Extended version of paper published in Proc. AAMAS 2025, pages 1079--1087, IFAAMAS/ACM, 2025.

A View of the Certainty-Equivalence Method for PAC RL as an Application of the Trajectory Tree Method

A Preprint

Shivaram Kalyanakrishnan
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai, 400076
shivaram@cse.iitb.ac.in

Sheel Shah
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, 400076
19D070052@iitb.ac.in

Santhosh Kumar Guguloth
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai, 400076
santhoshkg@iitb.ac.in

November 2, 2025

Abstract

Reinforcement learning (RL) enables an agent interacting with an unknown MDP M to optimise its behaviour by observing transitions sampled from M. A natural entity that emerges in the agent's reasoning is $\widehat{M}$, the maximum likelihood estimate of M based on the observed transitions. The well-known *certainty-equivalence* method (CEM) dictates that the agent update its behaviour to $\widehat{\pi}$, which is an optimal policy for $\widehat{M}$. Not only is CEM intuitive, it has been shown to enjoy minimax-optimal sample complexity in some regions of the parameter space for PAC RL with a generative model (Agarwal et al., 2020).

A seemingly unrelated algorithm is the "trajectory tree method" (TTM) (Kearns et al., 1999), originally developed for efficient decision-time planning in large POMDPs. This paper presents a theoretical investigation that stems from the surprising finding that CEM may indeed be viewed as an application of TTM. The qualitative benefits of this view are (1) new and simple proofs of sample complexity upper bounds for CEM, in fact under a (2) weaker assumption on the rewards than is prevalent in the current literature. Our analysis applies to both non-stationary and stationary MDPs. Quantitatively, we obtain (3) improvements in the sample-complexity upper bounds for CEM both for non-stationary and stationary MDPs, in the regime that the "mistake probability" δ is small. Additionally, we show (4) a lower bound on the sample complexity for finite-horizon MDPs, which establishes the minimax-optimality of our upper bound for non-stationary MDPs in the small-δ regime.

1 Introduction

The principle of *certainty-equivalence* has been a recurring theme in the design of reinforcement learning (RL) algorithms (Azar et al., 2013; Agarwal et al., 2020). Concretely, consider an agent interacting with an unknown Markov Decision Problem (MDP) M. The agent gains information about M by repeatedly querying a generative model with an arbitrary (state, action) pair or (state, action, time-step) triple, and it is provided an

accordingly-sampled next state and reward. Based on this set of samples D, the agent must propose rewarding behaviour for M. The first step in applying the certainty-equivalence method (CEM) is to identify $\widehat{M}$, a maximum likelihood estimate of M based on D ($\widehat{M}$ is also called the "empirical model"). The agent then computes a policy $\widehat{\pi}$ that is optimal for $\widehat{M}$. In other words, the agent computes the same behaviour as it would *if it were certain* that $\widehat{M} = M$. The idea is intuitive since $\widehat{M}$ indeed approaches M as D grows larger.

A natural question is whether CEM is *optimal* in its sample complexity. A line of work that formalises the problem using the PAC framework has provided partially affirmative answers, although gaps remain. If M is a stationary MDP, the baseline for comparison has been a sample-complexity lower bound from Azar et al. (2013). These authors also provide a sample-complexity upper bound for an iterative implementation of CEM. Their upper bound matches the lower bound when restrictions are placed on some problem parameters—the tolerance ϵ and the discount factor γ. In subsequent work, Agarwal et al. (2020) partially relax the restriction. Interestingly, Li et al. (2023) show that minimax-optimality is possible over the full range of problem parameters by injecting randomness into CEM (hence, technically, the resulting algorithm is *not* CEM). They also provide an upper bound for CEM itself in the case that M is a non-stationary MDP, adopting the convention of using a finite horizon H in place of discount factor γ. Although the preceding analyses (Azar et al., 2013; Agarwal et al., 2020; Li et al., 2020) vary in approach, they have a common technical core that uses bounds on the variance of the long-term return.

1.1 Contribution

In this paper, we provide an alternative perspective on CEM, which offers a new template for analysis and new upper bounds.

1.1.1 New analytical framework

We illustrate a connection between CEM and the seemingly-unrelated trajectory tree method (TTM), proposed by Kearns et al. (1999) for decision-time planning in large MDPs and POMDPs. A trajectory tree is designed to provide unbiased estimates of the value function of *every* possible policy for the task. In TTM, Kearns et al. (1999) deliberately generate several *independent* trajectory trees, so that confident estimates of value functions can be obtained by averaging. Our main insight is that CEM *implicitly* performs the same kind of averaging. Consequently, we can reuse the proof structure accompanying TTM (summarised in Section 3), only now using a variant of Hoeffding's inequality for a sum of *dependent* random variables (Hoeffding, 1963, see Section 5). Otherwise, we only need elementary probability and counting, setting up simple, intuitive proofs. We also obtain quantitative gains.

1.1.2 Upper bound for non-stationary MDPs

The more straightforward case for us to analyse is when M is *non-stationary*: that is, its dynamics can change over time. Let CEM-NS denote the algorithm based on the certainty-equivalence principle for this setting. Under CEM-NS, the maximum likelihood MDP $\widehat{M}$ is also a non-stationary MDP, which estimates a separate transition probability distribution over next states for each (state, action, time-step) triple. If M has a set of states S, a set of actions A, and horizon H, our analysis shows that CEM-NS requires $O\left(\frac{|S||A|H^3}{\epsilon^2}\log\frac{1}{\delta} + \frac{|S|^2|A|H^4\log|A|}{\epsilon^2}\right)$ samples, where tolerance ϵ and mistake probability δ are the usual PAC parameters (formally specified in the next section). This bound is in general incomparable with the $O\left(\frac{|S||A|H^4}{\epsilon^2}\log\frac{|S||A|H}{\delta}\right)$ upper bound shown recently by Li et al. (2023), and is tighter by a factor of H in the regime of small δ. Interestingly, our upper bound holds with a weaker assumption on the rewards (explained in Section 2.1) than is common in the literature. We present the main elements of our analytical approach, situated in the context of the non-stationary setting, in Section 4.

1.1.3 Upper bound for stationary MDPs

If it is known that M is a stationary MDP, then the certainty-equivalence principle would imply constructing a *stationary* maximum likelihood MDP $\widehat{M}$ by pooling together all the samples for any (state, action) pair. Let CEM-S denote the algorithm that is consistent with this approach. In Section 5, we analyse CEM-S under the usual assumption that M is infinite-horizon, with discount factor $\gamma < 1$. A key technical difference emerges when we analyse CEM-S using the TTM toolkit. In the stationary setting, some trajectory trees—or equivalently, "worlds", as we shall denote them—use the *same* sample transition at different time steps, and therefore no longer provide *unbiased* value estimates of policies. We simply use the fact that such worlds constitute only a small fraction of the universe of worlds, and hence their influence is limited.

Our eventual sample-complexity upper bound for CEM-S is $\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^3\epsilon^2}\left(\log\frac{1}{\delta}+|S||A|\epsilon\right)\right)$, where $\tilde{O}$ suppresses factors that are logarithmic in $\frac{1}{\epsilon}$ and $\frac{1}{1-\gamma}$. This upper bound matches the lower bound from Azar et al. (2013) in the regime of small δ. By contrast, the upper bounds provided by Azar et al. (2013) and Agarwal et al. (2020) hold for all $\delta \in (0,1)$, but unlike ours, apply only to restricted ranges of ϵ.

1.1.4 Lower bound for finite-horizon MDPs

As an independent contribution, we adapt the lower bound of Azar et al. (2013) to the finite horizon setting, showing that $\Omega\left(\frac{|S||A|H^3}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples are necessary on some instances for any PAC algorithm in the finite-horizon setting. This result, presented in Section 6, establishes the new finding that within the small-δ regime, CEM is indeed a minimax-optimal algorithm for non-stationary MDPs.

In short, our paper furthers the understanding of CEM, a natural and intuitive algorithm, by bringing out its connection with TTM, itself a classical algorithm. Our analysis and results are significant to the theory of RL, which is a central paradigm for agent learning. Our work also motivates further analysis and algorithm design. We begin with a formal problem statement (Section 2) and a review of the relevant literature (Section 3) before presenting our analysis.

2 PAC RL: Problem Statement

We formalise the requirement of PAC RL with a generative model.

2.1 Markov Decision Problems

We adopt a definition of MDPs that covers both stationary and non-stationary tasks, with both finite and infinite horizons. An MDP $M = (S, A, T, R, H, \gamma)$ comprises a set of states S and a set of actions A. We assume S and A are finite. Positive integer H (possibly infinite) denotes the task horizon; let $[H]$ denote the set $\{0, 1, 2, \ldots, H-1\}$. The transition function $T : S \times A \times [H] \times S \to [0,1]$ assigns a probability $T(s,a,t,s')$ to change state from $s \in S$ to $s' \in S$ by taking action $a \in A$ at time step $t \in [H]$; hence $\sum_{s' \in S} T(s,a,t,s') = 1$ for $s \in S$, $a \in A$, $t \in [H]$. Taking action $a \in A$ from state $s \in S$ at time step $t \in H$ also earns a numeric reward $R(s,a,t)$. Hence, an agent's interaction with the MDP is a sequence $s^0, a^0, r^0, s^1, a^1, r^1, \ldots, s^{H-1}, a^{H-1}, r^{H-1}, s^H$ wherein for time step $t \in [H]$, the agent (1) takes action a^t from state s^t, (2) obtains reward $r^t = R(s^t, a^t, t)$, and (3) proceeds to state $s^{t+1} \sim T(s^t, a^t, t)$, with the convention that s^H is a terminal state. The discount factor $\gamma \in [0,1]$ is used to compute long-term values; we permit $\gamma = 1$ only when H is finite.

Previous work (Azar et al., 2013; Agarwal et al., 2020; Li et al., 2020) has typically assumed that *each* reward comes from a known, bounded range (taken by convention as $[0,1]$). However, we only enforce the weaker requirement that the discounted *sum* of rewards $\sum_{t\in[H]} \gamma^t r^t$ lie in a known interval (Jiang and Agarwal, 2018). For easy comparison with previous results, we take this interval as $[0, V_{\max}]$, where $V_{\max} \leq \min\left\{H, \frac{1}{1-\gamma}\right\}$, as would follow if each reward is at most 1. To simplify exposition, we assume that the rewards are deterministic,

and that the reward function is known to the agent. Approximating a stochastic reward function R from samples would not alter the asymptotic complexity of our upper bounds, as also observed by Agarwal et al. (2020).

Let $\pi : S \times [H] \to A$ be a non-stationary policy for M. Its value function $V^\pi : S \times [H] \to \mathbb{R}$ specifies the expected long-term discounted reward for each $(s,t) \in S \times [H]$, and is given by

$$V^\pi(s,t) = R(s,\pi(s,t),t) + \gamma \sum_{s' \in S} T(s,\pi(s,t),t,s')V^\pi(s',t+1),$$

with the convention that $V^\pi(\cdot, H) \stackrel{\text{def}}{=} 0$. It is well-known that every MDP has an *optimal* policy $\pi^\star : S \times [H] \to A$, which satisfies $V^{\pi^\star}(s,t) \geq V^\pi(s,t)$ for all $(s,t) \in S \times [H]$ and $\pi : S \times [H] \to A$. The value function of $\pi^\star$ is denoted $V^\star$. We may assume $\pi^\star$ to be *stationary* (that is, independent of time step $t \in [H]$) if M is also stationary (that is, T and R do not depend on t) *and* H is infinite.

2.2 Learning Algorithms

When learning with a generative model, an algorithm $\mathcal{L}$ can repeatedly query arbitrary $(s,a,t) \in S \times A \times [H]$, and is returned $r = R(s,a,t)$, $s' \sim T(s,a,t)$ by the environment. Hence, at any stage, the data D available with the algorithm is the sequence of samples so gathered. Based on D, the algorithm may either pick a new tuple to query, or stop and return a policy.

