Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning

Christoph Dann
Machine Learning Department
Carnegie Mellon University
`cdann@cdann.net`

Emma Brunskill
Computer Science Department
Carnegie Mellon University
`ebrun@cs.cmu.edu`

Abstract

Recently, there has been significant progress in understanding reinforcement learning in discounted infinite-horizon Markov decision processes (MDPs) by deriving tight sample complexity bounds. However, in many real-world applications, an interactive learning agent operates for a fixed or bounded period of time, for example tutoring students for exams or handling customer service requests. Such scenarios can often be better treated as episodic fixed-horizon MDPs, for which only looser bounds on the sample complexity exist. A natural notion of sample complexity in this setting is the number of episodes required to guarantee a certain performance with high probability (PAC guarantee). In this paper, we derive an upper PAC bound $\tilde{O}(\frac{|\mathcal{S}|^2|\mathcal{A}|H^2}{\epsilon^2}\ln\frac{1}{\delta})$ and a lower PAC bound $\tilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|H^2}{\epsilon^2}\ln\frac{1}{\delta+c})$ that match up to log-terms and an additional linear dependency on the number of states $|\mathcal{S}|$. The lower bound is the first of its kind for this setting. Our upper bound leverages Bernstein's inequality to improve on previous bounds for episodic finite-horizon MDPs which have a time-horizon dependency of at least H^3.

1 Introduction and Motivation

Consider test preparation software that tutors students for a national advanced placement exam taken at the end of a year, or maximizing business revenue by the end of each quarter. Each individual task instance requires making a sequence of decisions for a fixed number of steps H (e.g., tutoring one student to take an exam in spring 2015 or maximizing revenue for the end of the second quarter of 2014). Therefore, they can be viewed as a finite-horizon sequential decision making under uncertainty problem, in contrast to an infinite horizon setting in which the number of time steps is infinite. When the domain parameters (e.g. Markov decision process parameters) are not known in advance, and there is the opportunity to repeat the task many times (teaching a new student for each year's exam, maximizing revenue for each new quarter), this can be treated as episodic fixed-horizon reinforcement learning (RL). One important question is to understand how much experience is required to act well in this setting. We formalize this as the sample complexity of reinforcement learning [1], which is the number of time steps on which the algorithm may select an action whose value is not near-optimal. RL algorithms with a sample complexity that is a polynomial function of the domain parameters are referred to as Probably Approximately Correct (PAC) [2, 3, 4, 1]. Though there has been significant work on PAC RL algorithms for the infinite horizon setting, there has been relatively little work on the finite horizon scenario.

In this paper we present the first, to our knowledge, lower bound, and a new upper bound on the sample complexity of episodic finite horizon PAC reinforcement learning in discrete state-action spaces. Our bounds are tight up to log-factors in the time horizon H, the accuracy ϵ, the number of actions $|\mathcal{A}|$ and up to an additive constant in the failure probability δ. These bounds improve upon existing results by a factor of at least H. Our results also apply when the reward model is a function of the within-episode time step in addition to the state and action space. While we assume a stationary transition model, our results can be extended readily to time-dependent state-transitions. Our proposed `UCFH` (Upper-confidence fixed-horizon RL) algorithm that achieves our upper PAC guarantee can be applied directly to wide range of fixed-horizon episodic MDPs with known rewards.[1] It does not require additional structure such as assuming access to a generative model [8] or that the state transitions are sparse or acyclic [6].

The limited prior research on upper bound PAC results for finite horizon MDPs has focused on different settings, such as partitioning a longer trajectory into fixed length segments [4, 1], or considering a sliding time window [9]. The tightest dependence on the horizon in terms of the number of episodes presented in these approaches is at least H^3 whereas our dependence is only H^2. More importantly, such alternative settings require the optimal policy to be stationary, whereas in general in finite horizon settings the optimal policy is nonstationary (e.g. is a function of both the state and the within episode time-step).[2] Fiechter [10, 11] and Reveliotis and Bountourelis [12] do tackle a closely related setting, but find a dependence that is at least H^4.

Our work builds on recent work [6, 8] on PAC infinite horizon discounted RL that offers much tighter upper and lower sample complexity bounds than was previously known. To use an infinite horizon algorithm in a finite horizon setting, a simple change is to augment the state space by the time step (ranging over $1, \dots, H$), which enables the learned policy to be non-stationary in the original state space (or equivalently, stationary in the newly augmented space). Unfortunately, since these recent bounds are in general a quadratic function of the state space size, the proposed state space expansion would introduce at least an additional H^2 factor in the sample complexity term, yielding at least a H^4 dependence in the number of episodes for the sample complexity.

Somewhat surprisingly, we prove an upper bound on the sample complexity for the finite horizon case that only scales quadratically with the horizon. A key part of our proof is that the variance of the value function in the finite horizon setting satisfies a Bellman equation. We also leverage recent insights that state–action pairs can be estimated to different precisions depending on the frequency to which they are visited under a policy, extending these ideas to also handle when the policy followed is nonstationary. Our lower bound analysis is quite different than some prior infinite-horizon results, and involves a construction of parallel multi-armed bandits where it is required that the best arm in a certain portion of the bandits is identified with high probability to achieve near-optimality.

2 Problem Setting and Notation

We consider episodic fixed-horizon MDPs, which can be formalized as a tuple $M = (\mathcal{S}, \mathcal{A}, r, p, p_0, H)$. Both, the statespace $\mathcal{S}$ and the actionspace $\mathcal{A}$ are finite sets. The learning agent interacts with the MDP in episodes of H time steps. At time $t = 1 \dots H$, the agent observes a state s_t and choses an action a_t based on a policy π that potentially depends on the within-episode time step, i.e., $a_t = \pi_t(s_t)$ for $t = 1, \dots, H$. The next state is sampled from the stationary transition kernel $s_{t+1} \sim p(\cdot|s_t, a_t)$ and the initial state from $s_1 \sim p_0$. In addition the agent receives a reward drawn from a distribution[3] with mean $r_t(s_t)$ determined by the reward function. The reward function r is possibly time-dependent and takes values in $[0, 1]$. The quality of a policy π is evaluated by the *total expected reward* of an episode $R^\pi_M = \mathbb{E}\left[\sum_{t=1}^H r_t(s_t)\right]$. For simplicity,[1] we assume that the reward function r is known to the agent but the transition kernel p is unknown. The question we study is how many episodes does a learning agent follow a policy π that is not ϵ-optimal, i.e., $R^*_M - \epsilon > R^\pi_M$, with probability at least $1 - \delta$ for any chosen accuracy ϵ and failure probability δ.

Notation. In the following sections, we reason about the true MDP M, an empirical MDP $\hat{M}$ and an optimistic MDP $\tilde{M}$ which are identical except for their transition probabilities p, $\hat{p}$ and $\tilde{p}_t$. We will provide more details about these MDPs later. We introduce the notation explicitly only for M but the quantities carry over to $\tilde{M}$ and $\hat{M}$ with additional tildes or hats by replacing p with $\tilde{p}_t$ or $\hat{p}$.

---

[1] Previous works [5] have shown that the complexity of learning state transitions usually dominates learning reward functions. We therefore follow existing sample complexity analyses [6, 7] and assume known rewards for simplicity. The algorithm and PAC bound can be extended readily to the case of unknown reward functions.

[2] The best action will generally depend on the state and the number of remaining time steps. In the tutoring example, even if the student has the same state of knowledge, the optimal tutor decision may be to space practice if there is many days till the test and provide intensive short-term practice if the test is tomorrow.

[3] It is straightforward to have the reward depend on the state, or state/action or state/action/next state.

The (linear) operator $P_i^\pi f(s) := \mathbb{E}[f(s_{i+1})|s_i = s] = \sum_{s' \in \mathcal{S}} p(s'|s, \pi_i(s)) f(s')$ takes any function $f : \mathcal{S} \to \mathbb{R}$ and returns the expected value of f with respect to the next time step.[4] For convenience, we define the multi-step version as $P_{i:j}^\pi f := P_i^\pi P_{i+1}^\pi \ldots P_j^\pi f$. The value function from time i to time j is defined as $V_{i:j}^\pi(s) := \mathbb{E}\left[\sum_{t=i}^j r_t(s_t) | s_i = s\right] = \sum_{t=i}^j P_{i:t-1}^\pi r_t = \left(P_i^\pi V_{i+1:j}^\pi\right)(s) + r_i(s)$ and $V_{i:j}^*$ is the optimal value-function. When the policy is clear, we omit the superscript π.

We denote by $\mathcal{S}(s, a) \subseteq \mathcal{S}$ the set of possible successor states of state s and action a. The maximum number of them is denoted by $C = \max_{s,a \in \mathcal{S} \times \mathcal{A}} |\mathcal{S}(s, a)|$. In general, without making further assumptions, we have $C = |\mathcal{S}|$, though in many practical domains (robotics, user modeling) each state can only transition to a subset of the full set of states (e.g. a robot can't teleport across the building, but can only take local moves). The notation $\tilde{O}$ is similar to the usual O-notation but ignores log-terms. More precisely $f = \tilde{O}(g)$ if there are constants c_1, c_2 such that $f \leq c_1 g (\ln g)^{c_2}$ and analogously for $\tilde{\Omega}$. The natural logarithm is $\ln$ and $\log = \log_2$ is the base-2 logarithm.

3 Upper PAC-Bound

We now introduce a new model-based algorithm, `UCFH`, for RL in finite horizon episodic domains. We will later prove `UCFH` is PAC with an upper bound on its sample complexity that is smaller than prior approaches. Like many other PAC RL algorithms [3, 13, 14, 15], `UCFH` uses an optimism under uncertainty approach to balance exploration and exploitation. The algorithm generally works in phases comprised of optimistic planning, policy execution and model updating that take several episodes each. Phases are indexed by k. As the agent acts in the environment and observes (s, a, r, s') tuples, `UCFH` maintains a confidence set over the possible transition parameters for each state-action pair that are consistent with the observed transitions. Defining such a confidence set that holds with high probability can be be achieved using concentration inequalities like the Hoeffding inequality. One innovation in our work is to use a particular new set of conditions to define the confidence set that enables us to obtain our tighter bounds. We will discuss the confidence sets further below. The collection of these confidence sets together form a class of MDPs $\mathcal{M}_k$ that are consistent with the observed data. We define $\hat{M}_k$ as the maximum likelihood estimate of the MDP given the previous observations.

Given $\mathcal{M}_k$, `UCFH` computes a policy π^k by performing optimistic planning. Specifically, we use a finite horizon variant of extended value iteration (EVI) [5, 14]. EVI performs modified Bellman backups that are optimistic with respect to a given set of parameters. That is, given a confidence set of possible transition model parameters, it selects in each time step the model within that set that maximizes the expected sum of future rewards. Appendix A provides more details about fixed horizon EVI.

`UCFH` then executes π^k until there is a state-action pair (s, a) that has been visited often enough since its last update (defined precisely in the until-condition in `UCFH`). After updating the model statistics for this (s, a)-pair, a new policy π^{k+1} is obtained by optimistic planning again. We refer to each such iteration of planning-execution-update as a *phase* with index k. If there is no ambiguity, we omit the phase indices k to avoid cluttered notation.

`UCFH` is inspired by the infinite-horizon `UCRL`-γ algorithm by Lattimore and Hutter [6] but has several important differences. First, the policy can only be updated at the end of an episode, so there is no need for explicit delay phases as in `UCRL`-γ. Second, the policies π^k in `UCFH` are time-dependent. Finally, `UCFH` can directly deal with non-sparse transition probabilities, whereas `UCRL`-γ only directly allows two possible successor states for each (s, a)-pair ($C = 2$).

Confidence sets. The class of MDPs $\mathcal{M}_k$ consists of fixed-horizon MDPs M' with the known true reward function r and where the transition probability $p_t'(s'|s, a)$ from any $(s, a) \in \mathcal{S} \times \mathcal{A}$ to $s' \in \mathcal{S}(s, a)$ at any time t is in the confidence set induced by $\hat{p}(s'|s, a)$ of the empirical MDP $\hat{M}$. Solely for the purpose of computationally more efficient optimistic planning, we allow time-dependent transitions (allows choosing different transition models in different time steps to maximize reward), but this does not affect the theoretical guarantees as the true stationary MDP is still in $\mathcal{M}_k$ with high

---

[4] The definition also works for time-dependent transition probabilities.

