Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds

Andrea Zanette [1]   Emma Brunskill [2]

Abstract

Strong worst-case performance bounds for episodic reinforcement learning exist but fortunately in practice RL algorithms perform much better than such bounds would predict. Algorithms and theory that provide strong problem-dependent bounds could help illuminate the key features of what makes a RL problem hard and reduce the barrier to using RL algorithms in practice. As a step towards this we derive an algorithm and analysis for finite horizon discrete MDPs with state-of-the-art worst-case regret bounds and substantially tighter bounds if the RL environment has special features but without apriori knowledge of the environment from the algorithm. As a result of our analysis, we also help address an open learning theory question (Jiang & Agarwal, 2018) about episodic MDPs with a constant upper-bound on the sum of rewards, providing a regret bound function of the number of episodes with no dependence on the horizon.

1. Introduction

In reinforcement learning (RL) an agent must learn how to make good decision without having access to an exact model of the world. Most of the literature for provably efficient exploration in Markov decision processes (MDPs) (Jaksch et al., 2010; Osband et al., 2013; Lattimore & Hutter, 2014; Dann & Brunskill, 2015; Dann et al., 2017; Osband & Roy, 2017; Azar et al., 2017; Kakade et al., 2018) has focused on providing near-optimal worst-case performance bounds. Such bounds are highly

[1]Institute for Computational and Mathematical Engineering, Stanford University, USA [2]Department of Computer Science, Stanford University, USA. Correspondence to: Andrea Zanette <zanette@stanford.edu>, Emma Brunskill <ebrun@cs.stanford.edu>.

desirable as they do not depend on the structure of the particular environment considered and therefore hold for even extremely hard-to-learn MDPs.

Fortunately in practice reinforcement learning algorithms often perform far better than what these problem-independent bounds would suggest. While we may observe better or worse performance empirically on different MDPs, we would like to derive a more systematic understanding of what types of decision processes are inherently easier or more challenging for RL. This motivates our interest in deriving algorithms and theoretical analyses that provide problem-dependent bounds. Ideally, such algorithms will do as well as RL solutions designed for the worst case if the problem is pathologically difficult and otherwise match the performance bounds of algorithms specifically designed for a particular problem subclass. This exciting scenario might bring considerable saving in the time spent designing domain-specific RL solutions and in training a human expert to judge and recognize the complexity of different problems. An added benefit would include the robustness of the RL solution in case the actual model does not belong to the identified subclass, yielding increased confidence to deploying RL to high-stakes applications.

Towards this goal, in this paper we contribute with a new algorithm for episodic tabular reinforcement learning which automatically provides provably stronger regret bounds in many domains which have a small variance of the optimal value function (in the infinite horizon setting, this variance has been called the *environmental norm* (Maillard et al., 2014)). Indeed, there is good reason to believe that some features of the range or variability of the optimal value function should be a critical aspect of the hardness of reinforcement learning in a MDP. Many worst-case bounds for finite-state MDPs scale with a *worst case* bound on the range / magnitude of the value function, such as the diameter D for an infinite-horizon setting and the horizon H in an episodic problem. Note that here both D and H arise in the analyses as upper bounds on the (range of the) *optimistic* value function across the entire MDP[1]. As more samples are collected,

---

[1]Many RL algorithms with strong performance bounds rely on the principle of optimism under uncertainty and compute an optimistic value function.

one would hope that the agent's optimistic value function converges to the true optimal value function. Unfortunately this is not the case, see for example (Jaksch et al., 2010; Bartlett & Tewari, 2009; Zanette & Brunskill, 2018) for a discussion of this. As a result, most prior analyses bounded the optimistic value function by generic quantities like D or H regardless of the actual behaviour of the optimal value function.

While the majority of formal performance guarantees has focused on bounds for the worst case, there have been several contributions of algorithms and/or theoretical analyses focused on MDPs with particular structure. Such contributions have focused on the infinite horizon setting, which involves a number of subtleties that are not present in the finite horizon setting we consider, which is likely a cause of the less strong results in this setting which can require stronger input knowledge on a tighter range on the possible value function (Bartlett & Tewari, 2009; Fruit et al., 2018), or do not match in dominant terms strong bounds for the worst case setting (Maillard et al., 2014). We defer more detailed discussion of related work to Section 7, except to briefly highlight likely the most closely related recent result from (Talebi & Maillard, 2018). Like us, Talebi and Maillard provide a problem-dependent regret bound that scales as a function of the variance of the next state distribution. However, like the aforementioned references, their focus is on the infinite horizon setting. In this setting the authors achieve their resulting regret bound under an assumption that the mixing time of the MDP is such that all states are visited at a linear rate in expectation regardless of the agent's chosen policies. This mixing rate, that could be exponential in certain MDPs, appears in the regret bound. In our, arguably simpler finite horizon setting, we do not use an assumption on the mixing rate of the MDP and we instead pursue a different proof technique to obtain strong results for this setting.

More precisely, in this paper we derive an algorithm for finite horizon discrete MDPs and associated analysis that yields state-of-the art worst-case regret bounds of order $\tilde{O}(\sqrt{HSAT})$ in the leading term while improving if the environment has next-state value function variance (i.e., small environmental norm) or bounded total possible reward. Compared to the existing literature, our work

- Maintains state of the art worst-case guarantees (Azar et al., 2017) for episodic finite horizon settings,

- Improves the regret bounds of (Zanette & Brunskill, 2018) when deployed in the same settings,

- Provides demonstration that characterizing problems using environmental norm (Maillard et al., 2014) can yield substantially tighter theoretical guarantees in the finite horizon setting,

- Identifies problem classes with low environmental norm which are of significant interest, including deterministic domains, single-goal MDPs, and high stochasticity domains, and

- Helps address an open learning theory problem (Jiang & Agarwal, 2018), showing that for their setting, we obtain a regret bound that scales with no dependence on the planning horizon in the dominant terms.

The paper is organized as follows: we recall some basic definitions in Section 2 and describe the algorithm in Section 3. We state and comment the main result in Section 4, discuss how this helps address an open learning theory problem in Section 5 and then describe selected problem-dependent bounds in Section 6. The analysis is sketched in Section 4.1. Due to space constraints, most proofs are in the full report available at:

https://arxiv.org/abs/1901.00210.

2. Preliminaries and Definitions

In this section we introduce some notation and definitions. We consider undiscounted finite horizon MDPs (Sutton & Barto, 1998), which are defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, H \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces with cardinality S and A, respectively. We denote by $p(s' \mid s, a)$ the probability of transitioning to state s' after taking action a in state s while $r(s, a) \in [0, 1]$ is the average instantaneous reward collected. We label with $n_k(s, a)$ the visits to the (s, a) pair at the beginning of the k-th episode. The agent interacts with the MDP starting from arbitrary initial states in a sequence of episodes $k \in [K]$ (where $[K] = \{j \in \mathbb{N} : 1 \leq j \leq K\}$) of fixed length H by selecting a policy $\tilde{\pi}_k$ which maps states s and timesteps t to actions. Each policy identifies a value function for every state s and timestep $t \in [H]$ defined as $V_t^{\tilde{\pi}_k}(s_t) = \mathbb{E}_{(s,a) \sim \tilde{\pi}_k} \sum_{i=t}^{H} r(s, a)$ which is the expected return until the end of the episode (the conditional expectation is over the pairs (s, a) encountered in the MDP upon starting from s_t). The optimal policy is indicated with π^* and its value function as $V_t^{\pi^*}$. We indicate with $\underline{V}_{t+1k}^{\tilde{\pi}_k}$ and $\overline{V}_{t+1k}^{\tilde{\pi}_k}$, respectively, a pointwise underestimate, respectively, overestimate, of the optimal value function and with $\hat{p}_k(\cdot \mid s, a)$ and $\hat{r}_k(s, a)$ the MLE estimates of $p(\cdot \mid s, a)$ and $r(s, a)$. We focus on deriving a high probability upper bound on the REGRET$(K) \overset{def}{=} \sum_{k \in [K]} \left(V_1^{\pi^*}(s_k) - V_1^{\tilde{\pi}_k}(s_k) \right)$ to measure the agent's learning performance. We use the $\tilde{O}(\cdot)$ notation to indicate a quantity that depends on $(\cdot)$ up to a polylog expression of a quantity at most polynomial in $S, A, T, K, H, \frac{1}{\delta}$. We also use the $\lesssim, \gtrsim, \simeq$ notation to mean $\leq, \geq, =$, respectively, up to a numerical constant and indi-

cate with $\|X\|_{2,p}$ the 2-norm of a random variable[2] under p, i.e., $\|X\|_{2,p} \overset{def}{=} \sqrt{\mathbb{E}_p X^2} \overset{def}{=} \sqrt{\sum_{s'} p(s') X^2(s')}$ if $p(\cdot)$ is its probability mass function.

3. EULER

We define the maximum per-step conditional variance (conditioning is on the (s, a) pair) for a particular MDP as $\mathbb{Q}^*$:

$$\mathbb{Q}^* \overset{def}{=} \max_{s,a,t} \left(\text{Var} \, R(s, a) + \underset{s^+ \sim p(s,a)}{\text{Var}} V_{t+1}^{\pi^*}(s^+) \right) \quad (1)$$

where $R(s, a)$ is the reward random variable in (s, a). This definition is identical to the environmental norm (Maillard et al., 2014) but here we will generally refer to it as the maximum conditional value variance, in order to connect with other work which explicitly bounds the variance.

We introduce the algorithm *Episodic Upper Lower Exploration in Reinforcement learning* (EULER) which adopts the paradigm of "optimism under uncertainty" to conduct exploration. Recent work (Dann & Brunskill, 2015; Dann et al., 2017; Azar et al., 2017) has demonstrated how the choice of the exploration bonus is critical to enabling tighter *problem-independent* performance bounds. Indeed minimax worst case regret bounds have been obtained by using a Bernstein-Friedman-type reward bonus defined over an empirical quantity related very closely to the conditional value variance $\mathbb{Q}^*$, plus an additional correction term necessary to ensure optimism (Azar et al., 2017).

Similarly, in our algorithm we use a bonus that combines an empirical Bernstein type inequality for estimating the $\mathbb{Q}^*$ conditional variance, coupled with a different correction term which explicitly accounts for the value function uncertainty. We provide pseudocode for EULER which details the main procedure in Figure 1. Notice that EULER has the same computational complexity as value iteration.

4. Main Result

Now we present our main result, which is a problem-dependent high-probability regret upper bound for EULER in terms of the underlying max conditional variance $\mathbb{Q}^*$ and maximum return. Crucially, EULER is **not** provided with $\mathbb{Q}^*$ and the value of the max return. We also prove a worst-case guarantee that matches the established (Osband & Van Roy, 2016; Jaksch et al., 2010) lower bound of $\Omega(\sqrt{HSAT})$ in the dominant term. We introduce the following definition:

Definition 1 (Max Return). *We define as $\mathcal{G} \in \mathbb{R}$ the maximum (random) return in an episode upon following any*

---

[2]To be precise, this is a norm between classes of random variables that are almost surely the same

policy π from any starting state s_0, i.e., the deterministic upper bound to:

$$\sum_{t=1}^{H} R(s_t, \pi(s_t)) \le \mathcal{G}, \quad \forall \pi, s_0. \quad (2)$$

where the states $s_1, \ldots, s_H$ are the (random) states generated upon following the trajectory identified by the policy π from s_0.

Theorem 1 (Problem Dependent High Probability Regret Upper Bound for EULER). *With probability at least $1 - \delta$ the regret of EULER is bounded for any time $T \le KH$ by the minimum between*

$$\tilde{O} \left(\sqrt{\mathbb{Q}^* SAT} + \sqrt{S} SAH^2 (\sqrt{S} + \sqrt{H}) \right) \quad (3)$$

and

$$\tilde{O} \left(\sqrt{\frac{\mathcal{G}^2}{H} SAT} + \sqrt{S} SAH^2 (\sqrt{S} + \sqrt{H}) \right), \quad (4)$$

jointly for all episodes $k \in [K]$.

While the maximum conditional variance $\mathbb{Q}^*$ is always upper bounded by $\mathcal{G}$ if rewards are positive and bounded, we include both forms of regret bound for two reasons. First, the second bound is tighter than naively upper bounding $\mathbb{Q}^* \le \mathcal{G}^2$ by a factor of H. Second, we will shortly see that both quantities can provide insights into which instances of MDP domains can have lower regret.

In addition, since the rewards are in $[0, 1]$, we immediately have that $\mathcal{G}^2 \le H^2$, and thereby obtain a worst-case regret bound expressed in the following corollary:

Corollary 1.1. *With probability at least $1 - \delta$ the regret of EULER is bounded for any time $T \le KH$ by*

$$\tilde{O} \left(\sqrt{HSAT} + \sqrt{S} SAH^2 (\sqrt{S} + \sqrt{H}) \right). \quad (5)$$

This matches in the dominant term the minimax regret problem independent bounds for tabular episodic RL settings (Azar et al., 2017). Therefore, the importance of our theorem 1 lies in providing problem dependent bounds (equation 3,4) while simultaneously matching the existing best worst case guarantees (equation 5). We shall shortly show that our results help address a recent open question on the performance dependence of episodic MDPs on the horizon (Jiang & Agarwal, 2018).

4.1. Sketch of the Theoretical Analysis

We devote this section to the sketch of the main point of the regret analysis that yields problem dependent bounds. Readers that wish to focus on how our results yield insight into the complexity of solving different problems may skip ahead to the next section. Central to the analysis is the relation between the agent's optimistic MDP and the "true"

Algorithm 1 EULER for Stationary Episodic MDPs

1: **Input**: $\delta' = \frac{1}{7}\delta$, $b_k^r(s,a) = \sqrt{\frac{2\widehat{\mathrm{Var}}R(s,a)\ln\frac{4SAT}{\delta'}}{n_k(s,a)}} + \frac{7\ln\frac{4SAT}{\delta'}}{3(n_k(s,a)-1)}$, $\phi(s,a) = \sqrt{\frac{2\widehat{\mathrm{Var}}_{\hat{p}_k(s,a)}(\overline{V}_{t+1k}^{\tilde{\pi}_k})\ln\frac{4SAT}{\delta'}}{n_k(s,a)}} + \frac{H\ln\frac{4SAT}{\delta'}}{3(n_k(s,a)-1)}$,

$B_p = H\sqrt{\frac{2\ln(4SAT)}{\delta'}}$, $B_v = \sqrt{\frac{2\ln(4SAT)}{\delta'}}$, $J = \frac{H\ln(4SAT)/\delta'}{3}$.

2: **for** $k = 1, 2, \dots$ **do**
3:    **for** $t = H, H-1, \dots, 1$ **do**
4:       **for** $s \in \mathcal{S}$ **do**
5:          **for** $a \in \mathcal{A}$ **do**
6:             $\hat{p} = \frac{p_{sum}(\cdot, s, a)}{n_k(s,a)}$
7:             $b_k^{pv} = \phi(\hat{p}(s,a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s,a)}}\left(\frac{4J+B_p}{\sqrt{n_k(s,a)}} + B_v\|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}}\right)$
8:             $Q(a) = \min\{H-t, \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b_k^r(s,a) + \hat{p}^\top \overline{V}_{t+1} + b_k^{pv}\}$
9:          **end for**
10:          $\tilde{\pi}_k(s,t) = \arg\max_a Q(a)$
11:          $\overline{V}_t(s) = Q(\tilde{\pi}_k(s,t))$
12:          $b_k^{pv} = \phi(\hat{p}(s, \tilde{\pi}_k(s,t)), \underline{V}_{t+1}) + \frac{1}{\sqrt{n(s,\tilde{\pi}_k(s,t))}}\left(\frac{4J+B_p}{\sqrt{n_k(s,\tilde{\pi}_k(s,t))}} + B_v\|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}}\right)$
13:          $\underline{V}_t(s) = \max\{0, \hat{r}_k(s, \tilde{\pi}_k(s,t)) - b_k^r(s, \tilde{\pi}_k(s,t)) + \hat{p}^\top \underline{V}_{t+1} - b_k^{pv}\}$
14:       **end for**
15:    **end for**
16:    Evaluate policy $\tilde{\pi}_k$ and update MLE estimates $\hat{p}(\cdot, \cdot)$ and $\hat{r}(\cdot, \cdot)$
17: **end for**

MDP. A more detailed overview of the proof is given in section C of the appendix, while the rest of the appendix presents the detailed analysis under a more general framework.

Regret Decomposition Denote with $\mathbb{E}_{(s,a)\sim\tilde{\pi}_k}$ the expectation taken along the trajectories identified by the agent's policy $\tilde{\pi}_k$. A standard regret decomposition is given below (see (Dann et al., 2017; Azar et al., 2017)):

$$\mathrm{REGRET}(K) \le \sum_{\substack{k\in[K] \\ t\in[H] \\ (s,a)\in\mathcal{S}\times\mathcal{A}}} \mathbb{E}_{(s,a)\sim\tilde{\pi}_k}\left(\underbrace{\tilde{r}_k(s,a) - r(s,a)}_{\substack{\text{REWARD} \\ \text{ESTIMATION} \\ \text{AND OPTIMISM}}}\right.$$

$$+ \underbrace{(\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k}}_{\substack{\text{TRANSITION} \\ \text{DYNAMICS} \\ \text{OPTIMISM}}}$$

$$+ \underbrace{(\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a))^\top V_{t+1}^{\pi^*}}_{\substack{\text{TRANSITION} \\ \text{DYNAMICS} \\ \text{ESTIMATION}}}$$

$$\left. + \underbrace{(\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*}\right)}_{\substack{\text{LOWER} \\ \text{ORDER} \\ \text{TERM}}}\right) \quad (6)$$

Here, the "tilde" quantities $\tilde{r}$ and $\tilde{p}$ represent the agent's optimistic estimate. Of the terms in equation 6, the "Transition Dynamics Estimation" and "Transition Dynamics Op-

timism" are the leading terms to bound as far as the regret is concerned. The former is expressed through MDP quantities (i.e, the true transition dynamics $p(\cdot \mid s,a)$ and the optimal value function $V_{t+1}^{\pi^*}$) and hence it can be readily bounded using Bernstein Inequality, giving rise to a problem dependent regret contribution. More challenging is to show that a similar simplification can be obtained for the "Transition Dynamics Optimism" term which relies on the agent's optimistic estimates $\tilde{p}_k(\cdot \mid s,a)$ and $\overline{V}_{t+1k}^{\tilde{\pi}_k}$.

Optimism on the System Dynamics Said term $(\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k}$ represents the difference between the agent's imagined (i.e., optimistic) transition $\tilde{p}_k(\cdot \mid s,a)$ and the maximum likelihood transition $\hat{p}_k(\cdot \mid s,a)$ weighted by the next-state optimistic value function $\overline{V}_{t+1k}^{\tilde{\pi}_k}$. By construction, this is the exploration bonus which incorporates an estimate of the conditional variance over the value function. This bonus reads:

$$\underset{\substack{\text{TRANSITION} \\ \text{DYNAMICS} \\ \text{OPTIMISM}}}{} = \underset{\substack{\text{EXPLO-} \\ \text{RATION} \\ \text{BONUS}}}{} \approx \underbrace{\overbrace{\sqrt{\frac{\mathrm{Var}_{s\sim\hat{p}_k(\cdot\mid s,a)}\overline{V}_{t+1k}^{\tilde{\pi}_k}}{n_k(s,a)}}}^{\substack{\text{DOMINANT TERM} \\ \text{OF EXPLORATION BONUS}}} + \frac{H}{n_k(s,a)}}_{\substack{\text{EMPIRICAL BERNSTEIN evaluated} \\ \text{WITH EMPIRICAL VALUE FUNCTION}}}$$

$$(7)$$

$$+ \underbrace{\left(\frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{\hat{p}_k(\cdot|s,a)}}{\sqrt{n_k(s,a)}} + \frac{H}{n_k(s,a)} \right)}_{\text{CORRECTION BONUS}} \quad (8)$$

In the above expression the "Correction Bonus" is needed to ensure optimism because the "Empirical Bernstein" contribution is evaluated with the agent's estimate $\overline{V}_{t+1k}^{\tilde{\pi}_k}$ as opposed to the real $V_{t+1}^{\pi^*}$. If we assume that $\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{\hat{p}_k(\cdot|s,a)}$ shrinks quickly enough, then the "Dominant Term" in equation 7 is the most slowly decaying term with a rate $1/\sqrt{n}$. If that term involved the true transition dynamics $p(\cdot \mid s,a)$ and value function $V_{t+1}^{\pi^*}$ (as opposed to the agent's estimates $\hat{p}_k(\cdot \mid s,a)$ and $\overline{V}_{t+1k}^{\tilde{\pi}_k}$) then problem dependent bounds would follow in the same way as they could be proved for the "Transition Dynamics Estimation". Therefore we wish to study the relation between such "Dominant Term" evaluated with the agent's MDP estimates vs the MDP's true parameters.