In the PAC formulation, the other inputs to the learning algorithm are a tolerance parameter $\epsilon \in (0, V_{\max})$ and a mistake probability $\delta \in (0,1)$. The policy π returned by $\mathcal{L}$ is ϵ-optimal if for all $s \in S$, $V^\pi(s,0) \geq V^\star(s,0) - \epsilon$. We require that on every MDP M it is run, $\mathcal{L}$ stop and return an ϵ-optimal policy with probability at least $1-\delta$. The *sample complexity* of $\mathcal{L}$ on a run is the number of samples it has gathered before termination. In this paper, we restrict our attention to worst case sample-complexity upper bounds (across problem instances) for CEM. For simplicity, we assume that the algorithm samples each $(s,a,t) \in S \times A \times [H]$ the same number of times N, where N is a function of $|S|$, $|A|$, H, γ, V_{max}, ϵ, and δ. We seek upper bounds on N to ensure the PAC guarantee.

3 Related Work

In this section, we review sample-complexity bounds for PAC RL, and provide a sketch of TTM.

3.1 PAC RL with a Generative Model

The original PAC formulation of RL was put forth by Fiechter (1994), who established that its sample complexity is polynomial in the problem parameters. Kearns and Singh (1999) then demonstrated that model-free learning algorithms such as Q-learning can also achieve polynomial sample complexity. For a stationary, infinite-horizon MDP, the model size scales as $\Theta(|S|^2|A|)$, whereas Q-learning uses $\Theta(|S||A|)$ entries. Progress on PAC RL with a generative model has accelerated in the last decade, owing to the minimax-optimal bounds furnished by Azar et al. (2013). For stationary, infinite-horizon tasks having k (state, action) pairs, Azar et al. (2013) show a sample-complexity *lower bound* of $\Omega\left(\frac{k}{(1-\gamma)^3\epsilon^2}\log\frac{k}{\delta}\right)$ for obtaining an ϵ-approximation of the optimal action value function $Q^\star$. They construct an MDP instance on which every PAC algorithm must incur at least the specified sample complexity. They also provide an upper bound (applicable to all MDPs), which is is "minimax-optimal" in the sense that there exists an MDP on which the lower and upper bounds match up to a constant factor.

The tools proposed by Azar et al. (2013) have been the basis for many subsequent investigations. The essential idea is to construct the empirical model $\widehat{M}$, and to compute an output policy by running value iteration (or policy iteration) on $\widehat{M}$ for a finite number of iterations. If Q_k is the k-step action value function of the output policy on $\widehat{M}$, for $k \geq 1$, the analysis proceeds by inductively upper-bounding the difference between Q_k and $Q^\star$. Although the original algorithms of Azar et al. (2013) estimate $Q^\star$ with minimax-optimal sample

complexity, they do not automatically yield a near-optimal policy. Obtaining such a policy from the action value function would ordinarily require scaling the sample complexity by $\frac{1}{1-\gamma}$. A variance-reduction technique proposed by Sidford et al. (2018), while different from CEM, directly yields a near-optimal policy without this additional complexity. Yet, the minimax-optimal upper bounds given above do not apply to the entire range of $\epsilon \in (0, V_{\max})$. For instance, the upper bound given by Azar et al. (2013) only holds for $\epsilon \in (0, 1/\sqrt{(1-\gamma)|S|})$, and that of Sidford et al. (2018) only for $\epsilon \in (0, 1]$. The most recent advance in this line of work is due to Agarwal et al. (2020), who show that CEM itself can deliver a near-optimal policy for stationary MDPs with minimax-optimal sample complexity, under the constraint that $\epsilon \leq \sqrt{1/(1-\gamma)}$. The main components of their analysis are bounds on the variance of the return (introduced by Azar et al. (2013)), and an intermediate MDP designed to break the dependence among samples used to construct the empirical model. In contrast to all these approaches, our analysis only relies on a version of Hoeffding's inequality (Hoeffding, 1963). We obtain an upper bound for the entire range of problem parameters, whose ratio to the lower bound approaches a logarithmic term as $\delta \to 0$ (while keeping other parameters fixed).

Li et al. (2023) devise learning algorithms that are minimax-optimal for stationary MDPs for the entire range of parameters, including $\epsilon \in (0, 1/(1-\gamma))$. A key feature of their algorithms is the careful use of randomness for perturbing rewards or action-selection probabilities. The statistical guarantees of these algorithms kick in as soon as the sample size reaches $\Theta(|S||A|/(1-\gamma))$, whereas the so-called "sample barrier" in the guarantees of Agarwal et al. (2020) is $\Theta(|S||A|/(1-\gamma)^2)$. Li et al. (2023) do not need to randomise their algorithm for the non-stationary setting, and consequently it boils down to exactly CEM. Our upper bound for CEM in the non-stationary setting is tighter than theirs by a factor of H in the regime of small δ, although it can be looser for large δ.

Suppose we wish to estimate the action-value for some $(s, a) \in S \times A$, and this state-action pair gives reward r and transitions to a (random) next state s'. If the horizon H is finite, then the H-step action-value of (s, a) depends only on the $(H-1)$-step return from s'. Since our problem does not require us to explicitly estimate h-step returns for $h < H$, we make no independent assumption on the range of the h-step returns. We allow rewards obtained after visiting s' to be arbitrarily large or small (possibly negative), provided the sum of the first H rewards following (s, a) is bounded in $[0, V_{\max}]$. This distinction between the ranges of H-step and $(H-1)$-step rewards becomes inconsequential if H is infinite, and we anyway have to estimate action-values at all states. Mainly focused on stationary, infinite-horizon MDPs, the previous literature (Azar et al., 2013; Agarwal et al., 2020; Li et al., 2023) constructs concentration bounds by expressing the variance of the return from (s, a) in terms of the variance of the return from s'. We do not employ such a step. Rather, like in the analysis of TTM, we only apply Hoeffding's inequality to H-step returns.

3.2 Trajectory Tree Method

In *decision-time planning* (Kearns et al., 2002), the aim is to identify, a near-optimal action to take from the agent's current state s^0, with a given probability. A *trajectory tree* (Kearns et al., 1999) is a randomly-grown tree whose nodes correspond to states, starting with s^0 at the root. From any node s^t, $t \in [H]$, exactly one sample $s' \sim T(s^t, a, t)$ is drawn for each possible action $a \in A$, giving rise to a child node s'. This process results in a tree of size $|A|^H$ (but independent of $|S|$), as illustrated in Figure 1. Each transition has an associated reward. In POMDPs, each node additionally stores a randomly generated *observation*.

The rationale for building such a tree is that it can provide an unbiased estimate of the value of any arbitrary policy π (possibly history-dependent), starting from s^0. Observe that applying the policy takes us through a trajectory (fixed actions, random next states) with the same probability as in the true MDP or POMDP. Hence, $V^\pi(s)$ can be estimated by growing some m independent trajectory trees rooted at s^0, and averaging their value estimates. Crucially, the same m trees can be used to evaluate every policy π from the policy class Π being considered (which can be arbitrary). If Π is finite, then setting $m = O\left(\frac{V_{\max}}{\epsilon^2} \log \frac{|\Pi|}{\delta}\right)$ and selecting the *empirically-best* policy guarantees ϵ-optimality of the chosen action with probability at least $1-\delta$. This is because, by Hoeffding's inequality, each policy is estimated $\Theta(\epsilon)$-accurately with probability at least $1 - \frac{\delta}{|\Pi|}$ (Kakade, 2003, see Chapter 6).

TTM essentially arises from a view of any MDP as a distribution over deterministic MDPs (each represented as a trajectory tree from the current state). This same view also facilitates variance reduction in policy search (Ng and Jordan, 2000). To obtain bounds independent of $|S|$, Kearns et al. (1999) *branch* from every action sequence. On the other hand, to analyse CEM, we are happy with bounds that depend on $|S|$. Correspondingly, we represent each deterministic MDP as a collection of samples, one for each (state, action, time-step) triple. We call such a collection a "world".

4 Non-Stationary MDPs

We present our main ideas for the more general setting of non-stationary MDPs. First we summarise CEM in this setting.

4.1 Certainty-Equivalence: CEM-NS Algorithm

If the underlying MDP $M = (S, A, T, R, H, \gamma)$ is known to be non-stationary, then so is its maximum likelihood estimate $\widehat{M}$. It is sufficient for our purposes to assume that D contains the same number of samples, $N \geq 1$, for each tuple $(s, a, t) \in S \times A \times [H]$. Let count$(s, a, t, s')$ denote the number of observed transitions of (s, a, t) to $s' \in S$. The empirical transition function $\widehat{T}$ is set to

$$\widehat{T}(s, a, t, s') = \frac{\text{count}(s, a, t, s')}{N}.$$

$\widehat{M} = (S, A, \widehat{T}, R, H, \gamma)$ is a maximum likelihood estimate of M based on D. Let $V^\star_{\widehat{M}} : S \times [H] \to \mathbb{R}$ denote the optimal value function of $\widehat{M}$, and let $\widehat{\pi} : S \times [H] \to A$ be a corresponding optimal policy. $V^\star_{\widehat{M}}$ and $\widehat{\pi}$ are easily computed by dynamic programming. For $(s, t) \in S \times [H]$,

$$V^\star_{\widehat{M}}(s, t) = \max_{a \in A} \Big(R(s, a, t) + \gamma \sum_{s' \in S} \widehat{T}(s, a, t, s') V^\star_{\widehat{M}}(s', t + 1) \Big); \tag{1}$$

$$\widehat{\pi}(s, t) \in \operatorname*{argmax}_{a \in A} \Big(R(s, a, t) + \gamma \sum_{s' \in S} \widehat{T}(s, a, t, s') V^\star_{\widehat{M}}(s', t + 1) \Big). \tag{2}$$

We denote by CEM-NS ("NS" for "non-stationary") the algorithm that computes $\widehat{\pi}$ as its answer.

4.2 Set of Worlds

The unknowns in M are the transition probabilities for each $(s, a, t) \in S \times A \times [H]$. Hence the *minimum* amount of information required to build a complete estimate of M is exactly one transition for each (s, a, t) tuple. In our notation, the resulting estimate would be $\widehat{M}$ with $N = 1$—a deterministic MDP that is a "sample"

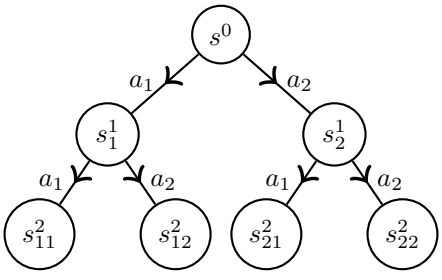

Figure 1: Example of trajectory tree for horizon $H = 2$, with starting state s^0, and actions a_1, a_2. Rewards are not shown.

of M. This estimate would allow the agent to evaluate any arbitrary behaviour, albeit with significant error. The conventional view is that as more transitions are observed, they make the point estimate $\widehat{M}$ more accurate. In our complementary view, larger N simply means more samples of M, each sample still an atomic (deterministic) MDP.

Recall that D contains N transitions for each $(s, a, t) \in S \times A \times [H]$. Take $[N] \stackrel{\text{def}}{=} \{1, 2, \ldots, N\}$, so each collected transition for (s, a, t) is indexed by some number $i \in [N]$. Let $x \in X \stackrel{\text{def}}{=} [N]^{|S||A|H}$ be a string of length $|S||A|H$ on the alphabet $[N]$. We view x as a code specifying a process to construct a deterministic MDP. The input to the process is the random data D; hence the resulting MDP M_x is a random variable. Concretely, x picks out a particular transition from the N collected in D for each (s, a, t) tuple. If $x(s, a, t) = i \in [N]$ for some $(s, a, t) \in S \times A \times [H]$, then the transition function T_x of M_x puts the entire transition probability from (s, a, t) on the state $s' \in S$ observed in the i-th sample of (s, a, t).

We refer to each $x \in X$ as a "world", defined by the code described above, and specifying a random deterministic MDP $M_x = (S, A, T_x, R, H, \gamma)$. Thus X is the "set of all worlds", of size $N^{|S||A|H}$. For any *fixed* D, the collection of $N^{|S||A|H}$ induced MDPs would generally be a multi-set, since multiple worlds $x \in X$ can induce the same MDP. Example 1 illustrates the definition of X and the process of sampling MDPs from D. A world is the semantic counterpart of a trajectory tree, since it allows for any policy to be evaluated. The syntactic difference is that a world associates a sample with every $(s, a, t) \in S \times A \times [H]$, whereas a trajectory tree associates a sample with each (state, action, state, action, ...) sequence visited while constructing the tree.