Algorithm 1: `UCFH`: **U**pper-**C**onfidence **F**ixed-**H**orizon episodic reinforcement learning algorithm

```
Input: desired accuracy $\epsilon \in (0,1]$, failure tolerance $\delta \in (0,1]$, fixed-horizon MDP $M$
Result: with probability at least $1-\delta$: $\epsilon$-optimal policy
$k := 1, \quad w_{\min} := \frac{\epsilon}{4H|\mathcal{S}|}, \quad \delta_1 := \frac{\delta}{2U_{\max}C}, \quad U_{\max} := |\mathcal{S}\times\mathcal{A}|\log_2 \frac{|\mathcal{S}|H}{w_{\min}}$;
$m := 512(\log_2\log_2 H)^2 \frac{CH^2}{\epsilon^2} \log^2\left(\frac{8H^2|\mathcal{S}|^2}{\epsilon}\right) \ln \frac{6|\mathcal{S}\times\mathcal{A}|C\log_2^2(4|\mathcal{S}|^2H^2/\epsilon)}{\delta}$;
$n(s,a) = v(s,a) = n(s,a,s') := 0 \quad \forall, s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}(s,a)$;
while do
    /* Optimistic planning */
    $\hat{p}(s'|s,a) := n(s,a,s')/n(s,a)$, for all $(s,a)$ with $n(s,a) > 0$ and $s' \in \mathcal{S}(s,a)$;
    $\mathcal{M}_k := \big\{\tilde{M} \in \mathcal{M}_{\text{nonst.}} : \forall (s,a) \in \mathcal{S}\times\mathcal{A}, t = 1 \dots H, s' \in \mathcal{S}(s,a)$
            $\tilde{p}_t(s'|s,a) \in$ ConfidenceSet $(\hat{p}(s'|s,a), n(s,a))\big\}$;
    $\tilde{M}_k, \pi^k :=$ FixedHorizonEVI$(\mathcal{M}_k)$;
    /* Execute policy */
    repeat
        SampleEpisode($\pi^k$) ; // from $M$ using $\pi^k$
    until there is a $(s,a) \in \mathcal{S}\times\mathcal{A}$ with $v(s,a) \geq \max\{mw_{\min}, n(s,a)\}$ and $n(s,a) < |\mathcal{S}|mH$;
    /* Update model statistics for one $(s,a)$-pair with condition above */
    $n(s,a) := n(s,a) + v(s,a)$;
    $n(s,a,s') := n(s,a,s') + v(s,a,s') \quad \forall s' \in \mathcal{S}(s,a)$;
    $v(s,a) := v(s,a,s') := 0 \quad \forall s' \in \mathcal{S}(s,a)$; $k := k+1$
Procedure SampleEpisode($\pi$)
    $s_0 \sim p_0$;
    for $t = 0$ to $H-1$ do
        $a_t := \pi_{t+1}(s_t)$ and $s_{t+1} \sim p(\cdot|s_t, a_t)$;
        $v(s_t, a_t) := v(s_t, a_t) + 1$ and $v(s_t, a_t, s_{t+1}) := v(s_t, a_t, s_{t+1}) + 1$;
Function ConfidenceSet($p$, $n$)
    $\mathcal{P} := \Big\{p' \in [0,1] :\text{if } n > 1 : |p'(1-p') - p(1-p)| \leq \frac{2\ln(6/\delta_1)}{n-1},$   (1)
            $|p - p'| \leq \min\left(\sqrt{\frac{\ln(6/\delta_1)}{2n}}, \sqrt{\frac{2p(1-p)}{n}} \ln(6/\delta_1) + \frac{2}{3n}\ln\frac{6}{\delta_1}\right)\Big\}$   (2)
    return $\mathcal{P}$
```

probability. Unlike the confidence intervals used by Lattimore and Hutter [6], we not only include conditions based on Hoeffding's inequality[5] and Bernstein's inequality (Eq. 2), but also require that the variance $p(1-p)$ of the Bernoulli random variable associated with this transition is close to the empirical one (Eq. 1). This additional condition (Eq. 1) is key for making the algorithm directly applicable to generic MDPs (in which states can transition to any number of next states, e.g. $C > 2$) while only having a linear dependency on C in the PAC bound.

3.1 PAC Analysis

For simplicity we assume that each episode starts in a fixed start state s_0. This assumption is not crucial and can easily be removed by additional notational effort.

Theorem 1. *For any $0 < \epsilon, \delta \leq 1$, the following holds. With probability at least $1-\delta$,* `UCFH` *produces a sequence of policies π^k, that yield at most*

$$\tilde{O}\left(\frac{H^2C|\mathcal{S}\times\mathcal{A}|}{\epsilon^2}\ln\frac{1}{\delta}\right)$$

episodes with $R^* - R^{\pi^k} = V^*_{1:H}(s_0) - V^{\pi^k}_{1:H}(s_0) > \epsilon$. The maximum number of possible successor states is denoted by $1 < C \leq |\mathcal{S}|$.

---

[5] The first condition in the min in Equation (2) is actually not necessary for the theoretical results to hold. It can be removed and all $6/\delta_1$ can be replaced by $4/\delta_1$.

Similarities to other analyses. The proof of Theorem 1 is quite long and involved, but builds on similar techniques for sample-complexity bounds in reinforcement learning (see e.g. Brafman and Tennenholtz [3], Strehl and Littman [16]). The general proof strategy is closest to the one of `UCRL`-γ [6] and the obtained bounds are similar if we replace the time horizon H with the equivalent in the discounted case $1/(1-\gamma)$. However, there are important differences that we highlight now briefly.

- A central quantity in the analysis by Lattimore and Hutter [6] is the local variance of the value function. The exact definition for the fixed-horizon case will be given below. The key insight for the almost tight bounds of Lattimore and Hutter [6] and Azar et al. [8] is to leverage the fact that these local variances satisfy a Bellman equation [17] and so the discounted sum of local variances can be bounded by $O((1-\gamma)^{-2})$ instead of $O((1-\gamma)^{-3})$. We prove in Lemma 4 that local value function variances $\sigma^2_{i:j}$ also satisfy a Bellman equation for fixed-horizon MDPs even if transition probabilities and rewards are time-dependent. This allows us to bound the total sum of local variances by $O(H^2)$ and obtain similarly strong results in this setting.
- Lattimore and Hutter [6] assumed there are only two possible successor states (i.e., $C = 2$) which allows them to easily relate the local variances $\sigma^2_{i:j}$ to the difference of the expected value of successor states in the true and optimistic MDP $(P_i - \tilde{P}_i)\tilde{V}_{i+1:j}$. For $C > 2$, the relation is less clear, but we address this by proving a bound with tight dependencies on C (Lemma C.6).
- To avoid super-linear dependency on C in the final PAC bound, we add the additional condition in Equation (1) to the confidence set. We show that this allows us to upper-bound the total reward difference $R^* - R^{\pi^k}$ of policy π^k with terms that either depend on $\sigma^2_{i:j}$ or decrease linearly in the number of samples. This gives the desired linear dependency on C in the final bound. We therefore avoid assuming $C = 2$ which makes `UCFH` directly applicable to generic MDPs with $C > 2$ without the impractical transformation argument used by Lattimore and Hutter [6].

We will now introduce the notion of *knownness* and *importance* of state-action pairs that is essential for the analysis of `UCFH` and subsequently present several lemmas necessary for the proof of Theorem 1. We only sketch proofs here but detailed proofs for all results are available in the appendix.

Fine-grained categorization of (s,a)-pairs. Many PAC RL sample complexity proofs [3, 4, 13, 14] only have a binary notion of "knownness", distinguishing between known (transition probability estimated sufficiently accurately) and unknown (s,a)-pairs. However, as recently shown by Lattimore and Hutter [6] for the infinite horizon setting, it is possible to obtain much tighter sample complexity results by using a more fine grained categorization. In particular, a key idea is that in order to obtain accurate estimates of the value function of a policy from a starting state, it is sufficient to have only a loose estimate of the parameters of (s,a)-pairs that are unlikely to be visited under this policy.

Let the *weight* of a (s,a)-pair given policy π^k be its expected frequency in an episode

$$w_k(s,a) := \sum_{t=1}^{H} \mathbb{P}(s_t = s, \pi^k_t(s_t) = a) = \sum_{t=1}^{H} P_{1:t-1}\mathbb{I}\{s = \cdot, a = \pi^k_t(s)\}(s_0).$$

The *importance* ι_k of (s,a) is its relative weight compared to $w_{\min} := \frac{\epsilon}{4H|\mathcal{S}|}$ on a log-scale

$$\iota_k(s,a) := \min\left\{z_i \;:\; z_i \geq \frac{w_k(s,a)}{w_{\min}}\right\} \quad \text{where } z_1 = 0 \text{ and } z_i = 2^{i-2}\ \forall i = 2,3,\dots.$$

Note that $\iota_k(s,a) \in \{0,1,2,4,8,16\dots\}$ is an integer indicating the influence of the state-action pair on the value function of π^k. Similarly, we define the *knownness*

$$\kappa_k(s,a) := \max\left\{z_i \;:\; z_i \leq \frac{n_k(s,a)}{m w_k(s,a)}\right\} \in \{0,1,2,4,\dots\}$$

which indicates how often (s,a) has been observed relative to its importance. The constant m is defined in Algorithm 1. We can now categorize (s,a)-pairs into subsets

$$X_{k,\kappa,\iota} := \{(s,a) \in X_k \;:\; \kappa_k(s,a) = \kappa, \iota_k(s,a) = \iota\} \quad \text{and} \quad \bar{X}_k = \mathcal{S} \times \mathcal{A} \setminus X_k$$

where $X_k = \{(s,a) \in \mathcal{S} \times \mathcal{A} \;:\; \iota_k(s,a) > 0\}$ is the *active set* and $\bar{X}_k$ the set of state-action pairs that are very unlikely under the current policy. Intuitively, the model of `UCFH` is accurate if only few

(s, a) are in categories with low knownness – that is, important under the current policy but have not been observed often so far. Recall that over time observations are generated under many policies (as the policy is recomputed), so this condition does not always hold. We will therefore distinguish between phases k where $|X_{k,\kappa,\iota}| \leq \kappa$ for all κ and ι and phases where this condition is violated. The condition essentially allows for only a few (s, a) in categories that are less known and more and more (s, a) in categories that are more well known. In fact, we will show that the policy is ϵ-optimal with high probability in phases that satisfy this condition.

We first show the validity of the confidence sets $\mathcal{M}_k$.

Lemma 1 (Capturing the true MDP whp.). *$M \in \mathcal{M}_k$ for all k with probability at least $1 - \delta/2$.*

Proof Sketch. By combining Hoeffding's inequality, Bernstein's inequality and the concentration result on empirical variances by Maurer and Pontil [18] with the union bound, we get that $p(s'|s, a) \in \mathcal{P}$ with probability at least $1 - \delta_1$ for a single phase k, fixed $s, a \in \mathcal{S} \times \mathcal{A}$ and fixed $s' \in \mathcal{S}(s, a)$. We then show that the number of model updates is bounded by $U_{\max}$ and apply the union bound. ☐

The following lemma bounds the number of episodes in which $\forall \kappa, \iota : \ |X_{k,\kappa,\iota}| \leq \kappa$ is violated with high probability.

Lemma 2. *Let E be the number of episodes k for which there are κ and ι with $|X_{k,\kappa,\iota}| > \kappa$, i.e. $E = \sum_{k=1}^{\infty} \mathbb{I}\{\exists(\kappa, \iota) : |X_{k,\kappa,\iota}| > \kappa\}$ and assume that $m \geq \frac{6H^2}{\epsilon} \ln \frac{2E_{\max}}{\delta}$. Then $\mathbb{P}(E \leq 6NE_{\max}) \geq 1 - \delta/2$ where $N = |\mathcal{S} \times \mathcal{A}|\, m$ and $E_{\max} = \log_2 \frac{H}{w_{\min}} \log_2 |\mathcal{S}|$.*

Proof Sketch. We first bound the total number of times a fixed pair (s, a) can be observed while being in a particular category $X_{k,\kappa,\iota}$ in all phases k for $1 \leq \kappa < |\mathcal{S}|$. We then show that for a particular (κ, ι), the number of episodes where $|X_{k,\kappa,\iota}| > \kappa$ is bounded with high probability, as the value of ι implies a minimum probability of observing each (s, a) pair in $X_{k,\kappa,\iota}$ in an episode. Since the observations are not independent we use martingale concentration results to show the statement for a fixed (κ, ι). The desired result follows with the union bound over all relevant κ and ι. ☐

The next lemma states that in episodes where the condition $\forall \kappa, \iota : \ |X_{k,\kappa,\iota}| \leq \kappa$ is satisfied and the true MDP is in the confidence set, the expected optimistic policy value is close to the true value. This lemma is the technically most involved part of the proof.

Lemma 3 (Bound mismatch in total reward). *Assume $M \in \mathcal{M}_k$. If $|X_{k,\kappa,\iota}| \leq \kappa$ for all (κ, ι) and $0 < \epsilon \leq 1$ and $m \geq 512 \frac{CH^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2 \left(\frac{8H^2|\mathcal{S}|^2}{\epsilon}\right) \ln \frac{6}{\delta_1}$. Then $|\tilde{V}^{\pi^k}_{1:H}(s_0) - V^{\pi^k}_{1:H}(s_0)| \leq \epsilon$.*

Proof Sketch. Using basic algebraic transformations, we show that $|p - \tilde{p}| \leq \sqrt{\tilde{p}(1 - \tilde{p})} O\left(\sqrt{\frac{1}{n} \ln \frac{1}{\delta_1}}\right) + O\left(\frac{1}{n} \ln \frac{1}{\delta_1}\right)$ for each $\tilde{p}, p \in \mathcal{P}$ in the confidence set as defined in Eq. 2. Since we assume $M \in \mathcal{M}_k$, we know that $p(s'|s, a)$ and $\tilde{p}(s'|s, a)$ satisfy this bound with $n(s, a)$ for all s,a and s'. We use that to bound the difference of the expected value function of the successor state in M and $\tilde{M}$, proving that $|(P_i - \tilde{P}_i)\tilde{V}_{i+1:j}(s)| \leq O\left(\frac{CH}{n(s,\pi(s))} \ln \frac{1}{\delta_1}\right) + O\left(\sqrt{\frac{C}{n(s,\pi(s))} \ln \frac{1}{\delta_1}}\right) \tilde{\sigma}_{i:j}(s)$, where the local variance of the value function is defined as $\sigma^2_{i:j}(s, a) := \mathbb{E}\left[(V^{\pi}_{i+1:j}(s_{i+1}) - P^{\pi}_i V^{\pi}_{i+1:j}(s_i))^2 | s_i = s, a_i = a\right]$ and $\sigma^2_{i:j}(s) := \sigma^2_{i:j}(s, \pi_i(s))$. This bound then is applied to $|\tilde{V}_{1:H}(s_0) - V_{1:H}(s_0)| \leq \sum_{t=0}^{H-1} P_{1:t} |(P_t - \tilde{P}_t)\tilde{V}_{t+1:H}(s)|$. The basic idea is to split the bound into a sum of two parts by partitioning of the (s, a) space by knownness, e.g. that is $(s_t, a_t) \in \bar{X}_{\kappa,\iota}$ for all κ and ι and $(s_t, a_t) \in \bar{X}$. Using the fact that $w(s_t, a_t)$ and $n(s_t, a_t)$ are tightly coupled for each (κ, ι), we can bound the expression eventually by ϵ. The final key ingredient in the remainder of the proof is to bound $\sum_{t=1}^{H} P_{1:t-1} \sigma_{t:H}(s)^2$ by $O(H^2)$ instead of the trivial bound $O(H^3)$. To this end, we show the lemma below. ☐

Lemma 4. *The variance of the value function defined as $\mathcal{V}^{\pi}_{i:j}(s) := \mathbb{E}\left[\left(\sum_{t=i}^{j} r_t(s_t) - V^{\pi}_{i:j}(s_i)\right)^2 | s_i = s\right]$ satisfies a Bellman equation $\mathcal{V}_{i:j} = P_i \mathcal{V}_{i+1:j} + \sigma^2_{i:j}$ which gives $\mathcal{V}_{i:j} = \sum_{t=i}^{j} P_{i:t-1} \sigma^2_{t:j}$. Since $0 \leq \mathcal{V}_{1:H} \leq H^2 r^2_{\max}$, it follows that $0 \leq \sum_{t=1}^{j} P_{i:t-1} \sigma^2_{t:j}(s) \leq H^2 r^2_{\max}$ for all $s \in \mathcal{S}$.*

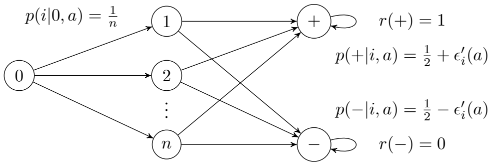
$p(i|0,a) = \frac{1}{n}$
0
1
2
⋮
n
+
−
$r(+) = 1$
$p(+|i,a) = \frac{1}{2} + \epsilon'_i(a)$
$p(-|i,a) = \frac{1}{2} - \epsilon'_i(a)$
$r(-) = 0$

Figure 1: Class of a hard-to-learn finite horizon MDPs. The function ϵ' is defined as $\epsilon'(a_1) = \epsilon/2$, $\epsilon'(a_i^*) = \epsilon$ and otherwise $\epsilon'(a) = 0$ where a_i^* is an unknown action per state i and ϵ is a parameter.

