MDP $\hat{M}^{(i)} = (S, \mathcal{A}, \hat{T}^{(i)}, H, D, R)$. Call this policy $\pi^{(i)}$.
4. Evaluate the utility of the policy $\pi^{(i)}$ on the real system $M$. More specifically, run the policy $\pi^{(i)}$ for $k_1$ trials on the system $M$. Let $\hat{U}_M(\pi^{(i)})$ be the average sum of rewards accumulated in the $k_1$ trials. Save the state-action trajectories encountered during these trials.
5. If $\hat{U}_M(\pi^{(i)}) \geq \hat{U}_M(\pi_T) - \alpha/2$, return $\pi^{(i)}$ and exit. Otherwise set $i = i + 1$ and go back to step 2.

In the $i^{\text{th}}$ iteration of the algorithm, a policy is found using an estimate $\hat{T}^{(i)}$ of the true system dynamics $T$. For the discrete dynamics, the estimate used in the algorithm is the maximum likelihood estimates for each of the multinomial distributions $P(\cdot|s, a)$. For the linearly parameterized dynamics, the model parameters $A, B$ are estimated via regularized linear regression. In particular the $k^{\text{th}}$ rows of $A$ and $B$ are estimated by[8] $\arg\min_{A_{k,:},B_{k,:}} \sum_j (x^{(j)}_{\text{next}} - (A_{k,:}\phi(x^{(j)}_{\text{curr}}) + B_{k,:}u^{(j)}_{\text{curr}}))^2 + \frac{1}{\kappa^2}(\|A_{k,:}\|_2^2 + \|B_{k,:}\|_2^2)$, where $j$ indexes over all state-action-state triples $\{(x^{(j)}_{\text{curr}}, u^{(j)}_{\text{curr}}, x^{(j)}_{\text{next}})\}_j$ occurring after each other in the trajectories observed for the system.

## 5. Main theorem

The following theorem gives performance and running time guarantees for the algorithm described in Section 4.[9]

**Theorem 3.** *Let an MDP $M = (S, \mathcal{A}, T, H, D, R)$ be given, except for its transition probabilities $T$. Let the system either be a discrete dynamics system or a linearly parameterized dynamical system as defined in Section 3. Let any $\alpha > 0, \delta > 0$ be given. Let $\pi_T$ be the teacher's policy, and let $\pi$ be the policy returned by the algorithm defined above. Let $N$ denote the number of iterations of the main loop of the algorithm until the exit condition is met. Let $\mathcal{T} = (H, R_{\max}, |S|, |\mathcal{A}|)$ for the discrete case, and let $\mathcal{T} = (H, R_{\max}, n_S, n_{\mathcal{A}}, \|A\|_{\text{F}}, \|B\|_{\text{F}})$ for the linearly parameterized dynamics case. Then for*

$$U_M(\pi) \;\geq\; U_M(\pi_T) - \alpha, \qquad (2)$$

$$N \;=\; O(\text{poly}(\tfrac{1}{\alpha}, \tfrac{1}{\delta}, \mathcal{T})) \qquad (3)$$

*to hold with probability at least $1 - \delta$, it suffices that*

$$N_T \;=\; \Omega(\text{poly}(\tfrac{1}{\alpha}, \tfrac{1}{\delta}, \mathcal{T})), \qquad (4)$$

$$k_1 \;=\; \Omega(\text{poly}(\tfrac{1}{\alpha}, \tfrac{1}{\delta}, \mathcal{T})). \qquad (5)$$

Note that Eqn. (2) follows from the termination condition of our algorithm and assuming we choose $k_1$ and $N_T$ large enough such that the utilities of the policies $\{\pi^{(i)}\}_i$ and $\pi_T$ are sufficiently accurately evaluated in $M$.

The proof of this theorem is quite lengthy, and will make up most of the remainder of this paper. We now give a high-level sketch of the proof ideas. Our proof is based on showing the following two facts:

1. After we have collected sufficient data from the teacher, the estimated model is accurate for evaluating the utility of the teacher's policy in every iteration of the algorithm. (Note this does not merely require that the model has to be accurate after the $N_T$ trials under the teacher's policy, but also has to stay accurate after extra data is collected from testing the policies $\{\pi^{(i)}\}_i$.)
2. One can visit inaccurately modeled state-action pairs only a "small" number of times until all state-action pairs are accurately modeled.

We now sketch how these two facts can be proved. After we have collected sufficient data from the teacher, the state-action pairs that are visited often under the teacher's policy are modeled well. From the Simulation Lemma we know that an accurate model of the state-action pairs visited often under the teacher's policy is sufficient for accurate evaluation of the utility of the teacher's policy. This establishes (1.). Every time an inaccurate state-action pair is visited, the data collected for that state-action pair can be used to improve the model. However the model can be improved only a "small" number of times until it is accurate for all state-action pairs. This establishes (2.).

We now explain how these two facts can be used to bound the number of iterations of our algorithm. Consider the policy $\pi^{(i)}$ found in iteration $i$ of the algorithm. This policy $\pi^{(i)}$ is the optimal policy[10] for the current model. When

[7] A policy $\pi_1$ is an $\epsilon$-optimal policy for an MDP $M$ if $U_M(\pi_1) \geq \max_\pi U_M(\pi) - \epsilon$.

[8] We use matlab-like notation. $A_{k,:}$ denotes the $k^{\text{th}}$ row of $A$.

[9] The performance guarantees in the theorem are stated with respect to the teacher's demonstrated performance. However, the proof requires only that the initial dynamical model be accurate for at least one good policy. Thus, for example, it is sufficient to observe a few good teacher demonstrations along with many bad demonstrations (ones generated via a highly sub-optimal policy); or even only bad demonstrations that manage to visit good parts of the state space.

[10] For simplicity of exposition in this informal discussion, we assume $\pi^{(i)}$ is optimal, rather than near-optimal. The formal results in this paper do not use this assumption.

finding this policy $\pi^{(i)}$ in the model we could have chosen the teacher's policy. So the policy $\pi^{(i)}$ performs at least as well as the teacher's policy in the current model. Now if in the real system the utility of the policy $\pi^{(i)}$ is significantly lower than the teacher's utility (which is the case as long as the algorithm continues), then the model incorrectly predicted that $\pi^{(i)}$ was better than the teacher's policy. From (1.) we have that the model correctly evaluates the utility of the teacher's policy. Thus the model must have evaluated the policy $\pi^{(i)}$ inaccurately. Using the (contrapositive of) the Simulation Lemma, we get that the policy $\pi^{(i)}$ must be visiting (with probability bounded away from 0) state-action pairs that are not very accurately modeled. So when running the policy $\pi^{(i)}$ we can collect training data that allow us to improve the model. Now from (2.) we have that visiting inaccurately modeled state-action pairs can only happen a small number of times until the dynamics is learned, thus giving us a bound on the number of iterations of the algorithm.

The theorem will be proved for the discrete dynamics case in Section 6 and for the linearly parameterized dynamics case in Section 7.

## 6. Discrete state space systems

In this section we prove Theorem 3 for the case of discrete dynamics.

The Hoeffding inequality gives a bound on the number of samples that are sufficient to estimate the expectation of a (bounded) random variable. In our algorithm, we want to guarantee that the model is accurate (for the teacher's policy) not only when we have seen the samples from the teacher, but also any time after additional samples are collected. The following lemma, which is a direct consequence of Hoeffding's inequality (as shown in the long version), gives such a bound.

**Lemma 4.** *Let any $\epsilon > 0, \delta > 0$ be given. Let $X_i$ be IID $k$-valued multinomial random variables, with distribution denoted by $P$. Let $\hat{P}_n$ denote the $n$ sample estimate of $P$. Then for $\max_{n \geq N} d_{\rm var}(P(\cdot), \hat{P}_n(\cdot)) \leq \epsilon$ to hold with probability $1-\delta$, it suffices that $N \geq \frac{k^2}{4\epsilon^2} \log \frac{k^2}{\delta \epsilon}$.*

Lemma 4 will serve two important purposes. In the proof of Lemma 5 it is used to bound the number of trajectories needed under the teacher's policy to guarantee that frequently visited state-action pairs are accurately modeled in all models $\{\hat{M}^{(i)}\}_i$. This corresponds to establishing Fact (1.) of the proof outline in Section 5. In the proof of Lemma 6 it is used to bound the total number of times a state-action pair can be visited that is not accurately modeled. This latter fact corresponds exactly to establishing Fact (2.) of the proof outline in Section 5.[11]

**Lemma 5.** *Let any $\alpha > 0, \delta > 0$ be given. Assume we use the algorithm as described in Section 4. Let $N_T$ satisfy the following condition $N_T \geq \frac{4096|S|^3|\mathcal{A}|H^5 R_{\rm max}^3}{\alpha^3} \log \frac{32 H^2 R_{\rm max}|S|^3|\mathcal{A}|}{\delta\alpha}$. Then with probability $1-\delta$ we have that $\forall i \geq N_T \;\; |U_{\hat{M}^{(i)}}(\pi_T) - U_M(\pi_T)| \leq \alpha/8$.*

*Proof (sketch).* Let $\epsilon > 0, \eta > 0$. Let $SA_\xi \subseteq S \times A$ be the set of state-action pairs such that the probability of seeing any specific state-action pair $(s,a) \in SA_\xi$ under the policy $\pi_T$ in a single trial of duration $H$ is at least $\frac{\eta}{|S||\mathcal{A}|}$. From Lemma 4 and Lemma 2 we have that for any $(s,a) \in SA_\xi$ for

$$\forall i \geq N_T \;\; d_{\rm var}(P(\cdot|s,a), \hat{P}^{(i)}(\cdot|s,a)) \leq \epsilon \qquad (6)$$

to hold with probability $1 - \delta' - \delta''$, it is sufficient to have

$$N_T \geq \tfrac{2|S||\mathcal{A}|}{\eta}\left(\tfrac{|S|^2}{4\epsilon^2} \log \tfrac{|S|^2}{\delta'\epsilon} + \log \tfrac{1}{\delta''}\right). \qquad (7)$$

Taking a union bound over all state-action pairs $(s,a) \in SA_\xi$ (note $|SA_\xi| \leq |S||\mathcal{A}|$) gives that for Eqn. (6) to hold for all $(s,a) \in SA_\xi$ with probability $1 - |S||\mathcal{A}|\delta' - |S||\mathcal{A}|\delta''$, it suffices that Eqn. (7) is satisfied. We also have from the definition of $SA_\xi$ that $\mathrm{P}(\{(s_t,a_t)\}_{t=0}^H \subseteq \overline{SA}_\xi|\pi_T) \geq 1-\eta$. Thus the Simulation Lemma gives us that

$$\forall i \;\; |U_{\hat{M}^{(i)}}(\pi_T) - U_M(\pi_T)| \leq H^2 \epsilon R_{\rm max} + \eta H R_{\rm max}.$$

Now choose $\epsilon = \frac{1}{2}\frac{\alpha/8}{H^2 R_{\rm max}}$, $\eta = \frac{1}{2}\frac{\alpha/8}{H R_{\rm max}}$ and $\delta' = \delta'' = \frac{\delta}{2|S||\mathcal{A}|}$ to get the lemma. □

Lemma 5 shows that, after having seen the teacher sufficiently often, the learned model will be accurate for evaluating the utility of the teacher's policy. Moreover, no later data collection (no matter under which policy the data is collected) can make the model inaccurate for evaluation of the utility of the teacher's policy. I.e., $U_{\hat{M}^{(i)}}(\pi_T)$ will be close to $U_M(\pi_T)$ for all $i$.

**Lemma 6.** *Let any $\alpha > 0, \delta > 0$ be given. Let*

$$\begin{aligned} N_{\rm ubound} = {} & \tfrac{32 H R_{\rm max}}{\alpha}(\log \tfrac{4}{\delta} + \\ & \tfrac{16^2 H^4 R_{\rm max}^2 |S|^3|\mathcal{A}|}{4\alpha^2} \log \tfrac{64 H^2 R_{\rm max}|S|^3|\mathcal{A}|}{\alpha\delta}). \end{aligned} \qquad (8)$$

*Assume in the algorithm described in Section 4 we use*

$$\begin{aligned} k_1 & \geq \tfrac{16^2 H^2 R_{\rm max}^2}{2\alpha^2} \log \tfrac{8 N_{\rm ubound}}{\delta}, & (9) \\ N_T & \geq \tfrac{4096|S|^3|\mathcal{A}|H^5 R_{\rm max}^3}{\alpha^3} \log \tfrac{256 H^2 R_{\rm max}|S|^3|\mathcal{A}|}{\delta\alpha}. & (10) \end{aligned}$$

*Let $N$ denote the number of iterations of the algorithm until it terminates. Then we have that with probability $1-\delta$ the following hold*

$$\begin{aligned} &\text{(i)} && N \leq N_{ubound}, & (11) \\ &\text{(ii)} && \forall i = 1:N \;\; |U_{\hat{M}^{(i)}}(\pi_T) - U_M(\pi_T)| \leq \tfrac{\alpha}{8}, & (12) \\ &\text{(iii)} && \forall i = 1:N \;\; |\hat{U}_M(\pi^{(i)}) - U_M(\pi^{(i)})| \leq \tfrac{\alpha}{16}, & (13) \\ &\text{(iv)} && |\hat{U}_M(\pi_T) - U_M(\pi_T)| \leq \tfrac{\alpha}{16}. & (14) \end{aligned}$$

[11] Fact (2.) follows completely straightforwardly from Lemma 4, so rather than stating it as a separate lemma, we will instead derive it within the proof of Lemma 6.

*Proof (sketch).* From Lemma 5 and from the Hoeffding inequality we have that for Eqn. (12), (13) and (14) to hold (for all $i \leq N_{\rm ubound}$) with probability $1 - \frac{\delta}{2}$, it suffices that Eqn. (10) and Eqn. (9) are satisfied.

Now since the algorithm only exits in iteration $N$, we must have for all $i = 1 : N - 1$ that

$$\hat{U}_M(\pi^{(i)}) < \hat{U}_M(\pi_T) - \alpha/2. \tag{15}$$

Combining Eqn. (15), (12), (13) and (14) and the fact that $\pi^{(i)}$ is $\alpha/8$-optimal for $\hat{M}^{(i)}$ we get

$$\forall i = 1 : N - 1 \;\; U_{\hat{M}^{(i)}}(\pi^{(i)}) \geq U_M(\pi^{(i)}) + \alpha/8. \tag{16}$$

In words: when the algorithm continues (in iterations $i = 1 : N - 1$), the model overestimated the utility of $\pi^{(i)}$. Using the contrapositive of the Simulation Lemma with $\epsilon = \frac{1}{2}\frac{\alpha/8}{H^2 R_{\max}}$ we get that for all $i = 1 : N - 1$ the policy $\pi^{(i)}$ must be visiting a state-action pair $(s, a)$ that satisfies

$$d_{\rm var}(P(\cdot|s,a), \hat{P}^{(i)}(\cdot|s,a)) > \frac{\alpha}{16H^2R_{\max}} \tag{17}$$

with probability at least $\frac{\alpha}{16HR_{\max}}$. From Lemma 2 and Lemma 4 we get that if the algorithm had run for a number of iterations $N_{\rm ubound}$ then with probability $1 - \frac{\delta}{2}$ all state-actions pairs would satisfy

$$d_{\rm var}(P(\cdot|s,a), \tilde{P}^{(N)}(\cdot|s,a)) \leq \frac{\alpha}{16H^2R_{\max}}. \tag{18}$$

On the other hand we showed above that if the algorithm does not exit in iteration $i$, there must be a state-action pair satisfying Eqn. (17), which contradicts Eqn. (18). Thus $N_{\rm ubound}$ gives an upper bound on the number of iterations of the algorithm. □

The proof of Theorem 3 for the case of discrete dynamics is a straightforward consequence of Lemma 6.

*Proof of Theorem 3 for discrete dynamics.* First note that the conditions on $N_T$ and $k_1$ of Lemma 6 are satisfied in Theorem 3. So Lemma 6 proves the bound on the number of iterations as stated in Eqn. (3). Now it only remains to prove that at termination, Eqn. (2) holds. We have from the termination condition that $\hat{U}(\pi) \geq \hat{U}(\pi_T) - \alpha/2$. Now using Eqn. (13) and Eqn. (14) we get $U_\pi \geq U_{\pi_T} - \frac{5}{8}\alpha$, which implies Eqn. (2). □

## 7. Linearly parameterized dynamical systems

In this section we prove Theorem 3 for the case of linearly parameterized dynamics described in Eqn. (1). As pointed out in Section 5, the performance guarantee of Eqn. (2) follows from the termination condition of our algorithm and assuming we choose $k_1$ and $N_T$ large enough such that the utility of the policies $\{\pi^{(i)}\}_i$ and $\pi_T$ are sufficiently accurately evaluated in $M$. This leaves us to prove the bound on the number of iterations of the algorithm. As explained more extensively in Section 5, there are two main parts to this proof. In Section 7.2 we establish the first part: the estimated model is accurate for evaluating the utility of the teacher's policy in every iteration of the algorithm. In Section 7.3 we establish the second part: one can visit inaccurately modeled states only a "small" number of times (since every such visit improves the model). In Section 7.4 we combine these two results to prove Theorem 3 for the case of linearly parameterized dynamical systems.

### 7.1. Preliminaries

The following proposition will allow us to relate accuracy of the expected value of the next-state to variational distance for the next-state distribution. This will be important for using the Simulation Lemma, which is stated in terms of variational distance.