.

Example 1. *Consider MDP M with states $S = \{s_0, s_1\}$, actions $A = \{a_0, a_1\}$, and horizon $H = 3$. The table below describes a possible configuration of data D resulting from sampling each (state, action, time-step) tuple $N = 3$ times.*

s		s_0	s_0	s_0	s_0	s_0	s_0	s_1	s_1	s_1	s_1	s_1	s_1
a		a_0	a_0	a_0	a_1	a_1	a_1	a_0	a_0	a_0	a_1	a_1	a_1
t		0	1	2	0	1	2	0	1	2	0	1	2
Samples of s'	$i = 1$	s_1	s_1	s_1	s_1	s_1	s_1	s_0	s_1	s_0	s_1	s_0	s_0
	$i = 2$	s_0	s_0	s_1	s_0	s_1	s_1	s_1	s_1	s_1	s_0	s_0	s_1
	$i = 3$	s_1	s_0	s_0	s_1	s_1	s_1	s_0	s_1	s_1	s_1	s_0	s_1

Each sample $i \in [N]$ contains the next state. Each world is specified by a 12-length string over the alphabet $\{1, 2, 3\}$. If we interpret this string in the sequence of the columns in the table, the world $x = 132121123211$ induces MDP M_x with transition probabilities $T_x(s_0, a_0, 0, s_1) = 1$, $T_x(s_0, a_0, 1, s_0) = 1$, $T_x(s_0, a_0, 2, s_1) = 1$, and so on. Notice that $x' = 122121123211$, which differs from x only in its second position, would induce the same MDP since the second and third samples of $(s_0, a_0, 1)$ both lead to s_0. The total number of worlds is $3^{|S||A|H} = 531441$; for D in our example the number of unique MDPs induced is $2^8 = 256$, since only 8 of the 12 (s, a, t) triples have samples with both possible next states.

4.3 Evaluating Policies on the Set of Worlds

The value of policy $\pi : S \times [H] \to A$ on MDP M_x corresponding to world $x \in X$ is given by

$$V_x^\pi(s, t) = R(s, \pi(s, t), t) + \gamma \sum_{s'} T_x(s, \pi(s, t), t, s') V_x^\pi(s', t + 1) \tag{3}$$

for $(s, t) \in S \times [H]$. We note V_x^π to be an unbiased estimator of V^π.

Lemma 2 (Worlds provide unbiased estimates)**.** *For $x \in X$, $\pi : S \times [H] \to A$, and $(s, t) \in S \times [H]$:*

$$\mathbb{E}[V_x^\pi(s, t)] = V^\pi(s, t).$$

Proof. Fix $x \in X$ and $\pi : S \times [H] \to A$. As base case of an inductive argument, note that for $s \in S$, $\mathbb{E}[V_x^\pi(s, H)] \stackrel{\text{def}}{=} \mathbb{E}[0] = 0 = V^\pi(s, H)$. Assume that for some $t \in [H]$, for $s \in S$, $\mathbb{E}[V_x^\pi(s, t + 1)] = V^\pi(s, t + 1)$.

Now, in (3), $T_x(s, \pi(s,t), t, s')$ is the outcome of a sample for time step t, but the samples for computing $V_x^\pi(s', t+1)$ are all from time steps $t+1$ and higher. Hence, random variables $T_x(s, \pi(s,t), t, s')$ and $V_x^\pi(s', t+1)$ are independent, implying that for $s \in S$,

$$\begin{aligned}\mathbb{E}[V_x^\pi(s,t)] &= \mathbb{E}[R(s,\pi(s,t),t)] + \\ &\qquad \gamma \sum_{s'} \mathbb{E}\Big[T_x(s,\pi(s,t),t,s')\Big]\mathbb{E}\Big[V_x^\pi(s',t+1)\Big] \\ &= R(s,\pi(s,t),t) + \gamma \sum_{s'} T(s,\pi(s,t),t,s') V^\pi(s',t+1),\end{aligned}$$

since (1) $T_x(s, \pi(s,t), t, s')$ is 1 with probability $T(s, \pi(s,t), t, s')$, and otherwise 0; and (2) from the induction hypothesis, $\mathbb{E}[V_x^\pi(s, t+1)] = V^\pi(s, t+1)$. The RHS is the same as in the Bellman equation on M for π; hence $\mathbb{E}[V_x^\pi(s,t)] = V^\pi(s,t)$. □

Our upcoming analysis will depend on generalising value functions to *sets* of worlds. We define the value function of a set as the average over its members.

Definition 3. *For $Z \subseteq X$, $\pi : S \times [H] \to A$, $(s,t) \in S \times [H]$,*

$$V_Z^\pi(s,t) \stackrel{def}{=} \frac{1}{|Z|} \sum_{z \in Z} V_z^\pi(s,t).$$

At this point, we can already conceive a recipe based on the classical TTM method to construct a near-optimal policy for M. To implement the idea of Kearns et al. (1999), consider the N-sized subset of worlds $X' \subseteq X$, given by $X' = \left\{1^{|S||A|H}, 2^{|S||A|H}, \ldots, N^{|S||A|H}\right\}$. By design, no two worlds in X' share any samples; hence they can provide N independent value function estimates for each policy. From Hoeffding's Inequality (Hoeffding, 1963), the value function of each policy would be ϵ-optimal with probability $1 - \delta/|\Pi|$ for $N = O\left(\frac{(V_{\max})^2}{\epsilon^2} \log \frac{|\Pi|}{\delta}\right)$. Thus, an algorithm that returns an "optimal policy" for X' from a set of policies Π would meet our PAC criterion with about $O\left(\frac{|S||A|H(V_{\max})^2}{\epsilon^2} \log \frac{|\Pi|}{\delta}\right)$ samples (Kearns et al., 1999; Kakade, 2003). Unfortunately, it is not easy to *compute* an optimal policy for X' if the policy class Π is the (usual) set of Markovian, non-stationary policies. The structure of X' is such that in general, *history-dependent* policies can perform strictly better than Markovian policies. On the other hand, it is straightforward to compute an optimal Markovian, non-stationary policy for the entire universe of worlds X. In fact, as formalised in the following lemma, value functions of policies turn out to be *identical* on $\widehat{M}$ and on X. Therefore, the output of CEM-NS—$\widehat{\pi}$—is itself an optimal policy for X!

Lemma 4 (Consistency of X and $\widehat{M}$)**.** *For $\pi : S \times [H] \to A$ and $(s,t) \in S \times [H]$,*

$$V_X^\pi(s,t) = V_{\widehat{M}}^\pi(s,t).$$

The proof of this important lemma is given in Appendix A. The idea is to expand V_X^π and use the fact that each sample in D occurs in exactly the same number of worlds $x \in X$, whereupon it emerges that V_X^π satisfies the Bellman equations for π on $\widehat{M}$.

The crux of our paper is in the contrast between X' and X. Although V_x^π is an unbiased estimate of V^π for each $x \in X$ and $\pi : S \times [H] \to A$, the deviation of their *average* V_X^π from V^π cannot be bounded directly using Hoeffding's inequality, since V_x^π and $V_{x'}^\pi$ could be *dependent* for worlds $x, x' \in X$. For example, the worlds $1^{|S||A|H}$ and $12^{|S||A|H-1}$ use the same sample for $(s_0, a_0, 0)$. In spite of this dependence, can we still piggyback on the analytical framework of TTM? Our answer is affirmative, and forms the basis of our view of CEM as an application of TTM.

4.4 Batches of Mutually-Disjoint Worlds

We consider N-sized "batches" within X that do lead to independent samples of M. Define worlds $x, x' \in X$ to be *disjoint* if for all $(s, a, t) \in S \times A \times [H]$, $x(s, a, t) \neq x'(s, a, t)$. In other words, x and x' are disjoint if they do not share any samples. A *batch* $b \subseteq X$ is a set of some N mutually disjoint elements of X. The set $\{132212312132, 221323121321, 313131233213\}$ is a batch in Example 1, as also is set X' from Section 4.3. For $x \in X$, let B_x be the set of all batches in which x is present, and let B be the set of all batches. Simple counting (provided in Appendix B) shows that for $x \in X$, $|B_x| = (N-1)!^{|S||A|H-1}$, and $|B| = N!^{|S||A|H-1}$. Recall that V_X^π is the average value function of π over worlds $x \in X$. At the heart of our proof is the following equation, which shows V_X^π also as the average of the value functions of π over batches $b \in B$.

$$\begin{aligned} V_X^\pi(s,t) &= \frac{1}{|X|}\sum_{x\in X} V_x^\pi(s,t) = \frac{1}{|X|}\sum_{b\in B}\sum_{x\in b}\frac{1}{|B_x|}V_x^\pi(s,t) \\ &= \frac{N}{|X|(N-1)!^{|S||A|H-1}}\sum_{b\in B}\frac{\sum_{x\in b}V_x^\pi(s,t)}{N} \\ &= \frac{1}{|B|}\sum_{b\in B} V_b^\pi(s,t). \end{aligned} \tag{4}$$

The significance of (4) is that for *each* batch $b \in B$, V_b^π is indeed an average of N *independent* random variables, whose deviation from their expected value can be bounded using Hoeffding's inequality. Since V_X^π is a convex combination of $V_b^\pi, b \in B$, we can apply Hoeffding's (less-used) result on the sums of dependent random variables (Hoeffding, 1963, see Section 5). We restate Hoeffding's result as the following lemma. The commonly-used version of Hoeffding's inequality for independent random variables (Hoeffding, 1963, see Theorem 2) is obtained by taking $m = 1$.

Lemma 5. *[Hoeffding's inequality for average of certain dependent random variables] Fix positive integers ℓ and m. For $i \in \{1, 2, \ldots \ell\}$, $j \in \{1, 2, \ldots, m\}$, let $U_{i,j}$ be a real-valued random variable supported on $[\alpha, \beta] \subset \mathbb{R}$; suppose $U_{i,j}$ and $U_{i,j'}$ are independent for $j, j' \in \{1, 2, \ldots, m\}$ if $j \neq j'$. Note that $U_{i,j}$ and $U_{i',j'}$ could be dependent if $i \neq i'$, for $i, i' \in \{1, 2, \ldots, \ell\}$, and $j, j' \in \{1, 2, \ldots, m\}$. Define*

$$U_i \stackrel{def}{=} \frac{1}{m}\sum_{j=1}^{m} U_{i,j} \text{ and } U \stackrel{def}{=} \sum_{i=1}^{\ell} p_i U_i$$

for some $p_1, p_2, \ldots, p_\ell \in [0, 1]$ satisfying $\sum_{i=1}^{\ell} p_i = 1$. For $\gamma > 0$,

$$\mathbb{P}\{U \geq \mathbb{E}[U] + \gamma\} \leq \exp\left(\frac{-2m\gamma^2}{(\beta-\alpha)^2}\right) \text{ and}$$
$$\mathbb{P}\{U \leq \mathbb{E}[U] - \gamma\} \leq \exp\left(\frac{-2m\gamma^2}{(\beta-\alpha)^2}\right).$$

For convenient reference, we give a proof of this lemma in Appendix C (the original proof is from Hoeffding (1963, see Section 5)). We are ready for our main result, which uses Lemma 5 to legitimise CEM's approach of optimising behaviour uniformly over every possible batch, in contrast with TTM's approach of doing so for a single, arbitrary batch.

