Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning

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

Tor Lattimore*
tor.lattimore@gmail.com

Emma Brunskill
Computer Science Department
Stanford University
ebrun@cs.stanford.edu

Abstract

Statistical performance bounds for reinforcement learning (RL) algorithms can be critical for high-stakes applications like healthcare. This paper introduces a new framework for theoretically measuring the performance of such algorithms called *Uniform-PAC*, which is a strengthening of the classical Probably Approximately Correct (PAC) framework. In contrast to the PAC framework, the uniform version may be used to derive high probability regret guarantees and so forms a bridge between the two setups that has been missing in the literature. We demonstrate the benefits of the new framework for finite-state episodic MDPs with a new algorithm that is Uniform-PAC and simultaneously achieves optimal regret *and* PAC guarantees except for a factor of the horizon.

1 Introduction

The recent empirical successes of deep reinforcement learning (RL) are tremendously exciting, but the performance of these approaches still varies significantly across domains, each of which requires the user to solve a new tuning problem [1]. Ultimately we would like reinforcement learning algorithms that simultaneously perform well empirically and have strong theoretical guarantees. Such algorithms are especially important for high stakes domains like health care, education and customer service, where non-expert users demand excellent outcomes.

We propose a new framework for measuring the performance of reinforcement learning algorithms called Uniform-PAC. Briefly, an algorithm is Uniform-PAC if with high probability it simultaneously for all $\varepsilon > 0$ selects an ε-optimal policy on all episodes except for a number that scales polynomially with $1/\varepsilon$. Algorithms that are Uniform-PAC converge to an optimal policy with high probability and immediately yield both PAC and high probability regret bounds, which makes them superior to algorithms that come with only PAC or regret guarantees. Indeed,

(a) Neither PAC nor regret guarantees imply convergence to optimal policies with high probability;

(b) (ε, δ)-PAC algorithms may be $\varepsilon/2$-suboptimal in every episode;

(c) Algorithms with small regret may be maximally suboptimal infinitely often.

*Tor Lattimore is now at DeepMind, London

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Uniform-PAC algorithms suffer none of these drawbacks. One could hope that existing algorithms with PAC or regret guarantees might be Uniform-PAC already, with only the analysis missing. Unfortunately this is not the case and modification is required to adapt these approaches to satisfy the new performance metric. The key insight for obtaining Uniform-PAC guarantees is to leverage time-uniform concentration bounds such as the finite-time versions of the law of iterated logarithm, which obviates the need for horizon-dependent confidence levels.

We provide a new optimistic algorithm for episodic RL called UBEV that is Uniform PAC. Unlike its predecessors, UBEV uses confidence intervals based on the law of iterated logarithm (LIL) which hold uniformly over time. They allow us to more tightly control the probability of failure events in which the algorithm behaves poorly. Our analysis is nearly optimal according to the traditional metrics, with a linear dependence on the state space for the PAC setting and square root dependence for the regret. Therefore UBEV is a Uniform PAC algorithm with PAC bounds and high probability regret bounds that are near optimal in the dependence on the length of the episodes (horizon) and optimal in the state and action spaces cardinality as well as the number of episodes. To our knowledge UBEV is the first algorithm with both near-optimal PAC and regret guarantees.

Notation and setup. We consider episodic fixed-horizon MDPs with time-dependent dynamics, which can be formalized as a tuple $M = (\mathcal{S}, \mathcal{A}, p_R, P, p_0, H)$. The statespace $\mathcal{S}$ and the actionspace $\mathcal{A}$ are finite sets with cardinality S and A. The agent interacts with the MDP in episodes of H time steps each. At the beginning of each time-step $t \in [H]$ the agent observes a state s_t and chooses an action a_t based on a policy π that may depend on the within-episode time step ($a_t = \pi(s_t, t)$). The next state is sampled from the tth transition kernel $s_{t+1} \sim P(\cdot|s_t, a_t, t)$ and the initial state from $s_1 \sim p_0$. The agent then receives a reward drawn from a distribution $p_R(s_t, a_t, t)$ which can depend on s_t, a_t and t with mean $r(s_t, a_t, t)$ determined by the reward function. The reward distribution p_R is supported on $[0, 1]$.[2] The value function from time step t for policy π is defined as

$$V_t^\pi(s) := \mathbb{E}\left[\sum_{i=t}^{H} r(s_i, a_i, i)\middle| s_t = s\right] = \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s,t), t) V_{t+1}^\pi(s') + r(s, \pi(s,t), t)\,.$$

and the optimal value function is denoted by $V_t^\star$. In any fixed episode, the quality of a policy π is evaluated by the *total expected reward* or *return*

$$\rho^\pi := \mathbb{E}\left[\sum_{i=t}^{H} r(s_i, a_i, i)\middle|\pi\right] = p_0^\top V_1^\pi\,,$$

which is compared to the *optimal return* $\rho^\star = p_0^\top V_1^\star$. For this notation p_0 and the value functions $V_t^\star$, V_1^π are interpreted as vectors of length S. If an algorithm follows policy π_k in episode k, then the optimality gap in episode k is $\Delta_k := \rho^\star - \rho^{\pi_k}$ which is bounded by $\Delta_{\max} = \max_\pi \rho^\star - \rho^\pi \leq H$. We let $N_\varepsilon := \sum_{k=1}^\infty \mathbb{I}\{\Delta_k > \varepsilon\}$ be the number of ε-errors and $R(T)$ be the regret after T episodes: $R(T) := \sum_{k=1}^T \Delta_k$. Note that T is the number of episodes and not total time steps (which is HT after T episodes) and k is an episode index while t usually denotes time indices within an episode. The $\tilde{O}$ notation is similar to the usual O-notation but suppresses additional polylog-factors, that is $g(x) = \tilde{O}(f(x))$ iff there is a polynomial p such that $g(x) = O(f(x)p(\log(x)))$.

2 Uniform PAC and Existing Learning Frameworks

We briefly summarize the most common performance measures used in the literature.

- (ε, δ)-*PAC:* There exists a polynomial function $F_{\text{PAC}}(S, A, H, 1/\varepsilon, \log(1/\delta))$ such that
$$\mathbb{P}\left(N_\varepsilon > F_{\text{PAC}}(S, A, H, 1/\varepsilon, \log(1/\delta))\right) \leq \delta\,.$$
- *Expected Regret:* There exists a function $F_{\text{ER}}(S, A, H, T)$ such that $\mathbb{E}[R(T)] \leq F_{\text{ER}}(S, A, H, T)$.
- *High Probability Regret:* There exists a function $F_{\text{HPR}}(S, A, H, T, \log(1/\delta))$ such that
$$\mathbb{P}\left(R(T) > F_{\text{HPR}}(S, A, H, T, \log(1/\delta))\right) \leq \delta\,.$$

[2]The reward may be allowed to depend on the next-state with no further effort in the proofs. The boundedness assumption could be replaced by the assumption of subgaussian noise with known subgaussian parameter.

- *Uniform High Probability Regret:* There exists a function $F_{\text{UHPR}}(S, A, H, T, \log(1/\delta))$ such that

$$\mathbb{P}\left(\text{exists } T : R(T) > F_{\text{UHPR}}(S, A, H, T, \log(1/\delta))\right) \le \delta\,.$$

In all definitions the function F should be polynomial in all arguments. For notational conciseness we often omit some of the parameters of F where the context is clear. The different performance guarantees are widely used (e.g. PAC: [2, 3, 4, 5], (uniform) high-probability regret: [6, 7, 8]; expected regret: [9, 10, 11, 12]). Due to space constraints, we will not discuss Bayesian-style performance guarantees that only hold in expectation with respect to a distribution over problem instances. We will shortly discuss the limitations of the frameworks listed above, but first formally define the Uniform-PAC criteria

Definition 1 (Uniform-PAC)**.** *An algorithm is Uniform-PAC for $\delta > 0$ if*

$$\mathbb{P}\left(\textit{exists } \varepsilon > 0 :\ N_\varepsilon > F_{\textit{UPAC}}\left(S, A, H, 1/\varepsilon, \log(1/\delta)\right)\right) \le \delta\,,$$

where $F_{\textit{UPAC}}$ is polynomial in all arguments.

All the performance metrics are functions of the distribution of the sequence of errors over the episodes $(\Delta_k)_{k\in\mathbb{N}}$. Regret bounds are the integral of this sequence up to time T, which is a random variable. The expected regret is just the expectation of the integral, while the high-probability regret is a quantile. PAC bounds are the quantile of the size of the superlevel set for a fixed level ε. Uniform-PAC bounds are like PAC bounds, but hold for all ε simultaneously.

Limitations of regret. Since regret guarantees only bound the integral of Δ_k over k, it does not distinguish between making a few severe mistakes and many small mistakes. In fact, since regret bounds provably grow with the number of episodes T, an algorithm that achieves optimal regret may still make infinitely many mistakes (of arbitrary quality, see proof of Theorem 2 below). This is highly undesirable in high-stakes scenarios. For example in drug treatment optimization in healthcare, we would like to distinguish between infrequent severe complications (few large Δ_k) and frequent minor side effects (many small Δ_k). In fact, even with an optimal regret bound, we could still serve infinitely patients with the worst possible treatment.

Limitations of PAC. PAC bounds limit the number of mistakes for a given accuracy level ε, but is otherwise non-restrictive. That means an algorithm with $\Delta_k > \varepsilon/2$ for all k almost surely might still be (ε, δ)-PAC. Worse, many algorithms designed to be (ε, δ)-PAC actually exhibit this behavior because they explicitly halt learning once an ε-optimal policy has been found. The less widely used TCE (total cost of exploration) bounds [13] and KWIK guarantees [14] suffer from the same issueand for conciseness are not discussed in detail.

Advantages of Uniform-PAC. The new criterion overcomes the limitations of PAC and regret guarantees by measuring the number of ε-errors at every level simultaneously. By definition, algorithms that are Uniform-PAC for a δ are (ε, δ)-PAC for all $\varepsilon > 0$. We will soon see that an algorithm with a non-trivial Uniform-PAC guarantee also has small regret with high probability. Furthermore, there is no loss in the reduction so that an algorithm with optimal Uniform-PAC guarantees also has optimal regret, at least in the episodic RL setting. In this sense Uniform-PAC is the missing bridge between regret and PAC. Finally, for algorithms based on confidence bounds, Uniform-PAC guarantees are usually obtained without much additional work by replacing standard concentration bounds with versions that hold uniformly over episodes (e.g. using the law of the iterated logarithms). In this sense we think Uniform-PAC is the new 'gold-standard' of theoretical guarantees for RL algorithms.

2.1 Relationships between Performance Guarantees

Existing theoretical analyses usually focus exclusively on either the regret or PAC framework. Besides occasional heuristic translations, Proposition 4 in [15] and Corollary 3 in [6] are the only results relating a notion of PAC and regret, we are aware of. Yet the guarantees there are not widely used[3]

[3]The average per-step regret in [6] is superficially a PAC bound, but does not hold over infinitely many time-steps and exhibits the limitations of a conventional regret bound. The translation to average loss in [15] comes at additional costs due to the discounted infinite horizon setting.

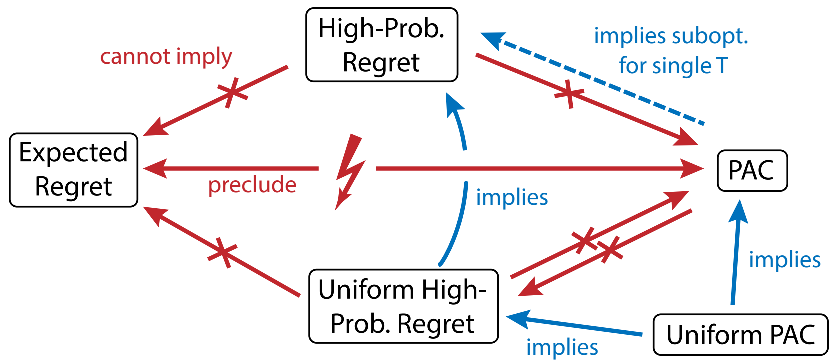

Figure 1: Visual summary of relationship among the different learning frameworks: Expected regret (ER) and PAC preclude each other while the other crossed arrows represent only a *does-not-implies* relationship. Blue arrows represent *imply* relationships. For details see the theorem statements.

unlike the definitions given above which we now formally relate to each other. A simplified overview of the relations discussed below is shown in Figure 1.

Theorem 1. *No algorithm can achieve*

- *a sub-linear expected regret bound for all T and*
- *a finite (ε, δ)-PAC bound for a small enough ε*

simultaneously for all two-armed multi-armed bandits with Bernoulli reward distributions. This implies that such guarantees also cannot be satisfied simultaneously for all episodic MDPs.

A full proof is in Appendix A.1, but the intuition is simple. Suppose a two-armed Bernoulli bandit has mean rewards $1/2 + \varepsilon$ and $1/2$ respectively and the second arm is chosen at most $F < \infty$ times with probability at least $1 - \delta$, then one can easily show that in an alternative bandit with mean rewards $1/2 + \varepsilon$ and $1/2 + 2\varepsilon$ there is a non-zero probability that the second arm is played finitely often and in this bandit the expected regret will be linear. Therefore, sub-linear expected regret is only possible if each arm is pulled infinitely often almost surely.

Theorem 2. *The following statements hold for performance guarantees in episodic MDPs:*

(a) If an algorithm satisfies a (ε, δ)-PAC bound with $F_{PAC} = \Theta(1/\varepsilon^2)$ then it satisfies for a specific $T = \Theta(\varepsilon^{-3})$ a $F_{HPR} = \Theta(T^{2/3})$ bound. Further, there is an MDP and algorithm that satisfies the (ε, δ)-PAC bound $F_{PAC} = \Theta(1/\varepsilon^2)$ on that MDP and has regret $R(T) = \Omega(T^{2/3})$ on that MDP for any T. That means a (ε, δ)-PAC bound with $F_{PAC} = \Theta(1/\varepsilon^2)$ can only be converted to a high-probability regret bound with $F_{HPR} = \Omega(T^{2/3})$.

(b) For any chosen $\varepsilon, \delta > 0$ and F_{PAC}, there is an MDP and algorithm that satisfies the (ε, δ)-PAC bound F_{PAC} on that MDP and has regret $R(T) = \Omega(T)$ on that MDP. That means a (ε, δ)-PAC bound cannot be converted to a sub-linear uniform high-probability regret bound.

(c) For any $F_{UHPR}(T, \delta)$ with $F_{UHPR}(T, \delta) \to \infty$ as $T \to \infty$, there is an algorithm that satisfies that uniform high-probability regret bound on some MDP but makes infinitely many mistakes for any sufficiently small accuracy level $\varepsilon > 0$ for that MDP. Therefore, a high-probability regret bound (uniform or not) cannot be converted to a finite (ε, δ)-PAC bound.

(d) For any $F_{UHPR}(T, \delta)$ there is an algorithm that satisfies that uniform high-probability regret bound on some MDP but suffers expected regret $\mathbb{E}R(T) = \Omega(T)$ on that MDP.

For most interesting RL problems including episodic MDPs the worst-case expected regret grows with $O(\sqrt{T})$. The theorem shows that establishing an optimal high probability regret bound does not imply any finite PAC bound. While PAC bounds may be converted to regret bounds, the resulting bounds are necessarily severely suboptimal with a rate of $T^{2/3}$. The next theorem formalises the claim that Uniform-PAC is stronger than both the PAC and high-probability regret criteria.

Theorem 3. *Suppose an algorithm is Uniform-PAC for some δ with $F_{UPAC} = \tilde{O}(C_1/\varepsilon + C_2/\varepsilon^2)$ where $C_1, C_2 > 0$ are constant in ε, but may depend on other quantities such as S, A, H, $\log(1/\delta)$, then the algorithm*

(a) converges to optimal policies with high probability: $\mathbb{P}(\lim_{k\to\infty} \Delta_k = 0) \geq 1 - \delta$.

(b) is (ε, δ)-PAC with bound $F_{PAC} = F_{UPAC}$ for all ε.

(c) enjoys a high-probability regret at level δ with $F_{UHPR} = \tilde{O}(\sqrt{C_2 T} + \max\{C_1, C_2\})$.

Observe that stronger uniform PAC bounds lead to stronger regret bounds and for RL in episodic MDPs, an optimal uniform-PAC bound implies a uniform regret bound. To our knowledge, there are no existing approaches with PAC or regret guarantees that are Uniform-PAC. PAC methods such as MBIE, MoRMax, UCRL-γ, UCFH, Delayed Q-Learning or Median-PAC all depend on advance knowledge of ε and eventually stop improving their policies. Even when disabling the stopping condition, these methods are not uniform-PAC as their confidence bounds only hold for finitely many episodes and are eventually violated according to the law of iterated logarithms. Existing algorithms with uniform high-probability regret bounds such as UCRL2 or UCBVI [16] also do not satisfy uniform-PAC bounds since they use upper confidence bounds with width $\sqrt{\log(T)/n}$ where T is the number of observed episodes and n is the number of observations for a specific state and action. The presence of $\log(T)$ causes the algorithm to try each action in each state infinitely often. One might begin to wonder if uniform-PAC is too good to be true. Can *any* algorithm meet the requirements? We demonstrate in Section 4 that the answer is yes by showing that UBEV has meaningful Uniform-PAC bounds. A key technique that allows us to prove these bounds is the use of finite-time law of iterated logarithm confidence bounds which decrease at rate $\sqrt{(\log\log n)/n}$.

3 The UBEV Algorithm

The pseudo-code for the proposed UBEV algorithm is given in Algorithm 1. In each episode it follows an optimistic policy π_k that is computed by backwards induction using a carefully chosen confidence interval on the transition probabilities in each state. In line 8 an optimistic estimate of the Q-function for the current state-action-time triple is computed using the empirical estimates of the expected next state value $\hat{V}_{\text{next}} \in \mathbb{R}$ (given that the values at the next time are $\tilde{V}_{t+1}$) and expected immediate reward $\hat{r}$ plus confidence bounds $(H-t)\phi$ and ϕ. We show in Lemma D.1 in the appendix that the policy update in Lines 3–9 finds an optimal solution to $\max_{P', r', V', \pi'} \ \mathbb{E}_{s\sim p_0}[V_1'(s)]$ subject to the constraints that for all $s \in \mathcal{S}, a \in \mathcal{A}, t \in [H]$,

$$
\begin{aligned}
&V_t'(s) = r(s, \pi'(s,t), t) + P'(s, \pi'(s,t), t)^\top V_{t+1}' \qquad \text{(Bellman Equation)} && (1)\\
&V_{H+1}' = 0, \quad P'(s,a,t) \in \Delta_S, \quad r'(s,a,t) \in [0,1] \\
&|[(P' - \hat{P}_k)(s,a,t)]^\top V_{t+1}'| \leq \phi(s,a,t)(H-t) \\
&|r'(s,a,t) - \hat{r}_k(s,a,t)| \leq \phi(s,a,t) && (2)
\end{aligned}
$$

where $(P' - \hat{P}_k)(s,a,t)$ is short for $P'(s,a,t) - \hat{P}_k(s,a,t) = P'(\cdot|s,a,t) - \hat{P}_k(\cdot|s,a,t)$ and

$$
\phi(s,a,t) = \sqrt{\frac{2\ln\ln\max\{e, n(s,a,t)\} + \ln(18SAH/\delta)}{n(s,a,t)}} = O\left(\sqrt{\frac{\ln(SAH\ln(n(s,a,t))/\delta)}{n(s,a,t)}}\right)
$$

is the width of a confidence bound with $e = \exp(1)$ and $\hat{P}_k(s'|s,a,t) = \frac{m(s',s,a,t)}{n(s,a,t)}$ are the empirical transition probabilities and $\hat{r}_k(s,a,t) = l(s,a,t)/n(s,a,t)$ the empirical immediate rewards (both at the beginning of the kth episode). Our algorithm is conceptually similar to other algorithms based on the optimism principle such as MBIE [5], UCFH [3], UCRL2 [6] or UCRL-γ [2] but there are several key differences:

- Instead of using confidence intervals over the transition kernel by itself, we incorporate the value function directly into the concentration analysis. Ultimately this saves a factor of S in the sample complexity, but the price is a more difficult analysis. Previously MoRMax [17] also used the idea of directly bounding the transition and value function, but in a very different algorithm that required discarding data and had a less tight bound. A similar technique has been used by Azar et al. [16].

