Beyond Value-Function Gaps: Improved Instance-Dependent Regret Bounds for Episodic Reinforcement Learning

Chris Dann
Google Research
chrisdann@google.com

Teodor V. Marinov*
Google Research
tvmarinov@google.com

Mehryar Mohri
Courant Institute and Google Research
mohri@google.com

Julian Zimmert
Google Research
zimmert@google.com

Abstract

We provide improved gap-dependent regret bounds for reinforcement learning in finite episodic Markov decision processes. Compared to prior work, our bounds depend on alternative definitions of gaps. These definitions are based on the insight that, in order to achieve a favorable regret, an algorithm does not need to learn how to behave optimally in states that are not reached by an optimal policy. We prove tighter upper regret bounds for optimistic algorithms and accompany them with new information-theoretic lower bounds for a large class of MDPs. Our results show that optimistic algorithms can not achieve the information-theoretic lower bounds even in deterministic MDPs unless there is a unique optimal policy.

1 Introduction

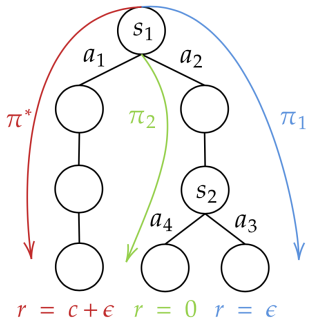
Reinforcement Learning (RL) is a general scenario where agents interact with the environment to achieve some goal. The environment and an agent's interactions are typically modeled as a Markov decision process (MDP) [29], which can represent a rich variety of tasks. But, for which MDPs can an agent or an RL algorithm succeed? This requires a theoretical analysis of the complexity of an MDP. This paper studies this question in the tabular episodic setting, where an agent interacts with the environment in episodes of fixed length H and where the size of the state and action space is finite (S and A respectively).

While the performance of RL algorithms in tabular Markov decision processes has been the subject of many studies in the past [e.g. 11, 22, 28, 7, 4, 20, 34, 6], the vast majority of existing analyses focuses on worst-case problem-independent regret bounds, which only take into account the size of the MDP, the horizon H and the number of episodes K.

Recently, however, some significant progress has been achieved towards deriving more optimistic (problem-dependent) guarantees. This includes more refined regret bounds for the tabular episodic setting that depend on structural properties of the specific MDP considered [30, 25, 21, 13, 17]. Motivated by instance-dependent analyses in multi-armed bandits [24], these analyses derive gap-dependent regret-bounds of the form $O\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{H\log(K)}{\mathrm{gap}(s,a)}\right)$, where the sum is over state-actions pairs (s,a) and where the gap notion is defined as the difference of the optimal value function V^* of the Bellman optimal policy π^* and the Q-function of π^* at a sub-optimal action: $\mathrm{gap}(s,a) =$

*Author was at Johns Hopkins University during part of this work.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).



	Value-function gap (prior)	Return gap (ours)
General Regret bounds	$O\Big(\sum_{s,a} \frac{H \log(K)}{\mathrm{gap}(s,a)}\Big)$ $\Omega\Big(\sum_{s,a:\ s\in\pi^*} \frac{\log(K)}{\mathrm{gap}(s,a)}\Big)$	$O\Big(\sum_{s,a} \frac{\log(K)}{\overline{\mathrm{gap}}(s,a)}\Big)$ $\Omega\Big(\sum_{s,a} \frac{\log(K)}{H\overline{\mathrm{gap}}(s,a)}\Big)$
Example on the left	$\mathrm{gap}(s_1,a_2) = c$ $\mathrm{gap}(s_2,a_4) = \epsilon$ $O\Big(\frac{SH\log(K)}{\epsilon}\Big)$	$\overline{\mathrm{gap}}(s_1,a_2) = c$ $\overline{\mathrm{gap}}(s_2,a_4) = \frac{c+\epsilon}{H} \approx c$ $O\Big(\frac{SH\log(K)}{c}\Big)$

Figure 1: Comparison of our contributions in MDPs with deterministic transitions. Bounds only include the main terms and all sums over (s,a) are understood to only include terms where the respective gap is nonzero. $\overline{\mathrm{gap}}$ is our alternative *return gap* definition introduced later (Definition 3.1).

$V^*(s) - Q^*(s,a)$. We will refer to this gap definition as *value-function gap* in the following. We note that a similar notion of gap has been used in the infinite horizon setting to achieve instance-dependent bounds [1, 31, 2, 12, 27], however, a strong assumption about irreducibility of the MDP is required.

While regret bounds based on these value function gaps generalize the bounds available in the multi-armed bandit setting, we argue that they have a major limitation. The bound at each state-action pair depends only on the gap at the pair and treats all state-action pairs equally, ignoring their topological ordering in the MDP. This can have a major impact on the derived bound. In this paper, we address this issue and formalize the following key observation about the difficulty of RL in an episodic MDP through improved instance-dependent regret bounds:

> *Learning a policy with optimal return does not require an RL agent to distinguish between actions with similar outcomes (small value-function gap) in states that can only be reached by taking highly suboptimal actions (large value-function gap).*

To illustrate this insight, consider autonomous driving, where each episode corresponds to driving from a start to a destination. If the RL agent decides to run a red light on a crowded intersection, then a car crash is inevitable. Even though the agent could slightly affect the severity of the car crash by steering, this effect is small and, hence, a good RL agent does not need to learn how to best steer after running a red light. Instead, it would only need a few samples to learn to obey the traffic light in the first place as the action of disregarding a red light has a very large value-function gap.

To understand how this observation translates into regret bounds, consider the toy example in Figure 1. This MDP has deterministic transitions and only terminal rewards with $c \gg \epsilon > 0$. There are two decision points, s_1 and s_2, with two actions each, and all other states have a single action. There are three policies which govern the regret bounds: π^* (red path) which takes action a_1 in state s_1; π_1 which takes action a_2 at s_1 and a_3 at s_2 (blue path); and π_2 which takes action a_2 at s_1 and a_4 at s_2 (green path). Since π^* follows the red path, it never reaches s_2 and achieves optimal return $c+\epsilon$, while π_1 and π_2 are both suboptimal with return ϵ and 0 respectively. Existing value-function gaps evaluate to $\mathrm{gap}(s_1,a_2) = c$ and $\mathrm{gap}(s_2,a_4) = \epsilon$ which yields a regret bound of order $H\log(K)(1/c + 1/\epsilon)$. The idea behind these bounds is to capture the necessary number of episodes to distinguish the value of the optimal policy π^* from the value of any other sub-optimal policy *on all states*. However, since π^* will never reach s_2 it is not necessary to distinguish it from any other policy at s_2. A good algorithm only needs to determine that a_2 is sub-optimal in s_1, which eliminates both π_1 and π_2 as optimal policies after only $\log(K)/c^2$ episodes. This suggests a regret of order $O(\log(K)/c)$. The bounds presented in this paper achieve this rate up to factors of H by replacing the gaps at every state-action pair with the average of all gaps along certain paths containing the state action pair. We call these averaged gaps *return gaps*. The return gap at (s,a) is denoted as $\overline{\mathrm{gap}}(s,a)$. Our new bounds replace $\mathrm{gap}(s_2,a_4) = \epsilon$ by $\overline{\mathrm{gap}}(s_2,a_4) \approx \frac{1}{2}\mathrm{gap}(s_1,a_2) + \frac{1}{2}\mathrm{gap}(s_2,a_4) = \Omega(c)$. Notice that ϵ and c can be selected arbitrarily in this example. In particular, if we take $c = 0.5$ and $\epsilon = 1/\sqrt{K}$ our bounds remain logarithmic $O(\log(K))$, while prior regret bounds scale as $\sqrt{K}$.

This work is motivated by the insight just discussed. First, we show that improved regret bounds are indeed possible by proving a tighter regret bound for StrongEuler, an existing algorithm

based on the optimism-in-the-face-of-uncertainty (OFU) principle [30]. Our regret bound is stated in terms of our new return gaps that capture the problem difficulty more accurately and avoid explicit dependencies on the smallest value function gap $\mathrm{gap}_{\min}$. Our technique applies to optimistic algorithms in general and as a by-product improves the dependency on episode length H of prior results. Second, we investigate the difficulty of RL in episodic MDPs from an information-theoretic perspective by deriving regret lower-bounds. We show that existing value-function gaps are indeed sufficient to capture difficulty of problems but only when each state is visited by an optimal policy with some probability. Finally, we prove a new lower bound when the transitions of the MDP are deterministic that depends only on the difference in return of the optimal policy and suboptimal policies, which is closely related to our notion of return gap.

2 Problem setting and notation

We consider reinforcement learning in episodic tabular MDPs with a fixed horizon. An MDP can be described as a tuple $(\mathcal{S}, \mathcal{A}, P, R, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are state- and action-space of size S and A respectively, P is the state transition distribution with $P(\cdot|s,a) \in \Delta^{S-1}$ the next state probability distribution, given that action a was taken in the current state s. R is the reward distribution defined over $\mathcal{S} \times \mathcal{A}$ and $r(s,a) = \mathbb{E}[R(s,a)] \in [0,1]$. Episodes admit a fixed length or *horizon* H.

We consider *layered* MDPs: each state $s \in \mathcal{S}$ belongs to a layer $\kappa(s) \in [H]$ and the only non-zero transitions are between states s, s' in consecutive layers, with $\kappa(s') = \kappa(s) + 1$. This common assumption [see e.g. 23] corresponds to MDPs with time-dependent transitions, as in [20, 7], but allows us to omit an explicit time-index in value-functions and policies. For ease of presentation, we assume there is a unique start state s_1 with $\kappa(s_1) = 1$ but our results can be generalized to multiple (possibly adversarial) start states. Similarly, for convenience, we assume that all states are reachable by some policy with non-zero probability, but not necessarily all policies or the same policy.

We denote by K the number of episodes during which the MDP is visited. Before each episode $k \in [K]$, the agent selects a deterministic policy $\pi_k \colon \mathcal{S} \to \mathcal{A}$ out of a set of all policies Π and π_k is then executed for all H time steps in episode k. For each policy π, we denote by $w^\pi(s,a) = \mathbb{P}(S_{\kappa(s)} = s, A_{\kappa(s)} = a \mid A_h = \pi(S_h)\ \forall h \in [H])$ and $w^\pi(s) = \sum_a w^\pi(s,a)$ probability of reaching state-action pair (s,a) and state s respectively when executing π. For convenience, $supp(\pi) = \{s \in \mathcal{S} \colon w^\pi(s) > 0\}$ is the set of states visited by π with non-zero probability. The Q- and value function of a policy π are

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{h=\kappa(s)}^{H} r(S_h, A_h) \,\middle|\, S_{\kappa(s)} = s, A_{\kappa(s)} = a\right], \quad \text{and} \quad V^\pi(s) = Q^\pi(s, \pi(s))$$

and the regret incurred by the agent is the sum of its regret over K episodes

$$\mathfrak{R}(K) = \sum_{k=1}^{K} v^* - v^{\pi_k} = \sum_{k=1}^{K} V^*(s_1) - V^{\pi_k}(s_1), \tag{1}$$

where $v^\pi = V^\pi(s_1)$ is the expected total sum of rewards or *return* of π and V^* is the optimal value function $V^*(s) = \max_{\pi \in \Pi} V^\pi(s)$. Finally, the set of optimal policies is denoted as $\Pi^* = \{\pi \in \Pi : V^\pi = V^*\}$. Note that we only call a policy optimal if it satisfies the Bellman equation in every state, as is common in literature, but there may be policies outside of Π^* that also achieve maximum return because they only take suboptimal actions outside of their support. The variance of the Q function at a state-action pair (s,a) of the optimal policy is $\mathcal{V}^*(s,a) = \mathbb{V}[R(s,a)] + \mathbb{V}_{s' \sim P(\cdot|s,a)}[V^*(s')]$, where $\mathbb{V}[X]$ denotes the variance of the r.v. X. The maximum variance over all state-action pairs is $\mathcal{V}^* = \max_{(s,a)} \mathcal{V}^*(s,a)$. Finally, our proofs will make use of the following clipping operator $\mathrm{clip}[a|b] = \chi(a \geq b)a$ that sets a to zero if it is smaller than b, where χ is the indicator function.

3 Novel upper bounds for optimistic algorithms

In this section, we present tighter regret upper-bounds for optimistic algorithms through a novel analysis technique. Our technique can be generally applied to model-based optimistic algorithms such as StrongEuler [30], UCBVI [3], ORLC [9] or Euler [34]. In the following, we will first

give a brief overview of this class of algorithms (see Appendix B for more details) and then state our main results for the StrongEuler algorithm [30]. We focus on this algorithm for concreteness and ease of comparison.

Optimistic algorithms maintain estimators of the Q-functions at every state-action pair such that there exists at least one policy π for which the estimator, $\bar{Q}^\pi$, overestimates the Q-function of the optimal policy, that is $\bar{Q}^\pi(s,a) \geq Q^*(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. During episode $k \in [K]$, the optimistic algorithm selects the policy π_k with highest optimistic value function $\bar{V}_k$. By definition, it holds that $\bar{V}_k(s) \geq V^*(s)$. The optimistic value and Q-functions are constructed through finite-sample estimators of the true rewards $r(s,a)$ and the transition kernel $P(\cdot|s,a)$ plus bias terms, similar to estimators for the UCB-I multi-armed bandit algorithm. Careful construction of these bias terms is crucial for deriving min-max optimal regret bounds in S, A and H [4]. Bias terms which yield the tightest known bounds come from concentration of martingales results such as Freedman's inequality [14] and empirical Bernstein's inequality for martingales [26].

The StrongEuler algorithm not only satisfies optimism, i.e., $\bar{V}_k \geq V^*$, but also a stronger version called *strong optimism*. To define strong optimism we need the notion of *surplus* which roughly measures the optimism at a fixed state-action pair. Formally the surplus at (s,a) during episode k is defined as

$$E_k(s,a) = \bar{Q}_k(s,a) - r(s,a) - \langle P(\cdot|s,a), \bar{V}_k \rangle \,. \tag{2}$$

We say that an algorithm is strongly optimistic if $E_k(s,a) \geq 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, k \in [K]$. Surpluses are also central to our new regret bounds and we will carefully discuss their use in Appendix F.

As hinted to in the introduction, the way prior regret bounds treat value-function gaps independently at each state-action pair can lead to excessively loose guarantees. Bounds that use value-function gaps [30, 25, 21] scale at least as

$$\sum_{s,a:\ \mathrm{gap}(s,a)>0} \frac{H \log(K)}{\mathrm{gap}(s,a)} + \sum_{s,a:\ \mathrm{gap}(s,a)=0} \frac{H \log(K)}{\mathrm{gap}_{\min}},$$

where state-action pairs with zero gap appear, with $\mathrm{gap}_{\min} = \min_{s,a:\ \mathrm{gap}(s,a)>0} \mathrm{gap}(s,a)$, the smallest positive gap. To illustrate where these bounds are loose, let us revisit the example in Figure 1. Here, these bounds evaluate to $\frac{H \log(K)}{c} + \frac{H \log(K)}{\epsilon} + \frac{SH \log(K)}{\epsilon}$, where the first two terms come from state-action pairs with positive value-function gaps and the last term comes from all the state-action pairs with zero gaps. There are several opportunities for improvement:

- **O.1 State-action pairs that can only be visited by taking optimal actions:** We should not pay the $1/\mathrm{gap}_{\min}$ factor for such (s,a) as there are no other suboptimal policies π to distinguish from π^* in such states.
- **O.2 State-action pairs that can only be visited by taking at least one suboptimal action:** We should not pay the $1/\mathrm{gap}(s_2,a_3)$ factor for state-action pair (s_2,a_3) and the $1/\mathrm{gap}_{\min}$ factor for (s_2,a_4) because no optimal policy visits s_2. Such state-action pairs should only be accounted for with the price to learn that a_2 is not optimal in state s_1. After all, learning to distinguish between π_1 and π_2 is unnecessary for optimal return.

Both opportunities suggest that the price $\frac{1}{\mathrm{gap}(s,a)}$ or $\frac{1}{\mathrm{gap}_{\min}}$ that each state-action pair (s,a) contributes to the regret bound can be reduced by taking into account the regret incurred by the time (s,a) is reached. Opportunity **O.1** postulates that if no regret can be incurred up to (and including) the time step (s,a) is reached, then this state-action pair should not appear in the regret bound. Similarly, if this regret is necessarily large, then the agent can learn this with few observations and stop reaching (s,a) earlier than $\mathrm{gap}(s,a)$ may suggest. Thus, as claimed in **O.2**, the contribution of (s,a) to the regret should be more limited in this case.

Since the total regret incurred during one episode by a policy π is simply the expected sum of value-function gaps visited (Lemma F.1 in the appendix),

$$v^* - v^\pi = \mathbb{E}_\pi \left[\sum_{h=1}^{H} \mathrm{gap}(S_h, A_h) \right], \tag{3}$$

we can measure the regret incurred up to reaching (S_t, A_t) by the sum of value function gaps $\sum_{h=1}^{t} \mathrm{gap}(S_h, A_h)$ up to this point t. We are interested in the regret incurred up to visiting a certain

state-action pair (s,a) which π may visit only with some probability. We therefore need to take the expectation of such gaps conditioned on the event that (s,a) is actually visited. We further condition on the event that this regret is nonzero, which is exactly the case when the agent encounters a positive value-function gap within the first $\kappa(s)$ time steps. We arrive at

$$\mathbb{E}_\pi\left[\sum_{h=1}^{\kappa(s)} \operatorname{gap}(S_h, A_h) \;\middle|\; S_{\kappa(s)} = s, A_{\kappa(s)} = a, B \le \kappa(s)\right],$$

where $B = \min\{h \in [H+1] \colon \operatorname{gap}(S_h, A_h) > 0\}$ is the first time a non-zero gap is visited. This quantity measures the regret incurred up to visiting (s,a) through suboptimal actions. If this quantity is large for all policies π, then a learner will stop visiting this state-action pair after few observations because it can rule out all actions that lead to (s,a) quickly. Conversely, if the event that we condition on has zero probability under any policy, then (s,a) can only be reached through optimal action choices (including a in s) and incurs no regret. This motivates our new definition of gaps that combines value function gaps with the regret incurred up to visiting the state-action pair:

Definition 3.1 (Return gap)**.** *For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ define $\mathcal{B}(s,a) \equiv \{B \le \kappa(s), S_{\kappa(s)} = s, A_{\kappa(s)} = a\}$, where B is the first time a non-zero gap is encountered. $\mathcal{B}(s,a)$ denotes the event that state-action pair (s,a) is visited and that a suboptimal action was played at any time up to visiting (s,a). We define the return gap as*

$$\overline{\operatorname{gap}}(s,a) \equiv \operatorname{gap}(s,a) \vee \min_{\substack{\pi \in \Pi \colon \\ \mathbb{P}_\pi(\mathcal{B}(s,a)) > 0}} \frac{1}{H}\, \mathbb{E}_\pi\left[\sum_{h=1}^{\kappa(s)} \operatorname{gap}(S_h, A_h) \;\middle|\; \mathcal{B}(s,a)\right]$$

if there is a policy $\pi \in \Pi$ with $\mathbb{P}_\pi(\mathcal{B}(s,a)) > 0$ and $\overline{\operatorname{gap}}(s,a) \equiv 0$ otherwise.

The additional $1/H$ factor in the second term is a required normalization suggesting that it is the average gap rather than their sum that matters. We emphasize that Definition 3.1 is independent of the choice of RL algorithm and in particular does not depend on the algorithm being optimistic. Thus, we expect our main ideas and techniques to be useful beyond the analysis of optimistic algorithms. Equipped with this definition, we are ready to state our main upper bound which pertains to the StrongEuler algorithm proposed by Simchowitz and Jamieson [30].

Theorem 3.2 (Main Result (Informal))**.** *The regret $\mathfrak{R}(K)$ of* StrongEuler *is bounded with high probability for all number of episodes K as*

$$\mathfrak{R}(K) \lessapprox \sum_{\substack{(s,a) \in \mathcal{S} \times \mathcal{A} \colon \\ \overline{\operatorname{gap}}(s,a) > 0}} \frac{\mathcal{V}^*(s,a)}{\overline{\operatorname{gap}}(s,a)} \log K.$$

In the above, we have restricted the bound to only those terms that have inverse polynomial dependence on the gaps.

Comparison with existing gap-dependent bounds. We now compare our bound to the existing gap-dependent bound for StrongEuler by Simchowitz and Jamieson [30, Corollary B.1]

$$\mathfrak{R}(K) \lessapprox \sum_{\substack{(s,a) \in \mathcal{S} \times \mathcal{A} \colon \\ \operatorname{gap}(s,a) > 0}} \frac{H\mathcal{V}^*(s,a)}{\operatorname{gap}(s,a)} \log K + \sum_{\substack{(s,a) \in \mathcal{S} \times \mathcal{A} \colon \\ \operatorname{gap}(s,a) = 0}} \frac{H\mathcal{V}^*}{\operatorname{gap}_{\min}} \log K. \tag{4}$$

We here focus only on terms that admit a dependency on K and an inverse-polynomial dependency on gaps as all other terms are comparable. Most notable is the absence of the second term of (4) in our bound in Theorem 3.2. Thus, while state-action pairs with $\overline{\operatorname{gap}}(s,a) = 0$ do not contribute to our regret bound, they appear with a $1/\operatorname{gap}_{\min}$ factor in existing bounds. Therefore, our bound addresses **O.1** because it does not pay for state-action pairs that can only be visited through optimal actions. Further, state-action pairs that do contribute to our bound satisfy $\frac{1}{\overline{\operatorname{gap}}(s,a)} \le \frac{1}{\operatorname{gap}(s,a)} \wedge \frac{H}{\operatorname{gap}_{\min}}$ and thus never contribute more than in the existing bound in (4). Therefore, our regret bound is never worse. In fact, it is significantly tighter when there are states that are only reachable by taking severely suboptimal actions, i.e., when the average value-function gaps are much larger than $\operatorname{gap}(s,a)$ or

$\mathrm{gap}_{\min}$. By our definition of return gaps, we only pay the inverse of these larger gaps instead of $\mathrm{gap}_{\min}$. Thus, our bound also addresses **O.2** and achieves the desired $\log(K)/c$ regret bound in the motivating example of Figure 1 as opposed to the $\log(K)/\epsilon$ bound of prior work.

One of the limitations of optimistic algorithms is their $S/\mathrm{gap}_{\min}$ dependence even when there is only one state with a gap of $\mathrm{gap}_{\min}$ [30]. We note that even though our bound in Theorem 3.2 improves on prior work, our result does not aim to address this limitation. Very recent concurrent work [32] proposed an action-elimination based algorithm that avoids the $S/\mathrm{gap}_{\min}$ issue of optimistic algorithm but their regret bounds still suffer the issues illustrated in Figure 1 (e.g. **O.2**). We therefore view our contributions as complementary. In fact, we believe our analysis techniques can be applied to their algorithm as well and result similar improvements as for the example in Figure 1.

Regret bound when transitions are deterministic. We now interpret Definition 3.1 for MDPs with deterministic transitions and derive an alternative form of our bound in this case. Let $\Pi_{s,a}$ be the set of all policies that visit (s,a) and have taken a suboptimal action up to that visit, that is,

$$\Pi_{s,a} \equiv \left\{\pi \in \Pi : s^\pi_{\kappa(s)} = s, a^\pi_{\kappa(s)} = a, \exists\, h \le \kappa(s), \mathrm{gap}(s^\pi_h, a^\pi_h) > 0\right\}.$$

where $(s^\pi_1, a^\pi_1, s^\pi_2, \dots, s^\pi_H, a^\pi_H)$ are the state-action pairs visited (deterministically) by π. Further, let $v^*_{s,a} = \max_{\pi\in\Pi_{s,a}} v^\pi$ be the best return of such policies. Definition 3.1 now evaluates to $\overline{\mathrm{gap}}(s,a) = \mathrm{gap}(s,a) \vee \frac{1}{H}(v^* - v^*_{s,a})$ and the bound in Theorem 3.2 can be written as

$$\mathfrak{R}(K) \lessapprox \sum_{s,a:\,\Pi_{s,a}\neq\varnothing} \frac{H\log(K)}{v^* - v^*_{s,a}}\,. \tag{5}$$

We show in Appendix F.7, that it is possible to further improve this bound when the optimal policy is unique by only summing over state-action pairs which are not visited by the optimal policy.