Theorem 6 (Sample complexity of CEM-NS)**.** *The CEM-NS algorithm provides the relevant PAC guarantee for non-stationary MDP M with parameters $\epsilon \in (0, V_{max})$, $\delta \in (0, 1)$ if run with*

$$N = \left\lceil \frac{2(V_{\max})^2}{\epsilon^2} \ln \frac{|S||A|^{|S|H}}{\delta} \right\rceil.$$

Proof. Recall that CEM-NS returns $\widehat{\pi}$, which depends on the data D, and hence is random. Lemma 4 gives us that $\widehat{\pi}$ is optimal for X. Now, if $\widehat{\pi}$ is not ϵ-optimal for M, it means that either (i) X under-estimates $V^{\pi^\star}(s, 0)$

by at least $\frac{\epsilon}{2}$, or (ii) X over-estimates $V^\pi(s,0)$ by at least $\frac{\epsilon}{2}$ for some non-ϵ-optimal policy $\pi : S \times [H] \to A$ and state $s \in S$. From (4), we have that $V_X^\pi(s,0) = \sum_{b \in B} \frac{1}{|B|} V_b^\pi(s,0)$, where $V_b^\pi(s,0)$ for each $b \in B$ is a sum on N independent random variables with mean $V^\pi(s,0)$ (from Lemma 2). Define $\delta' \stackrel{\text{def}}{=} \frac{\delta}{|S||A|^{|S|H}}$. We apply Lemma 5 to get $\mathbb{P}\left\{V_X^{\pi^\star}(s,0) \le V^{\pi^\star}(s,0) - \frac{\epsilon}{2}\right\} \le \delta'$ and $\mathbb{P}\left\{V_X^{\pi}(s,0) \ge V^{\pi}(s,0) + \frac{\epsilon}{2}\right\} \le \delta'$ for $s \in S$ and $\pi : S \times H \to A$. Since there are $|S|$ states and $|A|^{|S|H}$ policies, a union bound establishes that $\widehat{\pi}$ is ϵ-optimal with probability at least $|S||A|^{|S|H}\delta' = \delta$. □

Since each $(s,a,t) \in S \times A \times [H]$ is sampled N times by CEM-NS, and since $V_{\max} \le H$, the algorithm's overall sample complexity is

$$O\left(\frac{|S||A|H^3}{\epsilon^2}\left(\log\frac{1}{\delta} + |S|H\log|A|\right)\right).$$

Recall that Li et al. (2023) show a bound of $\tilde{O}\left(\frac{|S||A|H^4}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples for **CEM-NS**. In the regime that δ is made small after fixing other parameters, our bound is tighter by a factor of H. This is a significant result since the coefficient of $\log\frac{1}{\delta}$ now has a cubic dependence on the horizon—which we show is unavoidable by providing an explicit lower bound in Section 6.

5 Stationary MDPs

In this section, we analyse CEM when applied to *stationary* MDP M. We now use $\widehat{M}$, $\widehat{\pi}$, and X to denote corresponding objects in the stationary setting.

5.1 Certainty-Equivalence: CEM-S Algorithm

We continue with the same definition of $M = (S, A, T, R, H, \gamma)$, only now assuming that T and R do not depend on the time step t (which we drop from our notation). Consistent with previous literature, we also assume $H = \infty$. Since there is no time-dependence, we take that each tuple $(s,a) \in S \times A$ is sampled N times, $N \ge 1$, in the data D. For $(s,a,s') \in S \times A \times S$, let $\text{count}(s,a,s')$ denote the number of transitions observed in D to reach s' by taking a from s. The empirical MDP $\widehat{M} = (S, A, \widehat{T}, R, H, \gamma)$ therefore satisfies

$$\widehat{T}(s,a,s') = \frac{\text{count}(s,a,s')}{N}$$

for $s, s' \in S, a \in A$. It is well-known that every stationary, infinite-horizon MDP admits a deterministic optimal policy. The optimal value function $V^\star_{\widehat{M}}$ and optimal policy $\widehat{\pi} : S \to A$ satisfy

$$V^\star_{\widehat{M}}(s) = \max_{a \in A}\left(R(s,a) + \gamma\sum_{s' \in S}\widehat{T}(s,a,s')V^\star_{\widehat{M}}(s')\right);$$
$$\widehat{\pi}(s) \in \operatorname*{argmax}_{a \in A}\left(R(s,a) + \gamma\sum_{s' \in S}\widehat{T}(s,a,s')V^\star_{\widehat{M}}(s')\right)$$

for $s \in S$. $V^\star_{\widehat{M}}$ and $\widehat{\pi}$ may be computed from D by value iteration, policy iteration, or linear programming (Littman et al., 1995). Our upcoming sample-complexity bound would only get scaled by a constant factor, say, if an $\frac{\epsilon}{2}$-optimal policy is computed for $\widehat{M}$—and this computation needs only a polynomial number of arithmetic operations in $|S|$, $|A|$, $\frac{1}{1-\gamma}$, and $\log\frac{1}{\epsilon}$. To keep the exposition uncluttered, we assume that our certainty-equivalence implementation—denoted CEM-S ("S" for "stationary")—indeed computes and returns $\widehat{\pi}$ exactly.

5.2 Truncated Horizon

The assumption of a finite horizon H in the non-stationary setting meant that our worlds would also have this same horizon H. Since we have taken $H = \infty$ for stationary M, we require an intermediate step to apply the framework of a set of worlds. Consider a finite horizon MDP $M_{\overline{H}} = (S, A, T, R, \overline{H}, \gamma)$ that is identical to M other than for having a *finite* horizon $\overline{H} \stackrel{\text{def}}{=} \left\lceil \frac{1}{1-\gamma} \ln\left(\frac{4V_{\max}}{\epsilon}\right) \right\rceil$. The corresponding empirical MDP is $\widehat{M}_{\overline{H}} = (S, A, \widehat{T}, R, \overline{H}, \gamma)$. Since the infinite-discounted sum from each state is constrained to $[0, V_{\max}]$, the truncation loss $\mathbb{E}_\pi[\sum_{t=\overline{H}}^{\infty} \gamma^t r^t]$ must lie in $[0, \gamma^{\overline{H}} V_{\max}] \subseteq [0, \frac{\epsilon}{4}]$ on both M and $\widehat{M}$.

Proposition 7 (Bounded truncation loss)**.** *For $\pi : S \to A$, $s \in S$:*

$$V^\pi(s) - \frac{\epsilon}{4} \leq V^\pi_{M_{\overline{H}}}(s, 0) \leq V^\pi(s); V^\pi_{\widehat{M}}(s) - \frac{\epsilon}{4} \leq V^\pi_{\widehat{M}_{\overline{H}}}(s, 0) \leq V^\pi_{\widehat{M}}(s).$$

The truncated horizon $\overline{H}$ has no relevance to the CEM-S *algorithm* itself; the algorithm is based on the true (infinite) horizon of M. However, our *analysis* works on $M_{\overline{H}}$, using worlds of length $|S||A|\overline{H}$ to encode samples. Proposition 7 enables us to relate the extent of sub-optimality over a finite horizon $\overline{H}$ with that on an infinite horizon.

5.3 Set of Worlds

Each world x in our set of worlds X is an $|S||A|\overline{H}$-length string on the alphabet $[N]$. It associates a transition sample from D for each $(s, a, t) \in S \times A \times [\overline{H}]$. However, since $M_{\overline{H}}$ is stationary, samples are not distinguished based on time step in the data D. Hence, if D in Example 1 had come from a *stationary* MDP, we would ignore t and pool together all $N = 9$ samples for each (state, action) pair. Thus, for the pair (s_0, a_0), the sequence of samples (read row by row from top to bottom, and left to right within each row) would be $s_1, s_1, s_1, s_0, s_0, s_1, s_1, s_0, s_0$. The world $x = 571634978542$ would induce a deterministic MDP with probabilities of 1 for the twelve transitions $(s_0, a_0, 0, s_0)$, $(s_0, a_0, 1, s_1)$, $(s_0, a_0, 2, s_1)$, $(s_0, a_1, 0, s_1)$, $(s_0, a_1, 1, s_1)$, $(s_0, a_1, 2, s_0)$, $(s_1, a_0, 0, s_1)$, $(s_1, a_0, 1, s_0)$, $(s_1, a_0, 2, s_1)$, $(s_1, a_1, 0, s_0)$, $(s_1, a_1, 1, s_0)$, and $(s_1, a_1, 2, s_0)$. In general there are $N^{|S||A|\overline{H}}$ worlds $x \in X$.

In the stationary setting, it is seen that X evaluates policies identical to $\widehat{M}_{\overline{H}}$.

Lemma 8 (Consistency of X and $\widehat{M}_{\overline{H}}$)**.** *For $\pi : S \times [\overline{H}] \to A$ and $(s, t) \in S \times [\overline{H}]$,*

$$V^\pi_X(s, t) = V^\pi_{\widehat{M}_{\overline{H}}}(s, t).$$

The proof of Lemma 8 is identical to that of Lemma 4, and given in Appendix A. The lemma and upcoming results also apply to stationary policies (the "t" comes from the finite horizon of $\widehat{M}_{\overline{H}}$).

5.4 Biased and Unbiased Worlds

Recall that in the non-stationary setting, not all $x \in X$ were mutually disjoint—which means their induced MDPs had dependent transitions. We resolved this issue by partitioning X into N-sized *batches* of mutually-disjoint worlds. In the stationary setting, we could encounter an issue of dependence even *within* a single world $x \in X$. Consider the world $x = 4416823295$ from Example 1: this world is constrained to set both $T_x(s_0, a_0, 0, s_0)$ and $T_x(s_0, a_0, 1, s_0)$ based on the the *same* sample, namely the 4^{th} one collected for (s_0, a_0). Consequently, $T_x(s_0, a_0, 0, s_0)$ is *dependent* on $V^\pi_x(s_0, 1)$ for any policy π that takes a_0 from s_0 at time step 1. We can no longer claim $\mathbb{E}[V^\pi_x] = V^\pi_{M_{\overline{H}}}$ (like we did while analysing CEM-NS, in the proof of Lemma 2).

To proceed, we partition X into sets X_{biased} and X_{unbiased}. The set X_{biased} contains all worlds $x \in X$ for which there exist $(s, a, t, t') \in S \times A \times [\overline{H}] \times [\overline{H}]$, $t \neq t'$, such that $x(s, a, t) = x(s, a, t')$. Such worlds induce MDPs that provide possibly biased value estimates. The complementary set $X_{\text{unbiased}} \stackrel{\text{def}}{=} X \setminus X_{\text{biased}}$ contains worlds that do provide an unbiased estimate of the value function of each policy $\pi : S \times [\overline{H}] \to A$ on $M_{\overline{H}}$.

Lemma 9 (Worlds in X_{unbiased} provide unbiased estimates)**.** *For $x \in X_{\text{unbiased}}$, $\pi : S \times [\overline{H}] \to A$, $(s,t) \in S \times [\overline{H}]$,*

$$\mathbb{E}[V_x^\pi(s,t)] = V_{M_{\overline{H}}}^\pi(s,t).$$

The proof is identical to the one of Lemma 2, relying on the independence of random variables $T_x(s,\pi(s),t,s')$ and $V_x^\pi(s',t+1)$ for $x \in X_{\text{unbiased}}$.

Without any useful handle on worlds $x \in X_{\text{biased}}$, our strategy is to show that the size of X_{biased} as a fraction of $|X|$ vanishes with N, implying that $V_{X_{\text{biased}}}^\pi$ influences V_X^π only marginally when N is sufficiently large. The following lemma is proven in Appendix D.

Lemma 10 (Error from biased worlds vanishes with N)**.** *For $\pi : S \times [\overline{H}] \to A$, $(s,t) \in S \times [\overline{H}]$:*

$$\left|V_X^\pi(s,t) - V_{X_{\text{unbiased}}}^\pi(s,t)\right| \leq \frac{|S||A|\overline{H}(\overline{H}-1)V_{\max}}{N}.$$

Finally, just as we grouped $x \in X$ into mutually-disjoint batches in Section 4, we do the same for $x \in X_{\text{unbiased}}$ in the stationary setting. We do not consider worlds in X_{biased} for this grouping. Recall that worlds x and x' are disjoint if and only if $x(s,a,t) \neq x'(s,a,t)$ for all $(s,a,t) \in S \times A \times [\overline{H}]$. Assume for simplicity that N is a multiple of $\overline{H}$, and define $N' = N/\overline{H}$. Calculations provided in Appendix B show that (1) $|X_{\text{unbiased}}| = \frac{N!^{|S||A|}}{(N-\overline{H})!^{|S||A|}}$, (2) the set of all batches B (each batch containing N' mutually-disjoint worlds $x \in X_{\text{unbiased}}$) is of size $\frac{N!^{|S||A|}}{N'!}$, and (3) the set of all batches B_x that contain any particular world $x \in X_{\text{unbiased}}$ is of size $\frac{(N-\overline{H})!^{|S||A|}}{(N'-1)!}$. Substituting into a working similar to (4), we observe

$$V_{X_{\text{unbiased}}}^\pi(s,t) = \frac{1}{|B|}\sum_{b \in B} V_b^\pi(s,t) \tag{5}$$

for $\pi : S \times [\overline{H}] \to A$, $(s,t) \in S \times [\overline{H}]$, which facilitates the use of Lemma 5 on $V_{X_{\text{unbiased}}}^\pi$.