Convergence of the System Dynamics in the Dominant Term of the Exploration Bonus Theorem 10 of (Maurer & Pontil, 2009) gives the high probability statement:

$$\left| \sqrt{\underset{\hat{p}_k(\cdot|s,a)}{\text{Var}} V_{t+1}^{\pi^*}} - \sqrt{\underset{p(\cdot|s,a)}{\text{Var}} V_{t+1}^{\pi^*}} \right| \lessapprox \frac{H}{n_k(s,a)} \quad (9)$$

to quantify the rate of convergence of the empirical variance using the true value function (this leads to the empirical version of Bernstein's inequality). Next, two basic computations yield:

$$\left| \sqrt{\underset{\hat{p}_k(\cdot|s,a)}{\text{Var}} V_{t+1}^{\pi^*}} - \sqrt{\underset{\hat{p}_k(\cdot|s,a)}{\text{Var}} \overline{V}_{t+1k}^{\tilde{\pi}_k}} \right|$$
$$\leq \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*}\|_{\hat{p}_k(\cdot|s,a)} \leq \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{\hat{p}_k(\cdot|s,a)} \quad (10)$$

Together, equation 9 and 10 quantify the rate of convergence of $\text{Var}_{s \sim \hat{p}_k(\cdot|s,a)} \overline{V}_{t+1k}^{\tilde{\pi}_k}$ to $\text{Var}_{s \sim p(\cdot|s,a)} V_{t+1}^{\pi^*}$, yielding the following upper bound for the dominant term of the exploration bonus:

$$\underset{\substack{\text{DOMINANT} \\ \text{TERM OF} \\ \text{EXPLORATION} \\ \text{BONUS}}}{} = \sqrt{\frac{\text{Var}_{\hat{p}_k(\cdot|s,a)} \overline{V}_{t+1k}^{\tilde{\pi}_k}}{n_k(s,a)}} \lessapprox \underbrace{\sqrt{\frac{\text{Var}_{p(\cdot|s,a)} V_{t+1}^{\pi^*}}{n_k(s,a)}}}_{\substack{\text{GIVES PROBLEM} \\ \text{DEPENDENT BOUNDS}}}$$

$$+ \underbrace{\frac{H}{n_k(s,a)} + \frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{\hat{p}_k(\cdot|s,a)}}{\sqrt{n_k(s,a)}}}_{\text{SHRINKS FASTER}} \quad (11)$$

In words, we have decomposed the "Dominant Term of the Exploration Bonus" (which is constructed using the agent's

available knowledge) as a problem-dependent contribution (that is equivalent to Bernstein Inequality evaluated as if the model was known) and a term that accounts for the distance between the the true and empirical model, expressed as (computable) upper and lower bounds on the value function. This additional term shrinks faster that the former. It is precisely this "Correction Bonus" that we use in equation 7 and in the definition of the Algorithm itself.

What gives rise to problem dependent bounds? Our analysis highlights EULER uses a Bernstein inequality on the empirical estimate of the conditional variance of the next state values, with a correction term $\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{\hat{p}_k(\cdot|s,a)}$ function of the inaccuracy of the value function estimate at the next-step states re-weighted by their relative importance as encoded in the experienced transitions $\hat{p}_k(\cdot \mid s,a)$. Said correction term is of high value only if the successor states do not have an accurate estimate for the value function *and* they are going to be visited with high probability. A pigeonhole argument guarantees that this situation cannot happen for too long ensuring fast decay of $\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{\hat{p}_k(\cdot|s,a)}$ and therefore of the whole "Correction Bonus" of eq. 7.

Our primary analysis yields a regret bound that scales directly with the (unknown to the algorithm) problem-dependent $\mathbb{Q}^*$ max conditional variance of the next state values. We further extend this to a bound directly in terms of the max returns $\mathcal{G}$ by using a law of total variance argument.

Notice that such considerations and results would not be achievable by a naive application of an Hoeffding-like inequality as the latter would put equal weight on all successor states, but the accuracy in the estimation of $V_{t+1}^{\pi^*}(\cdot)$ only shrinks in a way that depends on the visitation frequency of said successor states as encoded in $\hat{p}_k(\cdot \mid s,a)$. The key to enable problem dependent bound is, therefore, to re-weight the importance of the uncertainty on the value function of the successor states by the corresponding visitation probability, which Bernstein Inequality implicitly does.

There exist other algorithms (e.g. (Dann & Brunskill, 2015; Azar et al., 2017)) which are based on Bernstein's inequality but to our knowledge they have not been analyzed in a way that provably yields problem dependent bounds as those presented here.

5. Horizon Dependence in Dominant Term

In this section we show that our result can help address a recently posed open question in the learning theory community (Jiang & Agarwal, 2018). The question posed centers on the whether there should exist a necessary dependence of sample complexity and regret lower bounds on the plan-

ning horizon H for episodic tabular MDP reinforcement learning tasks. Existing lower bound results for sample complexity (Dann & Brunskill, 2015) depend on the horizon, as do the best existing minimax regret bounds under asymptotic assumptions (Azar et al., 2017). However, such results have been derived under the common assumption of reward uniformity, that per-time-step rewards are bounded between 0 and 1, yielding a total value bounded by 0 and H. Jiang & Agarwal (2018) instead pose a more general setting, in which they assume that the rewards are positive and $\sum_{h=1}^{H} r_h \in [0, 1]$ holds almost surely: note the standard setting of reward uniformity can be expressed in this setting by first normalizing all rewards by dividing by H. The authors then ask that if in this new, more general setting of tabular episodic RL there is necessarily a dependence on the planning horizon in the lower bounds. Note that in this setting, the prior existing lower bounds on the sample complexity (Dann & Brunskill, 2015) would yield no dependence on the horizon.

For our work, the setting of Jiang and Agarwal immediately implies that

$$0 \le V_t^{\pi^*}(s) \le \mathcal{G} \le 1, \ \forall (s, t) \in \mathcal{S} \times [H]. \quad (12)$$

Further, since $V_t^{\pi^*}(s) \le 1$ and $r(s, a) \ge 0$ we must have $r(s, a) \in [0, 1]$, which is the assumption of this work. Therefore our main result (theorem 1) applies here. Recalling that $T = KH$, we obtain an upper bound on regret as

$$\tilde{O}\left(\sqrt{SAK} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})\right). \quad (13)$$

Note that the planning/episodic horizon H does not appear in the dominant regret term which scales polynomially with the number of episodes[3] K, and only appears in transient lower order terms that are independent of K.

In other words, **up to logarithmic dependency and transient terms, we have an upper bound on episodic regret that is independent of the horizon** H. This result answers part of Jiang and Agarwal's open question: for their setting, the regret primarily scales independently of the horizon.

Surprisingly, while EULER uses a common problem-agnostic bound on the maximum possible optimal value function (H), it does not need to be provided with information about the domain-dependent maximum possible value function to attain the improved bound in the setting of the COLT conjecture of Jiang & Agarwal (2018).

It remains an open question whether we could further avoid either a dependence on the planning horizon in the transient terms as well as obtaining a PAC result. In Appendix B we further discuss this direction. However, these results are promising: they suggest that the hardness of learning in sparse reward, and long horizon episodic MDPs may not be

---

[3]This is stronger than scaling polynomially with the time T

fundamentally much harder than shorter horizon domains if the total reward is bounded.

6. Problem dependent bounds

We now focus on deriving regret bounds for selected MDP classes that are very common in RL. We emphasize that such setting-dependent guarantees are obtained with the same algorithm that is not informed of a particular MDP's values of $\mathbb{Q}^*$ and $\mathcal{G}$. Although the described settings share common features and are sometimes subclasses of one another, they are in separate subsections due to their important relation to the past literature and their practical relevance. Importantly, they are *all characterized by low $\mathbb{Q}^*$*.

6.1. Bounds using the range of optimal value function

To improve over the worst case bound in infinite horizon RL there have been approaches that aim at obtaining stronger problem dependent bounds if the value function does not vary much across different states of the MDP. If $\mathrm{rng}\, V^{\pi^*}$ is smaller than the worst-case (either H or D for the fixed horizon vs recurrent RL), the reduced variability in the expected return suggests that performance can benefit from constructing tighter confidence intervals. This is achieved by Bartlett & Tewari (2009) by providing this range to their algorithm REGAL, achieving a regret bound:

$$\tilde{O}(\Phi S \sqrt{AT}) \quad (14)$$

where $\Phi \ge \mathrm{rng}\, V^{\pi^*}$ is an overestimate of the optimal value function range and is an input to the algorithm described in that paper. This means that if domain knowledge is available and is supplied to the algorithm the regret can be substantially reduced. This line of research was followed in (Fruit et al., 2018) which derived a computationally-tractable variant of REGAL. However, they still require knowledge of a value function range upper bound $\Phi \ge \mathrm{rng}\, V^{\pi^*}$. Specifying a too high value for Φ would increase the regret and a too low value would cause the algorithm to fail.

Our analysis shows that, in the episodic setting, it is possible to achieve at least the same but potentially much better level of performance *without knowing the optimal value function range*. This follows as an easy corollary of our main regret upper bound (Theorem 1) after bounding the environmental norm, as we discuss below.

Let $\mathcal{S}_{s,a}$ be the set of immediate successor states after one transition from state s upon taking action a there, that is, the states in the support of $p(\cdot \mid s, a)$ and define

$$\Phi_{succ} \stackrel{def}{=} \max_{s,a} \ \mathrm{rng}_{s^+ \in \mathcal{S}_{s,a}} \ V_{t+1}^{\pi^*}(s^+) \quad (15)$$

as the maximum value function range *when restricted to the immediate successor states*. Since the variance is upper

bounded by (one fourth of) the square range of a random variable we have that:

$$\mathbb{Q}^* \overset{def}{=} \max_{s,a,t}(\text{Var}\,(R(s,a)\mid(s,a)) + \underset{s^+\sim p(s,a)}{\text{Var}}\,V_{t+1}^{\pi^*}(s^+))$$

$$\leq \max_{s,a,t}\left(1 + (\underset{s^+\in\mathcal{S}_{s,a}}{\text{rng}}\,V_{t+1}^{\pi^*}(s^+))^2\right) \leq 1 + \Phi_{succ}^2.$$

This immediately yields:

Corollary 1.2 (Bounded Range of V^{π^*} Among Successor States). *With probability at least $1-\delta$, the regret of* EULER *is bounded by:*

$$\tilde{O}(\Phi_{succ}\sqrt{SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})). \qquad (16)$$

A few remarks are in order:

- EULER does not need to know the value of Φ_{succ} or of the environmental norm or of the value function range to attain the improved bound;

- Φ_{succ} can be much smaller than $\text{rng}\,V^{\pi^*}$ because it is the range of V^{π^*} restricted to few successor states as opposed to across the whole domain, and therefore it is always smaller than Φ, in other words: $\Phi \geq \text{rng}\,V^{\pi^*} \geq \Phi_{succ}$.

- (Bartlett & Tewari, 2009; Fruit et al., 2018) consider the more challenging infinite horizon setting, while our results holds for fixed horizon RL.

6.2. Bounds on the next-state variance of V^{π^*} and empirical benchmarks

The environmental norm also can empirically characterize the hardness of RL in single problem instances. This was one of the key contributions of the work that introduced the environmental norm (Maillard et al., 2014), which evaluated the environmental norm for a number of common RL benchmark simulation tasks including mountain car, pinball, taxi, bottleneck, inventory and red herring In these domains the environmental norm is correlated with the complexity of reinforcement learning in these environments, as evaluated empirically. Indeed, in these settings, the environmental norm is often much smaller then the maximum value function range, which can itself be much smaller than the worst-case bound D or H. Our new results provide solid theoretical justification for the observed empirical savings.

This measure of MDP complexity also intriguingly allows us to gain more insight on another important simulation domain, chain MDPs like that in Figure 1. Chain MDPS have been considered a canonical example of challenging hard-to-learn RL domains, since naive strategies like ϵ greedy can take exponential time to attain satisfactory performance. By setting for simplicity $N \overset{def}{=} S = H$ EULER

provides an upper regret bound of $\tilde{O}(\sqrt{NAK}+\dots)$ that is substantially tighter than a worst case bound $\tilde{O}(\sqrt{N^3AK}+\dots)$, at least for large K. This is intriguing because it suggests pathological MDPs may be even less common than expected. More details about this example are in appendix A.1.
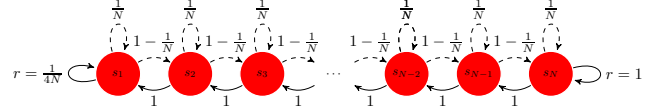

Figure 1. Classical hard-to-learn MDP

6.3. Stochasticity in the system dynamics

In this section we consider two important opposite classes of problems: deterministic MDPs and MDPs that are highly stochastic in that the successor state is sampled from a fixed distribution. These bounds are also a direct consequence of Theorem 1 and can be deduced from Corollary 1.2.

Deterministic domains Many problems of practical interest, for example in robotics, have low stochasticity, and this immediately yields low value for $\mathbb{Q}^*$. As a limit case, we consider domains with deterministic rewards and dynamics models. An agent designed to learn deterministic domains only needs to experience every transition *once* to reconstruct the model, which can take up to $O(SA)$ episodes with a regret at most $O(SAH)$(Wen & Van Roy, 2013).

Note in deterministic domains $\mathbb{Q}^* = 0$. Therefore if EULER is run on *any* deterministic MDP then the regret expression exhibits a $\log(T)$ dependence. This is a substantial improvement over prior RL regret bounds for problem-independent settings all have at least a $\sqrt{T}$ dependence. Further, a refined analysis (Appendix Section H.3) shows EULER is close to the lower bound except for a factor in the horizon and logarithmic terms:

Proposition 1. *If* EULER *is run on a deterministic MDP then the regret is bounded by* $\tilde{O}(SAH^2)$.

Highly mixing domains Recently, (Zanette & Brunskill, 2018) show that it is possible to design an algorithm that can switch between the MDP and the contextual bandit framework while retaining near-optimal performance in both without being informed of the setting. They consider mapping contextual bandit to an MDP whose transitions to different states (or contexts) are sampled from a fixed underlying distribution over which the agent has no control.

The Bandit-MDP considered in Zanette & Brunskill (2018) is an environment with high stochasticity (the MDP is

highly mixing since every state can be reached with some probability in one step). Since the transition function is unaffected by the agent, an easy computation yields $\mathrm{rng}\, V_t^{\pi^*} \leq 1$, as replicated in Appendix A.2. A regret guarantee in the leading order term of order $\tilde{O}(\sqrt{SAT})$ for EULER which matches the established lower bound for tabular contextual bandits (Bubeck & Cesa-Bianchi, 2012) follows from corollary 1.2. This is useful since in many practical applications it is unclear in advance if the domain is best modeled as a bandit or a sequential RL problem. Our results improve over (Zanette & Brunskill, 2018) since EULER has better worst-case guarantees by a factor of $\sqrt{H}$. Our approach is also feasible with next-state distributions that have zero or near zero mass over some of the next states: in contrast to prior work, the inverse minimum visitation probability does not show up in our analysis.

7. Related Literature

In infinite horizon RL, prior empirical evaluation of $\mathbb{Q}^*$ in (Maillard et al., 2014) has shown encouraging performance in a number of common benchmarks that $\mathbb{Q}^*$ has small value and its size relates to the hardness of solving the RL task. The theoretical results provides a regret bound whose leading order term is $\tilde{O}\left(\frac{1}{p_0}DS\sqrt{\mathbb{Q}^*AT}\right)$ (where p_0 is the minimum (non-zero) transition probability), and generally does not improve over worst case analysis for the infinite horizon setting. Our algorithm operates in an easier setting (finite horizon) where it can improve over the worst case, but it is an open question whether an improvement is possible in infinite horizon.

Our connection with (Bartlett & Tewari, 2009; Fruit et al., 2018; Zanette & Brunskill, 2018) has already been described. Here we focus on the remaining literature. We again note that the infinite horizon setting offers a number of important complexities and comparisons to the finite horizon setting (as considered here) cannot be directly made; however, as some of the closest related work lies in the infinite horizon setting, we briefly discuss it here.

- Bounds that depend on gap between policies: In the infinite horizon setting, (Even-Dar et al., 2006) has bounds dependent on the minimum gap in the optimal state action values between the best and second best action, and UCRL2 (Jaksch et al., 2010) has bounds as function of the gap in the difference in the average reward between the best and second best policies. Such gaps reflect an interesting alternate structure in the problem domain: note that in prior work as these gaps become arbitrarily small, the bound approaches infinity: even in such settings, if the next state variance is small, our bound will stay bounded. An interesting future direction is to consider bounds that consider both forms of structure.

- Regret bounds with value function approximation: In finite horizon settings, (Osband & Roy, 2014) uses the Eluder dimension as a measure of the dimensionality of the problem and (Jiang et al., 2017) proposes the Bellman rank to measure the learning complexity. Such measures capture a different notion of hardness than ours and do not match the lower bound in tabular settings.

- Infinite horizon results with additional properties of the transition model: the most closely related work to ours is (Talebi & Maillard, 2018) who also develop tighter regret bounds as a function of the next-state variance, but for infinite horizon settings. Exploration in such settings is nontrivial and the authors leverage an important assumption of ergodicity (which has also been considered in (Auer & Ortner, 2006). Specifically the agent will visit every state regardless of the current policy, and the rate of this mixing appears in the regret bound. An interesting and nontrivial question is whether our results can be extended to this setting without additional assumptions on the mixing structure of the domain.

A natural additional question is whether prior algorithms also inherit strong problem dependent rounds. Indeed, recent work by (Dann & Brunskill, 2015; Azar et al., 2017) has also used the variance of the value function at the next state in their analysis, though their final results are expressed as worst case bounds. However, the actual bonus terms used in their algorithms are distinct from our bonus terms, perhaps most significantly in that we maintain and leverage point-wise upper and lower bounds on the value function. While it is certainly possible that their algorithms or others already attain some form of problem dependent performance, they have not been analyzed in a way that yields problem dependent bounds. This is a technical area, and performing such analyses is a non-trivial deviation from a worst-case analysis. For example, the current worst case bounds from Azar et al. (2017) yield a regret bound that scales as $\tilde{O}(\sqrt{HSAT} + \sqrt{H^2T} + S^2AH^2)$ and it is a non-trivial extension to analyze how each of these terms might change to reflect problem-dependent quantities. One of our key contributions is an analysis of the rate of convergence of the empirical quantities about properties of the underlying MDP to the real ones in determining the regret bound.