3.1 Regret analysis with improved clipping: from minimum gap to average gap

In this section, we present the main technical innovations of our tighter regret analysis. Our framework applies to *optimistic* algorithms that maintain a Q-function estimate, $\bar{Q}_k(s,a)$, which overestimates the optimal Q-function $Q^*(s,a)$ with high probability in all states s, actions a and episodes k. We first give an overview of gap-dependent analyses and then describe our approach.

Overview of gap-dependent analyses. A central quantity in regret analyses of optimistic algorithms are the surpluses $E_k(s,a)$, defined in (2), which, roughly speaking, quantify the local amount of optimism. Worst-case regret analyses bound the regret in episode k as $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_{\pi_k}(s,a)E_k(s,a)$, the expected surpluses under the optimistic policy π_k executed in that episode. Instead, gap-dependent analyses rely on a tighter version and bound the instantaneous regret by the *clipped surpluses* [e.g. Proposition 3.1 30]

$$V^*(s_1) - V^{\pi_k}(s_1) \le 2e\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[E_k(s,a)\;\middle|\;\frac{1}{4H}\,\mathrm{gap}(s,a) \vee \frac{\mathrm{gap}_{\min}}{2H}\right]. \tag{6}$$

Sharper clipping with general thresholds. Our main technical contribution for achieving a regret bound in terms of return gaps $\overline{\mathrm{gap}}(s,a)$ is the following improved surplus clipping bound:

Proposition 3.3 (Improved surplus clipping bound)**.** *Let the surpluses $E_k(s,a)$ be generated by an optimistic algorithm. Then the instantaneous regret of π_k is bounded as follows:*

$$V^*(s_1) - V^{\pi_k}(s_1) \le 4\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[E_k(s,a)\;\middle|\;\frac{1}{4}\,\mathrm{gap}(s,a) \vee \epsilon_k(s,a)\right],$$

where $\epsilon_k\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}^+_0$ is any clipping threshold function that satisfies

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\epsilon_k(S_h,A_h)\right] \le \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H}\mathrm{gap}(S_h,A_h)\right].$$

Compared to previous surplus clipping bounds in (6), there are several notable differences. First, instead of $\mathrm{gap}_{\min}/2H$, we can now pair $\mathrm{gap}(s,a)$ with more general clipping thresholds $\epsilon_k(s,a)$, as long as their expected sum over time steps after the first non-zero gap was encountered is at most half the expected sum of gaps. We will provide some intuition for this condition below. Note that $\epsilon_k(s,a) \equiv \frac{\mathrm{gap}_{\min}}{2H}$ satisfies the condition because the LHS is bounded between $\frac{\mathrm{gap}_{\min}}{2H}\mathbb{P}_{\pi_k}(B \le H)$ and $\mathrm{gap}_{\min}\mathbb{P}_{\pi_k}(B \le H)$, and there must be at least one positive gap in the sum $\sum_{h=1}^{H}\mathrm{gap}(S_h,A_h)$ on the RHS in event $\{B \le H\}$. Thus our bound recovers existing results. In addition, the first term in our clipping thresholds is $\frac{1}{4}\mathrm{gap}(s,a)$ instead of $\frac{1}{4H}\mathrm{gap}(s,a)$. Simchowitz and Jamieson [30] are able to remove this spurious H factor only if the problem instance happens to be a bandit instance and the algorithm satisfies a condition called *strong optimism* where surpluses have to be non-negative. Our analysis does not require such conditions and therefore generalizes these existing results.[2]

Choice of clipping thresholds for return gaps. The condition in Proposition 3.3 suggests that one can set $\epsilon_k(S_h,A_h)$ to be proportional to the average expected gap under policy π_k:

$$\epsilon_k(s,a) = \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H}\mathrm{gap}(S_h,A_h)\,\middle|\,\mathcal{B}(s,a)\right]. \tag{7}$$

if $\mathbb{P}_{\pi_k}(\mathcal{B}(s,a)) > 0$ and $\epsilon_k(s,a) = \infty$ otherwise. Lemma F.5 in Appendix F shows that this choice indeed satisfies the condition in Proposition 3.3. If we now take the minimum over all policies for π_k, then we can proceed with the standard analysis and derive our main result in Theorem 3.2. However, by avoiding the minimum over policies, we can derive a stronger policy-dependent regret bound which we discuss in the appendix.

4 Instance-dependent lower bounds

We here shed light on what properties on an episodic MDP determine the statistical difficulty of RL by deriving information-theoretic lower bounds on the asymptotic expected regret of any (good) algorithm. To that end, we first derive a general result that expresses a lower bound as the optimal value of a certain optimization problem and then derive closed-form lower-bounds from this optimization problem that depend on certain notions of gaps for two special cases of episodic MDPs.

Specifically, in those special cases, we assume that the rewards follow a Gaussian distribution with variance $1/2$. We further assume that the optimal value function is bounded in the same range as individual rewards, e.g. as $0 \le V^*(s) < 1$ for all $s \in \mathcal{S}$. This assumption is common in the literature [e.g. 23, 19, 8] and can be considered harder than a normalization of $V^*(s) \in [0,H]$ [18].

4.1 General instance-dependent lower bound as an optimization problem

The idea behind deriving instance-dependent lower bounds for the stochastic MAB problem [24, 5, 15] and infinite horizon MDPs [16, 27] are based on first assuming that the algorithm studied is *uniformly good*, that is, on any instance of the problem and for any $\alpha > 0$, the algorithm incurs regret at most $o(T^\alpha)$, and then argue that, to achieve that guarantee, the algorithm must select a certain policy or action at least some number of times as it would otherwise not be able to distinguish the current MDP from another MDP that requires a different optimal strategy.

Since comparison between different MDPs is central to lower-bound constructions, it is convenient to make the problem-instance explicit in the notation. To that end, let Θ be the problem class of possible MDPs and we use subscripts θ and λ for value functions, return, MDP parameters etc., to denote specific problem instances $\theta, \lambda \in \Theta$ of those quantities. Further, for a policy π and MDP θ, $\mathbb{P}_\theta^\pi$ denotes the law of one episode, i.e., the distribution of $(S_1, A_1, R_1, S_2, A_2, R_2, \ldots, S_{H+1})$. To state the general regret lower-bound we need to introduce the set of *confusing* MDPs. This set consists of all MDPs λ in which there is at least one optimal policy π such that $\pi \notin \Pi_\theta^*$, i.e., π is not optimal for the original MDP and no policy in Π_θ^* has been changed.

Definition 4.1. *For any problem instance $\theta \in \Theta$ we define the set of confusing MDPs $\Lambda(\theta)$ as*

$$\Lambda(\theta) := \{\lambda \in \Theta \colon \Pi_\lambda^* \setminus \Pi_\theta^* \neq \varnothing \text{ and } KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) = 0 \ \forall \pi \in \Pi_\theta^*\}.$$

[2] Our layered state space assumption changes H factors in lower-order terms of our final regret compared to Simchowitz and Jamieson [30]. However, Proposition 3.3 directly applies to their setting with no penalty in H.

We are now ready to state our general regret lower-bound for episodic MDPs:

Theorem 4.2 (General instance-dependent lower bound for episodic MDPs)**.** *Let ψ be a uniformly good RL algorithm for Θ, that is, for all problem instances $\theta \in \Theta$ and exponents $\alpha > 0$, the regret of ψ is bounded as $\mathbb{E}[\mathfrak{R}_\theta(K)] \leq o(K^\alpha)$, and assume that $v_\theta^* < H$. Then, for any $\theta \in \Theta$, the regret of ψ satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathfrak{R}_\theta(K)]}{\log K} \geq C(\theta),$$

where $C(\theta)$ is the optimal value of the following optimization problem

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize}} \quad & \sum_{\pi\in\Pi} \eta(\pi)\,(v_\theta^* - v_\theta^\pi) \\ \text{s. t.} \quad & \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1 \qquad \textit{for all} \;\; \lambda \in \Lambda(\theta). \end{aligned} \tag{8}$$

The optimization problem in Theorem 4.2 can be interpreted as follows. The variables $\eta(\pi)$ are the (expected) number of times the algorithm chooses to play policy π which makes the objective the total expected regret incurred by the algorithm. The constraints encode that any uniformly good algorithm needs to be able to distinguish the true instance θ from all confusing instances $\lambda \in \Lambda(\theta)$, because otherwise it would incur linear regret. To do so, a uniformly good algorithm needs to play policies π that induce different behavior in λ and θ which is precisely captured by the constraints $\sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1$.

Although Theorem 4.2 has the flavor of results in the bandit and RL literature, there are a few notable differences. Compared to lower-bounds in the infinite-horizon MDP setting [16, 31, 27], we for example do not assume that the Markov chain induced by an optimal policy π^* is irreducible. That irreducibility plays a key role in converting the semi-infinite linear program (8), which typically has uncountably many constraints, into a linear program with only $O(SA)$ constraints. While for infinite horizon MDPs, irreducibility is somewhat necessary to facilitate exploration, this is not the case for the finite horizon setting and in general we cannot obtain a convenient reduction of the set of constraints $\Lambda(\theta)$ (see also Appendix E.2).

4.2 Gap-dependent lower bound when optimal policies visit all states

To derive closed-form gap-dependent bounds from the general optimization problem (8), we need to identify a finite subset of confusing MDPs $\Lambda(\theta)$ that each require the RL agent to play a distinct set of policies that do not help to distinguish the other confusing MDPs. To do so, we restrict our attention to the special case of MDPs where every state is visited with non-zero probability by some optimal policy, similar to the irreducibility assumptions in the infinite-horizon setting [31, 27]. In this case, it is sufficient to raise the expected immediate reward of a suboptimal (s, a) by $\mathrm{gap}_\theta(s, a)$ in order to create a confusing MDP, as shown in Lemma 4.3:

Lemma 4.3. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1 and let $\theta \in \Theta$ be an MDP in this class. Then for any suboptimal state-action pair (s, a) with $\mathrm{gap}_\theta(s, a) > 0$ such that s is visited by some optimal policy with non-zero probability, there exists a confusing MDP $\lambda \in \Lambda(\theta)$ with*

- *λ and θ only differ in the immediate reward at (s, a)*
- *$KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \leq \mathrm{gap}_\theta(s, a)^2$ for all $\pi \in \Pi$.*

By relaxing the problem in (8) to only consider constraints from the confusing MDPs in Lemma 4.3 with $KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \leq \mathrm{gap}_\theta(s, a)^2$, for every (s, a), we can derive the following closed-form bound:

Theorem 4.4 (Gap-dependent lower bound when optimal policies visit all states)**.** *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1. Let $\theta \in \Theta$ be an instance where every state is visited by some optimal policy with non-zero probability. Then any uniformly good algorithm on Θ has expected regret on θ that satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathfrak{R}_\theta(K)]}{\log K} \geq \sum_{s,a\,:\;\mathrm{gap}_\theta(s,a)>0} \frac{1}{\mathrm{gap}_\theta(s, a)}.$$

Theorem 4.4 can be viewed as a generalization of Proposition 2.2 in Simchowitz and Jamieson [30], which gives a lower bound of order $\sum_{s,a:\ \mathrm{gap}_\theta(s,a)>0} \frac{H}{\mathrm{gap}_\theta(s,a)}$ for a certain set of MDPs.[3] While our lower bound is a factor of H worse, it is significantly more general and holds in any MDP where optimal policies visit all states and with appropriate normalization of the value function. Theorem 4.4 indicates that value-function gaps characterize the instance-optimal regret when optimal policies cover the entire state space.

4.3 Gap-dependent lower bound for deterministic-transition MDPs

We expect that optimal policies do not visit all states in most MDPs of practical interest (e.g. because certain parts of the state space can only be reached by making an egregious error). We therefore now consider the general case where $\bigcup_{\pi\in\Pi^*_\theta} supp(\pi) \subsetneq \mathcal{S}$ but restrict our attention to MDPs with deterministic transitions where we are able to give an intuitive closed-form lower bound. Note that deterministic transitions imply $\forall\pi, s, a:\ w^\pi(s,a) \in \{0,1\}$. Here, a confusing MDP can be created by simply raising the reward of any (s,a) by

$$v^*_\theta - \max_{\pi:\ w^\pi_\theta(s,a)>0} v^\pi_\theta\ , \tag{9}$$

the regret of the best policy that visits (s,a), as long as it is positive and (s,a) is not visited by any optimal policy. (9) is positive when no optimal policy visits (s,a) in which case suboptimal actions have to be taken to reach (s,a) and $\overline{\mathrm{gap}}_\theta(s,a) > 0$. Let $\pi^*_{(s,a)}$ be any maximizer in (9), which has to act optimally after visiting (s,a). From the regret decomposition in (3) and the fact that $\pi^*_{(s,a)}$ visits (s,a) with probability 1, it follows that $v^*_\theta - v^{\pi^*_{(s,a)}}_\theta \geq \mathrm{gap}_\theta(s,a)$. We further have $v^*_\theta - v^{\pi^*_{(s,a)}}_\theta \leq H\overline{\mathrm{gap}}_\theta(s,a)$. Equipped with the subset of confusing MDPs λ that each raise the reward of a single (s,a) as $r_\lambda(s,a) = r_\theta(s,a) + v^*_\theta - v^{\pi^*_{(s,a)}}_\theta$, we can derive the following gap-dependent lower bound:

Theorem 4.5. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1. Let $\theta \in \Theta$ be an instance with deterministic transitions. Then any uniformly good algorithm on Θ has expected regret on θ that satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathfrak{R}_\theta(K)]}{\log K} \geq \sum_{s,a\in\mathcal{Z}_\theta:\ \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{1}{H\cdot(v^*_\theta - v^{\pi^*_{(s,a)}}_\theta)} \geq \sum_{s,a\in\mathcal{Z}_\theta:\ \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{1}{H^2\cdot\overline{\mathrm{gap}}_\theta(s,a)},$$

where $\mathcal{Z}_\theta = \{(s,a)\in\mathcal{S}\times\mathcal{A}\colon \forall\pi^*\in\Pi^*_\theta\ \ w^{\pi^*}_\theta(s,a) = 0\}$ is the set of state-action pairs that no optimal policy in θ visits.

We now compare the above lower bound to the upper bound guaranteed by StrongEuler in (5). The comparison is only with respect to number of episodes and gaps[4]

$$\sum_{s,a\in\mathcal{Z}_\theta:\ \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{\log(K)}{H^2\overline{\mathrm{gap}}_\theta(s,a)} \leq \mathbb{E}_\theta[\mathfrak{R}(K)] \leq \sum_{s,a:\ \overline{\mathrm{gap}}_\theta(s,a)>0} \frac{\log(K)}{\overline{\mathrm{gap}}_\theta(s,a)}.$$

The difference between the two bounds, besides the extra H^2 factor, is the fact that (s,a) pairs that are visited by any optimal policy ($s,a\neq\mathcal{Z}_\theta$) do not appear in the lower-bound while the upper-bound pays for such pairs if they can also be visited after playing a suboptimal action. This could result in cases where the number of terms in the lower bound is $O(1)$ but the number of terms in the upper bound is $\Omega(SA)$ leading to a large discrepancy. In Theorem E.11 in the appendix we show that there exists an MDP instance on which it is information-theoretically possible to achieve $O(\log(K)/\epsilon)$ regret, however, any optimistic algorithm with confidence parameter δ will incur expected regret of at least $\Omega(S\log(1/\delta)/\epsilon)$. Theorem E.11 has two implications for optimistic algorithms in MDPs with deterministic transitions. Specifically, optimistic algorithms

- cannot be asymptotically optimal if confidence parameter δ is tuned to the time horizon K;
- cannot have an anytime bound that matches the information-theoretic lower bound.

[3] We translated their results to our setting where $V^* \leq 1$ which reduces the bound by a factor of H.

[4] We carry out the comparison in expectation, since our lower bounds do not apply with high probability.

5 Conclusion

In this work, we prove that optimistic algorithms such as StrongEuler, can suffer substantially less regret compared to what prior work had shown. We do this by introducing a new notion of gap, while greatly simplifying and generalizing existing analysis techniques. We further investigated the information-theoretic limits of learning episodic layered MDPs. We provide two new closed-form lower bounds in the special case where the MDP has either deterministic transitions or the optimal policy is supported on all states. These lower bounds suggest that our notion of gap better captures the difficulty of an episodic MDP for RL.

References

[1] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007.

[2] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2009.

[3] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1707–1714. Omnipress, 2012.

[4] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

[5] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2017.

[6] Christoph Dann. *Strategic Exploration in Reinforcement Learning - New Algorithms and Learning Guarantees*. PhD thesis, Carnegie Mellon University, 2019.

[7] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[8] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC reinforcement learning with rich observations. *arXiv preprint arXiv:1803.00606*, 2018.

[9] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. *International Conference on Machine Learning*, 2019.

[10] Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic Q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity. *arXiv preprint arXiv:2002.07125*, 2020.

[11] Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 88–97. ACM, 1994.

[12] Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. IEEE, 2010.

[13] Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

[14] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[15] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

[16] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[17] Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. *arXiv preprint arXiv:2011.11566*, 2020.

[18] Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398, 2018.

[19] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713, 2017.

[20] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

[21] Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. *arXiv preprint arXiv:2006.05606*, 2020.

[22] Sham Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

[23] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

[24] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[25] Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*, 2019.

[26] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[27] Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8874–8882, 2018.

[28] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[29] Martin Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1994.

[30] Max Simchowitz and Kevin Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. *arXiv preprint arXiv:1905.03814*, 2019.

[31] Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems*, pages 1505–1512, 2008.

[32] Haike Xu, Tengyu Ma, and Simon S Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

[33] Kunhe Yang, Lin F Yang, and Simon S Du. Q-learning with logarithmic regret. *arXiv preprint arXiv:2006.09118*, 2020.

[34] A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *https://arxiv.org/abs/1901.00210*, 2019.

[35] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pages 1583–1591, 2013.

Contents of main article and appendix

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3, Section 4 and corresponding sections in the appendix.
   (b) Did you describe the limitations of your work? [Yes] See lower bounds, discussion after Equation 5, Appendix E.3.3
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is theoretical and we do not see any potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix E for lower bounds and Appendix F for upper bounds.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Related work

We now discuss related work carefully. Instance dependent regret lower bounds for the MAB were first introduced in Lai and Robbins [24]. Later Graves and Lai [16] extend such instance dependent lower bounds to the setting of controlled Markov chains, while assuming infinite horizon and certain properties of the stationary distribution of each policy. Building on their work, more recently Combes et al. [5] establish instance dependent lower bounds for the Structured Stochastic Bandit problem. Very recently, in the stochastic MAB, Garivier et al. [15] generalize and simplify the techniques of Lai and Robbins [24] to completely characterize the behavior of uniformly good algorithms. The work of Ok et al. [27] builds on these ideas to provide an instance dependent lower bound for infinite horizon MDPs, again under assumptions of how the stationary distributions of each policy will behave and

irreducibility of the Markov chain. The idea behind deriving the above bounds is to use the uniform goodness of the studied algorithm to argue that the algorithm must select a certain policy or action at least a fixed number of times. This number is governed by a change of environment under which said policy/action is now the best overall. The reasoning now is that unless the algorithm is able to distinguish between these two environments it will have to incur linear regret asymptotically. Since the algorithm is uniformly good this can not happen.

For infinite horizon MDPs with additional assumptions the works of Auer and Ortner [1], Tewari and Bartlett [31], Auer et al. [2], Filippi et al. [12], Ok et al. [27] establish logarithmic in horizon regret bounds of the form $O(D^2S^2A\log(T)/\delta)$, where δ is a gap-like quantity and D is a diameter measure. We now discuss the works of [31, 27], which should give more intuition about how the infinite horizon setting differs from our setting. Both works consider the non-episodic problem and therefore make some assumptions about the MDP $\mathcal{M}$. The main assumption, which allows for computationally tractable algorithms is that of irreducibility. Formally both works require that under any policy the induced Markov chain is irreducible. Intuitively, the notion of irreducibility allows for coming up with exploration strategies, which are close to min-max optimal and are easy to compute. In [27] this is done by considering the same semi-infinite LP 8 as in our work. Unlike our work, however, assuming that the Markov chain induced by the optimal policy π^* is irreducible allows for a nice characterization of the set $\Lambda(\theta)$ of "confusing" environments. In particular the authors manage to show that at every state s it is enough to consider the change of environment which makes the reward of any action $a : (s, a) \notin \pi^*$ equal to the reward of $a' : (s, a') \in \pi^*$. Because of the irreducability assumption we know that the support of $P(\cdot|s, a)$ is the same as the support of $P(\cdot|s, a')$ and this implies that the above change of environment makes the policy π which plays (s, a) and then coincides with π^* optimal. Some more work shows that considering only such changes of environment is sufficient for an equivalent formulation to the LP8. Since this is an LP with at most $S \times A$ constraints it is solvable in polynomial time and hence a version of the algorithm in [5] results in asymptotic min-max rates for the problem. The exploration in [31] is also based on a similar LP, however, slightly more sophisticated.

Very recently there has been a renewed interest in proposing instance dependent regret bounds for finite horizon tabular MDPs [30, 25, 21]. The works of [30, 25] are based on the OFU principle and the proposed regret bounds scale as $O(\sum_{(s,a)\notin\pi^*} H\log(T)/\operatorname{gap}(s, a) + SH\log(T)/\operatorname{gap}_{\min})$, disregarding variance terms and terms depending only poli-logarithmically on the gaps. The setting in [25] also considers adversarial corruptions to the MDP, unknown to the algorithm, and their bound scales with the amount of corruption. Jin and Luo [21] derive similar upper bounds, however, the authors assume a known transition kernel and take the approach of modelling the problem as an instance of Online Linear Optimization, through using occupancy measures [35]. For the problem of Q-learning, Yang et al. [33], Du et al. [10], also propose algorithms with regret scaling as $O(SAH^6\log(T)/\operatorname{gap}_{\min})$. All of these bounds scale at least as $\Omega(SH\log(T)/\operatorname{gap}_{\min})$. Simchowitz and Jamieson [30] show an MDP instance on which no optimistic algorithm can hope to do better.

B Model-based optimistic algorithms for tabular RL

This section is a general discussion of optimistic algorithms for the tabular setting. Our regret upper bounds can be extended to other model based optimistic algorithms or in general any optimistic algorithm for which we can show a meaningful bound on the surpluses in terms of the number of times a state-action pair has been visited throughout the K episodes.

Pseudo-code for a generic algorithm can be found in Algorithm 1. The algorithm begins by initializing an empirical transition kernel $\hat{P} \in [0, 1]^{S\times A\times S}$, empirical reward kernel $\hat{r} \in [0, 1]^{S\times A}$, and bonuses $b \in [0, 1]^{S\times A}$. If we let $n_k(s, a)$ be the number of times we have observed state-action pair (s, a) up to episode k and $n_k(s', s, a)$ the number of times we have observed state s' after visiting (s, a) then one standard way to define the empirical kernels at episode k are as follows:

$$\hat{r}(s,a) = \frac{1}{n_k(s,a)}\sum_{j=1}^{k} R_j(s,a), \qquad \hat{P}(s'|s,a) = \begin{cases} \frac{n_k(s',s,a)}{n_k(s,a)} & \text{if } n_k(s,a) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $R_j(s, a)$ is a sample from $r(s, a)$ at episode j if (s, a) was visited and 0 otherwise. At every episode the generic algorithm constructs an policy π_k using the empirical model together with bonus

Algorithm 1 Generic Model-Based Optimistic Algorithm for Tabular RL

Require: Number of episodes K, horizon H, number of states S, number of actions A, probability of failure δ.