Algorithm 1: UBEV (**U**pper **B**ounding the **E**xpected Next State **V**alue) Algorithm

Input : failure tolerance $\delta \in (0, 1]$

$n(s, a, t) = l(s, a, t) = m(s', s, a, t) = 0; \quad \tilde{V}_{H+1}(s') := 0 \quad \forall s, s' \in \mathcal{S}, a \in \mathcal{A}, t \in [H]$

for $k = 1, 2, 3, \ldots$ **do**

  /* Optimistic planning */

  for $t = H$ **to** 1 **do**

    for $s \in \mathcal{S}$ **do**

      for $a \in \mathcal{A}$ **do**

        $\phi := \sqrt{\frac{2 \ln\ln(\max\{e, n(s,a,t)\}) + \ln(18SAH/\delta)}{n(s,a,t)}}$ // confidence bound

        $\hat{r} := \frac{l(s,a,t)}{n(s,a,t)}; \quad \hat{V}_{\text{next}} := \frac{m(\cdot,s,a,t)^\top \tilde{V}_{t+1}}{n(s,a,t)}$ // empirical estimates

        $Q(a) := \min\{1, \hat{r} + \phi\} + \min\left\{\max \tilde{V}_{t+1}, \hat{V}_{\text{next}} + (H - t)\phi\right\}$

      $\pi_k(s, t) := \arg\max_a Q(a), \quad \tilde{V}_t(s) := Q(\pi_k(s, t))$

  /* Execute policy for one episode */

  $s_1 \sim p_0$;

  for $t = 1$ **to** H **do**

    $a_t := \pi_k(s_t, t), \; r_t \sim p_R(s_t, a_t, t)$ and $s_{t+1} \sim P(s_t, a_t, t)$

    $n(s_t, a_t, t)$++; $\quad m(s_{t+1}, s_t, a_t, t)$++; $\quad l(s_t, a_t, t)$+= r_t // update statistics

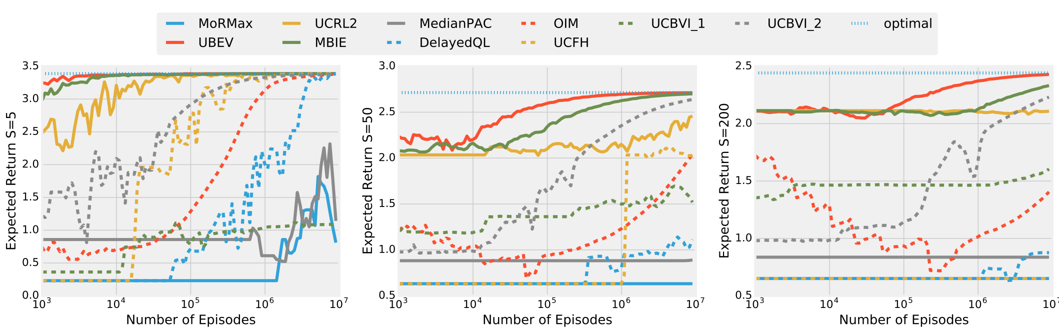

Figure 2: Empirical comparison of optimism-based algorithms with frequentist regret or PAC bounds on a randomly generated MDP with 3 actions, time horizon 10 and $S = 5, 50, 200$ states. All algorithms are run with parameters that satisfy their bound requirements. A detailed description of the experimental setup including a link to the source code can be found in Appendix B.

- Many algorithms update their policy less and less frequently (usually when the number of samples doubles), and only finitely often in total. Instead, we update the policy after every episode, which means that UBEV immediately leverages new observations.
- Confidence bounds in existing algorithms that keep improving the policy (e.g. Jaksch et al. [6], Azar et al. [16]) scale at a rate $\sqrt{\log(k)/n}$ where k is the number of episodes played so far and n is the number of times the specific (s, a, t) has been observed. As the results of a brief empirical comparison in Figure 2 indicate, this leads to slow learning (compare UCBVI_1 and UBEV's performance which differ essentially only by their use of different rate bounds). Instead the width of UBEV's confidence bounds ϕ scales at rate $\sqrt{\ln\ln(\max\{e, n\})/n} \approx \sqrt{(\log\log n)/n}$ which is the best achievable rate and results in significantly faster learning.

4 Uniform PAC Analysis

We now discuss the Uniform-PAC analysis of UBEV which results in the following Uniform-PAC and regret guarantee.

Theorem 4. *Let π_k be the policy of UBEV in the kth episode. Then with probability at least $1-\delta$ for all $\varepsilon > 0$ jointly the number of episodes k where the expected return from the start state is not ε-optimal (that is $\Delta_k > \varepsilon$) is at most*

$$O\left(\frac{SAH^4}{\varepsilon^2}\min\left\{1{+}\varepsilon S^2 A, S\right\}\operatorname{polylog}\left(A, S, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\right)\right).$$

Therefore, with probability at least $1-\delta$ UBEV converges to optimal policies and for all episodes T has regret

$$R(T) = O\left(H^2(\sqrt{SAT} + S^3A^2)\operatorname{polylog}(S, A, H, T)\right).$$

Here $\operatorname{polylog}(x\dots)$ is a function that can be bounded by a polynomial of logarithm, that is, $\exists k, C :$ $\operatorname{polylog}(x\dots) \le \ln(x\dots)^k{+}C$. In Appendix C we provide a lower bound on the sample complexity that shows that if $\varepsilon < 1/(S^2A)$, the Uniform-PAC bound is tight up to log-factors and a factor of H. To our knowledge, UBEV is the first algorithm with both near-tight (up to H factors) high probability regret and (ε, δ) PAC bounds as well as the first algorithm with any nontrivial uniform-PAC bound.

Using Theorem 3 the convergence and regret bound follows immediately from the uniform PAC bound. After a discussion of the different confidence bounds allowing us to prove uniform-PAC bounds, we will provide a short proof sketch of the uniform PAC bound.

4.1 Enabling Uniform PAC With Law-of-Iterated-Logarithm Confidence Bounds

To have a PAC bound for all ε jointly, it is critical that UBEV continually make use of new experience. If UBEV stopped leveraging new observations after some fixed number, it would not be able to distinguish with high probability among which of the remaining possible MDPs do or do not have optimal policies that are sufficiently optimal in the other MDPs. The algorithm therefore could potentially follow a policy that is not at least ε-optimal for infinitely many episodes for a sufficiently small ε. To enable UBEV to incorporate all new observations, the confidence bounds in UBEV must hold for an infinite number of updates. We therefore require a proof that the total probability of all possible failure events (of the high confidence bounds not holding) is bounded by δ, in order to obtain high probability guarantees. In contrast to prior (ε, δ)-PAC proofs that only consider a finite number of failure events (which is enabled by requiring an RL algorithm to stop using additional data), we must bound the probability of an infinite set of possible failure events.

Some choices of confidence bounds will hold uniformly across all sample sizes but are not sufficiently tight for uniform PAC results. For example, the recent work by Azar et al. [16] uses confidence intervals that shrink at a rate of $\sqrt{\frac{\ln T}{n}}$, where T is the number of episodes, and n is the number of samples of a (s, a) pair at a particular time step. This confidence interval will hold for all episodes, but these intervals do not shrink sufficiently quickly and can even increase. One simple approach for constructing confidence intervals that is sufficient for uniform PAC guarantees is to combine bounds for fixed number of samples with a union bound allocating failure probability δ/n^2 to the failure case with n samples. This results in confidence intervals that shrink at rate $\sqrt{1/n \ln n}$. Interestingly we know of no algorithms that do such in our setting.

We follow a similarly simple but much stronger approach of using law-of-iterated logarithm (LIL) bounds that shrink at the better rate of $\sqrt{1/n \ln\ln n}$. Such bounds have sparked recent interest in sequential decision making [18, 19, 20, 21, 22] but to the best of our knowledge we are the first to leverage them for RL. We prove several general LIL bounds in Appendix F and explain how we use these results in our analysis in Appendix E.2. These LIL bounds are both sufficient to ensure uniform PAC bounds, and much tighter (and therefore will lead to much better performance) than $\sqrt{1/n \ln T}$ bounds. Indeed, LIL have the tightest possible rate dependence on the number of samples n for a bound that holds for all timesteps (though they are not tight with respect to constants).

4.2 Proof Sketch

We now provide a short overview of our uniform PAC bound in Theorem 4. It follows the typical scheme for optimism based algorithms: we show that in each episode UBEV follows a policy that is optimal with respect to the MDP $\tilde{M}_k$ that yields highest return in a set of MDPs $\mathcal{M}_k$ given by the constraints in Eqs. (1)–(2) (Lemma D.1 in the appendix). We then define a failure event F (more details see below) such that on the complement F^C, the true MDP is in $\mathcal{M}_k$ for all k.

Under the event that the true MDP is in the desired set, the $V_1^\pi \leq V_1^\star \leq \tilde{V}_1^{\pi_k}$, i.e., the value $\tilde{V}_1^{\pi_k}$ of π_k in MDP $\tilde{M}_k$ is higher than the optimal value function of the true MDP M (Lemma E.16). Therefore, the optimality gap is bounded by $\Delta_k \leq p_0^\top(\tilde{V}_1^{\pi_k} - V_1^{\pi_k})$. The right hand side this expression is then decomposed via a standard identity (Lemma E.15) as

$$\sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_{tk}(s,a)((\tilde{P}_k - P)(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k} + \sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_{tk}(s,a)(\tilde{r}_k(s,a,t) - r(s,a,t)),$$

where $w_{tk}(s,a)$ is the probability that when following policy π_k in the true MDP we encounter $s_t = s$ and $a_t = a$. The quantities $\tilde{P}_k$, $\tilde{r}_k$ are the model parameters of the optimistic MDP $\tilde{M}_k$ For the sake of conciseness, we ignore the second term above in the following which can be bounded by $\varepsilon/3$ in the same way as the first. We further decompose the first term as

$$\sum_{\substack{t\in[H]\\(s,a)\in L_{tk}^c}} w_{tk}(s,a)((\tilde{P}_k - P)(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k} \tag{3}$$

$$+\sum_{\substack{t\in[H]\\(s,a)\in L_{tk}}} w_{tk}(s,a)((\tilde{P}_k - \hat{P}_k)(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k} + \sum_{\substack{t\in[H]\\(s,a)\in L_{tk}}} w_{tk}(s,a)((\hat{P}_k - P)(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k} \tag{4}$$

where $L_{tk} = \left\{(s,a)\in\mathcal{S}\times\mathcal{A} : w_{tk}(s,a) \geq w_{\min} = \frac{\varepsilon}{3HS^2}\right\}$ is the set of state-action pairs with non-negligible visitation probability. The value of $w_{\min}$ is chosen so that (3) is bounded by $\varepsilon/3$. Since $\tilde{V}^{\pi_k}$ is the optimal solution of the optimization problem in Eq. (1), we can bound

$$|((\tilde{P}_k - \hat{P}_k)(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| \leq \phi_k(s,a,t)(H-t) = O\left(\sqrt{\frac{H^2 \ln(\ln(n_{tk}(s,a))/\delta)}{n_{tk}(s,a)}}\right), \tag{5}$$

where $\phi_k(s,a,t)$ is the value of $\phi(s,a,t)$ and $n_{tk}(s,a)$ the value of $n(s,a,t)$ right before episode k. Further we decompose

$$|((\hat{P}_k - P)(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| \leq \|(\hat{P}_k - P)(s,a,t)\|_1 \|\tilde{V}_{t+1}^{\pi_k}\|_\infty \leq O\left(\sqrt{\frac{SH^2 \ln\frac{\ln n_{tk}(s,a)}{\delta}}{n_{tk}(s,a)}}\right), \tag{6}$$

where the second inequality follows from a standard concentration bound used in the definition of the failure event F (see below). Substituting this and (5) into (4) leads to

$$(4) \leq O\left(\sum_{t=1}^{H} \sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{\frac{SH^2\ln(\ln(n_{tk}(s,a))/\delta)}{n_{tk}(s,a)}}\right). \tag{7}$$

On F^C it also holds that $n_{tk}(s,a) \geq \frac{1}{2}\sum_{i<k} w_{ti}(s,a) - \ln\frac{9SAH}{\delta}$ and so on *nice episodes* where each $(s,a)\in L_{tk}$ with significant probability $w_{tk}(s,a)$ also had significant probability in the past, i.e., $\sum_{i<k} w_{ti}(s,a) \geq 4\ln\frac{9SA}{\delta}$, it holds that $n_{tk}(s,a) \geq \frac{1}{4}\sum_{i<k} w_{ti}(s,a)$. Substituting this into (7), we can use a careful pidgeon-hole argument laid out it Lemma E.3 in the appendix to show that this term is bounded by $\varepsilon/3$ on all but $O(AS^2H^4/\varepsilon^2 \operatorname{polylog}(A,S,H,1/\varepsilon,1/\delta))$ nice episodes. Again using a pidgeon-hole argument, one can show that all but at most $O(S^2AH^3/\varepsilon \ln(SAH/\delta))$ episodes are nice. Combining both bounds, we get that on F^C the optimality gap Δ_k is at most ε except for at most $O(AS^2H^4/\varepsilon^2 \operatorname{polylog}(A,S,H,1/\varepsilon,1/\delta))$ episodes.

We decompose the failure event into multiple components. In addition to the events F_k^N that a (s,a,t) triple has been observed few times compared to its visitation probabilities in the past, i.e., $n_{tk}(s,a) < \frac{1}{2}\sum_{i<k} w_{ti}(s,a) - \ln\frac{9SAH}{\delta}$ as well as a conditional version of this statement, the failure event F contains events where empirical estimates of the immediate rewards, the expected optimal value of the successor states and the individual transition probabilites are far from their true

expectations. For the full definition of F see Appendix E.2. F also contains event F^{L1} we used in Eq. (6) defined as

$$\left\{\exists k, s, a, t \,:\, \|\hat{P}_k(s,a,t) - P(s,a,t)\|_1 \geq \sqrt{\tfrac{4}{n_{tk}(s,a)}\left(2\,\mathrm{llnp}(n_{tk}(s,a)) + \ln\tfrac{18SAH(2^S-2)}{\delta}\right)}\right\}.$$

It states that the L1-distance of the empirical transition probabilities to the true probabilities for any (s,a,t) in any episode k is too large and we show that $\mathbb{P}(F^{L1}) \leq 1 - \delta/9$ using a uniform version of the popular bound by Weissman et al. [23] which we prove in Appendix F. We show in similar manner that the other events in F have small probability uniformly for all episodes k so that $\mathbb{P}(F) \leq \delta$. Together this yields the uniform PAC bound in Thm. 4 using the second term in the $\min$.

With a more refined analysis that avoids the use of Hölder's inequality in (6) and a stronger notion of nice episodes called friendly episodes we obtain the bound with the first term in the $\min$. However, since a similar analysis has been recently released [16], we defer this discussion to the appendix.

4.3 Discussion of UBEV Bound

The (Uniform-)PAC bound for UBEV in Theorem 4 is never worse than $\tilde{O}(S^2AH^4/\varepsilon^2)$, which improves on the similar MBIE algorithm by a factor of H^2 (after adapting the discounted setting for which MBIE was analysed to our setting). For $\varepsilon < 1/(S^2A)$ our bound has a linear dependence on the size of the state-space and depends on H^4, which is a tighter dependence on the horizon than MoRMax's $\tilde{O}(SAH^6/\varepsilon^2)$, the best sample-complexity bound with linear dependency S so far.

Comparing UBEV's regret bound to the ones of UCRL2 [6] and REGAL [24] requires care because (a) we measure the regret over entire episodes and (b) our transition dynamics are time-dependent within each episode, which effectively increases the state-space by a factor of H. Converting the bounds for UCRL2/REGAL to our setting yields a regret bound of order $SH^2\sqrt{AHT}$. Here, the diameter is H, the state space increases by H due to time-dependent transition dynamics and an additional $\sqrt{H}$ is gained by stating the regret in terms of episodes T instead of time steps. Hence, UBEV's bounds are better by a factor of $\sqrt{SH}$. Our bound matches the recent regret bound for episodic RL by Azar et al. [16] in the S, A and T terms but not in H. Azar et al. [16] has regret bounds that are optimal in H but their algorithm is not uniform PAC, due to the characteristics we outlined in Section 2.

5 Conclusion

The Uniform-PAC framework strengthens and unifies the PAC and high-probability regret performance criteria for reinforcement learning in episodic MDPs. The newly proposed algorithm is Uniform-PAC, which as a side-effect means it is the first algorithm that is both PAC and has sublinear (and nearly optimal) regret. Besides this, the use of law-of-the-iterated-logarithm confidence bounds in RL algorithms for MDPs provides a practical and theoretical boost at no cost in terms of computation or implementation complexity.

This work opens up several immediate research questions for future work. The definition of Uniform-PAC and the relations to other PAC and regret notions directly apply to multi-armed bandits and contextual bandits as special cases of episodic RL, but not to infinite horizon reinforcement learning. An extension to these non-episodic RL settings is highly desirable. Similarly, a version of the UBEV algorithm for infinite-horizon RL with linear state-space sample complexity would be of interest. More broadly, if theory is ever to say something useful about practical algorithms for large-scale reinforcement learning, then it will have to deal with the unrealizable function approximation setup (unlike the tabular function representation setting considered here), which is a major long-standing open challenge.

Acknowledgements. We appreciate the support of a NSF CAREER award and a gift from Yahoo.

References

[1] Vincent François-Lavet, Raphaël Fonteneau, and Damien Ernst. How to discount deep reinforcement learning: Towards new dynamic strategies. In *NIPS 2015 Workshop on Deep*

Reinforcement Learning, 2015.

[2] Tor Lattimore and Marcus Hutter. Near-optimal PAC bounds for discounted MDPs. In *Theoretical Computer Science*, volume 558, 2014.

[3] Christoph Dann and Emma Brunskill. Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning. In *Neural Information Processing Systems*, 2015.

[4] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable. In *International Conference on Machine Learning*, 2017.

[5] 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.

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

[7] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *Journal of Machine Learning Research*, volume 32, 2014.

[8] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. In *IEEE Transactions on Information Theory*, volume 58, 2012.

[9] Jean Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

[10] Peter Auer. Using upper confidence bounds for online learning. *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 270–293, 2000.

[11] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *arXiv.org*, cs.LG(1):138, 2012.

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

[13] Jason Pazis and Ronald Parr. Efficient PAC-optimal Exploration in Concurrent , Continuous State MDPs with Delayed Updates. In *AAAI Conference on Artificial Intelligence*, 2016.

[14] Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: A framework for self-aware learning. *Machine Learning*, 82(3):399–443, nov 2011.

[15] 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, 2008.

[16] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, 2017.

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

[18] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. 2013.

[19] Akshay Balsubramani and Aaditya Ramdas. Sequential Nonparametric Testing with the Law of the Iterated Logarithm. In *Uncertainty in Artificial Intelligence*, 2016.

[20] Aurélien Garivier, Emilie Kaufmann, and Tor Lattimore. On Explore-Then-Commit Strategies. In *Advances in Neural Information Processing Systems*, 2016.

[21] Pascal Massart. Concentration inequalities and model selection. *Lecture Notes in Mathematics*, 1896, 2007.

[22] Aurelien Garivier and Olivier Cappe. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Conference on Learning Theory*, 2011.

[23] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L 1 Deviation of the Empirical Distribution. Technical report, 2003. URL http://www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf?origin=publicationDetail.