8. Future Work and Conclusion

In this paper we have proposed EULER, an algorithm for episodic finite MDPs that matches the best known worst-case regret guarantees while provably obtaining much tighter guarantees if the domain has a small variance of the value function over the next-state distribution $\mathbb{Q}^*$ or a small bound in the possible achievable reward. EULER does not need to know these MDP-specific quantities in ad-

vance. We show that $\mathbb{Q}^*$ is low for a number of important subclasses of MDPs, including: MDPs with sparse rewards, (near) deterministic MDPs, highly mixing MDPs (such as those closer to bandits) and some classical empirical benchmarks. We also show how our result helps answer a recent open learning theory question about the necessary dependence of regret results on the episode horizon. Possible interesting directions for future work would be to examine problem-dependent bounds in the infinite horizon setting, incorporate a gap-dependent analysis, or see if such ideas could be extended to the continuous state setting.

Acknowledgments

References

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. In *NIPS*, 2006.

Azar, M., Munos, R., and Kappen, H. J. On the sample complexity of reinforcement learning with a generative model. In *ICML*, 2012.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *ICML*, 2017.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *NIPS*, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *NIPS*, 2017.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *ICML*, 2019.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit

and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. https://arxiv.org/abs/1802.04020, 2018.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

Jiang, N. and Agarwal, A. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pp. 3395–3398, 2018.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langforda, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *ICML*, 2017.

Kakade, S., Wang, M., and Yang, L. F. Variance reduction methods for sublinear reinforcement learning. *Arxiv*, 2018.

Lattimore, T. and Hutter, M. Near-optimal pac bounds for discounted mdps. In *Theoretical Computer Science*, 2014.

Maillard, O.-A., Mann, T. A., and Mannor, S. "how hard is my mdp?" the distribution-norm to the rescue. In *NIPS*, 2014.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. In *COLT*, 2009.

Osband, I. and Roy, B. V. Model-based reinforcement learning and the eluder dimension. In *NIPS*, 2014.

Osband, I. and Roy, B. V. Why is posterior sampling better than optimism for reinforcement learning? In *ICML*, 2017.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. In *Arxiv*, 2016. URL https://arxiv.org/pdf/1608.02732.pdf. https://arxiv.org/pdf/1608.02732.pdf.

Osband, I., Van Roy, B., and Russo, D. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, 2013.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Talebi, M. and Maillard, O. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, 2018.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. Technical report, Hewlett-Packard Labs, 2003.

Wen, Z. and Van Roy, B. Efficient exploration and value function generalization in deterministic systems. In *NIPS*, 2013.

Zanette, A. and Brunskill, E. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *ICML*, 2018.

Contents

Remark on constants: throughout the appendix we use numerical constants $c_{i,j,k}$ and $\tilde{L} = \log(SAHT/\delta)$, leaving their computation implicit.

A. Short Proofs Missing from the Main Text

A.1. Euler on Chain

Chain MDPs are commonly given as examples of challenging exploration domains because simple strategies like ϵ-greedy can take an exponential time to learn. We now discuss an intriguing result for the chain shown in figure 1 which is nearly identical to a previously introduced one (Osband & Roy, 2017). Like in that domain, each episode is $N = H = S$ timesteps long. The optimal policy is to go right which yields a reward of 1 for the episode. The transition probabilities are assigned in way that scales the optimal value function so that it is of order 1. Since the rewards are deterministic we immediately obtain (up to a constant that does not depend on N):

$$\mathbb{Q}^* \le \max_{s,a,t} \mathrm{Var}\left(\mathbb{Q}(s,a) \mid (s,a)\right) \le \max_{s,a,t} \mathrm{Var}\left(V_{t+1}^{\pi^*}(s^+) \mid (s,a)\right) \lesssim \frac{1}{N}\left(1 - \frac{1}{N}\right) \le \frac{1}{N} \tag{17}$$

since $V_{t+1}^{\pi^*}(s^+)$ is essentially dominated by a Bernoulli random variable with success parameter $(1 - 1/N)$ times an appropriate scaling factor of order one. Therefore EULER's regret[4] is dominated by a term which is $\tilde{O}\left(\sqrt{\mathbb{Q}^* SAT}\right) = \tilde{O}\left(\sqrt{\frac{1}{N} \times N \times A \times T}\right) = \tilde{O}\left(\sqrt{AT}\right)$. Notice that the lower order term should be added to the above expression and this is likely to be significant particularly for small T. However, we remark that the result above follows directly from Theorem 1 whose proof does not attempt to make the lower order term problem-dependent. Our result is substantially smaller than the typically reported bounds for this case, which are dominated by a $\tilde{O}(\sqrt{HSAT})$ term.

There are two main factors that lead to the above simplification for this class of MDPs: V^{π^*} is of order 1 and not $N = H$ like in hard-to-learn MDPs which yields the lower bounds (Dann & Brunskill, 2015) and also the variance decreases as we let N increase, each of which "removes" a factor of $\sqrt{N}$ from the known worst-case bound $\tilde{O}(\sqrt{HSAT}) = \tilde{O}(\sqrt{N^2 AT})$ after substituting $N = S = H$.

To enable EULER's level of performance on this (and other) MDPs one needs to carefully control both the confidence intervals of the rewards and of the transition probabilities (as EULER does), since Hoeffding-like concentration inequalities for the rewards alone would already induce a $\Theta(\sqrt{SAT}) = \Theta(\sqrt{NT})$ contribution to the regret expression.

This result is intriguing because it suggests that truly pathological MDP classes (which induce $\Omega(\sqrt{HSAT})$ regret) are even more uncommon than previously thought.

A.2. EULER on Tabular Contextual Bandits

Contextual multi-armed bandits are a generalization of the multiarmed bandit problem, see for example (Bubeck & Cesa-Bianchi, 2012) for a survey. In their simplest possible formulation they entail a discrete set of contexts or states $\{s = 1, 2, \dots\}$ and actions $\{a = 1, 2, \dots\}$ and the expected reward $r(s, a)$ depends on both the state and action. After playing an action, the agent transitions to the next states according to some fixed distribution $\mu \in \mathbb{R}^{|\mathcal{S}|}$ over which the agent has no control.

In principle, such problem can be recast as an MDP in which the next state is independent of the prior state and action. Consider an H-horizon MDP which maps to a tabular contextual bandit problem: the transition probability is identical $p(s'|s,a) = \mu(s')$ for all states and actions, where μ is a fixed distribution from which the next states are sampled. In such MDP Define the "best" and "worst" context at time t, respectively: $\overline{s}_t \stackrel{def}{=} \arg\max V_t^*(s)$ and $\underline{s}_t \stackrel{def}{=} \arg\min V_t^*(s)$ and recall that the transition dynamics $p(\cdot \mid s, a) = \mu$ depends nor on the action a nor on the current state s. We have that:

$$\begin{cases} V_t^*(\overline{s}_t) = \max_a \left(r(\overline{s}_t, a) + \mu^\top V_{t+1}^*\right) \\ V_t^*(\underline{s}_t) = \max_a \left(r(\underline{s}_t, a) + \mu^\top V_{t+1}^*\right) \end{cases} \tag{18}$$

---

[4]This is valid for any class of MDP which shares these properties, implying that the regret for the MDP in figure 1 can be even smaller

which immediately yields a bound on the range of the value function of the successor states:

$$\text{rng } V_{t+1}^{\pi^*} = V_t^*(\overline{s}_t) - V_t^*(\underline{s}_t) = \max_a r(\overline{s}_t, a) - \max_a r(\underline{s}_t, a) \leq 1 \tag{19}$$

where the last inequality follows from the fact that rewards are bounded $r(\cdot, \cdot) \in [0, 1]$. Therefore $\Phi \leq 1$ and corollary 1.2 yields a high probability regret upper bound of order

$$\tilde{O}\left(\sqrt{SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})\right) \tag{20}$$

for EULER. This means that EULER automatically attains the lower bound in the dominant term for tabular contextual bandits (Bubeck & Cesa-Bianchi, 2012) if deployed in such setting. We did not try to improve the lower order terms for this specific setting, which may give an improved bound.

B. Average Per-Episode Sample Complexity for the Setting of (Jiang & Agarwal, 2018)

Jiang and Agarwal (2018) also ask about the dependence of a lower bound on the sample complexity on the planning horizon. While our work focuses on a regret analysis, and does not provide PAC sample complexity results, we can use our regret results to bound with high probability the number of episodes needed to ensure that the *average* per-episode regret is less than ϵ To do so, we obtain the average per-episode loss of EULER by dividing by the number of episodes K:

$$\tilde{O}\left(\left(\frac{\sqrt{SA}}{\sqrt{K}} + \frac{\sqrt{S}SAH^2(\sqrt{S} + \sqrt{H}))}{K}\right)\right). \tag{21}$$

From here we can seek for the smallest K such that the average error is smaller than ϵ, obtaining:

$$\tilde{O}\left(\left(\frac{SA}{\epsilon^2} + \frac{\sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})}{\epsilon}\right)\right) \tag{22}$$

episodes before the *average per-episode* error is smaller than ϵ. For small $\epsilon << SA$, the first term dominates, which is again independent of a polynomial dependence on H. Of course, in order to formally obtain a PAC result (a worst case upper bound on the number of ϵ-suboptimal episodes) the algorithm would need to be modified to be PAC. In particular, the exploration bonus for a given (s, a) pair should be designed so it does not increase with time T if (s, a) is not visited. In practice this means replacing the $\log(T)$ factor of the exploration bonuses with something like $\log(n)$ where n is the visit count to a specific state-action pair and adjusting the numerical constant to make sure the exploration bonuses / confidence intervals are still valid with high probability. Please see (Dann et al., 2017) for a detailed explanation of how to proceed with the algorithm design and analysis in this case[5].

C. Appendix Overview and Proof Preview

We start by giving an overview of the proof that leads to the main result. This setting is more general than the one presented in the main text. In particular we 1) define a class of concentration inequalities for the transition dynamics 2) show that EULER achieves strong problem dependent regret bounds with any concentration inequality satisfying these assumptions. In particular, the main result presented in the main text follows as a corollary of the potentially more general analysis presented. We now give a preview of the proof of the main result, assuming rewards are known, though we later relax this assumption. We start by recalling EULER with yet-to-specify confidence intervals on the transition dynamics.

C.1. Algorithm

The algorithm is presented in figure 2.

C.2. Optimism

The goal of this section is to show that EULER guarantees optimism with high probability. As is well known, optimism is the "driver" of exploration as it allows to overestimate the regret by a concentration term with high probability.

---

[5]Indeed, the algorithm described in (Dann et al., 2017) is structurally similar to ours

---

Algorithm 2 EULER for Stationary Episodic MDPs

1: **Input**: confidence interval $b_k^r(\cdot, \cdot)$ $\phi(\cdot, \cdot)$ with failure probability δ, constants B_p, B_v.
2: $n(s,a) = r_{sum}(s,a) = p_{sum}(s', s, a) = 0, \quad \forall s, a \in \mathcal{S} \times \mathcal{A}; \quad \overline{V}_{H+1}(s) = 0, \quad \forall s \in \mathcal{S}$
3: **for** $k = 1, 2, \ldots$ **do**
4:    **for** $t = H, H-1, \ldots, 1$ **do**
5:       **for** $s \in \mathcal{S}$ **do**
6:          **for** $a \in \mathcal{A}$ **do**
7:             $\hat{p} = \frac{p_{sum}(\cdot, s, a)}{n(s,a)}$
8:             $b_k^{pv} = \phi(\hat{p}(s,a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s,a)}} \left(\frac{4J + B_p}{\sqrt{n_k(s,a)}} + B_v \|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}} \right)$
9:             $Q(a) = \min\{H - t, \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b_k^r(s,a) + \hat{p}^\top \overline{V}_{t+1} + b_k^{pv}\}$
10:          **end for**
11:          $\tilde{\pi}_k(s,t) = \arg\max_a Q(a)$
12:          $\overline{V}_t(s) = Q(\tilde{\pi}_k(s,t))$
13:          $b_k^{pv} = \phi(\hat{p}(s, \tilde{\pi}_k(s,t)), \underline{V}_{t+1}) + \frac{1}{\sqrt{n(s, \tilde{\pi}_k(s,t))}} \left(\frac{4J + B_p}{\sqrt{n_k(s, \tilde{\pi}_k(s,t))}} + B_v \|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}} \right)$
14:          $\underline{V}_t(s) = \max\{0, \hat{r}_k(s, \tilde{\pi}_k(s,t)) - b_k^r(s, \tilde{\pi}_k(s,t)) + \hat{p}^\top \underline{V}_{t+1} - b_k^{pv}\}$
15:       **end for**
16:    **end for**
17:    $s_1 \sim p_0$
18:    **for** $t = 1, \ldots H$ **do**
19:       $a_t = \tilde{\pi}_k(s_t, t); \quad r_t \sim p_R(s_t, a_t); \quad s_{t+1} \sim p_P(s_t, a_t)$
20:       $n(s_t, a_t) + +; \quad p_{sum}(s_{t+1}, s_t, a_t) + +$
21:    **end for**
22: **end for**

Let's consider the planning process at the beginning of episode k. This is detailed in lines 4 to 16 of algorithm 1. In order to guarantee finding a pointwise optimistic value function $\overline{V}_{tk}^{\tilde{\pi}_k} \geq V_t^{\pi^*}$ a bonus is added to account for "bad luck" in the system dynamics experienced up to episode k. If the agent knew the value of the confidence interval for the system dynamics $\phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*})$ then optimism could be inductively guaranteed (i.e., assuming that, by induction, $\overline{V}_{t+1k}^{\tilde{\pi}_k} \geq V_{t+1}^{\pi^*}$ holds pointwise) if said confidence interval holds:

$$\overline{V}_{tk}^{\tilde{\pi}_k} = r(s,a) + \hat{p}_k(\cdot \mid s,a)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} + \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*}) \tag{23}$$

$$\geq r(s,a) + \hat{p}_k(\cdot \mid s,a)^\top V_{t+1}^{\pi^*} + \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*}) \geq r(s,a) + p(\cdot \mid s,a)^\top V_{t+1}^{\pi^*} \geq V_{t+1}^{\pi^*} \tag{24}$$

If the above conclusion is true for every action then it is true for the maximizer as well. Unfortunately the agent knows nor the real transition dynamics nor the optimal value function to evaluate ϕ. Instead, it only has access to the estimated transition dynamics $\hat{p}_k(\cdot \mid s,a)$ and to an overestimate of the value function $\overline{V}_{t+1k}^{\tilde{\pi}_k}$. Unfortunately the confidence interval ϕ evaluated with such quantities $\phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k})$ is not guaranteed to overestimate $\phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*})$ and optimism may not be guaranteed. To remedy this the agent can try to estimate the difference

$$|\phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) - \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*})| \tag{25}$$

and add a correction term to account for that difference. A similar problem is faced in (Azar et al., 2017) where the authors propose an optimistic bonus which guarantees optimism when using the empirical Bernstein Inequality. By distinction, our way of constructing the bonus works with *any* concentration inequality satisfying assumption 1 and 2, as described in the Appendix Section D.2. Precisely, optimism is dealt with in Appendix section E; in lemma 4 we show how to bound equation 25 obtaining the result below:

$$|\phi(\hat{p}_k(\cdot \mid s,a), V) - \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*})| \leq \frac{B_v \|V - V_{t+1}^{\pi^*}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} + \frac{B_p + 4J}{n_k(s,a)}. \tag{26}$$

This is essentially a consequence of the definition of admissible bonus, i.e., Definition 3 (Appendix Section D.2). Unfortunately the upper bound in equation 26 depends on $V_{t+1}^{\pi^*}$ which is not known, so the problem is still unsolved. However, as we show in lemma 5 in the appendix it suffices to (pointwise) overestimate $\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*}$. To this aim, the algorithm maintains an underestimate of $V_{t+1}^{\pi^*}$ which we call $\underline{V}_{t+1k}^{\tilde{\pi}_k}$. Equipped with this underestimate, we define the *Exploration Bonus* in definition 5 (Appendix Section E.2), which we report below:

$$b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \stackrel{def}{=} \phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{4J + B_p}{n_k(s,a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}}. \tag{27}$$

Importantly, Equation 27 only uses quantities that are known to the agent: the functional form of $\phi(\cdot,\cdot)$, the maximum likelihood estimate $\hat{p}_k(\cdot \mid s,a)$, the overestimate and underestimate, $\overline{V}_{t+1k}^{\tilde{\pi}_k}$ and $\underline{V}_{t+1k}^{\tilde{\pi}_k}$ respectively, of the optimal value function at the next timestep, the visit count $n_k(s,a)$ and the constants J, B_p, B_v. Notice that the norm $\|\cdot\|_{2,\hat{p}}$ is computed using $\hat{p}_k(\cdot \mid s,a)$ which is known to the agent as opposed to $p(\cdot \mid s,a)$. If $\overline{V}_{t+1k}^{\tilde{\pi}_k}$ and $\underline{V}_{t+1k}^{\tilde{\pi}_k}$ bracket $V_{t+1}^{\pi^*}$ then we have that the bonus of equation 27 overestimates the admissible confidence interval $\phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*})$ that we could construct if we knew $p(\cdot \mid s,a)$ and $V_{t+1}^{\pi^*}$, that is:

$$b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \geq \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*}) \tag{28}$$

This is proved in Proposition 3 in the appendix. At this point we have all the elements to show optimism. In fact, we need a little more effort than simply optimism because the algorithm has to maintain a valid bracket of the optimal value function:

$$\underline{V}_{tk}^{\tilde{\pi}_k} \leq V_t^{\pi^*} \leq \overline{V}_{tk}^{\tilde{\pi}_k} \quad \text{(pointwise)} \tag{29}$$

This is done in Proposition 4 in the appendix and it simply relies on an induction argument.

At this point we have guaranteed optimism but we relied on the construction of confidence intervals for the value function, to which we turn our attention next.

C.3. Confidence Interval for the Value Function

During its execution, EULER implicitly construct confidence interval for the value function with the property defined by equation 29. Precisely in Proposition 5 we relate the distance $\overline{V}_{tk}^{\tilde{\pi}_k}(s) - \underline{V}_{tk}^{\tilde{\pi}_k}(s)$ to the number of visits to the (s,a) pairs in the trajectories originated upon following policy $\tilde{\pi}_k$ *on the true MDP* with high probability. In other words, assuming that confidence intervals hold we provide a way to relate the accuracy of the agent's estimate of the value function to a concentration term that depends on the number of visits to the (s,a) pairs that the agent is expected to encounter by following that policy, obtaining up to a constant:

$$\overline{V}_{tk}^{\tilde{\pi}_k}(s) - \underline{V}_{tk}^{\tilde{\pi}_k}(s) \lesssim \sum_{\tau=t}^{H} \mathbb{E}\left(\min\left\{\frac{F+D}{\sqrt{n_k(s_\tau,a)}}, H\right\} \mid s, \tilde{\pi}_k\right) \tag{30}$$

for some F, D defined in Proposition 5.

This serves as an estimate of the confidence interval for the optimal value function itself. The importance of the lemma lies in connecting a property of the algorithm (the difference between the "optimistic" and the "pessimistic" value function) to the uncertainty in the various states encountered in the MDP (upon following $\tilde{\pi}_k$) weighted by the *true* visitation probability.