Ensure: A sequence of policies $(\pi_k)_{k=1}^K$ with low regret.

Initialize empirical transition kernel $\hat{P} \in [0,1]^{S\times A\times S}$, empirical reward kernel $\hat{r} \in [0,1]^{S\times A}$, bonuses $b \in [0,1]^{S\times A}$.
for $k \in [K]$ **do**
  $h = H$, $Q_k(s_{H+1}, a_{H+1}) = 0, \forall (s,a) \in \mathcal{S}\times\mathcal{A}$.
  while $h > 0$ **do**
    $Q_k(s,a) = \hat{r}(s,a) + \langle \hat{P}(\cdot|s,a), V_k\rangle + b(s,a)$.
    $\pi_k(s) := \text{argmax}_a\, Q_k(s,a)$.
    $h- = 1$
  Play π_k, collect observations from transition kernel P and reward kernel r and update $\hat{P}$, $\hat{r}$, b.

terms $b(s,a), \forall (s,a) \in \mathcal{S}\times\mathcal{A}$. Bonuses are constructed by using concentration of measure results relating $\hat{r}(s,a)$ to $r(s,a)$ and $\hat{P}(\cdot|s,a)$ to $P(\cdot|s,a)$. These bonuses usually scale inversely with the empirical visitations $n_k(s,a), \forall (s,a) \in \mathcal{S}\times\mathcal{A}$, as $O(1/\sqrt{n_k(s,a)})$. Further, depending on the type of concentration of measure result, the bonuses could either have a direct dependence on K, H, S, A, δ (following from Azuma-Hoeffding style concentration bounds) or replace H with the empirical estimator (following Freedman style concentration bounds). The bonus terms ensure that optimism is satisfied for π_k, that is $Q_k(s,a) \geq Q^{\pi_k}(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all episodes $k \in [K]$ with probability at least $1-\delta$. Algorithms such as UCBVI [4], EULER [34] and STRONGEULER [30] are all versions of Algorithm 1 with different instantiations of the bonus terms.

The greedy choice of π_k together with optimism also ensures that $V_k(s) \geq V^*(s)$. This has been key in prior work as it is what allows to bound the instantaneous regret by the sum of surpluses and ultimately relate the regret upper bound back to the bonus terms and the number of visits of each state-action pair respectively. Our regret upper bounds are also based on this decomposition and as such are not really tied to the STRONGEULER algorithm but would work with any model-based optimistic algorithm for the tabular setting. The main novelty in this work is a way to control the surpluses by clipping them to a gap-like quantity which better captures the sub-optimality of π_k compared to π^*. We remark that our analysis can be extended to any algorithm which follows Algorithm 1 so as long as we can control the bonus terms sufficiently well.

C Experimental results

In this section we present experiments based on the following deterministic LP which can be found in Figure 2. In short the MDP has only deterministic transitions and 3 layers. The starting state is

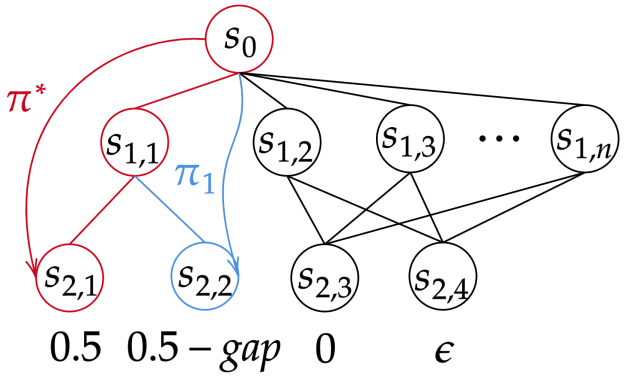
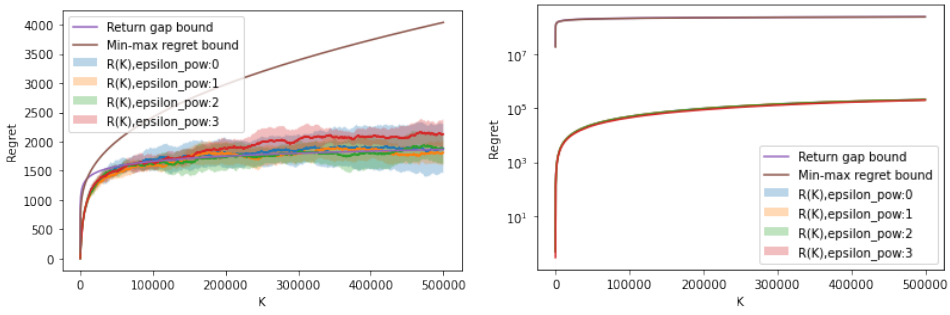


Figure 2: Deterministic MDP used in experiments

denoted by s_0 and the j-th state at layer i by $s_{i,j}$. There are $n+1$ possible actions at s_0, two possible actions at $s_{1,j}, \forall j \in [n+1]$, and a single possible action at $s_{2,j}, \forall j \in [4]$. The only non-negative rewards are at state-action pairs in the final layer. The unique optimal policy reaches state $s_{2,1}$ and has return equal to 0.5. We distinguish between two types of sub-optimal policies given by π_1 which visists $s_{1,1}$ and all other sub-optimal policies which visit $s_{1,j}, j \geq 2$. The return of policy π_1 determines the gap parameter in our experiments and the reward at state $s_{2,4}$ determines the ϵ parameter.

We run two sets of experiments using the UCBVI algorithm [4]. The chose this algorithm over Strong-EULER since UCBVI is slightly easier to implement and their differences are orthogonal to the issues studied here. The rewards in both experiments are Bernoulli with the respective mean provided below the state in Figure 2. In the first set of experiments we let the gap parameter to be equal to 0.5 and in the second set of experiments we let the gap parameter to be $\sqrt{\frac{S}{K}}$. We let $\epsilon = \frac{4^{\epsilon_{pow}}}{\sqrt{K}}$, where ϵ_{pow} takes integer values between 0 and $\lfloor 0.5 * \log_4(K) \rfloor$. We have two settings for n (respectively S) which are $n = 1$ and $n = 250$. In all experiments we have set $K = 500000$ and the topology of the MDP implies $H = 3$. Each experiment is repeated 5 times and we report the average regret of the algorithm, together with standard deviation of the regret. We note that in the first set of experiments we should observe regret which is close to $\Theta(\frac{SA\log(T)}{gap})$, this is because with our parameter choices the return gap is $gap/2$ for all settings of ϵ. In the second set of experiments we should observe regret which is close to $\Theta(\sqrt{SAK})$ as the min-max regret bounds dominate.



(a) $n = 1$ (b) $n = 250$

Figure 3: Large gap experiments

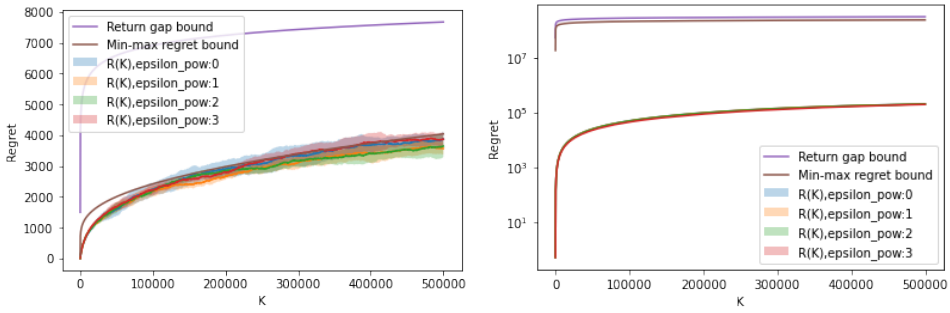
The first set of experiments can be found in Figure 3. We plot $S^2A + \frac{SA\log(T)}{gap}$ in purple and $S^2A + \sqrt{SAK}$ in brown for reference. We include the additive term of S^2A as this is what the theoretical regret bounds suggest. We see that for $n = 1$ our experiments almost perfectly match theory, including the observations made regarding Opportunity **O.1** and Opportunity **O.2**. In particular there is no obvious dependence on $1/\operatorname{gap}_{\min} = 1/\epsilon$, especially when $\epsilon = O(1/\sqrt{K})$, which in the plot is reflected by $\epsilon_{pow} = 0$. In the case for $n = 250$ the algorithm performs better than what our theory suggests. We expect that our bounds do not accurately capture the dependence on S and A, at least for deterministic transition MDPs. The second set of experiments can be found in Figure 4. Similar observations hold as in the large gap experiment.

D Additional Notation

We use the shorthand $(s,a) \in \pi$ to indicate that π admits a non-zero probability of visiting the state-action pair (s,a) and abusively use π as the set of such state-action pairs, when convenient.

E Proofs and extended discussion for regret lower-bounds

Let $N_{\psi,\pi}(k)$ be the random variable denoting the number of times policy π has been chosen by the strategy ψ. Let $N_{\psi,(s,a)}(k)$ be the number of times the state-action pair has been visited up to time k by the strategy ψ.



(a) $n = 1$ (b) $n = 250$

Figure 4: Small gap experiments

E.1 Lower bound as an optimization problem

We begin by formulating an LP characterizing the minimum regret incurred by any uniformly good algorithm ψ.

Theorem E.1. *Let ψ be a uniformly good RL algorithm for Θ, that is, for all problem instances $\theta \in \Theta$ and exponents $\alpha > 0$, the regret of ψ is bounded as $\mathbb{E}[\mathfrak{R}_\theta(K)] \leq o(K^\alpha)$. Then, for any $\theta \in \Theta$, the regret of ψ satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathfrak{R}_\theta(K)]}{\log K} \geq C(\theta),$$

where $C(\theta)$ is the optimal value of the following optimization problem

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize}} \quad & \sum_{\pi\in\Pi} \eta(\pi)\,(v_\theta^* - v_\theta^\pi) \\ \text{s. t.} \quad & \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1 \qquad \textit{for all} \;\; \lambda \in \Lambda(\theta), \end{aligned} \tag{11}$$

where $\Lambda'(\theta) = \{\lambda \in \Theta \colon \Pi_\lambda^* \cap \Pi_\theta^* = \varnothing, KL(\mathbb{P}_\theta^{\pi_\theta^*}, \mathbb{P}_\lambda^{\pi_\theta^*}) = 0\}$ are all environments that share no optimal policy with θ and do not change the rewards or transition kernel on π^*.

Proof. We can write the expected regret as $\mathbb{E}[\mathfrak{R}_\theta(K)] = \sum_{\pi\in\Pi} \mathbb{E}_\theta[N_{\psi,\pi}(K)](v_\theta^* - v_\theta^\pi)$. We will show that $\eta(\pi) = \mathbb{E}_\theta[N_{\psi,\pi}(K)]/\log K$ is feasible for the optimization problem in (8). This is sufficient to prove the theorem. To do so we follow the techniques of [15]. With slight abuse of notation, let $\mathbb{P}_\theta^{I_k}$ be the law of all trajectories up to episode k, where I_k is the history up to and including time k. Let Y_k be the random variable which is the value function of the policy, $\psi(I_k)$, selected at episode k. We have

$$\begin{aligned} KL(\mathbb{P}_\theta^{I_{k+1}}, \mathbb{P}_\lambda^{I_{k+1}}) &= KL(\mathbb{P}_\theta^{Y_{k+1},I_k}, \mathbb{P}_\lambda^{Y_{k+1},I_k}) \\ &= KL(\mathbb{P}_\theta^{I_k}, \mathbb{P}_\lambda^{I_k}) + \mathbb{E}\left[\mathbb{E}_{\mathbb{P}_\theta^{\psi(I_k)}}\left[\log \frac{\mathbb{P}_\theta^{\psi(I_k)}(Y_{k+1})}{\mathbb{P}_\lambda^{\psi(I_k)}(Y_{k+1})} \,\middle|\, I_k\right]\right] \\ &= KL(\mathbb{P}_\theta^{I_k}, \mathbb{P}_\lambda^{I_k}) + \mathbb{E}\left[\sum_{\pi\in\Pi} \chi(\psi(I_k) = \pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi)\right]. \end{aligned} \tag{12}$$

Iterating the argument we arrive at $\sum_{\pi\in\Pi} \mathbb{E}_\theta[N_{\psi,\pi}(K)] KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) = KL(\mathbb{P}_\theta^{I_K}, \mathbb{P}_\lambda^{I_K})$ where $\mathbb{E}_\theta$ denotes expectation in problem instance θ. Next one shows that for any measurable $Z \in [0,1]$, with respect to the natural sigma-algebra induced by I_K, it holds that $KL(\mathbb{P}_\theta^{I_K}, \mathbb{P}_\lambda^{I_K}) \geq kl(\mathbb{E}_\theta[Z], \mathbb{E}_\lambda[Z])$ where $kl(p,q) = p\log p/q + (1-p)\log(1-p)/(1-q)$ denotes the KL-divergence between two Bernoulli random variables p and q. This follows directly from Lemma 1 by Garivier et al. [15]. Finally we choose $Z = N_{\psi,\Pi_\lambda^*}(K)/K$ as the fraction of episodes where an optimal policy for λ

was played (here we use the short-hand notation $N_{\psi,\Pi^*_\lambda}(K) = \sum_{\pi\in\Pi^*_\lambda} N_{\psi,\pi}(K)$). Evaluating the kl-term we have

$$kl\left(\frac{\mathbb{E}_\theta[N_{\psi,\Pi^*_\lambda}(K)]}{K}, \frac{\mathbb{E}_\lambda[N_{\psi,\Pi^*_\lambda}(K)]}{K}\right) \geq \left(1 - \frac{\mathbb{E}_\theta[N_{\psi,\Pi^*_\lambda}(K)]}{K}\right)\log\frac{K}{K - \mathbb{E}_\lambda[N_{\psi,\Pi^*_\lambda}(K)]} - \log 2.$$

Since ψ is a uniformly good algorithm it follows that for any $\alpha > 0$, $K - \mathbb{E}_\lambda[N_{\psi,\Pi^*_\lambda}(K)] = o(K^\alpha)$. By assuming that $\Pi^*_\theta \cap \Pi^*_\lambda = \varnothing$, we get $\mathbb{E}_\theta[N_{\psi,\Pi^*_\lambda}(K)] = o(K)$. This implies that for K sufficiently large and all $1 \geq \alpha > 0$

$$kl\left(\frac{\mathbb{E}_\theta[N_{\psi,\Pi^*_\lambda}(K)]}{K}, \frac{\mathbb{E}_\lambda[N_{\psi,\Pi^*_\lambda}(K)]}{K}\right) \geq \log K - \log K^\alpha = (1-\alpha)\log K \xrightarrow{\alpha\to 0} \log K.$$

□

The set $\Lambda'(\theta)$ is uncountably infinite for any reasonable Θ we consider. What is worse the constraints of LP 8 will not form a closed set and thus the value of the optimization problem will actually be obtained on the boundary of the constraints. To deal with this issue it is possible to show the following.

Theorem 4.2 (General instance-dependent lower bound for episodic MDPs)**.** *Let ψ be a uniformly good RL algorithm for Θ, that is, for all problem instances $\theta \in \Theta$ and exponents $\alpha > 0$, the regret of ψ is bounded as $\mathbb{E}[\mathfrak{R}_\theta(K)] \leq o(K^\alpha)$, and assume that $v^*_\theta < H$. Then, for any $\theta \in \Theta$, the regret of ψ satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathfrak{R}_\theta(K)]}{\log K} \geq C(\theta),$$

where $C(\theta)$ is the optimal value of the following optimization problem

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize}} \quad & \sum_{\pi\in\Pi} \eta(\pi)\,(v^*_\theta - v^\pi_\theta) \\ \text{s. t.} \quad & \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) \geq 1 \qquad \textit{for all} \;\; \lambda \in \Lambda(\theta). \end{aligned} \tag{8}$$

Proof. For the rest of this proof we identify $\Lambda'(\theta) = \{\lambda \in \Theta : \Pi^*_\lambda \cap \Pi^*_\theta = \emptyset, KL(\mathbb{P}^{\pi^*_\theta}_\theta, \mathbb{P}^{\pi^*_\theta}_\lambda) = 0, \forall \pi^*_\theta \in \Pi^*_\theta\}$ as the set from Theorem E.1 and $\tilde{\Lambda}(\theta) = \{\lambda \in \Theta : v^{\pi^*_\lambda}_\lambda \geq v^{\pi^*_\theta}_\theta, \pi^*_\lambda \notin \Pi^*_\theta, KL(\mathbb{P}^{\pi^*_\theta}_\theta, \mathbb{P}^{\pi^*_\theta}_\lambda) = 0\}$. From the proof of Theorem E.1 it is clear that we can rewrite $\Lambda'(\theta)$ as the union $\bigcup_{\pi\in\Pi}\Lambda_\pi(\theta)$, where $\Lambda_\pi(\theta) = \{\lambda \in \Theta : KL(\mathbb{P}^{\pi^*_\theta}_\theta, \mathbb{P}^{\pi^*_\theta}_\lambda) = 0, v^{\pi^*_\lambda} > v^{\pi^*_\theta}_\theta, \pi^*_\lambda = \pi\}$ is the set of all environments which make π the optimal policy. This implies that we can equivalently write LP 8 as

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize}} \quad & \sum_{\pi\in\Pi} \eta(\pi)\,(v^*_\theta - v^\pi_\theta) \\ \text{s. t.} \quad & \inf_{\lambda\in\Lambda_{\pi'}(\theta)} \sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) \geq 1 \qquad \text{for all} \;\; \pi' \in \Pi. \end{aligned} \tag{13}$$

The above formulation now minimizes a linear function over a finite intersection of sets, however, these sets are still slightly inconvenient to work with. We are now going to try to make these sets more amenable to the proof techniques we would like to use for deriving specific lower bounds. We begin by noting that $\Lambda_\pi(\theta)$ is bounded in the following sense. We identify each λ with a vector in $[0,1]^{S^2A} \times [0,1]^{SA}$ where the first S^2A coordinates are transition probabilities and the last SA coordinates are the expected rewards. From now on we work with the natural topology on $[0,1]^{S^2A} \times [0,1]^{SA}$, induced by the ℓ_1 norm. Further, we claim that we can assume that $KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda)$ is a continuous function over $\Lambda_{\pi'}(\theta)$. The only points of discontinuity are at λ for which the support of the transition kernel induced by λ does not match the support of the transition kernel induced by θ. At such points the $KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) = \infty$. This implies that such λ does not achieve the infimum in the set of constraints so we can just restrict $\Lambda_{\pi'}(\theta)$ to contain only λ for which $KL(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_\lambda) < \infty$. With this restriction in hand the KL-divergence is continuous in λ.

Fix a π' and consider the set $\{\eta : \inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1\}$ corresponding to one of the constraints in LP 13. Denote $\tilde{\Lambda}_{\pi'}(\theta) = \{\lambda \in \Theta : KL(\mathbb{P}_\theta^{\pi_\theta^*}, \mathbb{P}_\lambda^{\pi_\theta^*}) = 0, v_\lambda^{\pi_\lambda^*} \geq v_\theta^{\pi_\theta^*}, \pi_\lambda^* \notin \Pi_\theta^*, \pi_\lambda^* = \pi'\}$. $\tilde{\Lambda}_{\pi'}(\theta)$ is closed as $KL(\mathbb{P}_\theta^{\pi_\theta^*}, \mathbb{P}_\lambda^{\pi_\theta^*})$ and $v_\lambda^{\pi_\lambda^*} - v_\theta^{\pi_\theta^*}$ are both continuous in λ. To see the statement for $v_\lambda^{\pi_\lambda^*}$, notice that this is the maximum over the continuous functions v_λ^π over $\pi \in \Pi$. Take any $\eta \in \tilde{\Lambda}_{\pi'}(\theta)$ and let $\{\lambda_j\}_{j=1}^\infty, \lambda_j \in \Lambda_{\pi'}(\theta)$ be a sequence of environments such that $\sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_{\lambda_j}^\pi) \geq 1 + 2^{-j}$. If there is no convergent subsequence of $\{\lambda_j\}_{j=1}^\infty$ in $\Lambda_{\pi'}(\theta)$ we claim it is because of the constraint $v_\lambda^{\pi_\lambda^*} > v_\theta^{\pi_\theta^*}$. Take the limit λ of any convergent subsequence of $\{\lambda_j\}_{j=1}^\infty$ in the closure of $\Lambda_{\pi'}(\theta)$. Then by continuity of the divergence we have $0 = \lim_{j \to \infty} KL(\mathbb{P}_\theta^{\pi_\theta^*}, \mathbb{P}_{\lambda_j}^{\pi_\theta^*}) = KL(\mathbb{P}_\theta^{\pi_\theta^*}, \mathbb{P}_\lambda^{\pi_\theta^*})$, thus it must be the case that $v_\lambda^{\pi_\lambda^*} \leq v_\theta^{\pi_\theta^*}$. This shows that $\tilde{\Lambda}_{\pi'}(\theta)$ is a subset of the closure of $\Lambda_{\pi'}(\theta)$ which implies it is the closure of $\Lambda_{\pi'}(\theta)$, i.e., $\bar{\Lambda}_{\pi'}(\theta) = \tilde{\Lambda}_{\pi'}(\theta)$.