**Proposition 7.** *We have*

$$d_{\rm var}(\mathcal{N}(\mu_1, \sigma^2 I_n), \mathcal{N}(\mu_2, \sigma^2 I_n)) \leq \frac{1}{\sqrt{2\pi}\sigma}\|\mu_1 - \mu_2\|_2.$$

### 7.2. Accuracy of the model for the teacher's policy

Given a set of state-action trajectories, the system matrices $A, B$ are estimated by solving $n_S$ separate regularized linear regression problems, one corresponding to each row of $A$ and $B$. After appropriately relabeling variables and data, each of these regularized linear regression problems is of the form

$$\min_\theta \sum_i (y^{(i)} - \theta^\top z^{(i)})^2 + \frac{\|\theta\|_2^2}{\kappa^2}. \tag{19}$$

Here $\theta \in \mathbb{R}^{n_S + n_\mathcal{A}}$ corresponds to a row in $A$ and $B$, and the norm bounds on $u$ and $\phi(x)$ result in $\|z\|_2 \leq \sqrt{2}$. The relabeled data points are kept in the same order as they were collected. The training data collected from the teacher's demonstration is indexed from 1 to $m = N_T H$. The additional training data collected when testing the policies $\{\pi^{(j)}\}_{j=1}^N$ is indexed from $m + 1$ to $\tilde{m} = N_T H + k_1 N H$. The data is generated according to a true model $M$ as described in Section 4. In the notation of Eqn. (19), this means there is some $\theta^*$ such that

$$\forall i \;\; y^{(i)} = \theta^{*\top} z^{(i)} + w^{(i)}, \tag{20}$$

where the $\{w^{(i)}\}_i$ are IID, with $w^{(i)} \sim \mathcal{N}(0, \sigma^2)$. The data generation process that we just described will be referred to as "data generated according to Eqn. (20)" from here on. Note that the training data $\{z^{(i)}\}_i$ in this setup are *non-IID*. The teacher's policy $\pi_T$ induces a distribution over states $x_t$ and inputs $u_t$ at all times $t$. However these distributions need not be the same for different times $t$, making the data non-IID. Moreover, the training data indexed from $m+1$ to $\tilde{m}$ is obtained from various policies and the resulting data generation process is very difficult to model. As a consequence, our analysis will consider the worst-case scenario where an adversary can choose the additional training data indexed from $m+1$ to $\tilde{m}$.

For $1 \leq k \leq N_T H + k_1 NH$ let the following equations define $\hat{\theta}^{(k)}$ and $\mathrm{loss}^{(k)}(\theta)$:

$$\begin{aligned} \mathrm{loss}^{(k)}(\theta) &= \textstyle\sum_{i=1}^{k}(y^{(i)} - \theta^\top z^{(i)})^2 + \frac{1}{\kappa^2}\|\theta\|_2^2, \\ \hat{\theta}^{(k)} &= \arg\min_\theta \ \mathrm{loss}^{(k)}(\theta). \end{aligned} \tag{21}$$

The following lemma establishes that a "small" number of samples from the teacher's policy $\pi_T$ is sufficient to guarantee an accurate model $\hat{\theta}^{(k)}$ for all time steps $k = N_T H$ to $N_T H + k_1 NH$.

**Lemma 8.** *Let any $\delta > 0, \epsilon > 0, \eta > 0$ be given. Consider data $\{y^{(i)}, z^{(i)}\}_{i=1}^{N_T H + k_1 NH}$ generated as described in Eqn. (20). Let $\{\hat{\theta}^{(k)}\}_k$ be defined as in Eqn. (21). Let $\{\tilde{y}^{(t)}, \tilde{z}^{(t)}\}_{t=1}^{H}$ be data generated from one trial under $\pi_T$ (and appropriately relabeled as described in paragraph above). Then for*

$$P(\max_{t \in 1:H} |\theta^\top \tilde{z}^{(t)} - \theta^{*\top} \tilde{z}^{(t)}| > \epsilon) \leq \eta \tag{22}$$

*to hold with probability $1-\delta$ for all $\theta \in \{\hat{\theta}^{(k)}\}_{k=N_T H}^{N_T H + k_1 NH}$, it suffices that*

$$N_T = \Omega\left(\mathrm{poly}(\tfrac{1}{\epsilon}, \tfrac{1}{\eta}, \tfrac{1}{\delta}, H, \|\theta^*\|_2, n_S, n_{\mathcal{A}}, k_1, N)\right).$$

If $\theta$ satisfies Eqn. (22) then it is accurate for data generated under the teacher's policy and we refer to it as accurate in the discussion below; otherwise it is referred to as inaccurate. We now sketch the key ideas in the proof of Lemma 8. A full proof is provided in the appendix. The proof proceeds in four steps.

**Step 1.** For any inaccurate parameter $\theta$ we establish that with high probability the following holds

$$\mathrm{loss}^{(N_T H)}(\theta) > \mathrm{loss}^{(N_T H)}(\theta^*) + \Omega(N_T). \tag{23}$$

I.e., the true parameter $\theta^*$ outperforms an inaccurate parameter $\theta$ by a margin of $\Omega(N_T)$ after seeing $N_T$ trajectories from the teacher. The key idea is that the expected value of the loss difference $\mathrm{loss}^{(N_T H)}(\theta) - \mathrm{loss}^{(N_T H)}(\theta^*)$ is of order $N_T$ for inaccurate $\theta$. Our proof establishes the concentration result for this non-IID setting by looking at a martingale over the differences in loss at every step and uses Azuma's inequality to prove the sum of these differences is close to its expectated value with high probability.

**Step 2.** Let $\mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta) = \sum_{i=N_T H+1}^{k}(y^{(i)} - \theta^\top z^{(i)})^2$ be the additional loss incurred over the additional data points $\{z^{(i)}\}_{i=N_T H+1}^{k}$. We establish that for any $-a < 0$,

$$P(\exists k > N_T H : \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta) < \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta^*) - a) \leq \exp(-\frac{a}{\sigma^2}).$$

In words, the probability of $\theta$ ever outperforming $\theta^*$ by a margin $a$ on the additional data is exponentially small in $a$. The proof considers the random walk $\{Z_k\}_k$

$$Z_k = \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta) - \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta^*).$$

Crudely speaking we exploit the fact that no matter how an adversary chooses each additional data point $z^{(i)}$ as a function of the history up to time $i-1$, the random walk $\{Z_k\}_k$ has a positive bias. More precisely, we use the Optional Stopping Theorem on the martingale $Y_k = \exp(\frac{-1}{2\sigma^2} Z_k)$.[12]

**Step 3.** Let $\theta$ be an inaccurate parameter. From Step 1 we have that the optimal $\theta^*$ outperforms $\theta$ by a margin $\Omega(N_T)$ after having seen the initial data points $\{z^{(i)}, y^{(i)}\}_{i=1}^{N_T H}$. Step 2 says that the probability for $\theta$ to ever make up for this margin $\Omega(N_T)$ is exponentially small in $N_T$. Our proof combines these two results to show that a "small" number of samples $N_T$ from the teacher is sufficient to guarantee (with high probability) that $\theta^*$ has a smaller loss than $\theta$ in every iteration, and thus $\theta \notin \{\hat{\theta}^{(k)}\}_{k=N_T H}^{N_T H + k_1 NH}$.

**Step 4.** Our proof uses a covering argument to extend the result that $\theta \notin \{\hat{\theta}^{(k)}\}_{k=N_T H}^{N_T H + k_1 NH}$ for one specific inaccurate $\theta$ from Step 3 to hold for all inaccurate parameters $\theta$ simultaneously. As a consequence, the estimated parameters $\hat{\theta}^{(k)}$ throughout all iterations $k$ $(N_T H \leq k \leq N_T H + k_1 NH)$ must be accurate. Which establishes Lemma 8.

**Theorem 9.** *Let any $\delta > 0, \alpha > 0$ be given. Let $\{\hat{M}^{(i)}\}_{i=1}^{N}$ be the models estimated throughout $N$ iterations of the algorithm for the linearly parameterized dynamics case, as described in Section 4. Then for $|U_{\hat{M}^{(i)}}(\pi_T) - U_M(\pi_T)| \leq \alpha$ to hold for all $i \in 1 : N$ with probability $1-\delta$, it suffices that $N_T = \Omega\left(\mathrm{poly}(\frac{1}{\alpha}, \frac{1}{\delta}, H, R_{\max}, \|A\|_{\mathrm{F}}, \|B\|_{\mathrm{F}}, n_S, n_{\mathcal{A}}, k_1, N)\right)$.*

*Proof (idea).* From Prop. 7 and Lemma 8 we conclude that the estimated models $\{\hat{M}^{(i)}\}_{i=1}^{N}$ are close to the true model in variational distance with high probability for states visited under the teacher's policy. Using the Simulation Lemma gives the resulting accuracy of utility evaluation. □

Theorem 9 shows that a "small" number of samples from the teacher's policy $\pi_T$ is sufficient to guarantee accurate models $\hat{M}_i^{(i)}$ throughout all iterations of the algorithm. An accurate model here means that the utility of the teacher's policy $\pi_T$ is accurately evaluated in that model, i.e., $U_{\hat{M}^{(i)}}(\pi_T)$ is close to $U_M(\pi_T)$.

### 7.3. Bound on the number of inaccurate states visits

Based on the online learning results for regularized linear regression in Kakade and Ng (2005), we can show the following result.

[12] **Definition**(Martingale.) Let $(\Omega, \mathcal{F}, P)$ be a probability space with a a filtration $\mathcal{F}_0, \mathcal{F}_1, \cdots$. Suppose that $X_0, X_1, \cdots$ are random variables such that for all $i \geq 0$, $X_i$ is $\mathcal{F}_i$-measurable. The sequence $X_0, X_1, \cdots$ is a martingale provided, for all $i \geq 0$, we have that $\mathrm{E}[X_{i+1}|\mathcal{F}_i] = X_i$. Due to space constraints we can not expand on these concepts here. We refer the reader to, e.g., (Durrett, 1995; Billingsley, 1995; Williams, 1991), for more details on martingales and stopping times.

**Lemma 10.** *Let any $\mu > 0, \delta > 0$ be given. For the algorithm described in Section 4 we have with probability $1 - \delta$ that the number of times a state-action pair $(x, u)$ is encountered such that $\|(A\phi(x) + Bu) - (\hat{A}^{(i)}\phi(x) + \hat{B}^{(i)}u)\|_2 > \mu$ is bounded by $N_\mu = O(k_1\sqrt{k_1 N}(\log k_1 N)^3 \text{poly}(\|A\|_F, \|B\|_F, n_S, n_{\mathcal{A}}, \log\frac{1}{\delta}, H, \frac{1}{\mu}))$.*

Lemma 10 is proved in the appendix. Lemma 10 is key to proving the bound on the number of iterations in the algorithm.

### 7.4. Proof of Theorem 3 for linearly parameterized dynamical systems

*Proof (rough sketch).* The conditions in Eqn. (4), (5) ensure that $\hat{U}_M(\pi_T), \{\hat{U}_M(\pi^{(i)}\}_i$ are accurately evaluated with high probability (by the Hoeffding inequality) and Eqn. (4) also ensures that $\{U_{\hat{M}^{(i)}}(\pi_T)\}_i$ are accurate estimates of $U_M(\pi_T)$ (by Theorem 9). Using the Simulation Lemma and the same reasoning as in the proof of Lemma 6 gives us that if the algorithm does not terminate in step 4 of the algorithm, then the policy $\pi^{(i)}$ must be visiting a state-action pair $(x, u)$ that satisfies

$$d_{\text{var}}(P(\cdot|x, u), \hat{P}^{(i)}(\cdot|x, u)) > \frac{\alpha}{16H^2R_{\max}} \qquad (24)$$

with probability at least $\frac{\alpha}{16HR_{\max}}$. If $(x, u)$ satisfies Eqn. (24) then we must have (using Prop. 7) that

$$\|(A\phi(x) + Bu) - (\hat{A}^{(i)}\phi(x) + \hat{B}^{(i)}u)\|_2 > \frac{\sqrt{2\pi}\sigma\alpha}{16H^2R_{\max}}.$$

From Lemma 10 this can happen only

$$\begin{aligned} &O(k_1\sqrt{k_1 N}(\log k_1 N)^3\text{poly}(\|A\|_F, \|B\|_F, n_S, n_{\mathcal{A}}, \\ &\quad \log\frac{1}{\delta}, H, R_{\max}, \frac{1}{\alpha})) \qquad (25) \end{aligned}$$

times in $N$ iterations of the algorithm. On the other hand, if the algorithm continues, we have from above that such an error must be encountered (with high probability)

$$\Omega(\frac{\alpha}{HR_{\max}}N) \qquad (26)$$

times. Note that the lower bound on the number of state-action pairs encountered with large error in Eqn. (26) grows faster in $N$ than the upper bound in Eqn. (25).[13] Once the lower bound is larger than the upper bound we have a contradiction. Thus from Eqn. (26) and (25) we can conclude that after a number of iterations as given by Eqn. (3) the algorithm must have terminated with high probability. Also, since we chose $k_1, N_T$ such that $\{\hat{U}_M(\pi^{(i)}\}_i$ and $\hat{U}_M(\pi_T)$ are accurately evaluated, Eqn. (2) holds when the algorithm terminates. □

[13] In this proof sketch we ignore a dependence of $k_1$ on $N$. See the long version for a formal proof.

## References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proc. ICML*.

Abbeel, P., & Ng, A. Y. (2005). Exploration and apprenticeship learning in reinforcement learning. *Proc. ICML*.

Amit, R., & Mataric, M. (2002). Learning movement sequences from demonstration. *Proc. ICDL*.

Anderson, B., & Moore, J. (1989). *Optimal control: Linear quadratic methods*. Prentice-Hall.

Bertsekas, D. P., & Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Athena Scientific.

Billingsley, P. (1995). *Probability and Measure*. Wiley Interscience.

Brafman, R. I., & Tennenholtz, M. (2002). R-max, a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*.

Demiris, J., & Hayes, G. (1994). A robot controller using learning by imitation.

Durrett, R. (1995). *Probability:Theory and Examples*. Duxbury Press.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *JAIR*.

Kakade, S., Kearns, M., & Langford, J. (2003). Exploration in metric state spaces. *Proc. ICML*.

Kakade, S., & Ng, A. Y. (2005). Online bounds for Bayesian algorithms. *NIPS 17*.

Kearns, M., & Koller, D. (1999). Efficient reinforcement learning in factored MDPs. *Proc. IJCAI*.

Kearns, M., & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning journal*.

Kuniyoshi, Y., Inaba, M., & Inoue, H. (1994). Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *T-RA, 10,* 799–822.

Ng, A. Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E., & Liang, E. (2004). Inverted autonomous helicopter flight via reinforcement learning. *International Symposium on Experimental Robotics*.

Pomerleau, D. (1989). Alvinn: An autonomous land vehicle in a neural network. *NIPS 1*.

Sammut, C., Hurst, S., Kedzier, D., & Michie, D. (1992). Learning to fly. *Proc. ICML*.

Schaal, S., & Atkeson, C. G. (1994). Robot learning by nonparametric regression. *Proc. IROS*.

Smart, W. D., & Kaelbling, L. P. (2000). Practical reinforcement learning in continuous spaces. *Proc. ICML*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. MIT Press.

Williams, D. (1991). *Probability with Martingales*. Cambridge Mathematical Textbooks.

## A. Some properties of the variational distance

In this section we state some known results involving the variational distance between probability distributions. (We include proofs to keep the paper self-contained.) We prove (and sometimes state) the propositions for probability distributions over a discrete domain only. The proofs (and statements) for continuous domains are very similar.