Proof Sketch. The proof works by induction and uses fact that the value function satisfies the Bellman equation and the tower-property of conditional expectations. ☐

Proof Sketch for Theorem 1. The proof of Theorem 1 consists of the following major parts:

1. The true MDP is in the set of MDPs $\mathcal{M}_k$ for all phases k with probability at least $1-\frac{\delta}{2}$ (Lemma 1).
2. The `FixedHorizonEVI` algorithm computes a value function whose optimistic value is higher than the optimal reward in the true MDP with probability at least $1 - \delta/2$ (Lemma A.1).
3. The number of episodes with $|X_{k,\kappa,\iota}| > \kappa$ for some κ and ι are bounded with probability at least $1 - \delta/2$ by $\tilde{O}(|\mathcal{S} \times \mathcal{A}|\, m)$ if $m = \tilde{\Omega}\left(\frac{H^2}{\epsilon} \ln \frac{|\mathcal{S}|}{\delta}\right)$ (Lemma 2).
4. If $|X_{k,\kappa,\iota}| \leq \kappa$ for all κ, ι, i.e., relevant state-action pairs are sufficiently known and $m = \tilde{\Omega}\left(\frac{CH^2}{\epsilon^2} \ln \frac{1}{\delta_1}\right)$, then the optimistic value computed is ϵ-close to the true MDP value. Together with part 2, we get that with high probability, the policy π^k is ϵ-optimal in this case.
5. From parts 3 and 4, with probability $1 - \delta$, there are at most $\tilde{O}\left(\frac{C|\mathcal{S}\times\mathcal{A}|H^2}{\epsilon^2} \ln \frac{1}{\delta}\right)$ episodes that are not ϵ-optimal.

4 Lower PAC Bound

Theorem 2. *There exist positive constants c_1, c_2, δ_0, ϵ_0 such that for every $\delta \in (0, \delta_0)$ and $\epsilon \in (0, \epsilon_0)$ and for every algorithm A that satisfies a PAC guarantee for (ϵ, δ) and outputs a deterministic policy, there is a fixed-horizon episodic MDP M_{hard} with*

$$\mathbb{E}[n_A] \geq \frac{c_1(H-2)^2(|\mathcal{A}|-1)(|\mathcal{S}|-3)}{\epsilon^2} \ln\left(\frac{c_2}{\delta + c_3}\right) = \Omega\left(\frac{|\mathcal{S} \times \mathcal{A}|H^2}{\epsilon^2} \ln\left(\frac{c_2}{\delta + c_3}\right)\right) \tag{3}$$

where n_A is the number of episodes until the algorithm's policy is (ϵ, δ)-accurate. The constants can be set to $\delta_0 = \frac{e^{-4}}{80} \approx \frac{1}{5000}$, $\epsilon_0 = \frac{H-2}{640e^4} \approx H/35000$, $c_2 = 4$ and $c_3 = e^{-4}/80$.

The ranges of possible δ and ϵ are of similar order than in other state-of-the-art lower bounds for multi-armed bandits [19] and discounted MDPs [14, 6]. They are mostly determined by the bandit result by Mannor and Tsitsiklis [19] we build on. Increasing the parameter limits δ_0 and ϵ_0 for bandits would immediately result in larger ranges in our lower bound, but this was not the focus of our analysis.

Proof Sketch. The basic idea is to show that the class of MDPs shown in Figure 1 require at least a number of observed episodes of the order of Equation (3). From the start state 0, the agent ends up in states 1 to n with equal probability, independent of the action. From each such state i, the agent transitions to either a good state + with reward 1 or a bad state − with reward 0 and stays there for the rest of the episode. Therefore, each state $i = 1, \ldots, n$ is essentially a multi-armed bandit with binary rewards of either 0 or $H - 2$. For each bandit, the probability of ending up in + or − is equal except for the first action a_1 with $p(s_{t+1} = +|s_t = i, a_t = a_1) = 1/2 + \epsilon/2$ and possibly an unknown optimal action a_i^* (different for each state i) with $p(s_{t+1} = +|s_t = i, a_t = a_i^*) = 1/2 + \epsilon$.

In the episodic fixed-horizon setting we are considering, taking a suboptimal action in one of the bandits does not necessarily yield a suboptimal episode. We have to consider the average over all bandits instead. In an ϵ-optimal episode, the agent therefore needs to follow a policy that would solve at least a certain portion of all n multi-armed bandits with probability at least $1 - \delta$. We show that the best strategy for the agent to achieve this is to try to solve all bandits with equal probability. The number of samples required to do so then results in the lower bound in Equation (3). ☐

Similar MDPs that essentially solve multiple of such multi-armed bandits have been used to prove lower sample-complexity bounds for discounted MDPs [14, 6]. However, the analysis in the infinite horizon case as well as for the sliding-window fixed-horizon optimality criterion considered by Kakade [4] is significantly simpler. For these criteria, every time step the agent follows a policy that is not ϵ-optimal counts as a ”mistake”. Therefore, every time the agent does not pick the optimal arm in any of the multi-armed bandits counts as a mistake. This contrasts with our fixed-horizon setting where we must instead consider taking an average over all bandits.

5 Related Work on Fixed-Horizon Sample Complexity Bounds

We are not aware of any lower sample complexity bounds beyond multi-armed bandit results that directly apply to our setting. Our upper bound in Theorem 1 improves upon existing results by at least a factor of H. We briefly review those existing results in the following.

Timestep bounds. Kakade [4, Chapter 8] proves upper and lower PAC bounds for a similar setting where the agent interacts indefinitely with the environment but the interactions are divided in segments of equal length and the agent is evaluated by the expected sum of rewards until the end of each segment. The bound states that there are not more than $\tilde{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H^6}{\epsilon^3}\ln\frac{1}{\delta}\right)$[6] time steps in which the agents acts ϵ-suboptimal. Strehl et al. [1] improves the state-dependency of these bounds for their delayed Q-learning algorithm to $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^5}{\epsilon^4}\ln\frac{1}{\delta}\right)$. However, in episodic MDP it is more natural to consider performance on the entire episode since suboptimality near the end of the episode is no issue as long as the total reward on the entire episode is sufficiently high. Kolter and Ng [9] use an interesting sliding-window criterion, but prove bounds for a Bayesian setting instead of PAC. Timestep-based bounds can be applied to the episodic case by augmenting the original statespace with a time-index per episode to allow resets after H steps. This adds H dependencies for each $|\mathcal{S}|$ in the original bound which results in a horizon-dependency of at least H^6 of these existing bounds. Translating the regret bounds of `UCRL2` in Corollary 3 by Jaksch et al. [20] yields a PAC-bound on the number of episodes of at least $\tilde{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H^3}{\epsilon^2}\ln\frac{1}{\delta}\right)$ even if one ignores the reset after H time steps. Timestep-based lower PAC-bounds cannot be applied directly to the episodic reward criterion.

Episode bounds. Similar to us, Fiechter [10] uses the value of initial states as optimality-criterion, but defines the value w.r.t. the γ-discounted infinite horizon. His results of order $\tilde{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H^7}{\epsilon^2}\ln\frac{1}{\delta}\right)$ episodes of length $\tilde{O}(1/(1-\gamma)) \approx \tilde{O}(H)$ are therefore not directly applicable to our setting. Auer and Ortner [5] investigate the same setting as we and propose a UCB-type algorithm that has no-regret, which translates into a basic PAC bound of order $\tilde{O}\left(\frac{|\mathcal{S}|^{10}|\mathcal{A}|H^7}{\epsilon^3}\ln\frac{1}{\delta}\right)$ episodes. We improve on this bound substantially in terms of its dependency on H, $|\mathcal{S}|$ and ϵ. Reveliotis and Bountourelis [12] also consider the episodic undiscounted fixed-horizon setting and present an efficient algorithm in cases where the transition graph is acyclic and the agent knows for each state a policy that visits this state with a known minimum probability q. These assumptions are quite limiting and rarely hold in practice and their bound of order $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^4}{\epsilon^2 q}\ln\frac{1}{\delta}\right)$ explicitly depends on $1/q$.

6 Conclusion

We have shown upper and lower bounds on the sample complexity of episodic fixed-horizon RL that are tight up to log-factors in the time horizon H, the accuracy ϵ, the number of actions $|\mathcal{A}|$ and up to an additive constant in the failure probability δ. These bounds improve upon existing results by a factor of at least H. One might hope to reduce the dependency of the upper bound on $|\mathcal{S}|$ to be linear by an analysis similar to `Mormax` [7] for discounted MDPs which has sample complexity linear in $|\mathcal{S}|$ at the penalty of additional dependencies on H. Our proposed `UCFH` algorithm that achieves our PAC bound can be applied to directly to a wide range of fixed-horizon episodic MDPs with known rewards and does not require additional structure such as sparse or acyclic state transitions assumed in previous work. The empirical evaluation of `UCFH` is an interesting direction for future work.

Acknowledgments: We thank Tor Lattimore for the helpful suggestions and comments. This work was supported by an NSF CAREER award and the ONR Young Investigator Program.

---

[6] For comparison we adapt existing bounds to our setting. While the original bound stated by Kakade [4] only has H^3, an additional H^3 comes in through ϵ^{-3} due to different normalization of rewards.

References

[1] Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC Model-Free Reinforcement Learning. In *International Conference on Machine Learning*, 2006.

[2] Michael J Kearns and Satinder P Singh. Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms. In *Advances in Neural Information Processing Systems*, 1999.

[3] Ronen I Brafman and Moshe Tennenholtz. R-MAX – A General Polynomail Time Algorithm for Near-Optimal Reinforcement Learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[4] Sham M. Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

[5] Peter Auer and Ronald Ortner. Online Regret Bounds for a New Reinforcement Learning Algorithm. In *Proceedings 1st Austrian Cognitive Vision Workshop*, 2005.

[6] Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, 2012.

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

[8] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. On the Sample Complexity of Reinforcement Learning with a Generative Model. In *International Conference on Machine Learning*, 2012.

[9] J Zico Kolter and Andrew Y Ng. Near-Bayesian exploration in polynomial time. In *International Conference on Machine Learning*, 2009.

[10] Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Conference on Learning Theory*, 1994.

[11] Claude-Nicolas Fiechter. Expected Mistake Bound Model for On-Line Reinforcement Learning. In *International Conference on Machine Learning*, 1997.

[12] Spyros Reveliotis and Theologos Bountourelis. Efficient PAC learning for episodic tasks with acyclic state spaces. *Discrete Event Dynamic Systems: Theory and Applications*, 17(3):307–327, 2007.

[13] Alexander L Strehl, Lihong Li, and Michael L Littman. Incremental Model-based Learners With Formal Learning-Time Guarantees. In *Conference on Uncertainty in Artificial Intelligence*, 2006.

[14] Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement Learning in Finite MDPs : PAC Analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

[15] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2010.

[16] Alexander L. Strehl and Michael L. Littman. An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, dec 2008.

[17] Matthew J Sobel. The Variance of Markov Decision Processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

[18] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample-Variance Penalization. In *Conference on Learning Theory*, 2009.

[19] Shie Mannor and John N Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

[20] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[21] Fan Chung and Linyuan Lu. Concentration Inequalities and Martingale Inequalities: A Survey. *Internet Mathematics*, 3(1):79–127, 2006.