Next, take $\eta \in \{\eta : \min_{\lambda \in \bar{\Lambda}_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1\}$ and let $\lambda_{\pi',\eta}$ be the environment on which the minimum is achieved. Such $\lambda_{\pi',\eta}$ exists because we just showed that $\bar{\Lambda}_{\pi'}(\theta)$ is closed and bounded and hence compact and the sum consists of a finite number of continuous functions. If $\lambda_{\pi',\eta} \in \Lambda_{\pi'}(\theta)$ then $\eta \in \{\eta : \inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1\}$. If $\lambda_{\pi',\eta} \notin \Lambda_{\pi'}(\theta)$ then $\lambda_{\pi',\eta}$ must be a limit point of $\Lambda_{\pi'}(\theta)$. By definition we can construct a convergent sequence of $\{\lambda_j\}_{j=1}^\infty, \lambda_j \in \Lambda_{\pi'}(\theta)$ to $\lambda_{\pi',\eta}$ such that $\sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_{\lambda_j}^\pi) \geq 1$. This implies $\sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_{\lambda_j}^\pi) \geq \inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi)$. Using the continuity of the KL term and taking limits, the above implies that the minimum upper bounds the infimum. Since we argued that $\Lambda_{\pi'}(\theta)$ is bounded and $\sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_{\lambda_j}^\pi)$ is also bounded from below this implies $\bar{\Lambda}_{\pi'}(\theta)$ contains the infimum $\inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi)$. This implies $\inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq \min_{\lambda \in \bar{\Lambda}_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi)$, and so the infimum over $\Lambda_\pi(\theta)$ equals the minimum over $\bar{\Lambda}_\pi(\theta)$. Which finally implies that $\eta \in \{\eta : \inf_{\lambda \in \Lambda_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1\}$. This shows that LP 13 is equivalent to

$$\begin{aligned} \underset{\eta(\pi) \geq 0}{\text{minimize}} \quad & \sum_{\pi \in \Pi} \eta(\pi) \left(v_\theta^* - v_\theta^\pi\right) \\ \text{s. t.} \quad & \min_{\lambda \in \bar{\Lambda}_{\pi'}(\theta)} \sum_{\pi \in \Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \geq 1 \qquad \text{for all } \pi' \in \Pi, \end{aligned}$$

or equivalently that we can consider the closure of $\Lambda(\theta)$ in LP 8, $\bar{\Lambda}(\theta) = \{\lambda \in \Theta \colon v_\lambda^{\pi_\lambda^*} \geq v_\theta^{\pi_\theta^*}, \pi_\lambda^* \notin \Pi_\theta^*, KL(\mathbb{P}_\theta^{\pi_\theta^*}, \mathbb{P}_\lambda^{\pi_\theta^*}) = 0\}$ i.e. the set of environments which makes any π optimal without changing the environment on state-action pairs in π_θ^*. □

E.2 Lower bounds for full support optimal policy

Lemma 4.3. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1 and let $\theta \in \Theta$ be an MDP in this class. Then for any suboptimal state-action pair (s, a) with $\text{gap}_\theta(s, a) > 0$ such that s is visited by some optimal policy with non-zero probability, there exists a confusing MDP $\lambda \in \Lambda(\theta)$ with*

- *λ and θ only differ in the immediate reward at (s, a)*
- *$KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \leq \text{gap}_\theta(s, a)^2$ for all $\pi \in \Pi$.*

Proof. Let λ be the environment that is identical to θ except for the immediate reward for state-action pair for (s, a). Specifically, let $R_\lambda(s, a)$ so that $r_\lambda(s, a) = r_\theta(s, a) + \Delta$ with $\Delta = \text{gap}_\theta(s, a)$. Since we assume that rewards are Gaussian, it follows that

$$\begin{aligned} KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) &= w_\lambda^\pi(s, a) KL(R_\theta(s, a), R_\lambda(s, a)) \leq KL(R_\theta(s, a), R_\lambda(s, a)) \\ &\leq \text{gap}_\theta(s, a)^2 \end{aligned}$$

for any policy $\pi \in \Pi$. We now show that the optimal value function (and thus return) of λ is uniformly upper-bounded by the optimal value function of θ. To that end, consider their difference in any state

s', which we will upper-bound by their difference in s as

$$\begin{aligned} V^*_\lambda(s') - V^*_\theta(s') &\leq \chi(\kappa(s) > \kappa(s'))\mathbb{P}^{\pi^*_\lambda}_\theta(s_{\kappa(s)} = s | s_{\kappa(s')} = s')[V^*_\lambda(s) - V^*_\theta(s)] \\ &\leq V^*_\lambda(s) - V^*_\theta(s). \end{aligned}$$

Further, the difference in s is exactly

$$\begin{aligned} V^*_\lambda(s) - V^*_\theta(s) &= r_\lambda(s,a) + \langle P_\theta(\cdot|s,a), V^*_\theta \rangle - V^*_\theta(s) \\ &= r_\theta(s,a) + \langle P_\theta(\cdot|s,a), V^*_\theta \rangle + \mathrm{gap}_\theta(s,a) - V^*_\theta(s) = 0. \end{aligned}$$

Hence, $V^*_\lambda = V^*_\theta \leq 1$ and thus $\lambda \in \Theta$. We will now show that there is a policy that is optimal in λ but not in θ. Let $\pi^* \in \Pi^*_\theta$ be any optimal policy for θ that has non-zero probability of visiting s and consider the policy

$$\tilde{\pi}(\tilde{s}) = \begin{cases} \pi^*(\tilde{s}) & \text{if } s \neq \tilde{s} \\ a & \text{if } s = \tilde{s} \end{cases}$$

that matches π^* on all states except s. We will now show that $\tilde{\pi}$ achieves the same return as π^* in λ. Consider their difference

$$\begin{aligned} v^{\tilde{\pi}}_\lambda - v^{\pi^*}_\lambda &\overset{(i)}{=} w^{\tilde{\pi}}_\lambda(s, \tilde{\pi}(s))[r_\lambda(s, \tilde{\pi}(s)) + \langle P_\lambda(\cdot|s, \tilde{\pi}(s)), V^{\tilde{\pi}}_\lambda \rangle] \\ &\quad - w^{\pi^*}_\lambda(s, \pi^*(s))[r_\lambda(s, \pi^*(s)) + \langle P_\lambda(\cdot|s, \pi^*(s)), V^{\pi^*}_\lambda \rangle] \\ &\overset{(ii)}{=} w^{\pi^*}_\lambda(s, \pi^*(s))[r_\lambda(s, \tilde{\pi}(s)) - r_\lambda(s, \pi^*(s)) + \langle P_\lambda(\cdot|s, \tilde{\pi}(s)) - P_\lambda(\cdot|s, \pi^*(s)), V^{\pi^*}_\lambda \rangle] \\ &\overset{(iii)}{=} w^{\pi^*}_\theta(s, \pi^*(s))[\Delta + r_\theta(s, \tilde{\pi}(s)) - r_\theta(s, \pi^*(s)) + \langle P_\theta(\cdot|s, \tilde{\pi}(s)) - P_\theta(\cdot|s, \pi^*(s)), V^*_\theta \rangle] \\ &\overset{(iv)}{=} w^{\pi^*}_\theta(s, \pi^*(s))[\Delta - \mathrm{gap}_\theta(s, \tilde{\pi}(s))] \end{aligned}$$

where (i) and (ii) follow from the fact that $\tilde{\pi}$ and π^* only differ on s and hence, their probability at arriving at s and their value for any successor state of s is identical. Step (iii) follows from the fact that λ and θ only differ on (s,a) which is not visited by π^*. Finally, step (iv) applies the definition of optimal value functions and value-function gaps. Since $\Delta = \mathrm{gap}_\theta(s, \tilde{\pi}(s))$, it follows that $v^{\tilde{\pi}}_\lambda = v^{\pi^*}_\lambda = v^{\pi^*}_\theta = v^*_\theta$. As we have seen above, the optimal value function (and return) is identical in θ and λ and, hence, $\tilde{\pi}$ is optimal in λ.

Note that the we can apply the chain of equalities above in the same manner to $v^{\tilde{\pi}}_\theta - v^{\pi^*}_\theta$ if we consider $\Delta = 0$. This yields

$$v^{\tilde{\pi}}_\theta - v^{\pi^*}_\theta = -w^{\pi^*}_\theta(s, \pi^*(s)) \, \mathrm{gap}_\theta(s,a) < 0$$

because $w^{\pi^*}_\theta(s, \pi^*(s)) > 0$ and $\mathrm{gap}_\theta(s,a) < 0$ by assumption. Hence $\tilde{\pi}$ is not optimal in θ, which completes the proof. □

Lemma E.2 (Optimization problem over $\mathcal{S} \times \mathcal{A}$ instead of Π)**.** *Let optimal value $C(\theta)$ of the optimization problem (8) in Theorem 4.2 is lower-bound by the optimal value of the problem*

$$\begin{aligned} \underset{\eta(s,a) \geq 0}{\text{minimize}} \quad & \sum_{s,a} \eta(s,a) \, \mathrm{gap}_\theta(s,a) \\ \textit{s.t.} \quad & \sum_{s,a} \eta(s,a) KL(R_\theta(s,a), R_\lambda(s,a)) \\ & + \sum_{s,a} \eta(s,a) KL(P_\theta(\cdot|s,a), P_\lambda(\cdot|s,a)) \geq 1 \qquad \textit{for all} \;\; \lambda \in \Lambda(\theta) \end{aligned} \tag{14}$$

Proof. First, we rewrite the objective of (8) as

$$\sum_{\pi \in \Pi} \eta(\pi)(v^*_\theta - v^\pi_\theta) \overset{(i)}{=} \sum_{\pi \in \Pi} \eta(\pi) \sum_{s,a} w^\pi_\theta(s,a) \, \mathrm{gap}_\theta(s,a) = \sum_{s,a} \left(\sum_{\pi \in \Pi} \eta(\pi) w^\pi_\theta(s,a) \right) \mathrm{gap}_\theta(s,a)$$

where step (i) applies Lemma F.1 proved in Appendix F. Here, $w_\theta^\pi(s,a)$ is the probability of reaching s and taking a when playing policy π in MDP θ. Similarly, the LHS of the constraints of (8) can be decomposed as

$$\begin{aligned}
&\sum_{\pi\in\Pi} \eta(\pi) KL(\mathbb{P}_\theta^\pi, \mathbb{P}_\lambda^\pi) \\
&= \sum_{\pi\in\Pi} \eta(\pi) \sum_{s,a} w_\theta^\pi(s,a) \left(KL(R_\theta(s,a), R_\lambda(s,a)) + KL(P_\theta(\cdot|s,a), P_\lambda(\cdot|s,a))\right) \\
&= \sum_{s,a} \left[\sum_{\pi\in\Pi} \eta(\pi) w_\theta^\pi(s,a)\right] \left(KL(R_\theta(s,a), R_\lambda(s,a)) + KL(P_\theta(\cdot|s,a), P_\lambda(\cdot|s,a))\right)
\end{aligned}$$

where the first equality follows from writing out the definition of the KL divergence. Let now $\eta(\pi)$ be a feasible solution to the original problem (8). Then the two equalities we just proved show that $\eta(s,a) = \sum_{\pi\in\Pi} \eta(\pi) w_\theta^\pi(s,a)$ is a feasible solution for the problem in (14) with the same value. Hence, since (14) is a minimization problem, its optimal value cannot be larger than $C(\theta)$, the optimal value of (8). □

Theorem 4.4 (Gap-dependent lower bound when optimal policies visit all states)**.** *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1. Let $\theta \in \Theta$ be an instance where every state is visited by some optimal policy with non-zero probability. Then any uniformly good algorithm on Θ has expected regret on θ that satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathfrak{R}_\theta(K)]}{\log K} \geq \sum_{s,a:\ \mathrm{gap}_\theta(s,a)>0} \frac{1}{\mathrm{gap}_\theta(s,a)}.$$

Proof. Let $\bar{\Lambda}(\theta)$ be a set of all confusing MDPs from Lemma 4.3, that is, for every suboptimal (s,a), $\bar{\Lambda}(\theta)$ contains exactly one confusing MDP that differs with θ only in the immediate reward at (s,a). Consider now the relaxation of Theorem 4.2 from Lemma E.2 and further relax it by reducing the set of constraints induced by $\Lambda(\theta)$ to only the set of constraints induced by $\bar{\Lambda}(\theta)$:

$$\begin{aligned}
\underset{\eta(s,a)\geq 0}{\text{minimize}} \quad & \sum_{s,a} \eta(s,a)\,\mathrm{gap}_\theta(s,a) \\
\text{s.t.} \quad & \sum_{s,a} \eta(s,a) KL(R_\theta(s,a), R_\lambda(s,a)) \geq 1 \qquad \text{for all } \ \lambda \in \bar{\Lambda}(\theta)
\end{aligned}$$

Since all confusing MDPs only differ in rewards, we dropped the KL-term for the transition probabilities. We can simplify the constraints by noting that for each λ, only one KL-term is non-zero and it has value $\mathrm{gap}_\theta(s,a)^2$. Hence, we can write the problem above equivalently as

$$\begin{aligned}
\underset{\eta(s,a)\geq 0}{\text{minimize}} \quad & \sum_{s,a} \eta(s,a)\,\mathrm{gap}_\theta(s,a) \\
\text{s.t.} \quad & \eta(s,a)\,\mathrm{gap}_\theta(s,a)^2 \geq 1 \qquad \text{for all } \ (s,a) \in \mathcal{S}\times\mathcal{A} \text{ with } \ \mathrm{gap}_\theta(s,a) > 0
\end{aligned}$$

Rearranging the constraint as $\eta(s,a) \geq 1/\,\mathrm{gap}_\theta(s,a)^2$, we see that the value is lower-bounded by

$$\sum_{s,a} \eta(s,a)\,\mathrm{gap}_\theta(s,a) \geq \sum_{s,a:\ \mathrm{gap}_\theta(s,a)>0} \eta(s,a)\,\mathrm{gap}_\theta(s,a) \geq \sum_{s,a:\ \mathrm{gap}_\theta(s,a)>0} \frac{1}{\mathrm{gap}_\theta(s,a)},$$

which completes the proof. □

We note that because the relaxation in Lemma E.2 essentially allows the algorithm to choose which state-action pairs to play instead of just policies, the final lower bound in Theorem 4.4 may be loose, especially in factors of H. However, it is unlikely that the $\mathrm{gap}_{\min}$ term arising in the upper bound of Simchowitz and Jamieson [30] can be recovered. We conjecture that such a term can be avoided by algorithms, which do not construct optimistic estimators for the Q-function at each state-action pair but rather just work with a class of policies and construct only optimistic estimators of the return.

E.3 Lower bounds for deterministic MDPs

We will show that we can derive lower bounds in two cases:

1. We show that if the graph induced by the MDP is a tree, then we can formulate a finite LP which has value at most a polynomial factor of H away from the value of LP 8.
2. We show that if we assume that the value function for any policy is at most 1 and the rewards of each state-action pair are at most 1, then we can derive a closed form lower bound. This lower bound is also at most a polynomial factor of H away from the solution to LP 8.

We begin by stating a helpful lemma, which upper and lower bounds the KL-divergence between two environments on any policy π. Since we consider Gaussian rewards with $\sigma = 1/\sqrt{2}$ it holds that $KL(R_\theta(s,a), R_\lambda(s,a)) = (r_\theta(s,a) - r_\lambda(s,a))^2$. Further for any π and λ it holds that $KL(\theta(\pi), \lambda(\pi)) = \sum_{(s,a)\in\pi} KL(R_\theta(s,a), R_\lambda(s,a)) = \sum_{(s,a)\in\pi}(r_\theta(s,a) - r_\lambda(s,a))^2$. We can now show the following lower bound on $KL(\theta(\pi), \lambda(\pi))$.

Lemma E.3. *Fix π and suppose λ is such that $\pi^*_\lambda = \pi$. Then $(v^* - v^\pi)^2 \geq KL(\theta(\pi), \lambda(\pi)) \geq \frac{(v^* - v^\pi)^2}{H}$.*

Proof. The second inequality follows from the fact that the optimization problem

$$\begin{aligned} \underset{\theta,\lambda\in\Lambda(\theta):\pi^*_\lambda=\pi}{\text{minimize}} \quad & \sum_{(s,a)\in\pi} (r_\theta(s,a) - r_\lambda(s,a))^2 \\ \text{s. t.} \quad & \sum_{(s,a)\in\pi} r_\lambda(s,a) - r_\theta(s,a) \geq v^* - v^\pi, \end{aligned}$$

admits a solution at θ, λ for which $r_\lambda(s,a) - r_\theta(s,a) = \frac{v^* - v^\pi}{H}, \forall (s,a) \in \pi$. The first inequality follows from considering the optimization problem

$$\begin{aligned} \underset{\theta,\lambda\in\Lambda(\theta):\pi^*_\lambda=\pi}{\text{maximize}} \quad & \sum_{(s,a)\in\pi} (r_\theta(s,a) - r_\lambda(s,a))^2 \\ \text{s.t.} \quad & \sum_{(s,a)\in\pi} r_\lambda(s,a) - r_\theta(s,a) \geq v^* - v^\pi, \end{aligned}$$

and the fact that it admits a solution at θ, λ for which there exists a single state-action pair $(s,a) \in \pi$ such that $r_\theta(s,a) - r_\lambda(s,a) = v^* - v^\pi$ and for all other (s,a) it holds that $r_\lambda(s,a) = r_\theta(s,a)$. □

Using the above Lemma E.3 we now show that we can restrict our attention only to environments $\lambda \in \Lambda(\theta)$ which make one of $\pi^*_{(s,a)}$ optimal and derive an upper bound on $C(\theta)$ which we will try to match, up to factors of H, later. Define the set $\tilde{\Lambda}(\theta) = \{\lambda \in \Lambda(\theta) : \exists (s,a) \in \mathcal{S}\times\mathcal{A}, \pi^*_\lambda = \pi^*_{(s,a)}\}$ and $\Pi^* = \{\pi \in \Pi, \pi \neq \pi^*_\theta : \exists (s,a) \in \mathcal{S}\times\mathcal{A}, \pi = \pi^*_{(s,a)}\}$. We have

Lemma E.4. *Let $\tilde{C}(\theta)$ be the value of the optimization problem*

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize}} \quad & \sum_{\pi\in\Pi^*} \eta(\pi)(v^* - v^\pi) \\ \text{s. t.} \quad & \sum_{\pi\in\Pi^*} \eta(\pi) KL(\theta(\pi), \lambda(\pi)) \geq 1, \forall \lambda \in \tilde{\Lambda}(\theta). \end{aligned} \tag{15}$$

Then $\sum_{\pi\in\Pi^*} \frac{H}{v^* - v^\pi} \geq C(\theta) \geq \frac{\tilde{C}(\theta)}{H}$.

Proof. We begin by showing $C(\theta) \geq \frac{\tilde{C}(\theta)}{H}$ holds. Fix a $\pi \notin \Pi^*$ s.t. the solution of LP 8 implies $\eta(\pi) > 0$. Let $\lambda \in \tilde{\Lambda}(\theta)$ be a change of environment for which $KL(\theta(\pi), \lambda(\pi)) > 0$. We can now shift all of the weight of $\eta(\pi)$ to $\eta(\pi^*_\lambda)$ while still preserving the validity of the constraint. Further doing so to all $\pi^*_{(s,a)}$ for which $\pi^*_{(s,a)} \cap \pi \neq \emptyset$ will not increase the objective by more than a factor of H as $v^* - v^\pi \geq \frac{1}{H}\sum_{(s,a)\in\pi} v^* - v^{\pi^*_{(s,a)}}$. Thus, we have converted the solution to LP 8 to a feasible solution to LP 15 which is only a factor of H larger.

Next we show that $\sum_{\pi\in\Pi^*}\frac{H}{v^*-v^\pi}\geq C(\theta)$. Set $\eta(\pi)=0,\forall\pi\in\Pi\setminus\Pi^*$ and set $\eta(\pi)=\frac{H}{(v^*-v^\pi)^2},\forall\pi\in\Pi^*$. If π is s.t. $\eta(\pi)>0$ then for any λ which makes π optimal it holds that

$$\begin{aligned}1&\leq\frac{H}{(v^*-v^{\pi^*_\lambda})^2}\times\frac{(v^*-v^{\pi^*_\lambda})^2}{H}\leq\frac{H}{(v^*-v^{\pi^*_\lambda})^2}KL(\theta(\pi^*_\lambda),\lambda(\pi^*_\lambda))\\&=\eta(\pi^*_\lambda)KL(\theta(\pi^*_\lambda),\lambda(\pi^*_\lambda))\leq\sum_{\pi'\in\Pi}\eta(\pi')KL(\theta(\pi'),\lambda(\pi')),\end{aligned}$$

where the second inequality follows from Lemma E.3. Next, if π is s.t. $\eta(\pi)=0$ then for any λ which makes π optimal it holds that

$$\begin{aligned}\sum_{\pi'\in\Pi}\eta(\pi')KL(\theta(\pi'),\lambda(\pi'))&\geq\sum_{(s,a)\in\pi^*_\lambda}\eta(\pi^*_{(s,a)})KL(\theta(\pi^*_{(s,a)}),\lambda(\pi^*_{(s,a)}))\\&=\sum_{(s,a)\in\pi^*_\lambda}\frac{H}{(v^*-v^{\pi^*_{(s,a)}})^2}KL(\theta(\pi^*_{(s,a)}),\lambda(\pi^*_{(s,a)}))\\&\geq\frac{H}{(v^*-v^{\pi^*_\lambda})^2}\sum_{(s,a)\in\pi^*_\lambda}KL(\theta(\pi^*_{(s,a)}),\lambda(\pi^*_{(s,a)}))\\&\geq\frac{H}{(v^*-v^{\pi^*_\lambda})^2}\sum_{(s,a)\in\pi^*_\lambda}KL(R_\theta(s,a),R_\lambda(s,a))\\&=\frac{H}{(v^*-v^{\pi^*_\lambda})^2}KL(\theta(\pi^*_\lambda),\lambda(\pi^*_\lambda))\geq1,\end{aligned}$$

where the second inequality follows from the fact that $v^{\pi^*_\lambda}\leq v^{\pi^*_{(s,a)}},\forall(s,a)\in\pi^*_\lambda$. □

E.3.1 Lower bound for Markov decision processes with bounded value function

Lemma E.5. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1. Consider an MDP $\theta\in\Theta$ with deterministic transitions. Then, for any reachable state-action pair (s,a) that is not visited by any optimal policy, there exists a confusing MDP $\lambda\in\Lambda(\theta)$ with*

- *λ and θ only differ in the immediate reward at (s,a)*
- *$KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_\lambda)=w^\pi_\theta(s,a)(v^*_\theta-v^{\pi^*_{(s,a)}}_\theta)^2$ for all $\pi\in\Pi$ where $v^{\pi^*_{(s,a)}}_\theta=\max_{\pi\colon w^\pi(s,a)>0}v^\pi_\theta$.*

Proof. Let $(s,a)\in\mathcal{S}\times\mathcal{A}$ be any state-action pair that is not visited by any optimal policy. Then $v^{\pi^*_{(s,a)}}_\theta=\max_{\pi\colon w^\pi(s,a)>0}v^\pi_\theta\leq v^*_\theta$ is strictly suboptimal in θ. Let $\tilde\pi$ be any policy that visits (s,a) and achieves the highest return $v^{\pi^*_{(s,a)}}_\theta$ in θ possible among such policies.

Define λ to be the MDP that matches θ except in the immediate reward at (s,a), which we set as $R_\lambda(s,a)=\mathcal{N}(r_\theta(s,a)+\Delta,1/2)$ with $\Delta=v^*_\theta-v^{\pi^*_{(s,a)}}_\theta$. That is, the expected reward of λ in (s,a) is raised by Δ. For any policy π, it then holds

$$\begin{aligned}KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_\lambda)&=w^\pi_\theta(s,a)KL(R_\theta(s,a),R_\lambda(s,a))\\v^\pi_\lambda&=w^\pi_\theta(s,a)\Delta+v^\pi_\theta\end{aligned}$$

due to the deterministic transitions. Hence, while $v^*_\lambda=v^*_\theta$ and all optimal policies of θ are still optimal in λ, now policy $\tilde\pi$, which is not optimal in θ is optimal in λ.

By the choice of Gaussian rewards with variance $1/2$, we have $KL(R_\theta(s,a),R_\lambda(s,a))=(v^*_\theta-v^{\pi^*_{(s,a)}}_\theta)^2$ and thus $KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_\lambda)=w^\pi_\theta(s,a)(v^*_\theta-v^{\pi^*_{(s,a)}}_\theta)^2$ for all $\pi\in\Pi$.

It only remains to show that $\lambda\in\Theta$, i.e., that all immediate rewards and optimal value function is bounded by 1. For rewards, we have

$$r_\lambda(s,a)=r_\theta(s,a)+\Delta=r_\theta(s,a)+v^*_\theta-v^{\pi^*_{(s,a)}}_\theta=v^*_\theta-\underbrace{(v^{\pi^*_{(s,a)}}_\theta-r_\theta(s,a))}_{\geq0}\leq v^*_\theta\leq1$$

for (s,a) and for all other (s',a'), $r_\lambda(s',a') = r_\theta(s',a') \leq 1$. Finally, the value function at any reachable state is bounded by the optimal return $v^*_\lambda = v^*_\theta \leq 1$ and for any unreachable state, the optimal value function of λ is identical to the optimal value function of θ. Hence, $\lambda \in \Theta$. □

Theorem 4.5. *Let Θ be the set of all episodic MDPs with Gaussian immediate rewards and optimal value function uniformly bounded by 1. Let $\theta \in \Theta$ be an instance with deterministic transitions. Then any uniformly good algorithm on Θ has expected regret on θ that satisfies*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathfrak{R}_\theta(K)]}{\log K} \geq \sum_{s,a\in\mathcal{Z}_\theta\,:\,\overline{\text{gap}}_\theta(s,a)>0} \frac{1}{H\cdot(v^*_\theta - v^{\pi^*_{(s,a)}}_\theta)} \geq \sum_{s,a\in\mathcal{Z}_\theta\,:\,\overline{\text{gap}}_\theta(s,a)>0} \frac{1}{H^2\cdot\overline{\text{gap}}_\theta(s,a)},$$

where $\mathcal{Z}_\theta = \{(s,a)\in\mathcal{S}\times\mathcal{A}\colon \forall\pi^*\in\Pi^*_\theta \quad w^{\pi^*}_\theta(s,a) = 0\}$ is the set of state-action pairs that no optimal policy in θ visits.