C.4. Regret Bound

We are finally ready to discuss the regret bounds that leads to the main result of Theorem 2 which is proved in Appendix Section H along with the related lemmata. We recall the following regret decomposition which is standard in recent analysis (Dann et al., 2017; Osband & Van Roy, 2016):

$$\text{Regret}(K) \stackrel{def}{=} \sum_{k}^{K} V_1^*(s_{1k}) - V_1^{\tilde{\pi}_k}(s_{1k}) \tag{31}$$

$$\leq \sum_{k=1}^{K} \sum_{t \in [H]} \sum_{(s,a)} w_{tk}(s,a) \left(\underbrace{(\tilde{r}_k(s,a) - r(s,a))}_{\text{Reward Estimation and Optimism}} + \underbrace{(\tilde{p}_k(s,a) - \hat{p}(s,a))^\top \overline{V}_{t+1}^{\tilde{\pi}_k}}_{\text{Transition Dynamics Optimism}} \right. \tag{32}$$

$$\left. + \underbrace{(\hat{p}_k(s,a) - p(s,a))^\top V_{t+1}^*}_{\text{Transition Dynamics Estimation}} + \underbrace{(\hat{p}_k(s,a) - p(s,a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*} \right)}_{\text{Lower Order Term}} \right) \tag{33}$$

$$\tag{34}$$

In later sections we bound each term individually; here we just touch on the order of magnitude of the leading order term which is the "Transition Dynamics Optimism." We begin by using the bonus added during the planning step (Definition 5):

$$\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a) \right)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \overset{def}{\leq} \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \tag{35}$$

where the inequality follows from the fact that the we "cap" the backup term $\tilde{p}_k(\cdot \mid s,a)\overline{V}_{t+1k}^{\tilde{\pi}_k} \leq H$ (see the $\min$ in the main algorithm). By definition of the bonus (definition 5) we get:

$$c_{200} \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} + \frac{B_p + J}{n_k(s,a)} \right), \tag{36}$$

for some constant c_{200}. Equation 26 ensures that $\phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k})$ and $\phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*})$ are close, leading to essentially the same upper bound up to a constant:

$$\lesssim \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*}) + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} + \frac{B_p + J}{n_k(s,a)} \right) \tag{37}$$

Using the functional form for ϕ we obtain the upper bound below (c):

$$\overset{c}{\lesssim} \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\underbrace{\frac{g(p, V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}}}_{\text{Leading Order Term}} + \underbrace{\frac{J + B_p}{n_k(s,a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}}}_{\text{Lower Order Term}} \right) \tag{38}$$

An induction argument coupled with equation 30 shows that $\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}$ shrinks at a rate $\frac{1}{\sqrt{n_k(s,a)}}$. Thus, the "Lower Order Term" shrinks at a rate $\frac{1}{n}$, and ultimately gives a regret contribution independent on T except for a log factor. The leading order term shrinks at a rate $\frac{1}{\sqrt{n}}$, giving the $\tilde{O}\left(\sqrt{\mathbb{Q}^* SAT}\right)$ contribution which is the leading order term. Further, since $g(\cdot, \cdot)$ here depends on $V_{t+1}^{\pi^*}$, for $t \in [H]$, this is a problem-dependent (and concentration-inequality-dependent) bound, as we wanted.

In the full proof as follows, we also include uncertainty over the reward function.

D. Failure Events and their Probabilities

We now discuss the failure events and the assumption for the concentration inequalities that lead to the definition of EULER. We then verify that Bernstein Inequality satisfies these assumptions, leading to a practical algorithm.

D.1. Empirical Bernstein Inequality for the Rewards

We recall the Empirical Bernstein Inequality[6] (Maurer & Pontil, 2009) for estimating the rewards:

---

[6]Note the change of $n_k(s,a) - 1$ to $n_k(s,a)$ compared to (Maurer & Pontil, 2009) in the lower order term, and the doubling of the constant for that term since $\frac{1}{n_k(s,a)-1} \leq \frac{2}{n_k(s,a)}$ for $n_k(s,a) \geq 2$.

Definition 2 (Reward Empirical Bernstein). *Let $R(s,a) \in [0,1]$ be the reward random variable in state s upon taking action a and let $\widehat{\mathrm{Var}} R(s,a)$ be its sample variance. The following holds true with probability at least $1 - \delta'$:*

$$\left| \hat{r}_k(s,a) - r(s,a) \right| \leq \sqrt{\frac{2\widehat{\mathrm{Var}} R(s,a) \ln \frac{4SAT}{\delta'}}{n_k(s,a)}} + \frac{14 \ln \frac{4SAT}{\delta'}}{3n_k(s,a)}. \tag{39}$$

This concentrates fast to the actual reward variance :

Lemma 1 (Delta ϕ_r). *With probability at least $1 - \delta'$ it holds that:*

$$|\sqrt{\widehat{\mathrm{Var}} R(s,a)} - \sqrt{\mathrm{Var}\, R(s,a)}| \leq \sqrt{\frac{4 \ln(2SAT/\delta')}{n_k(s,a)}}. \tag{40}$$

jointly for all states, actions, and timesteps.

Proof. Analogous to Theorem 10 in (Maurer & Pontil, 2009) with a union bound argument over the states, the actions and the the timesteps. Note the change of $n_k(s,a) - 1$ to $n_k(s,a)$ compared to (Maurer & Pontil, 2009) and the doubling of the constant since $\frac{1}{n_k(s,a)-1} \leq \frac{2}{n_k(s,a)}$ for $n_k(s,a) \geq 2$. □

D.2. Admissible Confidence Intervals on the Transition Dynamics

In this section we define a class of confidence intervals that are admissible for EULER for which our analysis holds. The aim is to ensure that $| (\hat{p}_k(s,a) - p(s,a))^\top V_{t+1}^{\pi^*} |$ is bounded with high probability throughout the execution of the algorithm. Said concentration inequality should be tight so that successor states with low visitation probability have low impact. The former requirement is formalized in equation 41 and the latter in equation 43, which we report below.

Assumption 1 (Confidence Intervals). *With probability at least $1 - \delta'$ it holds that:*

$$| (\hat{p}_k(s,a) - p(s,a))^\top V_{t+1}^{\pi^*} | \leq \phi(p(s,a), V_{t+1}^{\pi^*}) \tag{41}$$

jointly for all timesteps t, episodes k, states s and actions a. We assume that $\phi(p,V)$ takes the following functional form:

$$\phi(p,V) = \frac{g(p,V)}{\sqrt{n_k(s,a)}} + \frac{j(p,V)}{n_k(s,a)} \tag{42}$$

where $j(p,v) \leq J \in \mathbb{R}$. In particular we assume the following constraint on the functional form of $g(\cdot,\cdot)$:

$$|g(p,V_1) - g(p,V_2)| \leq B_v \|V_1 - V_2\|_{2,p} \tag{43}$$

and if the value function is uniform then:

$$g(p, \alpha \mathbb{1}) = 0, \quad \forall \alpha \in \mathbb{R}. \tag{44}$$

Equation 42 refers to the functional form of $\phi(\cdot,\cdot)$ which is the concentration inequality on the transition dynamics. Equation 42 identifies two contributions: a leading order term which scales with $1/\sqrt{n}$ and a lower order term that scales with $1/n$. Equation 43 plays a crucial role. It posits a requirement on the functional form of the leading order term of the concentration inequality when the coefficients $V_{t+1}^{\pi^*}(s')$ are changed. Precisely, it quantifies how the concentration inequality changes if we change $V_{t+1}^{\pi^*}$, formalizing the intuition that if $p(s' \mid s,a)$ is small then changing $V_{t+1}^{\pi^*}(s')$ should have little impact on the bound given by the concentration inequality. It also implies that $g(p,V)$ depends on V only through the entries that correspond to the support of p, that is, it depends on $V(s)$ if $p(s) \neq 0$. Practically speaking, if a successor s' cannot be visited from s then the value function at s' does not directly impact s, as one would hope.

The next assumption deals with the rate of convergence of the leading order term seen as a function of $\hat{p}$. Under mild assumptions, as $\hat{p}$ converges to p, a function of $\hat{p}$ converges as well. The assumption below is the corresponding non-asymptotic requirement:

Assumption 2 (Finite Time Bonus Bound). *With probability at least $1 - \delta'$ it holds that:*

$$|g(\hat{p}_k(\cdot \mid s,a), V_{t+1}^{\pi^*}) - g(p(\cdot \mid s,a), V_{t+1}^{\pi^*})| \leq \frac{B_p}{\sqrt{n_k(s,a)}} \tag{45}$$

jointly for all episodes k, timesteps t, states s, actions a and some constant B_p that does not depend on $n_k(s,a)$.

In both assumption 1 and 2 the constants B_v and B_p can depend on the input parameters (e.g., $S, A, H, T, \frac{1}{\delta}$ etc...).

We pose the following definition:

Definition 3 (Admissible ϕ). *If ϕ satisfies assumption 1 and 2 then we say that ϕ is admissible for EULER.*

Corollary 1.3 (Max ϕ). *Let the function $g(\cdot, \cdot)$ be defined as in equation 42. Combining equations 43 and 44 (where $V_2 = \vec{0}$) and recalling monotonicity of norms of random variables one immediately obtains:*

$$|g(p, V)| \leq B_v \|V_1\|_{2,p} \leq B_v \|V_1\|_\infty \leq B_v H. \tag{46}$$

D.3. Bernstein's Inequality

We now show that Bernstein Inequality is admissible for EULER.

Proposition 2 (Bernstein Is Admissible). *Bernstein Inequality as presented in equation 47 satisfies assumption 1 and 2 and is therefore admissible for EULER with coefficients $J = \frac{H \ln \frac{2SAT}{\delta'}}{3} = \tilde{O}(H)$, $B_v = \sqrt{2 \ln \frac{2SAT}{\delta'}} = \tilde{O}(1)$ and $B_p = \sqrt{2} H \sqrt{\ln \frac{2SAT}{\delta'}} = \tilde{O}(H)$.*

Proof. Bernstein's inequality guarantees that with probability at least $1 - \delta'$ we have that:

$$\left| (\hat{p}_k(s,a) - p(s,a))^\top V_{t+1}^{\pi^*} \right| \leq \sqrt{\frac{2 \operatorname{Var}_p V_{t+1}^{\pi^*} \ln \frac{2SAT}{\delta'}}{n_k(s,a)}} + \frac{H \ln \frac{2SAT}{\delta'}}{3n_k(s,a)} \stackrel{def}{=} \phi(p(s,a), V_{t+1}^{\pi^*}) \tag{47}$$

jointly for all timesteps, states s and actions a after a union bound argument over the states s, actions a and timesteps. Thus, equation 41 holds. Here $J = \frac{H \ln \frac{2SAT}{\delta'}}{3}$ and $\sqrt{2 \ln \frac{2SAT}{\delta'}} \stackrel{def}{\leq} L$ so that

$$g(p(s,a), V_{t+1}^{\pi^*}) \stackrel{def}{=} \sqrt{\operatorname{Var}_p V_{t+1}^{\pi^*}} \times \sqrt{2 \ln \frac{2SAT}{\delta'}} \leq L \sqrt{\operatorname{Var}_p V_{t+1}^{\pi^*}}.$$

Thus equation 42 holds. Consider the mean-centered random variables $\overline{V}_1 = V_1 - \mathbb{E} V_1$ and $\overline{V}_2 = V_2 - \mathbb{E} V_2$. Then:

$$\sqrt{\operatorname{Var}(V_1)} = \sqrt{\operatorname{Var}(\overline{V}_1)} = \sqrt{\mathbb{E}(\overline{V}_1)^2} = \|\overline{V}_1\|_{2,p} = \|\overline{V}_2 + \overline{V}_1 - \overline{V}_2\|_{2,p} \tag{48}$$

$$\leq \|\overline{V}_2\|_{2,p} + \|\overline{V}_1 - \overline{V}_2\|_{2,p} = \sqrt{\mathbb{E}(\overline{V}_2)} + \sqrt{\mathbb{E}(\overline{V}_1 - \overline{V}_2)^2} \tag{49}$$

$$= \sqrt{\operatorname{Var}(\overline{V}_2)} + \sqrt{\mathbb{E}(V_1 - V_2)^2 - (\mathbb{E}(V_1 - V_2))^2} \tag{50}$$

$$= \sqrt{\operatorname{Var}(V_2)} + \sqrt{\operatorname{Var}(V_1 - V_2)}. \tag{51}$$

where the inequality is Minkowski's inequality (i.e., the triangle inequality for norm of random variables). Rearranging we get:

$$|g(p, V_1) - g(p, V_2)| \leq L |\sqrt{\operatorname{Var}(V_1)} - \sqrt{\operatorname{Var}(V_2)}| \leq L \sqrt{\operatorname{Var}(V_2 - V_1)} \leq L \|V_2 - V_1\|_{2,p} \tag{52}$$

and so $B_v = L$ in equation 43.

Finally, a variation[7] of theorem 10 from (Maurer & Pontil, 2009) ensures that:

$$\left| \|V_t^{\pi^*}\|_{2,\hat{p}} - \|V_t^{\pi^*}\|_{2,p} \right| \leq H \sqrt{\frac{4 \ln \frac{2SAT}{\delta'}}{n_k(s,a)}} = \frac{B_p}{\sqrt{n_k(s,a)}} \tag{53}$$

with probability at least $1 - \delta'$ jointly for all states s, actions a and possible values for n after a union bound on these quantities. Hence $B_p = \sqrt{2} HL$ and assumption 2 is satisfied as well. This concludes the verification that Bernstein's inequality satisfies both 1 and 2 and is thus admissible for the algorithm.

$\square$

---

[7] Note the change of $n_k(s,a) - 1$ to $n_k(s,a)$ compared to (Maurer & Pontil, 2009) and the doubling of the constant for that term since $\frac{1}{n_k(s,a)-1} \leq \frac{2}{n_k(s,a)}$ for $n_k(s,a) \geq 2$.

D.4. Other Failure Events and Their Probabilities

An independent use of Bernstein inequality also gives with probability at least $1 - \delta'$ jointly for all states s, successors s', actions s and values for $n_k(s, a)$ the following component-wise bound on the failure event (see (Azar et al., 2017) for a derivation):

$$|\hat{p}_k(s' \mid s, a) - p(s' \mid s, a)| \leq \sqrt{\frac{p(s' \mid s, a)(1 - p(s' \mid s, a)) \ln \frac{2TS^2A}{\delta'}}{n_k(s, a)}} + \frac{\ln \frac{2TS^2A}{\delta'}}{3n_k(s, a)}. \tag{54}$$

Moreover, (Weissman et al., 2003) gives the following high probability bound on the one norm of the Maximum Likelihood Estimate ; in particular, with probability at least $1 - \delta'$ it holds that:

$$\|\hat{p}_k(\cdot \mid s, a) - p(\cdot \mid s, a)\|_1 \leq \sqrt{\frac{2S \ln \frac{2SAT}{\delta'}}{n_k(s, a)}} \tag{55}$$

jointly for all states s, actions a and possible values for $n_k(s, a)$ after a union bound argument on these quantities. Finally, with probability at least $1 - \delta'$ the following holds for every state-action pair, timestep and episode (see for example (Dann et al., 2019), failure event F^N in section B.1, for the proof):

$$n_k(s, a) \geq \frac{1}{2} \sum_{j<k} w_j(s, a) - H \ln \frac{SAH}{\delta'} \tag{56}$$

where $w_{tj}(s, a)$ is the probability of visiting the (s, a) pair in timestep t of episode j under the chosen policy and $\sum_{\tau \in H} w_{\tau j}(s, a)$ is the sum of the probabilities of visiting the (s, a) pair in episode j.

Lemma 2 (Failure Probability). *If $\delta' = \frac{1}{7}\delta$ then equation 39, 40,47,53, 54,55, 56 hold jointly with probability at least $1 - \delta$. When this happens we say that we EULER is outside of the failure event.*

Proof. By union bound. □

E. Optimism

In this section we show that EULER computes optimistic bounds on Q.

E.1. Rewards

In view of the empirical Bernstein Inequality for the rewards in lemma 2 we define as reward bonus:

Definition 4 (Reward Bonus).

$$b_k^r(s, a) \overset{def}{=} \sqrt{\frac{2\widehat{\mathrm{Var}}R(s, a) \ln \frac{4SAT}{\delta'}}{n_k(s, a)}} + \frac{14 \ln \frac{4SAT}{\delta'}}{3n_k(s, a)}. \tag{57}$$

Lemma 3 (Reward Bonus is Optimistic). *Outside of the failure event it holds that:*

$$\hat{r}_k(s, a) + b_k^r(s, a) \geq r(s, a) \tag{58}$$
$$\hat{r}_k(s, a) - b_k^r(s, a) \leq r(s, a) \tag{59}$$

Proof. By Definition 4 and lemma 2. □

E.2. Transition Dynamics

Lemma 4 (Delta ϕ). *If ϕ is admissible for EULER then for all $V \in \mathbb{R}^S$:*

$$|\phi(\hat{p}_k(\cdot \mid s, a), V) - \phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*})| \leq \frac{B_v \|V - V_{t+1}^{\pi^*}\|_{2,\hat{p}}}{\sqrt{n_k(s, a)}} + \frac{B_p + 4J}{n_k(s, a)}. \tag{60}$$

Outside of the failure event the above lemma deals with the functional form of ϕ; there are no "failure events" or probabilities to be considered here.

Proof. From the LHS of Equation 60 by adding and subtracting $\phi(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*})$ we get to an expression equivalent to the LHS of Equation 60:

$$(60) = |\phi(\hat{p}_k(\cdot \mid s, a), V) - \phi(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*}) + \phi(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*}) - \phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*})| \tag{61}$$

The triangle inequality allows to split the above equation into the upper bound below:

$$\leq |\phi(\hat{p}_k(\cdot \mid s, a), V) - \phi(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*})| + |\phi(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*}) - \phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*})| \tag{62}$$

Next we can use the constraint on ϕ. In particular, condition 42 implies that the above equation can be upper bounded as below:

$$\leq \left| \frac{g(\hat{p}_k(\cdot \mid s, a), V) - g(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s, a)}} \right| \tag{63}$$

$$+ \left| \frac{g(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*}) - g(p(\cdot \mid s, a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s, a)}} \right| + \frac{4J}{n_k(s, a)} \tag{64}$$

Finally, the functional constraints on g of Equation 43 together with Assumption 2, respectively, bound each of the above terms:

$$\left| \frac{g(\hat{p}_k(\cdot \mid s, a), V) - g(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s, a)}} \right| \leq \frac{B_v \|V - V_{t+1}^{\pi^*}\|_{2,\hat{p}}}{\sqrt{n_k(s, a)}} \tag{65}$$

$$\left| \frac{g(\hat{p}_k(\cdot \mid s, a), V_{t+1}^{\pi^*}) - g(p(\cdot \mid s, a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s, a)}} \right| \leq \frac{B_p}{n_k(s, a)} \tag{66}$$

completing the proof. $\qquad\square$

Lemma 4 is crucial in that it allows to relate how far off is the concentration inequality ϕ (computed using the empirical estimates for p and V) from the one computed using the "real" values, which would guarantee optimism. In other words, if one can compute $\|\overline{V}_{tk}^{\tilde{\pi}_k} - V_t^{\pi^*}\|_{2,\hat{p}}$ then this estimate can be used with lemma 4 to derive a bonus b_k^{pv}, function of the empirical quantities $\hat{p}_k(\cdot \mid s, a)$ and $\overline{V}_{t+1k}^{\tilde{\pi}_k}$, which is guaranteed to overestimate $\phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*})$. Ultimately, the purpose is to construct a "bonus" that overestimates ϕ without being "much larger" than ϕ. This is the motivation behind the following definition, which will eventually lead to optimism of EULER while ensuring a regret that is problem-dependent.