**Proposition 11.** *Let $Q, Q^*$ be probability distributions over a domain $\mathcal{X}$, let $f$ be a bounded random variable over $\mathcal{X}$. Then*

$$|E_Q f - E_{Q^*} f| \leq (\sup_{x \in \mathcal{X}} f(x) - \inf_{x \in \mathcal{X}} f(x)) d_{\mathrm{var}}(Q, Q^*).$$

*Proof.* Let $c = \inf_{x \in \mathcal{X}} f(x)$. Then we have

$$\begin{aligned}
& E_Q f - E_{Q^*} f \\
&= E_Q(f-c) - E_{Q^*}(f-c) \\
&= \sum_{x:Q(x)>Q^*(x)} (f(x)-c)(Q(x)-Q^*(x)) \\
&\quad + \sum_{x:Q(x)\leq Q^*(x)} (f(x)-c)(Q(x)-Q^*(x)) \\
&\leq \sum_{x:Q(x)>Q^*(x)} (f(x)-c)(Q(x)-Q^*(x)) \\
&\leq \sup_{x\in\mathcal{X}} (f(x)-c) \sum_{x:Q(x)>Q^*(x)} (Q(x)-Q^*(x)) \\
&= (\sup_{x\in\mathcal{X}} f(x) - \inf_{x\in\mathcal{X}} f(x)) d_{\mathrm{var}}(Q,Q^*). \qquad (27)
\end{aligned}$$

Here we used in order: adding and subtracting $c$; splitting the summation into positive and negative terms; dropping the negative terms from the summation; $f(x)$ is bounded by $\sup_{x\in\mathcal{X}} f(x)$; definition of $d_{\mathrm{var}}$[14] and $c$. The same argument with roles of $Q$ and $Q^*$ interchanged gives us:

$$E_Q f - E_{Q^*} f \leq (\sup_{x\in\mathcal{X}} f(x) - \inf_{x\in\mathcal{X}} f(x)) d_{\mathrm{var}}(Q,Q^*). \qquad (28)$$

Eqn. (27) and (28) combined prove the proposition. □

**Proposition 12.** *Let $Q_0(\cdot)$ and $Q_0^*(\cdot)$ be probability distributions over a domain $\mathcal{X}$. Let $\forall x \in \mathcal{X}$ $P(\cdot|x)$ be a probability distribution over $\mathcal{X}$. Let $Q_1(\cdot) = \sum_{x_0\in\mathcal{X}} P(\cdot|x_0)Q_0(x_0)$ and let $Q_1^*(\cdot) = \sum_{x_0\in\mathcal{X}} P(\cdot|x_0)Q_0^*(x_0)$, then*

$$d_{\mathrm{var}}(Q_1(\cdot), Q_1^*(\cdot)) \leq d_{\mathrm{var}}(Q_0(\cdot), Q_0^*(\cdot)).$$

*Proof.* We have

$$\begin{aligned}
& d_{\mathrm{var}}(Q_1(\cdot), Q_1^*(\cdot)) \\
&= \frac{1}{2} \sum_{x_1\in\mathcal{X}} \Big| \sum_{x_0\in\mathcal{X}} P(x_1|x_0)Q_0(x_0) - P(x_1|x_0)Q_0^*(x_0) \Big| \\
&= \frac{1}{2} \sum_{x_1\in\mathcal{X}} \Big| \sum_{x_0:Q_0(x_0)>Q_0^*(x_0)} P(x_1|x_0)(Q_0(x_0)-Q_0^*(x_0)) \\
&\quad + \sum_{x_0:Q_0(x_0)<Q_0^*(x_0)} P(x_1|x_0)(Q_0(x_0)-Q_0^*(x_0)) \Big| \\
&\leq \frac{1}{2} \sum_{x_1\in\mathcal{X}} \Big( \sum_{x_0:Q_0(x_0)>Q_0^*(x_0)} P(x_1|x_0)(Q_0(x_0)-Q_0^*(x_0)) \\
&\quad + \sum_{x_0:Q_0(x_0)<Q_0^*(x_0)} P(x_1|x_0)(-Q_0(x_0)+Q_0^*(x_0)) \Big) \\
&= \frac{1}{2} \sum_{x_0:Q_0(x_0)>Q_0^*(x_0)} (Q_0(x_0)-Q_0^*(x_0)) \\
&\quad + \frac{1}{2} \sum_{x_0:Q(x_0)<Q_0^*(x_0)} (-Q_0(x_0)+Q_0^*(x_0)) \\
&= d_{\mathrm{var}}(Q_0(\cdot), Q_0^*(\cdot)).
\end{aligned}$$

Here we used in order: the definition of $d_{\mathrm{var}}$; splitting the summation into two summations; triangle inequality (all terms are now positive); switching order of summation; the definition of $d_{\mathrm{var}}$. □

**Proposition 13.** *Let $Q_0(\cdot)$ and $Q_0^*(\cdot)$ be probability distributions over a domain $\mathcal{X}$. Let $\forall x \in \mathcal{X}$ $P^*(\cdot|x)$ and $P(\cdot|x)$ be probability distributions over $\mathcal{X}$. Let $Q_1(\cdot) = \sum_{x_0\in\mathcal{X}} P(\cdot|x_0)Q_0(x_0)$ and let $Q_1^*(\cdot) = \sum_{x_0\in\mathcal{X}} P^*(\cdot|x_0)Q_0^*(x_0)$, then*

$$\begin{aligned}
d_{\mathrm{var}}(Q_1(\cdot), Q_1^*(\cdot)) \leq\ & d_{\mathrm{var}}(Q_0(\cdot), Q_0^*(\cdot)) \\
& + \sup_{x\in\mathcal{X}} d_{\mathrm{var}}(P(\cdot|x), P^*(\cdot|x)).
\end{aligned}$$

*Proof.* Let $\bar{Q}_1(\cdot) = \sum_{x_0\in\mathcal{X}} P(\cdot|x_0)Q_0^*(x_0)$. Due to triangle inequality we have

$$d_{\mathrm{var}}(Q_1(\cdot), Q_1^*(\cdot)) \leq d_{\mathrm{var}}(Q_1(\cdot), \bar{Q}_1(\cdot)) + d_{\mathrm{var}}(\bar{Q}_1(\cdot), Q_1^*(\cdot)). \qquad (29)$$

For the first term Proposition 12 gives us

$$d_{\mathrm{var}}(Q_1(\cdot), \bar{Q}_1(\cdot)) \leq d_{\mathrm{var}}(Q_0(\cdot), Q_0^*(\cdot)). \qquad (30)$$

[14] Let $P(\cdot), Q(\cdot)$ be two probability distributions over a set $\mathcal{X}$, then in the main body we defined the variational distance $d_{\mathrm{var}}(P,Q)$ as follows: $d_{\mathrm{var}}(P,Q) = \frac{1}{2}\int_{x\in\mathcal{X}} |P(x) - Q(x)|$. It is well known the following definition is equivalent: $d_{\mathrm{var}}(P,Q) = \int_{x\in\mathcal{X}:P(x)>Q(x)} |P(x)-Q(x)|$.

For the second term we have

$$\begin{aligned} & d_{\text{var}}(\bar{Q}_1(\cdot), Q_1^*(\cdot)) \\ =\ & \frac{1}{2}\sum_{x_1\in\mathcal{X}} \Big| \sum_{x_0\in\mathcal{X}} P(x_1|x_0)Q_0^*(x_0) \\ & \qquad - \sum_{x_0} P^*(x_1|x_0)Q_0^*(x_0)\Big| \\ \le\ & \frac{1}{2} \sum_{x_1\in\mathcal{X},x_0\in\mathcal{X}} \left|P(x_1|x_0)-P^*(x_1|x_0)\right| Q_0^*(x_0) \\ =\ & \sum_{x_0\in\mathcal{X}} d_{\text{var}}(P(\cdot|x_0),P^*(\cdot|x_0))Q_0^*(x_0)\\ \le\ & \sup_{x_0\in X} d_{\text{var}}(P(\cdot|x_0),P^*(\cdot|x_0)). \end{aligned} \tag{31}$$

Combining Eqn. (29), (30) and (31) gives us the statement of the proposition. □

## B. Proofs for Section 2

### B.1. Proof of Lemma 1

We first prove the following lemma.

**Lemma 14.** *Let an MDP $M = (S, \mathcal{A}, T, H, D, R)$ be given. Let another MDP $\hat{M} = (S, \mathcal{A}, \hat{T}, H, D, R)$ – which only differs from $M$ in its transition probabilities – be given. Let any $\epsilon > 0$ be given. If $\hat{T} = \{\hat{P}(\cdot|s,a)\}_{s,a}$ satisfies*

$$\forall s \in S, a \in A \;\; d_{\text{var}}(P(\cdot|s,a), \hat{P}(\cdot|s,a)) \le \epsilon,$$

*then we have for any policy $\pi$ mapping from $S$ to (probability distributions over) $A$ that*

$$|U_M(\pi) - U_{\hat{M}}(\pi)| \le H^2 \epsilon R_{\max}.$$

*Proof.* With some abuse of notation, let $P_t$ and $\hat{P}_t$ denote the distributions over states at time $t$ induced by the policy $\pi$, the initial state distribution $D$ and the transition probabilities $T$ and $\hat{T}$ respectively. Then from using Prop. 13 inductively we have for all $t \in 0:H$ that

$$d_{\text{var}}(P_t(\cdot), \hat{P}_t(\cdot)) \le t\epsilon. \tag{32}$$

So we get that

$$\begin{aligned} |U_M(\pi) - U_{\hat{M}}(\pi)| &= \Big| \sum_{t=0}^{H} \mathrm{E}_{P_t}[R(s_t)] - \sum_{t=0}^{H} \mathrm{E}_{\hat{P}_t}[R(s_t)] \Big| \\ &= \Big| \sum_{t=0}^{H} \mathrm{E}_{P_t}[R(s_t)] - \mathrm{E}_{\hat{P}_t}[R(s_t)] \Big| \\ &\le \sum_{t=0}^{H} \Big| \mathrm{E}_{P_t} R(s_t) - \mathrm{E}_{\hat{P}_t} R(s_t) \Big| \\ &\le \sum_{t=0}^{H} d_{\text{var}}(P_t, \hat{P}_t) R_{\max} \\ &\le \sum_{t=0}^{H} t\epsilon R_{\max} \\ &\le H^2 \epsilon R_{\max}. \end{aligned}$$

Where we used in order the definition of $U$; reordering terms; standard inequality for the absolute value of a sum; Prop. 11; Eqn. (32); simple algebra. □

*Proof of Lemma 1.* Consider the auxiliary MDP $M^{(1)} = (S, \mathcal{A}, T^{(1)}, H, D, R)$, where $P^{(1)}(\cdot|s,a) = P(\cdot|s,a)$ if $(s,a) \in \overline{SA}_\eta$ and $P^{(1)}(\cdot|s,a) = \hat{P}(\cdot|s,a)$ otherwise. Then we have $\forall s \in S, a \in \mathcal{A} \;\; d_{\text{var}}(\hat{P}(\cdot|s,a), P^{(1)}(\cdot|s,a)) \le \epsilon$. Using Lemma 14 gives us that

$$|U_{\hat{M}}(\pi) - U_{M^{(1)}}(\pi)| \le H^2 \epsilon R_{\max}. \tag{33}$$

Also, using condition (ii) of Lemma 1 and the definitions of $M, M^{(1)}$ we trivially get

$$|U_{M^{(1)}}(\pi) - U_M(\pi)| \le \eta H R_{\max}. \tag{34}$$

Combining Eqn. (33) and Eqn. (34) using the triangle inequality gives the statement of the lemma. □

### B.2. Proof of Lemma 2

We prove the following slightly stronger lemma, of which the statement in Lemma 2 is a subset.

**Lemma 15.** *Let any $\delta > 0$ and $a > 0$ be given. Let $\{X_i\}_{i=1}^m$ be IID Bernoulli($\phi$) random variables. Then for*

$$P(\sum_{i=1}^{m} X_i \ge a) \tag{35}$$

*to hold with probability $1-\delta$ it suffices that*

$$m \ge \frac{1}{\phi}\left( a + \log\frac{1}{\delta} + \sqrt{(a + \log\frac{1}{\delta})^2 - a^2} \right). \tag{36}$$

*We can relax the condition above to get the following less tight, but simpler sufficient condition on $m$:*

$$m \ge \frac{2}{\phi}(a + \log\frac{1}{\delta}).$$

*If $a \geq 2$ and $\log \frac{1}{\delta} \geq 2$ we have that the following is a sufficient condition on $m$*

$$m \geq \frac{2}{\phi} a \log \frac{1}{\delta}.$$

*Proof.* We start from the multiplicative 1-sided Chernoff bound for a Bernoulli distribution, with expectation $\phi$. Let $\epsilon > 0$ then

$$P(\sum_{i=1}^{m} X_i < m(\phi - \epsilon\sqrt{\phi})) \leq \exp(\frac{-m\epsilon^2}{2}). \tag{37}$$

Now let $a = m(\phi - \epsilon\sqrt{\phi})$, which implies $\epsilon = \frac{m\phi - a}{m\sqrt{\phi}}$. Substituting this into Eqn. (37) gives

$$\begin{aligned} P(\sum_{i=1}^{m} X_i < a) &\leq \exp(\frac{-m}{2}(\frac{m\phi - a}{m\sqrt{\phi}})^2) \\ &= \exp(\frac{-1}{2}(\frac{m\phi - a}{\sqrt{m\phi}})^2). \qquad (38) \end{aligned}$$

Note for Eqn. (37) to be valid, we needed $\epsilon > 0$. So Eqn. (38) is only valid if

$$m\phi - a > 0. \tag{39}$$

So for $P(\sum_{i=1}^{m} X_i < a) \leq \delta$ to hold, it is sufficient that Eqn. (39) holds and that

$$\exp(\frac{-1}{2}(\frac{m\phi - a}{\sqrt{m\phi}})^2) \leq \delta.$$

By taking $\log$ on both sides, and multiplying with $-2$ we get the equivalent condition

$$(\frac{m\phi - a}{\sqrt{m\phi}})^2 \geq 2 \log \frac{1}{\delta}.$$

Algebraic manipulation gives the equivalent condition

$$m^2\phi^2 - 2(a + \log \frac{1}{\delta}) m\phi + a^2 \geq 0. \tag{40}$$

Eqn. (40) is satisfied if

$$\begin{aligned} m\phi \in & \Big( -\infty, a + \log \frac{1}{\delta} - \sqrt{(a + \log \frac{1}{\delta})^2 - a^2} \Big] \\ & \bigcup \Big[ a + \log \frac{1}{\delta} + \sqrt{(a + \log \frac{1}{\delta})^2 - a^2}, \infty \Big). \end{aligned}$$

Now recall that the Chernoff bound was only meaningful for $\epsilon > 0$, which corresponds to the condition given by Eqn. (39). So we get that for $P(\sum_{i=1}^{m} X_i < a) \leq \delta$ to hold it suffices that

$$m\phi \geq a + \log \frac{1}{\delta} + \sqrt{(a + \log \frac{1}{\delta})^2 - a^2}.$$

□

# C. Proofs for Section 6

## C.1. Proof of Lemma 4

We first prove the following lemma.

**Lemma 16.** *Let any $\epsilon, \delta > 0$ be given. Let $X_i \sim \text{Bernoulli}(\phi)$ be IID random variables. Let $\hat{\phi}_n$ be the estimate for $\phi$ after $n$ observations. I.e., we have $\hat{\phi}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for*

$$\text{max}_{n \geq N} |\phi - \hat{\phi}_n| \leq \epsilon$$

*to hold with probability $1 - \delta$, it suffices that*

$$N \geq \frac{1}{\epsilon^2} \log(\frac{2}{\delta\epsilon}).$$

*Proof.*

$$\begin{aligned} P(\text{max}_{n \geq N} |\phi - \hat{\phi}_n| > \epsilon) &= P(\bigcup_{n \geq N} \{|\phi - \hat{\phi}_n| > \epsilon\}) \\ &\leq \sum_{n \geq N} P(|\phi - \hat{\phi}_n| > \epsilon) \\ &\leq \sum_{n \geq N} 2e^{-2\epsilon^2 n} \\ &= \frac{2(e^{-2\epsilon^2})^N}{1 - e^{-2\epsilon^2}} \end{aligned}$$

Hence, in order to guarantee that $P(\text{max}_{n \geq N} |\phi - \hat{\phi}_n| > \epsilon) \leq \delta$, it is sufficient for $N$ to satisfy

$$\begin{aligned} \frac{2(e^{-2\epsilon^2})^N}{1 - e^{-2\epsilon^2}} &\leq \delta \\ (e^{-2\epsilon^2})^N &\leq \delta(1 - e^{-2\epsilon^2})/2 \\ \log((e^{-2\epsilon^2})^N) &\leq \log(\delta(1 - e^{-2\epsilon^2})/2) \\ (-2\epsilon^2)N &\leq \log\delta + \log(1 - e^{-2\epsilon^2}) - \log 2 \\ N &\geq \frac{1}{2\epsilon^2}(-\log\delta - \log(1 - e^{-2\epsilon^2}) + \log 2) \\ N &\geq \frac{1}{2\epsilon^2}(\log\frac{1}{\delta} + \log(\frac{1}{1 - e^{-2\epsilon^2}}) + \log 2). \end{aligned}$$

Now using the fact that for any $x : -\xi < x < 0$ we have that $\exp(x) < 1 - \frac{|x|}{\xi}(1 - \exp(-\xi))$, and more specifically, for any $x : -2 < x < 0$ we have that $\exp(x) < 1 - \frac{|x|}{2}(1 - \exp(-2)) < 1 - |x|/4$. Thus for $\epsilon \in (0, 1)$ it is sufficient to have

$$\begin{aligned} N &\geq \frac{1}{2\epsilon^2}(\log\frac{1}{\delta} + \log\frac{1}{1 - (1 - (2\epsilon^2)/4)} + \log 2) \\ N &\geq \frac{1}{2\epsilon^2}(\log\frac{1}{\delta} + \log\frac{2}{\epsilon^2} + \log 2) \\ N &\geq \frac{1}{\epsilon^2}\log\frac{2}{\delta\epsilon}. \end{aligned}$$

In case $\epsilon \notin (0,1)$, more specifically when $\epsilon \geq 1$, the lemma holds with probability one independent of the number of samples. □

*Proof of Lemma 4.* Let the multinomial distribution $P$ be parameterized by the $k$ parameters $(\phi^{(1)}, \cdots, \phi^{(k)})$. Let $\hat{P}_n$ be parameterized by $(\hat{\phi}_n^{(1)}, \cdots, \hat{\phi}_n^{(k)})$ which are the respective $n$-sample estimates for all parameters $\phi^{(\cdot)}$. Then from Lemma 16 and the union bound we have that for

$$\forall i \in 1:k \;\; \max_{n \geq N} |\phi^{(i)} - \hat{\phi}_n^{(i)}| \leq \epsilon'$$

to hold with probability $1 - k\delta'$, it suffices that

$$N \geq \frac{1}{\epsilon'^2} \log(\frac{2}{\delta' \epsilon'}). \tag{41}$$

Now we use the fact that

$$d_{\mathrm{var}}(P, \hat{P}_n) \;\; = \;\; \frac{1}{2} \sum_{i=1}^{k} |\phi^{(i)} - \hat{\phi}_n^{(i)}|.$$

We get that condition (41) is sufficient to ensure that

$$\max_{n \geq N} d_{\mathrm{var}}(P, \hat{P}_n) \leq \frac{k}{2} \epsilon'$$

holds with probability at least $1 - k\delta'$. Choosing $\epsilon = \frac{k}{2}\epsilon'$ and $\delta = k\delta'$ proves the lemma. □

### C.2. Proof of Lemma 5

We first show that state-action pairs that are visited with probability sufficiently bounded away from zero under a policy $\pi$ will be accurately modeled after observing sufficient trials under that policy $\pi$.