Proof. The proof works by first relaxing the general LP 8 and then considering its dual. We now define the set $\check{\Lambda}(\theta)$ which consists of all changes of environment which make $\pi^*_{(s,a)}$ optimal by only changing the distribution of the reward at (s,a) by making it $v^*_\theta - v^{\pi^*_{(s,a)}}_\theta$ larger. Formally, the set is defined as

$$\check{\Lambda}(\theta) = \big\{\lambda_{(s,a)}\colon \lambda\in\Lambda(\theta), KL(R_\theta(s,a),R_\lambda(s,a)) = (v^*_\lambda - v^{\pi^*_{(s,a)}})^2,$$
$$KL(R_\theta(s',a'),R_\lambda(s',a')) = 0, KL(P_\theta(s',a'),P_\lambda(s',a')) = 0, \forall(s',a')\neq(s,a)\big\}.$$

This set is guaranteed to be non-empty (for any reasonable MDP) by Lemma E.5. The relaxed LP is now give by

$$\begin{aligned} \underset{\eta(\pi)\geq 0}{\text{minimize}} \quad & \sum_{\pi\in\Pi}\eta(\pi)(v^*_\theta - v^\pi_\lambda) \\ \text{s. t.} \quad & \sum_{\pi\in\Pi}\eta(\pi)KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_\lambda)\geq 1 \qquad \text{for all } \lambda\in\check{\Lambda}(\theta). \end{aligned} \tag{16}$$

The dual of the above LP is given by

$$\begin{aligned} \underset{\mu(\lambda)\geq 0}{\text{maximize}} \quad & \sum_{\lambda\in\check{\Lambda}(\theta)}\mu(\lambda) \\ \text{s.t.} \quad & \sum_{\lambda\in\check{\Lambda}(\theta)}\mu(\lambda)KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_\lambda)\leq v^*_\theta - v^\pi_\theta \qquad \text{for all } \pi\in\Pi. \end{aligned} \tag{17}$$

By weak duality, the value of any feasible solution to (17) produces a lower bound on $C(\theta)$ in Theorem 4.2. Let

$$\mathcal{X} = \{(s,a)\in\mathcal{S}\times\mathcal{A}\colon w^\pi_\theta(s,a) = 0 \text{ for all } \pi\in\Pi^*_\theta \text{ and } w^\pi_\theta(s,a) > 0 \text{ for some } \pi\in\Pi\setminus\Pi^*_\theta\}$$

be the set of state-action pairs that are reachable in θ but no optimal policy visits. Then consider a dual solution μ that puts 0 on all confusing MDPs except on the $|\mathcal{X}|$ many MDPs from Lemma E.5. Since each such confusing MDP is associated with an $(s,a)\in\mathcal{X}$, we can rewrite μ as a mapping from $\mathcal{X}$ to $\mathbb{R}$ sending $(s,a)\to\lambda_{(s,a)}$. Specifically, we set

$$\mu(s,a) = \frac{1}{H}\left(v^*_\theta - v^{\pi^*_{(s,a)}}_\theta\right)^{-1} \qquad \text{for all } (s,a)\in\mathcal{X}.$$

To show that this μ is feasible, consider the LHS of the constraints in (17)

$$\begin{aligned} \sum_{\lambda\in\check{\Lambda}(\theta)}\mu(\lambda)KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_\lambda) &= \sum_{(s,a)\in\mathcal{X}}\frac{1}{H}\left(v^*_\theta - v^{\pi^*_{(s,a)}}_\theta\right)^{-1} KL(\mathbb{P}^\pi_\theta,\mathbb{P}^\pi_{(s,a)}) \\ &= \sum_{(s,a)\in\mathcal{X}}\frac{1}{H}\left(v^*_\theta - v^{\pi^*_{(s,a)}}_\theta\right)^{-1} w^\pi_\theta(s,a)(v^*_\theta - v^{\pi^*_{(s,a)}}_\theta)^2 \\ &= \sum_{(s,a)\in\mathcal{X}}\frac{1}{H}w^\pi_\theta(s,a)(v^*_\theta - v^{\pi^*_{(s,a)}}_\theta) \end{aligned}$$

where the first equality applies our definition of μ and the second uses the expression for the KL-divergence from Lemma E.5. By definition of $v_\theta^{\pi^*_{(s,a)}}$, we have $v_\theta^{\pi^*_{(s,a)}} \geq v_\theta^\pi$ for all policies π with $w_\theta^\pi(s,a) > 0$. Thus,

$$\begin{aligned}\sum_{(s,a)\in\mathcal{X}} \frac{1}{H} w_\theta^\pi(s,a)(v_\theta^* - v_\theta^{\pi^*_{(s,a)}}) &\leq \sum_{(s,a)\in\mathcal{X}} \frac{1}{H} w_\theta^\pi(s,a)(v_\theta^* - v_\theta^\pi) \\ &\leq v_\theta^* - v_\theta^\pi\end{aligned}$$

where the second inequality holds because each policy visits at most H states. Thus proves that μ defined above is indeed feasible. Hence, its objective value

$$\sum_{\lambda\in\Lambda(\theta)} \mu(\lambda) = \sum_{(s,a)\in\mathcal{X}} \frac{1}{H}\left(v_\theta^* - v_\theta^{\pi^*_{(s,a)}}\right)$$

is a lower-bound for $C(\theta)$ from Theorem 4.2 which finishes the proof. □

E.3.2 Tree-structured MDPs

Even though Lemma E.4 restricts the set of confusing environments from $\Lambda(\theta)$ to $\tilde{\Lambda}(\theta)$, this set could still have exponential or even infinite cardinality. In this section we show that for a type of special MDPs we can restrict ourselves to a finite subset of $\tilde{\Lambda}(\theta)$ of size at most SA.

Arrange $\pi^*_{(s,a)}, (s,a) \in \mathcal{S}\times\mathcal{A}$ according to the value functions $v^{\pi^*_{(s,a)}}$. Under this arrangement let $\pi_1 \succeq \pi_2 \succeq, \ldots, \succeq \pi_m$. Let $\pi_0 = \pi_\theta^*$. We will now construct m environments $\lambda_1, \ldots, \lambda_m$, which will constitute the finite subset. We begin by constructing λ_1 as follows. Let $\mathcal{B}_1$ be the set of all $(s_h, a_h) \in \pi_1$ and $(s_h, a_h) \notin \pi_0$. Arrange the elements in $\mathcal{B}_1$ in inverse dependence on horizon $(s_{h_1}, a_{h_1}) \preceq (s_{h_2}, a_{h_2}) \preceq \ldots \preceq (s_{h_{H_1}}, a_{h_{H_1}})$, where $H_1 = |\mathcal{B}_1|$, so that $h_1 > h_2 >, \ldots, h_{H_1}$. Let λ_1 be the environment which sets

$$\begin{aligned}
R_{\lambda_1}(s_{h_1}, a_{h_1}) &= \min(1, v^{\pi_0} - v^{\pi_1}) \\
R_{\lambda_1}(s_{h_2}, a_{h_2}) &= \min(1, \max(R_\theta(s_{h_2}, a_{h_2}), R_\theta(s_{h_2}, a_{h_2}) + v^{\pi_0} - (v^{\pi_1} - R_\theta(s_{h_1}, a_{h_1})) - 1))) \\
&\vdots \\
R_{\lambda_1}(s_{h_i}, a_{h_i}) &= \min(1, \max(R_\theta(s_{h_i}, a_{h_i}), R_\theta(s_{h_i}, a_{h_i}) + v^{\pi_0} - (v^{\pi_1} - \sum_{\ell=1}^{i} R_\theta(s_{h_\ell}, a_{h_\ell})) - i)) \\
&\vdots
\end{aligned}$$

Clearly λ_1 makes π_1 optimal and also does not change the value of any state-action pair which belongs to π_0 so it agrees with θ on π_0. Further $\pi_2, \pi_3, \ldots, \pi_m$ are still suboptimal policies under λ_1. This follows from the fact that for any $i > 1$, $v^{\pi_1} > v^{\pi_i}$ and there exists (s,a) such that $(s,a) \in \pi_i$ but $(s,a) \notin \pi_1$ so $R_{\lambda_1}(s,a) = R_\theta(s,a)$. Further λ_1 only increases the rewards for state-action pairs in π_1 and hence $v_{\lambda_1}^{\pi_1} > v_{\lambda_1}^{\pi_i}$. Notice that there exists an index $\tilde{H}_1$ at which $R_{\lambda_1}(s_{h_{\tilde{H}_1}}, a_{h_{\tilde{H}_1}}) = v^{\pi_0} - (v^{\pi_1} - \sum_{\ell=1}^{\tilde{H}_1} R_\theta(s_{h_\ell}, a_{h_\ell})) - \tilde{H}_1) \geq R_\theta(a_{\tilde{H}_1}, s_{\tilde{H}_1})$. For this index it holds that for $h < \tilde{H}_1$, $R_{\lambda_1}(s_h, a_h) = 1$ and for $h > \tilde{H}_1$, $R_{\lambda_1}(s_h, a_h) = R_\theta(s_h, a_h)$.

Let

$$\begin{aligned}
\mathcal{B}_i &= \{(s,a) \in \pi_i : (s,a) \notin \bigcup_{\ell<i} \pi_\ell\} \\
\tilde{\mathcal{B}}_i &= \{(s,a) \in \pi_i : (s,a) \in \bigcup_{\ell<i} \pi_\ell\}.
\end{aligned}$$

We first define an environment $\tilde{\lambda}_i$ on $(s,a) \in \tilde{\mathcal{B}}_i$ as follows. $R_{\lambda_i}(s,a) = R_{\lambda_\ell}(s,a)$, where $\ell < i$ is such that $(s,a) \in \mathcal{B}_\ell$. Let $v_{\tilde{\lambda}_i}^{\pi_i}$ be the value function of π_i with respect to $\tilde{\lambda}_i$.

Lemma E.6. *It holds that* $v_{\tilde{\lambda}_i}^{\pi_i} \leq v^{\pi_0}$.

Proof. Let $\tilde{H}_i$ be the index for which it holds that for $\ell \leq \tilde{H}_i$, $(s_{h_\ell}, a_{h_\ell}) \in \pi_i \iff (s_{h_\ell}, a_{h_\ell}) \in \mathcal{B}_i$. Such a $\tilde{H}_i$ exists as there is a unique sub-tree $\mathcal{M}_i$, of maximal depth, for which it holds that if $\pi_j \bigcap \mathcal{M}_i \neq \emptyset \iff \pi_i \succeq \pi_j$. The root of this subtree is exactly at depth $H - h_{\tilde{H}_i}$. Let π_j be any policy such that $\pi_j \succeq \pi_i$ and $\exists (s_{h_{\tilde{H}_i}}, a_{h_{\tilde{H}_i}}) \in \pi_j$. By the maximality of $\mathcal{M}_i$ such a π_j exists. Because of the tree structure it holds that for any $h' > h_{\tilde{H}_i}$ if $(s_{h'}, a_{h'}) \in \pi_i \implies (s_{h'}, a_{h'}) \in \pi_j$ and hence $\tilde{\lambda}_i = \lambda_j$ up to depth $h_{\tilde{H}_i}$. Since π_i and π_j match up to depth $H - h_{\tilde{H}_i}$ and $\pi_j \succeq \pi_i$ it also holds that

$$\sum_{\ell \leq \tilde{H}_i} R_{\lambda_j}(s_{h_\ell}^{\pi_j}, a_{h_\ell}^{\pi_j}) \geq \sum_{\ell \leq \tilde{H}_i} R_\theta(s_{h_\ell}^{\pi_j}, a_{h_\ell}^{\pi_j}) \geq \sum_{\ell \leq \tilde{H}_i} R_\theta(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i}) = \sum_{\ell \leq \tilde{H}_i} R_{\tilde{\lambda}_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i}).$$

Since π_j is optimal under λ_j the claim holds. □

For all $(s_{h_j}, a_{h_j}) \in \mathcal{B}_i$ we now set

$$R_{\lambda_i}(s_{h_j}, a_{h_j}) = \min(1, \max(R_\theta(s_{h_j}, a_{h_j}), R_\theta(s_{h_j}, a_{h_j}) + v^{\pi_0} - (v_{\tilde{\lambda}_i}^{\pi_i} - \sum_{\ell=1}^{j} R_{\tilde{\lambda}_i}(s_{h_\ell}, a_{h_\ell})) - j)), \tag{18}$$

and for all $(s_h, a_h) \in \tilde{\mathcal{B}}_i$ we set $R_{\lambda_i}(s_h, a_h) = R_{\tilde{\lambda}_i}(s_h, a_h)$. From the definition of $\tilde{\mathcal{B}}_i$ it follows that λ_i agrees with all λ_j for $j \leq i$ on state-action pairs in π_i. Finally we need to show that the construction in Equation 18 yields an environment λ_i for which π_i is optimal.

Lemma E.7. *Under λ_i it holds that π_i is optimal.*

Proof. Let $\tilde{H}_i$ and π_j be as in the proof of Lemma E.6. We now show that $\sum_{\ell \leq \tilde{H}_i} R_{\lambda_j}(s_{h_\ell}^{\pi_j}, a_{h_\ell}^{\pi_j}) \leq \sum_{\ell \leq \tilde{H}_i} R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i})$. We only need to show that $\sum_{\ell \leq \tilde{H}_i} R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i}) \geq v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i}$. From Equation 18 we have $R_{\lambda_i}(s_{h_1}, a_{h_1}) = \min(1, v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i})$. If $R_{\lambda_i}(s_{h_1}, a_{h_1}) = v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i}$ then the claim is complete. Suppose $R_{\lambda_i}(s_{h_1}, a_{h_1}) = 1$. This implies $v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i} \geq 1 - R_\theta(s_{h_1}, a_{h_1})$. Next the construction adds the remaining gap of $v^{\pi_0} - v_{\tilde{\lambda}_i}^{\pi_i} + R_\theta(s_{h_1}, a_{h_1}) - 1$ to $R_\theta(s_{h_2}, a_{h_2})$ and clips $R_{\lambda_i}(s_{h_2}, a_{h_2})$ to 1 if necessary. Continuing in this way we see that if ever $R_{\lambda_i}(s_{h_j}, a_{h_j}) = R_\theta(s_{h_j}, a_{h_j}) + v^{\pi_0} - (v_{\tilde{\lambda}_i}^{\pi_i} - \sum_{\ell=1}^{j} R_{\tilde{\lambda}_i}(s_{h_\ell}, a_{h_\ell})) - j$ then $v^{\pi_0} - V_{\tilde{\lambda}_i}^{\pi_i} \leq \sum_{\ell \leq \tilde{H}_i} R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i})$. On the other hand if this never occurs, we must have $R_{\lambda_i}(s_{h_\ell}^{\pi_i}, a_{h_\ell}^{\pi_i}) = 1 \geq R_{\lambda_j}(s_{h_\ell}^{\pi_j}, a_{h_\ell}^{\pi_j})$ which concludes the claim. □

Let $\hat{\Lambda}(\theta) = \{\lambda_1, \ldots, \lambda_m\}$ be the set of the environments constructed above. We now show that the value of the optimization problem is not too much smaller than the value of Problem 8.

Theorem E.8. *The value $\hat{C}(\theta)$ of the LP*

$$\begin{aligned} \underset{\eta(\pi) \geq 0}{\text{minimize}} \quad & \sum_{\pi \in \Pi^*} \eta(\pi)(v^* - v^\pi) \\ \text{s. t.} \quad & \sum_{\pi \in \Pi^*} \eta(\pi) KL(\theta(\pi), \lambda(\pi)) \geq 1, \forall \lambda \in \hat{\Lambda}(\theta), \end{aligned}$$

satisfies $\hat{C}(\theta) \geq \frac{C(\theta)}{H^2}$ and $C(\theta) \geq \frac{\hat{C}(\theta)}{H}$.

Proof. The inequality $C(\theta) \geq \frac{\hat{C}(\theta)}{H}$ follows from Lemma E.4 and the fact that the above optimization problem is a relaxation to LP 15.

To show the first inequality we consider the following relaxed LP

$$\begin{aligned} \underset{\eta(\pi) \geq 0}{\text{minimize}} \quad & \sum_{\pi \in \Pi} \eta(\pi)(v^* - v^\pi) \\ \text{s. t.} \quad & \sum_{\pi \in \Pi} \eta(\pi) KL(\theta(\pi), \lambda(\pi)) \geq 1, \forall \lambda \in \hat{\Lambda}(\theta). \end{aligned}$$

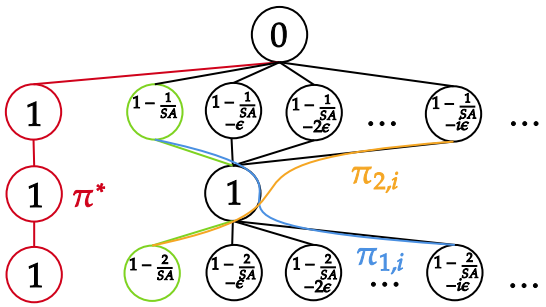


Figure 5: Issue with restricting LP to Π^*

Any solution to the LP in the statement of the theorem is feasible for the above LP and thus the value of the above LP is no larger. We now show that the value of the above LP is greater than or equal to $\frac{C(\theta)}{H^2}$. Fix $\lambda \in \hat{\Lambda}(\theta)$. We show that for any $\lambda' \in \Lambda(\theta)$ such that $\pi^*_\lambda = \pi^*_{\lambda'}$ it holds that $KL(\theta(\pi), \lambda(\pi)) \leq H^2 KL(\theta(\pi), \lambda'(\pi)), \forall \pi \in \Pi$. This would imply that if η is a solution to the above LP, then $H^2\eta$ is feasible for LP 8 and therefore $\hat{C}(\theta) \geq \frac{C(\theta)}{H^2}$.

Arrange $\pi \in \Pi : KL(\theta(\pi), \lambda(\pi)) > 0$ according to $KL(\theta(\pi), \lambda(\pi))$ so that

$$\pi_i \preceq \pi_j \iff KL(\theta(\pi_i), \lambda(\pi_i)) \geq KL(\theta(\pi_j), \lambda(\pi_j)).$$

Consider the optimization problem

$$\begin{aligned} &\underset{\lambda' \in \Lambda(\theta)}{\text{minimize}} && KL(\theta(\pi_i), \lambda'(\pi_i)) \\ &\text{s. t.} && \pi^*_{\lambda'} = \pi^*_\lambda. \end{aligned}$$

If we let $\Delta_{\lambda'}(s_h, a_h), (s_h, a_h) \in \pi^*_\lambda$ denote the change of reward for (s_h, a_h) under environment λ', then the above optimization problem can be equivalently written as

$$\begin{aligned} &\underset{\lambda' \in \Lambda(\theta)}{\text{minimize}} && \sum_{h=1}^{h_{\tilde{H}_i}} \Delta_{\lambda'}(s_h, a_h)^2 \\ &\text{s. t.} && \sum_{h=1}^{H} r(s_h, a_h) + \Delta_{\lambda'}(s_h, a_h) \geq v^*. \end{aligned}$$

It is easy to see that the solution to the above optimization problem is to set $r(s_h, a_h) + \Delta_{\lambda'}(s_h, a_h) = 1$ for all $h \in [h_{\tilde{H}_i} + 1, H]$ and spread the remaining mass of $v^* - \tilde{H}_i - (v^{\pi^*_\lambda} - \sum_{\ell=1}^{\tilde{H}_i} R_\theta(s_{h_\ell}, a_{h_\ell})$ as uniformly as possible on $\Delta_{\lambda'}(s_h, a_h)$, $h \in [1, h_{\tilde{H}_i}]$. Notice that under this construction the solution to the above optimization problem and λ match for $h \in [h_{\tilde{H}_i} + 1, H]$. Since the remaining mass is now the same it now holds that for any λ', $\sum_{h=1}^{h_{\tilde{H}_i}} \Delta_{\lambda'}(s_h, a_h)^2 \geq \frac{1}{h^2_{\tilde{H}_i}} \sum_{h=1}^{h_{\tilde{H}_i}} \Delta_\lambda(s_h, a_h)^2$. This implies $KL(\theta(\pi_i), \lambda'(\pi_i)) \geq \frac{1}{\tilde{H}_i^2} KL(\theta(\pi), \lambda(\pi))$ and the result follows as $\tilde{H}_i \leq H, \forall i \in [H]$. $\square$

E.3.3 Issue with deriving a general bound

We now try to give some intuition regarding why we could not derive a generic lower bound for deterministic transition MDPs. We have already outlined our general approach of restricting the set Π and $\Lambda(\theta)$ to finite subsets of manageable size and then showing that the value of the LP on these restricted sets is not much smaller than the value of the original LP. One natural restriction of Π is the set Π^* from Theorem 4.5. Suppose we restrict ourselves to the same set and consider only environments making policies in Π^* optimal as the restriction for $\Lambda(\theta)$. We now give an example of an MDP for which such a restriction will lead to an $\Omega(SA)$ multiplicative discrepancy between the value of the original semi-infinite LP and the restricted LP. The MDP can be found in Figure 5. The rewards for each action for a fixed state s are equal and are shown in the vertices corresponding to the states. The number of states in the second and last layer of the MDP are equal to $(SA - 3)/2$.

The optimal policy takes the red path and has value $V^{\pi^*} = 3$. The set Π^* consists of all policies $\pi_{j,i}$ which visit one of the states in green. The policies $\pi_{1,i}$, in blue, visit the green state in the second layer of the MDP and one of the states in the final layer, following the paths in blue. Similarly the policies $\pi_{2,i}$, in orange, visit one of the state in the second layer and the green state in the last layer, following the orange paths. The value function of $\pi_{j,i}$ is $V^{\pi_{j,i}} = 3 - \frac{3}{SA} - i\epsilon$, where $0 \leq i \leq (SA-4)/2$. We claim that playing each $\pi_{j,i}$ $\eta(\pi_{j,i}) = \Omega(SA)$ times is a feasible solution to the LP restricted to Π^*. Fix i, the $\lambda_{\pi_{1,i}}$ must put weight at least $1/SA$ on the green state in layer 2. Coupling with the fact that for all i' the rewards $\pi_{1,i'}$ are also changed under this environment we know that the constraint of the restricted LP with respect to $\lambda_{\pi_{1,i}}$ is lower bounded by $\sum_{i'} \eta(\pi_{1,i'})/(SA)^2$. Since there are $\Omega(SA)$ policies $\{\pi_{1,i'}\}_{i'}$, this implies that $\eta(\pi_{1,i}) = \Omega(SA)$ is feasible. A similar argument holds for any $\pi_{2,i}$. Thus the value of the restricted LP is at most $O(SA)$, for any $\epsilon \ll SA$.

However, we claim that the value of the semi-infinite LP which actually characterizes the regret is at least $\Omega(S^2A^2)$. First, to see that the above assignment of η is not feasible for the semi-infinite LP, consider any policy $\pi \notin \Pi^*$, e.g. take the policy which visits the state in layer 2 with reward $1 - 1/SA - \epsilon$ and the state in layer 4 with reward $1 - 2/SA - \epsilon$. Each of these states have been visited $O(SA)$ times and $\eta(\pi) = 0$ hence the constraint for the environment λ_π is upper bounded by $SA\left(\left(\frac{1}{SA} + \epsilon\right)^2 + \left(\left(\frac{2}{SA} + \epsilon\right)^2\right)\right) \approx 1/SA$. In general each of the states in black in the second layer and the fourth layer have been visited $1/SA$ times less than what is necessary to distinguish any $\pi \notin \Pi^*$ as sub-optimal. If we define the i-th column of the MDP as the pair consisting of the states with rewards $1 - 1/SA - i\epsilon$ and $1 - 2/SA - i\epsilon$ then to distinguish the policy visiting both of these states as sub-optimal we need to visit at least one of these $\Omega(S^2A^2)$ times. This implies we need to visit each column of the MDP $\Omega(S^2A^2)$ times and thus any strategy must incur regret at least $\Omega\left(\sum_i S^2A^2\frac{1}{SA}\right) = \Omega(S^2A^2)$, leading to the promised multiplicative gap of $\Omega(SA)$ between the values of the two LPs.