We have all the elements ready for an upper bound on the sample complexity of CEM-S.

Theorem 11 (Sample complexity of CEM-S)**.** *The CEM-S algorithm provides the relevant PAC guarantee for stationary MDP M with parameters $\epsilon \in (0, V_{max})$, $\delta \in (0,1)$ if run with*

$$N = \max\left(\left\lceil \frac{32(V_{\max})^2}{\epsilon^2} \ln \frac{|S||A|^{|S|}}{\delta} \right\rceil, \left\lceil \frac{8|S||A|(\overline{H}-1)V_{\max}}{\epsilon} \right\rceil\right)\overline{H}.$$

The proof (given in detail in Appendix E) follows the same core structure as of the non-stationary case in Theorem 6, but requires additional steps to account for the truncated horizon $\overline{H}$ and the partition of X into sets X_{biased} and X_{unbiased}. We infer that the sample complexity of CEM-S is

$$O\left(\frac{|S||A|}{(1-\gamma)^3\epsilon^2}\left(\log\frac{1}{(1-\gamma)\epsilon}\right)\left(\log\frac{1}{\delta} + |S||A|\epsilon\log\frac{1}{1-\gamma}\right)\right).$$

Unlike existing upper bounds (Azar et al., 2013; Agarwal et al., 2020) that only hold for restricted ranges of ϵ, this bound applies to the entire range of problem parameters. Observe that the coefficient of $\log(\frac{1}{\delta})$ is $\tilde{O}\left(\frac{|S||A|}{(1-\gamma)^3\epsilon^2}\right)$. Thus, we have the novel result that for arbitrary, fixed values of $|S|$, $|A|$, ϵ, and γ, CEM-S is optimal up to logarithmic factors as $\delta \to 0$. The notion of optimality in the limit as $\delta \to 0$ has also been applied in other PAC learning contexts (Garivier and Kaufmann, 2016).

6 Sample-Complexity Lower Bound For Finite-Horizon MDPs

In this section, we furnish a lower bound on the sample complexity required for any PAC algorithm on finite-horizon MDPs. This bound, shown by constructing a family of stationary MDPs, applies to both stationary

and to non-stationary MDPs. The basic structure is taken from Azar et al. (2013), who provide a similar lower bound for infinite-horizon MDPs with discounting. The main change required is to substitute terms depending on discount factor γ with terms depending on horizon H. We also note that the lower bound of Azar et al. (2013) applies only to a restricted class of algorithms that make "independent" predictions for "independent" state-action pairs (Azar et al., 2013, see Lemma 18). To obtain a more general result, we also borrow from the proof structure used by Mannor and Tsitsiklis (2004) for best-arm identification in stochastic multi-armed bandits.

At each step, algorithm $\mathcal{L}$ has a history of (state, action, time step) triples that have already been sampled, along with their observed outcomes (next states). The algorithm must either (1) stop and publish $\hat{Q}(s,a,0)$ for each $(s,a) \in S \times A$, or (2) specify a probability distribution over all $(s,a,t) \in S \times A \times [H]$. If the latter, an (s,a,t) triple is sampled, its outcome recorded, and the process moves to the next step. For $\epsilon > 0$, the output of $\mathcal{L}$ is ϵ-correct if for all $(s,a) \in S \times A$, $|Q^\star(s,a,0) - \hat{Q}(s,a,0)| \leq \epsilon$. In turn, for $\delta > 0$, $\mathcal{L}$ is an (ϵ,δ)-PAC algorithm if on each input MDP, the probability that $\mathcal{L}$ stops and returns an ϵ-correct output is at least $1-\delta$.

Observe that we provide a lower-bound for accurately estimating Q-values; this is mainly for being consistent with Azar et al. (2013). It is easily shown that the same lower bound holds (up to a constant factor) for estimating an ϵ-optimal policy with probability $1-\delta$. In our working, we use $Q^\star(s,a)$ and $\hat{Q}(s,a)$ to denote $Q^\star(s,a,0)$ and $\hat{Q}(s,a,0)$, respectively. Below is our formal statement.

Theorem 12. *Fix set of states S, set of actions A arbitrarily, and let horizon $H > 200$. There exist an MDP $(S,A,T,R,H,\gamma=1)$, constants $c_1 > 0, c_2 > 0$ such that for all $\epsilon \in (0,1)$, $\delta \in (0,0.5)$, any $(\epsilon,\delta) - PAC$ algorithm $\mathcal{L}$ has an expected sample complexity of at least*

$$\frac{c_1 |S||A| H^3}{\epsilon^2} \ln\left(\frac{c_2}{\delta}\right)$$

on this MDP.

The full proof of this theorem is provided in Appendix F. The proof relies on a purposefully-designed family of MDPs constructed by Azar et al. (2013). These authors consider two specific MDPs, M_0 and M_1, whose Q-functions are more than 2ϵ apart, and hence any estimate cannot simultaneously be ϵ-correct for both MDPs. On the other hand, since the MDPs are sufficiently close, an agent cannot distinguish them based on observed samples alone unless the sample size is sufficiently large. Consequently, unless a sufficient number of samples are observed, an algorithm must necessarily be non-(ϵ,δ)-PAC.

Our proof follows this same structure, except (1) we use a finite horizon H for the MDP family, and (2) we consider $\frac{k}{3}+1$ MDPs where k is the number of state-action pairs. The latter adaptation lets us extend the lower bound to arbitrary algorithms. In our proof, there is a base MDP M_0 and for each state-action pair i that represents a choice, an MDP M_i that differs from M_0 only at this state-action pair. We lower-bound the ratio of the likelihood of observing the same data from these MDPs in terms of the number of samples. If the sample complexity is "small", the likelihood ratio bound allows us to show that if a certain event has a high probability under MDP M_0, it also has a high probability under M_i. By choosing this event to be the event that $|\hat{Q} - Q_0^*| < \epsilon$, we argue that if an algorithm gives an ϵ-correct answer with probability at least $1-\delta$ on M_0, it gives this same (but now *not* ϵ-correct) answer on M_i with probability at least $1-\delta$. Hence an algorithm cannot be PAC unless the sample complexity is "large"—as quantified in Theorem 12.

7 Conclusion

In this paper, we bring to light a surprising connection between the well known certainty-equivalence method (CEM) for PAC RL, and the trajectory tree method (TTM) for decision-time planning. We show that CEM implicitly computes a policy that simultaneously optimises over all possible "batches" of worlds, whereas TTM explicitly sets up a single batch of trajectory trees (functionally akin to worlds) to compute its policy. Noticing this connection, we establish upper bounds for CEM using only Hoeffding's inequality, yet which improve upon current bounds in the regime of small δ. Our results are especially significant in the finite-

horizon (non-stationary) setting, where in spite of making a weaker assumption on the rewards, we show the minimax-optimality of CEM in the small-δ regime.

Our new perspective sets up several possible directions for future work, including (1) the derivation of instance-specific upper bounds for sequential PAC RL algorithms, and (2) generalising the idea to formalisms that use function approximation. Finally, (3) it would be worth investigating the applicability of our analytical framework to related on-line learning problems, such as exploration in continuing tasks without "reset" access (Brafman and Tennenholtz, 2003; Strehl and Littman, 2008), the episodic off-policy setting (Yin et al., 2021), and regret minimisation (Auer and Ortner, 2006; Dann et al., 2017).

References

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal. In *Proceedings of 33rd Conference on Learning Theory*, pages 67–83. PMLR.

Auer, P. and Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC Bounds on the Sample Complexity of Reinforcement Learning with a Generative Model. *Machine Learning*, 91:325–349.

Brafman, R. I. and Tennenholtz, M. (2003). R-Max - a General Polynomial Time Algorithm for near-Optimal Reinforcement Learning. *Journal of Machine Learning Research*, 3:213–231.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Fiechter, C.-N. (1994). Efficient Reinforcement Learning. In *Proceedings of the Seventh Annual Conference on Computational Learning Theory*, pages 88–97. ACM Press.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027. PMLR.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Jiang, N. and Agarwal, A. (2018). Open Problem: The Dependence of Sample Complexity Lower Bounds on Planning Horizon. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 3395–3398. PMLR.

Kakade, S. M. (2003). *On the Sample Complexity of Reinforcement Learning*. Phd thesis, University College London.

Kearns, M., Mansour, Y., and Ng, A. (2002). A Sparse Sampling Algorithm for Near-Optimal Planning in Large Markov Decision Processes. *Machine Learning*, 49:193–208.

Kearns, M., Mansour, Y., and Ng, A. Y. (1999). Approximate planning in large pomdps via reusable trajectories. In *Advances in Neural Information Processing Systems 12*, pages 1001–1007. MIT Press.

Kearns, M. and Singh, S. (1999). Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, page 996–1002. MIT Press.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. In *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. Curran Associates, Inc.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2023). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221.

Littman, M. L., Dean, T. L., and Kaelbling, L. P. (1995). On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 394–402. Morgan Kaufmann.

Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648.

Ng, A. Y. and Jordan, M. (2000). Pegasus: A policy search method for large mdps and pomdps. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, page 406–415. Morgan Kaufmann Publishers Inc.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018). Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model. In *Advances in Neural Information Processing Systems*, volume 31, page 5192–5202. Curran Associates, Inc.

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74:1309–1331.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1567–1575. PMLR.

A Proofs of Lemma 4 and Lemma 8

For proving Lemma 4, we have to show that for $\pi : S \times [H] \to A$, $(s,t) \in S \times [H]$: $V_X^\pi(s,t) = V_{\widehat{M}}^\pi(s,t)$. We use induction on t. For the base case of $t = H-1$, notice that $V_X^\pi(s,H-1) = R(s,\pi(s,H-1),H-1) = V_{\widehat{M}}^\pi(s,H-1)$. Suppose that the claim is true for $t+1$, where $t \in \{0,1,\ldots,H-2\}$. We have

$$\begin{aligned} V_X^\pi(s,t) &= \frac{1}{|X|}\sum_{x\in X} V_x^\pi(s,t) \\ &= \frac{1}{|X|}\sum_{x\in X}\left(R(s,\pi(s,t),t) + \gamma \sum_{s'\in S} T_x(s,\pi(s,t),t,s')V_x^\pi(s',t+1)\right). \end{aligned}$$

Recall that x is a string of length $|S||A|H$ over $[N]$, organised as H segments, with each $|S||A|$-length segment specifying the index of the sample from $[N]$ for each state-action pair. Let

- x_- be x's prefix of length $|S||A|t$ (empty if $t=0$);
- $x_\circ$ be the subsequent "middle portion" of length $|S||A|$, and
- x_+ be the suffix of length $|S||A|(H-t-1)$.