Definition 5 (Transition Bonus). *Define the bonus* $b_k^{pv}(\cdot, \cdot, \cdot)$:

$$b_k^{pv}(\hat{p}_k(\cdot \mid s, a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \overset{def}{=} \phi(\hat{p}_k(\cdot \mid s, a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{4J + B_p}{n_k(s, a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s, a)}}. \tag{67}$$

En-route to showing optimism we first show the value of having an overestimate *and* and underestimate of $V_{t+1}^{\pi^*}$ in lemma 5. This allows to relate the bonus of definition 5 to the concentration inequality identified by $\phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*})$. The result of the first lemma is summarized below:

Lemma 5 (Optimism Overestimate). *For any transition probability vector p, (i.e., such that $\|p\|_1 = 1$) and any $V \in \mathbb{R}^S$ if:*

$$\underline{V}_{t+1k}^{\tilde{\pi}_k} \leq V \leq \overline{V}_{t+1k}^{\tilde{\pi}_k} \tag{68}$$

holds pointwise then

$$\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V\|_{2,p} \leq \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p} \tag{69}$$

$$\|V - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p} \leq \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p} \tag{70}$$

holds.

Proof. The hypothesis ensures:

$$0 \leq \overline{V}_{t+1k}^{\tilde{\pi}_k}(s') - V(s') \leq \overline{V}_{t+1k}^{\tilde{\pi}_k}(s') - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s'). \tag{71}$$

Since these are positive quantities we can square them and preserve the order of the inequality:

$$0 \leq \left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s') - V(s')\right)^2 \leq \left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s') - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s')\right)^2. \tag{72}$$

A linear combination of the above terms, weighted by the component of p, i.e., $p(s')$ gives the second moment squared:

$$0 \leq \sum_{s'} p(s') \left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s') - V(s')\right)^2 \leq \sum_{s'} p(s') \left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s') - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s')\right)^2. \tag{73}$$

Taking square-root yields:

$$\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V\|_{2,p} \leq \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p}. \tag{74}$$

Equation 70 is proved analogously. □

The above lemma ensures the result below:

Proposition 3 (Transition Bonus is Optimistic). *If the following condition hold:*

1. *ϕ is admissible*

2. *$\underline{V}_{t+1k}^{\tilde{\pi}_k} \leq V_{t+1}^{\pi^*} \leq \overline{V}_{t+1k}^{\tilde{\pi}_k}$ pointwise*

then the following holds true:

$$b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \geq \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*}) \tag{75}$$

$$b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \underline{V}_{t+1k}^{\tilde{\pi}_k}, \overline{V}_{t+1k}^{\tilde{\pi}_k}) \geq \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*}). \tag{76}$$

If condition in equation 75 is satisfied then we say that the bonus $b_k^{pv}(\cdot, \cdot)$ is optimistic *for* EULER.

This is the key result that will ensure optimism of the algorithm later.

Proof. Condition 2) coupled with lemma 5 ensures:

$$\frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} \geq \frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}}. \tag{77}$$

Together, equation 77 and assumption 2 imply the first of the following inequalities on the bonus:

$$b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \overset{def}{=} \phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{4J + B_p}{n_k(s,a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} \tag{78}$$

$$\geq \phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{4J + B_p}{\sqrt{n_k(s,a)}} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} \tag{79}$$

$$\geq \phi(p(\cdot \mid s,a), V_{t+1}^{\pi^*}). \tag{80}$$

while the second inequality is ensured by lemma 4, completing the proof of equation 75. Equation 76 is proved analogously. □

Proposition 3 states that the bonus defined in Definition 5, which was constructed out of the admissible confidence interval for ϕ, overestimates ϕ. This is enough to guarantee optimism of the algorithm:

E.3. Algorithm is Optimistic

Proposition 4 (Algorithm Brackets $V_t^{\pi^*}$). *Outside of the failure event if* EULER *is run with an admissible ϕ then:*

$$\underline{V}_{tk}^{\tilde{\pi}_k} \leq V_t^{\pi^*} \leq \overline{V}_{tk}^{\tilde{\pi}_k} \quad \text{(pointwise)} \tag{81}$$

holds for all timesteps t and episodes k.

Proof. We proceed by induction. Suppose equation 81 holds for all states s in timestep $t+1$. If

$$\hat{r}_k(s,t) + b_k^r(s,a) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} + b_k^{pv}(\hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t)), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \geq H - t \tag{82}$$

holds then we are done. If the above does not hold then maximization over the actions in the optimistic MDP justifies the last inequality below, while the first inequality is justified by lemma 3:

$$\overline{V}_{tk}^{\tilde{\pi}_k} = \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b_k^r(s, \tilde{\pi}_k(s,t)) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} + b_k^{pv}(\hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t)), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \tag{83}$$

$$\geq r(s, \tilde{\pi}_k(s,t)) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} + b_k^{pv}(\hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t)), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \tag{84}$$

$$\geq r(s, \pi^*(s,t)) + \hat{p}_k(\cdot \mid s, \pi^*(s,t))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} + b_k^{pv}(\hat{p}_k(\cdot \mid s, \pi^*(s,t)), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}). \tag{85}$$

Next, the inductive hypothesis $V_{t+1}^{\pi^*} \leq \overline{V}_{t+1k}^{\tilde{\pi}_k}$ yields the following lower bound:

$$\geq r(s, \pi^*(s,t)) + \hat{p}_k(\cdot \mid s, \pi^*(s,t))^\top V_{t+1}^{\pi^*} + b_k^{pv}(\hat{p}_k(\cdot \mid s, \pi^*(s,t)), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}). \tag{86}$$

Proposition 3 finally gives:

$$\geq r(s, \pi^*(s,t)) + \hat{p}_k(\cdot \mid s, \pi^*(s,t))^\top V_{t+1}^{\pi^*} + \phi(p(\cdot \mid s, \pi^*(s,t)), V_{t+1}^{\pi^*}) \tag{87}$$

Since ϕ is admissible we get:

$$\geq r(s, \pi^*(s,t)) + p(\cdot \mid s, \pi^*(s,t))^\top V_{t+1}^{\pi^*} = V_t^{\pi^*}(s) \tag{88}$$

This holds for every state s, completing the proof that EULER is "optimistic". It remains to show "pessimism", again by induction. If

$$\hat{r}_k(s, \tilde{\pi}_k(s,t)) - b_k^r(s, \tilde{\pi}_k(s,t)) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top \underline{V}_{t+1k}^{\tilde{\pi}_k} - b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \underline{V}_{t+1k}^{\tilde{\pi}_k}, \overline{V}_{t+1k}^{\tilde{\pi}_k}) \leq 0 \tag{89}$$

we are done. If this is not the case then an upper bound is given by proposition 3 and lemma 3:

$$\underline{V}_{tk}^{\tilde{\pi}_k}(s) = \hat{r}_k(s, \tilde{\pi}_k(s,t)) - b_k^r(s, \tilde{\pi}_k(s,t)) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top \underline{V}_{t+1k}^{\tilde{\pi}_k} - b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \underline{V}_{t+1k}^{\tilde{\pi}_k}, \overline{V}_{t+1k}^{\tilde{\pi}_k}) \tag{90}$$

$$\leq r(s, \tilde{\pi}_k(s,t)) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top \underline{V}_{t+1k}^{\tilde{\pi}_k} - b_k^{pv}(\hat{p}_k(\cdot \mid s,a), \underline{V}_{t+1k}^{\tilde{\pi}_k}, \overline{V}_{t+1k}^{\tilde{\pi}_k}) \tag{91}$$

$$\leq r(s, \tilde{\pi}_k(s,t)) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top \underline{V}_{t+1k}^{\tilde{\pi}_k} - \phi(p(\cdot \mid s, \tilde{\pi}_k(s,t)), V_{t+1}^{\pi^*}) \tag{92}$$

$$\tag{93}$$

The inductive hypothesis $\underline{V}_{t+1k}^{\tilde{\pi}_k} \leq V_{t+1}^{\pi^*}$ justifies the following upper bound:

$$\leq r(s, \tilde{\pi}_k(s,t)) + \hat{p}_k(\cdot \mid s, \tilde{\pi}_k(s,t))^\top V_{t+1}^{\pi^*} - \phi(p(\cdot \mid s, \tilde{\pi}_k(s,t)), V_{t+1}^{\pi^*}) \tag{94}$$

$$\tag{95}$$

Finally, since ϕ is admissible we get:

$$\leq r(s, \tilde{\pi}_k(s,t)) + p(\cdot \mid s, \tilde{\pi}_k(s,t))^\top V_{t+1}^{\pi^*} \tag{96}$$

By definition the optimal policy π^* must achieve a higher value:

$$\leq r(s, \pi^*(s,t)) + p(\cdot \mid s, \pi^*(s,t))^\top V_{t+1}^{\pi^*} = V_t^{\pi^*}(s) \tag{97}$$

completing the proof. $\square$

F. Delta Optimism

Proposition 5 (Delta Optimism). *Outside of the failure event for* EULER *it holds that:*

$$\overline{V}_{tk}^{\tilde{\pi}_k}(s) - \underline{V}_{tk}^{\tilde{\pi}_k}(s) \le c_{5,3,2} \sum_{\tau=t}^{H} \mathbb{E}\left(\min\left\{\frac{g(p(\cdot \mid s_\tau, a), V_{\tau+1}^{\pi^*})}{\sqrt{n_k(s_\tau, a)}} + \frac{F}{\sqrt{n_k(s_\tau, a)}} + \frac{D}{n_k(s_\tau, a)}, H\right\} \Big| s, \tilde{\pi}_k\right) \tag{98}$$

$$\le c_{5,3,3} \sum_{\tau=t}^{H} \mathbb{E}\left(\min\left\{\frac{F+D}{\sqrt{n_k(s_\tau, a)}}, H\right\} \Big| s, \tilde{\pi}_k\right) \tag{99}$$

where

$$F \stackrel{def}{=} (H\sqrt{S} + B_v H)L \tag{100}$$

$$D \stackrel{def}{=} (J + B_p)L \tag{101}$$

and the conditional expectation $\mathbb{E}(\cdot \mid s, \tilde{\pi}_k)$ is with respect to the states s_τ encountered during the k-th episode upon following policy $\tilde{\pi}_k$ after visiting state s in timestep t. We use the convention that the terms in RHS corresponding to $n_k(s_\tau, a) = 0$ are bounded by H, which is the maximum difference in $\overline{V}_{tk}^{\tilde{\pi}_k}(s') - \underline{V}_{tk}^{\tilde{\pi}_k}(s')$.

Proof. By the planning step for the action a chosen by EULER it holds that:

$$\begin{cases} \overline{V}_{tk}^{\tilde{\pi}_k}(s) \le \hat{r}_k(s, a) + b_k^r(s, a) + \hat{p}_k(\cdot \mid s, a)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} + b_k^{pv}(\hat{p}_k(\cdot \mid s, a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \\ \underline{V}_{tk}^{\tilde{\pi}_k}(s) \ge \hat{r}_k(s, a) - b_k^r(s, a) + \hat{p}_k(\cdot \mid s, a)^\top \underline{V}_{t+1k}^{\tilde{\pi}_k} - b_k^{pv}(\hat{p}_k(\cdot \mid s, a), \underline{V}_{t+1k}^{\tilde{\pi}_k}, \overline{V}_{t+1k}^{\tilde{\pi}_k}) \end{cases} \tag{102}$$

Notice that the exploration bonuses b_k^r, b_k^{pv} are bounded by $\tilde{O}(H)$ by construction in the algorithm, as well as the value function. This justifies the "hard bound" of H that appears in equation 98, which we drop for the rest of the proof to simplify the notation. Subtraction yields:

$$\overline{V}_{tk}^{\tilde{\pi}_k}(s) - \underline{V}_{tk}^{\tilde{\pi}_k}(s) \le \hat{p}_k(\cdot \mid s, a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) + b_k^{pv}(\hat{p}_k(\cdot \mid s, a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) + b_k^{pv}(\hat{p}_k(\cdot \mid s, a), \underline{V}_{t+1k}^{\tilde{\pi}_k}, \overline{V}_{t+1k}^{\tilde{\pi}_k}) + 2b_k^r(s, a). \tag{103}$$

Now we substitute the definition of bonus (Definition 5) to obtain:

$$\overline{V}_{tk}^{\tilde{\pi}_k}(s) - \underline{V}_{tk}^{\tilde{\pi}_k}(s) \le 2b_k^r(s, a) + \hat{p}_k(\cdot \mid s, a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) \tag{104}$$

$$+ \phi(\hat{p}_k(\cdot \mid s, a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{4J + B_p}{n_k(s, a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s, a)}} \tag{105}$$

$$+ \phi(\hat{p}_k(\cdot \mid s, a), \underline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{4J + B_p}{n_k(s, a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s, a)}}. \tag{106}$$

With the help of lemma 4 we relate ϕ evaluated at the empirical quantities to the "real" $\phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*})$, leading to the following upper bound of the above equation:

$$\le 2b_k^r(s, a) + \hat{p}_k(\cdot \mid s, a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) + 2\phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*}) + 4\left(\frac{4J + B_p}{n_k(s, a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s, a)}}\right). \tag{107}$$

Now, we want $p(\cdot \mid s, a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)$ to appear instead of $\hat{p}_k(\cdot \mid s, a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)$ to do induction on the "true" MDP and so we add and subtract the former to obtain:

$$= p(\cdot \mid s, a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) + 2b_k^r(s, a) + (\hat{p}_k(\cdot \mid s, a) - p(\cdot \mid s, a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) \tag{108}$$

$$+ 2\phi(p(\cdot \mid s, a), V_{t+1}^{\pi^*}) + 4\left(\frac{4J + B_p}{n_k(s, a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s, a)}}\right). \tag{109}$$

and using the definition of ϕ we finally have:

$$\leq 2b_k^r(s,a) + p(\cdot \mid s,a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) + (\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)$$

$$+ 2\frac{g(p(\cdot \mid s,a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}} + 2\frac{J}{n_k(s,a)} + 4\left(\frac{4J + B_p}{n_k(s,a)} + \frac{B_v\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}}\right). \tag{110}$$

. To deal with term $(\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)$ we use Holder's inequality and the fact that we are outside of the failure event so that equation 55 holds:

$$\leq 2b_k^r(s,a) + p(\cdot \mid s,a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) + \|\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s, \tilde{\pi}_k(s,t))\|_1 \|V_{t+1}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_\infty \tag{111}$$

$$+ 2\frac{g(p(\cdot \mid s,a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}} + 2\frac{J}{n_k(s,a)} + 4\left(\frac{4J + B_p}{n_k(s,a)} + \frac{B_v\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}}\right) \tag{112}$$

$$\leq 2b_k^r(s,a) + p(\cdot \mid s,a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right) + H\sqrt{\frac{S}{n_k(s,a)}} \times L \tag{113}$$

$$+ 2\frac{g(p(\cdot \mid s,a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}} + 2\frac{J}{n_k(s,a)} + 4\left(\frac{4J + B_p}{n_k(s,a)} + \frac{B_v H}{\sqrt{n_k(s,a)}}\right). \tag{114}$$

$$\tag{115}$$

Induction gives the statement when coupled with the fact that $\overline{V}_{t+1k}^{\tilde{\pi}_k}(s) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s) \leq H$ and that $b_k^r(s,a) \leq c_{5,3,1}\frac{L}{\sqrt{n_k(s,a)}}$ from the definition of Bernstein inequality. The second inequality in the theorem statement is given by $\sqrt{n} \leq n$ for $n \geq 1$ coupled with corollary 1.3. Log factors are incorporated into L. $\qquad\square$

G. The "Good" Set L_k

We now introduce the set L_k. The construction is similar to (Dann et al., 2017) although we modify it here for our to handle the regret framework (as opposed to PAC) under stationary dynamics. The idea is to partition the state-action space at each episode into two sets, the set of episodes that have been visited sufficiently often (so that we can lower bound these visits by their expectations using standard concentration inequalities) and the set of (s,a) that were not visited often enough to cause high regret. In particular:

Definition 6 (The Good Set). *The set L_k is defined as:*

$$L_k \stackrel{def}{=} \left\{(s,a) \in \mathcal{S} \times \mathcal{A} : \frac{1}{4}\sum_{j<k} w_j(s,a) \geq H\ln\frac{SAH}{\delta'} + H\right\}. \tag{116}$$

The above definition enables the following lemma that relates the realized number of visits to a state to their visit probabilities:

Lemma 6 (Visitation Ratio). *Outside the failure event if $(s,a) \in L_k$ then*

$$n_k(s,a) \geq \frac{1}{4}\sum_{j\leq k} w_j(s,a) \tag{117}$$

holds.

Proof. Outside the failure event equation 56 justifies the first passage below:

$$n_k(s,a) \geq \frac{1}{2}\sum_{j<k} w_j(s,a) - H\ln\frac{SAH}{\delta'} \tag{118}$$

$$= \frac{1}{4}\sum_{j<k} w_j(s,a) + \frac{1}{4}\sum_{j<k} w_j(s,a) - H\ln\frac{SAH}{\delta'} \geq \frac{1}{4}\sum_{j<k} w_j(s,a) + H \geq \frac{1}{4}\sum_{j<k} w_j(s,a) + w_k(s,a) \geq \frac{1}{4}\sum_{j\leq k} w_j(s,a)$$

(119)

while the second inequality holds because $(s,a) \in L_k$ by assumption and the third because $w_k(s,a) \leq H$. $\qquad\square$

Finally, the following corollary ensures that if $(s,a) \notin L_k$ then it will contribute very little to the regret:

Lemma 7 (Minimal Contribution). *It holds that:*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\notin L_k} w_{tk}(s,a) \leq c_{7,3,1} SAHL.$$

(120)

Proof. By definition 6, if $(s,a) \notin L_k$ then

$$\frac{1}{4}\sum_{j\leq k} w_j(s,a) < H\ln\frac{SAH}{\delta'} + H$$

(121)

holds. Now sum over the (s,a) pairs not in L_k, the timesteps t and episodes k to obtain:

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\notin L_k} w_{tk}(s,a) = \sum_{s,a}\sum_{k=1}^{K} w_k(s,a)\mathbb{1}\{(s,a)\notin L_k\} \leq \sum_{s,a}\left(4H\ln\frac{SAH}{\delta'} + 4H\right) \leq c_{7,3,1} SAH\tilde{L}$$

(122)

$\qquad\square$

H. Regret Analysis

We begin our regret analysis of EULER. We will carry out the analysis outside of the failure event to derive a high probability regret bound.

H.1. Main Result

Theorem 2 (Main Result). *If ϕ is admissible then with probability at least $1 - \delta$ the cumulated regret of EULER up to timestep T is upper bounded by the minimum between:*

$$\tilde{O}(\sqrt{(\mathbb{C}_r^* + \mathbb{C}^*)SAT} + \sqrt{S}SAH(F + D + H^{\frac{3}{2}}))$$

(123)

and

$$\tilde{O}(\sqrt{(\mathbb{C}_r^* + \mathbb{C}^\pi)SAT} + \sqrt{S}SAH(F + D + H^{\frac{3}{2}}) + B_v^2 SAH^2)$$

(124)

where

$$F \stackrel{def}{=} \tilde{O}(H\sqrt{S} + B_v H)$$

(125)

$$D \stackrel{def}{=} \tilde{O}(J + B_p)$$

(126)

and $\mathbb{C}^*$ and $\mathbb{C}^\pi$ are problem dependent upper bounds on the following quantities:

$$\mathbb{C}^* \geq \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s,a} w_{tk}(s,a)g(p, V_{t+1}^{\pi^*})^2 \stackrel{def}{=} \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{\tilde{\pi}_k} g(p, V_{t+1}^{\pi^*})^2$$

(127)

and

$$\mathbb{C}^\pi \geq \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s,a} w_{tk}(s,a)g(p, V_{t+1}^{\tilde{\pi}_k})^2 \stackrel{def}{=} \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{\tilde{\pi}_k} g(p, V_{t+1}^{\tilde{\pi}_k})^2.$$

(128)

while $\mathbb{C}_r^*$ is defined in lemma 8.

Note that F and D are identical to the definitions given in Equations 100 and 101 respectively.