Why does such a gap arise and how can we hope to fix it this issue? Any feasible solution to the LP restricted to Π^* essentially needs to visit the states in green $\Theta(S^2A^2)$ times. This is sufficient to distinguish the green states as sub-optimal to visit and hence any strategy visiting these states would be also deemed sub-optimal. This is achievable by playing each strategy in Π^* in the order of $\Theta(SA)$ times as already discussed. Now, even though Π^* covers all other states, from our argument above we see that we need to play each $\pi \in \Pi^*$ in the order of $\Theta(S^2A^2)$ times to be able to determine all sub-optimal states. To solve this issue, we either have to increase the size of Π^* to include for example all policies visiting each column of the MDP or at the very least include changes of environments in the constraint set which make such policies optimal. This is clearly computationally feasible for the MDP in Figure 5, however, it is not clear how to proceed for general MDPs, without having to include exponentially many constraints. This begs the question about the computational hardness of achieving both upper and lower regret bounds in a factor of $o(SA)$ from what is optimal.

E.4 Lower bounds for optimistic algorithms in MDPs with deterministic transitions

In this section we prove a lower bound on the regret of optimistic algorithms, demonstrating that optimistic algorithms can not hope to achieve the information-theoretic lower bounds even if the MDPs have deterministic transitions. While the result might seem similar to the one proposed by Simchowitz and Jamieson [30] (Theorem 2.3) we would like to emphasize that the construction of Simchowitz and Jamieson [30] does not apply to MDPs with deterministic transitions, and that the idea behind our construction is significantly different.

Consider the MDP in Figure 6. This MDP has $2n + 9$ states and $4n + 8$ actions. The rewards for each action are either $1/12$ or $1/12 + \epsilon/2$ and can be found next to the transitions from the respective states. We are going to label the states according to their layer and their position in the layer so that the first state is $s_{1,1}$ the state which is to the left of $s_{1,1}$ in layer 2 is $s_{2,1}$ and to the right $s_{2,2}$. In general the i-th state in layer h is denoted as $s_{h,i}$. The rewards in all states are deterministic, with a single exception of a Bernoulli reward from state $s_{4,1}$ to $s_{5,2}$ with mean $1/12$. From the construction it is clear that $V^*(s_{1,1}) = 1/2 + \epsilon$. Further there are two sets of optimal policies with the above value function – the n optimal policies which visit state $s_{2,2}$ and the n optimal policies which visit $s_{5,1}$. Notice that the information-theoretic lower bound for this MDP is in $O(\log(K)/\epsilon)$ as only the transition from state $s_{4,1}$ to $s_{5,2}$ does not belong to an optimal policy. In particular, there is no

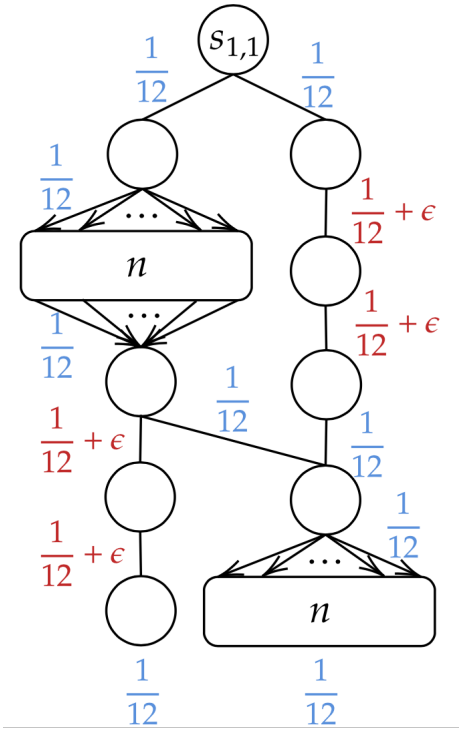


Figure 6: Deterministic MDP instance for optimistic lower bound

dependence on n. Next we try to show that the class of optimistic algorithms will incur regret at least $\Omega(n \log(\delta^{-1})/\epsilon)$.

Class of algorithms. We adopt the class of algorithms from Section G.2 in [30] with an additional assumption which we clarify momentarily. Recall that the class of algorithms assumes access to an optimistic value function $\bar{V}_k(s) \geq V^*(s)$ and optimistic Q-functions. In particular the algorithms construct optimistic Q and value functions as

$$\bar{V}_k(s) = \max_{a \in \mathcal{A}} \bar{Q}_k(s,a)$$
$$Q_k(s,a) = \hat{r}_k(s,a) + b_k^{rw}(s,a) + \hat{p}_k(s,a)^\top \bar{V}_k + b_k(s,a).$$

We assume that there exists a $c \geq 1$ such that

$$\frac{c}{2}\sqrt{\frac{\log(M(1 \vee n_k(s,a)))/\delta}{(1 \vee n_k(s,a))}} \leq b_k^{rw}(s,a) \leq c\sqrt{\frac{\log(M(1 \vee n_k(s,a)))/\delta}{(1 \vee n_k(s,a))}},$$

where $M = \theta(n)$ and $b_k(s,a) \sim \sqrt{S} f_k(s,a) b_k^{rw}(s,a)$, where f_k is a decreasing function in the number of visits to (s,a) given by $n_k(s,a)$. For $n_k(s,a) = \Omega(n \log(n))$, we assume $b_k(s,a) \leq b_k^{rw}(s,a)$. One can verify that this is true for the the Q and value functions of StrongEuler.

Lower bound. Let $\epsilon > 0$ be sufficiently small to be specified later and let N be such that

$$N = \lfloor \frac{c^2 n \log(MN/(n\delta))}{16\epsilon^2} \rfloor.$$

Lemma E.9. *There exists n_0, ϵ_0 such that for any pair of $n \geq n_0$ and $\epsilon \leq \epsilon_0$ and any $k \leq N$, with probability at least $1 - \delta$, it holds that either $n_k(s_{5,1}) < N/4$, or $\bar{Q}_k(s_{4,1}, 1) < \bar{Q}_k(s_{4,1}, 2)$.*

Proof. Assume $n_k(s_{5,1}) \geq N/4$, then we have

$$\begin{aligned}\bar{Q}_k(s_{4,1},1) &= \frac{1}{4} + \epsilon + \sum_{i=4}^{6} b_k^{rw}(s_{i,1},1) + b_k(s_{i,1}) \\ &\leq \frac{1}{4} + \epsilon + 6c\sqrt{\frac{\log(MN/(4\delta))}{N/4}} \leq \frac{1}{4} + \epsilon + \frac{48\epsilon}{\sqrt{n}}\,,\end{aligned}$$

where we assume ϵ is sufficiently small such that $b_k(s,a) \leq b_k^{rw}(s,a)$ for $n_k(s,a) \geq N/4$.

On the other hand, we have have with probability at least 1-δ, that $\forall k : \hat{r}_k(s_{4,1},2) + b_k^{rw}(s_{4,1},2) \geq 1/12$. Hence conditioned under that event, we have

$$\begin{aligned}\bar{Q}_k(s_{4,1},2) &= \frac{1}{4} + b_k^{rw}(s_{4,1},2) + b_k(s_{4,1},2) + \max_{j\in\{2,\dots n+1} \sum_{i=5}^{6} b_k^{rw}(s_{i,j},1) + b_k(s_{i,j},1) \\ &\geq \frac{1}{4} + c\sqrt{\frac{\log(MN/(n\delta))}{N/n}} \geq \frac{1}{4} + 4\epsilon\,.\end{aligned}$$

The proof is completed for $n_0 = 48^2$. □

We can show the same for the upper part of the MDP.

Lemma E.10. *There exists n_0, ϵ_0 such that for any pair of $n \geq n_0$ and $\epsilon \leq \epsilon_0$ and any $k \leq N$, with probability at least $1-\delta$, it holds that either $n_k(s_{1,2}) < N/4$, or $\bar{Q}_k(s_{1,1},2) < \bar{Q}_k(s_{1,1},1)$.*

Proof. First we split $\bar{Q}_k(s_{1,1},2)$ into the observed sum of mean rewards and bonuses from $s_{1,1}$ to $s_{5,2}$ and the value $\bar{V}_k(s_{5,2})$. Then we upper bound $\bar{Q}_k(s_{1,1},1)$ by $\bar{V}_k(s_{5,2})$ and the maximum observed sum of mean rewards and bonuses along the paths passing by $s_{3,j}$ for $j \in [n]$. Finally analogous to the proof of Lemma E.9, it is straightforward show that the latter is always larger as long as the visitation count for $s_{2,2}$ exceeds $N/4$. □

Theorem E.11. *There exists an MDP instance with deterministic transitions on which any optimistic algorithm with confidence parameter δ will incur expected regret of at least $\Omega(S\log(\delta^{-1})/\epsilon))$ while it is asymptotically possible to achieve $\Omega(\log(K)/\epsilon)$ regret.*

Proof. Taking the MDP from Figure 6. Applying Lemma E.9 and E.10 shows that after N episodes with probability at least $1-2\delta$, the visitation count of $s_{2,2}$ and $s_{5,1}$ each do not exceed $N/4$. Hence there are at least $N/2$ episodes in which neither of them is visited, which means an ϵ-suboptimal policy is taken. Hence the expected regret after N episodes is at least

$$(1-2\delta)\epsilon N/2 = \Omega\left(\frac{S\log(\delta^{-1})}{\epsilon}\right).$$

□

Theorem E.11 has two implications for optimistic algorithms in MDPs with deterministic transitions.

- It is impossible to be asymptotically optimal if the confidence parameter δ is tuned to the time horizon K.
- It is impossible to have an anytime bound matching the information-theoretic lower bound.

F Proofs and extended discussion for regret upper-bounds

F.1 Further discussion on Opportunity O.2

The example in Figure 1 does not illustrate **O.2** to its fullest extent. We now expand this example and elaborate why it is important to address Opportunity **O.2**. Our example can be found in Figure 7. The MDP is an extension of the one presented in Figure 1 with the new addition of actions a_5 and a_6 in

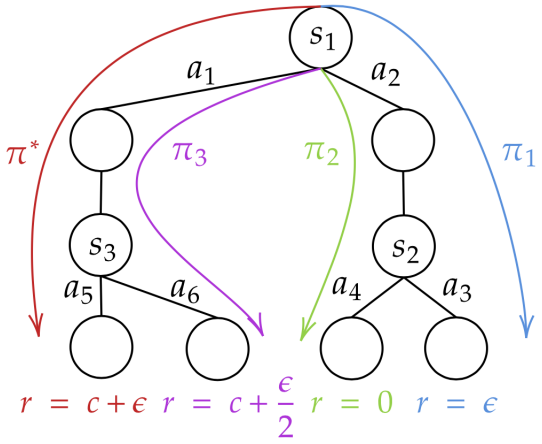


Figure 7: Example for Opportunity **O.2**

state s_3 and the new state following action a_6. Again there is only a single action available at all other states than s_1, s_2, s_3. The reward of the state following action a_6 is set as $r = c + \epsilon/2$. This defines a new sub-optimal policy π_3 and the gap $\mathrm{gap}(s_3, a_6) = \frac{\epsilon}{2}$. Information theoretically it is impossible to distinguish π_3 as sub-optimal in less than $\Omega(\log(K)/\epsilon^2)$ rounds and so any uniformly good algorithm would have to pay at least $O(\log(K)/\epsilon)$ regret. However, what we observed previously still holds true, i.e., we should not have to play more than $\log(K)/c^2$ rounds to eliminate both π_1 and π_2 as sub-optimal policies. Prior work now suffers Opportunity **O.2** as it would pay $\log(K)/\epsilon$ regret for all zero gap state-action pairs belonging to either π_1 or π_2, essentially evaluating to $SA\log(K)/\epsilon$. On the other hand our bounds will only pay $\log(K)/\epsilon$ regret for zero gap state-action pairs belonging to π_3.

F.2 Useful decomposition lemmas

We start by providing the following lemma that establishes that the instantaneous regret can be decomposed into gaps defined w.r.t. any optimal (and not necessarily Bellman optimal) policy.

Lemma F.1 (General policy gap decomposition)**.** *Let* $\mathrm{gap}^{\hat{\pi}}(s,a) = V^{\hat{\pi}}(s) - Q^{\hat{\pi}}(s,a)$ *for any optimal policy* $\hat{\pi} \in \Pi^*$*. Then the difference in values of* $\hat{\pi}$ *and any policy* $\pi \in \Pi$ *is*

$$V^{\hat{\pi}}(s) - V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} \mathrm{gap}^{\hat{\pi}}(S_h, A_h) \,\middle|\, S_{\kappa(s)} = s\right] \tag{19}$$

and, further, the instantaneous regret of π *is*

$$v^* - v^{\pi} = \sum_{s,a} w^{\pi}(s,a)\,\mathrm{gap}^{\hat{\pi}}(s,a). \tag{20}$$

Proof. We start by establishing a recursive bound for the value difference of π and $\hat{\pi}$ for any s

$$\begin{aligned} V^{\hat{\pi}}(s) - V^{\pi}(s) &= V^{\hat{\pi}}(s) - Q^{\hat{\pi}}(s,\pi(s)) + Q^{\hat{\pi}}(s,\pi(s)) - V^{\pi}(s) \\ &= \mathrm{gap}^{\hat{\pi}}(s,\pi(s)) + Q^{\hat{\pi}}(s,\pi(s)) - Q^{\pi}(s,\pi(s)) \\ &= \mathrm{gap}^{\hat{\pi}}(s,\pi(s)) + \sum_{s'} P_{\theta}(s'|s,\pi(s))[V^{\hat{\pi}}(s') - V^{\pi}(s')]. \end{aligned}$$

Unrolling this recursion for all layers gives

$$V^{\hat{\pi}}(s) - V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} \mathrm{gap}^{\hat{\pi}}(S_h, A_h) \,\middle|\, S_{\kappa(s)} = s\right].$$

To show the second identity, consider $s = s_1$ and note that $v^{\pi} = V^{\pi}(s_1)$ and $v^* = v^{\hat{\pi}} = V^{\hat{\pi}}(s_1)$ because $\hat{\pi}$ is an optimal policy. □

For the rest of the paper we are going to focus only on the Bellman optimal policy from each state and hence only consider $\mathrm{gap}^{\hat{\pi}}(s,a) = \mathrm{gap}(s,a)$. All of our analysis will also go through for arbitrary $\mathrm{gap}^{\hat{\pi}}, \hat{\pi} \in \Pi^*$, however, this did not provide us with improved regret bounds.

We now show the following technical lemma which generalizes the decomposition of value function differences and will be useful in the surplus clipping analysis.

Lemma F.2. *Let $\Psi : \mathcal{S} \to \mathbb{R}$, $\Delta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be functions satisfying $\Psi(s) = 0$ for any s with $\kappa(s) = H+1$ and $\pi \colon \mathcal{S} \to \mathcal{A}$ a deterministic policy. Further, assume that the following relation holds*

$$\Psi(s) = \Delta(s, \pi(s)) + \langle P(\cdot|s,\pi(s)), \Psi \rangle,$$

and let $\mathcal{A}$ be any event that is $\mathcal{H}_h$-measurable where $\mathcal{H}_h = \sigma(S_1, A_1, R_1, \dots, S_h)$ is the sigma-field induced by the episode up to the state at time h. Then, for any $h \in [H]$ and $h' \in \mathbb{N}$ with $h \le h' \le H+1$, it holds that

$$\mathbb{E}_\pi[\chi(\mathcal{A})\,\Psi(S_h))] = \mathbb{E}_\pi\left[\chi(\mathcal{A})\left(\sum_{t=h}^{h'-1}\Delta(S_t,A_t) + \Psi(S_{h'+1})\right)\right] = \mathbb{E}_\pi\left[\chi(\mathcal{A})\sum_{t=h}^{H}\Delta(S_t,A_t)\right].$$

Proof. First apply the assumption of Ψ recursively to get

$$\Psi(s) = \mathbb{E}_\pi\left[\sum_{t=\kappa(s)}^{h'-1}\Delta(S_t,A_t) + \Psi(S_{h'}) \,\middle|\, S_{\kappa(s)} = s\right].$$

Plugging this identity into $\mathbb{E}_\pi[\chi(\mathcal{A})\,\Psi(S_h))]$ yields

$$\begin{aligned}
\mathbb{E}_\pi[\chi(\mathcal{A})\,\Psi(S_h))] &= \mathbb{E}_\pi\left[\chi(\mathcal{A})\,\mathbb{E}_\pi\left[\sum_{t=h}^{h'-1}\Delta(S_t,A_t) + \Psi(S_{h'}) \,\middle|\, S_h\right]\right] \\
&\overset{(i)}{=} \mathbb{E}_\pi\left[\chi(\mathcal{A})\,\mathbb{E}_\pi\left[\sum_{t=h}^{h'-1}\Delta(S_t,A_t) + \Psi(S_{h'}) \,\middle|\, \mathcal{H}_h\right]\right] \\
&\overset{(ii)}{=} \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\chi(\mathcal{A})\left(\sum_{t=h}^{h'-1}\Delta(S_t,A_t) + \Psi(S_{h'})\right) \,\middle|\, \mathcal{H}_h\right]\right] \\
&\overset{(iii)}{=} \mathbb{E}_\pi\left[\chi(\mathcal{A})\left(\sum_{t=h}^{h'-1}\Delta(S_t,A_t) + \Psi(S_{h'})\right)\right]
\end{aligned}$$

where $\mathcal{H}_h = \sigma(S_1, A_1, R_1, \dots, S_h)$ is the sigma-field induced by the episode up to the state at time h. Identity (i) holds because of the Markov-property and (ii) holds because $\mathcal{A}$ is $\mathcal{H}_h$-measurable. The final identity (iii) uses the tower-property of conditional expectations. □

F.3 General surplus clipping for optimistic algorithms

Clipped operators. One of the main arguments to derive instance dependent bounds is to write the instantaneous regret in terms of the surpluses which are clipped to the minimum positive gap. We now define the clipping threshold $\epsilon_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_0^+$ and associated clipped surpluses

$$\ddot{E}_k(s,a) = \mathrm{clip}\left[E_k(s,a) \mid \epsilon_k(s,a)\right] = \chi\left(E_k(s,a) \ge \epsilon_k(s,a)\right) E_k(s,a). \tag{21}$$

Next, define the clipped Q- and value-function as

$$\ddot{Q}_k(s,a) = \ddot{E}_k(s,a) + r(s,a) + \langle P(\cdot|s,a), \ddot{V}_k, \rangle \quad \text{and} \quad \ddot{V}_k(s) = \ddot{Q}_k(s,\pi_k(s)). \tag{22}$$

The random variable which is the state visited by π_k at time h throughout episode k is denoted by S_h and A_h is the action at time h.

Events about encountered gaps Define the event $\mathcal{E}_h = \{\text{gap}(S_h, A_h) > 0\}$ that at time h an action with a positive gap played, the $\mathcal{P}_{1:h} = \bigcap_{h'=1}^{h-1} \mathcal{E}_{h'}^c$ that only actions with zero gap have been played until h and the event $\mathcal{A}_h = \mathcal{E}_h \cap \mathcal{P}_{1:h}$ that the first positive gap was encountered at time h. Let $\mathcal{A}_{H+1} = \mathcal{P}_{1:H}$ be the event that only zero gaps were encountered. Further, let

$$B = \min\{h \in [H+1] \colon \text{gap}(S_h, A_h) > 0\}$$

be the first time a non-zero gap is encountered. Note that B is a stopping time w.r.t. the filtration $\mathcal{F}_h = \sigma(S_1, A_1, \dots, S_h, A_h)$.

The proof of Simchowitz and Jamieson [30] consists of two main steps. First show that for their definition of clipped value functions one can bound $\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \frac{1}{2}(\bar{V}_k(s_1) - V^{\pi_k}(s_1))$. Next, using optimism together with the fact that π_k has highest value function at episode k it follows that $\bar{V}_k(s_1) - V^{\pi_k}(s_1) \geq V^*(s_1) - V^{\pi_k}(s_1)$. The second main step is to use a high-probability bound on the clipped surpluses to relate them to the probability to visit the respective state-action pair and the proof is finished via an integration lemma. We now show that the first step can be carried out in greater generality by defining a less restrictive clipping operator. This operator is independent of the details in the definition of gap at each state-action pair but rather only uses a certain property which allows us to decompose the episodic regret as a sum over gaps. We will also further show that one does not need to use an integration lemma for the second step but can rather reformulate the regret bound as an optimization problem. This will allow us to clip surpluses at state-action pairs with zero gaps beyond the $\text{gap}_{\min}$ rate.

Clipping with an arbitrary threshold. Recall the definition of the clipped surpluses and clipped value function in Equation 21 and Equation 22. We begin by showing a general relation between the clipped value function difference and the non-clipped surpluses for any clipping threshold $\epsilon_k : \mathcal{S} \to \mathbb{R}$. This will help in establishing $\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \frac{1}{2}(\bar{V}_k(s_1) - V^{\pi_k}(s_1))$.