In other words, $x = x_- x_\circ x_+$ for some $x_- \in X_- \stackrel{\text{def}}{=} [N]^{|S||A|t}$, $x_\circ \in X_\circ \stackrel{\text{def}}{=} [N]^{|S||A|}$, $x_+ \in X_+ \stackrel{\text{def}}{=} [N]^{|S||A|(H-t-1)}$. We continue based on this decomposition:

$$V_X^\pi(s,t) = R(s,\pi(s,t),t) + \frac{\gamma}{|X|}\sum_{s'\in S}\sum_{x_-\in X_-}\sum_{x_\circ\in X_\circ}\sum_{x_+\in X_+} T_{x_-x_\circ x_+}(s,\pi(s,t),t,s')V^\pi_{x_-x_\circ x_+}(s',t+1).$$

Now, $T_{x_-x_\circ x_+}(s,\pi(s,t),t,s')$ does not depend on x_- or x_+, so we can simply denote it $T_{x_\circ}(s,\pi(s,t),t,s')$. Similarly, $V^\pi_{x_-x_\circ x_+}(s',t+1)$ does not depend on x_- or $x_\circ$, and so we may denote it $V^\pi_{x_+}(s',t+1)$. With this observation, we see that

$$\begin{aligned} V_X^\pi(s,t) = R(s,\pi(s,t),t)+ \\ \frac{\gamma}{|X|}\sum_{s'\in S}|X_-|\left(\sum_{x_\circ\in X_\circ} T_{x_\circ}(s,\pi(s,t),t,s')\right)\left(\sum_{x_+\in X_+} V^\pi_{x_+}(s',t+1)\right). \end{aligned}$$

The term $\sum_{x_\circ\in X_\circ} T_{x_\circ}(s,\pi(s,t),t,s')$ is seen to be

$$N^{|S||A|-1}\text{count}(s,\pi(s,t),t,s') = N^{|S||A|}\widehat{T}(s,\pi(s,t),t,s');$$

the $N^{|S||A|-1}$ factor comes from other state-action pairs for time step t taking every possible index from $[N]$ in the set of worlds. Since $V^\pi_{x_+}(s',t+1)$ does not depend on x_- and $x_\circ$, but x_- and $x_\circ$ take all possible values for each $x_+ \in X_+$, we substitute $\sum_{x_+\in X_+} V^\pi_{x_+}(s',t+1)$ with

$$\frac{|X_+|}{|X|}\sum_{x'\in X} V^\pi_{x'}(s',t+1) = |X_+|V_X^\pi(s',t+1) = |X_+|V^\pi_{\widehat{M}}(s',t+1),$$

where the last step applies the induction hypothesis. In aggregate, we now have

$$\begin{aligned} V_X^\pi(s,t) &= R(s,\pi(s,t),t) + \frac{\gamma}{|X|}\sum_{s'\in S}|X_-|N^{|S||A|}\widehat{T}(s,\pi(s,t),t,s')|X_+|V^\pi_{\widehat{M}}(s',t+1) \\ &= R(s,\pi(s,t),t) + \gamma\sum_{s'\in S}\widehat{T}(s,\pi(s,t),t,s')V^\pi_{\widehat{M}}(s',t+1) = V^\pi_{\widehat{M}}(s,t), \end{aligned}$$

which completes the proof of Lemma 4.

The proof of Lemma 8 is identical: we only have to ignore the time-independence of R and T, and take $\overline{H}$ as horizon instead of H.

B Counting with Batches

We compute the sizes of the set of all batches B, and the set of all batches containing some fixed world x. For the non-stationary setting, the set of worlds relevant to this exercise is the universe X, whereas for the stationary setting, counting is only done on X_{unbiased}.

B.1 Non-Stationary Setting

A batch contains N worlds such that for any pair of distinct worlds x, x' in X, there is no $(s, a, t) \in S \times A \times [H]$ such that $x(s, a, t) = x'(s, a, t)$. To count the total number of such batches possible, it is helpful to visualise a table with N rows and $|S||A|H$ columns. In how many ways can we fill up the cells with numbers from $[N]$ such that no two rows share a common element in any column?

	1	2	3	...	$\|S\|\|A\|H$
1					
2					
3					
$\vdots$					
N					

We proceed row by row. The first row can be filled up in $N^{|S||A|H}$ possible ways. With the first row filled, the second row can be filled in $(N-1)^{|S||A|H}$ possible ways. We proceed in this manner, until entries for the last row are fixed by those from the preceding $N-1$ rows. Hence, the number of possible tables we could have created is

$$N^{|S||A|H} \times (N-1)^{|S||A|H} \times (N-2)^{|S||A|H} \times 1^{|S||A|H} = N!^{|S||A|H}.$$

A batch is a *set* rather than a sequence—so any two tables that contain the same contents in each row, even if the rows are permuted, fall in the same equivalence class of a batch. Since there are $N!$ possible ways to permute the rows of the table, we observe that the number of unique batches is $|B| = \frac{N!^{|S||A|H}}{N!} = N!^{|S||A|H-1}$.

If a particular world (here row) x is fixed, the same reasoning implies the number of batches containing x is $|B_x| = (N-1)!^{|S||A|H-1}$.

B.2 Stationary Setting

Recall that world $x \in X$ is *biased* if there exist $(s, a, t, t') \in S \times A \times [\overline{H}] \times [\overline{H}]$, $t \neq t'$, such that $x(s, a, t) = x(s, a, t')$. Hence, in an unbiased world $x \in X_{\text{unbiased}}$, for each $(s, a) \in S \times A$, we must pick distinct samples for each $t \in [\overline{H}]$. The number of ways to select an $\overline{H}$-sized permutation from $[N]$ is $N(N-1)(N-2)\dots(N-\overline{H}+1) = \frac{N!}{(N-\overline{H})!}$.Since we gather such a permutation for each $(s, a) \in S \times A$, the size of X_{unbiased} is $\left(\frac{N!}{(N-\overline{H})!}\right)^{|S||A|}$.

In the stationary setting, we consider batches of size $N' = N/\overline{H}$, which are mutually-disjoint sets of worlds from X_{unbiased}. In a world $x \in X_{\text{unbiased}}$, no index from $[N]$ can repeat within the sub-table for each state-action pair. Other than for that constraint, we can calculate the size of B as before.

To fill up the first row of the table, we must pick some $\overline{H}$ indices from $[N]$ for each state-action pair. The number of ways we can do this is

$$\left(N(N-1)(N-2)\dots(N-\overline{H}+1)\right)^{|S||A|}.$$

None of the $\overline{H}$ indices used for any state-action pair in the first row can be used subsequently. Correspondingly, the number of ways to fill up the second row becomes

$$\left((N-\overline{H})(N-\overline{H}-1)(N-\overline{H}-2)\dots(N-2\overline{H}+1)\right)^{|S||A|}.$$

Proceeding similarly, the number of ways to fill up the last (N'-th) row is

$$\big((N - k\overline{H})(N - k\overline{H} - 1)(N - k\overline{H} - 2)\dots(N - (k+1)\overline{H} + 1)\big)^{|S||A|}$$

for $k = N' - 1$. Taking a product over rows, we observe that the total number of ways to fill it up is $N!^{|S||A|}$. Once again, we must account for permutations of the rows, which are N' in number. Hence, the number of batches $|B| = \frac{N!^{|S||A|}}{N'!}$.

If we fix a particular row with world x, it leaves $N' - 1$ rows to fill out, but only $N - \overline{H}$ indices available for each state-action pair. Repeating the same counting argument, we get $|B_x| = \frac{(N-\overline{H})!^{|S||A|}}{(N'-1)!}$.

C Proof of Lemma 5

The lemma and the proof are from Hoeffding (1963, see Section 5). The proof is rephrased here for the reader's convenience. For arbitrary $h > 0$, we have

$$\begin{aligned}\mathbb{P}\{U \geq \mathbb{E}[U] + \gamma\} &= \mathbb{P}\{h(U - \mathbb{E}[U]) \geq h\gamma\} \\ &= \mathbb{P}\{\exp(h(U - \mathbb{E}[U])) \geq \exp(h\gamma))\}. \end{aligned} \tag{6}$$

Applying Markov's Inequality to the non-negative random variable $\exp(h(U - \mathbb{E}[U]))$, we observe

$$\begin{aligned}\mathbb{P}\{\exp(h(U - \mathbb{E}[U])) \geq \exp(h\gamma))\} &\leq \exp(-h\gamma)\mathbb{E}[\exp(h(U - \mathbb{E}[U]))] \\ &= \exp(-h\gamma)\mathbb{E}\left[\exp\left(h\sum_{i=1}^{\ell} p_i(U_i - \mathbb{E}[U_i])\right)\right]. \end{aligned} \tag{7}$$

In turn, since the exponential function is convex, by Jensen's inequality,

$$\exp\left(h\sum_{i=1}^{\ell} p_i(U_i - \mathbb{E}[U_i])\right) \leq \sum_{i=1}^{\ell} p_i \exp(h(U_i - \mathbb{E}[U_i])). \tag{8}$$

Combining (6), (7), and (8), we have

$$\mathbb{P}\{U \geq \mathbb{E}[U] + \gamma\} \leq \exp(-h\gamma)\sum_{i=1}^{\ell} p_i\mathbb{E}[\exp(h(U_i - \mathbb{E}[U_i]))]. \tag{9}$$

The application of Jensen's inequality to obtain (8) above is not needed for the common application of Hoeffding's Inequality to averages of *independent* random variables, which is the case for us if $\ell = 1$. For the general case, (8) enables us to upper-bound the deviation probability by a convex combination of expectations, one corresponding to each $i \in \{1, 2, \ldots, \ell\}$. Since each U_i is indeed an average of independent random variables, each expectation can be upper-bounded exactly as in the proof of the common variant (Hoeffding, 1963, see Theorem 2). For $i \in \{1, 2, \ldots, \ell\}$,

$$\begin{aligned}\mathbb{E}[\exp(h(U_i - \mathbb{E}[U_i]))] &= \mathbb{E}\left[\exp\left(\frac{h}{m}\sum_{j=1}^{m}(U_{i,j} - \mathbb{E}[U_{i,j}]\right)\right] \\ &= \mathbb{E}\left[\prod_{j=1}^{m}\exp\left(\frac{h}{m}(U_{i,j} - \mathbb{E}[U_{i,j}])\right)\right] \\ &= \prod_{j=1}^{m}\mathbb{E}\left[\exp\left(\frac{h}{m}(U_{i,j} - \mathbb{E}[U_{i,j}])\right)\right], \end{aligned} \tag{10}$$

where the last step follows from the independence of $U_{i,j}$ and $U_{i,j'}$ for $j, j' \in \{1, 2, \ldots, m\}$, $j \neq j'$. At this point, we apply a mathematical fact that is sometimes called Hoeffding's lemma (Hoeffding, 1963, see Section 4 for proof). Noting that $U_{i,j}$ for $i \in \{1, 2, \ldots, \ell\}, j \in \{1, 2, \ldots, m\}$ is bounded in $[\alpha, \beta]$, we have

$$\mathbb{E}\left[\exp\left(\frac{h}{m}(U_{i,j} - \mathbb{E}[U_{i,j}])\right)\right] \leq \exp\left(\frac{h^2(\beta - \alpha)^2}{8m^2}\right). \tag{11}$$

Combining (9), (10), and (11), we obtain

$$\begin{aligned}\mathbb{P}\{U \geq \mathbb{E}[U] + \gamma\} &\leq \exp(-h\gamma)\sum_{i=1}^{\ell} p_i \prod_{j=1}^{m}\exp\left(\frac{h^2(\beta - \alpha)^2}{8m^2}\right) \\ &= \sum_{i=1}^{\ell} p_i \exp\left(-h\gamma + \frac{h^2(\beta - \alpha)^2}{8m}\right). \end{aligned} \tag{12}$$

Substituting the particular choice of $h = \frac{4\gamma m}{(\beta-\alpha)^2}$ in (12), we conclude

$$\begin{aligned}\mathbb{P}\{U \geq \mathbb{E}[U] + \gamma\} &\leq \sum_{i=1}^{\ell} p_i \exp\left(\frac{-2m\gamma^2}{(\beta-\alpha)^2}\right) \\ &= \exp\left(\frac{-2m\gamma^2}{(\beta-\alpha)^2}\right). \end{aligned} \tag{13}$$

For $i \in \{1, 2, \ldots, \ell\}$ and $j, j' \in \{1, 2, \ldots, m\}$, $j \neq j'$, our precondition that $U_{i,j}$ and $U_{i,j'}$ are independent implies that $-U_{i,j}$ and $-U_{i,j'}$ are also independent. Also note that $-U_{i,j}$ is supported on $[-\beta, -\alpha]$. We can therefore present the same proof as above to obtain

$$\mathbb{P}\{-U \geq \mathbb{E}[-U] + \gamma\} \leq \exp\left(\frac{-2m\gamma^2}{(-\alpha+\beta)^2}\right),$$

or equivalently,

$$\mathbb{P}\{U \leq \mathbb{E}[U] - \gamma\} \leq \exp\left(\frac{-2m\gamma^2}{(\beta-\alpha)^2}\right). \tag{14}$$

(13) and (14) complete the proof of the lemma.

D Proof of Lemma 10

We need to show that for $\pi : S \times [\overline{H}] \to A$, $(s,t) \in S \times [\overline{H}]$: $|V_X^\pi(s,t) - V_{X_{\text{unbiased}}}^\pi(s,t)| \leq \frac{|S||A|\overline{H}(\overline{H}-1)V_{\max}}{N}$.