Proof. Outside of the failure event proposition 4 guarantees optimism and thus:

$$V_t^{\pi^*}(s) - V_t^{\tilde{\pi}_k}(s) \leq \overline{V}_{tk}^{\tilde{\pi}_k}(s) - V_t^{\tilde{\pi}_k}(s) \tag{129}$$

holds for any state and time, and in particular in particular for $t = 1$. Lemma E.15 in (Dann et al., 2017) is a standard decomposition that allows us to claim:

$$\text{REGRET}(K) \stackrel{def}{=} \sum_{k=1}^{K} V_1^{\pi^*}(s) - V_1^{\tilde{\pi}_k}(s) \stackrel{Optimism}{\leq} \sum_{k=1}^{K} \overline{V}_1^{\tilde{\pi}_k}(s) - V_1^{\tilde{\pi}_k}(s) \tag{130}$$

$$\leq \sum_{k=1}^{K} \sum_{t=1}^{H} \mathbb{E}\left(\tilde{r}_k(s_t, a) - r(s_t, a) + (\tilde{p}_k(\cdot \mid s_t, a) - p(\cdot \mid s_t, a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \right) \tag{131}$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s, a) \left((\tilde{r}_k(s, a) - r(s, a)) + (\tilde{p}_k(\cdot \mid s, a) - p(\cdot \mid s, a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \right) \tag{132}$$

$$+ \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \notin L_k} w_{tk}(s, a) \underbrace{\left((\tilde{r}_k(s, a) - r(s, a)) + (\tilde{p}_k(\cdot \mid s, a) - p(\cdot \mid s, a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \right)}_{\leq H} \tag{133}$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s, a) \left((\tilde{r}_k(s, a) - r(s, a)) + (\tilde{p}_k(\cdot \mid s, a) - p(\cdot \mid s, a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \right) + c_{200,1} LSAH^2, \tag{134}$$

for some constant $c_{200,1}$, where the bound in the last passage follows from Lemma 7. By adding and subtracting $\hat{p}_k(\cdot \mid s, a)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k}$ and also $p(\cdot \mid s, a)^\top V_{t+1}^{\pi^*}$ to the above we get the upper bound below:

$$\leq \sum_{k=1}^{K} \sum_{t \in [H]} \sum_{(s,a)} w_{tk}(s, a) \left(\underbrace{(\tilde{r}_k(s, a) - r(s, a))}_{\text{Reward Estimation and Optimism}} + \underbrace{(\tilde{p}_k(s, a) - \hat{p}(s, a))^\top \overline{V}_{t+1}^{\tilde{\pi}_k}}_{\text{Transition Dynamics Optimism}} \right.$$

$$\left. + \underbrace{(\hat{p}_k(s, a) - p(s, a))^\top V_{t+1}^*}_{\text{Transition Dynamics Estimation}} + \underbrace{(\hat{p}_k(s, a) - p(s, a))^\top \left(\overline{V}_{t+1}^{\tilde{\pi}_k} - V_{t+1}^* \right)}_{\text{Lower Order Term}} \right) + c_{200,1} LSAH^2. \tag{135}$$

Here $w_{tk}(s, a)$ is the visitation probability to the (s, a) pair at timestep t in episode k. Each term is bounded in lemmata 8,9,10 and 11 to obtain

$$\leq c_{2,1,3} \left(\left(\sqrt{\mathbb{C}_r^* SAT} + SA + \sqrt{\mathbb{C}^* SAT} + (J + B_p) SA + SAH(F + D + H^{\frac{3}{2}}) + \right. \right. \tag{136}$$

$$\sqrt{S} SAH(F + D + H^{\frac{3}{2}}) + S^2 AH + SAH^2 \Big) \tilde{L}^2 \Big) \tag{137}$$

$$\leq c_{2,1,4} \left((\sqrt{\mathbb{C}_r^* SAT} + \sqrt{\mathbb{C}^* SAT} + JSA + B_p SA + \sqrt{S} SAH(F + D + H^{\frac{3}{2}})) \tilde{L}^3 \right) \tag{138}$$

after simplification. Cauchy-Schwartz immediately implies the following bound:

$$\leq c_{2,1,5} \left((\sqrt{(\mathbb{C}_r^* + \mathbb{C}^*) SAT} + \sqrt{S} SAH(F + D + H^{\frac{3}{2}})) \tilde{L}^3 \right) \tag{139}$$

after absorbing the constant into the lower order term.

We now do the same argument but instead use the variants of Lemmas 9 and 10 that express their bounds as a function of $\mathbb{C}^\pi$. This yields:

$$\leq c_{2,1,6} \left((\sqrt{(\mathbb{C}_r^* + \mathbb{C}^\pi) SAT} + \sqrt{S} SAH(F + D + H^{\frac{3}{2}})) \tilde{L}^3 + B_v \sqrt{SAH^2 \mathcal{R}(K) \tilde{L}} \right). \tag{140}$$

We can re-express this regret bound as follows. Outside of the failure event, and where s_{1k} is the arbitrary starting state in episode k), we have:

$$\sum_{k=1}^{K} \left(V_1^{\pi^*}(s_{1k}) - V_{1k}^{\tilde{\pi}_k}(s_{1k}) \right) \overset{def}{=} \mathcal{R}(K) \leq c_{2,1,6}Y + c_{2,1,6}M\sqrt{\mathcal{R}(k)} \tag{141}$$

with

$$Y = (\sqrt{(\mathbb{C}_r^* + \mathbb{C}^\pi)SAT} + \sqrt{S}SAH(F + D + H^{\frac{3}{2}}))\tilde{L}^3 \tag{142}$$

$$M = B_v\sqrt{SAH^2\tilde{L}}. \tag{143}$$

This is satisfied as long as:

$$\mathcal{R}(K) - c_{2,1,6}M\sqrt{\mathcal{R}(k)} - c_{2,1,6}Y \leq 0. \tag{144}$$

We can solve the quadratic equation (quadratic in $\sqrt{\mathcal{R}(K)}$). This implies that the largest that $\sqrt{\mathcal{R}(K)}$ can be is:

$$\sqrt{\mathcal{R}(K)} \leq \frac{1}{2}\left(c_{2,1,6}M + \sqrt{c_{2,1,6}^2 M^2 + 4c_{2,1,6}Y} \right). \tag{145}$$

By squaring and applying Cauchy-Schwartz we obtain

$$\mathcal{R}(K) \leq c_{2,1,6}^2 M^2 + c_{2,1,6}^2 M^2 + 4c_{2,1,6}Y, \tag{146}$$

completing the proof of the main result. $\qquad\square$

H.2. Regret Bounds with Bernstein Inequality

We now specialize the result of Theorem 2 when Bernstein Inequality is used. First we check that Bernstein's Inequality satisfies assumption 1 and 2. Bernstein's inequality guarantees that with probability at least $1 - \delta'$ we have that:

$$|(\hat{p}_k(s,a) - p(s,a))^\top V_{t+1}^{\pi^*}| \leq \sqrt{\frac{2\operatorname{Var}_p V_{t+1}^{\pi^*} \ln \frac{2SAT}{\delta'}}{n_k(s,a)}} + \frac{H\ln\frac{2SAT}{\delta'}}{3n_k(s,a)} \overset{def}{=} \phi(p(s,a), V_{t+1}^{\pi^*}). \tag{147}$$

after a union bound on the number of states S, actions A and visits $1, ..., T$ to the specific state-action pair (s, a).

Proposition 2 combined with Theorem 2 and a recursive application of the law of total variance is the proof of the following proposition:

Proposition 6 (Problem Independent Bound for EULER with Bernstein Inequality). *If EULER is run with Bernstein Inequality defined in equation 47 with B_p and B_v and ϕ defined in proposition 2 then with probability at least $1 - \delta$ the regret of EULER at timestep T is bounded by the minimum between*

$$\tilde{O}\left(\sqrt{\mathbb{Q}^*SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H}) \right) \tag{148}$$

and

$$\tilde{O}\left(\sqrt{\frac{(\mathcal{G})^2}{H}SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H}) \right). \tag{149}$$

jointly for all episodes $k \in [K]$.

Proof. Proposition 2 shows that Bernstein Inequality of equation 47 is admissible with $B_p = \tilde{O}(H)$, $B_v = \tilde{O}(1)$, $J = \tilde{O}(H)$ so that $F + D = c_{200,3}\tilde{L}H\sqrt{S}$, for some constant $c_{200,3}$, by direct computation. This allows us to apply Theorem 2 and compute an explicit form for the lower order term and the constants $\mathbb{C}^*$, $\mathbb{C}^\pi$.

To obtain the problem dependent bound notice that with the definition of $\mathbb{Q}^*$ in the main text in equation 1 and of the $\mathbb{Q}_t(\cdot,\cdot)$ random variables in the same section in the main text:

$$\mathbb{C}_r^* + \mathbb{C}^* \overset{def}{=} \tag{150}$$

$$= \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim\tilde{\pi}_k} \left(\text{Var}\left(R(s,a) \mid (s,a)\right)\right) + \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim\tilde{\pi}_k} \text{Var}\left(V_{t+1}^{\pi^*}(s^+) \mid (s,a)\right) \tag{151}$$

$$= \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim\tilde{\pi}_k} \left(\text{Var}\left(R(s,a) \mid (s,a)\right) + \text{Var}\left(V_{t+1}^{\pi^*}(s^+) \mid (s,a)\right)\right) \tag{152}$$

$$\overset{(a)}{=} \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim\tilde{\pi}_k} \left(\text{Var}\left(R(s,a) + V_{t+1}^{\pi^*}(s^+) \mid (s,a)\right)\right) \tag{153}$$

$$= \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim\tilde{\pi}_k} \left(\text{Var}\left(\mathbb{Q}_t(s,a) \mid (s,a)\right)\right) \overset{def}{\leq} \mathbb{Q}^* \tag{154}$$

Notice that (a) follows by independence of the sampled reward and transition given an (s,a) pair. This gives the problem dependent bound (also in the main text, Theorem 1).

To obtain the problem-independent worst case guarantee we use a Law of Total Variance argument. Using the variant given by equation 124 in Theorem 2 we need to bound :

$$\mathbb{C}^\pi = \frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}_{\tilde{\pi}_k} \left(\underset{\tilde{\pi}_k}{\text{Var}}\left(V_{t+1}^{\tilde{\pi}_k}(s_{t+1}) \middle| s_t\right) \middle| s_1\right) \tag{155}$$

$$= \frac{1}{T}\sum_{k=1}^{K} \mathbb{E}_{\tilde{\pi}_k} \left(\left(\sum_{t=1}^{H} r(s_t, \tilde{\pi}_k(s_t,t)) - V_1^{\tilde{\pi}_k}(s_1)\right)^2 \middle| s_1\right) \leq \frac{1}{T}K\mathcal{G}^2 = \frac{\mathcal{G}^2}{H} \tag{156}$$

where the second equality follows from a law of total variance argument (see (Azar et al., 2017) for example) reproduced in lemma 15 yielding the stated worst case bound. Expression $\mathbb{E}_{\tilde{\pi}_k}\left(\left(\sum_{t=1}^{H} r(s_t, \tilde{\pi}_k(s_t,t)) - V_1^{\tilde{\pi}_k}(s_1)\right)^2 \middle| s_1\right)$ is the variance of the returns (with fixed rewards) induced by the MDP dynamics upon starting from s_1 and following $\tilde{\pi}_k$. Since the random per episode return $\sum_{t=1}^{H} R(s_t, \tilde{\pi}_k(s_t)) \leq \mathcal{G}$, it must be that $\sum_{t=1}^{H} r(s_t, \tilde{\pi}_k(s_t)) \leq \mathcal{G}$ as well. The variance of a random variable is upper bounded by the range square, justifying the inequality. Finally, plugging in $\mathbb{C}^\pi$ and $\mathbb{C}_r^*$ into equation 124 in Theorem 2 concludes the proof of the result.

This proposition is also restated in the main text as Theorem 1. $\qquad\square$

H.3. Regret Bound in Deterministic Domain with Bernstein Inequality

We now examine the regret of EULER when used with Bernstein Inequality in deterministic domains.

Proposition 7. *If* EULER *is run on a deterministic MDP then the regret is bounded by* $\tilde{O}(SAH^2)$.

Proof. Define as $\mathcal{N}$ as the set of episodes in which the agent visits an (s,a) that is not in L_k. Since the domain is deterministic, each time an (s,a) pair is visited we have $w_{tk}(s,a) = 1$ and hence there can be at most $\tilde{O}(H)$ episodes in which (s,a) is visited but $(s,a) \notin L_k$. Since there are at most SA state and action pairs we have that there are at most $\tilde{O}(SAH)$ such episodes, with a regret at most $\tilde{O}(SAH^2)$. Therefore for any starting state s_k:

$$\sum_{k\in\mathcal{N}} V_1^{\pi^*}(s_k) - V_1^{\tilde{\pi}_k}(s_k) \leq c_{1,2,3}SAH^2. \tag{157}$$

Under the episodes not in $\mathcal{N}$ there is zero probability of visiting a new (s,a) pair, and therefore the maximum likelihood estimate the transition probability is exact. That is, using optimism (a):

$$\text{REGRET}(K) = \sum_{k=1}^{K} V_1^{\pi^*}(s_k) - V_1^{\tilde{\pi}_k}(s_k) \overset{(a)}{\leq} \sum_{k=1}^{K} \overline{V}_{1k}^{\tilde{\pi}_k}(s_k) - V_1^{\tilde{\pi}_k}(s_k) \tag{158}$$

$$= \sum_{k \notin \mathcal{N}} \overline{V}_{1k}^{\tilde{\pi}_k}(s_k) - V_1^{\tilde{\pi}_k}(s_k) + \sum_{k \in \mathcal{N}} \overline{V}_{1k}^{\tilde{\pi}_k}(s_k) - V_1^{\tilde{\pi}_k}(s_k) \tag{159}$$

$$\leq \sum_{k \notin \mathcal{N}} V_1^{\pi^*}(s_k) - V_1^{\tilde{\pi}_k}(s_k) + c_{1,2,3} SAH^2 \tag{160}$$

$$\leq \sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{s,a} w_{tk}(s,a) \left(\tilde{r}(s,a) - r(s,a) + (\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \right) + c_{1,2,3} SAH^2. \tag{161}$$

Bounding the Rewards An application of lemma 8 yields:

$$\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{s,a} w_{tk}(s,a) \left(\tilde{r}(s,a) - r(s,a) \right) \leq c_{1,2,1} SA\tilde{L}^3 \tag{162}$$

since $\mathbb{C}_r^* = 0$. The lemma can be applied because if an episode is not in $\mathcal{N}$ then all $(s,a) \in L_k$.

For the rest of the proof we focus on bounding the exploration bonus:

$$\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{s,a} w_{tk}(s,a) \left(\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a) \right)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \tag{163}$$

$$\leq \sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{s,a} w_{tk}(s,a) b(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}, \underline{V}_{t+1k}^{\tilde{\pi}_k}) \tag{164}$$

$$\leq \sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{s,a} w_{tk}(s,a) \left(\phi(\hat{p}_k(\cdot \mid s,a), \overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{4J + B_p}{n_k(s,a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} \right). \tag{165}$$

using the definition of bonus 5 (here $\phi(\cdot, \cdot)$ is the true Bernstein Inequality evaluated with the empirical quantities). Before bounding the above term we need to understand how the correction term behaves.

Bounding the Delta Optimism in Deterministic Domains We wish to show that

$$\overline{V}_{tk}^{\tilde{\pi}_k}(s) - \underline{V}_{tk}^{\tilde{\pi}_k}(s) \leq C_1 \sum_{\tau=t}^{H} \frac{H}{n_k(s_\tau, a)} \times \tilde{L} \tag{166}$$

where C_1 is some absolute numeric constant and s_t are the states encountered upon following the agent chosen policy. To achieve this proceed as in proposition 5 until equation 110 to get:

$$\overline{V}_{tk}^{\tilde{\pi}_k}(s) - \underline{V}_{tk}^{\tilde{\pi}_k}(s) \leq 2b_k^r(s,a) + p(\cdot \mid s,a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k} \right) + (\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k} \right) \tag{167}$$

$$+ 2\frac{g(p(\cdot \mid s,a), V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}} + 2\frac{J}{n_k(s,a)} + 4\left(\frac{4J + B_p}{n_k(s,a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} \right) \tag{168}$$

$$\overset{(a)}{=} 2b_k^r(s,a) + p(\cdot \mid s,a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k} \right) + 2\frac{J}{n_k(s,a)} + 4\left(\frac{4J + B_p}{n_k(s,a)} + \frac{B_v \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} \right) \tag{169}$$

where (a) follows from the fact that the maximum likelihood is exact for episodes not in $\mathcal{N}$ and so the relevant terms above vanish from the expression. If Bernstein Inequality is used then as explained in proposition 2 $B_v = \tilde{O}(1), B_p = \tilde{O}(H), J = \tilde{O}(H)$ and also $b_k^r(s,a) = C_1/n_k(s,a) \times$ polylog since both the variance and the empirical variance are zero. Therefore for appropriate constants $C_1, C_2, \dots$ the above inequality can be written as:

$$= \left(\frac{C_1 + C_2 H}{n_k(s,a)} + (\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1})) + C_3 \sqrt{\frac{(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}))^2}{n_k(s,a)}} \right) \tilde{L} \tag{170}$$

$$= \left(\frac{C_1 + C_2 H}{n_k(s,a)} + (\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1})) + C_3 \frac{\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1})}{\sqrt{n_k(s,a)}} \right) \tilde{L} \tag{171}$$

$$\leq \left(\frac{C_1 + C_2 H}{\sqrt{n_k(s,a)}} + (\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1})) + C_3 \frac{H}{\sqrt{n_k(s,a)}} \right) \tilde{L} \tag{172}$$

$$\leq (\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1})) + C_4 \frac{H}{\sqrt{n_k(s,a)}} \tilde{L} \tag{173}$$

$$\leq \sum_{\tau=t}^{H} C_4 \frac{H}{\sqrt{n_k(s_\tau,a)}} \tilde{L} \tag{174}$$

$$\tag{175}$$

The last passage follows by induction and completes the proof of equation 166 for episodes $\notin \mathcal{N}$. Equipped with this it remains to bound the bonus on the transition dynamics.