**Lemma 17.** *Let an MDP $(S, \mathcal{A}, T, H, D, R)$ be given. Let $\pi$ be a policy for this MDP. Let any $\epsilon, \delta, \xi > 0$ be given. Let $\{\hat{P}_n(\cdot|s,a)\}_{s,a}$ be the maximum likelihood transition probability estimates based upon observing the policy $\pi$ for $n$ trials of duration $H$. Let $SA_\xi \subseteq S \times A$ be the set of state-action pairs such that the probability of seeing any specific state-action pair $(s,a) \in SA_\xi$ under the policy $\pi$ in a single trial of duration $H$ is at least $\xi$. Then for*

$$\forall n \geq N, \; \forall (s,a) \in SA_\xi, d_{\mathrm{var}}(P(\cdot|s,a), \hat{P}_n(\cdot|s,a)) \leq \epsilon$$

*to hold with probability $1 - \delta$, it suffices that*

$$N \geq \frac{2}{\xi} ( \frac{|S|^2}{4\epsilon^2} \log \frac{2|S|^3|\mathcal{A}|}{\delta\epsilon} + \log \frac{2|S||\mathcal{A}|}{\delta} ).$$

*Proof.* Let $\tilde{P}_k(\cdot|s,a)$ denote the transition probability estimate after observing the state-action pair $(s,a)$ $k$ times. From Lemma 4, for

$$\forall k \geq K, \;\; d_{\mathrm{var}}(P(\cdot|s,a), \tilde{P}_k(\cdot|s,a)) \leq \epsilon$$

to hold with probability $1 - \delta'$, it suffices that

$$K \geq \frac{|S|^2}{4\epsilon^2} \log \frac{|S|^2}{\delta' \epsilon}.$$

Now for $(s,a) \in SA_\xi$ we combine this result with Lemma 2. This gives that for

$$\forall n \geq N \;\; d_{\mathrm{var}}(P(\cdot|s,a), \hat{P}_n(\cdot|s,a)) \leq \epsilon \tag{42}$$

to hold with probability $1 - \delta' - \delta''$, it is sufficient to have

$$N \geq \frac{2}{\xi} ( \frac{|S|^2}{4\epsilon^2} \log \frac{|S|^2}{\delta' \epsilon} + \log \frac{1}{\delta''} ). \tag{43}$$

Taking a union bound over all state-action pairs $(s,a) \in SA_\xi$ (note $|SA_\xi| \leq |S||\mathcal{A}|$) gives that for Eqn. (42) to hold for all $(s,a) \in SA_\xi$ with probability $1 - |S||\mathcal{A}|\delta' - |S||\mathcal{A}|\delta''$, it suffices that Eqn. (43) is satisfied. Choosing $\delta' = \delta'' = \frac{\delta}{2|S||\mathcal{A}|}$ gives the lemma. □

The above lemma tells us only a polynomial number of sample trajectories under the teacher's policy are necessary, to guarantee that the state-action pairs frequently visited under the teacher's policy are accurate in all models $\{M^{(i)}\}_i$. Now we use the Simulation Lemma to translate this into accurate evaluation of the utility of the teacher's policy in the models $\{M^{(i)}\}_i$.

**Lemma 5** (restated). Let any $\alpha, \delta > 0$ be given. Assume we use the algorithm as described in Section 4. Let $N_T$ satisfy the following condition

$$\begin{aligned} N_T \;\; \geq \;\; & \frac{32|S||\mathcal{A}|HR_{\max}}{\alpha} \Big( \log \frac{2|S||\mathcal{A}|}{\delta} \\ & + \frac{64|S|^2 H^4 R_{\max}^2}{\alpha^2} \log \frac{32H^2 R_{\max}|S|^3|\mathcal{A}|}{\delta\alpha} \Big), \end{aligned}$$

or simplified and less tight (Notice that if $\alpha/8 > HR_{\max}$, the statement is trivially true with probability 1. So we can simplify using the fact that $\alpha/8 \leq HR_{\max}$.)

$$N_T \geq \frac{4096|S|^3|\mathcal{A}|H^5 R_{\max}^3}{\alpha^3} \log \frac{32H^2 R_{\max}|S|^3|\mathcal{A}|}{\delta\alpha}$$

Then with probability $1 - \delta$ we have that

$$\forall i \;\; |U_{\hat{M}^{(i)}}(\pi_T) - U_M(\pi_T)| \leq \alpha/8.$$

*Proof.* Let

$$\epsilon = \frac{1}{2} \frac{\alpha/8}{H^2 R_{\max}}$$

and let

$$\eta = \frac{1}{2} \frac{\alpha/8}{HR_{\max}}.$$

From the Simulation Lemma we have that if there exists a set of state-action pairs $\overline{SA}_\eta \subseteq S \times \mathcal{A}$ such that the following holds

$$\begin{array}{rl}\text{(i)} & \forall (s,a) \in \overline{SA}_\eta,\ \ d_{\mathrm{var}}(P(\cdot|s,a), \hat{P}(\cdot|s,a)) \leq \epsilon, \\ \text{(ii)} & P(\{(s_t,a_t)\}_{t=0}^{H} \subseteq \overline{SA}_\eta | \pi_T, M) \geq 1-\eta.\end{array}$$

Then we have

$$|U_M(\pi_T) - U_{\hat{M}}(\pi_T)| \leq H^2 \epsilon R_{\max} + \eta H R_{\max} = \frac{\alpha}{8}.$$

Where the last inequality follows from our choice of $\epsilon$ and $\eta$ above. Now choose $\overline{SA}_\eta = \{(s,a) \in S \times \mathcal{A} :$ $\mathrm{P}((s,a)$ is visited in a trial of length $H$under $\pi_T) \geq \frac{\eta}{|S||\mathcal{A}|}\}$, then $\overline{SA}_\eta$ satisfies condition (ii). So it remains to show that condition (i) is satisfied. From Lemma 17 we have that for all $(s,a) \in \overline{SA}_\eta$ for

$$d_{\mathrm{var}}(P(\cdot|s,a), \hat{P}_n(s,a)) \leq \epsilon$$

to hold with probability $1-\delta$ it suffices that

$$N \geq \frac{2|S||\mathcal{A}|}{\eta}\left(\frac{|S|^2}{4\epsilon^2}\log\frac{2|S|^3|\mathcal{A}|}{\delta\epsilon} + \log\frac{2|S||\mathcal{A}|}{\delta}\right).$$

Filling in the choices of $\epsilon$ and $\eta$ into this condition gives the sufficient condition on $N_T$ as stated in this lemma. □

### C.3. Proof of Lemma 6

We now give a formal proof of Lemma 6.

*Proof of Lemma 6.* From Lemma 5 we have that for Eqn. (12) to hold with probability $1-\delta'$ it suffices that

$$N_T \geq \frac{4096|S|^3|\mathcal{A}|H^5 R_{\max}^3}{\alpha^3}\log\frac{32H^2 R_{\max}|S|^3|\mathcal{A}|}{\delta'\alpha}. \tag{44}$$

Eqn. (14) is trivially true when $\alpha > 16HR_{\max}$. If $\alpha \leq 16HR_{\max}$, the Hoeffding inequality gives us that for Eqn. (14) to hold with probability at least $1-\delta'$, it is (more than) sufficient that $N_T$ satisfies Eqn. (44).

The Hoeffding inequality also gives us that for Eqn. (13) to hold with probability $1-N\delta''$ it suffices that

$$k_1 \geq \frac{16^2 H^2 R_{\max}^2}{2\alpha^2}\log\frac{2}{\delta''}. \tag{45}$$

Now since the algorithm only exits in iteration $N$, we must have for all $i = 1 : N-1$ that

$$\hat{U}_M(\pi^{(i)}) < \hat{U}_M(\pi_T) - \alpha/2. \tag{46}$$

Combining Eqn. (46), (12), (13) and (14) and the fact that $\pi^{(i)}$ is $\alpha/8$-optimal for $\hat{M}^{(i)}$ we get

$$\forall i\,(1 \leq i < N-1)\ U_{\hat{M}^{(i)}}(\pi^{(i)}) \geq U_M(\pi^{(i)}) + \alpha/8. \tag{47}$$

So far we have shown that for Eqn. (12), (13), (14) and (47) to hold with probability $1-2\delta'-N\delta''$, it suffices that Eqn. (44) and Eqn. (45) are satisfied.

Now using the contrapositive of the Simulation Lemma and choosing $\epsilon = \frac{1}{2}\frac{\alpha/8}{H^2 R_{\max}}$ we get from Eqn. (47) that the policy $\pi^{(i)}$ must be visiting a state-action pair $(s,a)$ that satisfies

$$d_{\mathrm{var}}(P(\cdot|s,a), \hat{P}^{(i)}(\cdot|s,a)) > \frac{\alpha}{16H^2 R_{\max}} \tag{48}$$

with probability at least $\frac{\alpha}{16HR_{\max}}$ in every trial of horizon $H$.

[High-level proof intuition. Eqn. (47) states that the models $\hat{M}^{(i)}$ are inaccurate; they overestimate the utility of the policies $\{\pi^{(i)}\}_{i=1:N-1}$. We used the contrapositive of the Simulation Lemma to show this implies that an “inaccurately” modeled state-action pair must be visited with probability $\frac{\alpha}{16HR_{\max}}$ by such a policy $\pi^{(i)}$. This means that data collected under such a policy will improve the model. It remains to show that this can only happen a limited number of times until the model has become a good model.]

Let $\tilde{P}_k(\cdot|s,a)$ be the estimate of $P(\cdot|s,a)$ based upon $k$ observations of the state-action pair $(s,a)$. From Lemma 4 we have that for any state-action pair $(s,a)$ for

$$\forall k > K,\ \ d_{\mathrm{var}}(P(\cdot|s,a), \tilde{P}_k(\cdot|s,a)) \leq \frac{\alpha}{16H^2 R_{\max}}$$

to hold with probability $1-\delta'''$, it suffices that

$$K \geq \frac{16^2 H^4 R_{\max}^2 |S|^2}{4\alpha^2}\log\frac{16H^2 R_{\max}|S|^2}{\alpha\delta'''}.$$

The above equation bounds the number of times a state-action pair can be visited until it is “accurately” modeled with probability $1-\delta'''$. So with probability $1-|S||\mathcal{A}|\delta'''$ one can encounter state-action pairs $(s,a)$ that—at the moment of encounter—satisfy Eqn. (48) (i.e., that are inaccurately modeled) at most

$$|S||\mathcal{A}|\frac{16^2 H^4 R_{\max}^2 |S|^2}{4\alpha^2}\log\frac{16H^2 R_{\max}|S|^2}{\alpha\delta'''} \tag{49}$$

times. Since (from above) such a state-action pair is encountered with probability at least $\frac{\alpha}{16HR_{\max}}$ in each iteration of the algorithm before it exits, Lemma 2 gives that w.p. $1-\delta''''$ after a number of iterations

$$\begin{aligned} N_{\mathrm{ubound}} \ &= \ \frac{32HR_{\max}}{\alpha}\Big(\log\frac{1}{\delta''''} \\ &+ |S||\mathcal{A}|\frac{16^2 H^4 R_{\max}^2 |S|^2}{4\alpha^2}\log\frac{16H^2 R_{\max}|S|^2}{\alpha\delta'''}\Big) \end{aligned} \tag{50}$$

such a state-action pair has been encountered as many times as stated in Eqn. (49). So $N_{\mathrm{ubound}}$ is an upperbound on the number of iterations of the algorithm with probability

$1 - 2\delta' - N\delta'' - |S||\mathcal{A}|\delta''' - \delta''''$. Choose $\delta', \delta'', \delta''', \delta''''$ such that

$$2\delta' = N_{\text{ubound}}\delta'' = |S||\mathcal{A}|\delta''' = \delta'''' = \frac{1}{4}\delta. \tag{51}$$

Substituting these choices into Eqn. (50), Eqn. (45) and Eqn. (44) gives us Eqn. (8), Eqn. (9) and Eqn. (10). □

## D. Proofs for Section 7.1

*Proof of Proposition 7.* W.l.o.g. assume a coordinate system $x_{1:n}$ such that

$$\forall i \in 2:n, (\mu_1)_i = (\mu_2)_i = 0,$$

and

$$(\mu_1)_1 \leq (\mu_2)_1.$$

Also, let $a = (\mu_1)_1$ and $b = (\mu_2)_1$. Let

$$f(z;\mu) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left(\frac{-\|z-\mu\|_2^2}{2\sigma^2}\right).$$

Then we have that

$$\begin{aligned}
& d_{\text{var}}(\mathcal{N}(\mu_1,\sigma^2), \mathcal{N}(\mu_2,\sigma^2)) \\
= \ & \frac{1}{2}\int_{x_{2:n}} dx_{2:n} \int_{x_1=-\infty}^{x_1=+\infty} \big|f(x;\mu_1) - f(x;\mu_2)\big| dx_1 \\
= \ & \frac{1}{2}\int_{x_{2:n}} dx_{2:n} \int_{x_1=-\infty}^{x_1=(a+b)/2} \big(f(x;\mu_1) - f(x;\mu_2)\big) dx_1 \\
+ & \frac{1}{2}\int_{x_{2:n}} dx_{2:n} \int_{x_1=(a+b)/2}^{x_1=+\infty} \big(f(x;\mu_2) - f(x;\mu_1)\big) dx_1 \\
= \ & \frac{1}{2}\left(1 - 2\int_{x_{2:n}} dx_{2:n} \int_{x_1=(a+b)/2}^{x_1=+\infty} f(x;\mu_1) dx_1\right) \\
+ & \frac{1}{2}\left(1 - 2\int_{x_{2:n}} dx_{2:n} \int_{x_1=-\infty}^{x_1=(a+b)/2} f(x;\mu_2) dx_1\right) \\
= \ & 1 - 2\frac{1}{\sqrt{2\pi}\sigma}\int_{z=|a-b|/2}^{\infty} \exp(\frac{-z^2}{2\sigma^2}) dz \\
= \ & \frac{1}{\sqrt{2\pi}\sigma}\int_{z=-|a-b|/2}^{|a-b|/2} \exp(\frac{-z^2}{2\sigma^2}) dx_1 \\
\leq \ & \frac{1}{\sqrt{2\pi}\sigma}|a-b| \\
= \ & \frac{1}{\sqrt{2\pi}\sigma}\|\mu_2 - \mu_1\|_2.
\end{aligned}$$

□

In our setting, we have Gaussian noise contributing additively to the next-state given current-state and action. As a consequence, the random variables (and their increments over time) are not bounded. The following proposition shows that we can "essentially" treat Gaussian random variables as bounded random variables with high probability by truncating the tails.

**Proposition 18.** *Let any $N > 0, \delta > 0, \sigma > 0$ be given. Let $\{w^{(i)}\}_{i=1}^N$ be Gaussian random variables, namely $w^{(i)} \sim \mathcal{N}(0,\sigma^2)$. Then with probability $1-\delta$ we have*

$$\max_{i\in 1:N}|w^{(i)}| \leq \sigma \log \frac{4N}{\sqrt{2\pi}\delta}.$$

*Proof.*

$$\begin{aligned}
\mathrm{P}(|w^{(i)}| \geq K) &= 2\frac{1}{\sqrt{2\pi}\sigma}\int_{z=K}^{\infty} \exp(-\frac{z^2}{2\sigma^2}) dz \\
&\leq 2\frac{1}{\sqrt{2\pi}\sigma}\int_{z=K}^{\infty} 2\exp(-\frac{z}{\sigma}) dz \\
&= 4\frac{1}{\sqrt{2\pi}}\exp(-K/\sigma)
\end{aligned}$$

So we have

$$\mathrm{P}(\max_{i\in 1:N}|w^{(i)}| \geq K) \leq N4\frac{1}{\sqrt{2\pi}}\exp(-K/\sigma),$$

which is equivalent to the proposition. □

Note the bound we use for the tail is fairly loose, but it's a simple form and enough for our purposes. Let $K > 0$, here is an another bound (based upon integration by parts):

$$\begin{aligned}
\int_{z=K}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\frac{z}{z}\exp(-\frac{z^2}{2\sigma^2}) dz &= \left[\frac{-\sigma}{z\sqrt{2\pi}}\exp(-\frac{z^2}{2\sigma^2})\right]_K^{\infty} \\
& \quad - \int_{z=K}^{\infty} \frac{\sigma}{\sqrt{2\pi}}\frac{1}{z^2}\exp(-\frac{z^2}{2\sigma^2}) dz \\
&\leq \frac{\sigma}{\sqrt{2\pi}K}\exp(-\frac{K^2}{2\sigma^2}).
\end{aligned}$$

## E. Proofs for Section 7.2

In this section, let $n = n_S + n_{\mathcal{A}}$. In this section we establish several helper lemmas, and then prove Lemma 8. Proofs of helper lemmas are given as sketches or even left out in this section. The proofs are given in the following sections.

We will prove Lemma 8 in the following two steps.

- In Section E.1 we establish that for any inaccurate parameter $\theta$, we have that $\text{loss}^{(N_T H)}(\theta) > \text{loss}^{(N_T H)}(\theta^*) + \Omega(N_T)$. I.e., the true parameter $\theta^*$ outperforms all inaccurate parameters by a margin of $\Omega(N_T)$.

- In Section E.2 we establish that no matter how the additional data $\{z^{(i)}\}_{i=N_T H+1}^{N_T H + k_1 NH}$ are chosen, the probability that $\theta^*$ ever gets outperformed by an inaccurate parameter $\theta$ is exponentially small in the margin $\Omega(N_T)$. As a consequence, a "small" number of samples $N_T$ from the teacher is sufficient to guarantee that throughout all iterations an accurate parameter has the smallest loss.