First we decompose $V_X^\pi(s,t)$ for $\pi \in S \times [\overline{H}] \to A$ and $(s,t) \in S \times [\overline{H}]$ as

$$V_X^\pi(s,t) = \frac{1}{|X|}\sum_{x \in X} V_x^\pi(s,t) = \frac{|X_{\text{unbiased}}|}{|X|} V_{X_{\text{unbiased}}}^\pi(s,t) + \frac{|X_{\text{biased}}|}{|X|} V_{X_{\text{biased}}}^\pi(s,t)$$
$$= V_{X_{\text{unbiased}}}^\pi(s,t) + \frac{|X_{\text{biased}}|}{|X|}\left(V_{X_{\text{biased}}}^\pi(s,t) - V_{X_{\text{unbiased}}}^\pi(s,t)\right),$$

which implies

$$|V_X^\pi(s,t) - V_{X_{\text{unbiased}}}^\pi(s,t)| = \frac{|X_{\text{biased}}|}{|X|}|V_{X_{\text{biased}}}^\pi(s,t) - V_{X_{\text{unbiased}}}^\pi(s,t)|.$$

$|V_{X_{\text{biased}}}^\pi(s,t) - V_{X_{\text{unbiased}}}^\pi(s,t)|$ is at most $V_{\max}$. The proof is completed by observing

$$\frac{|X_{\text{biased}}|}{|X|} = 1 - \frac{|X_{\text{unbiased}}|}{|X|} = 1 - \frac{(N(N-1)\ldots(N-\overline{H}+1))^{|S||A|}}{N^{|S||A|\overline{H}}}$$
$$\leq 1 - \left(1 - \frac{\overline{H}-1}{N}\right)^{|S||A|\overline{H}} \leq \frac{|S||A|\overline{H}(\overline{H}-1)}{N}.$$

E Proof of Theorem 11

CEM-S returns $\widehat{\pi} : S \to A$, which is an optimal policy for $\widehat{M}$. If $\widehat{\pi}$ is not ϵ-optimal (for M), then either (i) $\widehat{M}$ has under-estimated $\pi^\star$ by at least $\frac{\epsilon}{2}$ for some state $s \in S$, or (ii) $\widehat{M}$ has over-estimated some other policy $\pi : S \to A$ by at least $\frac{\epsilon}{2}$ for some state $s \in S$. We upper-bound the probability of these events, in turn.

Suppose for state $s \in S$, we have $V^{\pi^\star}_{\widehat{M}}(s) \leq V^\star(s) - \frac{\epsilon}{2}$. Proposition 7 then implies $V^{\pi^\star}_{\widehat{M}_{\overline{H}}}(s,0) \leq V^\star(s) - \frac{\epsilon}{2}$. Again from Proposition 7, $V^{\pi^\star}_{\widehat{M}_{\overline{H}}}(s,0) \leq V^\star_{M_{\overline{H}}}(s.0) - \frac{\epsilon}{4}$. Equivalently, from Lemma 8, we get $V^{\pi^\star}_{X}(s,0) \leq V^\star_{M_{\overline{H}}}(s,0) - \frac{\epsilon}{4}$. If we now apply Lemma 10 using the specified value of N, we conclude that $V^{\pi^\star}_{X_{\text{unbiased}}}(s,0) \leq V^\star_{M_{\overline{H}}}(s,0) - \frac{\epsilon}{8}$. We upper-bound the probability of this consequent using Lemma 5, after observing that $V^{\pi^\star}_{X_{\text{unbiased}}}(s,0)$ can be rewritten as a convex combination over batches (from (5)), and its expected value is $V^{\pi^\star}_{M_{\overline{H}}}(s,0)$ (from Lemma 9). Defining $\delta' \stackrel{\text{def}}{=} \frac{\delta}{|S||A|^{|S|}}$ and recalling that each batch has $N' = N/\overline{H}$ worlds, it follows from Lemma 5 that the probability of $\pi^\star$ being under-estimated by at least $\frac{\epsilon}{2}$ for state s is at most δ'.

Now fix state $s \in S$ and a policy $\pi : S \to A$ other than $\pi^\star$. By a symmetric working to the preceding one for $\pi^\star$, we obtain that the probability π is over-estimated by at least $\frac{\epsilon}{2}$ on s is again at most δ'. Since we have considered $|A|^{|S|}$ policies and $|S|$ states, a union bound restricts the mistake probability to $|S||A|^{|S|}\delta' = \delta$.

F Proof of Lower Bound from Section 6

In this section, we fill in the full proof of Theorem 12. First we describe a family of MDPs constructed to achieve this lower bound, and then follow with the proof. Both the construction and the proof are closely based on those from Azar et al. (2013).

F.1 Family of MDPs

Consider an MDP $M_{ab}(p, \alpha, H)$ defined as follows (see Figure 2).

1. K initial states $X = \{x_i : i \in [1, K]\}$.
2. L initial actions $A = \{a_j : j \in [1, L]\}$ from each initial state x_i. These actions are deterministic and yield a reward of 0. Taking action a_j from state x_i leads us to state y_{ij}.
3. KL secondary states $Y^1 = \{y_{ij} : i \in [1, K], j \in [1, L]\}$, each with only one action. This is the only set of states in the MDP with a non-deterministic action. With probability p_{ij} we stay at the same state, and with $1 - p_{ij}$ we go to a corresponding state in $Y^2 = \{y^2_{ij} : i \in [1, K], j \in [1, L]\}$. The reward of this action is always 1.
4. $$p_{ij} = \begin{cases} p + \alpha, & \text{if } i = a, j = b, \\ p, & \text{otherwise.} \end{cases}$$
5. KL "terminal" states Y^2, each with only one action—looping back to the state and having a reward of 0 (not shown).
6. A horizon of H, after which the episode ends.
7. No discounting of rewards (that is, $\gamma = 1$).

We also define another MDP $M_0(p, \alpha, H)$ where $p_{ij} = p$ for all i, j, and then define the set of MDPs

$$\mathcal{M}(p, \alpha, H) = \{M_0(p, \alpha, H)\} \cup \{M_{ab}(p, \alpha, H) : a \in [1, K], b \in [1, L]\}.$$

We use $i \in [1, KL]$ interchangeably with $\{i, j\} : i \in [1, K], j \in [1, L]$. Note that the total number of state-action pairs is $k \stackrel{\text{def}}{=} 3KL$.

We establish some properties of our family of MDPs before proceeding to the proof of Theorem 12.

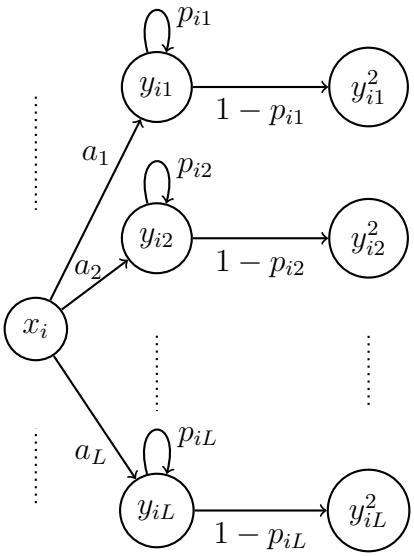

Figure 2: The MDP setup used to derive the lower bound.

Lemma 13. *The optimal value function for MDPs in the class $\mathcal{M}(p, \alpha, H)$ is given by:*

$$V_i^*(y_j, 0) = \begin{cases} \dfrac{1-(p+\alpha)^H}{1-(p+\alpha)}, & \text{if } i = j, \\ \dfrac{1-p^H}{1-p}, & \text{otherwise.} \end{cases} \tag{15}$$

This fact is easily verified. Note that the definition also holds for M_0 because j is never equal to 0.

Lemma 14. *The optimal value functions for the MDPs $M_i(p = 1 - \frac{1}{H}, \alpha = \frac{40\epsilon}{H^2}, H)$ and $M_0(p = 1 - \frac{1}{H}, \alpha = \frac{40\epsilon}{H^2}, H)$ (with $\epsilon < 1, H > 200$) are far apart: for $i \in [1, KL]$,*

$$Q_i^*(y_i) - Q_0^*(y_i) > 2\epsilon.$$

Proof.

$$\begin{aligned} (1-p)\left((p+\alpha)^H - p^H\right) &= \frac{1}{H}\left((p+\alpha)^H - p^H\right) \\ &= \frac{1}{H}\left((p+\alpha) - p\right)\left((p+\alpha)^{H-1} + (p+\alpha)^{H-2}p + \ldots + p^{H-1}\right) \\ &\leq \frac{\alpha}{H} H (p+\alpha)^{H-1} = \alpha(p+\alpha)^{H-1}. \end{aligned} \tag{16}$$

Moreover, it can be shown that $\left(1 - \frac{1}{n} + \frac{40}{n^2}\right)^{n-1}$ is a series that decreases monotonically for $n > 5$ and converges to $\frac{1}{e}$. Evaluating for $n = 200$ gives us that $\left(1 - \frac{1}{H} + \frac{40}{H^2}\right)^{H-1} < \frac{1.25}{e}$ $\forall H > 200$. Therefore,

$$\begin{aligned} p^H < p^{H-1} &< (p+\alpha)^{H-1} \\ &= \left(1 - \frac{1}{H} + \frac{40\epsilon}{H^2}\right)^{H-1} \\ &< \left(1 - \frac{1}{H} + \frac{40}{H^2}\right)^{H-1} < \frac{1.25}{e}. \end{aligned}$$

Now,

$$\begin{aligned}
Q_i^*(y_i) - Q_0^*(y_i) &= \frac{1-(p+\alpha)^H}{(1-(p+\alpha))} - \frac{1-p^H}{1-p} \\
&= \frac{(1-p)-(1-p)(p+\alpha)^H-(1-(p+\alpha))+(1-(p+\alpha))p^H}{(1-p)(1-(p+\alpha))} \\
&= \frac{\alpha-(1-p)(p+\alpha)^H+(1-(p+\alpha))p^H}{(1-p)(1-(p+\alpha))} \\
&= \frac{\alpha-(1-p)(p+\alpha)^H+(1-p)p^H-\alpha p^H}{(1-p)(1-(p+\alpha))} \\
&= \frac{\alpha-(1-p)((p+\alpha)^H-p^H)-\alpha p^H}{(1-p)(1-(p+\alpha))} \\
&\geq \frac{\alpha-\alpha(p+\alpha)^{H-1}-\alpha p^H}{(1-p)(1-(p+\alpha))} \\
&> \frac{\alpha(1-\frac{2.5}{e})}{(1-p)(1-(p+\alpha))} \\
&> \frac{\alpha\frac{1}{20}}{(1-p)(1-(p+\alpha))} \\
&> \frac{\frac{\alpha}{20}}{(1-p)^2} \\
&= 2\epsilon.
\end{aligned}$$

□

F.2 Proof of Theorem 12 (Lower Bound)

Fix $H > 200$, $\epsilon < 1$, $\frac{1}{2} < p = 1 - \frac{1}{H} < 1$, $0 < \alpha = \frac{40\epsilon}{H^2} < (1-p)/2$, $c_1' = 20$, $c_2' = 6$. Let random variable τ denote the sample complexity of our algorithm on the chosen instance. We prove the following statement:

$$E[\tau] < \frac{c_1 k H^3}{\epsilon^2} \ln\left(\frac{c_2}{\delta}\right) \implies \exists M_i \in \mathcal{M}\left(1-\frac{1}{H}, \frac{40\epsilon}{H^2}, H\right) : P_i(|Q_i - \hat{Q}| > \epsilon) > \delta.$$

Lemma 15 (Chernoff Hoeffding's bound for Bernoulli random variables)**.** *Define s as the sum of l i.i.d. Bernoulli(p) tosses $\left(p > \frac{1}{2}\right)$. Define $\theta = \exp\left(-\frac{c_1'\alpha^2 l}{p(1-p)}\right)$, $\Delta = \sqrt{2p(1-p)l\ln\frac{c_2'}{2\theta}}$, and $\mathcal{E} = \{s \leq pl + \Delta\}$. Then:*

$$P(\mathcal{E}) > 1 - \frac{2\theta}{c_2'}.$$

Proof.

$$\begin{aligned}
P(\mathcal{E}) &> 1 - \exp\left(-\frac{KL(p+\Delta||p)}{l}\right) \\
&\geq 1 - \exp\left(-\frac{\Delta^2}{2lp(1-p)}\right) \\
&= 1 - \exp\left(-\ln\frac{c_2'}{2\theta}\right) = 1 - \frac{2\theta}{c_2'}.
\end{aligned}$$

□

Lemma 16 (Change of measure using likelihood ratios)**.**

$$P_i(E) = E_i[\mathbf{1}_E] = E_j\left[\frac{L_i(W)}{L_j(W)}\mathbf{1}_E\right] \tag{17}$$

where W is a sample path (for example, the actual sequence of Bernoulli tosses) that controls the event E, and $L_i(w) = P_i(W = w)$ is the likelihood of observing this sample path under MDP M_i.