Bounding the Transition Dynamics   Proceed as in lemma 10 up to equation 224. Since we are using Bernstein Inequality, $\mathbb{C}^* = 0, B_p = \tilde{O}(H), J = \tilde{O}(H)$ in deterministic domains and so the regret reads:

$$\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{s,a} w_{tk}(s,a) (\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a))^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \leq c_{1,2,6} \Big(SAH+ \tag{176}$$

$$+ B_v \underbrace{\sqrt{\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s,a)}{n_k(s,a)}}}_{\tilde{O}(SA)} \times \sqrt{\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}^2} \tag{177}$$

$$c_{1,2,7} SAH + c_{1,2,8} \sqrt{SA\tilde{L}} \times \sqrt{\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}^2} \Big) \tag{178}$$

We focus on the last factor:

$$\sqrt{\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \hat{p}_k(\cdot \mid s,a) (\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k})^2} = \sqrt{\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} (\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}) - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1}))^2} \tag{179}$$

$$\leq c_{1,2,20} \sqrt{\sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \left(\sum_{\tau=t}^{H} \frac{H}{\sqrt{n_k(s_\tau,a)}} \right)^2} \tag{180}$$

$$\leq c_{1,2,21} \sqrt{H \sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \sum_{\tau=t}^{H} \left(\frac{H}{\sqrt{n_k(s_\tau,a)}} \right)^2} \tag{181}$$

$$\leq c_{1,2,22} \sqrt{H^3 \sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \sum_{\tau=t}^{H} \left(\frac{1}{\sqrt{n_k(s_\tau,a)}} \right)^2} \tag{182}$$

$$\leq c_{1,2,23} \sqrt{H^3 \sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \sum_{\tau=1}^{H} \left(\frac{1}{\sqrt{n_k(s_\tau,a)}} \right)^2} \tag{183}$$

$$\leq c_{1,2,24} \sqrt{H^4 \sum_{k \notin \mathcal{N}} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s,a)}{n_k(s,a)}} \tag{184}$$

$$\leq c_{1,2,25}\sqrt{SA\tilde{L}}H^2 \tag{185}$$

Thus

$$\sum_{k\notin\mathcal{N}}\sum_{t=1}^{H}\sum_{s,a}w_{tk}(s,a)\left(\tilde{p}_k(\cdot\mid s,a)-\hat{p}_k(\cdot\mid s,a)\right)^{\top}\overline{V}_{t+1k}^{\tilde{\pi}_k}\leq c_{1,2,7}SAH\,\text{polylog}+c_{1,2,9}\sqrt{SA\tilde{L}}\times c_{1,2,10}\sqrt{SA\tilde{L}}H^2 \tag{186}$$

$$\leq c_{1,2,11}SAH^2\tilde{L}. \tag{187}$$

Concluding the Proof of the Regret Bound on Deterministic Domain Summing the regret for episodes not in $\mathcal{N}$, the reward optimism and the transition dynamics optimism one obtains the final regret bound of order:

$$\leq c_{1,2,30}SAH^2\tilde{L}^3. \tag{188}$$

Notice that there are no failure events to consider, so this is a deterministic statement. $\qquad\square$

H.4. Reward Estimation and Optimism

Lemma 8 (Reward Estimation and Optimism). *Outside of the failure event it holds that:*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\left(\tilde{r}_k(s,a)-r(s,a)\right)\quad\leq\quad c_{8,3,1}\tilde{L}^3\left(\left(\sqrt{\mathbb{C}_r^*SAT}+SA\right)\right) \tag{189}$$

$$=\quad c_{8,3,2}\tilde{L}^3\times\left(\sqrt{\frac{\mathcal{G}}{H}SAT}+SA\right). \tag{190}$$

where

$$\mathbb{C}_r^*=\frac{1}{T}\left(\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\,\text{Var}\,R(s,a)\right)\leq\frac{\mathcal{G}}{H} \tag{191}$$

Proof. The optimistic reward is obtained by adding the reward bonus the empirical reward estimate:

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\left(\tilde{r}_k(s,a)-r(s,a)\right)\leq\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)b_k^r(s,a) \tag{192}$$

$$\leq c_{8,3,4}\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\left(\sqrt{\frac{2\widehat{\text{Var}}R(s,a)\ln\left(\frac{4SAT}{\delta'}\right)}{n_k(s,a)}}+\frac{7\ln\left(\frac{4SAT}{\delta'}\right)}{3n_k(s,a)}\right) \tag{193}$$

$$\leq c_{8,3,5}\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\left(\sqrt{\frac{\widehat{\text{Var}}R(s,a)}{n_k(s,a)}}+\frac{1}{n_k(s,a)}\right)\times 3\ln\left(\frac{4SAT}{\delta'}\right) \tag{194}$$

$$\leq c_{8,3,6}\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\left(\sqrt{\frac{\left(\sqrt{\text{Var}\,R(s,a)}+\sqrt{2\ln(2SAT/\delta')/n_k(s,a)}\right)^2}{n_k(s,a)}}+\frac{1}{n_k(s,a)}\right)\times 3\ln\left(\frac{4SAT}{\delta'}\right) \tag{195}$$

$$\leq c_{8,3,7}\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\left(\sqrt{\frac{\text{Var}\,R(s,a)}{n_k(s,a)}}+\sqrt{\frac{2\ln(2SAT/\delta')/n_k(s,a)}{n_k(s,a)}}+\frac{1}{n_k(s,a)}\right)\times 3\ln\left(\frac{4SAT}{\delta'}\right) \tag{196}$$

$$\leq c_{8,3,8} \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\sqrt{\frac{\operatorname{Var} R(s,a)}{n_k(s,a)}} + \frac{\sqrt{2\ln(2SAT/\delta')}}{n_k(s,a)} + \frac{1}{n_k(s,a)} \right) \times 3\ln\left(\frac{4SAT}{\delta'} \right) \tag{197}$$

$$\leq c_{8,3,9} \tilde{L}^2 \times \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\sqrt{\frac{\operatorname{Var} R(s,a)}{n_k(s,a)}} + \frac{1}{n_k(s,a)} \right) \tag{198}$$

$$\leq c_{8,3,10} \tilde{L}^2 \times \left(\sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s,a)}{n_k(s,a)}} \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \operatorname{Var} R(s,a)} + \right. \tag{199}$$

$$\left. \left(\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s,a)}{n_k(s,a)} \right) \right. \tag{200}$$

where the fourth line follows from lemma 1 that bounds the difference between the empirical and estimated variances, and the following inequalities come from algebraic manipulations and consolidating the $\operatorname{polylog}(S, A, H, T, 1/\delta')$ terms into a single expression and moving this to outside the sum (since they are independent of the variables in the sum). The final inequality follows from Cauchy Schwartz, yielding the result after the application of lemma 13.

To compute the upper bound of equation 191 we proceed as follows. The $w_{tk}(s,a)$ are the probability of visiting state s and taking action a there in timestep t of episode k given the policy selected by the agent in episode k. The core idea is that $\mathbb{C}_r^*$ is a per-step average of the reward variance within an episode. Regardless of the policy followed by the agent, the sum of reward random variables $R(\cdot, \cdot)$ cannot exceed $\mathcal{G}$.

$$\mathbb{C}_r^* = \frac{1}{T} \left(\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \operatorname{Var} R(s,a) \right) \tag{201}$$

$$\overset{(a)}{\leq} \frac{1}{T} \left(\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a)} w_{tk}(s,a) \operatorname{Var} R(s,a) \right) \tag{202}$$

$$\overset{(b)}{=} \frac{1}{T} \left(\sum_{k=1}^{K} \sum_{t=1}^{H} \mathbb{E}_{s_t \sim \tilde{\pi}_k} \operatorname{Var}\left(R(s_t, \tilde{\pi}_k(s_t)) \right) \right) \tag{203}$$

$$\overset{(c)}{\leq} \frac{1}{T} \left(\sum_{k=1}^{K} \sum_{t=1}^{H} \operatorname{Var}\left(R(s_t, \tilde{\pi}_k(s_t)) \right) \right) \tag{204}$$

$$\overset{(d)}{\leq} \frac{1}{T} \left(\sum_{k=1}^{K} \operatorname{Var}\left(\sum_{t=1}^{H} R(s_t, \tilde{\pi}_k(s_t)) \right) \right) \tag{205}$$

$$\overset{(e)}{\leq} \frac{1}{T} \left(\sum_{k=1}^{K} (\mathcal{G})^2 \right) \tag{206}$$

$$= \frac{(\mathcal{G})^2}{H} \tag{207}$$

In the above inequalities, (a) is an upper bound because it drops the condition $(s,a) \in L_k$, (b) is rewriting $w_{tk}(s,a)$ as an expectation over the trajectories visited by $\tilde{\pi}_k$ in timestep t and episode k. Next, (c) is due to the law of total variance, also known as Eve's law (notice that the variance is over the encountered states here as well). Since the reward random variables are independent, we can bring the sum inside in step (d). By definition, this is the variance of the return, which is at most the maximum return squared $(\mathcal{G})^2$ (step (e)).

$\square$

H.5. Transition Dynamics Estimation

Lemma 9 (Transition Dynamics Estimation). *Outside of the failure event if ϕ is admissible then it holds that:*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\hat{p}_k(\cdot\mid s,a)-p(\cdot\mid s,a)\right)^T V_{t+1}^{\pi^*} \le c_{9,3,0}\left(\sqrt{\mathbb{C}^* SAT}+JSA\right)\tilde{L}. \tag{208}$$

The following bound also holds:

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\hat{p}_k(\cdot\mid s,a)-p(\cdot\mid s,a)\right)^T V_{t+1}^{\pi^*} \tag{209}$$

$$\le c_{9,3,1}\left(\left(\sqrt{\mathbb{C}^{\pi} SAT}+JSA+SAH(F+D+H^{\frac{3}{2}})\right)\tilde{L}+B_v\sqrt{SAH^2\mathcal{R}(K)\tilde{L}}\right). \tag{210}$$

Proof. Using the definition of ϕ, outside of the failure event it holds that:

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\hat{p}_k(\cdot\mid s,a)-p(\cdot\mid s,a)\right)^T V_{t+1}^{\pi^*} \le \sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\frac{g(p,V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}}+\frac{J}{n_k(s,a)}\right). \tag{211}$$

Next, Cauchy-Schwartz justifies the following upper bound:

$$\le \sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(g(p,V_{t+1}^{\pi^*})\right)^2}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}\frac{w_{tk}(s,a)}{n_k(s,a)}}+J\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}\frac{w_{tk}(s,a)}{n_k(s,a)} \tag{212}$$

Finally, using lemma 13 and definition 5 of $\mathbb{C}^*$ we can obtain the statement:

$$\le c_{9,3,2}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(g(p,V_{t+1}^{\pi^*})\right)^2}\times c_{9,3,3}\left(\sqrt{SA\tilde{L}}\right)+c_{9,3,4}\left(JSA\right)\tilde{L} \le c_{9,3,5}\left(\sqrt{\mathbb{C}^* SAT}+JSA\right)\tilde{L}. \tag{213}$$

To obtain the second bound, we use a similar argument coupled with lemma 14. $\qquad\square$

H.6. Transition Dynamics Optimism

Lemma 10 (Transition Dynamics Optimism). *Outside of the failure event if ϕ is admissible it holds that:*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\tilde{p}_k(\cdot\mid s,a)-\hat{p}_k(\cdot\mid s,a)\right)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} = \tag{214}$$

$$\le c_{10,3,1}\left(\sqrt{\mathbb{C}^* SAT}+(J+B_p)SA+SAH(F+D+H^{\frac{3}{2}})\right)\tilde{L}^2. \tag{215}$$

The bound below also hold:

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\tilde{p}_k(\cdot\mid s,a)-\hat{p}_k(\cdot\mid s,a)\right)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} = \tag{216}$$

$$\le c_{10,3,2}\left(\left(\sqrt{\mathbb{C}^{\pi} SAT}+(J+B_p)SA+SAH(F+D+H^{\frac{3}{2}})\right)\tilde{L}^2+B_v\sqrt{SAH^2\mathcal{R}(K)\tilde{L}}\right). \tag{217}$$

Proof. We begin by using definition 5 (for the bonus) to justify (a):

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\tilde{p}_k(\cdot\mid s,a)-\hat{p}_k(\cdot\mid s,a)\right)^\top \overline{V}_{t+1k}^{\tilde{\pi}_k} \tag{218}$$

$$\overset{a}{\leq} c_{10,3,3} \sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\phi(\hat{p}_k(\cdot\mid s,a),\overline{V}_{t+1k}^{\tilde{\pi}_k}) + \frac{B_v\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} + \frac{B_p + J}{n_k(s,a)}\right) \tag{219}$$

$$\overset{b}{\leq} c_{10,3,4} \sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\phi(p(\cdot\mid s,a),V_{t+1}^{\pi^*}) + \frac{B_v\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}} + \frac{B_p + J}{n_k(s,a)}\right) \tag{220}$$

while (b) is justified by lemma 4 and 5. Using the functional form for ϕ we obtain the upper bound below (c):

$$\overset{c}{\leq} c_{10,3,4} \sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k} w_{tk}(s,a)\left(\underbrace{\frac{g(p,V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}} + \frac{J + B_p}{n_k(s,a)}}_{\approx\text{Transition Dynamics Estimation}} + \underbrace{\frac{B_v\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}}{\sqrt{n_k(s,a)}}}_{\text{Lower Order Term}}\right) \tag{221}$$

The term "$\approx$ Transition Dynamics Estimation" is nearly identical to what appears in the proof of lemma 9 and can be bounded in the same way. That is, apply Cauchy-Schwartz first and use lemma 13 along with the definition of $\mathbb{C}^*$ to get to the bound below:

$$\overset{d}{\leq} c_{10,3,5}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}\frac{w_{tk}(s,a)}{n_k(s,a)}}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)g(p,V_{t+1}^{\pi^*})^2} + (J+B_p)\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}\frac{w_{tk}(s,a)}{n_k(s,a)} \tag{222}$$

$$\leq c_{10,3,6}\left(\sqrt{\mathbb{C}^*SAT} + (J+B_p)SA\right)\tilde{L}. \tag{223}$$

Now we turn our attention to the "Lower Order Term" and apply Cauchy-Schwartz to get:

$$\leq c_{10,3,7}B_v\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}\frac{w_{tk}(s,a)}{n_k(s,a)}} \times \sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}^2} \tag{224}$$

The first factor is bounded by $c_{200,5}B_v\sqrt{SA\tilde{L}}$ (for some constant $c_{200,5}$ by Lemma 13. Notice that

$$\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,\hat{p}}^2 = \hat{p}_k(\cdot\mid s,\tilde{\pi}_k(s,t))^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2 \tag{225}$$

$$= p(\cdot\mid s,\tilde{\pi}_k(s,t))^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2 + (\hat{p}_k(\cdot\mid s,\tilde{\pi}_k(s,t)) - p(\cdot\mid s,\tilde{\pi}_k(s,t)))^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2 \tag{226}$$

$$= \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p}^2 + (\hat{p}_k(\cdot\mid s,\tilde{\pi}_k(s,t)) - p(\cdot\mid s,\tilde{\pi}_k(s,t)))^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2 \tag{227}$$

The above inequality and $\sqrt{a+b} \leq c_{10,3,9}(\sqrt{a}+\sqrt{b})$ for real a,b allows us to write the following upper bound:

$$B_v\sqrt{SA\tilde{L}}c_{10,3,10}\times\left(\sqrt{\underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p}^2}_{(a)}}\right. \tag{228}$$

$$\left. + \sqrt{\underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)\left|\hat{p}_k(\cdot\mid s,a) - p(\cdot\mid s,a)\right)^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2\right|}_{(b)}}\right) \tag{229}$$

To bound (a) we use lemma 12:

$$(a) \leq c_{10,3,11}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)p(\cdot\mid s,a)^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2} \leq c_{12,3,12}\sqrt{SAH^2(F+D)^2 + SAH^5}\tilde{L}. \tag{230}$$

We now bound (b), which is a lower order term. We don't leverage this fact here and bound it trivially by:

$$(b) \le c_{10,3,13} \sqrt{H} \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) (\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k} \right)} \tag{231}$$

The same computation as in Lemma 11 now gives:

$$(b) \le c_{10,3,14} \sqrt{H} \times \sqrt{\sqrt{\bar{S}} SAH(F + D + H^{\frac{3}{2}}) + S^2 AH \tilde{L}} \tag{232}$$

This concludes the proof for the first bound. For the second bound proceed analogously but use the variant given by lemma 14 when bounding the term "$\approx$ Transition Dynamics Estimation" in lemma 9 to obtain:

$$\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\underbrace{\frac{g(p, V_{t+1}^{\pi^*})}{\sqrt{n_k(s,a)}} + \frac{J + B_p}{n_k(s,a)}}_{\approx \text{Transition Dynamics Estimation}} \right) \tag{233}$$

$$\le c_{10,3,15} \left(\left(\sqrt{\mathbb{C}^\pi SAT} + (J + B_p)SA + SAH(F + D + H^{\frac{3}{2}}) \right) \tilde{L}^2 + B_v \sqrt{SAH^2 \mathcal{R}(K)\tilde{L}} \right). \tag{234}$$

This concludes the proof. $\qquad \square$

H.7. Lower Order Term

Lemma 11 (Lower Order Term). *Outside of the failure event for* EULER *it holds that :*

$$\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left| (\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a))^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*} \right) \right| = \tag{235}$$

$$\le c_{11,3,1} \left(\sqrt{\bar{S}} SAH(F + D + H^{\frac{3}{2}}) + S^2 AH \right) \tilde{L}^2. \tag{236}$$

Proof. Using the concentration inequality on equation 54 we get:

$$\le \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \sum_{s'} \sqrt{\frac{p(s' \mid s,a)(1 - p(s' \mid s,a))}{n_k(s,a)}} \left| V_{t+1}^{\pi^*}(s') - \overline{V}_{t+1k}^{\tilde{\pi}_k}(s') \right| \tilde{L}^{0.5} \tag{237}$$

$$+ \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \sum_{s'} \frac{H}{n_k(s,a)} \tilde{L} \tag{238}$$

Since $\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\pi^*} \le \overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}$ pointwise by Proposition 4 and by bounding the second term with Lemma 13 and using $(1 - p) \le 1$ for $p \in [0, 1]$ we obtain:

$$\le \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \sum_{s'} \sqrt{\frac{p(s' \mid s,a)}{n_k(s,a)}} \left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s') - \underline{V}_{t+1k}^{\tilde{\pi}_k}(s') \right) \tilde{L}^{0.5} + c_{11,3,4} S^2 AH \tilde{L}^2). \tag{239}$$

Cauchy-Schwartz leads to the following upper bound:

$$\le c_{11,3,5} \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \left(\sqrt{\frac{S \times p(\cdot \mid s,a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k} \right)^2}{n_k(s,a)}} \right) \tilde{L}^{0.5} + c_{11,3,5a} S^2 AH \tilde{L}^2 \tag{240}$$

One more application of Cauchy-Schwartz gives us:

$$\leq c_{11,3,6}\sqrt{S}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}\frac{w_{tk}(s,a)}{n_k(s,a)}}\times\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)\in L_k}w_{tk}(s,a)p(\cdot\mid s,a)^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}-\underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2}\tilde{L}^{0.5}+c_{11,3,7}S^2AH\tilde{L}^2 \tag{241}$$

Recalling lemma 13 and lemma 12 we obtain:

$$\leq c_{11,3,8}\left(\sqrt{S}\times\sqrt{SA\tilde{L}}\right)\times c_{11,3,9}(\sqrt{SAH^2(F+D)^2+SAH^5})\tilde{L}^{0.5}+c_{11,3,10}S^2AH\tilde{L}^2 \tag{242}$$

which can be simplified to obtain the statement. $\qquad\square$

Lemma 12 (Cumulative Delta Optimism). *Outside of the failure event it holds that:*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s,a}w_{tk}(s,a)p(\cdot\mid s,a)^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}-\underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2=c_{12,3,1}(SAH^2(F+D)^2+SAH^5)\tilde{L}. \tag{243}$$

where F and D are defined in proposition 5.