[24] Peter L. Bartlett and a. Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42, 2009.

[25] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5.

[26] Rick Durrett. *Probability - Theory and Examples*. Cambridge University Press, 4 edition, 2010. ISBN 978-0-521-76539-8.

Appendices of Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning

A Framework Relation Proofs

A.1 Proof of Theorem 1

Proof. We will use two episodic MDPs, M_1 and M_2, which are essentially 2-armed bandits and hard to distinguish to prove this statement. Both MDPs have one state, horizon $H = 1$, and two actions $\mathcal{A} = \{1, 2\}$. For a fixed $\alpha > 0$, the rewards are Bernoulli$(1/2 + \alpha/2)$ distributed for actions 1 in both MDPs. Playing action 2 in M_1 gives Bernoulli$(1/2)$ rewards and action 2 in M_2 gives Bernoulli$(1/2 + \alpha)$ rewards.

Assume now that an algorithm in MDP M_1 with nonzero probability plays the suboptimal action only at most N times in total, i.e., $\mathbb{P}_{M_1}(n_2 \le N) \ge \beta$ where n_2 is the number of times action 2 is played and $\infty > N > 0, \beta > 0$. Then

$$\mathbb{P}_{M_1}(n_2 \le N) = \mathbb{E}_{M_1}\left[\mathbb{I}\{n_2 \le N\}\right] = \mathbb{E}_{M_2}\left[\frac{\mathbb{P}_{M_1}(Y_\infty)}{\mathbb{P}_{M_2}(Y_\infty)}\mathbb{I}\{n_2 \le N\}\right]$$

where $Y_k = (A_1, R_1, A_2, R_2, \dots A_k, R_k)$ denotes the entire sequence of observed rewards R_i and action indices A_i after k episodes. Since $\mathbb{P}_{M_1}(A_k|Y_{k-1}) = \mathbb{P}_{M_2}(A_k|Y_{k-1})$ and $\mathbb{P}_{M_1}(R_k|A_k = 1, Y_{k-1}) = \mathbb{P}_{M_2}(R_k|A_k = 1, Y_{k-1})$ and

$$\frac{\mathbb{P}_{M_1}(R_k|A_k = 2, Y_{k-1})}{\mathbb{P}_{M_2}(R_k|A_k = 2, Y_{k-1})} \le \max\left\{\frac{1/2}{1/2+\alpha}, \frac{1/2}{1/2-\alpha}\right\} = \frac{1}{1-2\alpha}$$

the likelihood ratio of Y_∞ is upper bounded by $(1+2\alpha)^N$ if the second action has been chosen at most N times. Hence

$$\begin{aligned}\mathbb{P}_{M_2}\left[n_2 \le N\right] = \frac{(1-2\alpha)^N}{(1-2\alpha)^N}\mathbb{E}_{M_2}\left[\mathbb{I}\{n_2 \le N\}\right] \ge (1-2\alpha)^N \mathbb{E}_{M_2}\left[\frac{\mathbb{P}_{M_1}(Y_\infty)}{\mathbb{P}_{M_2}(Y_\infty)}\mathbb{I}\{n_2 \le N\}\right] \\ \ge (1-2\alpha)^N \beta > 0\end{aligned}$$

Therefore, the regret for M_2 is for T large enough $\mathbb{E}_{M_2} R(T) \ge (T - N)\beta(1-2\alpha)^N \alpha/2 = O(T)$. Hence, for the algorithm to ensure sublinear regret for M_2, it has to play the suboptimal action for M_1 infinitely often with probability 1. This however implies that the algorithm cannot satisfy any finite PAC bound for accuracy $\varepsilon < \alpha/2$. □

A.2 Proof of Theorem 2

Proof. **PAC Bound to high-probability regret bound:** Consider a fixed $\delta > 0$ and PAC bound with $F_{\text{PAC}} = \Theta(1/\varepsilon^2)$. Then there is a $C > 0$ such that the following algorithm satisfies the PAC bound. The algorithm uses the worst possible policy with optimality gap H in all episodes on some event E and in the first C/ε^2 episodes on the complimentary event E^C. For the remaining episodes on E^C it follows a policy with optimality gap ε. The probability of E is δ. The regret of the algorithm on E is $R(T) = TH$ and on E^C it is $R(T) = \min\{T, C/\varepsilon^2\}H + \min\{T - C/\varepsilon^2, 0\}\varepsilon$. For $T \ge C/\varepsilon^2$, on any event the regret of this algorithm is at least

$$R(T) = \frac{CH}{\varepsilon^2} + \left(T - \frac{C}{\varepsilon^2}\right)\varepsilon = T\varepsilon + \frac{C(H-\varepsilon)}{\varepsilon^2}. \tag{8}$$

The quantity

$$\frac{R(T)}{T^{2/3}} = \frac{C(H-\varepsilon)}{T^{2/3}\varepsilon^2} + \varepsilon T^{1/3}$$

takes its minimum at $T = \frac{C(H-\varepsilon)}{\varepsilon^3}$ with a positive value and hence $R(T) = \Omega(T^{2/3})$. Therefore a PAC bound with rate $1/\varepsilon^2$ implies at best a high-probability regret bound of order $O(T^{2/3})$ and is only tight at $T = \Theta(1/\varepsilon^3)$. Furthermore, by looking at Equation (8), we see that for any fixed ε, there is an algorithm that has uniform high-probability regret that is $\Omega(T)$.

PAC Bound to uniform high-probability regret bound: Consider a fixed $\delta > 0$ and $\varepsilon > 0$ and a PAC bound F_{PAC} that evaluates to some value N for parameter ε. The algorithm uses the worst possible policy with optimality gap H in all episodes on some event E and in the first N episodes on

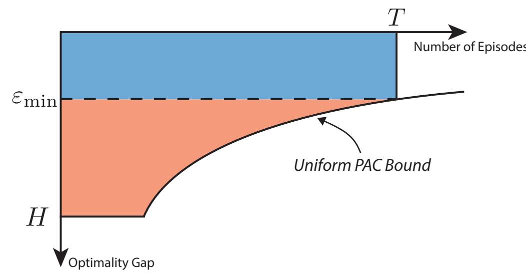

Figure 3: Relation of PAC-bound and Regret; The area of the shaded regions are a bound on the regret after T episodes.

the complimentary event E^C. For the remaining episodes on E^C it follows a policy with optimality gap ε. The probability of E is δ. The regret of the algorithm on E is $R(T) = TH$ and on E^C it is $R(T) = \min\{T, N\}H + \min\{T - N, 0\}\varepsilon$. For $T \geq N$, on any event the regret of this algorithm is at least

$$R(T) = NH + (T - N)\,\varepsilon = T\varepsilon + H(T - N) = \Omega(T).$$

Uniform high-probability regret bound to PAC bound: Consider an MDP such that at least one suboptimal policy exists with optimality gap $\varepsilon > 0$. Further let $L(T)$ be a nondecreasing function with $F_{\text{UHPR}}(T) \geq L(T)$ and $L(T) \to \infty$ as $T \to \infty$. Then the algorithm plays the optimal policy except for episodes k where $\lfloor L(k-1)/\varepsilon \rfloor \neq \lfloor L(k)/\varepsilon \rfloor$. This algorithm satisfies the regret bound but makes infinitely many $\varepsilon/2$-mistakes with probability 1.

Uniform high-probability regret bound to expected regret bound: Consider an MDP such that at least one suboptimal policy exists with optimality gap $\varepsilon > 0$. Consider an algorithm that with probability δ always plays the suboptimal policy and with probability $1 - \delta$ always plays the optimal policy. This algorithm satisfies the uniform high-probability regret bound but suffers regret $\mathbb{E}R(T) = \delta\varepsilon T = \Omega(T)$. $\square$

A.3 Proof of Theorem 3

Proof. **Convergence to optimal policies:** The convergence to the set of optimal policies follows directly by using the definition of limits on the Δ_k sequence for each outcome in the high-probability event where the bound holds.
(ε, δ)**-PAC:** Due to sub-additivity of probabilities, we have

$$\mathbb{P}\left(N_\varepsilon > F_{\text{PAC}}\left(\frac{1}{\varepsilon}, \log\frac{1}{\delta}\right)\right) \leq \mathbb{P}\left(\bigcup_{\varepsilon'}\left\{N_{\varepsilon'} > F_{\text{PAC}}\left(\frac{1}{\varepsilon'}, \log\frac{1}{\delta}\right)\right\}\right)$$
$$= \mathbb{P}\left(\exists \varepsilon' \,:\, N_{\varepsilon'} > F_{\text{PAC}}\left(\frac{1}{\varepsilon'}, \log\frac{1}{\delta}\right)\right) \leq \delta.$$

High-Probability Regret Bound: This part is proved separately in Theorem A.1 below. $\square$

Theorem A.1 (Uniform-PAC to Regret Conversion Theorem)**.** *Assume on some event E an algorithm follows for all ε an ε-optimal policy π_k, i.e., $\Delta_k \leq \varepsilon$, on all but at most*

$$\frac{C_1}{\varepsilon}\left(\ln\frac{C_3}{\varepsilon}\right)^k + \frac{C_2}{\varepsilon^2}\left(\ln\frac{C_3}{\varepsilon}\right)^{2k}$$

episodes where $C_1 \geq C_2 \geq 2$ and $C_3 \geq \max\{H, e\}$ and C_1, C_2, C_3 do not depend on ε . Then this algorithm has on this event a regret of

$$R(T) \leq (\sqrt{C_2 T} + C_1)\,\mathrm{polylog}(T, C_3, C_1) = O(\sqrt{C_2 T}\,\mathrm{polylog}(T, C_3, C_1, H))$$

for all number of episodes T.

Proof. The mistake bound $g(\varepsilon) = \frac{C_1}{\varepsilon}\left(\ln\frac{C_3}{\varepsilon}\right)^k + \frac{C_2}{\varepsilon^2}\left(\ln\frac{C_3}{\varepsilon}\right)^{2k} \leq T$ is monotonically decreasing for $\varepsilon \in (0, H]$. For a given T large enough, we can therefore find an $\varepsilon_{\min} \in (0, H]$ such that $g(\varepsilon) \leq T$ for all $\varepsilon \in (\varepsilon_{\min}, H]$. The regret $R(T)$ of the algorithm can then be bounded as follows

$$R(T) \leq T\varepsilon_{\min} + \int_{\varepsilon_{\min}}^{H} g(\varepsilon)d\varepsilon.$$

This bound assumes the worst case where first the algorithm makes the worst mistakes possible with regret H and subsequently less and less severe mistakes controlled by the mistake bound. For a better intuition, see Figure 3.

We first find a suitable $\varepsilon_{\min}$. Define $y = \frac{1}{\varepsilon}\left(\ln\frac{C_3}{\varepsilon}\right)^k$ then since g is monotonically decreasing, it is sufficient to find a ε with $g(\varepsilon) \leq T$. That is equivalent to $C_1 y + C_2 y^2 \leq T$ for which

$$\frac{1}{\varepsilon}\left(\ln\frac{C_3}{\varepsilon}\right)^k = y \leq \frac{C_1}{2C_2} + \frac{\sqrt{C_1^2 + 4TC_2}}{2C_2} =: a$$

is sufficient. We set now

$$\varepsilon_{\min} = \frac{\ln(C_3 a)^k}{a} = \frac{2C_2}{C_1 + \sqrt{C_1^2 + 4TC_2}}\left(\ln\frac{(C_1 + \sqrt{C_1^2 + 4TC_2})C_3}{2C_2}\right)^k$$

which is a valid choice as

$$\begin{aligned}
\frac{1}{\varepsilon_{\min}}\left(\ln\frac{C_3}{\varepsilon_{\min}}\right)^k =& \frac{a}{\ln(C_3 a)^k}\left(\ln\frac{C_3 a}{\ln(C_3 a)^k}\right)^k = \frac{a}{\ln(C_3 a)^k}\left(\ln(C_3 a) - k\ln\ln(C_3 a)\right)^k \\
\leq& \frac{a}{\ln(C_3 a)^k}\left(\ln(C_3 a)\right)^k = a.
\end{aligned}$$

We now first bound the regret further as

$$\begin{aligned}
R(T) \leq& T\varepsilon_{\min} + \int_{\varepsilon_{\min}}^{H} g(\varepsilon)d\varepsilon \leq T\varepsilon_{\min} + C_1\left(\ln\frac{C_3}{\varepsilon_{\min}}\right)^k \int_{\varepsilon_{\min}}^{H}\frac{1}{\varepsilon}d\varepsilon + C_2\left(\ln\frac{C_3}{\varepsilon_{\min}}\right)^{2k}\int_{\varepsilon_{\min}}^{H}\frac{1}{\varepsilon^2}d\varepsilon \\
=& T\varepsilon_{\min} + C_1\left(\ln\frac{C_3}{\varepsilon_{\min}}\right)^k \ln\frac{H}{\varepsilon_{\min}} + C_2\left(\ln\frac{C_3}{\varepsilon_{\min}}\right)^{2k}\left[\frac{1}{\varepsilon_{\min}} - \frac{1}{H}\right]
\end{aligned}$$

and then use the choice of $\varepsilon_{\min}$ from above to look at each of the terms in this bound individually. In the following bounds we extensively use the fact $\ln(a+b) \leq \ln(a) + \ln(b) = \ln(ab)$ for all $a, b \geq 2$ and that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ which holds for all $a, b \geq 0$.

$$\begin{aligned}
T\varepsilon_{\min} =& \frac{2TC_2}{C_1 + \sqrt{C_1^2 + 4TC_2}}\left(\ln\frac{C_3(C_1 + \sqrt{C_1^2 + 4TC_2})}{2C_2}\right)^k \\
\leq& \frac{2TC_2}{\sqrt{4TC_2}}\left(\ln C_3 + \ln C_1 + \ln C_1 + \ln\frac{2\sqrt{TC_2}}{2C_2}\right)^k \\
\leq& \sqrt{TC_2}\left(\ln(C_3 C_1^2\sqrt{T})\right)^k
\end{aligned}$$

Now for a $C \geq 0$ we first look at

$$\begin{aligned}
\ln\frac{C}{\varepsilon_{\min}} =& \ln C + \ln\frac{C_1 + \sqrt{C_1^2 + 4TC_2}}{2C_2} - k\ln\ln\frac{C_3(C_1 + \sqrt{C_1^2 + 4TC_2})}{2C_2} \\
\leq& \ln C + \ln\frac{C_1 + \sqrt{C_1^2 + 4TC_2}}{2C_2} \\
\leq& \ln C + \ln C_1 + \ln C_1 + \ln\frac{\sqrt{4TC_2}}{2C_2} \\
\leq& \ln(CC_1^2\sqrt{T})
\end{aligned}$$

where the first inequality follows from the fact that $\frac{C_3(C_1+\sqrt{C_1^2+4TC_2})}{2C_2} \geq \frac{C_3 2C_1}{2C_2} \geq e$. Hence, we can bound

$$C_1 \left(\ln \frac{C_3}{\varepsilon_{\min}}\right)^k \ln \frac{H}{\varepsilon_{\min}} \leq C_1 \left(\ln(C_3 C_1^2 \sqrt{T})\right)^k \ln(HC_1^2\sqrt{T}).$$

Now since

$$\frac{1}{\varepsilon_{\min}} = \frac{C_1 + \sqrt{C_1^2 + 4TC_2}}{2C_2} \left(\ln \frac{C_3(C_1 + \sqrt{C_1^2 + 4TC_2})}{2C_2}\right)^{-k} \leq \frac{C_1}{C_2} + \sqrt{\frac{T}{C_2}}$$

we get

$$\begin{aligned} C_2 \left(\ln \frac{C_3}{\varepsilon_{\min}}\right)^{2k} \left[\frac{1}{\varepsilon_{\min}} - \frac{1}{H}\right] &\leq C_2 \left(\ln(C_3 C_1^2 \sqrt{T})\right)^{2k} \left[\frac{C_1}{C_2} + \sqrt{\frac{T}{C_2}}\right] \\ &\leq \left(\ln(C_3 C_1^2 \sqrt{T})\right)^{2k} \left[C_1 + \sqrt{TC_2}\right]. \end{aligned}$$

As a result we can conclude that $R(T) \leq (\sqrt{C_2 T} + C_1) \operatorname{polylog}(T, C_3, C_1, H) = O(\sqrt{C_2 T} \operatorname{polylog}(T, C_3, C_1, H))$. ☐

B Experimental Details

We generated the MDPs with $S = 5, 50, 200$ states, $A = 3$ actions and $H = 10$ timesteps as follows: The transition probabilities $P(s, a, t)$ were sampled independently from Dirichlet $\left(\frac{1}{10}, \dots \frac{1}{10}\right)$ and the rewards were all deterministic with their value $r(s, a, t)$ set to 0 with probability 85% and set uniformly at random in $[0, 1]$ otherwise. This construction results in MDPs that have concentrated but non-deterministic transition probabilities and sparse rewards.

Since some algorithms have been proposed assuming the rewards $r(s, a, t)$ are known and we aim for a fair comparison, we assumed for all algorithms that the immediate rewards $r(s, a, t)$ are known and adapted the algorithms accordingly. For example, in UBEV, the $\min\left\{1, \frac{l(s,a,t)}{\max\{1, n(s,a,t)\}} + \phi\right\}$ term was replaced by the true known rewards $r(s, a, t)$ and the δ parameter in ϕ was scaled by $9/7$ accordingly since the concentration result for immediate rewards is not necessary in this case. We used $\delta = \frac{1}{10}$ for all algorithms and $\varepsilon = \frac{1}{10}$ if they require to know ε beforehand.

We adapted MoRMax, UCRL2, UCFH, MBIE, MedianPAC, Delayed Q-Learning and OIM to the episodic MDP setting with time-dependent transition dynamics by using allowing them to learn time-dependent dynamics and use finite-horizon planning. We did adapt the confidence intervals and but did not re-derive the constants for each algorithm. When in doubt we opted for smaller constants typically resulting better performance of the competitors. We further replaced the range of the value function $O(H)$ by the observed range of the optimistic next state values in the confidence bounds. We also reduced the number of episodes used in the delays by a factor of $\frac{1}{1000}$ for MoRMax and Delayed Q-Learning and by 10^{-6} for UCFH because they would otherwise not have performed a single policy update even for $S = 5$ within the 10 million episodes we considered. This scaling violates their theoretical guarantees but at least shows that the methods work in principle.

The performance reported in Figure 2 are the expected return of the current policy of each algorithm averaged over 1000 episodes. The figure shows a single run of the same randomly generated MDP but the results are representative. We reran this experiments with different random seeds and consistently obtained qualitatively similar results.

Source code for the experiments including concise but efficient implementations of the algorithms is available at `https://github.com/chrodan/FiniteEpisodicRL.jl`.

C PAC Lower Bound

Theorem C.1. *There exist positive constants c, $\delta_0 > 0$, $\varepsilon_0 > 0$ such that for every $\varepsilon \in (0, \varepsilon_0)$, $S \geq 4$, $A \geq 2$ and for every algorithm A that and $n \leq \frac{cASH^3}{\varepsilon^2}$ there is a fixed-horizon episodic MDP*

M_{hard} with time-dependent transition probabilities and S states and A actions so that returning an ε-optimal policy after n episodes is at most $1-\delta_0$. That implies that no algorithm can have a PAC guarantee better than $\Omega\left(\frac{ASH^3}{\varepsilon^2}\right)$ for sufficiently small ε.

Note that this lower bound on the sample complexity of any method in episodic MDPs with time-dependent dynamics applies to the arbitrary but fixed ε PAC bound and therefore immediately to the stronger uniform-PAC bounds. This theorem can be proved in the same way as Theorem 5 by Jiang et al. [4], which itself is a standard construction involving a careful layering of difficult instances of the multi-armed bandit problem.[4] For simplicity, we omitted the dependency on the failure probability δ, but using the techniques in the proof of Theorem 26 by Strehl et al. [5], a lower bound of order $\Omega\left(\frac{ASH^3}{\varepsilon^2}\log(SA/\delta)\right)$ can be obtained. The lower bound shows for small ε the sample complexity of UBEV given in Theorem 4 is optimal except for a factor of H and logarithmic terms.