Appendices

A Fixed-Horizon Extended Value Iteration

We want to find a policy π^k and optimistic $\tilde{M}_k \in \mathcal{M}_k$ which have the highest total reward $R^{\pi^k}_{\tilde{M}_k} = \max_{\pi, M' \in \mathcal{M}_k} R^{\pi}_{M'}$. Note that π^k is an optimal policy for M_k but not necessarily for M. We can find such a policy by dynamic programming similar to extended value iteration [16, 5]. The optimal Q-function can be computed as $\tilde{Q}^*_{H:H}(s,a) = r_H(s)$ and for $i = H-1, \dots, 2, 1$ as

$$\tilde{Q}^*_{i:H}(s,a) = r_i(s) + \max_{\tilde{p}_i \in \mathcal{P}_{s,a}} \left\{ \sum_{s' \in \mathcal{S}(s,a)} \tilde{p}_i \max_{b \in \mathcal{A}} \tilde{Q}^*_{i+1:H}(s', b) \right\}$$

The feasible set is defined as $\mathcal{P}_{s,a} := \{p \in [0,1]^{|\mathcal{S}(s,a)|} \mid \|p\|_1 = 1, \forall s' \in \mathcal{S}(s,a) : p(s') \in \texttt{ConfidenceSet}(\hat{p}(s'|s,a), n(s,a))\}$. The optimal policy $\pi^k_t(s)$ at time t is then simply the maximizer of the inner $\max$ operator and the transition probability $\tilde{p}_t(\cdot|s,a)$ is the maximizer of the outer maximum. The inner $\max$ can be solved efficiently by enumeration and the outer maximum similar to extended value iteration [16]. The basic idea is to put as much probability mass as possible to successor states with highest value. See the following algorithm for the implementation details.

```
Function FixedHorizonEVI($\mathcal{M}$)
$\tilde{Q}^*_{H:H}(s,a) = r_H(s) \quad \forall s,a \in \mathcal{S} \times \mathcal{A}$ ;                      // $O(|\mathcal{S}||\mathcal{A}|)$
for $t = H-1$ to $1$ do                                                   // $O(H|\mathcal{S}|\log|\mathcal{S}| + H|\mathcal{S}||\mathcal{A}|C))$
    $\pi_{t+1}(s) := \arg\max_{a \in \mathcal{A}} \tilde{Q}^*_{t+1:H}(s,a) \quad \forall s \in \mathcal{S}$ ;                // $O(|\mathcal{S}||\mathcal{A}|)$
    sort states $s^{(1)}, \dots s^{(|\mathcal{S}|)}$ such that
        $\tilde{Q}^*_{t+1:H}(s^{(i)}, \pi_{t+1}(s^{(i)})) \geq \tilde{Q}^*_{t+1:H}(s^{(i+1)}, \pi_{t+1}(s^{(i+1)}))$ ;      // $O(|\mathcal{S}|\log|\mathcal{S}|)$
    for $s,a \in \mathcal{S} \times \mathcal{A}$ do                                          // $O(|\mathcal{S}||\mathcal{A}|C)$
        $\tilde{p}_t(s'|s,a) := \min \texttt{ConfidenceSet}(\hat{p}(s'|s,a), n(s,a)) \quad \forall s' \in \mathcal{S}(s,a)$ ;   // $O(C)$
        $\Delta := 1 - \sum_{s' \in \mathcal{S}(s,a)} \tilde{p}_t(s'|s,a)$ ;                               // $O(C)$
        $i := 1$ ;                                                            // $O(1)$
        while $\Delta > 0$ do                                                  // $O(C)$
            $s' := s^{(i)}$;
            $\Delta' := \min\{\Delta, \max \texttt{ConfidenceSet}(\hat{p}(s'|s,a), n(s,a)) - \tilde{p}_t(s'|s,a)\}$;
            $\tilde{p}_t(s'|s,a) := \tilde{p}_t(s'|s,a) + \Delta$;
            $\Delta := \Delta - \Delta'; i := i + 1$;
        $\tilde{Q}^*_{t:H}(s,a) = \sum_{s' \in \mathcal{S}(s,a)} \tilde{p}_t(s'|s,a) \tilde{Q}^*_{t+1:H}(s', \pi_{t+1}(s'))$ ;         // $O(C)$
$\pi_1(s) := \arg\max_{a \in \mathcal{A}} \tilde{Q}^*_{1:H}(s,a) \quad \forall s \in \mathcal{S}$ ;                        // $O(|\mathcal{S}||\mathcal{A}|$
return MDP with transition probabilities $\tilde{p}_t$, optimal policy $\pi$
```

Note that due to the nonlinear constraint in Equation (1), $\texttt{ConfidenceSet}(\hat{p}(s'|s,a), n(s,a))$ may be the union of two disjoint intervals instead of one interval. Still, $\min$- and $\max$-operations on the confidence sets can be computed readily in constant time. Therefore, the transition probabilities $\tilde{p}_t(\cdot|s,a)$ for a single time step t and state-action pair s,a can be computed in $O(|\mathcal{S}||\mathcal{A}|C)$ given sorted states. Sorting the states takes $O(|\mathcal{S}|\log|\mathcal{S}|)$ which results in $O(H|\mathcal{S}|\log|\mathcal{S}| + H|\mathcal{S}||\mathcal{A}|C))$ runtime complexity of `FixedHorizonEVI` (see comments in Function FixedHorizonEVI). The Algorithm requires $O(H|\mathcal{S}||\mathcal{A}|C)$ additional space besides the storage requirements of the input MDP $\mathcal{M}$ as the transition probabilities $\tilde{p}_t$ are returned by the algorithm. If those are not required and only the optimal policy is of interest, the additional space can be reduced to $O(|\mathcal{S}||\mathcal{A}|)$.

Lemma A.1 (Validity of optimistic planning)**.** `FixedHorizonEVI`$(\mathcal{M}_k)$ *returns* $\tilde{M}, \pi^k = \arg\max_{M \in \mathcal{M}_k, \pi} R^{\pi}_{M}$.

Proof Sketch. This result can be proved straight-forwardly by showing that π^k is optimal in the last time step H with highest possible reward and then subsequently for all previous time steps inductively. It follows directly from the definition of the algorithm in Function FixedHorizonEVI that the returned MDP is in $\mathcal{M}_k$. ☐

B Runtime- and Space-Complexity of `UCFH`

Sampling one episode and updating the respective v variables has $O(H)$ runtime. Theorem 1 states that after at most $\tilde{O}\left(\frac{H^2 C|\mathcal{S}\times\mathcal{A}|}{\epsilon^2}\ln\frac{1}{\delta}\right)$ observed episodes, the current policy is ϵ-optimal with sufficiently high probability. This results in a total runtime for sampling of $\tilde{O}\left(\frac{H^3 C|\mathcal{S}\times\mathcal{A}|}{\epsilon^2}\ln\frac{1}{\delta}\right)$.

Each update of the policy involves updating the n variables and $\mathcal{M}_k$ which takes runtime $O(C)$ and a call of `FixedHorizonEVI` with runtime cost $O(H|\mathcal{S}||\mathcal{A}|C + H|\mathcal{S}|\log|\mathcal{S}|)$. From Lemma C.1 below, we know that the policy can be updated at most $U_{\max}$ times which a gives total runtime for policy updates of

$$\begin{aligned} O(U_{\max}H|\mathcal{S}|(|\mathcal{A}|C + \log|\mathcal{S}|)) =& O\left(H|\mathcal{S}|^2|\mathcal{A}|(|\mathcal{A}|C + \log|\mathcal{S}|)\log\frac{|\mathcal{S}|^2H^2}{\epsilon}\right) \\ =& \tilde{O}\left(H|\mathcal{S}|^2|\mathcal{A}|^2 C\log\frac{1}{\epsilon}\right). \end{aligned}$$

The total runtime of `UCFH` before the policy is ϵ-optimal with probability at least $1-\delta$ is therefore

$$\tilde{O}\left(\frac{H^3|\mathcal{S}|^2|\mathcal{A}|^2 C}{\epsilon^2}\ln\frac{1}{\delta}\right).$$

The space complexity of `UCFH` is dominated by the requirement to store statistics for each possible transition which gives $O(|\mathcal{S}||\mathcal{A}|C)$ complexity.

C Detailed Proofs for the Upper PAC Bound

C.1 Bound on the Number of Policy Changes of `UCFH`

Lemma C.1. *The total number of updates is bounded by* $U_{\max} = |\mathcal{S}\times\mathcal{A}|\log_2\frac{|\mathcal{S}|H}{w_{\min}}$.

Proof. First note that $n(s,a)$ is never never decreasing and no updates happen once $n(s,a) \geq |\mathcal{S}|mH$ for all (s,a). In each update, the $n(s,a)$ of exactly one (s,a) pair increases by $\max\{mw_{\min}, n(s,a)\}$. For a single (s,a) pair, such updates can happen only $\log_2(|\mathcal{S}|mH) - \log_2(mw_{\min})$ times. Hence, there are at most $|\mathcal{S}\times\mathcal{A}|\log_2\frac{|\mathcal{S}|mH}{w_{\min}m}$ updates in total. ☐

C.2 Proof of Lemma 1 – Capturing the true MDP

Proof. For a single (s,a) pair, $s' \in \mathcal{S}(s,a)$ and k, we can treat the event that s' is the successor state of s when chosing action a as a Bernoulli random variable with probability $p(s'|s,a)$. Using Hoeffding's inequality,[7] we then realize that

$$|p(s'|s,a) - \hat{p}(s'|s,a)| \leq \sqrt{\frac{\ln(6/\delta_1)}{2n}}$$

and by Bernstein's inequality

$$|p(s'|s,a) - \hat{p}(s'|s,a)| \leq \sqrt{\frac{2p(s'|s,a)(1-p(s'|s,a))\ln(6/\delta_1)}{n}} + \frac{2}{3n}\ln(6/\delta_1)$$

with probability at least $1-\delta_1/3$ respectively. Using both inequalities of Theorem 10 by Maurer and Pontil [18][8] and the fact that $(a+b)^2 \geq a^2+b^2$, we have

$$|p(s'|s,a)(1-p(s'|s,a)) - \tilde{p}(s'|s,a)(1-\tilde{p}(s'|s,a))| \leq \frac{2\ln(6/\delta_1)}{n-1}$$

---

[7] While the considered random variables are strictly speaking not necessarily independent, they can be treated as such for the concentration inequalities applied here. See Appendix A of Strehl and Littman [16] for details.

[8] The empirical variance denoted by $V_n(\mathbf{X})$ by Maurer and Pontil [18] is $\tilde{p}(s'|s,a)(1-\tilde{p}(s'|s,a))$ in our case and $\mathbb{E}V_n$ is the true variance which amounts to $p(s'|s,a)(1-p(s'|s,a))$ for us.

for $n > 1$ with probability at least $1 - \delta_1/3$. All three inequalities hold with probability $1 - \delta_1$ by the union bound. By Lemma C.1, there are at most $U_{\max}$ updates and so there are at most $U_{\max}$ different k to consider. Since in each update, only a single (s, a) pair with at most C successor states is updated, for all k and (s, a), there are only $U_{\max}C$ different $\hat{p}(s'|s, a)$ to consider. Applying the union bound, we get that $M \notin \mathcal{M}_k$ for any k with probability at most $U_{\max}C\delta_1$. By setting $\delta_1 = \frac{\delta}{2CU_{\max}}$ we get the desired result. ☐

C.3 Bounding the number of episodes with $\kappa > |X_{k,\kappa,\iota}|$ for some κ, ι

Before presenting the proof of Lemma 2 which bounds the total number of episodes where there is a κ and ι such that $\kappa > |X_{k,\kappa,\iota}|$, we establish a bound for each individual κ and ι in the following two additional lemmas.

Lemma C.2 (Bound on observations of $X_{\cdot,\kappa,\iota}$)**.** *The total number of observations of* $(s, a) \in X_{k,\kappa,\iota}$ *where* $\kappa \in [1, |\mathcal{S}| - 1]$ *and* $\iota > 0$ *over all phases* k *is at most* $3|\mathcal{S} \times \mathcal{A}|mw_\iota\kappa$*. The variable* w_ι *is the smallest possible weight of a* (s, a)*-pair that has importance* ι*.*

Proof. We denote the smallest possible weight for any (s, a) pair such that $\iota(s, a) = \iota$ by $w_\iota := \min\{w(s, a) : \iota_k(s, a) = \iota\}$. Note that $w_{\iota+1} = 2w_\iota$ for $\iota > 0$. Consider any phase k and fix $(s, a) \in X_{k,\kappa,\iota}$ with $\iota > 0$. Since we assumed $\iota_k(s, a) = \iota > 0$, we have $w_\iota \leq w_k(s, a) < 2w_\iota$. From $\kappa_k(s, a) = \kappa$, it follows that

$$\frac{n_k(s, a)}{2mw_k(s, a)} \leq \kappa \leq \frac{n_k(s, a)}{mw_k(s, a)}$$

which implies that

$$mw_\iota\kappa \leq mw_k(s, a)\kappa \leq n_k(s, a) \leq 2mw_k(s, a)\kappa \leq 4mw_\iota\kappa. \tag{4}$$

Hence, each state can only be observed $3mw_\iota$ times while being in $\{(s, a) \in X_{k,\kappa,\iota} : k \in \mathbb{N}\}$. ☐

Lemma C.3. *The number of episodes* $E_{\kappa,\iota}$ *in phases with* $|X_{k,\kappa,\iota}| > \kappa$ *is bounded for every* $\alpha \geq 3$ *with high probability,*

$$P(E_{\kappa,\iota} > \alpha N) \leq \exp\left(-\frac{\beta w_\iota(\kappa + 1)N}{H}\right)$$

where $N = |\mathcal{S} \times \mathcal{A}|m$ *and* $\beta = \frac{\alpha(3/\alpha - 1)^2}{7/3 - 1/\alpha}$.

Proof. Let $\nu_i := \sum_{t=1}^{H} \mathbb{I}\{(s_t, a_t) \in X_{k,\kappa,\iota}\}$ be the number of observations of (s, a) in $X_{k,\kappa,\iota}$ in the ith epsiode with $X_{k,\kappa,\iota} > \kappa$. We have $i \in \{1, \dots E_{\kappa,\iota}\})$ and k is the phase that episode i belongs to.

Since $X_{k,\kappa,\iota} \geq \kappa + 1$ and all states in partition (κ, ι) have $w_k(s, a) \geq w_\iota$, we get

$$\mathbb{E}[\nu_i|\nu_1, \dots \nu_{i-1}] \geq (\kappa + 1)w_\iota. \tag{5}$$

Also $\mathbb{V}[\nu_i|\nu_1 \dots \nu_{i-1}] \leq \mathbb{E}[\nu_i|\nu_1, \dots \nu_{i-1}]H$ as $\nu_i \in [0, H]$.