Lemma F.3. *Let $\epsilon_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_0^+$ be arbitrary. Then for any optimistic algorithm it holds that*

$$\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \left(\text{gap}(S_h, A_h) - \epsilon_k(S_h, A_h)\right)\right]. \tag{23}$$

Proof. We use $W_k(s) = \ddot{V}_k(s) - V^{\pi_k}(s)$ in the following and first show that $W(s_1) \geq \mathbb{E}_{\pi_k}[W_k(S_B)]$. As a precursor, we prove

$$\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_h)\right] \geq \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_{h+1}\right) W_k(S_{h+1})\right] + \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h+1}\right) W_k(S_{h+1})\right]. \tag{24}$$

To see this, plug the definitions into $W_k(s)$ which gives $W_k(s) = \ddot{V}_k(s) - V^{\pi_k}(s) = \ddot{E}_k(s, \pi_k(s)) + \langle P(\cdot|s, \pi_k(s)), W_k\rangle$ and use this in the LHS of (24) as

$$\begin{aligned}
\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_h)\right] &= \mathbb{E}_{\pi_k}\Big[\chi\left(\mathcal{P}_{1:h}\right) \underbrace{\ddot{E}_k(S_h, A_h)}_{\geq 0}\Big] + \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) \mathbb{E}[W_k(S_{h+1}) \mid S_h]\right] \\
&\overset{(i)}{\geq} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) \mathbb{E}_{\pi_k}[W_k(S_{h+1}) \mid \mathcal{H}_h]\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_{h+1}) \mid \mathcal{H}_h]\right] = \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_{h+1})\right]
\end{aligned}$$

where $\mathcal{H}_h = \sigma(S_1, A_1, R_1, \dots, S_h)$ is the sigma-field induced by the episode up to the state at time h. Step (i) follows from $\text{clip}[\cdot|c] \geq 0$ for any $c \geq 0$ and the Markov property and (ii) holds because $\mathcal{P}_{1:h}$ is $\mathcal{H}_h$-measurable. We now rewrite the RHS by splitting the expectation based on whether event $\mathcal{E}_{h+1}$ occurred as

$$\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h}\right) W_k(S_{h+1})\right] = \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{P}_{1:h+1}\right) W_k(S_{h+1})\right] + \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_{h+1}\right) W_k(S_{h+1})\right].$$

We have now shown (24), which we will now use to lower-bound $W_k(s_1)$ as

$$\begin{aligned}
W_k(s_1) &= \mathbb{E}_{\pi_k}[\chi\left(\mathcal{E}_1\right) W_1(S_1)] + \mathbb{E}_{\pi_k}[\chi\left(\mathcal{E}_1^c\right) W_1(S_1)] \\
&= \mathbb{E}_{\pi_k}[\chi\left(\mathcal{A}_1\right) W_1(S_1)] + \mathbb{E}_{\pi_k}[\chi\left(\mathcal{P}_{1:1}\right) W_1(S_1)] \\
&\geq \mathbb{E}_{\pi_k}[\chi\left(\mathcal{A}_1\right) W_1(S_1)] + \sum_{h=2}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) W_k(S_h)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) W_k(S_h)\right] = \mathbb{E}_{\pi_k}[W_k(S_B)].
\end{aligned}$$

Applying Lemma F.2 with $\mathcal{A} = \mathcal{A}_h$, $\Psi = W_k$ and $\Delta = \ddot{E}_k$ yields

$$\begin{aligned} W_k(s_1) &\geq \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \ddot{E}_k(S_{h'}, A_{h'})\right] \\ &\geq \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} E_k(S_{h'}, A_{h'})\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \epsilon_k(S_{h'}, A_{h'})\right], \end{aligned}$$

where we applied the definition clipped surpluses which gives $\ddot{E}_k(s,a) = \operatorname{clip}[E_k(s,a) \mid \epsilon_k(s,a)] \geq E_k(s,a) - \epsilon_k(s,a)$. It only remains to show that

$$\mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} E_k(S_{h'}, A_{h'})\right] \geq \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \operatorname{gap}(S_{h'}, A_{h'})\right].$$

To do so, we apply Lemma F.2 twice, first with $\mathcal{A} = \mathcal{A}_h$, $\Psi = \bar{V}_k - V^{\pi_k}$ and $\Delta = E_k$ and then again with $\mathcal{A} = \mathcal{A}_h$, $\Psi = V^* - V^{\pi_k}$ and $\Delta = \operatorname{gap}$ which gives

$$\begin{aligned} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} E_k(S_{h'}, A_{h'})\right] &= \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right)\left(\bar{V}_k(S_h) - V^{\pi_k}(S_h)\right)\right] \\ &\geq \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right)\left(V^*(S_h) - V^{\pi_k}(S_h)\right)\right] \\ &= \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \operatorname{gap}(S_{h'}, A_{h'})\right]. \end{aligned}$$

Thus, we have shown that

$$\begin{aligned} &\ddot{V}_k(s_1) - V^{\pi_k}(s_1) = W_k(s_1) \\ &\geq \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \operatorname{gap}(S_{h'}, A_{h'})\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \epsilon_k(S_{h'}, A_{h'})\right] \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\chi\left(\mathcal{A}_h\right) \sum_{h'=h}^{H} \left(\operatorname{gap}(S_{h'}, A_{h'}) - \epsilon_k(S_{h'}, A_{h'})\right)\right] \\ &= \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \left(\operatorname{gap}(S_h, A_h) - \epsilon_k(S_h, A_h)\right)\right] \end{aligned}$$

where the last equality uses the definition of B, the first time step at which a non-zero gap was encountered. □

Lemma F.4 (Optimism of clipped value function)**.** *Let the clipping thresholds $\epsilon_k \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_0^+$ used in the definition of $\ddot{V}_k$ satisfy*

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \epsilon_k(S_h, A_h)\right] \leq \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h)\right]$$

for some optimal policy $\hat{\pi}$. Then scaled optimism holds for the clipped value function, i.e.,

$$\ddot{V}_k(s_1) - V^{\pi_k}(s_1) \geq \frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1)).$$

Proof. The proof works by establishing the following chain of inequalities:

$$\begin{aligned}\frac{V^*(s_1) - V^{\pi_k}(s_1)}{2} &\overset{(a)}{=} \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h)\right] \overset{(b)}{=} \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_h))\right] \\ &\overset{(c)}{=} \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\left(\operatorname{gap}(S_h, A_h)) - \frac{1}{2}\operatorname{gap}(S_h, A_h))\right)\right] \\ &\overset{(d)}{\leq} \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\left(\operatorname{gap}(S_h, A_h)) - \epsilon_k(S_h, A_h))\right)\right] \\ &\overset{(e)}{\leq} \ddot{V}_k(s_1) - V^{\pi_k}(s_1).\end{aligned}$$

Here, (a) uses Lemma F.1 and (b) uses the definition of B. Step (c) is just algebra and step (d) uses the assumption on the threshold function. The last step (e) follows from Lemma F.3. □

Proposition 3.3 (Improved surplus clipping bound)**.** *Let the surpluses $E_k(s,a)$ be generated by an optimistic algorithm. Then the instantaneous regret of π_k is bounded as follows:*

$$V^*(s_1) - V^{\pi_k}(s_1) \leq 4\sum_{s,a} w^{\pi_k}(s,a)\operatorname{clip}\left[E_k(s,a) \,\middle|\, \frac{1}{4}\operatorname{gap}(s,a) \vee \epsilon_k(s,a)\right],$$

where $\epsilon_k \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_0^+$ is any clipping threshold function that satisfies

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \epsilon_k(S_h, A_h)\right] \leq \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \operatorname{gap}(S_h, A_h)\right].$$

Proof. Applying Lemma F.4 which ensures scaled optimism of the clipped value function gives

$$V^*(s_1) - V^{\pi_k}(s_1) \leq 2(\ddot{V}_k(s_1) - V^{\pi_k}(s_1)) = 2\sum_{s,a} w^{\pi_k}(s,a)\ddot{E}_k(s,a),$$

where the equality follows from the definition of $\ddot{V}_k(s_1)$ and Lemma F.2. Subtracting $\frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1))$ from both sides gives

$$\frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1)) \leq 2\sum_{s,a} w^{\pi_k}(s,a)\left(\ddot{E}_k(s,a) - \frac{\operatorname{gap}(s,a)}{4}\right)$$

because Lemma F.1 ensures that $\frac{1}{2}(V^*(s_1) - V^{\pi_k}(s_1)) = \frac{1}{2}\sum_{s,a} w^{\pi_k}(s,a)\operatorname{gap}(s,a)$. Reordering terms yields

$$\begin{aligned}V^*(s_1) - V^{\pi_k}(s_1) &\leq 4\sum_{s,a} w^{\pi_k}(s,a)\left(\ddot{E}_k(s,a) - \frac{\operatorname{gap}(s,a)}{4}\right) \\ &= 4\sum_{s,a} w^{\pi_k}(s,a)\left(\operatorname{clip}\left[E_k(s,a) \,\middle|\, \epsilon_k(s,a)\right] - \frac{\operatorname{gap}(s,a)}{4}\right) \\ &\leq 4\sum_{s,a} w^{\pi_k}(s,a)\operatorname{clip}\left[E_k(s,a) \,\middle|\, \epsilon_k(s,a) \vee \frac{\operatorname{gap}(s,a)}{4}\right],\end{aligned}$$

where the final inequality follows from the general properties of the clipping operator, which satisfies

$$\operatorname{clip}[a|b] - c = \begin{cases} a - c \leq a & \text{for } a \geq b \vee c \\ 0 - c \leq 0 & \text{for } a \leq b \\ a - c \leq 0 & \text{for } a \leq c \end{cases} \leq \operatorname{clip}[a|b \vee c].$$

□

F.4 Definition of valid clipping thresholds ϵ_k

Proposition 3.3 establishes a sufficient condition on the clipping thresholds ϵ_k that ensures that the penalized surplus clipping bounds holds. We now discuss several choices for this threshold that satisfy this condition.

Minimum positive gap $\mathrm{gap}_{\min}$: We now make the quick observation that taking $\epsilon_k \equiv \frac{\mathrm{gap}_{\min}}{2H}$ will satisfy the condition of Proposition 3.3, because on the event $\mathcal{B} \equiv \mathcal{A}_{H+1}^c$ there exists at least one positive gap in the sum $\sum_{h=1}^H \mathrm{gap}(S_h, A_h)$, which, by definition, is at least $\mathrm{gap}_{\min}$. This shows that our results already can recover the bounds in prior work, with significantly less effort.

Average gaps: Instead of the minimum gap which was used in existing analyses, we now show that we can also use the marginalized average gap which we will define now. Recall that $B = \min\{h \in [H+1] \colon \mathrm{gap}(S_h, A_h) > 0\}$ is the first time a non-zero gap is encountered. Note that B is a stopping time w.r.t. the filtration $\mathcal{F}_h = \sigma(S_1, A_1, \dots, S_h, A_h)$. Further let

$$\mathcal{B}(s,a) \equiv \{B \le \kappa(s), S_{\kappa(s)} = s, A_{\kappa(s)} = a\} \tag{25}$$

be the event that (s,a) was visited after a non-zero gap in the episode. We now define this clipping threshold

$$\epsilon_k(s,a) \equiv \begin{cases} \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^H \mathrm{gap}(S_h, A_h) \,\middle|\, \mathcal{B}(s,a)\right] & \text{if } \mathbb{P}_{\pi_k}(\mathcal{B}(s,a)) > 0 \\ \infty & \text{otherwise} \end{cases} \tag{26}$$

As the following lemma shows, this is a valid choice which satisfies the condition of Proposition 3.3.

Lemma F.5. *The expected sum of clipping thresholds in Equation (26) over all state-action pairs encountered after a positive gap is at most half the expected total gaps per episode. That is,*

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \epsilon_k(S_h, A_h)\right] \le \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^H \mathrm{gap}(S_h, A_h)\right].$$

Proof. We rewrite the LHS of the inequality to show as $\mathbb{E}_{\pi_k}\left[\sum_{h=1}^H \chi\left(B \le h\right) \epsilon_k(S_h, A_h)\right]$ and from now on consider the random variable $f_h(B, S_h, A_h) = \chi\left(B \le h\right) \epsilon_k(S_h, A_h)$ where $f_h(b, s, a) = \chi\left(b \le h\right) \epsilon_k(s,a)$ is a deterministic function[5]. We will show below that $\mathbb{E}_{\pi_k}\left[f_h(B, S_h, A_h)\right] \le \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \mathrm{gap}(S_h, A_h)\right]$. This is sufficient to prove the statement, because

$$\begin{aligned}
\mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \epsilon_k(S_h, A_h)\right] &= \sum_{h=1}^H \mathbb{E}_{\pi_k}\left[f_h(B, S_h, A_h)\right] \\
&\le \frac{1}{2H}\sum_{h=1}^H \mathbb{E}_{\pi_k}\left[\sum_{h'=B}^H \mathrm{gap}(S_{h'}, A_{h'})\right] \\
&= \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^H \mathrm{gap}(S_h, A_h)\right] = \frac{1}{2}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^H \mathrm{gap}(S_h, A_h)\right].
\end{aligned}$$

To bound the expected value of $f_h(B, S_h, A_h)$, we first write f_h for all triples b, s, a such that $\mathbb{P}_{\pi_k}(B = b, A_h = a, S_h = s) > 0$ as

$$\begin{aligned}
f_h(b,s,a) &\overset{(i)}{=} \chi\left(b \le h\right) \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=1}^H \mathrm{gap}(S_{h'}, A_{h'}) \,\middle|\, B \le h,\ S_h = s, A_h = a\right] \\
&\overset{(ii)}{=} \chi\left(b \le h\right) \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^h \mathrm{gap}(S_{h'}, A_{h'}) \,\middle|\, B \le h,\ S_h = s, A_h = a\right] \\
&\quad + \chi\left(b \le h\right) \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^H \mathrm{gap}(S_{h'}, A_{h'}) \,\middle|\, S_h = s, A_h = a\right],
\end{aligned}$$

where (i) expands the definition of ϵ_k and (ii) decomposes the sum inside the conditional expectation and uses the Markov-property to simplify the conditioning for terms after h. Before taking the

[5] It may still depend on the current policy π_k which is determined by observations in episodes 1 to $k-1$. But, crucially, f_h does not depend on any realization in the k-th episode

expectation of $f_h(B, S_h, A_h)$, we first rewrite the conditional expectation in the first term above, which will be useful later.

$$\begin{aligned}
&\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; B \le h,\ S_h = s, A_h = a\right] \\
&\overset{(i)}{=} \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \chi\left(A_h = a, S_h = s\right) \chi\left(B \le h\right)\right]}{\mathbb{P}_{\pi_k}\left[B \le h,\ S_h = s, A_h = a\right]} \\
&\overset{(ii)}{=} \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \chi\left(A_h = a, S_h = s\right)\right]}{\mathbb{P}_{\pi_k}\left[B \le h,\ S_h = s, A_h = a\right]} \\
&= \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h = s, A_h = a\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\middle|\; S_h = s, A_h = a\right]}.
\end{aligned}$$

Here, step (i) uses the property of conditional expectations with respect to an event with nonzero probability and (ii) follows from the definition of B: When $B > h$, the sum of gaps until h is zero. Consider now the expectation of $f_h(B, S_h, A_h)$

$$\begin{aligned}
&\mathbb{E}_{\pi_k}\left[f_h(B, S_h, A_h)\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right) \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\middle|\; S_h, A_h\right]}\right] \qquad (27) \\
&\quad + \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right) \mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]\right] \qquad (28)
\end{aligned}$$

The term in (28) can be bounded using the tower-property of expectations as

$$\begin{aligned}
&\frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right) \mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]\right] \\
&\le \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]\right] = \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=h+1}^{H} \operatorname{gap}(S_{h'}, A_{h'})\right].
\end{aligned}$$

For the term in (27), we also use the tower-property to rewrite it as

$$\begin{aligned}
&\frac{1}{2H}\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right) \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\middle|\; S_h, A_h\right]}\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right) \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\middle|\; S_h, A_h\right]} \;\middle|\; S_h, A_h\right]\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\chi\left(B \le h\right) \;\middle|\; S_h, A_h\right] \frac{\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]}{\mathbb{P}_{\pi_k}\left[B \le h \;\middle|\; S_h, A_h\right]}\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'}) \;\middle|\; S_h, A_h\right]\right] \\
&= \frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h'=B}^{h} \operatorname{gap}(S_{h'}, A_{h'})\right].
\end{aligned}$$

Summing both terms yields the required upper-bound $\frac{1}{2H}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H}\operatorname{gap}(S_h,A_h)\right]$ on the expectation $\mathbb{E}_{\pi_k}\left[f_h(B,S_h,A_h)\right]$. □

F.5 Policy-dependent regret bound for StrongEuler

We now show how to derive a regret bound for StrongEuler algorithm in Simchowitz and Jamieson [30] that depends on the gaps of the played policies throughout the K episodes.

To build on parts of the analysis in Simchowitz and Jamieson [30], we first define some useful notation analogous to Simchowitz and Jamieson [30] but adapted to our setting:

$$\begin{aligned}\bar{n}_k(s,a) &= \sum_{j=1}^{k} w^{\pi_k}(s,a),\\ M &= (SAH)^3,\\ \mathcal{V}^{\pi}(s,a) &= \mathbb{V}[R(s,a)] + \mathbb{V}_{s'\sim P(\cdot|s,a)}[V^{\pi}(s')],\\ \mathcal{V}_k(s,a) &= \mathcal{V}^{\pi_k}(s,a) \wedge \mathcal{V}^{*}(s,a)\end{aligned}$$

We will use their following results:

Proposition F.6 (Proposition F.1, F.9 and B.4 in Simchowitz and Jamieson [30])**.** *There is a good event $\mathcal{A}^{\mathrm{conc}}$ that holds with probability $1-\delta/2$. In this event,* StrongEuler *is strongly optimistic (as well as optimistic). Further, there is a universal constant $c\geq 1$ so that for all $k\geq 1$, $s\in\mathcal{S}$, $a\in\mathcal{A}$, the surpluses are bounded as*

$$0 \leq \frac{1}{c}E_k(s,a) \leq B_k^{\mathrm{lead}}(s,a) + \sum_{h=\kappa(s)}^{H}\mathbb{E}_{\pi_k}\left[B_k^{\mathrm{fut}}(S_h,A_h)\mid (S_{\kappa(s)},A_{\kappa(s)})=(s,a)\right],$$

where $B^{\mathrm{lead}}, B^{\mathrm{fut}}$ are defined as

$$\begin{aligned}B_k^{\mathrm{lead}}(s,a) &= H \wedge \sqrt{\frac{\mathcal{V}_k(s,a)\log(Mn_k(s,a)/\delta)}{n_k(s,a)}},\\ B_k^{\mathrm{fut}}(s,a) &= H^3 \wedge H^3\left(\sqrt{\frac{S\log(Mn_k(s,a)/\delta)}{n_k(s,a)}} + \frac{S\log(Mn_k(s,a)/\delta)}{n_k(s,a)}\right)^2.\end{aligned}$$

Lemma F.7 (Lemma B.3 in Simchowitz and Jamieson [30])**.** *Let $m\geq 2$, $a_1,\dots,a_m\geq 0$ and $\epsilon\geq 0$. Then* $\operatorname{clip}\left[\sum_{i=1}^{m}a_i\middle|\epsilon\right] \leq 2\sum_{i=1}^{m}\operatorname{clip}\left[a_i\middle|\frac{\epsilon}{2m}\right]$.

Equipped with these results and our improved surplus clipping proposition in Proposition F.6, we can now derive the following bound on the regret of StrongEuler

Lemma F.8. *In event $\mathcal{A}^{\mathrm{conc}}$, the regret of* StrongEuler *is bounded for all $k\geq 1$ as*

$$\begin{aligned}\mathfrak{R}(K) \leq & 8\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\operatorname{clip}\left[cB_k^{\mathrm{lead}}(s,a)\;\middle|\;\frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right]\\ &+ 16\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\operatorname{clip}\left[cB_k^{\mathrm{fut}}(s,a)\;\middle|\;\frac{\breve{\mathrm{gap}}_k(s,a)}{8SA}\right],\end{aligned}$$

with a universal constant $c\geq 1$ and $\breve{\mathrm{gap}}_k(s,a) = \frac{\operatorname{gap}(s,a)}{4}\vee\epsilon_k(s,a)$.

Proof. We now use our improved surplus clipping result from Proposition 3.3 as a starting point to bound the instantaneous regret of StrongEuler in the kth episode as

$$V^{*}(s_1) - V^{\pi_k}(s_1) \leq 4\sum_{s,a} w^{\pi_k}(s,a)\operatorname{clip}\left[E_k(s,a)\,\middle|\,\breve{\mathrm{gap}}_k(s,a)\right]. \tag{29}$$

Next, we write the bound on the surpluses from Proposition F.6 as

$$E_k(s,a) \le cB_k^{\mathrm{lead}}(s,a) + c\sum_{s',a'} \chi\left(\kappa(s') \ge \kappa(s)\right) \mathbb{P}^{\pi_k}\left[S_{\kappa(s')} = s', A_{\kappa(s')} = a' \mid (S_{\kappa(s)}, A_{\kappa(s)}) = (s,a)\right] B_k^{\mathrm{fut}}(s',a')$$

and plugging it in (29) and applying Lemma F.7 gives

$$V^*(s_1) - V^{\pi_k}(s_1) \le 8\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{lead}}(s,a) \;\middle|\; \frac{\check{\mathrm{gap}}_k(s,a)}{4}\right] + 16\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{fut}}(s,a) \;\middle|\; \frac{\check{\mathrm{gap}}_k(s,a)}{8SA}\right].$$

The statement to show follows now by summing over $k \in [K]$. The form of the second term in the previous display follows from the inequality

$$\sum_{s,a} w^{\pi_k}(s,a)\chi\left(\kappa(s') \ge \kappa(s)\right) \mathbb{P}^{\pi_k}\left[S_{\kappa(s')} = s', A_{\kappa(s')} = a' \mid (S_{\kappa(s)}, A_{\kappa(s)}) = (s,a)\right]$$
$$\le \sum_{s,a} w^{\pi_k}(s,a)\mathbb{P}^{\pi_k}\left[S_{\kappa(s')} = s', A_{\kappa(s')} = a' \mid (S_{\kappa(s)}, A_{\kappa(s)}) = (s,a)\right] = w^{\pi_k}(s',a').$$

□

We note that if $\pi_k \equiv \hat{\pi}$ for any $\hat{\pi} \in \Pi^*$ then $V^*(s_1) - V^{\pi_k}(s_1) = 0$, and WLOG we can disregard such terms in the total regret.

The next step is to relate $\bar{n}_k(s,a)$ to $n_k(s,a)$ via the following lemma.

Lemma F.9 (Lemma B.7 in Simchowitz and Jamieson [30])**.** *Define the event* $\mathcal{A}^{\mathrm{samp}}$

$$\mathcal{A}^{\mathrm{samp}} = \left\{\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall k \ge \tau(s,a)\colon n_k(s,a) \ge \frac{\bar{n}_k(s,a)}{4}\right\},$$

where $\tau(s,a) = \inf\{k : \bar{n}_k(s,a) \ge H_{\mathrm{samp}}\}$ *and* $H_{\mathrm{samp}} = c'\log(M/\delta)$ *for a universal constant* c'*. Then event* $\mathcal{A}^{\mathrm{samp}}$ *holds with probability* $1 - \delta/2$*.*

Proof. This can be proved analogously to Lemma B.7 in Simchowitz and Jamieson [30] and Lemma 6 in Dann et al. [9] with the difference that in our case, there can only be at most one observation of (s,a) per episode for each (s,a) due to our layered assumption. Thus, there is no need to sum over observations accumulated for each $h \in [H]$ and our $H_{\mathrm{samp}} = O(\log(H))$ as opposed to $O(H\log(H))$. □

Lemma F.10. *Let* $f_{s,a}\colon \mathbb{N} \to \mathbb{R}$ *be non-increasing with* $\sup_u f_{s,a}(u) \le \hat{f} < \infty$ *for all* $s,a \in \mathcal{S}\times\mathcal{A}$*. Then on event* $\mathcal{A}^{\mathrm{samp}}$ *in Lemma F.9, we have*

$$\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) \le SA\hat{f}H_{\mathrm{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4).$$

Proof.

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) \\
&= \sum_{s,a}\sum_{k=1}^{\tau(s,a)-1} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(n_k(s,a)) \\
&\le \sum_{s,a}\left(\sum_{k=1}^{\tau(s,a)-1} w^{\pi_k}(s,a)\right)\hat{f} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4) \\
&= \sum_{s,a} n_{\tau(s,a)}(s,a)\hat{f} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4) \\
&\le SAH_{\mathrm{samp}}\hat{f} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a) f_{s,a}(\bar{n}_k(s,a)/4).
\end{aligned}
$$

□

Theorem F.11 (Regret Bound for StrongEuler)**.** *With probability at least $1-\delta$, the regret of* StrongEuler *is bounded for all number of episodes $K \in \mathbb{N}$ as*

$$
\begin{aligned}
\mathfrak{R}(K) \lesssim \sum_{s,a} \min_{t\in[K_{(s,a)}]} \Bigg\{ & \frac{\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta, t, \breve{\mathrm{gap}}_t(s,a))}{\breve{\mathrm{gap}}_t(s,a)} \\
& + \sqrt{(K_{(s,a)} - t)\mathcal{LOG}(M/\delta, K_{(s,a)}, \breve{\mathrm{gap}}_{K_{(s,a)}}(s,a))} \Bigg\} \\
& + \sum_{s,a} SH^3 \log\frac{MK}{\delta} \min\left\{\log\frac{MK}{\delta}, \log\frac{MH}{\breve{\mathrm{gap}}_{\min}(s,a)}\right\} \\
& + SAH^3(S \vee H)\log\frac{M}{\delta}.
\end{aligned}
$$

Here, $K_{(s,a)}$ is the last round during which a policy π was played such that $w^{\pi}(s,a) > 0$, $\breve{\mathrm{gap}}_t(s,a) = \mathrm{gap}(s,a) \vee \epsilon_t(s,a)$, $\breve{\mathrm{gap}}_{\min}(s,a) = \min_{k\in[K]\colon \breve{\mathrm{gap}}_k(s,a)>0} \breve{\mathrm{gap}}_k(s,a)$ is the smallest gap encountered for each (s,a), and $\mathcal{LOG}(M/\delta, t, \breve{\mathrm{gap}}_t(s,a)) = \log\left(\frac{M}{\delta}\right)\log\left(t \wedge 1 + \frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\mathrm{gap}}_t(s,a)^2}\right)$.