D Planning Problem of UBEV

Lemma D.1 (Planning Problem)**.** *The policy update in Lines 3–9 of Algorithm 1 finds an optimal solution to the optimization problem*

$$\begin{aligned}
&\max_{P',V',\pi',r'} \mathbb{E}_{s\sim p_0}[V_1'(s)]\\
&\forall s\in\mathcal{S}, a\in\mathcal{A}, t\in[H]:\\
&\quad V'_{H+1}=0, \qquad P'(s,a,t)\in\Delta_S, \qquad r'(s,a,t)\in[0,1]\\
&\quad V'_t(s)=r'(s,\pi'(s,t),t)+\mathbb{E}_{s'\sim P'(s,\pi'(s,t),t)}[V'_{t+1}]\\
&\quad |(P'(s,a,t)-\hat{P}_k(s,a,t))^\top V'_{t+1}|\le\phi(s,a,t)(H-t)\\
&\quad |r'(s,a,t)-\hat{r}_k(s,a,t)|\le\phi(s,a,t)
\end{aligned}$$

where $\phi(s,a,t)=\sqrt{\frac{2\operatorname{llnp}(n(s,a,t))+\ln(18SAH/\delta)}{n(s,a,t)}}$ is a confidence bound and $\hat{P}_k(s'|s,a,t)=m(s',s,a,t)/n(s,a,t)$ are the empirical transition probabilities and $\hat{r}_k(s,a,t)=l(s,a,t)/n(s,a,t)$ the empirical average rewards.

Proof. Since $\tilde{V}_{H+1}(\cdot)$ is initialized with 0 and never changed, we immediately get that it is an optimal value for $V'_{H+1}(\cdot)$ which is constrained to be 0. Consider now a single time step t and assume V'_{t+1} are fixed to the optimal values $\tilde{V}_{t+1}$. Plugging in the computation of $Q(a)$ into the computation of $\tilde{V}_t(s)$, we get

$$\begin{aligned}
\tilde{V}_t(s)=\max_a Q(a)=\max_{a\in\mathcal{A}}\Big[&\min\{1,\hat{r}(s,a,t)+\phi(s,a,t)\}\\
&+\min\Big\{\max\tilde{V}_{t+1},\mathbb{I}\{n(s,a,t)>0\}(\hat{P}(s,a,t)^\top\tilde{V}_{t+1})+\phi(s,a,t)(H-t)\Big\}\Big]
\end{aligned}$$

using the convention that $\hat{r}(s,a,t)=0$ if $n(s,a,t)=0$. Assuming that $V'_{t+1}=\tilde{V}_{t+1}$, and that our goal for now is to maximize $\tilde{V}_t(s)$, this can be rewritten as

$$\begin{aligned}
\max_{P'(s,a,t),r'(s,a,t)}\tilde{V}_t(s)=&\max_{P'(s,a,t),r'(s,a,t),\pi'(s,t)}\left[r'(s,\pi'(s,t),t)+P'(s,\pi'(s,t),t)^\top\tilde{V}_{t+1}\right]\\
\text{s.t.}\qquad &\forall a\in\mathcal{A}: r'(s,a,t)\in[0,1], \qquad P'(s,a,t)\in\Delta_S\\
|(P'(s,a,t)-\hat{P}_k(s,a,t))^\top V'_{t+1}|\le&\phi(s,a,t)(H-t)\\
|r'(s,a,t)-\hat{r}_k(s,a,t)|\le&\phi(s,a,t)
\end{aligned}$$

since in this problem either $P'(s,\pi'(s,t),t)^\top\tilde{V}_{t+1}=\hat{P}(s,\pi'(s,t),t)^\top\tilde{V}_{t+1}+\phi(s,a,t)(H-t)$ if that does not violate $P'(s,\pi'(s,t),t)^\top\tilde{V}_{t+1}\le\max\tilde{V}_{t+1}$ and otherwise $P'(s',s,\pi'(s,t),t)=1$

[4]We here only use $H/2$ timesteps for bandits and the remaining $H/2$ time steps to accumulate a reward of $O(H)$ for each bandit

for one state s' with $\tilde{V}_{t+1}(s') = \max \tilde{V}_{t+1}$. Similarly, either $r'(s, \pi'(s,t), t) = \hat{r}(s, \pi'(s,t), t) + \phi(s, \pi'(s,t), t)$ if that does not violate $r'(s, \pi'(s,t), t) \leq 1$ or $r'(s, \pi'(s,t), t) = 1$ otherwise. Using induction for $t = H, H-1 \ldots 1$, we see that UBEV computes an optimal solution to

$$
\begin{aligned}
&\max_{P', V', \pi', r'} \; V_1'(\tilde{s}) \\
&\forall s \in \mathcal{S}, a \in \mathcal{A}, t \in [H]: \\
&\quad V_{H+1}' = 0, \qquad P'(s,a,t) \in \Delta_S, \qquad r'(s,a,t) \in [0,1] \\
&\quad V_t'(s) = r'(s, \pi'(s,t), t) + \mathbb{E}_{s' \sim P'(s, \pi'(s,t), t)}[V_{t+1}'] \\
&\quad |(P'(s,a,t) - \hat{P}_k(s,a,t))^\top V_{t+1}'| \leq \phi(s,a,t)(H-t) \\
&\quad |r'(s,a,t) - \hat{r}_k(s,a,t)| \leq \phi(s,a,t)
\end{aligned}
$$

for any fixed $\tilde{s}$. The intersection of all optimal solutions to this problem for all $\tilde{s} \in \mathcal{S}$ are also an optimal solution to

$$
\begin{aligned}
&\max_{P', V', \pi', r'} \; p_0^\top V_1' \\
&\forall s \in \mathcal{S}, a \in \mathcal{A}, t \in [H]: \\
&\quad V_{H+1}' = 0, \qquad P'(s,a,t) \in \Delta_S, \qquad r'(s,a,t) \in [0,1] \\
&\quad V_t'(s) = r'(s, \pi'(s,t), t) + \mathbb{E}_{s' \sim P'(s, \pi'(s,t), t)}[V_{t+1}'] \\
&\quad |(P'(s,a,t) - \hat{P}_k(s,a,t))^\top V_{t+1}'| \leq \phi(s,a,t)(H-t) \\
&\quad |r'(s,a,t) - \hat{r}_k(s,a,t)| \leq \phi(s,a,t).
\end{aligned}
$$

Hence, UBEV computes an optimal solution to this problem. □

E Details of PAC Analysis

In the analysis, we denote the value of $n(\cdot, t)$ after the planning in iteration k as $n_{tk}(\cdot)$. We further denote by $P(s'|s,a,t)$ the probability of sampling state s' as s_{t+1} when $s_t = s, a_t = a$. With slight abuse of notation, $P(s,a,t) \in [0,1]^S$ denotes the probability vector of $P(\cdot|s,a,t)$. We further use $\tilde{P}_k(s'|s,a,t)$ as conditional probability of $s_{t+1} = s'$ given $s_t = s, a_t = a$ but in the optimistic MDP $\tilde{M}$ computed in the optimistic planning steps in iteration k. We also use the following definitions:

$$
\begin{aligned}
w_{\min} =& w'_{\min} = \frac{\varepsilon c_\varepsilon}{H^2 S} \\
c_\varepsilon =& \frac{1}{3} \\
L_{tk} =& \{(s,a) \in \mathcal{S} \times \mathcal{A} \,:\, w_{tk}(s,a) \geq w_{\min}\} \\
\operatorname{llnp}(x) =& \ln(\ln(\max\{x, e\})) \\
\operatorname{rng}(x) =& \max(x) - \min(x) \\
\delta' =& \frac{\delta}{9}
\end{aligned}
$$

In the following, we provide the formal proof for Theorem 4 and then present all necessary lemmas:

E.1 Proof of Theorem 4

Proof of Theorem 4. Corollary E.5 ensures that the failure event has probability at most δ. Outside the failure event Lemma E.2 ensures that all but at most $\frac{48A^2S^3H^4}{\varepsilon}\operatorname{polylog}(A,S,H,1/\varepsilon,1/\delta)$ episodes are friendly. Finally, Lemma E.8 shows that all friendly episodes except at most $\left(\frac{9216}{\varepsilon} + 417S\right)\frac{ASH^4}{\varepsilon}\operatorname{polylog}(A,S,H,1/\varepsilon,1/\delta)$ are ε-optimal. The second bound follows from replacing AS^2 by $1/\varepsilon$ in the second term. Furthermore, outside the failure event Lemma E.2 ensures that all but at most $\frac{6AS^2H^3}{\varepsilon}\operatorname{polylog}(A,S,H,1/\varepsilon,1/\delta)$ episodes are nice. Finally, Lemma E.7 shows that all nice episodes except at most $(4+S)\,576\frac{ASH^4}{\varepsilon}\operatorname{polylog}(A,S,H,1/\varepsilon,1/\delta)$ are ε-optimal.

□

E.2 Failure Events and Their Probabilities

In this section, we define a failure event F in which we cannot guarantee the performance of UBEV. We then show that this event F only occurs with low probability. All our arguments are based on general uniform concentration of measure statements that we prove in Section F. In the following we argue how the apply in our setting and finally combine all concentration results to get $\mathbb{P}(F) \leq \delta$. The failure event is defined as

$$F = \bigcup_k \left[F_k^N \cup F_k^{CN} \cup F_k^P \cup F_k^V \cup F_k^{L1} \cup F_k^R\right]$$

where

$$F_k^N = \left\{ \exists s, a, t : \, n_{tk}(s,a) < \frac{1}{2} \sum_{i<k} w_{ti}(s,a) - \ln \frac{SAH}{\delta'} \right\}$$

$$F_k^{CN} = \left\{ \exists s, a, s', a', u < t : \, n_{tk}(s,a) < \frac{1}{2} n_{uk}(s',a') \sum_{i<k} w_{ui}^t(s,a|s',a') - \ln \left(\frac{S^2A^2H^2}{\delta'} \right) \right\}$$

$$F_k^V = \left\{ \exists s, a, t : \, |(\hat{P}_k(s,a,t) - P(s,a,t))^\top V_{t+1}^\star| \geq \sqrt{\frac{\operatorname{rng}(V_{t+1}^\star)^2}{n_{tk}(s,a)} \left(2 \operatorname{llnp}(n_{tk}(s,a)) + \ln \frac{3SAH}{\delta'} \right)} \right\}$$

$$F_k^P = \left\{ \exists s, s', a, t : \, |\hat{P}_k(s'|s,a,t) - P(s'|s,a,t)| \geq \sqrt{\frac{2P(s'|s,a,t)}{n_{tk}(s,a)} \left(2 \operatorname{llnp}(n_{tk}(s,a)) + \ln \frac{3S^2AH}{\delta'} \right)} \right.$$

$$\left. + \frac{1}{n_{tk}(s,a)} \left(2 \operatorname{llnp}(n_{tk}(s,a)) + \ln \frac{3S^2AH}{\delta'} \right) \right\}$$

$$F_k^{L1} = \left\{ \exists s, a, t : \, \|\hat{P}_k(s,a,t) - P(s,a,t)\|_1 \geq \sqrt{\frac{4}{n_{tk}(s,a)} \left(2 \operatorname{llnp}(n_{tk}(s,a)) + \ln \frac{3SAH(2^S-2)}{\delta'} \right)} \right\}$$

$$F_k^R = \left\{ \exists s, a, t : \, |\hat{r}_k(s,a,t) - r(s,a,t)| \geq \sqrt{\frac{1}{n_{tk}(s,a)} \left(2 \operatorname{llnp}(n_{tk}(s,a)) + \ln \frac{3SAH}{\delta'} \right)} \right\}.$$

We now bound the probability of each type of failure event individually:

Corollary E.1. *For any $\delta' > 0$, it holds that $\mathbb{P}\left(\bigcup_{k=1}^\infty F_k^V\right) \leq 2\delta'$ and $\mathbb{P}\left(\bigcup_{k=1}^\infty F_k^R\right) \leq 2\delta'$*

Proof. Consider a fix $s \in \mathcal{S}, a \in \mathcal{A}, t \in [H]$ and denote $\mathcal{F}_k$ the sigma-field induced by the first $k-1$ episodes and the k-th episode up to s_t and a_t but not s_{t+1}. Define τ_i to be the index of the episode where (s,a) was observed at time t the ith time. Note that τ_i are stopping times with respect to $\mathcal{F}_i$. Define now the filtration $\mathcal{G}_i = \mathcal{F}_{\tau_i} = \{A \in \mathcal{F}_\infty \, : \, A \cap \{\tau_i \leq t\} \in \mathcal{F}_t \; \forall \, t \geq 0\}$ and $X_k = (V_{t+1}^\star(s_k') - P(s,a,t)^\top V_{t+1}^\star)\mathbb{I}\{\tau_k < \infty\}$ where s_i' is the value of s_{t+1} in episode τ_i (or arbitrary, if $\tau_i = \infty$).

By the Markov property of the MDP, we have that X_i is a martingale difference sequence with respect to the filtration $\mathcal{G}_i$. Further, since $\mathbb{E}[X_i|\mathcal{G}_{i-1}] = 0$ and $|X_i| \in [0, \operatorname{rng}(V_{t+1}^\star)]$, X_i conditionally $\operatorname{rng}(V_{t+1}^\star)/2$-subgaussian due to Hoeffding's Lemma, i.e., satisfies $\mathbb{E}[\exp(\lambda X_i)|\mathcal{G}_{i-1}] \leq \exp(\lambda^2 \operatorname{rng}(V_{t+1}^\star)^2/2)$.

We can therefore apply Lemma F.1 and conclude that

$$\mathbb{P}\left(\exists k : |(\hat{P}_k(s,a,t) - P(s,a,t))^\top V_{t+1}^\star| \geq \sqrt{\frac{\operatorname{rng}(V_{t+1}^\star)^2}{n_{tk}(s,a)} \left(2 \operatorname{llnp}(n_{tk}(s,a)) + \ln \frac{3}{\delta'} \right)} \right) \leq 2\delta' .$$

Analogously

$$\mathbb{P}\left(\exists k : |\hat{r}_k(s,a,t) - r(s,a,t)| \geq \sqrt{\frac{1}{n_{tk}(s,a)} \left(2 \operatorname{llnp}(n_{tk}(s,a)) + \ln \frac{3}{\delta'} \right)} \right) \leq 2\delta' .$$

Applying the union bound over all $s \in \mathcal{S}, a \in \mathcal{A}$ and $t \in [H]$, we obtain the desired statement for F^V. In complete analogy using the same filtration, we can show the statement for F^R. $\square$

Corollary E.2. *For any $\delta' > 0$, it holds that $\mathbb{P}\left(\bigcup_{k=1}^{\infty} F_k^P\right) \leq 2\delta'$.*

Proof. Consider first a fix $s', s \in \mathcal{S}$, $t \in [H]$ and $a \in \mathcal{A}$. Let K denote the number of times the triple s, a, t was encountered in total during the run of the algorithm. Define the random sequence X_i as follows. For $i \leq K$, let X_i be the indicator of whether s' was the next state when s, a, t was encountered the ith time and for $i > K$, let $X_i \sim \text{Bernoulli}(P(s'|s,a,t))$ be drawn i.i.d. By construction this is a sequence of i.i.d. Bernoulli random variables with mean $P(s'|s,a,t)$. Further the event

$$\bigcup_k \left\{ \left|\hat{P}_k(s'|s,a,t) - P(s'|s,a,t)\right| \geq \sqrt{\frac{2P(s'|s,a,t)}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n(s,a,t)) + \ln\frac{3S^2AH}{\delta'}\right)} \right.$$
$$\left. + \frac{1}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a)) + \ln\frac{3S^2AH}{\delta'}\right) \right\}$$

is contained in the event

$$\bigcup_i \left\{ |\hat{\mu}_i - \mu| \geq \sqrt{\frac{2\mu}{i}\left(2\operatorname{llnp}(i) + \ln\frac{3}{\delta'}\right)} + \frac{1}{i}\left(2\operatorname{llnp}(i) + \ln\frac{3S^2AH}{\delta'}\right) \right\}$$

whose probability can be bounded by $2\delta'/S^2/A/H$ using Lemma F.2. The statement now follows by applying the union bound. $\square$

Corollary E.3. *For any $\delta' > 0$, it holds that $\mathbb{P}\left(\bigcup_{k=1}^{\infty} F_k^{L1}\right) \leq \delta'$*

Proof. Using the same argument as in the proof of Corollary E.2 the statement follows from Lemma F.3. $\square$

Corollary E.4. *It holds that*

$$\mathbb{P}\left(\bigcup_k F_k^N\right) \leq \delta' \quad \textit{and} \quad \mathbb{P}\left(\bigcup_k F_k^{CN}\right) \leq \delta'.$$

Proof. Consider a fix $s \in \mathcal{S}, a \in \mathcal{A}, t \in [H]$. We define $\mathcal{F}_k$ to be the sigma-field induced by the first $k-1$ episodes and X_k as the indicator whether s, a, t was observed in episode k. The probability $w_{tk}(s,a)$ pf whether $X_k = 1$ is F_k measurable and hence we can apply Lemma F.4 with $W = \ln\frac{SAH}{\delta'}$ and obtain that $\mathbb{P}\left(\bigcup_k F_k^N\right) \leq \delta'$ after applying the union bound.

For the second statement, consider again a fix $s, s' \in \mathcal{S}, a, a' \in \mathcal{A}, u, t \in [H]$ with $u < t$ and denote by $\mathcal{F}_k$ the sigma-field induced by the first $k-1$ episodes and the k-th episode up to s_u and a_u but not s_{u+1}. Define τ_i to be the index of the episode where (s', a') was observed at time u the ith time. Note that τ_i are stopping times with respect to $\mathcal{F}_i$. Define now the filtration $\mathcal{G}_i = \mathcal{F}_{\tau_i} = \{A \in \mathcal{F}_\infty : A \cap \{\tau_i \leq k\} \in \mathcal{F}_k \, \forall\, k \geq 0\}$ and X_i to be the indicator whether s, a, t and s', a', u was observed in episode τ_i. If $\tau_i = \infty$, we set $X_i = 0$. Note that the probablity $w_{ui}^t(s, a|s', a')\mathbb{I}\{\tau_i < \infty\}$ of $X_i = 1$ is $\mathcal{G}_i$-measureable.

By the Markov property of the MDP, we have that X_i is a martingale difference sequence with respect to the filtration $\mathcal{G}_i$. We can therefore apply Lemma F.4 with $W = \ln\frac{S^2A^2H^2}{\delta'}$ and using the union bound over all s, a, s', a', u, t, we get $\mathbb{P}\left(\bigcup_k F_k^{CN}\right) \leq \delta'$. $\square$

Corollary E.5. *The total failure probability of the algorithm is bounded by $\mathbb{P}(F) \leq 9\delta' = \delta$.*

Proof. Statement follows directly from Corollary E.1, Corollary E.2, Corollary E.3, Corollary E.4 and the union bound. $\square$

E.3 Nice and Friendly Episodes

We now define the notion of *nice* and the stronger *friendly* episodes. In nice episodes, all states either have low probability of occuring or the sum of probability of occuring in the previous episodes is large enough so that outside the failure event we can guarantee that

$$n_{tk}(s,a) \geq \frac{1}{4}\sum_{i<k} w_{ti}(s,a).$$

This allows us to then bound the number of nice episodes by the number of times terms of the form

$$\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{\frac{\operatorname{llnp}(n_{tk}(s,a))+D}{n_{tk}(s,a)}}$$

can exceed a chosen threshold (see Lemma E.3 below). In the next section, we will bound the optimality gap of an episode by terms of such form and use the results derived here to bound the number of nice episodes where the algorithm can follow a ε-suboptimal policy. Together with a bound on the number of non-nice episodes, we obtain the sample complexity of UBEV shown in Theorem 4.