To reason about $E_{\kappa,\iota}$, we define the continuation

$$\nu_i^+ := \begin{cases} \nu_i & \text{if } i \leq E_{\kappa,\iota} \\ w_\iota(\kappa + 1) & \text{otherwise} \end{cases}$$

and the centered auxiliary sequence

$$\bar{\nu}_i := \frac{\nu_i^+ w_\iota(\kappa + 1)}{\mathbb{E}[\nu_i^+|\nu_1^+, \dots \nu_{i-1}^+]}.$$

By construction

$$\mathbb{E}[\bar{\nu}_i|\bar{\nu}_1, \dots \bar{\nu}_{i-1}] = w_\iota(\kappa + 1)\frac{\mathbb{E}[\nu_i^+|\bar{\nu}_1, \dots, \bar{\nu}_{i-1}]}{\mathbb{E}[\nu_i^+|\nu_1^+, \dots \nu_{i-1}^+]} = w_\iota(\kappa + 1).$$

By Lemma C.2, we have that $E_{\kappa,\iota} > \alpha N$ only if

$$\sum_{i=1}^{\alpha N} \bar{\nu}_i \leq 3Nw_\iota \kappa \leq 3Nw_\iota(\kappa + 1).$$

Define now the martingale

$$B_i := \mathbb{E}\left[\sum_{j=1}^{\alpha N} \bar{\nu}_j | \bar{\nu}_1, \dots \bar{\nu}_i\right] = \sum_{j=1}^{i} \bar{\nu}_j + \sum_{j=i+1}^{\alpha N} \mathbb{E}[\bar{\nu}_j | \bar{\nu}_1 \dots \bar{\nu}_i]$$

which gives $B_0 = \alpha N w_\iota(\kappa+1)$ and $B_{\alpha N} = \sum_{i=1}^{\alpha N} \bar{\nu}_i$. Further, since $\nu_i^+ \in [0, H]$ and Equation (5), we have

$$\begin{aligned} |B_{i+1} - B_i| = |\bar{\nu}_i - \mathbb{E}[\bar{\nu}_i | \bar{\nu}_1, \dots, \bar{\nu}_{i-1}]| &= \left| \frac{w_\iota(\kappa+1)(\nu_i^+ - \mathbb{E}[\nu_i^+ | \bar{\nu}_1, \dots \bar{\nu}_{i-1}])}{\mathbb{E}[\nu_i^+ | \nu_1^+, \dots \nu_{i-1}^+]} \right| \\ &\leq \left| \nu_i^+ - \mathbb{E}[\nu_i^+ | \bar{\nu}_1, \dots \bar{\nu}_{i-1}] \right| \leq H. \end{aligned}$$

Using

$$\begin{aligned} \sigma^2 &:= \sum_{i=1}^{\alpha N} \mathbb{V}[B_i - B_{i-1} | B_1 - B_0, \dots B_{i-1} - B_{i-2}] \\ &= \sum_{i=1}^{\alpha N} \mathbb{V}[\bar{\nu}_i | \bar{\nu}_1, \dots \bar{\nu}_{i-1}] \leq \alpha N H w_\iota(\kappa+1) = HB_0 \end{aligned}$$

we can apply Theorem 22 by Chung and Lu [21] and obtain

$$\begin{aligned} \mathbb{P}(E_{\kappa,\iota} > \alpha N) &\leq \mathbb{P}\left(\sum_{i=1}^{\alpha N} \bar{\nu}_i \leq 3Nw_\iota(\kappa+1)\right) \\ &= \mathbb{P}(B_{\alpha N} - B_0 \leq 3B_0/\alpha - B_0) = \mathbb{P}(B_{\alpha N} - B_0 \leq -(1 - 3/\alpha) B_0) \\ &\leq \exp\left(-\frac{(3/\alpha - 1)^2 B_0^2}{2\sigma^2 + H(1/3 - 1/\alpha)B_0}\right) \end{aligned}$$

for $\alpha \geq 3$. We can further simplify the bound to

$$\begin{aligned} \mathbb{P}(E_{\kappa,\iota} > \alpha N) &\leq \exp\left(-\frac{(3/\alpha - 1)^2 B_0^2}{2HB_0 + H(1/3 - 1/\alpha)B_0}\right) \\ &\leq \exp\left(-\frac{(3/\alpha - 1)^2}{2 + (-1/\alpha + 1/3)} \frac{B_0}{H}\right) \\ &= \exp\left(-\frac{\alpha(3/\alpha - 1)^2}{7/3 - 1/\alpha} \frac{Nw_\iota(\kappa+1)}{H}\right). \end{aligned}$$

□

We are now ready to prove Lemma 2 by combining the bound in the previous lemma for all κ and ι.

Proof of Lemma 2. Since $w_k(s,a) \leq H$, we have that $\frac{w_k(s,a)}{w_{\min}} < \frac{H}{w_{\min}}$ and so $\iota_k(s,a) \leq H/w_{\min} = 4H^2|\mathcal{S}|/\epsilon$. In addition, $|X_{k,\kappa,\iota}| \leq |\mathcal{S}|$ for all k, κ, ι and so $|X_{k,\kappa,\iota}| > \kappa$ can only be true for $\kappa \leq |\mathcal{S}|$. Hence, only

$$E_{\max} = \log_2 \frac{H}{w_{\min}} \log_2 |\mathcal{S}|$$

possible values for (κ, ι) exists that can have $|X_{k,\kappa,\iota}| > \kappa$. Using the union bound over all (κ, ι) and Lemma C.3, we get that

$$\begin{aligned}\mathbb{P}(E \leq \alpha N E_{\max}) \geq& \mathbb{P}(\max_{(\kappa,\iota)} E_{\kappa,\iota} \leq \alpha N) \geq 1 - E_{\max} \exp\left(-\frac{\beta w_\iota (\kappa+1) N}{H}\right) \\ \geq& 1 - E_{\max} \exp\left(-\frac{\beta w_{\min} N}{H}\right) = 1 - E_{\max} \exp\left(-\frac{\beta w_{\min} m |\mathcal{S} \times \mathcal{A}|}{H}\right) \\ =& 1 - E_{\max} \exp\left(-\frac{\beta \epsilon m |\mathcal{S} \times \mathcal{A}|}{4H^2 |\mathcal{S}|}\right)\end{aligned}$$

Bounding the right hand-side by $1 - \delta/2$ and solving for m gives

$$1 - E_{\max} \exp\left(-\frac{\beta \epsilon m |\mathcal{S} \times \mathcal{A}|}{4H^2 |\mathcal{S}|}\right) \geq 1 - \delta/2 \quad \Leftrightarrow \quad m \geq \frac{4H^2 |\mathcal{S}|}{|\mathcal{S} \times \mathcal{A}| \beta \epsilon} \ln \frac{2E_{\max}}{\delta}$$

Hence, the condition

$$m \geq \frac{4H^2}{\beta \epsilon} \ln \frac{2E_{\max}}{\delta}$$

is sufficient for the desired result to hold. By plugging in $\alpha = 6$ and $\beta = \frac{\alpha(3/\alpha - 1)^2}{7/3 - 1/\alpha} = \frac{9}{13} \geq \frac{2}{3}$, we obtain the statement to show. ☐

C.4 Bound on the value function difference for episodes with $\forall \kappa, \iota : |X_{k,\kappa,\iota}| \leq \kappa$

To prove Lemma 3, it is sufficient to consider a fixed phase k. To avoid notational clutter, we therefore omit the phase indices k in this section.

For the proof, we reason about a sequence of MDPs M_d which have the same transition probabilities but different reward functions $r^{(d)}$. For $d = 0$, the reward function is the original reward function r of M, i.e. $r_t^{(0)} = r_t$ for all $t = 1 \dots H$. The following reward functions are then defined recursively as $r_t^{(2d+2)} = \sigma_{t:H}^{(d),2}$, where $\sigma_{t:H}^{(d),2}$ is the local variance of the value function w.r.t. the rewards $r^{(d)}$. Note that for every d and $t = 1 \dots H$ and $s \in \mathcal{S}$, we have $r_t^{(d)}(s) \in [0, H^d]$. In complete analogy, we define $\tilde{M}_d$ and $\hat{M}_d$.

We first prove a sequence of lemmas necessary for Lemma 3.

Lemma C.4.

$$V_{i,j} - \tilde{V}_{i,j} = \sum_{t=i}^{H-1} P_{i:t-1}(P_t - \tilde{P}_t)\tilde{V}_{t+1:j}$$

Proof.

$$\begin{aligned}V_{i,j}(s) - \tilde{V}_{i,j}(s) =& r(s) + P_i V_{i+1:j}(s) - r(s) - \tilde{P}_i \tilde{V}_{i+1:j}(s) + P_i \tilde{V}_{i+1,j}(s) - P_i \tilde{V}_{i+1:j}(s) \\ =& P_i(V_{i+1:j} - \tilde{V}_{i+1:j}) + (P_i - \tilde{P}_i)\tilde{V}_{i+1:j}(s)\end{aligned}$$

Since we have $V_{j:j}(s) = r(s) = \tilde{V}_{j:j}(s)$, we can recursively expand the first difference until $i = j$ and get

$$V_{i,j} - \tilde{V}_{i,j} = \sum_{t=i}^{j-1} P_{i:t-1}(P_t - \tilde{P}_t)\tilde{V}_{t+1:j}$$

☐

Lemma C.5. *Assume $p, \hat{p}, \tilde{p} \in [0,1]$ satisfy $\hat{p} \in \mathcal{P}$ and $\tilde{p} \in \mathcal{P}$ where*

$$\begin{aligned}\mathcal{P} := \Bigg\{ p' \in [0,1] : & |p - p'| \leq \sqrt{\frac{\ln(6/\delta_1)}{2n}}, \\ & |p - p'| \leq \sqrt{\frac{2p(1-p)}{n} \ln(6/\delta_1)} + \frac{2}{3n} \ln(6/\delta_1), \\ & \text{if } n > 1 : |p'(1-p') - p(1-p)| \leq \frac{2\ln(6/\delta_1)}{n-1} \Bigg\}.\end{aligned}$$

Then

$$|p-\tilde{p}| \le \sqrt{\frac{8\tilde{p}(1-\tilde{p})}{n}\ln(6/\delta_1)} + \frac{16}{3(n-1)}\ln(6/\delta_1).$$

Proof.

$$\begin{aligned}
|p-\tilde{p}| \le & |p-\hat{p}| + |\hat{p}-\tilde{p}| \le 2\sqrt{\frac{2\hat{p}(1-\hat{p})}{n}\ln(6/\delta_1)} + 2\frac{2}{3n}\ln(6/\delta_1) \\
\le & 2\sqrt{\frac{2}{n}\left(\tilde{p}(1-\tilde{p}) + \frac{2\ln(6/\delta_1)}{n-1}\right)\ln(6/\delta_1)} + \frac{4}{3n}\ln(6/\delta_1) \\
\le & 2\sqrt{\frac{2\tilde{p}(1-\tilde{p})}{n}\ln(6/\delta_1)} + 2\frac{2\ln(6/\delta_1)}{n-1} + \frac{4}{3n}\ln(6/\delta_1) \\
\le & 2\sqrt{\frac{2\tilde{p}(1-\tilde{p})}{n}\ln(6/\delta_1)} + \frac{16}{3(n-1)}\ln(6/\delta_1)
\end{aligned}$$

□

Lemma C.6. *Assume*

$$|p(s'|s,a) - \tilde{p}_i(s'|s,a)| \le c_1(s,a) + c_2(s,a)\sqrt{\tilde{p}_i(s'|s,a)(1-\tilde{p}_i(s'|s,a))}$$

for $a = \pi_i(s)$ *and all* $s', s \in \mathcal{S}$. *Then*

$$|(P_i - \tilde{P}_i)\tilde{V}_{i+1:j}(s)| \le c_1(s,a)|\mathcal{S}(s,a)|\|\tilde{V}_{i+1:j}\|_\infty + c_2(s,a)\sqrt{|\mathcal{S}(s,a)|}\tilde{\sigma}_{i:j}(s)$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ *where* $\mathcal{S}(s,a)$ *denotes the set of possible successor states of state* s *and action* a.

Proof. Let s and $a = \pi_i(s)$ be fixed and define for this fixed s the constant function $\bar{V}(s') = \tilde{P}_i\tilde{V}_{i+1:j}(s)$ *[sic]* as the expected value function of the successor states of s. Note that $\bar{V}(s')$ is a constant function and so $\bar{V} = \tilde{P}_i\bar{V} = P_i\bar{V}$.

$$\begin{aligned}
&|(P_i - \tilde{P}_i)\tilde{V}_{i+1:j}(s)| = |(P_i - \tilde{P}_i)\tilde{V}_{i+1:j}(s) + \bar{V}(s) - \bar{V}(s)| \\
=&|(P_i - \tilde{P}_i)(\tilde{V}_{i+1:j} - \bar{V})(s)|
\end{aligned}$$

$$\le \sum_{s' \in \mathcal{S}(s,a)} |p(s'|s,a) - \tilde{p}_i(s'|s,a)||\tilde{V}_{i+1:j}(s') - \bar{V}(s')| \tag{6}$$

$$\begin{aligned}
\le & \sum_{s' \in \mathcal{S}(s,a)} \left(c_1(s,a) + c_2(s,a)\sqrt{\tilde{p}_i(s'|s,a)(1-\tilde{p}_i(s'|s,a))}\right)|\tilde{V}_{i+1:j}(s') - \bar{V}(s')| \\
\le & |\mathcal{S}(s,a)|c_1(s,a)\|\tilde{V}_{i+1:j}\|_\infty + c_2(s,a)\sum_{s' \in \mathcal{S}(s,a)}\sqrt{\tilde{p}_i(s'|s,a)(1-\tilde{p}_i(s'|s,a))(\tilde{V}_{i+1:j}(s') - \bar{V}(s'))^2}
\end{aligned}$$

$$\le |\mathcal{S}(s,a)|c_1(s,a)\|\tilde{V}_{i+1:j}\|_\infty + c_2(s,a)\sqrt{|\mathcal{S}(s,a)|\sum_{s' \in \mathcal{S}(s,a)}\tilde{p}_i(s'|s,a)(1-\tilde{p}_i(s'|s,a))(\tilde{V}_{i+1:j}(s') - \bar{V}(s'))^2} \tag{7}$$

$$\begin{aligned}
\le & |\mathcal{S}(s,a)|c_1(s,a)\|\tilde{V}_{i+1:j}\|_\infty + c_2(s,a)\sqrt{|\mathcal{S}(s,a)|\sum_{s' \in \mathcal{S}(s,a)}\tilde{p}_i(s'|s,a)(\tilde{V}_{i+1:j}(s') - \bar{V}(s'))^2} \\
= & |\mathcal{S}(s,a)|c_1(s,a)\|\tilde{V}_{i+1:j}\|_\infty + c_2(s,a)\sqrt{|\mathcal{S}(s,a)|}\tilde{\sigma}_{i:j}(s)
\end{aligned}$$