This is a well-known result from probability, used exactly in this form by Azar et al. (2013).

Lemma 17 (Bound on the ratio of likelihoods)**.** *If $\alpha \leq (1-p)/2$, then, under $\mathcal{E}$, we have that*

$$\frac{L_i(W)}{L_0(W)} \geq \left(\frac{2\theta}{c_2'}\right)^{\frac{2}{c_1'}+\frac{2(1-p)}{pc_1'}+2\sqrt{\frac{2}{c_1'}}} \geq \frac{2\theta}{c_2'}.$$

Proof. We first note that $\alpha \leq \frac{1-p}{2} \leq p(1-p) \leq \frac{p}{2} \leq \frac{p}{\sqrt{2}}$ and $\alpha \leq \frac{1-p}{2} \implies \alpha^2 \leq \frac{(1-p)}{2}\frac{(1-p)}{2} \leq \frac{(1-p)}{2}\frac{1}{2} \leq \frac{p(1-p)}{2}$. Hence, $\alpha \leq (1-p)/2$ is a sufficient condition for the statements made in Eqn. 18, Eqn. 19, Eqn. 20, Eqn. 23 and Eqn. 24 (derived below). The first inequality simply follows by combining these equations. The second inequality follows by observing that we chose $p > \frac{1}{2}$, $c_1' = 20$, $c_2' = 6$, which imply that $\frac{2}{c_1'} + \frac{2(1-p)}{pc_1'} + 2\sqrt{\frac{2}{c_1'}} \leq 1$ and that $\frac{2\theta}{c_2'} < 1$.

Let s be the sum of l Bernoulli tosses. But there are two Bernoullis - p and $p+\alpha$. Let W be some sequence of tosses such that their sum was l. So we have:

$$\begin{aligned}
\frac{L_i(W)}{L_0(W)} &= \frac{(p+\alpha)^s(1-p-\alpha)^{l-s}}{p^s(1-p)^{l-s}} \\
&= \left(1+\frac{\alpha}{p}\right)^s\left(1-\frac{\alpha}{1-p}\right)^{l-s} \\
&= \left(1+\frac{\alpha}{p}\right)^s\left(1-\frac{\alpha}{1-p}\right)^{l-\frac{s}{p}}\left(1-\frac{\alpha}{1-p}\right)^{s\frac{1-p}{p}}.
\end{aligned}$$

Now, if $\alpha < (1-p)/2$,

$$\begin{aligned}
\left(1-\frac{\alpha}{1-p}\right)^{\frac{1-p}{p}} &= \exp\left(\ln\left(\left(1-\frac{\alpha}{1-p}\right)^{\frac{1-p}{p}}\right)\right) \\
&= \exp\left(\frac{1-p}{p}\ln\left(\left(1-\frac{\alpha}{1-p}\right)\right)\right) \\
(a) &\geq \exp\left(\frac{1-p}{p}\left(-\frac{\alpha}{1-p}-\left(\frac{\alpha}{1-p}\right)^2\right)\right) \\
&= \exp\left(-\left(\frac{\alpha}{p}+\frac{\alpha^2}{p(1-p)}\right)\right) \\
&= \exp\left(-\frac{\alpha}{p}\right)\exp\left(-\frac{\alpha^2}{p(1-p)}\right) \\
(b) &\geq \left(1-\frac{\alpha}{p}\right)\left(1-\frac{\alpha^2}{p(1-p)}\right)
\end{aligned} \tag{18}$$

where (a) follows from $\ln(1-u) \geq -u-u^2$ for $0 \leq u \leq 1/2$. Note that $\alpha < (1-p)/2 \implies \frac{\alpha}{1-p} \leq 1/2$. (b) follows from $\exp(-u) \geq 1-u$ for $0 \leq u \leq 1$, and the observation that $\alpha < (1-p)/2 < 1-p < p \implies \frac{\alpha}{p} \leq 1$.

Hence,

$$\begin{aligned}
\frac{L_i(W)}{L_0(W)} &= \left(1+\frac{\alpha}{p}\right)^s \left(1-\frac{\alpha}{1-p}\right)^{l-\frac{s}{p}} \left(1-\frac{\alpha}{1-p}\right)^{s\frac{1-p}{p}} \\
&\geq \left(1+\frac{\alpha}{p}\right)^s \left(1-\frac{\alpha}{1-p}\right)^{l-\frac{s}{p}} \left(1-\frac{\alpha}{p}\right)^s \left(1-\frac{\alpha^2}{p(1-p)}\right)^s \\
&= \left(1-\frac{\alpha^2}{p^2}\right)^s \left(1-\frac{\alpha}{1-p}\right)^{l-\frac{s}{p}} \left(1-\frac{\alpha^2}{p(1-p)}\right)^s \\
&\geq \left(1-\frac{\alpha^2}{p^2}\right)^l \left(1-\frac{\alpha}{1-p}\right)^{l-\frac{s}{p}} \left(1-\frac{\alpha^2}{p(1-p)}\right)^l . \qquad (19)
\end{aligned}$$

For $\alpha^2 \leq p^2/2$ and $\alpha^2 \leq p(1-p)/2$, we have:

$$\begin{aligned}
\left(1-\frac{\alpha^2}{p^2}\right)^l &= \exp\left(\ln\left(\left(1-\frac{\alpha^2}{p^2}\right)^l\right)\right) \\
&= \exp\left(l\ln\left(1-\frac{\alpha^2}{p^2}\right)\right) \\
(a) &\geq \exp\left(-2l\frac{\alpha^2}{p^2}\right) \\
&\geq \left(\frac{2\theta}{c_2'}\right)^{\frac{2(1-p)}{pc_1'}} \qquad (20)
\end{aligned}$$

where (a) follows from $\ln(1-u) \geq -2u$ for $0 \leq u \leq 1/2$, and the final step follows from the definition of θ and choice of c_2':

$$\exp\left(-2\frac{l\alpha^2}{p(1-p)}\right) = \theta^{\frac{2}{c_1'}} \geq \left(\frac{2\theta}{c_2'}\right)^{\frac{2}{c_1'}} \qquad (21)$$

$$\exp\left(-2\frac{l\alpha^2}{p^2}\right) = \theta^{\frac{2(1-p)}{c_1'p}} \geq \left(\frac{2\theta}{c_2'}\right)^{\frac{2(1-p)}{pc_1'}} \qquad (22)$$

Similarly,

$$\left(1-\frac{\alpha^2}{p(1-p)}\right)^l \geq \left(\frac{2\theta}{c_2'}\right)^{\frac{2}{c_1'}} . \qquad (23)$$

Finally, under $\mathcal{E}$, and for $\alpha \leq (1-p)/2$, we have:

$$l - \frac{s}{p} \leq \frac{\Delta}{p}. \text{ Therefore}$$

$$\begin{aligned}
\left(1 - \frac{\alpha}{1-p}\right)^{l-\frac{s}{p}} &\geq \exp\left(-2\frac{\Delta}{p}\frac{\alpha}{1-p}\right) \\
&= \exp\left(-2\frac{\alpha\sqrt{2p(1-p)l\log\frac{c_2'}{2\theta}}}{p(1-p)}\right) \\
&= \exp\left(-2\sqrt{\frac{2l\alpha^2}{p(1-p)}\log\frac{c_2'}{2\theta}}\right) \\
&\geq \exp\left(-2\sqrt{-\log\left(\frac{2\theta}{c_2'}^{\frac{2}{c_1'}}\right)\log\frac{c_2'}{2\theta}}\right) \\
&= \exp\left(-2\sqrt{\frac{2}{c_1'}\log\frac{c_2'}{2\theta}\log\frac{c_2'}{2\theta}}\right) \\
&= \log\left(\left(\frac{2\theta}{c_2'}\right)^{2\sqrt{\frac{2}{c_1'}}}\right). \qquad (24)
\end{aligned}$$

□

Lemma 18. *If an algorithm is (ϵ, δ)-PAC, then,*

$$E[\tau_i] > \tau_i^* \stackrel{def}{=} \frac{H^3}{64000\epsilon^2}\log\left(\frac{1}{6\delta}\right), \qquad (25)$$

where random variable τ_i denotes the number of samples of state-action pair i.

Proof. We choose $p = 1 - \frac{1}{H}$, $H > 200$, $\epsilon < 1$ and $\alpha = 40\epsilon(1-p)^2$ and construct the MDPs $M_i(p, \alpha, H), M_0(p, \alpha, H)$. Note that this choice gives us

$$\alpha = \frac{40\epsilon}{H^2} < \frac{40}{H^2} < \frac{1}{2H} = \frac{1-p}{2}.$$

Now let us assume that there is some (ϵ, δ)-PAC algorithm $\mathcal{L}$ with

$$E[\tau_i] \leq \tau_i^*.$$

We show that this leads to a contradiction.

First we bound θ as follows:

$$\begin{aligned}
\theta &= \exp\left(-\frac{c_1'\alpha^2\tau_i}{p(1-p)}\right) \\
&= \exp\left(-\frac{c_1' 1600\epsilon^2(1-p)^4\tau_i}{p(1-p)}\right) \\
&> \exp\left(-\frac{c_1' 3200\epsilon^2\tau_i}{H^3}\right) (\because p > 1/2).
\end{aligned}$$

This gives us (by applying Markov's inequality):

$$\tau_i \leq 10\tau_i^* \implies \frac{\theta}{c_2'} > \delta.$$

Let $E_i = \{|Q_i^* - \hat{Q}| \leq \epsilon\}$.

$$\therefore P_0(E_0^c) \leq \delta, P_i(E_i^c) \leq \delta,$$

by the assumption that $\mathcal{L}$ is (ϵ, δ)-PAC. Further, let $E_i' = E_0 \cap \mathcal{E} \cap \{\tau_i \leq 10\tau_i^*\}$.

$$\begin{aligned}
\therefore P_0(E_i') &\geq P_0(E_0)P_0(\mathcal{E})P_0\left(\{\tau_i \leq 10\tau_i^*\}\right) \\
&\geq (1-\delta)\left(1 - \frac{2\theta}{c_2'}\right)\frac{9}{10} \geq \left(1 - \frac{\theta}{c_2'}\right)\left(1 - \frac{2\theta}{c_2'}\right)\frac{9}{10} \\
&= \frac{5}{6}\frac{4}{6}\frac{9}{10} = \frac{1}{2},
\end{aligned}$$

where we invoke Lemma 15 to bound $P_0(\mathcal{E})$, Markov's inequality to bound $P_0\left(\{\tau_i \leq 10\tau_i^*\}\right)$ and the fact that $\theta < 1$ by definition. Then, using Lemma 17 and Lemma 16 gives us:

$$\begin{aligned}
P_i(E_0) \geq P_i(E_i') &= E_0\left[\frac{L_i(W)}{L_0(W)}\mathbf{1}_{E_i'}\right] \\
&\geq E_0\left[\frac{2\theta}{c_2'}\mathbf{1}_{E_i'}\right] \geq E_0\left[2\delta\mathbf{1}_{E_i'}\right] \\
&= 2\delta P_0(E_i') \geq \delta.
\end{aligned}$$

Using Lemma (14) gives us:

$$E_0 \subset E_i^c \implies P_i(E_i^c) > P_i(E_0) \geq \delta,$$

which contradicts that $\mathcal{L}$ is (ϵ, δ)-PAC.

□

By constructing the family $\mathcal{M}(p, \alpha, H)$, we can extend Lemma 18 to each τ_i and hence obtain a lower bound on the number of samples any (ϵ, δ)-PAC algorithm must observe. Note that $|\mathcal{M}| = KL = \frac{k}{3}$.