### E.1. Model estimated from teacher's data

In this section we establish that for any inaccurate parameter $\theta$, we have that $\mathrm{loss}^{(N_TH)}(\theta) > \mathrm{loss}^{(N_TH)}(\theta^*) + \Omega(N_T)$. I.e., the true parameter $\theta^*$ outperforms all inaccurate parameters by a margin of $\Omega(N_T)$. A standard way to prove this is to prove it for one specific inaccurate parameter $\theta$, and then use a cover of $\theta$-space and a union bound to prove it for the whole space. The following lemma shows that it is sufficient to cover only a "small" ball around the origin. More specifically, it shows that with high probability all the solutions $\{\hat{\theta}^{(k)}\}_{k=N_TH}^{N_TH+k_1NH}$ to the regularized linear regression problem lie in a small ball $S$ around the origin. Note that the solutions $\{\hat{\theta}^{(k)}\}_{k=N_TH}^{N_TH+k_1NH}$ to the regularized linear regression problems at steps $k = N_TH$ to $N_TH + k_1NH$ are by definition the only parameters we end up using in the algorithm. So it is sufficient to show $\mathrm{loss}^{(N_TH)}(\theta) > \mathrm{loss}^{(N_TH)}(\theta^*) + \Omega(N_T)$ for all inaccurate $\theta$ in the "small ball" $S$ that contains all $\{\hat{\theta}^{(k)}\}_{k=N_TH}^{N_TH+k_1NH}$.

**Lemma 19.** *Let any $\delta > 0$ be given. Let $\{y^{(i)}, z^{(i)}\}_{i=1}^{N_TH+k_1NH}$ be generated as described in Eqn. (20). Let $\{\hat{\theta}^{(k)}\}_k$ be defined as in Eqn. (21). Let $\tilde{m} = N_TH + k_1NH$. Then we have with probability $1-\delta$ that $\forall k\ (1 \le k \le \tilde{m})$*

$$\|\hat{\theta}^{(k)}\|_2 \quad \le \quad \kappa\sqrt{\tilde{m}}\left(\sqrt{2}\|\theta^*\|_2 + \sigma\log\frac{4\tilde{m}}{\sqrt{2\pi}\delta}\right)$$

*and that*

$$\max{}_{i=1:\tilde{m}}|y^{(i)}| \le \sqrt{2}\|\theta^*\|_2 + \sigma\log\frac{4\tilde{m}}{\sqrt{2\pi}\delta}.$$

Note Lemma 19 would not hold if we used unregularized linear regression. Instead of regularizing with a quadratic penalty, one could regularize by explicitly constraining $\theta$ to be norm-bounded. This would directly result in a ball that is sufficient to be covered and thus simplify some of the proofs. However, in practice regularized linear regression with a quadratic penalty (as used in our algorithm) is much more commonly used than linear regression with a norm constraint on $\theta$.

The following lemma establishes $\mathrm{loss}^{(N_TH)}(\theta) > \mathrm{loss}^{(N_TH)}(\theta^*) + \Omega(N_T)$ for all inaccurate $\theta \in S$.

**Lemma 20.** *Let any $\delta, \epsilon, \eta > 0$ be given. Let $\{y^{(i)}, z^{(i)}\}_{i=1}^{N_TH+k_1NH}$ be generated as described in Eqn. (20). Let $\tilde{m} = N_TH + k_1NH$. Let $y_{\max} = \max_{i=1:\tilde{m}}|y^{(i)}| \le \sqrt{2}\|\theta^*\|_2 + \sigma\log\frac{8\tilde{m}}{\sqrt{2\pi}\delta}$ and let $R \le \kappa\sqrt{\tilde{m}}\left(\sqrt{2}\|\theta^*\|_2 + \sigma\log\frac{8\tilde{m}}{\sqrt{2\pi}\delta}\right)$. Let $S = \{\theta : \|\theta\|_2 \le R\}$. Then for all $\theta \in S$ that do not satisfy Eqn. (22) we have that for $\mathrm{loss}^{(N_TH)}(\theta) - \mathrm{loss}^{(N_TH)}(\theta^*) \ge N_T\epsilon^2\eta/4$ to hold with probability $1-\delta/4$, it suffices that*

$$N_T = \Omega(\mathrm{poly}(\tfrac{1}{\epsilon}, \tfrac{1}{\eta}, \tfrac{1}{\delta}, H, \|\theta^*\|_2, n_S, n_{\mathcal{A}}, k_1, N)). \quad (52)$$

### E.2. Influence of data from policies $\{\pi^{(i)}\}_i$

In this section we study the power of an adversary to favor one specific $\theta$ over $\theta^*$ by choosing the samples $\{z^{(i)}\}_{i=N_TH+1}^{N_TH+k_1NH}$. More specifically the following lemma shows that the adversary's power is very limited. I.e., no matter what policy the adversary uses, the probability of ever making $\theta$ outperform $\theta^*$ by a margin $a \ge 0$ on this set of adversarially chosen samples is bounded by $\exp(-a/\sigma^2)$. Let

$$\mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta) = \textstyle\sum_{i=m+1}^{k}(y^{(i)} - \theta^\top z^{(i)})^2.$$

By convention let $\mathrm{loss}_{\mathrm{adv}}^{(k)} = 0$ for $k \le m$.

**Lemma 21.** *Let any $a \ge 0$ be given. Let any $\theta \in \mathbb{R}^n$ be given. Let all else be as defined above. Then we have*

$$\mathrm{P}(\exists k > m : \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta) \le \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta^*) - a) \le \exp(\tfrac{-a}{2\sigma^2}).$$

The following lemma uses a covering argument to extend Lemma 21 to hold for the set $\{\theta : \|\theta\|_2 \le R\}$.

**Lemma 22.** *Let any $\delta, \epsilon, \eta > 0$ be given. Let $\{y^{(i)}, z^{(i)}\}_{i=1}^{N_TH+k_1NH}$ be generated as described in Eqn. (20). Let $\tilde{m} = N_TH + k_1NH$. Let*

$$R = \kappa\sqrt{\tilde{m}}\left(\sqrt{2}\|\theta^*\|_2 + \sigma\log\frac{8\tilde{m}}{\sqrt{2\pi}\delta}\right).$$

*Let $S = \{\theta : \|\theta\|_2 \le R\}$. Then for all $\theta \in S$ that do not satisfy Eqn. (22) we have that for*

$$\mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta) - \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta^*) \ge -N_T\epsilon^2\eta/8 > 0$$

*to hold with probability $1 - \frac{\delta}{4}$ for all $k \in N_TH : k_1NH$, it suffices that*

$$N_T = \Omega\left(\mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\eta}, \frac{1}{\delta}, H, \|\theta^*\|_2, n_S, n_{\mathcal{A}}, k_1, N)\right). \quad (53)$$

### E.3. Proof of Lemma 8

*Proof of Lemma 8.* From Lemma 19 we have with probability $1 - \frac{\delta}{2}$ that all the estimates $\{\hat{\theta}^{(k)}\}_{k=N_TH+1}^{N_TH+k_1NH}$ lie in a bounded sphere $S = \{\theta : \|\theta\|_2 \le \kappa\sqrt{\tilde{m}}\left(\sqrt{2}\|\theta^*\|_2 + \sigma\log\frac{8\tilde{m}}{\sqrt{2\pi}\delta}\right)\}$ around the origin and that $\max_{i=1:\tilde{m}}|y^{(i)}| \le \sqrt{2}\|\theta^*\|_2 + \sigma\log\frac{8\tilde{m}}{\sqrt{2\pi}\delta}$. Lemma 20 gives us that for all $\theta$ in this sphere $S$ that do not satisfy the accuracy condition of Eqn. (22) we have that

$$\mathrm{loss}^{(N_TH)}(\theta) - \mathrm{loss}^{(N_TH)}(\theta^*) \ge N_T\epsilon^2\eta/4 \quad (54)$$

holds with probability $1 - \frac{\delta}{4}$ for large enough $N_T$ as quantified in Eqn. (52), which corresponds to the condition on $N_T$ in the lemma we are proving. From Lemma 22 we have that

$$\mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta) - \mathrm{loss}_{\mathrm{adv}}^{(k)}(\theta^*) \ge -N_T\epsilon^2\eta/8 \quad (55)$$

holds with probability $1 - \frac{\delta}{4}$ under the same condition on $N_T$. Now since for any $k \geq N_T H$ we have by definition $\text{loss}^{(k)}(\theta) = \text{loss}^{(N_T H)}(\theta) + \text{loss}^{(k)}_{\text{adv}}(\theta)$. We can combine Eqn. (54) and Eqn. (55) to obtain that

$$\text{loss}^{(k)}(\theta) \geq \text{loss}^{(k)}(\theta^*) + N_T \epsilon^2 \eta / 8 \tag{56}$$

holds with probability $1 - \delta$. This means that all $\theta \in S$ that do not satisfy the accuracy condition of Eqn. (22) are outperformed by the true parameter $\theta^*$ by a margin of at least $N_T \epsilon^2 \eta / 8$ for all $k \geq N_T H$. As a consequence all the $\theta$ estimates $\{\hat{\theta}^{(k)}\}_{k=N_T H+1}^{N_T H + k_1 N H}$ must satisfy the accuracy condition of Eqn. (22). □

### E.4. Proof of Lemma 19

*Proof of Lemma 19.* Let $y_{\max} = \max_{i=1:\tilde{m}} |y^{(i)}|$. For any $k : 1 \leq k \leq \tilde{m}$ we have $\text{loss}^{(k)}(\vec{0}) = \sum_{i=1}^{k} (y^{(i)})^2 \leq \tilde{m}(y_{\max})^2$. Since $\hat{\theta}^{(k)}$ achieves the minimal loss, we must have $\text{loss}^{(k)}(\hat{\theta}^{(k)}) \leq \tilde{m}(y_{\max})^2$, and thus

$$\tfrac{\|\hat{\theta}^{(k)}\|_2^2}{\kappa^2} \leq \tilde{m}(y_{\max})^2. \tag{57}$$

We also have (using $\|z^{(i)}\|_2 \leq \sqrt{2}$ and Proposition 18) that

$$\begin{aligned} y_{\max} &\leq \sqrt{2}\|\theta^*\|_2 + \max_{i=1:\tilde{m}} |w^{(i)}| \\ &\leq \sqrt{2}\|\theta^*\|_2 + \sigma \log \tfrac{8\tilde{m}}{\sqrt{2\pi}\delta} \text{w.p. } 1 - \tfrac{\delta}{2}, \end{aligned}$$

which combined with Eqn. (57) proves the lemma. □

### E.5. Proof of Lemma 20

Let

$$\Delta^{(k)}(\theta_1, \theta_2) = \frac{1}{\kappa^2} \big| \|\theta_1\|_2^2 - \|\theta_2\|_2^2 \big|$$
$$+ \max_{\{z^{(i)}, y^{(i)}\}_i} \sum_{i=1}^{k} \big| (y^{(i)} - \theta_1^\top z^{(i)})^2 - (y^{(i)} - \theta_2^\top z^{(i)})^2 \big|.$$

We first prove the following lemma.

**Lemma 23.** *Let any $\lambda > 0, \tilde{m} > 0$ be given. Let $y_{\max} = \max_{i=1:\tilde{m}} |y^{(i)}|$. There is a subset $\tilde{S}$ of the sphere $S = \{\theta : \|\theta\|_2 \leq R\}$ such that for all $\theta \in S$, there exists $\tilde{\theta} \in \tilde{S}$ such that for all $k (1 \leq k \leq \tilde{m})$ we have:*

$$|\text{loss}^{(k)}(\theta) - \text{loss}^{(k)}(\tilde{\theta})| \leq \Delta^{(k)}(\theta, \tilde{\theta}) \leq \lambda \tag{58}$$

*And we have the following bound on the number of points in $\tilde{S}$*

$$|\tilde{S}| \leq \left( \frac{2R\sqrt{n}(4\tilde{m} y_{\max} + (8\tilde{m} + 2\frac{1}{\kappa^2})R)}{\lambda} \right)^n.$$

*In the special case where $R \leq \kappa\sqrt{\tilde{m}}(\sqrt{2}\|\theta^*\|_2 + \sigma \log \frac{8\tilde{m}}{\sqrt{2\pi}\delta})$ and $y_{\max} \leq \sqrt{2}\|\theta^*\|_2 + \sigma \log \frac{8\tilde{m}}{\sqrt{2\pi}\delta}$ we get*

$$\log |\tilde{S}| = O\left( n \log \text{poly}(\frac{1}{\lambda}, \frac{1}{\delta}, \tilde{m}, \|\theta^*\|_2, n) \right).$$

*Proof.* Note that the first inequality in Eqn. (58) trivially holds by definition of $\Delta$. We now prove the second inequality. Consider the difference in loss for two weight vectors $\theta_1, \theta_2$. The contribution of one training sample is bounded by

$$\begin{aligned} &\sup_{z:\|z\|\leq\sqrt{2}, y:|y|\leq y_{\max}} \big| (y - \theta_1^T z)^2 - (y - \theta_2^T z)^2 \big| \\ = &\sup_{z:\|z\|\leq\sqrt{2}, y:|y|\leq y_{\max}} \big| (y - \theta_1^T z)^2 - (y - \theta_1^T z + (\theta_1 - \theta_2)^T z)^2 \big| \\ = &\sup_{z:\|z\|\leq\sqrt{2}, y:|y|\leq y_{\max}} \big| - ((\theta_1 - \theta_2)^T z)^2 - 2(y - \theta_1^T z)(\theta_1 - \theta_2)^T z \big| \\ \leq &\ 2\|\theta_1 - \theta_2\|_2^2 + 2\sqrt{2}(y_{\max} + \sqrt{2}R)\|\theta_1 - \theta_2\|_2 \\ \leq &\ 4R\|\theta_1 - \theta_2\|_2 + 2\sqrt{2}(y_{\max} + \sqrt{2}R)\|\theta_1 - \theta_2\|_2 \\ \leq &\ (4y_{\max} + 8R)\|\theta_1 - \theta_2\|_2. \end{aligned}$$

So we have

$$\begin{aligned} &\Delta^{(k)}(\theta_1, \theta_2) \\ \leq &\quad k(4y_{\max} + 8R)\|\theta_1 - \theta_2\|_2 + \frac{1}{\kappa^2} \big| \|\theta_1\|_2^2 - \|\theta_2\|_2^2 \big| \\ = &\quad k(4y_{\max} + 8R)\|\theta_1 - \theta_2\|_2 + \frac{1}{\kappa^2} \big| (\theta_1 + \theta_2)^\top (\theta_1 - \theta_2) \big| \\ \leq &\quad k(4y_{\max} + 8R)\|\theta_1 - \theta_2\|_2 + \frac{1}{\kappa^2} 2R\|\theta_1 - \theta_2\|_2 \\ \leq &\quad (4k y_{\max} + (8k + 2\frac{1}{\kappa^2})R)\|\theta_1 - \theta_2\|_2. \end{aligned}$$

To cover a sphere of radius $R$ up to $\|\cdot\|_2 \leq \gamma$, it is sufficient to have (cover the enclosing cube with side of $2R$ regularly) $(2R\sqrt{n}/\gamma)^n$ points. In our case we want to cover the set of considered $\theta$'s up to loss accuracy $\lambda$, so we have $\gamma = \frac{\lambda}{4k y_{\max} + (8k + 2\frac{1}{\kappa^2})R}$, resulting in a number of points

$$\left( \frac{2R\sqrt{n}(4k y_{\max} + (8k + 2\frac{1}{\kappa^2})R)}{\lambda} \right)^n.$$

This establishes the lemma for one specific $k$. It is easily seen that the cover used for $k = \tilde{m}$ can be used for all $k \leq \tilde{m}$. This proves the theorem. □

We will use $\tilde{S}(n, \frac{1}{\lambda}, \frac{1}{\delta}, \tilde{m}, \|\theta^*\|_2)$ if we want to explicitly show the dependence on the parameters.