In Inequality (6), we wrote out the definition of P_i and $\tilde{P}_i$ and applied the triangle inequality. We then applied the assumed bound and bounded $|\tilde{V}_{i+1:j}(s') - \bar{V}(s')|$ by $\|V_{i+1:j}\|_\infty$ as all value functions are nonnegative. In Inequality (7), we applied the Cauchy-Schwarz inequality and subsequently used the fact that each term is the sum is nonnegative and that $(1 - \tilde{p}_i(s'|s,a)) \le 1$. The final equality follows from the definition of $\tilde{\sigma}_{i:j}$. □

C.4.1 Bounding the difference in value function

Lemma C.7. *Assume $M \in \mathcal{M}_k$. If $|X_{\kappa,\iota}| \leq \kappa$ for all (κ, ι). Then*

$$|V^{(d)}_{1:H}(s_0) - \tilde{V}^{(d)}_{1:H}(s_0)| =: \Delta_d \leq \hat{A}_d + \hat{B}_d + \min\{\hat{C}_d, \hat{C}'_d + \hat{C}''\sqrt{\Delta_{2d+2}}\}$$

where

$$\hat{A}_d = \frac{\epsilon}{4}H^d, \qquad \hat{B}_d = \frac{32H^{d+1}\,|\mathcal{K}\times\mathcal{I}|\,C}{3m}\ln\frac{6}{\delta_1},$$

and

$$\hat{C}'_d = \sqrt{C\,|\mathcal{K}\times\mathcal{I}|\,\frac{8}{m}H^{2d+2}\ln\frac{6}{\delta_1}} \qquad \hat{C}_d = \hat{C}'_d\sqrt{H}, \qquad \hat{C}'' = \sqrt{C\,|\mathcal{K}\times\mathcal{I}|\,\frac{8}{m}\ln\frac{6}{\delta_1}}.$$

Proof.

$$\begin{aligned}
\Delta_d =& |V^{(d)}_{1:H}(s_0) - \tilde{V}^{(d)}_{1:H}(s_0)| = \left|\sum_{t=1}^{H-1} P_{1:t-1}(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}(s_0)\right| \\
&\leq \sum_{t=1}^{H-1} P_{1:t-1}|(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}|(s_0) \\
&= \sum_{t=1}^{H-1} P_{1:t-1}\left(\sum_{s,a\in\mathcal{S}\times\mathcal{A}} \mathbb{I}\{s=\cdot, a=\pi_t(s)\}|(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}|\right)(s_0) \\
&= \sum_{s,a\in\mathcal{S}\times\mathcal{A}}\sum_{t=1}^{H-1} P_{1:t-1}\left(\mathbb{I}\{s=\cdot, a=\pi_t(s)\}|(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}|\right)(s_0) \\
&= \sum_{s,a\in\mathcal{S}\times\mathcal{A}}\sum_{t=1}^{H-1} P_{1:t-1}\left(\mathbb{I}\{s=\cdot, a=\pi_t(s)\}|(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}(s)|\right)(s_0)
\end{aligned}$$

The first equality follows from Lemma C.4, the second step from the fact that $V_{t+1:H} \geq 0$ and $P_{1:t-1}$ being non-expansive. In the third, we introduce an indicator function which does not change the value as we sum over all (s, a) pairs. The fourth step relies on the linearity of the $P_{i:j}$ operators. In the fifth step, we realize that $\mathbb{I}\{s = \cdot, a = \pi_t(s)\}|(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}(\cdot)$ is a function that takes nonzero values only for input s. We can therefore replace the argument of the second term with s without changing the value. The term then becomes constant and by linearity of $P_{i:j}$, we can write

$$|V^{(d)}_{1:H}(s_0) - \tilde{V}^{(d)}_{1:H}(s_0)| = \Delta_d \leq \sum_{s,a\in\mathcal{S}\times\mathcal{A}}\sum_{t=1}^{H-1} |(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}(s)|(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)$$

$$\begin{aligned}
\leq& \sum_{s,a\notin X}\sum_{t=1}^{H-1} \|\tilde{V}^{(d)}_{t+1:H}\|_\infty (P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0) \\
&+ \sum_{s,a\in X}\sum_{t=1}^{H-1} |(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}(s)|(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0) \\
\leq& \sum_{s,a\notin X}\sum_{t=1}^{H-1} H^{d+1}(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0) \\
&+ \sum_{s,a\in X}\sum_{t=1}^{H-1} |(P_t - \tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}(s)|(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{s,a\notin X}\sum_{t=1}^{H-1} H^{d+1}(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)\\
&\quad+\sum_{s,a\in X}\sum_{t=1}^{H-1}\left||\mathcal{S}(s,a)|c_1(s,a)H^{d+1}+c_2(s,a)\sqrt{|\mathcal{S}(s,a)|}\tilde{\sigma}^{(d)}_{t:H}(s,a)\right|(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{s,a\notin X}\sum_{t=1}^{H} H^{d+1}(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)\\
&\quad+\sum_{s,a\in X}\sum_{t=1}^{H}\left||\mathcal{S}(s,a)|c_1(s,a)H^{d+1}\right|(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)\\
&\quad+\sum_{s,a\in X}\sum_{t=1}^{H-1}\left|c_2(s,a)\sqrt{|\mathcal{S}(s,a)|}\tilde{\sigma}^{(d)}_{t:H}(s,a)\right|(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)\\
&\leq \sum_{s,a\notin X} H^{d+1}w(s,a)+\sum_{s,a\in X}|\mathcal{S}(s,a)|c_1(s,a)H^{d+1}w(s,a)\\
&\quad+\sum_{s,a\in X}\sqrt{|\mathcal{S}(s,a)|}c_2(s,a)\sum_{t=1}^{H-1}\tilde{\sigma}^{(d)}_{t:H}(s,a)(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)\\
&\leq \sum_{s,a\notin X} H^{d+1}w(s,a)+\sum_{s,a\in X}Cc_1(s,a)H^{d+1}w(s,a)\\
&\quad+\sum_{s,a\in X}\sqrt{C}c_2(s,a)\sum_{t=1}^{H-1}\tilde{\sigma}^{(d)}_{t:H}(s,a)(P_{1:t-1}\mathbb{I}\{s=\cdot, a=\pi_t(s)\})(s_0)
\end{aligned}
$$

In the second inequality, we split the sum over all (s,a) pairs and used the fact that P_t and $\tilde{P}_t$ are non-expansive, i.e., $|(P_t-\tilde{P}_t)\tilde{V}^{(d)}_{t+1:H}(s)|\leq \|V^{(d)}_{t+1:H}\|_\infty$. The next step follows from $\|V^{(d)}_{t+1:H}\|_\infty \leq \|V^{(d)}_{1:H}\|_\infty \leq H^{d+1}$. We then apply Lemma C.6 and subsequently use that all terms are nonnegative and the definition of $w(s,a)$. Eventually, we use that $|\mathcal{S}(s,a)|\leq C$ for all s,a. Using the assumption that $M\in\mathcal{M}_k$ and $\tilde{M}\in\mathcal{M}_k$ from Lemma A.1, we can apply Lemma C.5 and get that

$$
c_2(s,a)=\sqrt{\frac{8}{n(s,a)}\ln\frac{6}{\delta_1}}\quad\text{and}\quad c_1(s,a)=\frac{16}{3(n(s,a)-1)}\ln\frac{6}{\delta_1}.
$$

Hence, we can bound

$$
\Delta_d \leq A(s_0)+B(s_0)+C(s_0)
$$

as a sum of three terms which we will consider individually in the following. The first term is

$$
A(s_0)=\sum_{s,a\notin X}H^{d+1}w(s,a)\leq w_{\min}|\mathcal{S}|H^{d+1}\leq\frac{\epsilon H^{d+1}|\mathcal{S}|}{4H|\mathcal{S}|}=\frac{\epsilon}{4}H^d=\hat{A}_d
$$

as $w(s,a)\leq w_{\min}$ for all s,a not in the active set and that the policy is deterministic, which implies that there are only $|\mathcal{S}|$ nonzero w. The next term is

$$
\begin{aligned}
B(s_0)=&C\sum_{s,a\in X}w(s,a)H^{d+1}\frac{16}{3(n(s,a)-1)}\ln\frac{6}{\delta_1}\\
=&H^{d+1}C\ln\frac{6}{\delta_1}\sum_{\kappa,\iota}\sum_{s,a\in X_{\kappa,\iota}}w(s,a)\frac{16}{3(n(s,a)-1)}\\
\leq&H^{d+1}\frac{16C}{3}\ln\frac{6}{\delta_1}\sum_{\kappa,\iota}\sum_{s,a\in X_{\kappa,\iota}}\frac{w(s,a)}{n(s,a)}\frac{n(s,a)}{n(s,a)-1}.
\end{aligned}
$$

For $s, a \in X_{\kappa,\iota}$, we have $n(s,a) \geq mw(s,a)\kappa$ (see Equation (4)) and so

$$\frac{w(s,a)}{n(s,a)} \leq \frac{1}{\kappa m}. \tag{8}$$

Further, for all relevant (s,a)-pairs, we have $n(s,a) > 1$ (follows from $|X_{\kappa,\iota}| \leq \kappa$) which implies

$$B(s_0) \leq H^{d+1} \frac{32C}{3} \ln \frac{6}{\delta_1} \sum_{\kappa,\iota} \frac{|X_{\kappa,\iota}|}{\kappa m}$$

and since we assumed $|X_{\kappa,\iota}| \leq \kappa$

$$B(s_0) \leq \frac{32H^{d+1} \left|\mathcal{K} \times \mathcal{I}\right| C}{3m} \ln \frac{6}{\delta_1} = \hat{B}_d$$

where $\mathcal{K} \times \mathcal{I}$ is the set of all possible (κ,ι)-pairs. The last term is

$$\begin{aligned}
C(s_0) =& \sqrt{C} \sum_{s,a \in X} c_2(s,a) \sum_{t=1}^{H-1} \tilde{\sigma}^{(d)}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\} \\
\leq& \sqrt{C} \sum_{s,a \in X} c_2(s,a) \sum_{t=1}^{H-1} \tilde{\sigma}^{(d)}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\} \\
\leq& \sqrt{C} \sum_{s,a \in X} c_2(s,a) \sqrt{\sum_{t=1}^{H-1} P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}} \sqrt{\sum_{t=1}^{H-1} \tilde{\sigma}^{(d),2}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}} \\
\leq& \sqrt{C} \sum_{s,a \in X} \sqrt{\frac{8w(s,a)}{n(s,a)} \ln \frac{6}{\delta_1} \sum_{t=1}^{H-1} \tilde{\sigma}^{(d),2}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}}
\end{aligned}$$

where we first applied the Cauchy-Schwarz inequality and then used the definition of $c_2(s,a)$ and $w(s,a)$.

$$\begin{aligned}
C(s_0) \leq& \sqrt{C} \sum_{\kappa,\iota} \sum_{s,a \in X_{\kappa,\iota}} \sqrt{\frac{8w(s,a)}{n(s,a)} \ln \frac{6}{\delta_1} \sum_{t=1}^{H-1} \tilde{\sigma}^{(d),2}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}(s_0)} \\
\leq& \sqrt{C} \sum_{\kappa,\iota} \sqrt{|X_{\kappa,\iota}| \sum_{s,a \in X_{\kappa,\iota}} \frac{8w(s,a)}{n(s,a)} \ln \frac{6}{\delta_1} \sum_{t=1}^{H-1} \tilde{\sigma}^{(d),2}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}(s_0)} \\
\leq& \sqrt{C} \sum_{\kappa,\iota} \sqrt{\sum_{s,a \in X_{\kappa,\iota}} \frac{8}{m} \ln \frac{6}{\delta_1} \sum_{t=1}^{H-1} \tilde{\sigma}^{(d),2}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}(s_0)} \\
\leq& \sqrt{C \left|\mathcal{K} \times \mathcal{I}\right| \frac{8}{m} \ln \frac{6}{\delta_1} \sum_{s,a \in X} \sum_{t=1}^{H-1} \tilde{\sigma}^{(d),2}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}(s_0)} \\
\leq& \sqrt{C \left|\mathcal{K} \times \mathcal{I}\right| \frac{8}{m} \ln \frac{6}{\delta_1} \sum_{s,a \in \mathcal{S} \times \mathcal{A}} \sum_{t=1}^{H-1} \tilde{\sigma}^{(d),2}_{t:H}(s,a)) P_{1:t-1} \mathbb{I}\{s = \cdot, a = \pi_t(s)\}(s_0)} \\
=& \sqrt{C \left|\mathcal{K} \times \mathcal{I}\right| \frac{8}{m} \ln \frac{6}{\delta_1} \sum_{t=1}^{H-1} P_{1:t-1} \tilde{\sigma}^{(d),2}_{t:H}(s_0)} \qquad (9) \\
\leq& \sqrt{C \left|\mathcal{K} \times \mathcal{I}\right| \frac{8H^{2d+3} \ln(6/\delta_1)}{m}} = \hat{C}_d
\end{aligned}$$

We first split the sum and applied the Cauchy-Schwarz inequality. Then we used again Inequality (8) and $|X_{\kappa,\iota}| \leq \kappa$. In the fourth step, we applied Cauchy-Schwarz and the final inequality follows from $\|\tilde{\sigma}_{t:H}^{(d),2}\|_\infty \leq H^{2d+2}$ and the fact that $P_{1:t-1}$ is non-expansive. Alternatively, we can rewrite the bound in Equation (9) as

$$\begin{aligned}
C(s_0) \leq & \sqrt{C\,|\mathcal{K} \times \mathcal{I}|\, \frac{8}{m} \ln \frac{6}{\delta_1} \sum_{t=1}^{H-1} P_{1:t-1} \tilde{\sigma}_{t:H}^{(d),2}(s_0)} \\
= & \sqrt{C\,|\mathcal{K} \times \mathcal{I}|\, \frac{8}{m} \ln \frac{6}{\delta_1} \sum_{t=1}^{H-1} P_{1:t-1} \tilde{\sigma}_{t:H}^{(d),2}(s_0) - \tilde{P}_{1:t-1} \tilde{\sigma}_{t:H}^{(d),2}(s_0) + \tilde{P}_{1:t-1} \tilde{\sigma}_{t:H}^{(d),2}(s_0)}.
\end{aligned}$$