Similarly, we use a more refined analysis of the optimality gap of friendly episodes together with Lemma E.4 below to obtain the tighter sample complexity linear-polylog in S.

Definition 2 (Nice and Friendly Episodes)**.** *An episode k is* nice *if and only if for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in [H]$ the following two conditions hold:*

$$w_{tk}(s,a) \leq w_{\min} \quad \vee \quad \frac{1}{4}\sum_{i<k} w_{ti}(s,a) \geq \ln\frac{SAH}{\delta'}$$

An episode k is friendly *if and only if it is nice and for all $s, s' \in \mathcal{S}$, $a, a' \in \mathcal{A}$ and $u, t \in [H]$ with $u < t$ the following two conditions hold:*

$$w^t_{uk}(s,a|s',a') \leq w'_{\min} \quad \vee \quad \frac{1}{4}\sum_{i<k} w^t_{ui}(s,a|s',a') \geq \ln\frac{S^2A^2H^2}{\delta'}.$$

We denote the set of all nice episodes by $N \subseteq \mathbb{N}$ and the set of all friendly episodes by $K \subseteq N$.

Lemma E.1 (Properties of nice and friendly episodes)**.** *If an episode k is nice, i.e., $k \in N$, then on F^c (outside the failure event) for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in [H]$ with $u < t$ the following statement holds:*

$$w_{tk}(s,a) \leq w_{\min} \quad \vee \quad n_{tk}(s,a) \geq \frac{1}{4}\sum_{i<k} w_{ti}(s,a).$$

If an episode k is friendly, i.e., $k \in K$, then on F^c (outside the failure event) for all $s, s' \in \mathcal{S}$, $a, a' \in \mathcal{A}$ and $u, t \in [H]$ with $u < t$ the above statement holds as well as

$$w^t_{uk}(s,a|s',a') \leq w'_{\min} \quad \vee \quad n_{tk}(s,a) \geq \frac{1}{4}n_{uk}(s',a')\sum_{i<k} w^t_{ui}(s,a|s',a').$$

Proof. Since we consider the event $F_k^{N^c}$, it holds for all s, a, t triples with $w_{tk}(s,a) > w_{\min}$

$$n_{tk}(s,a) \geq \frac{1}{2}\sum_{i<k} w_{ti}(s,a) - \ln\frac{SAH}{\delta'} \geq \frac{1}{4}\sum_{i<k} w_{ti}(s,a)$$

for $k \in N$ Further, since we only consider the event $F_k^{CN^c}$,we have for all $s, s' \in \mathcal{S}$, $a, a' \in \mathcal{A}$, $u, t \in [H]$ with $u < t$ and $w^t_{uk}(s,a|s',a') > w_{\min}$

$$n_{tk}(s,a) \geq \frac{1}{2}n_{uk}(s',a')\sum_{i<k} w^t_{ui}(s,a|s',a') - \ln\frac{S^2A^2H^2}{\delta'}$$

for $k \in E$. If $n_{uk}(s', a') = 0$ then $n_{tk}(s,a) \geq 0 = \frac{1}{4} n_{uk}(s', a') \sum_{i<k} w^t_{ui}(s, a | s', a')$ holds trivially. Otherwise $n_{uk}(s', a') \geq 1$ and therefore

$$
\begin{aligned}
n_{tk}(s,a) \geq & \frac{1}{2} n_{uk}(s', a') \sum_{i<k} w^t_{ui}(s, a|s', a') - \ln \frac{S^2 A^2 H^2}{\delta'} \\
\geq & \frac{1}{2} n_{uk}(s', a') \sum_{i<k} w^t_{ui}(s, a|s', a') - \frac{1}{4} \sum_{i<k} w^t_{ui}(s, a|s', a') \\
\geq & \frac{1}{4} n_{uk}(s', a') \sum_{i<k} w^t_{ui}(s, a|s', a')
\end{aligned}
$$

□

Lemma E.2 (Number of non-nice and non-friendly episodes)**.** *On the good event F^c, the number of episodes that are not friendly is at most*

$$48 \frac{S^3 A^2 H^4}{\varepsilon} \ln \frac{S^2 A^2 H^2}{\delta'}$$

and the number episodes that are not nice is at most

$$\frac{6 S^2 A H^3}{\varepsilon} \ln \frac{SAH}{\delta'}.$$

Proof. If an episode k is not nice, then there is s, a, t with $w_{tk}(s,a) > w_{\min}$ and $\sum_{i<k} w_{ti}(s,a) < 4 \ln \frac{SAH}{\delta'}$. Since the sum on the left-hand side of this inequality increases by at least $w_{\min}$ when this happens and the right hand side stays constant, this situation can occur at most

$$\frac{4SAH}{w_{\min}} \ln \frac{SAH}{\delta'} = \frac{24 S^2 A H^3}{\varepsilon} \ln \frac{SAH}{\delta'}$$

times in total. If an episode k is not friendly, it is either not nice or there is s, a, t and s', a', u with $u < t$ and $w^t_{uk}(s', a'|s, a) > w'_{\min}$ and $\sum_{i<k} w^t_{ui}(s, a|s', a') < 4 \ln \frac{S^2 A^2 H^2}{\delta'}$. Since the sum on the left-hand side of this inequality increases by at least $w'_{\min}$ each time this happens while the right hand side stays constant, this can happen at most $\frac{4 S^2 A^2 H^2}{w'_{\min}} \ln \frac{S^2 A^2 H^2}{\delta'}$ times in total. Therefore, there can only be at most

$$
\begin{aligned}
& \frac{4SAH}{w_{\min}} \ln \frac{SAH}{\delta'} + \frac{4 S^2 A^2 H^2}{w'_{\min}} \ln \frac{S^2 A^2 H^2}{\delta'} \\
= & \frac{4 S^2 A H^3}{c_\varepsilon \varepsilon} \ln \frac{SAH}{\delta'} + \frac{4 S^3 A^2 H^4}{c_\varepsilon \varepsilon} \ln \frac{S^2 A^2 H^2}{\delta'} \leq \frac{48 S^3 A^2 H^4}{\varepsilon^2} \ln \frac{S^2 A^2 H^2}{\delta'}
\end{aligned}
$$

non-friendly episodes. □

Lemma E.3 (Main Rate Lemma)**.** *Let $r \geq 1$ fix and $C > 0$ which can depend polynomially on the relevant quantities and $\varepsilon' > 0$ and let $D \geq 1$ which can depend poly-logarithmically on the relevant quantities. Then*

$$\sum_t \sum_{s,a \in L_{tk}} w_{tk}(s,a) \left(\frac{C(\operatorname{llnp}(n_{tk}(s,a)) + D)}{n_{tk}(s,a)} \right)^{1/r} \leq \varepsilon'$$

on all but at most

$$\frac{8CASH^r}{\varepsilon'^r} \operatorname{polylog}(S, A, H, \delta^{-1}, \varepsilon'^{-1}).$$

nice episodes.

Proof. Define

$$\begin{aligned}\Delta_k =&\sum_t \sum_{s,a\in L_{tk}} w_{tk}(s,a)\left(\frac{C(\mathrm{llnp}(n_{tk}(s,a))+D)}{n_{tk}(s,a)}\right)^{1/r}\\ =&\sum_t \sum_{s,a\in L_{tk}} w_{tk}(s,a)^{1-\frac{1}{r}}\left(w_{tk}(s,a)\frac{C(\mathrm{llnp}(n_{tk}(s,a))+D)}{n_{tk}(s,a)}\right)^{1/r}.\end{aligned}$$

We first bound using Hölder's inequality

$$\Delta_k \le \left(\sum_t \sum_{s,a\in L_{tk}} \frac{CH^{r-1}w_{tk}(s,a)(\mathrm{llnp}(n_{tk}(s,a))+D)}{n_{tk}(s,a)}\right)^{\frac{1}{r}}.$$

Using the property in Lemma E.1 of nice episodes as well as the fact that $w_{tk}(s,a) \le 1$ and $\sum_{i<k} w_{ti}(s,a) \ge 4\ln\frac{SAH}{\delta'} \ge 4\ln(2) \ge 2$, we bound

$$n_{tk}(s,a) \ge \frac{1}{4}\sum_{i<k} w_{ti}(s,a) \ge \frac{1}{8}\sum_{i\le k} w_{ti}(s,a).$$

The function $\frac{\mathrm{llnp}(x)+D}{x}$ is monotonically decreasing in $x \ge 0$ since $D \ge 1$ (see Lemma E.6). This allows us to bound

$$\begin{aligned}\Delta_k^r \le&\sum_t \sum_{s,a\in L_{tk}} \frac{CH^{r-1}w_{tk}(s,a)(\mathrm{llnp}(n_{tk}(s,a))+D)}{n_{tk}(s,a)}\\ \le&8CH^{r-1}\sum_t \sum_{s,a\in L_{tk}} \frac{w_{tk}(s,a)\left(\mathrm{llnp}\left(\frac{1}{8}\sum_{i\le k} w_{ti}(s,a)\right)+D\right)}{\sum_{i\le k} w_{ti}(s,a)}\\ \le&8CH^{r-1}\sum_t \sum_{s,a\in L_{tk}} \frac{w_{tk}(s,a)\left(\mathrm{llnp}\left(\sum_{i\le k} w_{ti}(s,a)\right)+D\right)}{\sum_{i\le k} w_{ti}(s,a)}.\end{aligned}$$

Assume now $\Delta_k > \varepsilon'$. In this case the right-hand side of the inequality above is also larger than ε'^r and there is at least one (s,a,t) with $w_{tk}(s,a) > w_{\min}$ and

$$\begin{aligned}\frac{8CSAH^r\left(\mathrm{llnp}\left(\sum_{i\le k} w_{ti}(s,a)\right)+D\right)}{\sum_{i\le k} w_{ti}(s,a)} >&\varepsilon'^r\\ \Leftrightarrow \frac{\mathrm{llnp}\left(\sum_{i\le k} w_{ti}(s,a)\right)+D}{\sum_{i\le k} w_{ti}(s,a)} >&\frac{\varepsilon'^r}{8CSAH^r}.\end{aligned}$$

Let us denote $C' = \frac{8CASH^r}{\varepsilon'^r}$. Since $\frac{\mathrm{llnp}(x)+D}{x}$ is monotonically decreasing and $x = C'^2 + 3C'D$ satisfies $\frac{\mathrm{llnp}(x)+D}{x} \le \frac{\sqrt{x}+D}{x} \le \frac{1}{C'}$, we know that if $\sum_{i\le k} w_{ti}(s,a) \ge C'^2 + 3C'D$ then the above condition cannot be satisfied for s,a,t. Since each time the condition is satisfied, it holds that $w_{tk}(s,a) > w_{\min}$ and so $\sum_{i\le k} w_{ti}(s,a)$ increases by at least $w_{\min}$, it can happen at most

$$m \le \frac{ASH(C'^2+3C'D)}{w_{\min}}$$

times that $\Delta_k > \varepsilon'$. Define $K = \{k : \Delta_k > \varepsilon'\} \cap N$ and we know that $|K| \le m$. Now we consider the sum

$$\begin{aligned}\sum_{k\in K}\Delta_k^r \le&\sum_{k\in K} 8CH^{r-1}\sum_t \sum_{s,a\in L_{tk}} \frac{w_{tk}(s,a)\left(\mathrm{llnp}\left(\sum_{i\le k} w_{ti}(s,a)\right)+D\right)}{\sum_{i\le k} w_{ti}(s,a)}\\ \le&8CH^{r-1}\left(\mathrm{llnp}\left(C'^2+3C'D\right)+D\right)\sum_t \sum_{s,a\in L_{tk}} \sum_{k\in K} \frac{w_{tk}(s,a)}{\sum_{i\le k} w_{ti}(s,a)\mathbb{I}\{w_{ti}(s,a)\ge w_{\min}\}}\end{aligned}$$

For every (s,a,t), we consider the sequence of $w_{ti}(s,a) \in [w_{\min}, 1]$ with $i \in I = \{i \in \mathbb{N} : w_{ti}(s,a) \geq w_{\min}\}$ and apply Lemma E.5. This yields that

$$\sum_{k \in K} \frac{w_{tk}(s,a)}{\sum_{i \leq k} w_{ti}(s,a)\mathbb{I}\{w_{ti}(s,a) \geq w_{\min}\}} \leq 1 + \ln(m/w_{\min}) = \ln\left(\frac{me}{w_{\min}}\right)$$

and hence

$$\sum_{k \in K} \Delta_k^r \leq 8CASH^r \ln\left(\frac{me}{w_{\min}}\right)\left(\operatorname{llnp}\left(C'^2 + 3C'D\right) + D\right)$$

Since each element in K has to contribute at least ε'^r to this bound, we can conclude that

$$\sum_{k \in N} \mathbb{I}\{\Delta_k \geq \varepsilon'\} \leq \sum_{k \in K} \mathbb{I}\{\Delta_k \geq \varepsilon'\} \leq |K| \leq \frac{8CASH^r}{\varepsilon'^r} \ln\left(\frac{me}{w_{\min}}\right)\left(\operatorname{llnp}\left(C'^2 + 3C'D\right) + D\right).$$

Since $\ln\left(\frac{me}{w_{\min}}\right)\left(\operatorname{llnp}\left(C'^2 + 3C'D\right) + D\right)$ is $\operatorname{polylog}(S, A, H, \delta^{-1}, \varepsilon'^{-1})$, the proof is complete. $\square$

Lemma E.4 (Conditional Rate Lemma)**.** *Let $r \geq 1$ fix and $C > 0$ which can depend polynomially on the relevant quantities and $\varepsilon' > 0$ and let $D \geq 1$ which can depend poly-logarithmically on the relevant quantities. Further $T \subset [H]$ is a subset of time-indices with $u < t$ for all $t \in T$. Then*

$$\sum_{t \in T} \sum_{s,a \in L_k^{ut}} w_{uk}^t(s,a|s',a') \left(\frac{C(\operatorname{llnp}(n_{tk}(s,a)) + D)}{n_{tk}(s,a)}\right)^{1/r} \leq \varepsilon' \left(\frac{\operatorname{llnp}(n_{uk}(s',a') + D + 1}{n_{uk}(s',a')}\right)^{1/r}$$

on all but at most

$$\frac{8CAS|T|^r}{\varepsilon'^r} \operatorname{polylog}(S, A, H, \delta^{-1}, \varepsilon'^{-1}).$$

friendly episodes E.

Proof. The proof follows mainly the structure of Lemma E.3. For the sake of completeness, we still present all steps here. Define

$$\begin{aligned}
\Delta_k =& \sum_{t \in T} \sum_{s,a \in L_k^{ut}} w_{uk}^t(s,a|s',a') \left(\frac{C(\operatorname{llnp}(n_{tk}(s,a)) + D)}{n_{tk}(s,a)}\right)^{1/r} \\
=& \sum_{t \in T} \sum_{s,a \in L_k^{ut}} w_{uk}^t(s,a|s',a')^{1-1/r} \left(w_{uk}^t(s,a|s',a') \frac{C(\operatorname{llnp}(n_{tk}(s,a)) + D)}{n_{tk}(s,a)}\right)^{1/r}.
\end{aligned}$$

We first bound using Hölder's inequality

$$\Delta_k \leq \left(\sum_{t \geq u} \sum_{s,a \in L_k^{ut}} w_{uk}^t(s,a|s',a') \frac{C|T|^{r-1}(\operatorname{llnp}(n_{tk}(s,a)) + D)}{n_{tk}(s,a)}\right)^{\frac{1}{r}}$$

Using the property in Lemma E.1 of friendly episodes as well as the fact that $w_{uk}^t(s,a|s',a') \leq 1$ and $\sum_{i<k} w_{ui}^t(s,a|s',a') \geq 4 \ln \frac{S^2A^2H^2}{\delta'} \geq 4\ln(2) \geq 2$, we bound

$$n_{tk}(s,a) \geq \frac{1}{4} n_{uk}(s',a') \sum_{i<k} w_{ui}^t(s,a|s',a') \geq \frac{1}{8} n_{uk}(s',a') \sum_{i \leq k} w_{ui}^t(s,a|s',a').$$

The function $\frac{\operatorname{llnp}(x)+D}{x}$ is monotonically decreasing in $x \geq 0$ since $D \geq 1$ (see Lemma E.6). This allows us to bound

$$
\begin{aligned}
\Delta_k^r \leq & \sum_{t \in T} \sum_{s,a \in L_k^{ut}} w_{uk}^t(s,a|s',a') \frac{C|T|^{r-1}(\operatorname{llnp}(n_{tk}(s,a)) + D)}{n_{tk}(s,a)} \\
\leq & 8C|T|^{r-1} \sum_{t \in T} \sum_{s,a \in L_k^{ut}} \frac{w_{uk}^t(s,a|s',a')(\operatorname{llnp}\left(\frac{1}{8} n_{uk}(s',a') \sum_{i \leq k} w_{ui}^t(s,a|s',a')\right) + D)}{n_{uk}(s',a') \sum_{i \leq k} w_{ui}^t(s,a|s',a')} \\
\leq & 8C|T|^{r-1} \sum_{t \in T} \sum_{s,a \in L_k^{ut}} \frac{w_{uk}^t(s,a|s',a')(\operatorname{llnp}\left(\sum_{i \leq k} w_{ui}^t(s,a|s',a')\right) + \operatorname{llnp}(n_{uk}(s',a')) + D + 1)}{n_{uk}(s',a') \sum_{i \leq k} w_{ui}^t(s,a|s',a')},
\end{aligned}
$$

where for the last line we used the first and last property in Lemma E.6. For notational convenience, we will use $D' = D + 1 + \operatorname{llnp}(n_{uk}(s',a'))$. Assume now $\Delta_k > \varepsilon' \left(\frac{D'}{n_{uk}(s',a')}\right)^{1/r}$. In this case the right-hand side of the inequality above is also larger than $\varepsilon'^r \left(\frac{D'}{n_{uk}(s',a')}\right)$ and there is at least one (s,a,t) with $w_{uk}^t(s,a|s',a') > w_{\min}$ and

$$
\begin{aligned}
& \frac{8CSA|T|^r \left(\operatorname{llnp}\left(\sum_{i \leq k} w_{ui}^t(s,a|s',a')\right) + D'\right)}{\sum_{i \leq k} w_{ui}^t(s,a|s',a')} > D'\varepsilon'^r \\
\Leftrightarrow & \frac{\left(\operatorname{llnp}\left(\sum_{i \leq k} w_{ui}^t(s,a|s',a')\right) + D'\right)}{\sum_{i \leq k} w_{ui}^t(s,a|s',a')} > \frac{D'\varepsilon'^r}{8CSA|T|^r}.
\end{aligned}
$$