*Proof of Lemma 20.* Let $\theta \in \mathbb{R}^n$. Let $e_\theta^{(i)} = \theta^{*\top} z^{(i)} - \theta^\top z^{(i)}$, and let $\tilde{e}_\theta^{(i)}$ be $e_\theta^{(i)}$ clipped to the interval $[-K_w, K_w]$. Let $K_w$ be such that $\max_{i \in 1:N_T H} |w^{(i)}| \leq K_w$. Then we have that

$$\begin{aligned} \text{loss}^{(N_T H)}(\theta) &= \sum_{t=1}^{N_T H} (w^{(i)} + e_\theta^{(i)})^2 + \frac{1}{\kappa^2}\|\theta\|_2^2 \\ &\geq \sum_{t=1}^{N_T H} (w^{(i)} + \tilde{e}_\theta^{(i)})^2 + \frac{1}{\kappa^2}\|\theta\|_2^2. \end{aligned}$$

Now consider the difference in loss for $\theta$ and the optimal $\theta^*$:

$$\begin{aligned}
&\mathrm{loss}^{(N_T H)}(\theta) - \mathrm{loss}^{(N_T H)}(\theta^*) \\
&= \sum_{i=1}^{N_T H} (w^{(i)} + e_\theta^{(i)})^2 + \frac{1}{\kappa^2}\|\theta\|_2^2 - \sum_{i=1}^{N_T H} (w^{(i)})^2 - \frac{1}{\kappa^2}\|\theta^*\|_2^2 \\
&\geq \sum_{i=1}^{N_T H} (w^{(i)} + \tilde{e}_\theta^{(i)})^2 + \frac{1}{\kappa^2}\|\theta\|_2^2 - \sum_{i=1}^{N_T H} (w^{(i)})^2 - \frac{1}{\kappa^2}\|\theta^*\|_2^2 \\
&= \sum_{i=1}^{N_T H} (\tilde{e}_\theta^{(i)})^2 + 2w^{(i)}\tilde{e}_\theta^{(i)} + \frac{1}{\kappa^2}(\|\theta\|_2^2 - \|\theta^*\|_2^2) \qquad (59)
\end{aligned}$$

Let $Z_t = \sum_{i=1}^t (\tilde{e}_\theta^{(i)})^2 + 2w^{(i)}\tilde{e}_\theta^{(i)} - \mathrm{E}(\tilde{e}_\theta^{(i)})^2$. Note that $\forall t, |Z_t - Z_{t-1}| \leq 4K_w^2$. Then applying Azuma's inequality to the martingale $\{Z_t\}_t$ gives us that

$$\sum_{i=1}^{N_T H} (\tilde{e}_\theta^{(i)})^2 + 2w^{(i)}\tilde{e}_\theta^{(i)} - \mathrm{E}(\tilde{e}_\theta^{(i)})^2 \geq -\lambda \qquad (60)$$

holds with probability $1 - \exp(-\lambda^2/(2N_T H 4K_w^2))$. Combining Eqn. (59) and (60) gives that

$$\begin{aligned}
&\mathrm{loss}^{(N_T H)}(\theta) - \mathrm{loss}^{(N_T H)}(\theta^*) \geq \\
&\quad \sum_{i=1}^{N_T H} \mathrm{E}(\tilde{e}_\theta^{(i)})^2 - \lambda + \frac{1}{\kappa^2}(\|\theta\|_2^2 - \|\theta^*\|_2^2)
\end{aligned}$$

holds with probability $1 - \exp(-\lambda^2/(8N_T H K_w^2))$. Now using Lemma 23 we get that with probability $1 - |\tilde{S}|\exp(-\lambda^2/(8N_T H K_w^2))$ the following holds for all $\theta(\|\theta\|_2 \leq R) : \exists\tilde{\theta} \in \tilde{S} \text{ s.t. } \Delta^{(N_T H)}(\theta, \tilde{\theta}) \leq \lambda$ and

$$\begin{aligned}
&\mathrm{loss}^{(N_T H)}(\theta) - \mathrm{loss}^{(N_T H)}(\theta^*) \\
&= \mathrm{loss}^{(N_T H)}(\theta) - \mathrm{loss}^{(N_T H)}(\tilde{\theta}) \\
&\quad + \mathrm{loss}^{(N_T H)}(\tilde{\theta}) - \mathrm{loss}^{(N_T H)}(\theta^*) \\
&\geq -\lambda + \mathrm{loss}^{(N_T H)}(\tilde{\theta}) - \mathrm{loss}^{(N_T H)}(\theta^*) \\
&\geq \sum_{i=1}^{N_T H} \mathrm{E}(\tilde{e}_{\tilde{\theta}}^{(i)})^2 - 2\lambda + \frac{1}{\kappa^2}(\|\tilde{\theta}\|_2^2 - \|\theta^*\|_2^2) \\
&= \sum_{i=1}^{N_T H} \mathrm{E}(\tilde{e}_\theta^{(i)})^2 - 2\lambda + \frac{1}{\kappa^2}(\|\theta\|_2^2 - \|\theta^*\|_2^2) \\
&\quad + \sum_{i=1}^{N_T H} \mathrm{E}(\tilde{e}_{\tilde{\theta}}^{(i)})^2 - \mathrm{E}(\tilde{e}_\theta^{(i)})^2 + \frac{1}{\kappa^2}(\|\tilde{\theta}\|_2^2 - \|\theta\|_2^2). \qquad (61)
\end{aligned}$$

We also have that

$$\begin{aligned}
&\Big|\sum_{i=1}^{N_T H} \mathrm{E}[(\tilde{e}_{\tilde{\theta}}^{(i)})^2] - \mathrm{E}[(\tilde{e}_\theta^{(i)})^2]\Big| \\
&\leq \sum_{i=1}^{N_T H} \Big|\mathrm{E}[(\tilde{e}_{\tilde{\theta}}^{(i)})^2] - \mathrm{E}[(\tilde{e}_\theta^{(i)})^2]\Big| \\
&\leq \sum_{i=1}^{N_T H} \Big|\mathrm{E}[(e_{\tilde{\theta}}^{(i)})^2] - \mathrm{E}[(e_\theta^{(i)})^2]\Big| \\
&= \sum_{i=1}^{N_T H} \Big|\mathrm{E}[(\theta^{*\top} z^{(i)} - \tilde{\theta}^\top z^{(i)})^2] \\
&\quad - \mathrm{E}[(\theta^{*\top} z^{(i)} - \theta^\top z^{(i)})^2]\Big| \\
&= \sum_{i=1}^{N_T H} \Big|\mathrm{E}[(\theta^{*\top} z^{(i)} + w^{(i)} - \tilde{\theta}^\top z^{(i)})^2] \\
&\quad - \mathrm{E}[(\theta^{*\top} z^{(i)} + w^{(i)} - \theta^\top z^{(i)})^2]\Big| \\
&\leq \sum_{i=1}^{N_T H} \max_{z^{(i)}, y^{(i)}} \Big|[(y^{(i)} - \tilde{\theta}^\top z^{(i)})^2] \\
&\quad - [(y^{(i)} - \theta^\top z^{(i)})^2]\Big|.
\end{aligned}$$

The second inequality uses the fact that for any $z^{(i)}$, we have $|(\tilde{e}_{\tilde{\theta}}^{(i)})^2 - (\tilde{e}_\theta^{(i)})^2| \leq |(e_{\tilde{\theta}}^{(i)})^2 - (e_\theta^{(i)})^2|$.
And thus we have

$$\begin{aligned}
&\Big|\sum_{i=1}^{N_T H} \mathrm{E}(\tilde{e}_{\tilde{\theta}}^{(i)})^2 - \mathrm{E}(\tilde{e}_\theta^{(i)})^2\Big| + \frac{1}{\kappa^2}\Big|\|\tilde{\theta}\|_2^2 - \|\theta\|_2^2\Big| \\
&\quad \leq \Delta^{(N_T H)}(\theta, \tilde{\theta}) \\
&\quad \leq \lambda. \qquad (62)
\end{aligned}$$

Combining Eqn. (61) and (62) gives us

$$\begin{aligned}
&\mathrm{loss}^{(N_T H)}(\theta) - \mathrm{loss}^{(N_T H)}(\theta^*) \\
&\quad \geq \sum_{i=1}^{N_T H} \mathrm{E}(\tilde{e}_\theta^{(i)})^2 - 3\lambda + \frac{1}{\kappa^2}(\|\theta\|_2^2 - \|\theta^*\|_2^2).
\end{aligned}$$

Now for any $\epsilon > 0, \eta > 0$, if $\theta$ satisfies

$$P(\max_{i \in 1:H}(e_\theta^{(i)}) > \epsilon) > \eta \qquad (63)$$

(note this corresponds to $\theta$ not satisfying Eqn. (22)) then we have that (let $\bar{\epsilon} = \min\{K_w, \epsilon\}$)

$$\begin{aligned}
&\mathrm{loss}^{(N_T H)}(\theta) - \mathrm{loss}^{(N_T H)}(\theta^*) \\
&\quad \geq N_T \bar{\epsilon}^2 \eta - 3\lambda + \frac{1}{\kappa^2}(\|\theta\|_2^2 - \|\theta^*\|_2^2)
\end{aligned}$$

holds w.p. $1 - |\tilde{S}|\exp(-\lambda^2/(8N_T H K_w^2))$.

Now choosing $\lambda = \frac{N_T\bar{\epsilon}^2\eta}{6}$ gives us that

$$\begin{aligned}&\text{loss}^{(N_TH)}(\theta) - \text{loss}^{(N_TH)}(\theta^*)\\ \geq\quad & \frac{N_T\bar{\epsilon}^2\eta}{2} + \frac{1}{\kappa^2}(\|\theta\|_2^2 - \|\theta^*\|_2^2).\end{aligned}$$

So if

$$N_T \geq \frac{4}{\bar{\epsilon}^2\eta\kappa^2}\|\theta^*\|_2^2 \tag{64}$$

then we have that

$$\text{loss}^{(N_TH)}(\theta) - \text{loss}^{(N_TH)}(\theta^*) \geq \frac{N_T\bar{\epsilon}^2\eta}{4}$$

holds w.p. $1 - |\tilde{S}|\exp(-(N_T\bar{\epsilon}^4\eta^2/(288HK_w^2))$.

We have from Prop. 18 that $\max_{i\in 1:N_TH} w^{(i)} \leq \sigma\log\frac{4N_TH}{\sqrt{2\pi}\delta'}$ with probability $1-\delta'$. So we can choose $K_w = \sigma\log\frac{4N_TH}{\sqrt{2\pi}\delta'}$ (and add in a failure probability of $\delta'$). Making the dependencies in $\tilde{S}$ (and recall we chose $\lambda = \frac{N_T\bar{\epsilon}^2\eta}{6}$) explicit, we have here $\tilde{S}(n, \frac{6}{N_T\bar{\epsilon}^2\eta}, \frac{1}{\delta}, \tilde{m}, \|\theta^*\|_2)$ and thus

$$\log|\tilde{S}| = O\left(n\log\text{poly}(n, \frac{1}{N_T\bar{\epsilon}^2\eta}, \frac{1}{\delta'}, \tilde{m}, \|\theta^*\|_2)\right),$$

Now we choose

$$\begin{aligned}\frac{\delta}{8} &= \delta'\\ &= |\tilde{S}|\exp(-(N_T\bar{\epsilon}^4\eta^2/(288H\sigma^2\log^2\frac{32N_TH}{\sqrt{2\pi}\delta})).\end{aligned}$$

This gives us the following conditions on $N_T$.

$$\begin{aligned}&\text{(i)} && \text{Eqn. (64)},\\ &\text{(ii)} && N_T \geq \frac{288H\sigma^2\log^2\frac{32N_TH}{\sqrt{2\pi}\delta}}{\bar{\epsilon}^4\eta^2}\log\frac{|\tilde{S}|}{\delta/8}.\end{aligned}$$

Recall $\bar{\epsilon} = \min\{K_w, \epsilon\}$. So we can replace conditions (i) and (ii) by the following conditions on $N_T$ (recall $K_w = \sigma\log\frac{4N_TH}{\sqrt{2\pi}\delta/8}$):

$$\begin{aligned}&\text{(ia)} && N_T \geq \frac{4}{\epsilon^2\eta\kappa^2}\|\theta^*\|_2^2\\ &\text{(ib)} && N_T \geq \frac{4}{(\sigma\log\frac{4N_TH}{\sqrt{2\pi}\delta/8})^2\eta\kappa^2}\|\theta^*\|_2^2\\ &\text{(iia)} && N_T \geq \frac{288H\sigma^2\log^2\frac{32N_TH}{\sqrt{2\pi}\delta}}{\epsilon^4\eta^2}\log\frac{|\tilde{S}|}{\delta/8},\\ &\text{(iib)} && N_T \geq \frac{288H\sigma^2\log^2\frac{32N_TH}{\sqrt{2\pi}\delta}}{(\sigma\log\frac{4N_TH}{\sqrt{2\pi}\delta/8})^4\eta^2}\log\frac{|\tilde{S}|}{\delta/8}.\end{aligned}$$

Combining the four conditions on $N_T$ with the expression for $|\tilde{S}|$ gives us the following condition on $N_T$ suffices:

$$N_T = \Omega(\text{poly}(\frac{1}{\epsilon}, \frac{1}{\eta}, \frac{1}{\delta}, H, \|\theta^*\|_2, n_S, n_{\mathcal{A}}, k_1, \log N)).$$

This proves the lemma. Note that in the statement of the lemma we have slightly weaker result, namely a polynomial dependence on $N$ rather than a polynomial dependence on $\log N$, which we proved here. □

### E.6. Proofs of Lemmas 21 and 22

We first prove the following lemma about a (possibly adversarial) biased random walk. We refer the reader to, e.g., (Durrett, 1995; Billingsley, 1995; Williams, 1991), for more details on martingales and stopping times.

**Lemma 24.** *Let $\{w^{(i)}\}_{i=1}^{\infty}$ be IID random variables with $w^{(i)} \sim \mathcal{N}(0, \tau^2)$. Let $\mathcal{F}_n = \sigma(w^{(1)}, \ldots, w^{(n)})$, the sigma algebra induced by these random variables. Let $\forall i, e^{(i)} \in \mathcal{F}_{i-1}$, i.e., $e^{(i)}$ has to be chosen based upon the past. Let $\forall n$,*

$$Z_n = \sum_{i=1}^{n}(e^{(i)})^2 + 2w^{(i)}e^{(i)}.$$

*Let any $a > 0$ be given. Let $T_a = \inf\{n : Z_n \leq -a\}$. Then we have*

$$P(T_a < \infty) \leq \exp(\frac{-a}{2\sigma^2}).$$

*Proof.* Let the martingale sequence $\{Y_n\}_n$ over the filtration $\{\mathcal{F}_n\}_n$ be defined as follows

$$\begin{aligned}Y_n &= \exp\left(\tfrac{-1}{2\sigma^2}Z_n\right)\\ &= \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(2w^{(i)}e^{(i)} + (e^{(i)})^2)\right).\end{aligned}$$

It is easily verified that $\{Y_n\}_n$ is adapted to $\{\mathcal{F}_n\}$, that $E|Y_n| < \infty$ and that $E(Y_{n+1}|\mathcal{F}_n) = Y_n$ for all $n$. Thus $Y_n$ is indeed a martingale with respect to $\{\mathcal{F}_n\}_n$. (Note this is true no matter what the adversary's policy is for choosing the $\mathcal{F}_{i-1}$-measurable functions $e^{(i)}$.)

Let any integer $K > 0$ be fixed. Let

$$\begin{aligned}T_b &= \inf\{n : Z_n \geq b\},\\ N &= \min\{T_a, T_b, K\}.\end{aligned}$$

Then $N$ is a finite stopping time. Thus we can apply the Optional Stopping Theorem[15] and get

$$\begin{aligned}1 &= \text{E}Y_0 = \text{E}Y_N\\ &= P(T_a < T_b, T_a < K)\text{E}[Y_N|T_a < T_b, T_a < K]\\ &\quad + P(T_b < T_a, T_b < K)\text{E}[Y_N|T_b < T_a, T_b < K]\\ &\quad + P(K \leq T_b, K \leq T_a)\text{E}[Y_N|K < T_b, K < T_a].\end{aligned}$$

Now since $\forall n, Y_n \geq 0$ we have

$$1 \geq P(T_a < T_b, T_a < K)\text{E}[Y_N|T_a < T_b, T_a < K].$$

[15] See, e.g., Durrett, 1995.

Using $\mathrm{E}[Y_N | T_a < T_b, T_a < K] \geq \exp(\frac{a}{2\sigma^2})$ we get

$$P(T_a < T_b, T_a < K) \leq \exp(\frac{-a}{2\sigma^2}).$$

Taking the limit for $K \to \infty$ (and using the monotone convergence theorem which allows us to interchange limit and expectation (probability)) gives us

$$P(T_a < T_b, T_a < \infty) \leq \exp(\frac{-a}{2\sigma^2}).$$

Since $b > 0$ was arbitrary, we get for $b \to \infty$ (and using the monotone convergence theorem which allows us to interchange limit and expectation (probability)) that

$$P(T_a < \infty) \leq \exp(\frac{-a}{2\sigma^2}).$$

□

*Proof of Lemma 21.* We have $y^{(i)} = \theta^{*\top} z^{(i)} + w^{(i)}$, where $w^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Let $e^{(i)} = \theta^{*\top} z^{(i)} - \theta^\top z^{(i)}$. Then

$$\begin{aligned} &\mathrm{loss_{adv}}^{(k)}(\theta) - \mathrm{loss_{adv}}^{(k)}(\theta^*) \\ &= \sum_{i=m+1}^{k} (\theta^{*\top} z^{(i)} + w^{(i)} - \theta^\top z^{(i)})^2 - (w^{(i)})^2 \\ &= \sum_{i=m+1}^{k} (w^{(i)} + e^{(i)})^2 - (w^{(i)})^2 \\ &= \sum_{i=m+1}^{k} (e^{(i)})^2 + 2w^{(i)} e^{(i)}. \end{aligned}$$

So we can apply Lemma 24 with $Z_n = \mathrm{loss_{adv}}^{(n+m)}(\theta) - \mathrm{loss_{adv}}^{(n+m)}(\theta^*)$, which proves the lemma. □

*Proof of Lemma 22.* Using a $\lambda = N_T \epsilon^2 \eta / 16$ cover $\tilde{S}$ for $S$ and Lemma 21 gives us that $\forall k \geq 1$ and for all $\theta \in S$ that

$$\mathrm{loss}^{(k)}_{\mathrm{adv}}(\theta) - \mathrm{loss}^{(k)}_{\mathrm{adv}}(\theta^*) \geq -\lambda - N_T \epsilon^2 \eta/16 = -N_T \epsilon^2 \eta / 8$$

holds w.p.