Proof. Starting from the right hand side

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)}w_{tk}(s,a)p(\cdot\mid s,a)^\top\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}-\underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2 \tag{244}$$

we unroll the inner product between the transition probability vector and the value function obtaining

$$=\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)}w_{tk}(s,a)\left(\sum_{s'}p(s'\mid s,a)\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s')-\underline{V}_{t+1k}^{\tilde{\pi}_k}(s')\right)^2\right). \tag{245}$$

Next, we move the summation operator (over s') outside

$$=\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)}\sum_{s'}w_{tk}(s,a)p(s'\mid s,a)\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s')-\underline{V}_{t+1k}^{\tilde{\pi}_k}(s')\right)^2 \tag{246}$$

and recall that $w_{t+1,k}(s',s,a)\overset{def}{=}w_{tk}(s,a)p(s'\mid s,a)$ is the probability of taking action a in s and then landing in s' at the next timestep.

$$=\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{(s,a)}\sum_{s'}w_{t+1,k}(s',s,a)\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s')-\underline{V}_{t+1k}^{\tilde{\pi}_k}(s')\right)^2. \tag{247}$$

Summing over all possible s,a pairs one obtains the probability of being in s' at timestep $t+1$

$$=\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s'}w_{t+1,k}(s')\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s')-\underline{V}_{t+1k}^{\tilde{\pi}_k}(s')\right)^2 \tag{248}$$

which can be interpreted as an expectation over trajectories identified by $\tilde{\pi}_k$

$$=\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{s_{t+1}\sim\tilde{\pi}_k}\left(\overline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1})-\underline{V}_{t+1k}^{\tilde{\pi}_k}(s_{t+1})\right)^2 \tag{249}$$

$$\leq\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{s_t\sim\tilde{\pi}_k}\left(\overline{V}_{tk}^{\tilde{\pi}_k}(s_t)-\underline{V}_{tk}^{\tilde{\pi}_k}(s_t)\right)^2. \tag{250}$$

The last upper bound follows because we are counting over the same quantities, but we add timestep $t = 1$ and drop timestep $t = H + 1$ for which the value functions are zero. Proposition 5 justifies the first inequality below where F and D are defined in said proposition (here the action a is the action taken by $\tilde{\pi}_k$ in s_τ):

$$\leq c_{12,3,2} \sum_{k=1}^{K} \sum_{t=1}^{H} \mathbb{E}_{s_t \sim \tilde{\pi}_k} \left(\sum_{\tau=t}^{H} \mathbb{E}_{s_\tau \sim \tilde{\pi}_k} \frac{F + D}{\sqrt{n_k(s_\tau, a)}} \,\Big|\, s_t \right)^2 \tag{251}$$

$$\overset{a}{\leq} c_{12,3,3} H \sum_{k=1}^{K} \sum_{t=1}^{H} \mathbb{E}_{s_t \sim \tilde{\pi}_k} \sum_{\tau=t}^{H} \left(\mathbb{E}_{s_\tau \sim \tilde{\pi}_k} \frac{F + D}{\sqrt{n_k(s_\tau, a)}} \,\Big|\, s_t \right)^2 \tag{252}$$

$$\overset{b}{\leq} c_{12,3,4} H \sum_{k=1}^{K} \sum_{t=1}^{H} \mathbb{E}_{s_t \sim \tilde{\pi}_k} \sum_{\tau=t}^{H} \mathbb{E}_{s_\tau \sim \tilde{\pi}_k} \frac{(F + D)^2}{n_k(s_\tau, a)} \,\Big|\, s_t \tag{253}$$

$$\leq c_{12,3,5} H \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{\tau=t}^{H} \mathbb{E}_{s_\tau \sim \tilde{\pi}_k} \frac{(F + D)^2}{n_k(s_\tau, a)} \tag{254}$$

$$\leq c_{12,3,6} H^2 \sum_{k=1}^{K} \sum_{t=1}^{H} \mathbb{E}_{s_t \sim \tilde{\pi}_k} \frac{(F + D)^2}{n_k(s_t, a)} \tag{255}$$

$$\leq c_{12,3,7} H^2 \left(\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s, a) \left(\frac{(F + D)^2}{n_k(s, a)} \right) + \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \notin L_k} w_{tk}(s, a) H^2 \right) \tag{256}$$

$$\leq c_{12,3,7} H^2 (F + D)^2 \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \left(\frac{w_{tk}(s, a)}{n_k(s, a)} \right) + H^4 \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \notin L_k} w_{tk}(s, a) \tag{257}$$

$$= c_{12,3,2} (SAH^2 (F + D)^2 + SAH^5) \tilde{L} \tag{258}$$

The last passage follows from lemma 13 and 7, while (a) and (b) follow from Cauchy-Schwartz and Jensen, respectively. To obtain the bound H^2 for the rightmost term (the one corresponding to states $(s, a) \notin L_k$) we used the hard bound discussed in proposition 5 that "caps" $\frac{(F+D)^2}{n_k(s,a)}$ at H^2. □

H.8. Auxiliary Lemmas

Lemma 13 (Visitation Ratio)**.**

$$\sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s, a)}{n_k(s, a)}} \leq c_{13,3,1} \sqrt{SA\tilde{L}} \tag{259}$$

Proof. Recall the definition $\sum_{t=1}^{H} w_{tk}(s, a) = w_k(s, a)$. Lemma 6 ensures step (a) below:

$$\sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s, a)}{n_k(s, a)}} = \sqrt{\sum_{k=1}^{K} \sum_{(s,a) \in L_k} \frac{w_k(s, a)}{n_k(s, a)}} \tag{260}$$

$$= \sqrt{\sum_{k=1}^{K} \sum_{s,a} \frac{w_k(s, a)}{n_k(s, a)} \mathbb{1}\{(s, a) \in L_k\}} \tag{261}$$

$$\overset{(a)}{\leq} c_{13,3,2} \sqrt{\sum_{k=1}^{K} \sum_{s,a} \frac{w_k(s, a)}{\sum_{\iota \leq k} w_\iota(s, a)} \mathbb{1}\{(s, a) \in L_k\}} \tag{262}$$

$$\tag{263}$$

It suffices to study

$$\sum_{k=1}^{K} \frac{w_k(s, a)}{\sum_{\iota \leq k} w_\iota(s, a)} \mathbb{1}\{(s, a) \in L_k\} \tag{264}$$

for a fixed (s, a). The above quantity is non-zero only if $(s, a) \in L_k$ for some k. Since $\sum_{\iota \le k} w_\iota(s, a)$ is strictly increasing with k, if $(s, a) \in L_k$ there must exist a critical episode $k_L \le k$ (that depends on the (s, a) pair) such that for all subsequent episodes $\iota \ge k_L$ we have that $(s, a) \in L_\iota$. Since by definition of k_L it must be that $(s, a) \in L_{k_L}(s, a)$, we must have $\sum_{\iota < k_L} w_\iota(s, a) + w_{k_L}(s, a) = \sum_{\iota \le k_L} w_\iota(s, a) > 2H$ by definition 6, implying $\sum_{\iota < k_L} w_\iota(s, a) > H$. This way we lower bound the summation in the denominator as:

$$\sum_{\iota \le k} w_\iota(s, a) = \sum_{\iota < k_L} w_\iota(s, a) + \sum_{k_L \le \iota \le k} w_\iota(s, a) > H + \sum_{k_L \le \iota \le k} w_\iota(s, a) \tag{265}$$

Therefore equation 264 can be upper bounded as:

$$\sum_{k=1}^{K} \frac{w_k(s, a)}{H + \sum_{k_L \le \iota \le k} w_\iota(s, a)} \mathbb{1}\{(s, a) \in L_k\}. \tag{266}$$

Since the indicator $\mathbb{1}\{(s, a) \in L_k\}$ is non-zero only when $k_L \le k \le K$, we can rewrite the above equation as:

$$\sum_{k_L \le k \le K} \frac{w_k(s, a)}{H + \sum_{k_L \le \iota \le k} w_\iota(s, a)}. \tag{267}$$

The above expression can be simplified in notation by setting $a_1 = w_{k_L}(s, a), a_2 = w_{k_L+1}(s, a), \dots, a_{K-k_L+1} = w_K(s, a)$. Now define the function $F(x) = \sum_{i=1}^{\lfloor x \rfloor} a_i + a_{\lceil x \rceil}(x - \lfloor x \rfloor)$, which is a function that coincides with the summation $\sum_{i=1}^{x} a_i$ for integer values of x and interpolates between them. Its derivative is $f(x) = a_{\lceil x \rceil}$. This way we can write:

$$\sum_{k=1}^{K-k_L+1} \frac{a_k}{H + \sum_{i=1}^{k} a_i} = \sum_{k=1}^{K-k_L+1} \frac{f(k)}{H + F(k)} \tag{268}$$

We have that $F(x) = \sum_{i=1}^{\lfloor x \rfloor} a_i + a_{\lceil x \rceil}(\lceil x \rceil - x) \le \sum_{i=1}^{\lfloor x \rfloor} a_i + a_{\lceil x \rceil} = \sum_{i=1}^{\lceil x \rceil} a_i = F(\lceil x \rceil)$ which justifies

$$\frac{f(\lceil x \rceil)}{H + F(\lceil x \rceil)} \le \frac{f(x)}{H + F(x)}. \tag{269}$$

Since the lhs is a step function, integrating the above yields:

$$\sum_{k=1}^{K-k_L+1} \frac{f(k)}{H + F(k)} = \int_0^{K-k_L+1} \frac{f(\lceil x \rceil)}{H + F(\lceil x \rceil)} dx \tag{270}$$

$$\le \int_0^{K-k_L+1} \frac{f(x)}{H + F(x)} dx \tag{271}$$

$$= \ln(H + F(K - k_L + 1)) - \ln(H + F(0)) \le \ln(2H + KH) \le \ln(T) \le \tilde{L}. \tag{272}$$

Summing over all the (s, a) pairs yields the result. $\qquad\square$

Lemma 14 (Bound Bridge).

$$\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s, a) \frac{g(p, V_{t+1}^{\pi^*})}{\sqrt{n_k(s, a)}} - \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s, a) \frac{g(p, V_{t+1}^{\tilde{\pi}_k})}{\sqrt{n_k(s, a)}} \tag{273}$$

$$\le c_{14,3,1} \left(SAH(F + D + H^{\frac{3}{2}})\tilde{L} + B_v \sqrt{SAH^2 \mathcal{R}(K)\tilde{L}} \right). \tag{274}$$

Proof. Equation 43 in assumption 1 ensures:

$$\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s, a) \frac{g(p, V_{t+1}^{\pi^*})}{\sqrt{n_k(s, a)}} - \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s, a) \frac{g(p, V_{t+1}^{\tilde{\pi}_k})}{\sqrt{n_k(s, a)}} \tag{275}$$

$$\leq B_v \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1}^{\tilde{\pi}_k}\|_{2,p}}{\sqrt{n_k(s,a)}}. \tag{276}$$

By adding and subtracting $V_{t+1,k}^{\pi^*}$ inside the norm operator we obtain:

$$= B_v \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1,k}^{\pi^*} + V_{t+1,k}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\|_{2,p}}{\sqrt{n_k(s,a)}} \tag{277}$$

$$\leq \underbrace{B_v \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - V_{t+1,k}^{\pi^*}\|_{2,p}}{\sqrt{n_k(s,a)}}}_{A} + \underbrace{B_v \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \frac{\|V_{t+1,k}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\|_{2,p}}{\sqrt{n_k(s,a)}}}_{B}. \tag{278}$$

In particular, the upper bound follows by the triangle inequality. Below we bound term A. Lemma 5 and proposition 4 ensure the upper bound below:

$$A \leq B_v \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \frac{\|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p}}{\sqrt{n_k(s,a)}} \tag{279}$$

from which Cauchy-Schwartz yields:

$$\leq B_v \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s,a)}{n_k(s,a)}} \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p}^2} \tag{280}$$

$$\leq c_{14,3,4} \sqrt{SA\tilde{L}} \times B_v \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \|\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\|_{2,p}^2} \tag{281}$$

$$\leq c_{14,3,4} \sqrt{SA\tilde{L}} \times B_v \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{s,a} w_{tk}(s,a) p(\cdot \mid s,a)^\top \left(\overline{V}_{t+1k}^{\tilde{\pi}_k} - \underline{V}_{t+1k}^{\tilde{\pi}_k}\right)^2} \tag{282}$$

$$\leq c_{14,3,4} \sqrt{SA\tilde{L}} \times \sqrt{SAH^2(F+D)^2 + SAH^5\tilde{L}} \leq c_{13,3,5} SAH(F+D+H^{\frac{3}{2}})\tilde{L} \tag{283}$$

where the bounds follow from lemma 13 and 12. It now remains to bound term B. By an identical argument using Cauchy-Schwartz we have that:

$$B = B_v \sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \frac{\|V_{t+1,k}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\|_{2,p}}{\sqrt{n_k(s,a)}} \tag{284}$$

$$\leq B_v \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} \frac{w_{tk}(s,a)}{n_k(s,a)}} \sqrt{\sum_{k=1}^{K} \sum_{t=1}^{H} \sum_{(s,a) \in L_k} w_{tk}(s,a) \|V_{t+1,k}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\|_{2,p}^2} \tag{285}$$

$$\leq c_{14,3,5} B_v \sqrt{SA\tilde{L}} \sqrt{H^2 \mathcal{R}(K)}. \tag{286}$$

The last passage follows from lemma 13 and 16. □

Lemma 15 (LTV). *The following inequality holds true:*

$$\mathbb{E}_{\tilde{\pi}_k}\left(\left(\sum_{t=1}^{H} r(s_t, \tilde{\pi}_k(s_t, t)) - V_1^{\tilde{\pi}_k}(s_1)\right)^2 \Big| s_1\right) = \mathbb{E}_{\tilde{\pi}_k}\left(\sum_{t=1}^{H} \underset{\tilde{\pi}_k}{\mathrm{Var}}\left(V_{t+1}^{\tilde{\pi}_k}(s_{t+1}) \Big| s_t\right) \Big| s_1\right). \tag{287}$$

where the expectation $\mathbb{E}_{\tilde{\pi}_k}(\cdot \mid s_1)$ is taken with respect to the trajectories followed by the agent upon following policy $\tilde{\pi}_k$ starting from s_1.

Proof.

$$\mathbb{E}_{\tilde{\pi}_k}\left(\left(\sum_{t=1}^{H} r(s_t, \tilde{\pi}_k(s_t, t)) - V_1^{\tilde{\pi}_k}(s_1)\right)^2 \bigg| s_1\right) = \tag{288}$$

$$= \mathbb{E}_{\tilde{\pi}_k}\left(\left(\left(r(s_1, \tilde{\pi}_k(s_1, t)) + \sum_{t=2}^{H} r(s_t, \tilde{\pi}_k(s_t, t))\right) - \left(r(s_1, \tilde{\pi}_k(s_1, t)) + \mathbb{E}_{\tilde{\pi}_k} V_2^{\tilde{\pi}_k}(s_2)\right)\right)^2 \bigg| s_1\right) \tag{289}$$

$$= \mathbb{E}_{\tilde{\pi}_k}\left(\left(\sum_{t=2}^{H} r(s_t, \tilde{\pi}_k(s_t, t)) - \mathbb{E}_{\tilde{\pi}_k} V_2^{\tilde{\pi}_k}(s_2)\right)^2 \bigg| s_1\right) \tag{290}$$

$$= \mathbb{E}_{\tilde{\pi}_k}\left(\left(\sum_{t=2}^{H} r(s_t, \tilde{\pi}_k(s_t, t)) - V_2^{\tilde{\pi}_k}(s_2) + V_2^{\tilde{\pi}_k}(s_2) - \mathbb{E}_{\tilde{\pi}_k} V_2^{\tilde{\pi}_k}(s_2)\right)^2 \bigg| s_1\right) \tag{291}$$

$$= \mathbb{E}_{\tilde{\pi}_k}\left(\left(\sum_{t=2}^{H} r(s_t, \tilde{\pi}_k(s_t, t)) - V_2^{\tilde{\pi}_k}(s_2)\right)^2 \bigg| s_1\right) + \mathbb{E}_{\tilde{\pi}_k}\left(V_2^{\tilde{\pi}_k}(s_2) - \mathbb{E}_{\tilde{\pi}_k} V_2^{\tilde{\pi}_k}(s_2)\bigg| s_1\right)^2 \tag{292}$$

$$= \mathbb{E}_{\tilde{\pi}_k}\left(\mathbb{E}\left(\left(\sum_{t=2}^{H} r(s_t, \tilde{\pi}_k(s_t, t)) - V_2^{\tilde{\pi}_k}(s_2)\right)^2 \bigg| s_2\right) \bigg| s_1\right) + \operatorname*{Var}_{\tilde{\pi}_k}\left(V_2^{\tilde{\pi}_k}(s_2)\bigg| s_1\right) \tag{293}$$

$$= \mathbb{E}_{\tilde{\pi}_k}\left(\sum_{t=2}^{H} \mathbb{E}_{\tilde{\pi}_k}\left(\sum_{t=2}^{H} \operatorname*{Var}_{\tilde{\pi}_k}\left(V_{t+1}^{\tilde{\pi}_k}(s_{t+1})\bigg| s_t\right) \bigg| s_2\right) \bigg| s_1\right) + \operatorname*{Var}_{\tilde{\pi}_k}\left(V_2^{\tilde{\pi}_k}(s_2)\bigg| s_1\right) \tag{294}$$

$$= \mathbb{E}_{\tilde{\pi}_k}\left(\sum_{t=1}^{H} \operatorname*{Var}_{\tilde{\pi}_k}\left(V_{t+1}^{\tilde{\pi}_k}(s_{t+1})\bigg| s_t\right) \bigg| s_1\right) \tag{295}$$

$$\tag{296}$$

See for example (Azar et al., 2012) for a proof equivalent to this. $\qquad\square$

Lemma 16 (Upper Bound on Partial Loss). *Define the regret (with the starting states $\{s_{1k}\}_{k=1,\ldots,K}$) up to episode K as:*

$$\mathcal{R}(K) \stackrel{def}{=} \sum_{k=1}^{K}\left(V_1^{\pi^*} - V_{1k}^{\tilde{\pi}_k}\right)(s_{1k}). \tag{297}$$

Then it holds that:

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s,a} w_{tk}(s,a)p(\cdot \mid s,a)^\top \left(V_{t+1}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\right)^2 \leq H^2 \mathcal{R}(K). \tag{298}$$

Proof.

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s,a} w_{tk}(s,a)p(\cdot \mid s,a)^\top \left(V_{t+1}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\right)^2 \leq H\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s,a} w_{tk}(s,a)p(\cdot \mid s,a)^\top \left(V_{t+1}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\right) \tag{299}$$

$$= H\sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s'} w_{t+1,k}(s')\left(V_{t+1}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\right)(s') \tag{300}$$

$$\stackrel{(a)}{\leq} H\sum_{k=1}^{K}\sum_{t=1}^{H}\left(V_1^{\pi^*} - V_{1k}^{\tilde{\pi}_k}\right)(s_{1k}) \tag{301}$$

$$= H^2 \underbrace{\sum_{k=1}^{K}\left(V_1^{\pi^*} - V_{1k}^{\tilde{\pi}_k}\right)(s_{1k})}_{\mathcal{R}(K)} \tag{302}$$

$$= H^2 \mathcal{R}(K). \tag{303}$$

Here (a) follows from lemma 17.

$\square$

Lemma 17. *Let s_{1k} be the starting state in episode k, and $w_{t+1,k}(s') = \sum_a w_{t+1,k}(s', a)$. It holds that:*

$$\sum_{s'} w_{t+1,k}(s') \left(V_{t+1}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\right)(s') \leq \left(V_1^{\pi^*} - V_{1k}^{\tilde{\pi}_k}\right)(s_{1k}) \tag{304}$$

Proof. Define the policy μ as the policy that follows $\tilde{\pi}_k$ up to timestep t and $\pi^\star$ afterwards (until the end of the episode). We have that for any starting state s_{1k}:

$$V_{1k}^{\pi^*}(s_{1k}) \geq V_{1k}^{\mu}(s_{1k}) \geq V_{1k}^{\tilde{\pi}_k}(s_{1k}). \tag{305}$$

The rightmost inequality is true because μ follows $\pi^\star$ once it gets to timestep $\geq t+1$. This argument also justifies the step below:

$$\sum_{s'} w_{t+1,k}(s') \left(V_{t+1}^{\pi^*} - V_{t+1}^{\tilde{\pi}_k}\right)(s') = \sum_{s'} w_{t+1,k}(s') \left(V_{t+1}^{\mu} - V_{t+1}^{\tilde{\pi}_k}\right)(s') = V_{1k}^{\mu}(s_{1k}) - V_{1k}^{\tilde{\pi}_k}(s_{1k}) \leq V_1^{\star}(s_{1k}) - V_{1k}^{\tilde{\pi}_k}(s_{1k}). \tag{306}$$

$\square$