Let us denote $C' = \frac{8CAS|T|^r}{\varepsilon'^r}$. Since $\frac{\operatorname{llnp}(x)+D'}{x}$ is monotonically decreasing and $x = C'^2 + 3C'$ satisfies $\frac{\operatorname{llnp}(x)+D'}{x} \leq \frac{\sqrt{x}+D'}{x} \leq D' \frac{\sqrt{x}+1}{x} \leq \frac{D'}{C'}$, we know that if $\sum_{i \leq k} w_{ui}^t(s,a|s',a') \geq C'^2 + 3C'$ then the above condition cannot be satisfied for s,a,t. Since each time the condition is satisfied, it holds that $w_{uk}^t(s,a|s',a') > w_{\min}$ and so $\sum_{i \leq k} w_{ui}^t(s,a|s',a')$ increases by at least $w_{\min}$, it can happen at most

$$
m \leq \frac{AS|T|(C'^2 + 3C')}{w_{\min}}
$$

times that $\Delta_k > \varepsilon' \left(\frac{D'}{n_{uk}(s',a')}\right)^{1/r}$. Define $K = \left\{k : \Delta_k > \varepsilon' \left(\frac{D'}{n_{uk}(s',a')}\right)^{1/r}\right\} \cap E$ and we know that $|K| \leq m$. Now we consider the sum

$$
\begin{aligned}
\sum_{k \in K} \Delta_k^r \leq & \sum_{k \in K} 8C|T|^{r-1} \sum_{t \in T} \sum_{s,a \in L_k^{ut}} \frac{w_{uk}^t(s,a|s',a')(\operatorname{llnp}\left(\sum_{i \leq k} w_{ui}^t(s,a|s',a')\right) + D')}{n_{uk}(s',a') \sum_{i \leq k} w_{ui}^t(s,a|s',a')} \\
\leq & \frac{8C|T|^{r-1}(\operatorname{llnp}\left(C'^2 + 3C'\right) + D')}{n_{uk}(s',a')} \sum_{t \in T} \sum_{s,a \in L_k^{ut}} \sum_{k \in K} \frac{w_{uk}^t(s,a|s',a')}{\sum_{i \leq k} w_{ui}^t(s,a|s',a')} \\
\leq & \frac{8C|T|^{r-1} D'(\operatorname{llnp}\left(C'^2 + 3C'\right) + 1)}{n_{uk}(s',a')} \sum_{t \in T} \sum_{s,a \in L_k^{ut}} \sum_{k \in K} \frac{w_{uk}^t(s,a|s',a')}{\sum_{i \leq k} w_{ui}^t(s,a|s',a') \mathbb{I}\{w_{ui}^t(s,a|s',a') \geq w_{\min}\}}
\end{aligned}
$$

For every (s,a,t), we consider the sequence of $w_{ui}^t(s,a|s',a') \in [w_{\min}, 1]$ with $i \in I = \{i \in \mathbb{N} : w_{ui}^t(s,a|s',a') \geq w_{\min}\}$ and apply Lemma E.5. This yields that

$$
\sum_{k \in K} \frac{w_{uk}^t(s,a|s',a')}{\sum_{i \leq k} w_{ui}^t(s,a|s',a') \mathbb{I}\{w_{ui}^t(s,a|s',a') \geq w_{\min}\}} \leq \ln\left(\frac{me}{w_{\min}}\right)
$$

and hence

$$\sum_{k\in K} \Delta_k^r \leq \frac{8CAS|T|^r D'(\mathrm{llnp}\left(C'^2+3C'\right)+1)}{n_{uk}(s',a')} \ln\left(\frac{me}{w_{\min}}\right)$$

Since each element in K has to contribute at least $\frac{D'\varepsilon'^r}{n_{uk}(s',a')}$ to this bound, we can conclude that

$$\begin{aligned}\sum_{k\in E} \mathbb{I}\{\Delta_k \geq \varepsilon'\} = &\sum_{k\in K} \mathbb{I}\{\Delta_k \geq \varepsilon'\}\\ &\leq |K| \leq \frac{8CAS|T|^r}{\varepsilon'^r} \ln\left(\frac{me}{w_{\min}}\right)\left(\mathrm{llnp}\left(C'^2+3C'\right)+1\right).\end{aligned}$$

Since $\ln\left(\frac{me}{w_{\min}}\right)\left(\mathrm{llnp}\left(C'^2+3C'\right)+1\right)$ is $\mathrm{polylog}(S, A, H, \delta^{-1}, \varepsilon'^{-1})$, the proof is complete. □

Lemma E.5. *Let a_i be a sequence taking values in $[a_{\min}, 1]$ with $a_{\min} > 0$ and $m > 0$, then*

$$\sum_{k=1}^{m} \frac{a_k}{\sum_{i=1}^k a_i} \leq \ln\left(\frac{me}{a_{\min}}\right).$$

Proof. Let f be a step-function taking value a_i on $[i-1, i)$ for all i. We have $F(t) := \int_0^t f(x)dx = \sum_{i=1}^t a_i$. By the fundamental theorem of Calculus, we can bound

$$\begin{aligned}\sum_{k=1}^{m} \frac{a_k}{\sum_{i=1}^k a_i} =& \frac{a_1}{a_1} + \int_1^m \frac{f(x)}{F(x)-F(0)}dx = 1 + \ln F(m) - \ln F(1)\\ \leq& 1 + \ln(m) - \ln a_{\min} = \ln\left(\frac{me}{a_{\min}}\right),\end{aligned}$$

where the inequality follows from $a_1 \geq a_{\min}$ and $\sum_{i=1}^m a_i \leq m$. □

Lemma E.6 (Properties of llnp). *The following properties hold:*

1. llnp *is continuous and nondecreasing.*
2. $f(x) = \frac{\mathrm{llnp}(nx)+D}{x}$ *with $n \geq 0$ and $D \geq 1$ is monotonically decreasing on $\mathbb{R}_+$.*
3. $\mathrm{llnp}(xy) \leq \mathrm{llnp}(x) + \mathrm{llnp}(y) + 1$ *for all $x, y \geq 0$.*

Proof.

1. For $x \leq e$ we have $\mathrm{llnp}(x) = 0$ and for $x \geq e$ we have $\mathrm{llnp}(x) = \ln(\ln(x))$ which is continuous and monotonically increasing and $\lim_{x\searrow e} \ln(\ln(x)) = 0$.
2. The function llnp is continuous as well as $1/x$ on $\mathbb{R}_+$ and therefore so it f. Further, f is differentiable except at $x = e/n$. For $x \in [0, e/n)$, we have $f(x) = D/x$ with derivative $-D/x^2 < 0$. Hence f is monotonically decreasing on $x \in [0, e/n)$. For $x > e/n$, we have $f(x) = \frac{\ln(\ln(nx))+D}{x}$ with derivative

$$-\frac{D+\ln(\ln(nx))}{x^2} + \frac{1}{x^2\ln(nx)} = \frac{1-\ln(nx)(D+\ln(\ln(nx)))}{x^2\ln(nx)}.$$

   The denominator is always positive in this range so f is monotonically decreasing if and only if $\ln(nx)(D - \ln(\ln(nx))) \geq 1$. Using $D \geq 1$, we have $\ln(nx)(D + \ln(\ln(nx))) \geq 1(1+0) = 1$.
3. First note that for $xy \leq e^e$ we have $\mathrm{llnp}(xy) \leq 1 \leq \mathrm{llnp}(x) + \mathrm{llnp}(y) + 1$ and therfore the statement holds for $x, y \leq e$.

   Then consider the case that $x, y \geq e$ and $\mathrm{llnp}(x) + \mathrm{llnp}(y) + 1 - \mathrm{llnp}(xy) = \ln\ln x + \ln\ln y + 1 - \ln(\ln(x) + \ln(y)) = -\ln(a+b) + 1 + \ln(a) + \ln(b)$ where $a = \ln x \geq 1$ and $b = \ln y \geq 1$. The function $g(a, b) = -\ln(a+b) + 1 + \ln(a) + \ln(b)$ is continuous and differentiable with $\frac{\partial g}{\partial a} = \frac{b}{a(a+b)} > 0$ and $\frac{\partial g}{\partial b} = \frac{a}{b(a+b)} > 0$. Therefore, g attains its minimum on $[1,\infty) \times [1,\infty)$ at $a = 1, b = 1$. Since $g(1,1) = 1 - \ln(2) \geq 0$, the statement also holds for $x, y \geq e$.

Finally consider the case where $x \leq e \leq y$. Then $\operatorname{llnp}(xy) \leq \operatorname{llnp}(ey) = \ln(1 + \ln y) \leq \ln\ln y + 1 \leq \operatorname{llnp}(x) + \operatorname{llnp}(y) + 1$. Due to symmetry this also holds for $y \leq e \leq x$.

□

E.4 Decomposition of Optimality Gap

In this section we decompose the optimality gap and then bound each term individually. Finally, both rate lemmas presented in the previous section are used to determine a bound on the number of nice / friendly episodes where the optimality gap can be larger than ε. The decomposition in the following lemma is a the simpler version bounding the number of ε-suboptimal nice episodes and eventually lead to the first bound in Theorem 4.

Lemma E.7 (Optimality Gap Bound On Nice Episodes)**.** *On the good event F^c it holds that $V_1^\star(s_0) - V_1^{\pi_k}(s_0) \leq \varepsilon$ on all nice episodes $k \in N$ except at most*

$$\frac{144(4 + 3H^2 + 4SH^2)ASH^2}{\varepsilon^2} \operatorname{polylog}(A, S, H, 1/\varepsilon, 1/\delta)$$

episodes.

Proof. Using optimism of the algorithm shown in Lemma E.16, we can bound

$$\begin{aligned}
&V_1^\star(s_0) - V_1^{\pi_k}(s_0) \\
\leq & |\tilde{V}_1^{\pi_k}(s_0) - V_1^{\pi_k}(s_0)| \\
\leq & \sum_{t=1}^{H}\sum_{s,a} w_{tk}(s,a)|(\tilde{P}_k(s,a,t) - P(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| + \sum_{t=1}^{H}\sum_{s,a} w_{tk}(s,a)|\tilde{r}_k(s,a,t) - r(s,a,t)| \\
\leq & \sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)|(\tilde{P}_k(s,a,t) - P(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| + \sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)|\tilde{r}_k(s,a,t) - r(s,a,t)| \\
& + \sum_{t=1}^{H}\sum_{s,a\notin L_{tk}} w_{tk}(s,a)|(\tilde{P}_k(s,a,t) - P(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| + \sum_{t=1}^{H}\sum_{s,a\notin L_{tk}} w_{tk}(s,a)|\tilde{r}_k(s,a,t) - r(s,a,t)| \\
\leq & \sum_{t=1}^{H}\sum_{s,a\notin L_{tk}} Hw_{\min} + \sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\Big[|(\tilde{P}_k(s,a,t) - \hat{P}_k(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| \\
& + |(\hat{P}_k(s,a,t) - P(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| + |\tilde{r}_k(s,a,t) - r(s,a,t)|\Big]
\end{aligned} \tag{9}$$

The first term is bounded by $c_\varepsilon \varepsilon = \frac{\varepsilon}{3}$. We now can use Lemma E.9, Lemma E.10 to bound the other terms by

$$\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{\frac{8(H + H\sqrt{S} + 2)^2}{n_{tk}(s,a)}\left(\operatorname{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{6SAH'}{\delta}\right)}.$$

We can then apply Lemma E.3 with $r = 2$, $C = 8(H + H\sqrt{S} + 2)^2$, $D = \frac{1}{2}\ln\frac{6SAH}{\delta'}$ (≥ 1 for any nontrivial setting) and $\varepsilon' = 2\varepsilon/3$ to bound this term by $\frac{2\varepsilon}{3}$ on all nice episodes except at most

$$\begin{aligned}
&\frac{64(H + \sqrt{S}H + 2)^2 ASH^2 3^2}{4\varepsilon^2} \operatorname{polylog}(A, S, H, 1/\varepsilon, 1/\delta) \\
\leq & \frac{144(4 + 3H^2 + 4SH^2)ASH^2}{\varepsilon^2} \operatorname{polylog}(A, S, H, 1/\varepsilon, 1/\delta)
\end{aligned}$$

Hence $V_1^\star(s_0) - V_1^{\pi_k}(s_0) \leq \varepsilon$ holds on all nice episodes except those. □

The lemma below is a refined version of the bound above and uses the stronger concept of friendly episodes to eventually lead to the second bound in Theorem 4.

Lemma E.8 (Optimality Gap Bound On Friendly Episodes)**.** *On the good event F^c it holds that $p_0^\top(V_1^\star - V_1^{\pi_k}) \le \varepsilon$ on all friendly episodes E except at most*

$$\left(\frac{9216}{\varepsilon} + 417S\right)\frac{ASH^4}{\varepsilon}\,\mathrm{polylog}(S, A, H, 1/\varepsilon, \delta)$$

episodes if $\delta' \le \frac{3AS^2H}{e^2}$.

Proof. We can further decompose the optimality gap bound in Equation (9) in the proof of Lemma E.7 as

$$\sum_{t=1}^{H}\sum_{s,a\notin L_{tk}}(H+1)w_{\min} + \sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\Big[|(\tilde{P}_k(s,a,t) - \hat{P}_k(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| + |\tilde{r}_k(s,a,t) - r(s,a,t)|$$

$$+ |(\hat{P}_k(s,a,t) - P(s,a,t))^\top V_{t+1}^\star| + |(\hat{P}_k(s,a,t) - P(s,a,t))^\top (V_{t+1}^\star - \tilde{V}_{t+1}^{\pi_k})|\Big].$$

$$\le c_\varepsilon \varepsilon + \sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\Big[|(\tilde{P}_k(s,a,t) - \hat{P}_k(s,a,t))^\top \tilde{V}_{t+1}^{\pi_k}| + |\tilde{r}_k(s,a,t) - r(s,a,t)|$$

$$+ \qquad |(\hat{P}_k(s,a,t) - P(s,a,t))^\top V_{t+1}^\star|\Big]$$

$$+ \sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)|(\hat{P}_k(s,a,t) - P(s,a,t))^\top (V_{t+1}^\star - \tilde{V}_{t+1}^{\pi_k})|.$$

The second term can be bounded using Lemmas E.11, E.10 and E.9 by

$$\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{\frac{32(H+1)^2}{n_{tk}(s,a)}\left(\mathrm{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{6SAH}{\delta'}\right)}.$$

which we bound by $\varepsilon/3$ using Lemma E.3 with $r = 2$, $C = 32(H+1)^2$, $D = \frac{1}{2}\ln\frac{6SAH}{\delta'}$ and $\varepsilon' = \varepsilon/3$ on all friendly episodes except at most

$$\frac{8CASH^2}{\varepsilon'^2}\,\mathrm{polylog}(S, A, H, 1/\varepsilon, 1/\delta) \le \frac{9216ASH^4}{\mathrm{polylog}}(S, A, H, 1/\varepsilon, 1/\delta).$$

Finally, we apply Lemma E.12 bound to bound the last term in Equation E.4 by $\varepsilon/3$ on all friendly epsiodes but at most

$$\frac{417AS^2H^4}{\varepsilon}\,\mathrm{polylog}(S, A, H, 1/\delta, 1/\varepsilon).$$

It hence follows that $p_0^\top(V_1^\star - V_1^{\pi_k}) \le \varepsilon$ on all friendly episodes but at most

$$\left(\frac{9216ASH^4}{\varepsilon^2} + \frac{417AS^2H^4}{\varepsilon}\right)\mathrm{polylog}(S, A, H, 1/\delta, 1/\varepsilon).$$

□

Lemma E.9 (Algorithm Learns Fast Enough)**.** *It holds for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in [H]$*

$$|(\hat{P}_k(s,a,t) - \tilde{P}_k(s,a,t))^\top \tilde{V}_{t+1}| \le \sqrt{\frac{2H^2}{n_{tk}(s,a)}\left(\mathrm{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{3SAH}{\delta'}\right)}.$$

Proof. Using the definition of the constraint in the planning step of the algorithm shown in Lemma D.1 we can bound

$$|(\hat{P}_k(s,a,t) - \tilde{P}_k(s,a,t))^\top \tilde{V}_{t+1}| \leq \sqrt{\frac{H^2}{n_{tk}(s,a)}\left(2\,\mathrm{llnp}(n_{tk}(s,a)) + \ln\frac{3SAH}{\delta'}\right)}.$$
$$\leq \sqrt{\frac{2H^2}{n_{tk}(s,a)}\left(\mathrm{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{3SAH}{\delta'}\right)}.$$

□

Lemma E.10 (Basic Decompsition Bound)**.** *On the good event F^c it holds for all $s \in \mathcal{S}, a \in \mathcal{A}$ and $t \in [H]$*

$$|(\hat{P}_k(s,a,t) - P(s,a,t))^\top \tilde{V}_{t+1}| \leq \sqrt{\frac{8H^2S}{n_{tk}(s,a)}\left(\mathrm{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{6SAH}{\delta'}\right)}$$
$$|\tilde{r}_k(s,a,t) - r(s,a,t)| \leq \sqrt{\frac{4}{n_{tk}(s,a)}\left(\mathrm{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{3SAH}{\delta'}\right)}.$$

Proof. On the good event $(F_k^{L1})c$ we have using Hölder's inequality

$$\begin{aligned}|(\hat{P}_k(s,a,t) - P(s,a,t))^\top \tilde{V}_{t+1}| \leq& \|\hat{P}_k(s,a,t) - P(s,a,t))\|_1 \|\tilde{V}_{t+1}\|_\infty \\ \leq& H\sqrt{\frac{4}{n_{tk}(s,a)}\left(2\,\mathrm{llnp}(n_{tk}(s,a)) + \ln\frac{3SAH(2^S-2)}{\delta'}\right)} \\ \leq& \sqrt{\frac{8H^2S}{n_{tk}(s,a)}\left(\mathrm{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{6SAH}{\delta'}\right)}.\end{aligned}$$

Further, on $(F_k^R)^c$ we have

$$\begin{aligned}|\tilde{r}_k(s,a,t) - r(s,a,t)| \leq& |\tilde{r}_k(s,a,t) - r(s,a,t)| + |\tilde{r}_k(s,a,t) - \hat{r}(s,a,t)| \\ \leq& 2\sqrt{\frac{1}{n_{tk}(s,a)}\left(2\,\mathrm{llnp}(n_{tk}(s,a)) + \ln\frac{3SAH}{\delta'}\right)}\end{aligned}$$

□

Lemma E.11 (Fixed V Term Confidence Bound)**.** *On the good event F^c it holds for all $s \in \mathcal{S}, a \in \mathcal{A}$ and $t \in [H]$*

$$|(\hat{P}_k(s,a,t) - P(s,a,t))^\top V^\star_{t+1}| \leq \sqrt{\frac{2H^2}{n_{tk}(s,a)}\left(\mathrm{llnp}\, n_{tk}(s,a) + \frac{1}{2}\ln\frac{3SAH}{\delta'}\right)}$$

Proof. Since we consider the event $(F_k^V)^c$, we can bound

$$|(\hat{P}_k(s,a,t) - P(s,a,t))^\top V^\star_{t+1}| \leq \sqrt{\frac{2H^2}{n_{tk}(s,a)}\left(\mathrm{llnp}\, n_{tk}(s,a) + \frac{1}{2}\ln\frac{3SAH}{\delta'}\right)}$$

□

Lemma E.12 (Lower Order Term)**.** *Assume $\delta' \leq \frac{3AS^2H}{e^2}$. On the good event F^c on all friendly episodes $k \in E$ except at most $\frac{417AS^2H^4}{\varepsilon}$* $\mathrm{polylog}(S,A,H,1/\delta,1/\varepsilon)$*. it holds that*

$$\sum_{t=1}^{H}\sum_{s,a \in L_{tk}} w_{tk}(s,a)|(\hat{P}_k(s,a,t) - P(s,a,t))^\top(\tilde{V}^{\pi_k}_{t+1} - V^\star_{t+1})| \leq \frac{\varepsilon}{3}.$$

Proof.

$$
\begin{aligned}
&\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)|(\hat{P}_k(s,a,t)-P(s,a,t))^\top(\tilde{V}^{\pi_k}_{t+1}-V^\star_{t+1})| \\
\leq&\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sum_{s'}\sqrt{\frac{2P(s'|s,a,t)}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a))+\ln\frac{3S^2AH}{\delta'}\right)}|\tilde{V}^{\pi_k}_{t+1}(s')-V^\star_{t+1}(s')| \\
&+\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sum_{s'}\frac{1}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a))+\ln\frac{3S^2AH}{\delta'}\right)|\tilde{V}^{\pi_k}_{t+1}(s')-V^\star_{t+1}(s')| \\
\leq&\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sum_{s'}\sqrt{\frac{2P(s'|s,a,t)}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a))+\ln\frac{3S^2AH}{\delta'}\right)\left(\tilde{V}^{\pi_k}_{t+1}(s')-V^\star_{t+1}(s')\right)^2} \\
&+\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} \frac{w_{tk}(s,a)HS}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a))+\ln\frac{3S^2AH}{\delta'}\right) \\
\leq&\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{\frac{2S}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a))+\ln\frac{3S^2AH}{\delta'}\right)P(s,a,t)^\top\left(\tilde{V}^{\pi_k}_{t+1}-V^\star_{t+1}\right)^2} \\
&+\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} \frac{w_{tk}(s,a)HS}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a))+\ln\frac{3S^2AH}{\delta'}\right)
\end{aligned}
$$