$$1 - |\tilde{S}| \exp(-N_T \epsilon^2 \eta / (32\sigma^2)).$$

The last term corresponds to the probability of the biased random walk reaching $-N_T \epsilon^2 \eta/16$. Now requiring that the last term is smaller than $\frac{\delta}{4}$ gives us the following requirement:

$$N_T \geq \frac{32\sigma^2}{\epsilon^2 \eta} \log \frac{|\tilde{S}(n, \frac{16}{N_T \epsilon^2 \eta}, \frac{2}{\delta}, \tilde{m}, \|\theta^*\|_2)|}{\delta/4},$$

which is satisfied when $N_T$ satisfies Eqn. (53). □

### E.7. Proof of Theorem 9

*Proof of Theorem 9.* From Lemma 8 we have that for a trial under the teacher's policy ($\{(x_t, u_t)\}_{t=1}^H$) for

$$\begin{aligned} P\big(&\max\nolimits_{t \in 1:H} \|A^{(i)} \phi(x_t) + B^{(i)} u_t \\ &-(A\phi(x_t) + B u_t)\|_2 > \epsilon\big) \leq \eta \end{aligned} \tag{65}$$

to hold with probability $1 - \delta$ for all $i \in 1 : N$, it suffices that

$$N_T = \Omega\Big(\mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\eta}, \frac{1}{\delta}, H, \|A\|_{\mathrm{F}}, \|B\|_{\mathrm{F}}, n_S, n_{\mathcal{A}}, k_1, N)\Big). \tag{66}$$

From Prop. 7 and Eqn. (65) we have that for

$$\begin{aligned} P \quad &\Big(\max\nolimits_{t \in 1:H} d_{\mathrm{var}}(P(\cdot\, | x_t, u_t), \\ &P^{(i)}(\cdot\, | x_t, u_t)) > \frac{1}{\sqrt{2\pi}\sigma}\epsilon\Big) \leq \eta \end{aligned} \tag{67}$$

to hold for all $i \in 1 : N$ with probability $1 - \delta$ it is sufficient that $N_T$ satisfies Eqn. (66). Let $\overline{SA}_\eta = \{(x, u) : \forall i \;\; d_{\mathrm{var}}(P(\cdot\, | x, u), P^{(i)}(\cdot\, | x, u)) \leq \frac{\epsilon}{\sqrt{2\pi}\sigma}\}$. Then using the Simulation Lemma combined with Eqn. (67) we obtain that that for

$$|U_M(\pi_T) - U_{M^{(i)}}(\pi_T)| \leq H^2 \frac{1}{\sqrt{2\pi}\sigma} \epsilon R_{\max} + \eta H R_{\max}$$

to hold for all $i \in 1 : N$ with probability $1 - \delta$, it suffices that $N_T$ satisfies Eqn. (66). Choosing $\epsilon = \frac{\sqrt{2\pi}\sigma}{2H^2 R_{\max}} \alpha$ and $\eta = \frac{1}{2HR_{\max}} \alpha$ proves the theorem. □

## F. More elaborate/detailed version of Section 7.3

### F.1. A result for Bayesian model averaging

Consider an adversary generating a data sequence $\{z^{(t)}\}_{t=1}^T$. For every time step $t$, $w^{(t)} \sim \mathcal{N}(0, \sigma^2)$, and $y^{(t)} = \theta^{*\top} z^{(t)}$. We assume $\|z^{(t)}\|_2^2 \leq 2$.[16] Now define a sequence $\{\hat{\theta}^{(t)}\}_{t=1}^T$ of estimates of $\theta^*$

$$\begin{aligned} \hat{\theta}^{(t)} &= \arg\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^{t} (y^{(i)} - \theta^\top z^{(i)})^2 + \frac{1}{\kappa^2}\|\theta\|_2^2 \\ &= \arg\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^{t} \frac{1}{\sigma^2}(y^{(i)} - \theta^\top z^{(i)})^2 + \frac{1}{\nu^2}\|\theta\|_2^2, \end{aligned}$$

here $\nu^2 = \kappa^2 \sigma^2$. Let $e^{(t)} = \theta^{*\top} z^{(t)} - \hat{\theta}^{(t)\top} z^{(t)}$. In this section, we will prove the following theorem

[16] Since $z^{(t)}$ later corresponds to the concatenation of $\phi(x^{(t)})$ and $u^{(t)}$, which both have norm smaller than one, this is the right choice. Kakade and Ng (2005) use $\|z^{(t)}\|_2 \leq 1$, which makes their results slightly different from the way we state their results in this paper.

**Theorem 25.** *Let everything be as above defined. Then no matter how the adversary chooses each $z^{(t)}$ (possibly based on everything seen up to time $t-1$), we have that with probability $1-\delta$*

$$N_\mu = \sum_{t=1}^{T} \mathbf{1}\{e^{(t)} \geq \mu\} \\ \leq \quad O(\sqrt{T}(\log T)^3 \mathrm{poly}(\|\theta^*\|_2, n, \log\frac{1}{\delta}, \frac{1}{\mu})).$$

This result will be obtained in the following three steps:

- prove an online log-loss bound for Bayesian model averaging (BMA),
- prove a bound on the variances $s_t^2$ used at every step in the BMA algorithm, in particular prove a bound on how often these variances can be 'large',
- prove a bound on the squared loss incurred for the time-steps when the variances $s_t^2$ are 'small'.

We now consider the Bayesian model averaging (BMA) algorithm, and give a bound on its worst-case online loss. In particular we consider the case of linear least squares regression. We have

$$p(y|z,\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(\theta^T z - y)^2}{2\sigma^2}\right), \qquad (68)$$

where $\sigma^2$ is a fixed, *known* constant that is not a parameter of our model. Note that Kakade and Ng (2005) give results for generalized linear models. This section reviews a subset of their results, where notation is specialized to the linear regression case.

Let $S = \{(z^{(1)}, y^{(1)}), (z^{(2)}, y^{(2)}), \ldots, (z^{(T)}, y^{(T)})\}$ be an arbitrary sequence of examples, possibly chosen by an adversary. We also use $S_t$ to denote the subsequence consisting of only the first $t$ examples. Unless otherwise stated, we will assume throughout this section that $||z^{(i)}||_2 \leq \sqrt{2}$ for all $i$.

Assume that we are going to use a Bayesian algorithm to make our online predictions. Specifically, assume that we have a Gaussian prior on the parameters:

$$p(\theta) = \mathcal{N}(\theta; \vec{0}, \nu^2 I_n),$$

where $I_n$ is the $n$-by-$n$ identity matrix, $\mathcal{N}(\cdot;\mu,\Sigma)$ is the density of a Gaussian with mean $\mu$ and covariance $\Sigma$, and $\nu^2 > 0$ is some fixed constant governing the variance in our prior. Also, let

$$p_t(\theta) = p(\theta|S_t) = \frac{\left(\prod_{i=1}^t p(y^{(i)}|x^{(i)},\theta)\right) p(\theta)}{\int_\theta \left(\prod_{i=1}^t p(y^{(i)}|x^{(i)},\theta)\right) p(\theta) d\theta}$$

be the posterior distribution over $\theta$ given the first $t$ training examples. We also have that $p_0(\theta) = p(\theta)$ is just the prior distribution.

On iteration $t$, we are given the input $x^{(t)}$, and the BMA algorithm makes a prediction using the posterior distribution over the outputs:

$$p(y|z^{(t)}, S_{t-1}) = \int_\theta p(y|z^{(t)},\theta) p(\theta|S_{t-1}) d\theta.$$

We are then given the true label $y^{(t)}$, and we suffer logloss $-\log p(y^{(t)}|z^{(t)}, S_{t-1})$. We define the cumulative loss of the BMA algorithm after $T$ rounds to be

$$L_{\mathrm{BMA}}(S) = \sum_{t=1}^{T} -\log p(y^{(t)}|z^{(t)}, S_{t-1}).$$

We will be interested in comparing against the loss of any expert that uses some fixed parameters $\theta \in \mathbb{R}^n$. Define $\ell_\theta(t) = -\log p(y^{(t)}|x^{(t)},\theta)$, and let

$$L_\theta(S) = \sum_{t=1}^{T} \ell_\theta(t) = \sum_{t=1}^{T} -\log p(y^{(t)}|z^{(t)},\theta).$$

A more general form of the following theorem has been proved in Kakade and Ng (2005), we specialized it to the linear regression case.

**Theorem 26.** *[(Kakade & Ng, 2005) Theorem 2.2, for $c = 1/\sigma^2$.] For all sequences $S$ of length $T$ and for all $\theta^*$*

$$L_{BMA}(S) \leq L_{\theta^*}(S) + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

Now we'll take a closer look at the predictions done by the BMA algorithm. Define $A_t = \frac{1}{\nu^2} I_n + \frac{1}{\sigma^2}\sum_{i=1}^t z^{(i)} {z^{(i)}}^T$, and $b_t = \frac{1}{\sigma^2}\sum_{i=1}^t z^{(i)} y^{(i)}$. We have that

$$p_t(\theta) = p(\theta|S_t) = \mathcal{N}\left(\theta; \hat{\theta}_t, \hat{\Sigma}_t\right), \qquad (69)$$

where $\hat{\theta}_t = A_t^{-1} b_t$, and $\hat{\Sigma}_t = A_t^{-1}$. Also, the predictions are given by

$$p(y^{(t+1)}|z^{(t+1)}, S_t) = \mathcal{N}\left(y^{(t+1)}; \hat{y}_{t+1}, s_{t+1}^2\right) \qquad (70)$$

where $\hat{y}_{t+1} = \hat{\theta}_t^T z^{(t+1)}$, $s_{t+1}^2 = {z^{(t+1)}}^T \hat{\Sigma}_t z^{(t+1)} + \sigma^2$. In contrast, the prediction of a fixed expert using parameter $\theta^*$ would be

$$p(y^{(t)}|z^{(t)}, \theta^*) = \mathcal{N}\left(y^{(t)}; y_t^*, \sigma^2\right), \qquad (71)$$

where $y_t^* = {\theta^*}^T z^{(t)}$.

Note that $s_t^2 \geq \sigma^2$, i.e., the BMA algorithm always predicts with a larger variance than a single expert.

**Lemma 27.** *The terms $s_t^2$ satisfy the following:*

$$\begin{aligned} \frac{s_t^2-\sigma^2}{\sigma^2} &\le \frac{2\nu^2}{\sigma^2}, \qquad (72)\\ \sum_{t=1}^T \frac{s_t^2-\sigma^2}{\sigma^2} &\le \frac{2\nu^2}{\sigma^2\log(1+\frac{2\nu^2}{\sigma^2})} n\log\left(1+\frac{2T\nu^2}{n\sigma^2}\right). \qquad (73)\end{aligned}$$

*Proof.* Let $m_t = \frac{s_t^2-\sigma^2}{\sigma^2}$. Consider a sequence of examples $\{(z^{(1)}, y^{(i)}), \ldots, (z^{(T)}, y^{(T)})\}$, where all the outputs $y^{(1)} = \cdots = y^{(T)} = 0$. Given this sequence, the BMA algorithm's predictions will also all be $\hat{y}_t = 0$. Thus, we have

$$\begin{aligned} L_{\mathrm{BMA}}(T) &= \sum_{t=1}^T -\log\mathcal{N}(0;0,s_t^2) \qquad (74)\\ &= \sum_{t=1}^T \left[-\log\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\log s_t^2\right]. \qquad (75)\end{aligned}$$

Now, consider the loss of an expert using the zero parameter vector $\theta = \vec{0}$. We have

$$\begin{aligned} L_{\theta^*}(T) &= \sum_{t=1}^T -\log\mathcal{N}(0;0,\sigma^2) \qquad (76)\\ &= \sum_{t=1}^T \left[-\log\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\log \sigma^2\right]. \qquad (77)\end{aligned}$$

Substituting Equations (75) and (77) into the main conclusion of Theorem 26, we get

$$\sum_{t=1}^T \frac{1}{2}\log s_t^2 \le \sum_{t=1}^T \frac{1}{2}\log\sigma^2 + \frac{n}{2}\log\left(1+\frac{2T\nu^2}{n\sigma^2}\right). \tag{78}$$

Using the definition $m_t = s_t^2/\sigma^2 - 1$, we get

$$\sum_{t=1}^T \log(1+m_t) \le n\log\left(1+\frac{2T\nu^2}{n\sigma^2}\right). \tag{79}$$

Finally, observe that for all $0 \le \epsilon \le K$, we have

$$\log(1+\epsilon) \ge \frac{\log(1+K)}{K}\cdot\epsilon \tag{80}$$

Also, we can bound $m_t$ as follows:

$$\begin{aligned} m_t &= \frac{s_t^2}{\sigma^2} - 1\\ &= \frac{{z^{(t+1)}}^T\hat{\Sigma}_t z^{(t+1)} + \sigma^2}{\sigma^2} - 1\\ &= \frac{1}{\sigma^2}{z^{(t+1)}}^T\left(\frac{1}{\nu^2}I_n + \frac{1}{\sigma^2}\sum_{i=1}^t z^{(t)}{z^{(t)}}^T\right)^{-1} z^{(t+1)}\\ &\le \frac{1}{\sigma^2}{z^{(t+1)}}^T\left(\frac{1}{\nu^2}I_n\right)^{-1} z^{(t+1)}\\ &\le \frac{2\nu^2}{\sigma^2}. \qquad (81)\end{aligned}$$

For the last step, we used the fact that $\|z^{(t+1)}\|_2 \le \sqrt{2}$. This shows (72).

Putting together (81) and (80) with $\epsilon = m_t$ and $K = 2\nu^2/\sigma^2$, we find that

$$m_t \le \frac{2\nu^2}{\sigma^2\log(1+\frac{2\nu^2}{\sigma^2})}\log(1+m_t). \tag{82}$$

Finally, Equations (82) and (79) together imply (73).

□

*Proof of Theorem 25.* As a direct consequence of Eqn. (73) of Lemma 27, we have the following bound for the number of times $N_{s_t^2>\sigma^2(1+\epsilon^2)}$ that $s_t^2 > \sigma^2 + \epsilon^2\sigma^2$:

$$N_{s_t^2>\sigma^2(1+\epsilon^2)} \le \frac{1}{\epsilon^2}\frac{2\nu^2}{\sigma^2\log(1+\frac{2\nu^2}{\sigma^2})}n\log(1+\frac{2T\nu^2}{n\sigma^2}). \tag{83}$$

If we let $\theta^*$ denote the "true" underlying parameter, and $w^{(t)}$ the noise at time $t$, then we have

$$y^{(t)} = \theta^{*\top}z^{(t)} + w^{(t)}.$$

Using the notation we just introduced we can then rewrite Theorem 26 as

$$\begin{aligned} &\sum_{t=1}^T\left(\frac{1}{2s_t^2}\left(w^{(t)} + \theta^{*\top}z^{(t)} - \hat{\theta}_{t-1}^\top z^{(t)}\right)^2 + \log\sqrt{2\pi}s_t\right) \le\\ &\quad\sum_{t=1}^T\left(\frac{1}{2\sigma^2}\left(w^{(t)}\right)^2 + \log\sqrt{2\pi}\sigma\right)\\ &\quad+\frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1+\frac{2T\nu^2}{n\sigma^2}\right).\end{aligned}$$

From now on let $e^{(t)} = \theta^{*\top}z^{(t)} - \hat{\theta}_{t-1}^\top z^{(t)}$.

Splitting up the summation into case where $s_t^2 > \sigma^2(1+\epsilon^2)$ and $s_t^2 \le \sigma^2(1+\epsilon^2)$, and leaving out some positive terms from the left-handside, and using the fact that $s_t \ge \sigma$ for all $t$ we get:

$$\begin{aligned} &\sum_{t:s_t^2\le\sigma^2(1+\epsilon^2)}\frac{1}{2s_t^2}\left(w^{(t)} + e^{(t)}\right)^2 \le\\ &\sum_{t=1}^T\frac{1}{2\sigma^2}\left(w^{(t)}\right)^2 + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1+\frac{2T\nu^2}{n\sigma^2}\right),\end{aligned}$$

which implies

$$\begin{aligned} &\sum_{t:s_t^2\le\sigma^2(1+\epsilon^2)}\frac{1}{2\sigma^2(1+\epsilon^2)}\left(w^{(t)} + e^{(t)}\right)^2 \le\\ &\sum_{t=1}^T\frac{1}{2\sigma^2}\left(w^{(t)}\right)^2 + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1+\frac{2T\nu^2}{n\sigma^2}\right).\end{aligned} \tag{84}$$

Now we would like to bound the number of times we can have that $e^{(t)} > \mu$ as a function of $\mu$. Since $e^{(t)}$ could depend on the whole history $\{y^{(i)}, z^{(i)}\}_{i=1}^{t-1}$, we use a martingale argument.

Let $K_w$ be such that for all $t$ we have $|w^{(t)}| \leq K_w$ and such that $\mu \leq K_w$. (The latter condition ensures that when clipping the errors, errors do not get clipped below $\mu$.) Let $\tilde{e}^{(t)}$ be defined as $e^{(t)}$ clipped to the interval $[-K_w, K_w]$. I.e., we define $\tilde{e}^{(t)} = \min\{K_w, \max\{-K_w, e^{(t)}\}\}$. Then we have

$$(w^{(t)} + \tilde{e}^{(t)})^2 \leq (w^{(t)} + e^{(t)})^2.$$

For the left-hand side of Eqn. (84) we apply Azuma's inequality to the martingale $Z_t = \sum_{i=1}^{t} (w^{(i)} + \tilde{e}^{(i)})^2 - \sigma^2 - (\tilde{e}^{(i)})^2$, which gives (the increment/decrement in one time step is at most $3K_w^2 + \sigma^2$)

$$P\left(\sum_{t:s_t^2 \leq \sigma^2(1+\epsilon^2)} (w^{(t)} + \tilde{e}^{(t)})^2 - (\tilde{e}^{(t)})^2 - \bar{T}\sigma^2 < -\bar{T}\lambda\right) \leq \exp\left(-\bar{T}\lambda^2/2(3K_w^2 + \sigma^2)^2\right), \tag{85}$$

where $\bar{T} = T - N_{s_t^2 > \sigma^2(1+\epsilon^2)}$.