Lemma 4 shows that the variance $\tilde{\mathcal{V}}_{1:H}^{(d)}$ also satisfies the Bellman equation with the local variances $\tilde{\sigma}_{i:j}^{(d),2}$. This insight allows us to bound $\sum_{t=1}^{H-1} \tilde{P}_{1:t-1} \tilde{\sigma}_{t:H}^{(d),2}(s_0) = \tilde{\mathcal{V}}_{1:H}^{(d)}(s_0) \leq H^{2d+2}$. Also, note that $\tilde{\sigma}_{t:H}^{(d),2} = r_t^{(2d+2)}$ which gives us

$$\begin{aligned}
C(s_0) \leq & \sqrt{C\,|\mathcal{K} \times \mathcal{I}|\, \frac{8}{m} \ln \frac{6}{\delta_1} \left(H^{2d+2} + \sum_{t=1}^{H-1} P_{1:t-1} r_t^{(2d+2)}(s_0) - \tilde{P}_{1:t-1} r_t^{(2d+2)}(s_0) \right)} \\
= & \sqrt{C\,|\mathcal{K} \times \mathcal{I}|\, \frac{8}{m} \ln \frac{6}{\delta_1} \left(H^{2d+2} + V_{1:H}^{(2d+2)}(s_0) - \tilde{V}_{1:H}^{(2d+2)}(s_0) \right)} \\
\leq & \sqrt{C\,|\mathcal{K} \times \mathcal{I}|\, \frac{8}{m} \ln \frac{6}{\delta_1} \left(H^{2d+2} + \Delta_{2d+2} \right)} \\
\leq & \sqrt{C\,|\mathcal{K} \times \mathcal{I}|\, \frac{8}{m} H^{2d+2} \ln \frac{6}{\delta_1}} + \sqrt{C\,|\mathcal{K} \times \mathcal{I}|\, \frac{8}{m} \Delta_{2d+2} \ln \frac{6}{\delta_1}} = \hat{C}'_d + \hat{C}'' \sqrt{\Delta_{2d+2}}
\end{aligned}$$

□

C.4.2 Proof of Lemma 4 (Bellman equation of local value function variances)

Proof of Lemma 4.

$$\begin{aligned}
\mathcal{V}_{i:j}(s) = & \mathbb{E}\left[\left(\sum_{t=i}^{j} r_t(s_t) - V_{i:j}(s_i) \right)^2 | s_i = s \right] \\
= & \mathbb{E}\left[\left(\sum_{t=i+1}^{j} r_t(s_t) - V_{i+1:j}(s_{i+1}) + V_{i+1:j}(s_{i+1}) + r_i(s_i) - V_{i:j}(s_i) \right)^2 | s_i = s \right] \\
= & \mathbb{E}\left[\left(\sum_{t=i+1}^{j} r(s_t) - V_{i+1:j}(s_{i+1}) \right)^2 | s_i = s \right] \\
& + 2\mathbb{E}\left[\left(\sum_{t=i+1}^{j} r_t(s_t) - V_{i+1:j}(s_{i+1}) \right) \left(V_{i+1:j}(s_{i+1}) + r_i(s_i) - V(s_i) \right) | s_i = s \right] \\
& + \mathbb{E}\left[\left(V_{i+1:j}(s_{i+1}) + r_i(s_i) - V_{i:j}(s_i) \right)^2 | s_i = s \right] \\
= & \mathbb{E}\left[\mathcal{V}_{i+1:j}(s_{i+1}) | s_i = s \right] \\
& + 2\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{t=i+1}^{j} r_t(s_t) - V_{i+1:j}(s_{i+1}) \right) \left(V_{i+1:j}(s_{i+1}) + r_i(s_i) - V_{i:j}(s_i) \right) | s_{i+1} \right] | s_i = s \right] \\
& + \mathbb{E}\left[\left(V_{i+1:j}(s_{i+1}) - P_i V_{i+1:j}(s_i) \right)^2 | s_i = s \right]
\end{aligned}$$

where the final equality follows from the tower property of conditional expectations, and the fact that $V_{i:j}(s_i) = P_i V_{i+1:j}(s_i) + r_i(s_i)$. Since by the definition of the value function

$$\mathbb{E}\left[\left(\sum_{t=i+1}^{j} r_t(s_t) - V_{i+1:j}(s_{i+1})\right) | s_{i+1}\right] = 0$$

the middle term vanishes and the last term is by definition $\sigma^2_{i:j}(s)$ we obtain

$$\mathcal{V}_{i:j}(s) = P_i \mathcal{V}_{i+1:j}(s) + \sigma^2_{i:j}(s).$$

Noting that $\mathcal{V}_{j:j}(s) = (r_j(s) - r_j(s))^2 = 0$, we can unroll the equation and obtain

$$\mathcal{V}_{i:j}(s) = \sum_{t=i}^{j} P_{i:t-1}\sigma^2_{t:j}(s).$$

From the definition of $\mathcal{V}_{1:H}$ and the fact that $0 \le r(\cdot) \le r_{\max}$, we see that $0 \le \mathcal{V}_{1:H} \le H^2 r^2_{\max}$ and the final statement of the lemma follows.

□

C.4.3 Proof of Lemma 3

Proof of Lemma 3. The recursive bound from Lemma C.7

$$\Delta_d \le \hat{A}_d + \hat{B}_d + \hat{C}'_d + \hat{C}''\sqrt{\Delta_{2d+2}}$$

has the form $\Delta_d \le Y_d + Z\sqrt{\Delta_{2d+2}}$. Expanding this form and using the triangle inequality gives

$$\begin{aligned}\Delta_0 \le & Y_0 + Z\sqrt{\Delta_2} \le Y_0 + Z\sqrt{Y_2 + Z\sqrt{\Delta_6}} \le Y_0 + Z\sqrt{Y_2} + Z^{3/2}\Delta_6^{1/4} \\ \le & Y_0 + Z\sqrt{Y_2} + Z^{3/2}Y_6^{1/4} + Z^{7/4}\Delta_{14}^{1/8} \le \dots\end{aligned}$$

and by doing this up to level $\gamma = \lceil \frac{\ln H}{2\ln 2} \rceil$, we obtain

$$\Delta_0 \le \sum_{d \in \mathcal{D}\setminus\{\gamma\}} Z^{\frac{2d}{2+d}} Y_d^{\frac{2}{2+d}} + Z^{\frac{2\gamma}{2+\gamma}} \Delta_\gamma^{\frac{2}{2+\gamma}}$$

where $\mathcal{D} = \{0, 2, 6, 14, \dots \gamma\}$. Note that the exponent of H compared to m is the larger in $\hat{C}'_d$ than in $\hat{B}_d$. Therefore, for sufficiently large m, $\hat{C}'_d$ dominates the other term. More precisely, for

$$m \ge \frac{128H}{9} C\,|\mathcal{K} \times \mathcal{I}| \ln \frac{6}{\delta_1} \tag{10}$$

we have $\hat{B}_d \le \hat{C}'_d$. We can therefore consider $Z = \hat{C}''$ and $Y_d = 2\hat{C}'_d + \hat{A}_d$. Also, since $\hat{C}_d \ge \hat{C}'_d$, we can bound $\Delta_\gamma \le \hat{A}_d + 2\hat{C}_d$. For notational simplicity, we will use the auxiliary variable

$$m_1 = \frac{8C|\mathcal{K} \times \mathcal{I}|H^2}{m\epsilon^2} \ln \frac{6}{\delta_1}.$$

and get

$$\begin{aligned} Z = \hat{C}'' = \sqrt{m_1}\frac{\epsilon}{H} \quad & \text{and} \\ Y_d = \hat{A}_d + 2\hat{C}'_d = (1/4 + 2\sqrt{m_1})H^d\epsilon \quad & \text{and} \\ \Delta_\gamma \le \hat{A}_\gamma + 2\hat{C}_\gamma = (1/4 + 2\sqrt{m_1 H})H^\gamma\epsilon. & \end{aligned}$$

Then

$$\left(Z^{2d}Y_d^2\right)^{(2+d)^{-1}} = \left(m_1^d \epsilon^{2d+2}(1/4 + 2\sqrt{m_1})^2\right)^{(2+d)^{-1}} = \epsilon\left(m_1^d \epsilon^d (1/4 + 2\sqrt{m_1})^2\right)^{(2+d)^{-1}}$$

and

$$\left(Z^{2\gamma}\Delta_\gamma\right)^{(2+\gamma)^{-1}} = \left(m_1^\gamma \epsilon^{2\gamma+2}(1/4 + 2\sqrt{m_1 H})^2\right)^{(2+\gamma)^{-1}} = \epsilon\left(m_1^\gamma \epsilon^{\gamma}(1/4 + 2\sqrt{m_1 H})^2\right)^{(2+\gamma)^{-1}} .$$

Putting these pieces together, we obtain

$$\begin{aligned}
\frac{\Delta_0}{\epsilon} \le & \sum_{d\in\mathcal{D}\setminus\{\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left(\frac{1}{4} + 2\sqrt{m_1}\right)^{\frac{2}{d+2}} + (\epsilon m_1)^{\frac{\gamma}{\gamma+2}} \left(\frac{1}{4} + 2\sqrt{Hm_1}\right)^{\frac{2}{\gamma+2}} \\
= & \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0,\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left(\frac{1}{4} + 2\sqrt{m_1}\right)^{\frac{2}{d+2}} + (\epsilon m_1)^{\frac{\gamma}{\gamma+2}} \left(\frac{1}{4} + 2\sqrt{Hm_1}\right)^{\frac{2}{\gamma+2}} \\
\le & \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0,\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left[\left(\frac{1}{4}\right)^{\frac{2}{d+2}} + (2\sqrt{m_1})^{\frac{2}{d+2}}\right] \\
& + (\epsilon m_1)^{\frac{\gamma}{\gamma+2}} \left[\left(\frac{1}{4}\right)^{\frac{2}{\gamma+2}} + \left(2\sqrt{Hm_1}\right)^{\frac{2}{\gamma+2}}\right]
\end{aligned}$$

where we used the fact that $(a+b)^\phi \le a^\phi + b^\phi$ for $a, b > 0$ and $0 < \phi < 1$. We now bound the $H^{1/(2+\gamma)}$ by using the definition of γ. Since

$$\frac{1}{2+\gamma} = \frac{2\ln 2}{4\ln 2 + \ln H} \le 2\log_H 2$$

and since $H \ge 1$, we have $H^{1/(2+\gamma)} \le 4$. Therefore

$$\begin{aligned}
\frac{\Delta_0}{\epsilon} \le & \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0,\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left[\left(\frac{1}{4}\right)^{\frac{2}{d+2}} + (2\sqrt{m_1})^{\frac{2}{d+2}}\right] \\
& + (\epsilon m_1)^{\frac{\gamma}{\gamma+2}} \left[\left(\frac{1}{4}\right)^{\frac{2}{\gamma+2}} + 4\,(2\sqrt{m_1})^{\frac{2}{\gamma+2}}\right] \\
\le & \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0\}} (\epsilon m_1)^{\frac{d}{2+d}} \left[\left(\frac{1}{4}\right)^{\frac{2}{d+2}} + 4\,(2\sqrt{m_1})^{\frac{2}{d+2}}\right] \\
\le & \frac{1}{4} + 2\sqrt{m_1} + \sum_{i=1}^{\log_2\gamma} (\epsilon m_1)^{1-2^{-i}} \left[\left(\frac{1}{4}\right)^{2^{-i}} + 4\,(2\sqrt{m_1})^{2^{-i}}\right] \\
\le & \frac{1}{4} + 2\sqrt{m_1} + \sum_{i=1}^{\log_2\gamma} m_1^{1-2^{-i}} \left[\left(\frac{1}{4}\right)^{2^{-i}} + 4\,(2\sqrt{m_1})^{2^{-i}}\right]
\end{aligned}$$

In the first inequality, we used the bound for $H^{1/(2+\gamma)}$ and in the second inequality we simplified the expression by noting that all terms are nonnegative. In the next step, we re-parameterized the sum. In the final inequality, we used the assumption that $0 < \epsilon \le 1$ and therefore $\epsilon^{1-2^{-i}} \le 1$.

$$\begin{aligned}
\frac{\Delta_0}{\epsilon} \le & \frac{1}{4} + 2\sqrt{m_1} + \frac{1}{4}\sum_{i=1}^{\log_2\gamma} (4m_1)^{1-2^{-i}} + 4\sum_{i=1}^{\log_2\gamma} (m_1)^{1-2^{-i}}\,(4m_1)^{2^{-i-1}} \\
\le & \frac{1}{4} + 2\sqrt{m_1} + \frac{1}{4}\sum_{i=1}^{\log_2\gamma} (4m_1)^{1-2^{-i}} + 16\sum_{i=1}^{\log_2\gamma} \left(\frac{m_1}{4}\right)^{1-2^{-i-1}} .
\end{aligned}$$

By requiring that

$$m_1 \le \frac{1}{4}$$

and noting that $1 - 2^{-i} \ge 1/2$ and $1 - 2^{-i-1} \ge 3/4$ for $i \ge 1$, we can bound the expression by

$$\frac{\Delta_0}{\epsilon} \le \frac{1}{4} + 2\sqrt{m_1} + \frac{1}{4}\log_2(\gamma)\sqrt{4m_1} + 16\log_2(\gamma)\left(\frac{m_1}{4}\right)^{3/4} .$$

By requiring that $m_1 \leq 1/64$ and $m_1 \leq (2\log_2 \gamma)^{-2}$ and $m_1 \leq 1/64(\log_2 \gamma)^{-4/3}$, we can assure that $\Delta_0 \leq \epsilon$. Taking all assumptions on m_1 we made above together, we realize that

$$m_1 \leq \left(\frac{1}{8\log_2\log_2 H}\right)^2 \leq \left(\frac{1}{8\log_2 \gamma}\right)^2$$

is sufficient for them to hold where we used $\log_2 \gamma = \log_2(\lceil \frac{1}{2}\log_2 H \rceil) \leq \log_2 \log_2 H$. This gives the following condition on m

$$m \geq 512 C (\log_2 \log_2 H)^2 |\mathcal{K} \times \mathcal{I}| \frac{H^2}{\epsilon^2} \ln \frac{6}{\delta_1}$$

which is a stronger condition that the one in Equation (10).