The first inequality follows since we only consider outcomes in the event $(F_k^P)^c$, the second from the fact that value function are in the range $[0,H]$ and the third is an application of the Cauchy-Schwarz inequality. Using of optimism of the algorithm (Lemma E.16), we now bound $P(s,a,t)^\top\left(\tilde{V}^{\pi_k}_{t+1}-V^\star_{t+1}\right)^2 \leq P(s,a,t)^\top\left(\tilde{V}^{\pi_k}_{t+1}-V^{\pi_k}_{t+1}\right)^2$ which we bound by $c_\varepsilon\varepsilon+\left(c_\varepsilon\varepsilon+\sqrt{\frac{C'^2}{n_{tk}(s,a)S}\left(\operatorname{llnp}(n_{tk}(s,a)+\frac{1}{2}\ln\frac{3AS^2H\varepsilon^4}{\delta'}\right)}\right)^2 \leq c_\varepsilon\varepsilon+(c_\varepsilon\varepsilon+\frac{C'}{\sqrt{S}}\sqrt{J(s,a,t)})^2$ using Lemma E.13. To keep the notation concise, we use here the shorthand $J(s,a,t)=\frac{1}{n_{tk}(s,a)}\left(\operatorname{llnp}(n_{tk}(s,a))+\frac{1}{2}\ln\frac{3e^4S^2AH}{\delta'}\right)$. This bound holds on all friendly episodes except at most $\left(32ASH^2+48AS^2H^3+AS^2H^4+16AS^2\right)\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$. Plugging this into the bound from above, we get the upper bound

$$
\begin{aligned}
&\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{4SJ(s,a,t)\left(c_\varepsilon\varepsilon+(c_\varepsilon\varepsilon+C'\sqrt{J(s,a,t)/S})^2\right)}+\sum_{t=1}^{H}\sum_{s,a\in L_{tk}}2w_{tk}(s,a)HSJ(s,a,t) \\
&\leq\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{4SJ(s,a,t)c_\varepsilon\varepsilon}+\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{4SJ(s,a,t)(c_\varepsilon\varepsilon+C'\sqrt{J(s,a,t)/S})^2} \\
&+\sum_{t=1}^{H}\sum_{s,a\in L_{tk}}2w_{tk}(s,a)HSJ(s,a,t) \\
=&\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)\sqrt{4(c_\varepsilon\varepsilon+c_\varepsilon^2\varepsilon^2)SJ(s,a,t)}+\sum_{t=1}^{H}\sum_{s,a\in L_{tk}}2w_{tk}(s,a)J(s,a,t)(C'+SH),
\end{aligned}
$$

where we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. We now bound the first term using Lemma E.3 with $r = 2, \varepsilon' = \varepsilon/6, D = \frac{1}{2}\ln\frac{3e^4S^2AH}{\delta'}, C = 4(c_\varepsilon\varepsilon + c_\varepsilon^2\varepsilon^2)S$ on all but $\frac{8CASH^2}{\varepsilon'^2}\operatorname{polylog}(\dots) = \frac{192c_\varepsilon(1+c_\varepsilon\varepsilon)AS^2H^2}{\varepsilon}\operatorname{polylog}(\dots)$ friendly episodes by $\varepsilon/6$.

Applying Lemma E.3 with $r = 1, \varepsilon' = \varepsilon/6, D = \frac{1}{2}\ln\frac{3e^4S^2AH}{\delta'}$ and $C = 2(C' + HS)$, we can bound the second term by $\varepsilon/6$ on all but $\frac{8CASH}{\varepsilon'}\operatorname{polylog}(\dots) = \frac{96AS(C'+HS)H^2}{\varepsilon}\operatorname{polylog}(\dots)$ friendly episodes. Hence, it holds

$$\sum_{t=1}^{H}\sum_{s,a\in L_{tk}} w_{tk}(s,a)|(\hat{P}_k(s,a,t) - P(s,a,t))^\top(\tilde{V}^{\pi_k}_{t+1} - V^\star_{t+1})| \leq \frac{\varepsilon}{3}$$

on all friendly episodes except at most

$$\left(\frac{96AS(C'+HS)H^2}{\varepsilon} + \frac{192c_\varepsilon(1+c_\varepsilon\varepsilon)AS^2H^2}{\varepsilon}\right.$$
$$\left. + 32ASH^2 + 48AS^2H^3 + AS^2H^4 + 16AS^2\right)\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$$

episodes. Since $C' = \operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$, this simplifies to

$$\left(\frac{96AS}{\varepsilon} + \frac{96AS^2H^3}{\varepsilon} + \frac{64AS^2H^2}{\varepsilon} + 64AS^2H^2\right.$$
$$\left. + 32ASH^2 + 48AS^2H^3 + AS^2H^4 + 16AS^2\right)\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$$
$$\leq ((64+32+48+1+16)AS^2H^4 + \frac{96+96+64}{\varepsilon}AS^2H^3)\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$$

failure episodes in E. We can finally bound the failure episodes by

$$\frac{417AS^2H^4}{\varepsilon}\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon).$$

□

Lemma E.13. *On the good event F^c for any $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in [H]$ with $\delta' \leq \frac{3AS^2H}{e^2}$ it holds*

$$P(s,a,t)^\top(\tilde{V}^{\pi_k}_{t+1} - V^{\pi_k}_{t+1})^2 \leq c_\varepsilon\varepsilon + \left(c_\varepsilon\varepsilon + \sqrt{\frac{1}{n_{tk}(s,a)S}\left(\operatorname{llnp}(n_{tk}(s,a) + \frac{1}{2}\ln\frac{3AS^2H\varepsilon^4}{\delta'}\right)}\right)^2$$

where $C' = 1 + \sqrt{\frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}}$ on all friendly episodes except for at most

$$\left(32ASH^2 + 48AS^2H^3 + AS^2H^4 + 16AS^2\right)\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$$

episodes.

Proof. Define $L' = \{s' : w^{t+1}_{tk}(s',a'|s,a) > w'_{\min}\}$ and $J(s') = \frac{\operatorname{llnp} n_{t+1k}(s',a') + \frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}}{n_{t+1k}(s',a')}$ where $a' = \pi_k(s',t+1)$ and $C' = 1 + \sqrt{\frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}}$. Using Lemma E.14, we bound

$$P(s,a,t)^\top(\tilde{V}^{\pi_k}_{t+1} - V^{\pi_k}_{t+1})^2 = \sum_{s'} P(s'|s,a,t)(\tilde{V}^{\pi_k}_{t+1}(s') - V^{\pi_k}_{t+1}(s'))^2$$
$$\leq S w_{\min} H^2 + \sum_{s'\in L'} P(s'|s,a,t)\left(c_\varepsilon\varepsilon + C'\sqrt{J(s')}\right)^2$$
$$\leq c_\varepsilon\varepsilon + C'^2\sum_{s'\in L'} P(s'|s,a,t)J(s') + c_\varepsilon^2\varepsilon^2 + 2c_\varepsilon\varepsilon C'\sum_{s'\in L'} P(s'|s,a,t)\sqrt{J(s')}$$

on all friendly episodes except at most $\left(32 + 48SH + SH^2\right) ASH^2 \operatorname{polylog}(S, A, H, 1/\delta, 1/\varepsilon)$. Define now $L'' = \{(s', a') : s' \in L', a' = \pi_k(s', t+1)\}$. We apply Lemma E.4 with $|T| = \{t+1\}, C = 1, D = \frac{1}{2} \ln \frac{3e^2 S^2 AH}{\delta'} \geq 1, r = 1$ and $\varepsilon' = 1/S$ to

$$
\begin{aligned}
\sum_{s' \in L'} P(s'|s,a,t) J(s') = & \sum_{s',a' \in L''} \frac{w_{tk}^{t+1}(s',a'|s,a)}{n_{t+1k}(s',a')} \left(\operatorname{llnp} n_{t+1k}(s',a') + \frac{1}{2} \ln \frac{3e^2 S^2 AH}{\delta'} \right) \\
\leq & \frac{1}{n_{tk}(s,a) S} \left(\operatorname{llnp}(n_{tk}(s,a)) + \frac{1}{2} \ln \frac{3 S^2 AH e^4}{\delta'} \right)
\end{aligned}
$$

on all but at most $8AS^2 \operatorname{polylog}(A, S, H, 1/\delta, 1/\varepsilon)$ friendly episodes. Similarly, we bound

$$
\begin{aligned}
& \sum_{s' \in L'} P(s'|s,a,t) \sqrt{J(s')} \\
= & \sum_{s',a' \in L''} w_{tk}^{t+1}(s',a'|s,a) \sqrt{\frac{1}{n_{t+1k}(s',a')} \left(\operatorname{llnp} n_{t+1k}(s',a') + \frac{1}{2} \ln \frac{3e^2 S^2 AH}{\delta'} \right)} \\
\leq & \sqrt{\frac{1}{n_{tk}(s,a) S} \left(\operatorname{llnp}(n_{tk}(s,a)) + \frac{1}{2} \ln \frac{3 S^2 AH e^4}{\delta'} \right)}
\end{aligned}
$$

on all but at most $8AS^2 \operatorname{polylog}(A, S, H, 1/\delta, 1/\varepsilon)$ friendly episodes. Hence on all friendly episodes except those failure episodes, we get

$$
P(s,a,t)^\top (\tilde{V}_{t+1}^{\pi_k} - V_{t+1}^{\pi_k})^2 \leq c_\varepsilon \varepsilon + \left(c_\varepsilon \varepsilon + \sqrt{\frac{C'^2}{n_{tk}(s,a) S} \left(\operatorname{llnp}(n_{tk}(s,a) + \frac{1}{2} \ln \frac{3AS^2 H \varepsilon^4}{\delta'} \right)} \right)^2.
$$

□

Lemma E.14. *Consider a fix $s' \in \mathcal{S}$ and $t \in [H]$, $\delta' \leq \frac{3AS^2H}{e^2}$ and the good event F^c. On all but at most*

$$
\left(32 + 48SH + SH^2\right) ASH^2 \operatorname{polylog}(S, A, H, 1/\delta, 1/\varepsilon)
$$

friendly episodes E it holds that

$$
V_t^{\pi_k}(s') - \tilde{V}_t^{\pi_k}(s') \leq c_\varepsilon \varepsilon + \left(1 + \sqrt{12 \ln \frac{3e^2 S^2 AH}{\delta'}} \right) \sqrt{\frac{1}{n_{tk}(s',a')} \left(\operatorname{llnp} n_{tk}(s',a') + \frac{1}{2} \ln \frac{3e^2 S^2 AH}{\delta'} \right)},
$$

where $a' = \pi_k(s', t)$.

Proof. For any t,s' and $a' = \pi_k(s', t)$ we use Lemma E.15 to write the value difference as

$$
\begin{aligned}
\tilde{V}_t^{\pi_k}(s') - V_t^{\pi_k}(s') = & \sum_{u=t}^{H} \sum_{s,a} w_{tk}^u(s,a|s',a') (\tilde{P}_k(s,a,u) - P(s,a,u))^\top \tilde{V}_{u+1} \\
& + \sum_{u=t}^{H} \sum_{s,a} w_{tk}^u(s,a|s',a') (\tilde{r}_k(s,a,u) - r(s,a,u))
\end{aligned}
$$

Let $L_k^{ut} = \{s, a \in \mathcal{S} \times \mathcal{A} : w_{tk}^u(s,a|s',a') \geq w_{\min}\}$ be the set of state-action pairs for which the conditional probability of observing is sufficiently large. Then we can bound the low-probability differences as

$$
\begin{aligned}
& \sum_{u=t}^{H} \sum_{s,a \in (L_k^{ut})^c} w_{tk}^u(s,a|s',a') [(\tilde{r}_k(s,a,u) - r(s,a,u)) + (P(s,a,u) - \tilde{P}_k(s,a,u))^\top \tilde{V}_{u+1}] \\
\leq & \sum_{u=t}^{H} \sum_{s,a \in (L_k^{ut})^c} w_{\min} H \leq w_{\min} H^2 S = c_\varepsilon \varepsilon.
\end{aligned}
$$

For the other terms with significant conditional probability, we can leverage the fact that we only consider events in $(F_k^R)^c$ and $(F_k^P)^c$ to bound

$$\begin{aligned}
&\sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a')(\tilde{r}_k(s,a,u) - r(s,a,u)) \\
\leq &\sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a')\sqrt{\frac{32}{n_{tk}(s,a)}\left(\operatorname{llnp}(n_{tk}(s,a)) + \frac{1}{2}\ln\frac{3SAH}{\delta'}\right)}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a')(P(s,a,u) - \tilde{P}_k(s,a,u))^\top \tilde{V}_{u+1} \\
\leq &\sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a') \sum_{s''} \tilde{V}_{u+1}(s'')\sqrt{\frac{2P(s''|s,a,u)}{n_{uk}(s,a)}\left(2\operatorname{llnp}(n_{uk}(s,a)) + \ln\frac{3S^2AH}{\delta'}\right)} \\
&+ \sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a') \sum_{s''} \frac{\tilde{V}_{u+1}(s'')}{n_{uk}(s,a)}\left(2\operatorname{llnp}(n_{uk}(s,a)) + \ln\frac{3S^2AH}{\delta'}\right) \\
\leq &\sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a')\sqrt{\frac{2SH^2}{n_{uk}(s,a)}\left(2\operatorname{llnp}(n_{uk}(s,a)) + \ln\frac{3S^2AH}{\delta'}\right)} \\
&+ \sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a') \frac{SH}{n_{uk}(s,a)}\left(2\operatorname{llnp}(n_{uk}(s,a)) + \ln\frac{3S^2AH}{\delta'}\right)
\end{aligned}$$

where we use Cauchy Schwarz for the last inequality. Combining these individual bounds, we can upper-bound the value difference as

$$\begin{aligned}
&\tilde{V}_t^{\pi_k}(s') - V_t^{\pi_k}(s') \\
\leq &c_\varepsilon \varepsilon + \sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a')\sqrt{\frac{(4\sqrt{2} + 2\sqrt{S}H)^2}{n_{uk}(s,a)}\left(\operatorname{llnp}(n_{uk}(s,a)) + \frac{1}{2}\ln\frac{3S^2AH}{\delta'}\right)} \\
&+ \sum_{u=t}^{H} \sum_{s,a \in L_k^{ut}} w_{tk}^u(s,a|s',a') \frac{2SH}{n_{uk}(s,a)}\left(\operatorname{llnp}(n_{uk}(s,a)) + \frac{1}{2}\ln\frac{3S^2AH}{\delta'}\right)
\end{aligned} \tag{10}$$

We now apply Lemma E.4 with $r = 2, D = \frac{1}{2}\ln\frac{3S^2AH}{\delta'}, C = (4\sqrt{2} + 2\sqrt{S}H)^2, T = \{t+1, t+2, \dots H\}$ and $\varepsilon' = 1$ and get that the second term above is bounded by

$$\sqrt{\frac{1}{n_{tk}(s',a')}\left(\operatorname{llnp} n_{tk}(s',a') + \frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}\right)}$$

on all friendly episodes but at most

$$\frac{8CASH^2}{\varepsilon'^2}\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon) = (32 + 16\sqrt{2S}H + SH^2)ASH^2\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$$

episodes. We apply Lemma E.4 again to the final term in Equation (10) above with $r = 1, D = \frac{1}{2}\ln\frac{3S^2AH}{\delta'} \geq 1, T = \{t+1, t+2, \dots H\}, C = 2SH$ and $\varepsilon' = 1$. Then the final term is bounded by $\frac{1}{n_{tk}(s',a')}\left(\operatorname{llnp} n_{tk}(s',a') + \frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}\right)$. on all friendly episodes but

$$\frac{8CASH^2}{\varepsilon'^2}\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon) = 16AS^2H^3\operatorname{polylog}(S,A,H,1/\delta,1/\varepsilon)$$

many. Combining these bounds, we arrive at

$$\begin{aligned}
&V_t^{\pi_k}(s') - \tilde{V}_t^{\pi_k}(s') \\
\le & c_\varepsilon \varepsilon + \sqrt{\frac{1}{n_{tk}(s',a')}\left(\operatorname{llnp} n_{tk}(s',a') + \frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}\right)} \\
&+ \frac{1}{n_{tk}(s',a')}\left(\operatorname{llnp} n_{tk}(s',a') + \frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}\right) \\
\le & c_\varepsilon \varepsilon + \left(1 + \sqrt{\frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}}\right)\sqrt{\frac{1}{n_{tk}(s',a')}\left(\operatorname{llnp} n_{tk}(s',a') + \frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}\right)},
\end{aligned}$$

where we bounded $\sqrt{\frac{1}{n_{tk}(s',a')}\left(\operatorname{llnp} n_{tk}(s',a') + \frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}\right)}$ by $\frac{1}{2}\ln\frac{3e^2S^2AH}{\delta'}$ since it is decreasing in $n_{tk}(s',a')$ and we therefore can simply use $n_{tk}(s',a') = 1$ (entire bound holds trivially for $n_{tk}(s',a') = 0$). $\square$

E.5 Useful Lemmas

Lemma E.15 (Value Difference Lemma)**.** *For any two MDPs M' and M'' with rewards r' and r'' and transition probabilities P' and P'', the difference in values with respect to the same policy π can be written as*

$$V_i'(s) - V_i''(s) = \mathbb{E}''\left[\sum_{t=i}^{H}(r'(s_t,a_t,t) - r''(s_t,a_t,t))\middle| s_i = s\right] + \mathbb{E}''\left[\sum_{t=i}^{H}(P'(s_t,a_t,t) - P''(s_t,a_t,t))^\top V'_{t+1}\middle| s_i = s\right]$$

where $V'_{H+1} = V''_{H+1} = \vec{0}$ and the expectation $\mathbb{E}'$ is taken w.r.t to P' and π and $\mathbb{E}''$ w.r.t. P'' and π.