For the right-hand side of Eqn. (84) we apply Azuma's inequality to the martingale $Z_t = \sum_{i=1}^{t} (w^{(i)})^2 - \sigma^2$, which gives us (the increment/decrement in one time step is bounded by $K_w^2 + \sigma^2$)

$$P\left(\sum_{t} (w^{(t)})^2 - T\sigma^2 > T\lambda\right) \leq \exp(-T\lambda^2/2(K_w^2 + \sigma^2)^2). \tag{86}$$

Combining the concentration results of Eqn. (85,86) with Eqn. (84) gives that with probability $1 - \exp(-\bar{T}\lambda^2/2(3K_w^2 + \sigma^2)^2) - \exp(-T\lambda^2/2(K_w^2 + \sigma^2)^2)$ the following holds:

$$\sum_{t:s_t^2 \leq \sigma^2(1+\epsilon^2)} \frac{1}{2\sigma^2(1+\epsilon^2)} (\tilde{e}^{(t)})^2 + \bar{T}\frac{\sigma^2 - \lambda}{2\sigma^2(1+\epsilon^2)} \leq T\frac{\sigma^2 + \lambda}{2\sigma^2} + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

Which is equivalent to

$$\sum_{t:s_t^2 \leq \sigma^2(1+\epsilon^2)} \frac{1}{2\sigma^2(1+\epsilon^2)} (\tilde{e}^{(t)})^2 \leq T\frac{\sigma^2 + \lambda}{2\sigma^2} - \bar{T}\frac{\sigma^2 - \lambda}{2\sigma^2(1+\epsilon^2)} + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

As a consequence we have

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \frac{\mu^2}{2\sigma^2(1+\epsilon^2)} = \sum_{t:s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \frac{1}{2\sigma^2(1+\epsilon^2)}\mu^2 \leq T\frac{\sigma^2 + \lambda}{2\sigma^2} - \bar{T}\frac{\sigma^2 - \lambda}{2\sigma^2(1+\epsilon^2)} + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

Substituting $\bar{T} = T - N_{s_t^2 > \sigma^2(1+\epsilon^2)}$ (and using $-\lambda < +\lambda$) we get

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \frac{\mu^2}{2\sigma^2(1+\epsilon^2)} \leq T\frac{\epsilon^2\sigma^2 + \epsilon^2\lambda + 2\lambda}{2\sigma^2(1+\epsilon^2)} + N_{s_t^2 > \sigma^2(1+\epsilon^2)} \frac{\sigma^2 + \lambda}{2\sigma^2(1+\epsilon^2)} + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

Substituting the bound for $N_{s_t^2 > \sigma^2(1+\epsilon^2)}$ from Eqn. (83) we get

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \frac{\mu^2}{2\sigma^2(1+\epsilon^2)} \leq T\frac{\epsilon^2\sigma^2 + \epsilon^2\lambda + 2\lambda}{2\sigma^2(1+\epsilon^2)} + \frac{\sigma^2 + \lambda}{2\sigma^2(1+\epsilon^2)} \frac{1}{\epsilon^2} \frac{2\nu^2}{\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})} n \log(1 + \frac{2T\nu^2}{n\sigma^2}) + \frac{1}{2\nu^2}\|\theta^*\|^2 + \frac{n}{2}\log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

Multiplying both sides with $2\sigma^2(1+\epsilon^2)$ gives

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \mu^2 \leq T(\epsilon^2\sigma^2 + \epsilon^2\lambda + 2\lambda) + (\sigma^2 + \lambda)\frac{1}{\epsilon^2} \frac{2\nu^2}{\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})} n \log(1 + \frac{2T\nu^2}{n\sigma^2}) + \frac{\sigma^2(1+\epsilon^2)}{\nu^2}\|\theta^*\|^2 + n\sigma^2(1+\epsilon^2)\log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

This all holds w.p.

$$1 - \exp(-\bar{T}\lambda^2/2(3K_w^2 + \sigma^2)^2) - \exp(-T\lambda^2/2(K_w^2 + \sigma^2)^2).$$

Now choose

$$\epsilon^2 = \sqrt{\frac{2\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{T\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}, \tag{87}$$

then we get

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \mu^2 \leq 2\lambda T$$
$$+2(\sigma^2 + \lambda)\sqrt{\frac{2T\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}$$
$$+ \left(1 + \sqrt{\frac{2\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{T\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}\right) \frac{\sigma^2}{\nu^2} \|\theta^*\|^2$$
$$+ \left(1 + \sqrt{\frac{2\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{T\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}\right) n\sigma^2 \log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

Substituting Eqn. (87) into Eqn. (83) gives

$$N_{s_t^2 > \sigma^2(1+\epsilon^2)} \leq \sqrt{T \frac{2\nu^2}{\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})} n \log(1 + \frac{2T\nu^2}{n\sigma^2})}.$$

So we have that there exists $T^* = O(\text{poly}(n))$ (note we assume $\nu, \sigma$ are fixed in our analysis) such that for all $T > T^*$ we have that $N_{s_t^2 > \sigma^2(1+\epsilon^2)} \leq \frac{1}{2}T$. Thus (for $T > T^*$) all the above holds w.p.

$$1 - 2\exp(-T\lambda^2/4(3K_w^2 + \sigma^2)^2).$$

Now setting $\frac{\delta}{2} = 2\exp(-T\lambda^2/4(3K_w^2 + \sigma^2)^2)$, or equivalently $\lambda^2 = \frac{4(3K_w^2+\sigma^2)^2}{T} \log\frac{4}{\delta}$ we get that

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \mu^2 \leq 2T\sqrt{\frac{4(3K_w^2 + \sigma^2)^2}{T} \log\frac{4}{\delta}}$$
$$+ 2(\sigma^2 + \sqrt{\frac{4(3K_w^2 + \sigma^2)^2}{T} \log\frac{4}{\delta}})\sqrt{\frac{2T\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}$$
$$+ \left(1 + \sqrt{\frac{2\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{T\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}\right) \frac{\sigma^2}{\nu^2} \|\theta^*\|^2$$
$$+ \left(1 + \sqrt{\frac{2\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{T\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}\right) n\sigma^2 \log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right). \tag{88}$$

holds with probability $1 - \frac{\delta}{2}$. Now recall that $K_w = \max\{\mu, \max_{t \in 1:T} |w^{(t)}|\}$. Thus from Prop. 18 we have that $K_w \leq \max\{\mu, \sigma \log \frac{8T}{\sqrt{2\pi\delta}}\}$ with probability $1 - \frac{\delta}{2}$. Thus we have that

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \mu^2 \leq$$
$$4(3(\max\{\mu, \sigma \log \frac{8T}{\sqrt{2\pi}\delta}\})^2 + \sigma^2)\sqrt{T \log\frac{4}{\delta}}$$
$$+ 2\left(\sigma^2 + 2(3(\max\{\mu, \sigma \log \frac{8T}{\sqrt{2\pi}\delta}\})^2 + \sigma^2)\sqrt{\frac{1}{T} \log\frac{4}{\delta}}\right)$$
$$\sqrt{\frac{2T\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}$$
$$+ \left(1 + \sqrt{\frac{2\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{T\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}\right) \frac{\sigma^2}{\nu^2} \|\theta^*\|^2$$
$$+ \left(1 + \sqrt{\frac{2\nu^2 n \log(1 + \frac{2T\nu^2}{n\sigma^2})}{T\sigma^2 \log(1 + \frac{2\nu^2}{\sigma^2})}}\right) n\sigma^2 \log\left(1 + \frac{2T\nu^2}{n\sigma^2}\right).$$

holds with probability $1 - \delta$.

After simplification we get that for any $T > T^* = O(\text{poly}(n))$ that

$$N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} =$$
$$O(\sqrt{T}(\log T)^3 \text{poly}(\|\theta^*\|_2, n, \log\frac{1}{\delta}, \frac{1}{\mu})) \tag{89}$$

holds with probability $1 - \delta$. The condition that $T > T^* = O(\text{poly}(n))$ is readily incorporated by adjusting the polynomial in Eqn. (89), and can thus be omitted.

Taking into account that we might have $e^{(t)} \geq \mu$ when $s_t^2 \geq (1 + \epsilon^2)\sigma^2$, we get that

$$\begin{aligned} N_{e^{(t)} \geq \mu} &\leq N_{s_t^2 \leq \sigma^2(1+\epsilon^2), e^{(t)} \geq \mu} \\ &\quad + N_{s_t^2 \leq \sigma^2(1+\epsilon^2)} \\ &= O(\sqrt{T}(\log T)^3 \text{poly}(\|\theta^*\|_2, n, \log\frac{1}{\delta}, \frac{1}{\mu})). \end{aligned}$$

Which proves the theorem. □

### F.2. Proof of Lemma 10

Lemma 10 considers the setting where the model is not updated for $k_1 H$ steps. And then updated for all these steps at once. The result we have from Theorem 25 applies only to the setting where the updates are done between every datapoint. Moreover Lemma 10 considers $n$ linear regression problems simultaneously.

**Lemma 28.** *Let any $l \in 1 : n$ be fixed. For the algorithm described in Section 4 we have that the number $N_\mu$ of times a state is encountered such that*

$$|(A_{l,:}\phi(x_t) + B_{l,:}u_t) - (\hat{A}_{l,:}^{(i)}\phi(x_t) - \hat{B}_{l,:}^{(i)}u_t)| > \mu \tag{90}$$

*satisfies*

$$\begin{aligned} N_\mu &= O(k_1 H \sqrt{N k_1 H}(\log N k_1 H)^3 \\ &\quad \text{poly}(\|A_{l,:}\|_F, \|B_{l,:}\|_F, n_S, n_{\mathcal{A}}, \log\frac{1}{\delta}, \frac{1}{\mu})) \end{aligned}$$

*Proof.* Consider $k_1 H$ versions of the data, permuted such that within each subsequence of length $k_1 H$ obtained under one policy, in every permutation a different data point comes first, but no data points are permutated accross trials under different policies. Then every prediction done in our algorithm is also done for (at least) one permutation in this new setup (the new setup includes more predictions than just these). Thus bounding the number of large errors in this new setup gives us a bound on the number of large errors encountered in our algorithm. This new setup consists of $k_1 H$ data sequences of length $N k_1 H$ each. Using Theorem 25 we get that

$$\begin{aligned} N_\mu &= O(k_1 H \sqrt{N k_1 H} (\log N k_1 H)^3 \\ &\quad \text{poly}(\sqrt{\|A_{l,:}\|_F^2, \|B_{l,:}\|_F^2}, n_S + n_{\mathcal{A}}, \log \frac{1}{\delta}, \frac{1}{\mu})) \end{aligned}$$

which can be simplified (and be made less tight) to the statement of the lemma. □

*Proof of Lemma 10.* When $x_t, u_t$ satisfy $\|(A\phi(x) + Bu) - (\hat{A}^{(i)}\phi(x) - \hat{B}^{(i)}u)\|_2 > \mu$, then there must be $l \in 1 : n_S$ such that

$$|(A_{l,:}\phi(x_t) + B_{l,:}u_t) - (\hat{A}_{l,:}^{(i)}\phi(x_t) - \hat{B}_{l,:}^{(i)}u_t)| > \mu/\sqrt{n_S}$$

is satisfied. From Lemma 28 we have that this can happen at most $N' = O(k_1 H \sqrt{N k_1 H} (\log N k_1 H)^3 \text{poly}(\|A_{l,:}\|_F, \|B_{l,:}\|_F, n_S, n_{\mathcal{A}}, \log \frac{1}{\delta}, \frac{1}{\mu}))$ times for each $l$ w.p. $1 - \delta$. So it can happen at most $n_S N'$ times w.p. $1 - n_S \delta$. This proves the lemma. □

### F.3. Proof of Theorem 3 for linearly parameterized dynamics

*Proof of Theorem 3.* Assume the algorithm runs for $N$ iterations. Using the Hoeffding inequality we have that for

$$\forall i = 1 : N \quad |\hat{U}_M(\pi^{(i)}) - U_M(\pi^{(i)})| \leq \frac{\alpha}{16} \tag{91}$$

to hold with probability $1 - \delta'$, it suffices that

$$k_1 \geq \frac{16^2 H^2 R_{\max}^2}{2\alpha^2} \log \frac{2N}{\delta'}. \tag{92}$$

Using the Hoeffding inequality, we also have that for

$$|\hat{U}_M(\pi_T) - U_M(\pi_T)| \leq \frac{\alpha}{16} \tag{93}$$

to hold with probability $1 - \delta''$, it suffices that

$$N_T \geq \frac{16^2 H^2 R_{\max}^2}{2\alpha^2} \log \frac{2}{\delta''}. \tag{94}$$

From Theorem 9 we have that for

$$|U_{\hat{M}^{(i)}}(\pi_T) - U_M(\pi_T)| \leq \frac{\alpha}{8} \tag{95}$$

to hold with probability $1 - \delta'''$, it suffices that $N_T =$

$$\Omega\left(\text{poly}(\frac{1}{\alpha}, \frac{1}{\delta'''}, H, R_{\max}, \|A\|_{\text{F}}, \|B\|_{\text{F}}, n_S, n_{\mathcal{A}}, k_1, N)\right). \tag{96}$$

Since the algorithm only exits in iteration $N$, we must have for all $i = 1 : N - 1$ that

$$\hat{U}_M(\pi^{(i)}) < \hat{U}_M(\pi_T) - \alpha/2. \tag{97}$$

Combining Eqn. (97, 91, 93, 95) and the fact that $\pi^{(i)}$ is $\alpha/8$-optimal for $\hat{M}^{(i)}$ we get

$$\forall i (1 \leq i \leq N-1) U_{\hat{M}^{(i)}}(\pi^{(i)}) \geq U_M(\pi^{(i)}) + \alpha/8. \tag{98}$$

Eqn. (98) states that for every iteration $i$ that the algorithm continues, the model is inaccurate in evaluating the utility of the policy $\pi^{(i)}$. Now using the contrapositive of the Simulation Lemma (choosing $\epsilon = \frac{1}{2}\frac{\alpha/8}{H^2 R_{\max}}$) we get from Eqn. (98) that the policy $\pi^{(i)}$ must be visiting a state-action pair $(x, u)$ that satisfies

$$d_{\text{var}}(P(\cdot|x,u), \hat{P}^{(i)}(\cdot|x,u)) > \tfrac{\alpha}{16H^2 R_{\max}} \tag{99}$$

with probability at least $\frac{\alpha}{16 H R_{\max}}$ in every trial of horizon $H$. If $(x, u)$ satisfies Eqn. (99) then we must have (using Prop. 7) that

$$\|(A\phi(x) + Bu) - (\hat{A}^{(i)}\phi(x) - \hat{B}^{(i)}u)\|_2 > \tfrac{\sqrt{2\pi}\sigma\alpha}{16H^2 R_{\max}}.$$

From Lemma 10 we have that with probability $1 - \delta''''$ this can happen only

$$\begin{aligned} N_\mu &= O(k_1 \sqrt{k_1 N} (\log k_1 N)^3 \text{poly}(\|A\|_F, \\ &\quad \|B\|_F, n_S, n_{\mathcal{A}}, \log \frac{1}{\delta''''}, \frac{1}{\alpha}, H, \frac{16H^2 R_{\max}}{\sqrt{2\pi}\sigma\alpha})) \end{aligned}$$

times in $N$ iterations of the algorithm. Substituting in the expression for $k_1$ from Eqn. (92) and simplifying gives us

$$\begin{aligned} N_\mu &= O(\sqrt{N} (\log N)^5 \text{poly}(\|A\|_F, \|B\|_F, n_S, n_{\mathcal{A}}, \\ &\quad \log \frac{1}{\delta''''}, \frac{1}{\alpha}, H, R_{\max}, \log \frac{1}{\delta'})). \end{aligned} \tag{100}$$

On the other hand, if the algorithm continues, we have from Eqn. (99) and Lemma 2 (choose $a = \frac{\alpha}{16 H R_{\max}} N/2 - \log \frac{1}{\delta'''''}$) that such an error must be encountered with probability $1 - \delta'''''$ at least

$$\frac{\alpha}{32 H R_{\max}} N - \log \frac{1}{\delta'''''} \tag{101}$$

times. From Eqn. (100) and Eqn. (101) we have that after a number of iterations

$$\begin{aligned} &O(\text{poly}(\|A\|_F, \|B\|_F, n_S, n_{\mathcal{A}}, \log \frac{1}{\delta''''}, \frac{1}{\alpha}, H, \\ &\qquad R_{\max}, \log \frac{1}{\delta'}, \log \frac{1}{\delta'''''})) \end{aligned}$$

the algorithm must have terminated with probability $1 - \delta' - \delta'' - \delta''' - \delta'''' - \delta'''''$. Now choose $\delta' = \delta'' = \delta''' = \delta'''' = \delta''''' = \frac{\delta}{5}$, to obtain the bound on the number of iterations of Eqn. (3). Given this bound on the number of iterations $N$, it is easily verified that the conditions of Eqn. (92, 94, 96) on $N_T$ and $k_1$ are met by Eqn. (4) and Eqn. (5) of Theorem 3. Also, since we chose $N_T, k_1$ such that $\hat{U}_M(\pi_T)$ and $\{\hat{U}_M(\pi^{(i)}\}_i$ are accurately evaluated (as specified in Eqn. (91,93)), we have that Eqn. (2) holds when the algorithm terminates.

□