By construction of $\iota(s,a)$, we have $\iota(s,a) \leq 2\frac{H}{w_{\min}} = \frac{8|\mathcal{S}|H^2}{\epsilon} = \frac{8H^2|\mathcal{S}|}{\epsilon}$. Also, $\kappa_k(s,a) \leq \frac{|\mathcal{S}|mH}{mw_{\min}} = \frac{4|\mathcal{S}|^2H^2}{\epsilon}$. Therefore

$$|\mathcal{K} \times \mathcal{I}| \leq \log_2 \frac{4|\mathcal{S}|^2H^2}{\epsilon} \log_2 \frac{8H^2|\mathcal{S}|}{\epsilon} \leq \log_2^2 \frac{8H^2|\mathcal{S}|^2}{\epsilon}$$

which let us conclude that

$$m \geq 512 \frac{CH^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2 \left(\frac{8H^2|\mathcal{S}|^2}{\epsilon}\right) \ln \frac{6}{\delta_1}$$

is a sufficient condition and thus, the statement to show, holds. ☐

C.5 Proof of Theorem 1

Proof of Theorem 1. By Lemma 2, we know that the number of episodes where $|X_{\kappa,\iota}| > \kappa$ for some κ, ι is bounded by $6E_{\max}|\mathcal{S} \times \mathcal{A}|m$ with probability at least $1 - \delta/2$. For all other episodes, we have by Lemma 3 that $|\tilde{R}^{\pi_k} - R^{\pi_k}| < \epsilon$. Since, with probability at least $1 - \delta/2$, we have by Lemma 1 $M \in \mathcal{M}_k$, we can use Lemma A.1 which gives $\tilde{R}^{\pi_k} > R^* \geq R^{\pi_k}$ to conclude that with probabilty at least $1 - \delta/2$, for all episodes with $|X_{\kappa,\iota}| \leq \kappa$ for all κ, ι, we have $R^* - R^{\pi_k} < \epsilon$. Applying the union bound, we get the desired result, if m satisfies

$$\begin{aligned} m \geq& 512 \frac{CH^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2 \left(\frac{8H^2|\mathcal{S}|^2}{\epsilon}\right) \ln \frac{6}{\delta_1} \quad \text{and} \\ m \geq& \frac{6H^2}{\epsilon} \ln \frac{2E_{\max}}{\delta}. \end{aligned}$$

From the definitions, we get

$$\ln \frac{6}{\delta_1} = \ln \frac{6CU_{\max}}{\delta} = \ln \frac{6|\mathcal{S} \times \mathcal{A}| C \log_2(|\mathcal{S}|H/w_{\min})}{\delta} = \ln \frac{6|\mathcal{S} \times \mathcal{A}| C \log_2(4|\mathcal{S}|^2H^2/\epsilon)}{\delta}$$

and

$$E_{\max} = \log_2 |\mathcal{S}| \log_2 \frac{4H^2|\mathcal{S}|}{\epsilon} \leq \log_2^2 \frac{4H^2|\mathcal{S}|}{\epsilon}$$

and

$$\begin{aligned} \ln \frac{2E_{\max}}{\delta} =& \ln \frac{2\log_2 |\mathcal{S}| \log_2(4H^2|\mathcal{S}|/\epsilon)}{\delta} \leq \ln \frac{2\log_2^2(4H^2|\mathcal{S}|/\epsilon)}{\delta} \\ \leq& \ln \frac{6\,|\mathcal{S} \times \mathcal{A}| \log_2^2(4|\mathcal{S}|^2H^2/\epsilon)}{\delta}. \end{aligned}$$

Setting

$$m = 512(\log_2 \log_2 H)^2 \frac{CH^2}{\epsilon^2} \log^2 \left(\frac{8H^2|\mathcal{S}|^2}{\epsilon}\right) \ln \frac{6\,|\mathcal{S} \times \mathcal{A}|\, C \log_2^2(4|\mathcal{S}|^2H^2/\epsilon)}{\delta}$$

is therefore a valid choice for m to ensure that with probability at least $1 - \delta$, there are at most

$$6mE_{\max} = 3072(\log_2\log_2 H)^2 \frac{CH^2\,|\mathcal{S}\times\mathcal{A}|}{\epsilon^2} \\ \times \log_2^2\left(\frac{4H^2|\mathcal{S}|}{\epsilon}\right)\log^2\left(\frac{8H^2|\mathcal{S}|^2}{\epsilon}\right)\ln\frac{6|\mathcal{S}\times\mathcal{A}|C\log_2^2(4|\mathcal{S}|^2H^2/\epsilon)}{\delta}$$

ϵ-suboptimal episodes.

□

D Proof of the Lower PAC Bound

Proof of Theorem 2. We consider the class of MDPs shown in Figure 1. The MDPs essentially consist of n parallel multi-armed bandits. For each bandit, there exist $m+1=|\mathcal{A}|$ possible instantiations, which we denote by $I_i = 0\dots m$. The instantiation, or *hypothesis*, $I_i = 0$ corresponds to $\epsilon_i(a) = \mathbb{I}\{a=a_0\}\epsilon'/2$, that is, only action a_0 has a small bias. The other hypotheses $I_i = j$ for $j=1\dots m$ correspond to $\epsilon_i(a) = \mathbb{I}\{a=a_0\}\epsilon'/2 + \mathbb{I}\{a=a_j\}\epsilon'$. We use $I=(I_1,\dots I_n)$ to indicate the instance of the entire MDP.

We define $G_i = \{\omega\in\Omega : \pi(i) = a_{I_i}\}$, the event that π, the policy generated by A chooses optimally in bandit i. For a given instance I, the difference between the optimal expected cumulative reward R^*_I and the expected cumulative reward R^π_I of policy π is at least

$$R^*_I - R^\pi_I \geq (H-2)\left(1-\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{G_i\}\right)\frac{\epsilon'}{2}.$$

For π to be ϵ-optimal, we therefore need

$$\begin{aligned}\epsilon \geq & R^*_I - R^\pi_I \geq (H-2)\left(1-\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{G_i\}\right)\frac{\epsilon'}{2},\\ \frac{2\epsilon}{(H-2)\epsilon'} \geq & \left(1-\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{G_i\}\right),\\ \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{G_i\} \geq & \left(1-\frac{2\epsilon}{(H-2)\epsilon'}\right),\\ \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{G_i\} \geq & \left(1-\frac{2\epsilon(H-2)\eta}{(H-2)16\epsilon e^4}\right) = 1-\frac{\eta}{8e^4}\end{aligned}$$

where we chose value $\epsilon' := \frac{16\epsilon e^4}{(H-2)\eta}$ for ϵ'. We will specify the exact value of parameter η later. The condition basically states that at least a fraction of $\phi := 1-\eta/(8e^4)$ bandits need to be solved optimally by A for the resulting policy π to be ϵ-accurate. For A to be (ϵ,δ)-correct, we therefore need

$$\mathbb{P}_I\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{G_i\} \geq \phi\right) \geq \mathbb{P}_I(R^*_I - R^\pi_I \geq \epsilon) \geq 1-\delta$$

for each instance I. Using Markov's inequality, we obtain

$$1-\delta \leq \mathbb{P}_I\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{G_i\} \geq \phi\right) \leq \frac{1}{n\phi}\sum_{i=1}^n \mathbb{E}_I[\mathbb{I}\{G_i\}] \leq \frac{1}{n\phi}\sum_{i=1}^n \mathbb{P}_I(G_i)$$

All G_i are independent of each other by construction of the MDP. In fact $\sum_{i=1}^n \mathbb{I}\{G_i\}$ is Poisson-binomial distributed as $\mathbb{I}\{G_i\}$ are independent Bernoulli random variables with potentially different mean. Therefore, upper bounds δ_i must exist such that $\delta_i \geq P_I(G_i^C)$ for all hypotheses I and such that $1-\delta \leq \frac{1}{n\phi}\sum_{i=1}^n(1-\delta_i)$ or equivalently $n(1+\delta\phi-\phi) \geq \sum_{i=1}^n \delta_i$. Since all G_i are independent of each other and

$$\epsilon' = \frac{16\epsilon e^4}{(H-2)\eta} \leq \frac{16(H-2)e^4\eta}{(H-2)64e^4\eta} = \frac{1}{4}$$

we can apply Theorem 1 by Mannor and Tsitsiklis [19] in cases where

$$\delta_i \leq \frac{1}{\eta}(1-\phi+\delta\phi) \leq \frac{1}{\eta}(1-\phi+\delta) \leq \frac{1}{8e^4} + \frac{\delta}{\eta} \leq \frac{2}{8e^4}.$$

This result gives us the minimum expected number of times $\mathbb{E}_I[n_i]$ we need to observe state i to ensure that $P_I(G_i^C) \leq \delta_i$

$$\mathbb{E}_I[n_i] \geq \left[\frac{c_1(|\mathcal{A}|-1)}{\epsilon'^2}\ln\left(\frac{c_2}{\delta_i}\right)\right]\mathbb{I}\{\eta\delta_i \leq 1-\phi+\phi\delta\},$$

for appropriate constants c_1 and c_2 (e.g. $c_1 = 400$ and $c_2 = 4$). We can find a valid lower bound for the total number of samples for any $\delta_1, \dots \delta_n$ by considering the worst bound over all $\delta_1, \dots \delta_n$. The following optimization problem encodes this idea

$$\min_{\delta_1,\dots\delta_n} \sum_{i=1}^n \ln\frac{1}{\delta_i}\mathbb{I}\{\eta\delta_i \leq 1-\phi+\phi\delta\} \tag{11}$$
$$\text{s.t.} \sum_{i=1}^n \delta_i \leq n(1+\phi\delta-\phi)$$

As shown in Lemma D.1 in the supplementary material, the optimal solution of the optimization problem in Equation (11) is $\delta_1 = \dots = \delta_n = c$ if $\eta(1-\ln c) \leq 1$ with $c = 1+\delta\phi-\phi$. Since the left-hand side of this condition is decreasing in c, we can plug in a lower bound of $c \geq 1-\phi = \frac{\eta}{8e^4}$ and get the sufficient condition

$$\eta(1-\ln\frac{\eta}{8e^4}) = \eta(1-\ln\eta+4+\ln 8) \leq 1.$$

It is easy to verify that $\eta = 1/10$ satisfies this condition. Hence $\delta_1 = \dots = \delta_n = c$ is the optimal solution to the problem in Equation (11). In each episode, we only observe a single state i and therefore, there need to be at least

$$\mathbb{E}_I[n_A] \geq \sum_{i=1}^n \mathbb{E}_I[n_i] \geq \frac{c_1(|\mathcal{A}|-1)n}{\epsilon'^2}\ln\left(\frac{c_2}{\delta_i}\right) \geq \frac{c_1(|\mathcal{A}|-1)n}{\epsilon'^2}\ln\left(\frac{c_2}{\delta+\frac{\eta}{8e^4}}\right)$$

observed episodes for appropriate constants c_1 and c_2. Plugging in ϵ' and $n = |\mathcal{S}|-3$, we obtain the desired statement.

□

Lemma D.1. *The optimization problem*

$$\min_{\delta_1\dots\delta_n\in[0,1]} \sum_{i=1}^n \ln\frac{1}{\delta_i}\mathbb{I}\{\eta\delta_i \leq c\}$$
$$\textit{s.t.} \sum_{i=1}^n \delta_i \leq nc$$

with $c \in [0,1]$ *and*

$$\eta(1-\ln c) \leq 1$$

has optimal solution $\delta_1 = \dots = \delta_n = c$.

Proof. Without the indicator part in the objective, we can show that $\delta_1 = \dots = \delta_n = c$ is an optimal solution by checking the KKT conditions and noting that the problem is convex. Let k denote the number of δ_j that are set such that the indicator function is 0. Without loss of generality we can assume that their value is $\delta_P := c/\eta$ and the remaining δ_j take the same value δ_A (for a fixed δ_P and k, the problem reduces to the one without the indicator functions). Then the problem transforms into

$$\min_{\delta_A\in(0,1),k\in\{0,1,\dots n\}} (n-k)\ln\frac{1}{\delta_A}$$
$$(n-k)\delta_A + k\delta_P \leq nc$$

We can rewrite the constraint as

$$\begin{aligned}(n-k)\delta_A + k\delta_P &\le nc\\ (n-k)\delta_A &\le nc - k\delta_P = \left(n - \frac{k}{\eta}\right) c\\ \delta_A &\le \frac{n-\frac{k}{\eta}}{n-k} c.\end{aligned}$$

Since the objective decreases with δ_A, it is optimal to choose δ_A as large as possible. The optimization problem then reduces to

$$\min_{k\in\{0,\dots\lfloor n/\gamma\rfloor\}} (n-k)\ln\left(\frac{n-k}{n-\gamma k}c^{-1}\right).$$

where we used for convenience $\gamma := 1/\eta$. We want to show that the optimal solution to this problem is $k = 0$. We can therefore relax the problem to the continuous domain without loss of generality

$$\min_{k\in[0,n/\gamma]} (n-k)\ln\left(\frac{n-k}{n-\gamma k}c^{-1}\right).$$

By reparameterizing the problem with $\alpha = k/n$, we get

$$\min_{\alpha\in[0,1/\gamma]} n(1-\alpha)\ln\left(\frac{1-\alpha}{c(1-\gamma\alpha)}\right).$$

We realize that the minimizer does not depend on n (while the value does). The second derivative of the objective function is

$$n\frac{(\gamma-1)^2}{(1-\gamma\alpha)^2(1-\alpha)},$$

which is nonnegative for $\alpha \in [0, 1/\gamma]$. Hence, the objective is convex in the feasible region and the minimizer of this problem is $\alpha = 0$ if the derivative of the objective is nonnegative in 0. The derivative of the objective in 0 is given by

$$n(\gamma - 1 + \ln(c)).$$

A sufficient condition for $\alpha = 0$ being optimal is therefore

$$\gamma \ge 1 - \ln c$$

or, in terms of the original problem with $\eta = 1/\gamma$, $\delta_1 = \dots \delta_n = c$ is optimal if

$$\eta(1-\ln c) \le 1$$

□