Proof. For $i = H+1$ the statement is trivially true. We assume now it holds for $i+1$ and show it holds also for i. Using only this induction hypothesis and basic algebra, we can write

$$\begin{aligned}
&V_i'(s) - V_i''(s) \\
=&\mathbb{E}_\pi[r'(s_i,a_i,i) + {V'_{i+1}}^\top P'(s_i,a_i,i) - r''(s_i,a_i,i) - {V''_{i+1}}^\top P''(s_i,a_i,i) | s_i = s] \\
=&\mathbb{E}_\pi[r'(s_i,a_i,i) - r''(s_i,a_i,i) | s_i = s] + \mathbb{E}_\pi\left[\sum_{s'\in\mathcal{S}} V'_{i+1}(s')(P'(s'|s_i,a_i,i) - P''(s'|s_i,a_i,i))\middle| s_i = s\right] \\
&+ \mathbb{E}_\pi\left[\sum_{s'\in\mathcal{S}} P''(s'|s_i,a_i,i)(V'_{i+1}(s') - V''_{i+1}(s'))\middle| s_i = s\right] \\
=&\mathbb{E}_\pi[r'(s_i,a_i,i) - r''(s_i,a_i,i) | s_i = s] + \mathbb{E}_\pi\left[\sum_{s'\in\mathcal{S}} V'_{i+1}(s')(P'(s'|s_i,a_i,i) - P''(s'|s_i,a_i,i))\middle| s_i = s\right] \\
&+ \mathbb{E}''\left[V'_{i+1}(s_{i+1}) - V''_{i+1}(s_{i+1}))\middle| s_i = s\right] \\
=&\mathbb{E}_\pi[r'(s_i,a_i,i) - r''(s_i,a_i,i) | s_i = s] + \mathbb{E}_\pi\left[\sum_{s'\in\mathcal{S}} V'_{i+1}(s')(P'(s'|s_i,a_i,i) - P''(s'|s_i,a_i,i))\middle| s_i = s\right] \\
&+ \mathbb{E}''\left[\mathbb{E}''\left[\sum_{t=i+1}^{H}(r'(s_t,a_t,t) - r''(s_t,a_t,t))\middle| s_{i+1}\right] + \mathbb{E}''\left[\sum_{t=i+1}^{H}(P'(s_t,a_t,t) - P''(s_t,a_t,t))^\top V'_{t+1}\middle| s_{i+1}\right]\middle| s_i = s\right]
\end{aligned}$$

$$=\mathbb{E}''\left[\sum_{t=i}^{H}(r'(s_t,a_t,t)-r''(s_t,a_t,t))\middle|s_i=s\right]+\mathbb{E}''\left[\sum_{t=i}^{H}(P'(s_t,a_t,t)-P''(s_t,a_t,t))^\top V'_{t+1}\middle|s_i=s\right]$$

where the last equality follows from law of total expectation ☐

Lemma E.16 (Algorithm ensures optimism)**.** *On the good event F^c it holds that for all episodes k, $t\in[H]$, $s\in\mathcal{S}$ that*

$$V_t^{\pi_k}(s)\le V_t^\star(s)\le \tilde V_t^{\pi_k}(s).$$

Proof. The first inequality follows simply from the definition of the optimal value function $V^\star$.

Since all outcome we consider are in the event $(F_k^V)^c$, we know that the true transition probabilities P, the optimal policy $\pi^\star$ and optimal policy $V^\star$ are a feasible solution for the optimistic planning problem in Lemma D.1 that UBEV solves. It therefore follows immediately that $p_0^\top \tilde V_1^{\pi_k}\ge p_0^\top V_1^\star$. ☐

F General Concentration Bounds

Lemma F.1. *Let $X_1,X_2,\dots$ be a martingale difference sequence adapted to filtration $\{\mathcal{F}_t\}_{t=1}^\infty$ with X_t conditionally σ^2-subgaussian so that $\mathbb{E}[\exp(\lambda(X_t-\mu))|\mathcal{F}_{t-1}]\le\exp(\lambda^2\sigma^2/2)$ almost surely for all $\lambda\in\mathbb{R}$. Then with $\hat\mu_t=\frac{1}{t}\sum_{i=1}^t X_i$ we have for all $\delta\in(0,1]$*

$$\mathbb{P}\left(\exists t:|\hat\mu_t-\mu|\ge\sqrt{\frac{4\sigma^2}{t}\left(2\operatorname{llnp}(t)+\ln\frac{3}{\delta}\right)}\right)\le 2\delta\,.$$

Proof. Let $S_t=\sum_{s=1}^t(X_s-\mu)$. Then

$$\begin{aligned}
&\mathbb{P}\left(\exists t:\hat\mu_t-\mu\ge\sqrt{\frac{4\sigma^2}{t}\left(2\operatorname{llnp}(t)+\ln\frac{3}{\delta}\right)}\right)\\
\le&\mathbb{P}\left(\exists t:S_t\ge\sqrt{4\sigma^2 t\left(2\operatorname{llnp}(t)+\ln\frac{3}{\delta}\right)}\right)\\
\le&\sum_{k=0}^\infty\mathbb{P}\left(\exists t\in[2^k,2^{k+1}]:S_t\ge\sqrt{4\sigma^2 t\left(2\operatorname{llnp}(t)+\ln\frac{3}{\delta}\right)}\right)\\
\le&\sum_{k=0}^\infty\mathbb{P}\left(\exists t\le 2^{k+1}:S_t\ge\sqrt{2\sigma^2 2^{k+1}\left(2\operatorname{llnp}(2^k)+\ln\frac{3}{\delta}\right)}\right)
\end{aligned}$$

We now consider $M_t=\exp(\lambda S_t)$ for $\lambda>0$ which is a nonnegative sub-martingale and use the short-hand $f=\sqrt{2\sigma^2 2^{k+1}\left(2\operatorname{llnp}(2^k)+\ln\frac{3}{\delta}\right)}$. Then by Doob's maximal inequality for nonnegative submartingales

$$\mathbb{P}\left(\exists t\le 2^{k+1}:S_t\ge f\right)=\mathbb{P}\left(\max_{t\le 2^{k+1}}M_t\ge\exp(\lambda f)\right)\le\frac{\mathbb{E}[M_{2^{k+1}}]}{\exp(\lambda f)}\le\exp\left(2^{k+1}\frac{\lambda^2\sigma^2}{2}-\lambda f\right).$$

Choosing the optimal $\lambda=\frac{f}{\sigma^2 2^{k+1}}$ we obtain the bound

$$\mathbb{P}\left(\exists t\le 2^{k+1}:S_t\ge f\right)\le\exp\left(-\frac{f^2}{2^{k+2}\sigma^2}\right)=\exp\left(-2\operatorname{llnp}(2^k)-\ln\frac{3}{\delta}\right)=\frac{\delta}{3}\exp\left(-2\operatorname{llnp}(2^k)\right)\tag{11}$$

$$\begin{aligned}
=&\frac{\delta}{3}\exp\left(-\max\{0,2\ln\max\{0,\ln 2^k\}\}\right)=\frac{\delta}{3}\min\left\{1,(k\ln 2)^{-2}\right\}\\
\le&\frac{\delta}{3}\min\left\{1,\frac{1}{k^2\ln 2}\right\}.
\end{aligned}$$

Plugging this back in the bound from above, we get

$$\mathbb{P}\left(\exists t : \hat{\mu}_t - \mu \geq \sqrt{\frac{4\sigma^2}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)}\right) \leq \frac{\delta}{3}\sum_{k=0}^{\infty}\min\left\{1, \frac{1}{k^2\ln(2)}\right\}$$
$$= \delta\,\frac{1}{3}\left(\frac{\pi^2}{6\ln 2} + 2 - 1/\ln(2)\right) \leq \delta\,. \quad (12)$$

For the other side, the argument follows completely analogously with

$$\begin{aligned}\mathbb{P}\left(\exists t \leq 2^{k+1} : S_t \leq -f\right) =& \mathbb{P}\left(\exists t \leq 2^{k+1} : -S_t \geq f\right)\\ =&\mathbb{P}\left(\max_{t\leq 2^{k+1}} \exp(-\lambda S_t) \geq \exp(\lambda f)\right)\\ \leq& \frac{\mathbb{E}[\exp(-\lambda S_{2^{k+1}})]}{\exp(\lambda f)} \leq \exp\left(2^{k+1}\frac{\lambda^2\sigma^2}{2} - \lambda f\right).\end{aligned}$$

□

Lemma F.2. *Let $X_1, X_2, \ldots$ be a sequence of Bernoulli random variables with bias $\mu \in [0,1]$. Then for all $\delta \in (0,1]$*

$$\mathbb{P}\left(\exists t : |\hat{\mu}_t - \mu| \geq \sqrt{\frac{2\mu}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)} + \frac{1}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)\right) \leq 2\delta$$

Proof.

$$\begin{aligned}&\mathbb{P}\left(\exists t : \hat{\mu}_t - \mu \geq \sqrt{\frac{2\mu}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)} + \frac{1}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)\right)\\ =&\mathbb{P}\left(\exists t : S_t \geq \sqrt{2\mu t\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)} + 2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)\\ \leq&\sum_{k=0}^{\infty}\mathbb{P}\left(\exists t \leq 2^{k+1} : S_t \geq \sqrt{2\mu 2^k\left(2\operatorname{llnp}(2^k) + \ln\frac{3}{\delta}\right)} + 2\operatorname{llnp}(2^k) + \ln\frac{3}{\delta}\right)\end{aligned}$$

Let $g = 2\operatorname{llnp}(2^k) + \ln\frac{3}{\delta}$ and $f = \sqrt{2^{k+1}\mu g} + g$. Further define $S_t = \sum_{i=1}^t X_i - t\mu$ and $M_t = \exp(\lambda S_t)$ which is by construction a nonnegative submartingale. Applying Doob's maximal inequality for nonnegative submartingales, we bound

$$\mathbb{P}\left(\exists t \leq 2^{k+1} : S_t \geq f\right) = \mathbb{P}\left(\max_{i\leq 2^{k+1}} M_i \geq \exp(\lambda f)\right) \leq \frac{\mathbb{E}[M_{2^{k+1}}]}{\exp(\lambda f)} = \exp\left(\ln\mathbb{E}[M_{2^{k+1}}] - \lambda f\right).$$

Since this holds for all $\lambda \in \mathbb{R}$, we can bound

$$\mathbb{P}\left(\exists t \leq 2^{k+1} : S_t \geq f\right) \leq \exp\left(-\sup_{\lambda\in\mathbb{R}}\left(\lambda f - \ln\mathbb{E}[M_{2^{k+1}}]\right)\right)$$

and using Corollary 2.11 by Boucheron et al. [25] (see also note below proof of Corollary 2.11) bound that by

$$\exp\left(-\frac{f^2}{2(2^{k+1}\mu + f/3)}\right)$$

We now argue that this quantity can be upper-bounded by $\exp(-g)$. This is equivalent to

$$
\begin{aligned}
-\frac{f^2}{2(2^{k+1}\mu + f/3)} \leq& - g \\
f^2 \geq& 2g(2^{k+1}\mu + f/3) = \frac{2}{3}gf + \frac{2^{k+2}}{3}\mu g \\
g^2 + 2\sqrt{2^{k+1}\mu g}g + 2^{k+1}\mu g \geq& \frac{2}{3}g^2 + \frac{2}{3}\sqrt{2^{k+1}\mu g}g + \frac{2^{k+2}}{3}\mu g \\
\frac{1}{3}g^2 + \frac{4}{3}\sqrt{2^{k+1}\mu g}g + \frac{1}{3}2^{k+1}\mu g \geq& 0.
\end{aligned}
$$

Each line is an equivalent inequality since $g, f \geq 0$ and each term on the left in the final inequality is nonnegative. Hence, we get $\mathbb{P}\left(\exists t \leq 2^{k+1} : S_t \geq f\right) \leq \exp(-g)$. Following now the arguments from the proof of Lemma F.1 in Equations (11)–(12), we obtain that

$$
\mathbb{P}\left(\exists t : \hat{\mu}_t - \mu \geq \sqrt{\frac{2\mu}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)} + \frac{1}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3}{\delta}\right)\right) \leq \delta.
$$

For the other direction, we proceed analogously to above and arrive at

$$
\mathbb{P}\left(\exists t \leq 2^{k+1} : -S_t \geq f\right) \leq \exp\left(-\sup_{\lambda\in\mathbb{R}}\left(-\lambda f - \ln\mathbb{E}[M_{2^{k+1}}]\right)\right)
$$

which we bound similarly to above by

$$
\exp\left(-\frac{f^2}{2(2^{k+1}\mu - f/3)}\right) \leq \exp\left(-\frac{f^2}{2(2^{k+1}\mu + f/3)}\right) \leq \exp(-g).
$$

□

Lemma F.3 (Uniform L1-Deviation Bound for Empirical Distribution)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. categorical variables on $[U]$ with distribution P. Then for all $\delta \in (0, 1]$*

$$
\mathbb{P}\left(\exists t \,:\, \|\hat{P}_t - P\|_1 \geq \sqrt{\frac{4}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3(2^U - 2)}{\delta}\right)}\right) \leq \delta
$$

where $\hat{P}_t$ is the empirical distribution based on samples $X_1 \ldots X_t$.

Proof. We use the identity $\|Q - P\|_1 = 2\max_{B\subseteq\mathcal{B}} Q(B) - P(B)$ which holds for all distributions P, Q defined on the finite set $\mathcal{B}$ to bound

$$
\begin{aligned}
&\mathbb{P}\left(\exists t \,:\, \|\hat{P}_t - P\|_1 \geq \sqrt{\frac{4}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3(2^U - 2)}{\delta}\right)}\right) \\
=&\mathbb{P}\left(\max_{t, B\subseteq[U]} \hat{P}_t(B) - P(B) \geq \frac{1}{2}\sqrt{\frac{4}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3(2^U - 2)}{\delta}\right)}\right) \\
\leq& \sum_{B\subseteq[U]} \mathbb{P}\left(\max_{t} \hat{P}_t(B) - P(B) \geq \sqrt{\frac{1}{t}\left(2\operatorname{llnp}(t) + \ln\frac{3(2^U - 2)}{\delta}\right)}\right).
\end{aligned}
$$

Define now $S_t = \sum_{i=1}^{t} \mathbb{I}\{X_1 \in B\} - tP(B)$ which is a martingale sequence. Then the last line above is equivalent to

$$\begin{aligned}
&\sum_{B \subseteq [U]} \mathbb{P}\left(\max_t S_t \geq \sqrt{t\left(2\,\mathrm{llnp}(t) + \ln \frac{3(2^U-2)}{\delta}\right)} \right) \\
&\leq \sum_{B \subseteq [U]} \mathbb{P}\left(\max_{k \in \mathbb{N}, t \in [2^k, 2^{k+1}]} S_t \geq \sqrt{t\left(2\,\mathrm{llnp}(t) + \ln \frac{3(2^U-2)}{\delta}\right)} \right) \\
&\leq \sum_{B \subseteq [U]} \sum_{k=0}^{\infty} \mathbb{P}\left(\max_{t \in [2^k, 2^{k+1}]} S_t \geq \sqrt{t\left(2\,\mathrm{llnp}(t) + \ln \frac{3(2^U-2)}{\delta}\right)} \right) \\
&\leq \sum_{B \subseteq [U]} \sum_{k=0}^{\infty} \mathbb{P}\left(\max_{t \leq 2^{k+1}} S_t \geq \sqrt{2^k\left(2\,\mathrm{llnp}(2^k) + \ln \frac{3(2^U-2)}{\delta}\right)} \right) \\
&= \sum_{B \subseteq [U]} \sum_{k=0}^{\infty} \mathbb{P}\left(\max_{t \leq 2^{k+1}} \exp(\lambda S_t) \geq \exp(\lambda f) \right) \\
&= \sum_{B \subseteq [U], B \neq \emptyset, B \neq [U]} \sum_{k=0}^{\infty} \mathbb{P}\left(\max_{t \leq 2^{k+1}} \exp(\lambda S_t) \geq \exp(\lambda f) \right)
\end{aligned}$$

where $f = \sqrt{2^k\left(2\,\mathrm{llnp}(2^k) + \ln \frac{3(2^U-2)}{\delta}\right)}$ and $\lambda \in \mathbb{R}$ and the last equality follows from the fact that for $B = \emptyset$ and $B = [U]$ the difference between the distributions has to be 0. Since $\mathbb{I}\{X_1 \in B\} - tP(B)$ is a centered Bernoulli variable it is $1/2$-subgaussian and so S_t satisfies $\mathbb{E}[\exp(\lambda S_t)] \leq \exp(\lambda^2 t/8)]$. Since S_t is a martingale, $\exp(\lambda S_t)$ is a nonnegative sub-martingale and we can apply the maximal inequality to bound

$$\mathbb{P}\left(\max_{t \leq 2^{k+1}} \exp(\lambda S_t) \geq \exp(\lambda f) \right) \leq \exp\left(\frac{1}{8}\lambda^2 2^{k+1} - \lambda f \right).$$

Choosing $\lambda = \frac{4f}{2^{k+1}}$, we get $\mathbb{P}\left(\max_{t \leq 2^{k+1}} \exp(\lambda S_t) \geq \exp(\lambda f)\right) \leq \exp\left(-\frac{f^2}{2^k}\right)$. Hence, using the same steps as in the proof of Lemma F.1, we get $\mathbb{P}\left(\max_{t \leq 2^{k+1}} \exp(\lambda S_t) \geq \exp(\lambda f)\right) \leq \frac{\delta}{3(2^{[U]}-2)} \min\left\{1, \frac{1}{k^2 \ln 2}\right\}$ and then

$$\begin{aligned}
&\mathbb{P}\left(\exists t : \|\hat{P}_t - P\|_1 \geq \sqrt{\frac{4}{t}\left(2\,\mathrm{llnp}(t) + \ln \frac{3(2^U-2)}{\delta}\right)} \right) \\
&\leq \sum_{B \subseteq [U], B \neq \emptyset, B \neq [U]} \frac{\delta}{3(2^{[U]}-2)} \sum_{k=0}^{\infty} \min\left\{1, \frac{1}{k^2 \ln 2}\right\} \leq \sum_{B \subseteq [U], B \neq \emptyset, B \neq [U]} \frac{\delta}{2^{[U]}-2} = \delta.
\end{aligned}$$

□

Lemma F.4. *Let $\mathcal{F}_i$ for $i = 1 \ldots$ be a filtration and $X_1, \ldots X_n$ be a sequence of Bernoulli random variables with $\mathbb{P}(X_i = 1 | \mathcal{F}_{i-1}) = P_i$ with P_i being $\mathcal{F}_{i-1}$-measurable and X_i being $\mathcal{F}_i$ measurable. It holds that*

$$\mathbb{P}\left(\exists n : \sum_{t=1}^{n} X_t < \sum_{t=1}^{n} P_t/2 - W \right) \leq e^{-W}$$

Proof. $P_t - X_t$ is a Martingale difference sequence with respect to the filtration $\mathcal{F}_t$. Since X_t is nonnegative and has finite second moment, we have for any $\lambda > 0$ that $\mathbb{E}\left[e^{-\lambda(X_t - P_t)} | \mathcal{F}_{t-1}\right] \leq e^{\lambda^2 P_t/2}$ (Exercise 2.9, Boucheron et al. [25]). Hence, we have

$$\mathbb{E}\left[e^{\lambda(P_t - X_t) - \lambda^2 P_t/2} | \mathcal{F}_{t-1}\right] \leq 1$$

and by setting $\lambda = 1$, we see that

$$M_n = e^{\sum_{t=1}^n (-X_t + P_t/2)}$$

is a supermartingale. It hence holds by Markov's inequality

$$\mathbb{P}\left(\sum_{t=1}^n (-X_t + P_t/2) \geq W\right) = \mathbb{P}\left(M_n \geq e^W\right) \leq e^{-W}\mathbb{E}[M_n] \leq e^{-W}$$

wich gives us the derised result

$$\mathbb{P}\left(\sum_{t=1}^n X_t \leq \sum_{t=1}^n P_t/2 - W\right) \leq e^{-W}$$

for a fixed n. We define now the stopping time $\tau = \min\{t \in \mathbb{N} : M_t > e^W\}$ and the sequence $\tau_n = \min\{t \in \mathbb{N} : M_t > e^W \vee t \geq n\}$. Applying the convergence theorem for nonnegative supermartingales (Theorem 5.2.9 in Durrett [26]), we get that $\lim_{t\to\infty} M_t$ is well-defined almost surely. Therefore, M_τ is well-defined even when $\tau = \infty$. By the optional stopping theorem for nonnegative supermartingales (Theorem 5.7.6 by Durrett [26]), we have $\mathbb{E}[M_{\tau_n}] \leq \mathbb{E}[M_0] \leq 1$ for all n and applying Fatou's lemma, we obtain $\mathbb{E}[M_\tau] = \mathbb{E}[\lim_{n\to\infty} M_{\tau_n}] \leq \liminf_{n\to\infty} \mathbb{E}[M_{\tau_n}] \leq 1$. Using Markov's inequality, we can finally bound

$$\mathbb{P}\left(\exists n : \sum_{t=1}^n X_t < \frac{1}{2}\sum_{t=1}^n P_t - W\right) \leq \mathbb{P}(\tau < \infty) \leq \mathbb{P}(M_\tau > e^W) \leq e^{-W}\mathbb{E}[M_\tau] \leq e^{-W}.$$

$\square$