Proof. We here consider the event $\mathcal{A}^{\mathrm{conc}} \cap \mathcal{A}^{\mathrm{samp}}$ which has probability at least $1-\delta$ by Proposition F.6 and Lemma F.9. We now start with the regret bound in Lemma F.8 and bound the two terms individually in the following:

Bounding the B^{lead} term We have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{lead}}(s,a) \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right] \\
&\overset{(i)}{\le} SAHH_{\mathrm{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a)\,\mathrm{clip}\left[c\sqrt{\frac{4\mathcal{V}_k(s,a)\log(M\bar{n}_k(s,a)/4\delta)}{\bar{n}_k(s,a)}} \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right] \\
&\overset{(ii)}{\le} SAHH_{\mathrm{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K_{(s,a)}} w^{\pi_k}(s,a)\mathrm{clip}\left[2c\sqrt{\mathcal{V}^*(s,a)\log\frac{M}{\delta}}\sqrt{\frac{\log(\bar{n}_k(s,a))}{\bar{n}_k(s,a)}} \,\middle|\, \frac{\breve{\mathrm{gap}}_k(s,a)}{4}\right],
\end{aligned}
\tag{30}
$$

where step (i) applies Lemma F.10 and (ii) follows from the definition of $\mathcal{V}_k(s,a)$, the definition of $K_{(s,a)}$ and

$$\log\left(\frac{M\bar{n}_k(s,a)}{4\delta}\right) = \log\left(\frac{M}{4\delta}\right) + \log\left(\bar{n}_k(s,a)\right)$$
$$\leq \left(\log\left(\frac{M}{4\delta}\right)+1\right)\log(\bar{n}_k(s,a)) = \log\left(\frac{Me}{4\delta}\right)\log(\bar{n}_k(s,a)) \leq \log(M/\delta)\log(\bar{n}_k(s,a)).$$

We now apply our optimization lemma (Lemma F.16) with $x_k = w^{\pi_k}(s,a)$, $v_k = 2c\sqrt{\mathcal{V}^*(s,a)\log(M/\delta)}$, and $\epsilon_k = \frac{\breve{\text{gap}}_k(s,a)}{4v_k}$ to bound each (s,a)-term in (30) for any $t \in [K]$ as

$$4\frac{v_t}{\epsilon_t}\log\left(t\wedge 1+\frac{1}{\epsilon_t^2}\right) + 4v_t\sqrt{\log\left(K\wedge 1+\frac{1}{\epsilon_K^2}(K-t)\right)}$$
$$= \frac{32c^2\mathcal{V}^*(s,a)\log\left(\frac{M}{\delta}\right)\log\left(t\wedge 1+\frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\text{gap}}_t(s,a)^2}\right)}{\breve{\text{gap}}_t(s,a)}$$
$$+8c\sqrt{(K-t)\mathcal{V}^*(s,a)\log\left(\frac{M}{\delta}\right)\log\left(K\wedge 1+\frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\text{gap}}_K(s,a)^2}\right)}.$$

Let $\mathcal{LOG}(M/\delta, t, \breve{\text{gap}}_t(s,a)) = \log\left(\frac{M}{\delta}\right)\log\left(t\wedge 1+\frac{16\mathcal{V}^*(s,a)\log(M/\delta)}{\breve{\text{gap}}_t(s,a)^2}\right)$. We have

$$\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a)\,\text{clip}\left[2c\sqrt{\mathcal{V}^*(s,a)\log(M/4\delta)}\sqrt{\frac{\log(\bar{n}_k(s,a))}{\bar{n}_k(s,a)}}\;\middle|\;\frac{\breve{\text{gap}}_k(s,a)}{4}\right]$$
$$\leq \frac{32c^2\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta,t,\breve{\text{gap}}_t(s,a))}{\breve{\text{gap}}_t(s,a)} + 8c\sqrt{(K-t)\mathcal{LOG}(M/\delta,K,\breve{\text{gap}}_K(s,a))}.$$

Plugging this bound back in (30) gives

$$\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\text{clip}\left[cB_k^{\text{lead}}(s,a)\;\middle|\;\frac{\breve{\text{gap}}_k(s,a)}{4}\right]$$
$$\lesssim SAH\log\frac{M}{\delta}$$
$$+\sum_{s,a}\min_{t\in[K_{(s,a)}]}\left\{\frac{\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta,t,\breve{\text{gap}}_t(s,a))}{\breve{\text{gap}}_t(s,a)} + \sqrt{(K_{(s,a)}-t)\mathcal{LOG}(M/\delta,K,\breve{\text{gap}}_{K_{(s,a)}}(s,a))}\right\}$$

where $\lesssim$ only ignores absolute constant factors.

Bounding the B^{fut} term Consider the second term in Lemma F.8 and event $\mathcal{A}^{\text{conc}}\cap\mathcal{A}^{\text{samp}}$. Then by Lemma F.10

$$\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\text{clip}\left[cB_k^{\text{fut}}(s,a)\;\middle|\;\frac{\breve{\text{gap}}_k(s,a)}{8SA}\right]$$
$$\leq SAH^3H_{\text{samp}} + \sum_{s,a}\sum_{k=\tau(s,a)}^{K} w^{\pi_k}(s,a)f_{s,a}(\bar{n}_k(s,a))$$

where $f_{s,a}$ is

$$f_{s,a}(\bar{n}_k(s,a)) = \text{clip}\left[2cH^3\wedge 2cH^3\left(\sqrt{\frac{S\log(M\bar{n}_k(s,a)/\delta)}{\bar{n}_k(s,a)}}+\frac{S\log(M\bar{n}_k(s,a)/\delta)}{\bar{n}_k(s,a)}\right)^2\;\middle|\;\frac{\breve{\text{gap}}_k(s,a)}{4}\right].$$

We now apply Lemma C.1 by Simchowitz and Jamieson [30] which gives

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{fut}}(s,a)\;\middle|\;\frac{\breve{\mathrm{gap}}_k(s,a)}{8SA}\right]\\
&\leq SAH^3H_{\mathrm{samp}} + \sum_{s,a} Hf_{s,a}(H) + \sum_{s,a}\int_{H}^{\bar{n}_K(s,a)} f_{s,a}(u)du\\
&\leq SAH^4c'\log(M/\delta) + \sum_{s,a}\int_{H}^{\bar{n}_K(s,a)} f_{s,a}(u)du.
\end{aligned}$$

The remaining integral term is bounded with Lemma B.9 (b) by Simchowitz and Jamieson [30] with $C' = S, C = H^3$ and $\epsilon = \breve{\mathrm{gap}}_{\min}(s,a) = \min_{k\in[K_{(s,a)}]:\,\breve{\mathrm{gap}}_k(s,a)>0}\breve{\mathrm{gap}}_k(s,a)$ as follows.

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{s,a} w^{\pi_k}(s,a)\,\mathrm{clip}\left[cB_k^{\mathrm{fut}}(s,a)\;\middle|\;\frac{\breve{\mathrm{gap}}_k(s,a)}{8SA}\right]\\
&\lesssim SAH^4\log\frac{M}{\delta} + \sum_{s,a}\left(SH^3\log\frac{M}{\delta} + SH^3\log\frac{MK}{\delta}\min\left\{\log\frac{MK}{\delta}, \log\frac{MH}{\breve{\mathrm{gap}}_{\min}(s,a)}\right\}\right)\\
&\lesssim SAH^3(S\vee H)\log\frac{M}{\delta} + \sum_{s,a} SH^3\log\frac{MK}{\delta}\min\left\{\log\frac{MK}{\delta}, \log\frac{MH}{\breve{\mathrm{gap}}_{\min}(s,a)}\right\}
\end{aligned}$$

□

Comparing with the bound in Simchowitz and Jamieson [30]. We now proceed to compare our bound directly to the one stated in Corollary B.1 [30]. We will ignore the factors with only poly-logarithmic dependence on gaps as they are are common between both bounds. We now recall the regret bound presented in Corollary B.1, modulo said factors:

$$\mathfrak{R}(K) \leq O\left(\sum_{(s,a)\in\mathcal{Z}_{sub}}\frac{\alpha H\mathcal{V}^*(s,a)}{\mathrm{gap}(s,a)}\mathcal{LOG}(M/\delta,K,\mathrm{gap}(s,a)) + |\mathcal{Z}_{opt}|\frac{H\mathcal{V}^*}{\mathrm{gap}_{\min}}\mathcal{LOG}(M/\delta,K,\mathrm{gap}_{\min})\right),$$

where $\mathcal{V}^* = \max_{(s,a)}\mathcal{V}(s,a)$, $\mathcal{Z}_{opt}$ is the set on which $\mathrm{gap}(s,\pi^*(s)) = 0$, i.e., the set of state-action pairs assigned to π^* according to the Bellman optimality condition, and $\mathcal{Z}_{sub}$ is the complement of $\mathcal{Z}_{opt}$. If we take $t = K$ in Theorem F.11, we have the following upper bound:

$$\mathfrak{R}(K) \leq O\left(\sum_{(s,a)\in\mathcal{Z}_{sub}}\frac{\mathcal{V}^*(s,a)\mathcal{LOG}(M/\delta,K,\mathrm{gap}(s,a))}{\mathrm{gap}(s,a)} + \frac{H\mathcal{V}^*|\mathcal{S}_{opt}|\mathcal{LOG}(M/\delta,K,\mathrm{gap}_{\min})}{\min_{k,s,a}\epsilon_k(s,a)}\right),$$

where $\mathcal{S}_{opt}$ is the set of all states for $s\in\mathcal{S}$ for which $\mathrm{gap}(s,\pi^*(s)) = 0$ and there exists at least one state s' with $\kappa(s') < s$ for which $\mathrm{gap}(s',\pi^*(s)) > 0$. We note that this set is no larger than the set $\mathcal{Z}_{opt}$ and further that even the smallest $\epsilon_k(s,a)$ can still be much larger than $\mathrm{gap}_{\min}$, as it is the conditional average of the gaps. In particular, this leads to an arbitrary improvement in our example in Figure 1 and an improvement of SA in the example in Figure 7.

F.6 Nearly tight bounds for deterministic transition MDPs

We recall that for deterministic MDPs, $\epsilon_k(s,a) = \frac{V^*(s_1)-V^{\pi_k}(s_1)}{2H}, \forall a$ and the definition of the set $\Pi_{s,a}$:

$$\Pi_{s,a} \equiv \{\pi\in\Pi : s^\pi_{\kappa(s)} = s, a^\pi_{\kappa(s)} = a, \exists\, h\leq\kappa(s), \mathrm{gap}(s^\pi_h,a^\pi_h) > 0\}.$$

We note that $\mathcal{V}(s,a)\leq 1$ as this is just the variance of the reward at (s,a). Theorem F.11 immediately yields the following regret bound by taking $t = K$.

Corollary F.12 (Explicit bound from (5))**.** *Suppose the transition kernel of the MDP consists only of point-masses. Then with probability* $1-\delta$*,* `StrongEuler`*'s regret is bounded as*

$$\mathfrak{R}(K) \leq O\Bigg(\sum_{(s,a):\Pi_{s,a}\neq\emptyset} \frac{H\mathcal{LOG}\left(M/\delta, K, \overline{\text{gap}}(s,a)\right)}{v^* - v^*_{s,a}} + \sum_{s,a} SH^3 \log\frac{MK}{\delta} \min\left\{\log\frac{MK}{\delta}, \log\frac{MH}{\overline{\text{gap}}(s,a)}\right\} + SAH^3(S\vee H)\log\frac{M}{\delta}\Bigg),$$

where $v^*_{s,a} = \max_{\pi\in\Pi_{s,a}} v^\pi$.

We now compare the above bound with the one in [30] again. For simplicity we are going to take K to be the smaller of the two quantities in the logarithm. To compare the bounds, we compare $\sum_{(s,a):\Pi_{s,a}\neq\emptyset} \frac{H(\log(KM/\delta)))}{v^*-v^*_{(s,a)}}$ to $\sum_{(s,a)\in\mathcal{Z}_{sub}} \frac{\alpha H\log(KM/\delta)}{\text{gap}(s,a)} + \frac{|\mathcal{Z}_{opt}|H}{\text{gap}_{\min}}$. Recall that $\alpha\in[0,1]$ is defined as the smallest value such that for all $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ it holds that

$$P(s'|s,a) - P(s'|s,\pi^*(s)) \leq \alpha P(s'|s,a).$$

For any deterministic transition MDP with more than one layer and one sub-optimal action it holds that $\alpha = 1$. We will compare $V^*(s_1) - V^{\pi^*_{(s,a)}}(s_1)$ to $\text{gap}(s,a) = Q^*(s,\pi^*(s)) - Q^*(s,a)$. This comparison is easy as by Lemma F.1 we can write

$$V^*(s_1) - V^{\pi^*_{(s,a)}}(s_1) = \sum_{(s',a')\in\pi^*_{(s,a)}} w_{\pi^*_{(s,a)}}(s',a')\,\text{gap}(s',a') = \sum_{(s',a')\in\pi^*_{(s,a)}} \text{gap}(s',a') \geq \text{gap}(s,a).$$

Hence, our bound in the worst case matches the one in Simchowitz and Jamieson [30] and can actually be significantly better. We would further like to remark that we have essentially solved all of the issues presented in the example MDP in Figure 1. In particular we do not pay any gap-dependent factors for states which are only visited by π^*, we do not pay a $\text{gap}_{\min}$ factor for any state and we never pay any factors for distinguishing between two suboptimal policies. Finally, we compare this bound to the lower bound derived Theorem 4.5 only with respect to number of episodes and gaps. Let $\mathcal{S}^*$ be the set of all states in the support of an optimal policy

$$\sum_{(s,a)\in\mathcal{S}\setminus\mathcal{S}^*\times\mathcal{A}} \frac{\log(K)}{H(v^* - v^{\pi^*_{(s,a)}}(s_1))} \leq \mathfrak{R}(K) \leq \sum_{(s,a):\Pi_{s,a}\neq\emptyset} \frac{H\log(K)}{v^* - v^*_{s,a}}.$$

The difference between the two bounds, outside of an extra H^2 factor, is in the sets $\mathcal{S}^*$ and the set $\{s,a:\Pi_{s,a}=\emptyset\}$. We note that $\{s,a:\Pi_{s,a}=\emptyset\}\subseteq\mathcal{S}^*$. Unfortunately there are examples in which $\{s,a:\Pi_{s,a}=\emptyset\}$ is $O(1)$ and $\mathcal{S}^* = \Omega(S)$ leading to a discrepancy between the upper and lower bounds of the order $\Omega(S)$. As we show in Theorem E.11 this discrepancy can not really be avoided by optimistic algorithms.

F.7 Tighter bounds for unique optimal policy.

If we further assume that the optimal policy is unique on its support, then we can show StrongEuler will only incur regret on sub-optimal state-action pairs. This matches the information theoretic lower bound up to horizon factors. We begin by showing a different type of upper bound on the expected gaps by the surpluses. Define the set $\beta_k = range(B)$ where B is the r.v. which is the stopping time with respect to π_k. For any π^*, define the set

$$\mathcal{O}_k(\pi^*) = \bigcup_{s_b\in\beta_k} \{(s,a)\in\mathcal{S}\times\mathcal{A} : \mathbb{P}_{\pi^*}((S_h,A_h)=(s,a)|S_{\kappa(s_b)}=s_b) \geq \mathbb{P}_{\pi_k}((S_h,A_h)=(s,a)|S_{\kappa(s_b)}=s_b)\}.$$

This set has the following intuitive definition – whenever $\mathcal{A}_B$ occurs we restrict our attention to the MDP with initial state S_B. On this restricted MDP, $\mathcal{O}_k$ is the set of state-action pairs which have greater probability to be visited by the optimal π^* than by π_k.

Lemma F.13. *Assume strong optimism and greedy $\bar{V}_k$, i.e., $\bar{V}_k(s) \geq \max_a \bar{Q}_k(s,a)$ for all $s \in \mathcal{S}$. Then there exists an optimal π^* for which*

$$\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \text{gap}(S_h, A_h)\right] \leq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \chi(S_h, A_h \notin \mathcal{O}_k(\pi^*)) E_k(S_h, A_h)\right].$$

Proof. One can write the optimistic value function for any s and π as follows

$$\begin{aligned}\bar{V}^{\pi}(s) &= \mathbb{E}_{\pi}\left[\sum_{h=\kappa(s)}^{H} E_k(S_h, A_h) + r(S_h, A_h) \big| S_{\kappa(s)} = s\right] \\ &= E_k(s, \pi(s)) + r(s, \pi(s)) + \langle P(\cdot|s, \pi(s)), \bar{V}^{\pi}\rangle.\end{aligned}$$

By backwards induction on H we show that for any s, $\kappa(s) \leq H$ $\bar{V}^{\pi} \leq \bar{V}_k$. The base case holds from the fact that on all $s : \kappa(s) = H$, $\bar{V}_k(s)$ is just the largest optimistic reward over all actions at s. For the induction step it holds that

$$\begin{aligned}\bar{V}^{\pi}(s) &= E_k(s, \pi(s)) + r(s, \pi(s)) + \langle P(\cdot|s, \pi(s)), \bar{V}^{\pi}\rangle \\ &\leq E_k(s, \pi(s)) + r(s, \pi(s)) + \langle P(\cdot|s, \pi(s)), \bar{V}_k\rangle \\ &= \bar{Q}_k(s, \pi(s)) \leq \bar{V}_k(s),\end{aligned}$$

where the first inequality holds from the induction hypothesis and the second inequality holds by definition of the value function. We now have

$$\begin{aligned}\mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \text{gap}(S_h, A_h)\right] &= \mathbb{E}_{\pi_k}\left[V^*(S_B) - V_k(S_B)\right] \\ &\leq \mathbb{E}_{\pi_k}\left[\bar{V}_k(S_B) - V_k(S_B)\right] - \mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right].\end{aligned}$$

Let us focus on the term $\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right]$

$$\begin{aligned}\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right] &= \mathbb{E}_{\pi_k}\left[\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B) | S_B\right]\right] \\ &= \mathbb{E}_{\pi_k}\left[\sum_s \frac{\bar{V}^*(s) - V^*(s)}{\mathbb{P}_{\pi_k}(S_B = s)} \chi(S_B = s)\right] \\ &= \mathbb{E}_{\pi_k}\left[\sum_s \frac{\mathbb{E}_{\pi^*}\left[\sum_{h=\kappa(s)}^{H} E_k(S_h, A_h) | S_{\kappa(s)} = s\right]}{\mathbb{P}_{\pi_k}(S_B = s)} \chi(S_B = s)\right].\end{aligned}$$

We can similarly expand the term $\mathbb{E}_{\pi_k}\left[\bar{V}_k(S_B) - V_k(S_B)\right]$. By the definition of $\mathcal{O}_k(\pi^*)$ it holds that for any $h \geq \kappa(s)$

$$\begin{aligned}&\mathbb{E}_{\pi_k}\left[E_k(S_h, A_h) | S_{\kappa(s)} = s\right] - \mathbb{E}_{\pi^*}\left[E_k(S_h, A_h) | S_{\kappa(s)} = s\right] \\ &\qquad \leq \mathbb{E}_{\pi_k}\left[\chi(S_h, A_h \notin \mathcal{O}_k(\pi^*)) E_k(S_h, A_h) | S_{\kappa(s)} = s\right].\end{aligned}$$

This implies

$$\begin{aligned}\mathbb{E}_{\pi_k}\left[\bar{V}^*(S_B) - V^*(S_B)\right] &\leq \mathbb{E}_{\pi_k}\left[\sum_s \frac{\mathbb{E}_{\pi_k}\left[\sum_{h=\kappa(s)}^{H} \chi(S_h, A_h \notin \mathcal{O}_k(\pi^*)) E_k(S_h, A_h) | S_{\kappa(s)} = s\right]}{\mathbb{P}_{\pi_k}(S_B = s)} \chi(S_B = s)\right] \\ &= \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \chi(S_h, A_h \notin \mathcal{O}_k(\pi^*)) E_k(S_h, A_h)\right].\end{aligned}$$

□

We next show a version of Lemma F.3 which takes into account the set $\mathcal{O}_k(\pi^*)$.

Lemma F.14. *With the same assumptions as in Lemma F.13, there exists an optimal π^* for which*

$$\ddot{V}_k(s_1) - V_k(s_1) \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_h) - \sum_{h=B}^{H} \chi(S_h, A_h \notin \mathcal{O}_k(\pi^*))\epsilon_k(S_h, A_h)\right],$$

where ϵ_k is arbitrary.

Proof. Since $\ddot{E}_k$ is non-negative on all state-action pairs we have

$$\begin{aligned}
\ddot{V}_k(s_1) - V^{\pi_k}(s_1) &= \mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \ddot{E}_k(S_h, A_h)\right] \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \ddot{E}_k(S_h, A_h)\right] \\
&\geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \chi\left((S_h, A_h) \notin \mathcal{O}_k\right) \ddot{E}_k(S_h, A_h)\right] \\
&\geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \chi((S_h, A_h) \notin \mathcal{O}_k) E_k(S_h, A_h)\right] - \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \chi((S_h, A_h) \notin \mathcal{O}_k)\epsilon_k(S_h, A_h)\right] \\
&\geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_H)\right] - \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \chi((S_h, A_h) \notin \mathcal{O}_k)\epsilon_k(S_h, A_h)\right],
\end{aligned}$$

where the second to last inequality follows from the definition of $\ddot{E}_k$ and the last inequality follows from Lemma F.13. □

Next, we define $\bar{\epsilon}_k$ in the following way. Let

$$\bar{\epsilon}_k(s, a) \equiv \begin{cases} \epsilon_k(s, a) & \text{if } (s, a) \notin \mathcal{O}_k(\pi^*) \\ \infty & \text{otherwise,} \end{cases} \tag{31}$$

where ϵ_k is the clipping function defined in Equation 26. Lemma F.14 now implies that

$$\ddot{V}_k(s_1) - V_k(s_1) \geq \mathbb{E}_{\pi_k}\left[\sum_{h=B}^{H} \operatorname{gap}(S_h, A_h) - \sum_{h=B}^{H} \bar{\epsilon}_k(S_h, A_h)\right].$$

This is sufficient to argue Lemma F.8 with $\breve{\text{gap}}_k(s, a) = \frac{\operatorname{gap}(s,a)}{4} \vee \bar{\epsilon}_k(s, a)$ and hence arrive at a version of Corollary F.12 which uses $\bar{\epsilon}_k$ as the clipping thresholds. Let us now argue that $\bar{\epsilon}_k(s, a) = \infty$ for all $(s, a) \in \pi^*$ whenever π^* is the unique optimal policy for the deterministic MDP. To do so consider $(s, a) \in \pi^*$ and $\pi_k \neq \pi^*$. Since the MDP is deterministic, β_k is a singleton and is the the first state s_b at which π_k differs from π^*. We now observe that if $\kappa(s) < \kappa(s_b)$, this implies $\epsilon_k(s, a) = \infty$ as $\mathcal{B}(s, a)$ does not occur. Further, the conditional probabilities $\mathbb{P}_{\pi^*}((S_h, A_h) = (s, a) | S_{\kappa(s_b)} = s_b)$ and $\mathbb{P}_{\pi_k}((S_h, A_h) = (s, a) | S_{\kappa(s_b)} = s_b)$ are both equal to 1 if $\kappa(s) > \kappa(s_b)$ and so $(s, a) \in \mathcal{O}_k(\pi^*)$ which implies $\bar{\epsilon}_k(s, a) = \infty$. Thus we can clip all gaps at $(s, a) \in \pi^*$ to infinity and they will never appear in the regret bound. With the notation from Corollary F.12 we have the following tighter bound.

Corollary F.15. *Suppose the transition kernel of the MDP consists only of point-masses and there exists a unique optimal π^*. Then with probability $1 - \delta$, `StrongEuler`'s regret is bounded as*

$$\begin{aligned}
\mathfrak{R}(K) \leq O\Bigg(& \sum_{(s,a) \notin \pi^*} \frac{\mathcal{LOG}\left(M/\delta, K, \overline{\operatorname{gap}}(s, a)\right)}{\overline{\operatorname{gap}}(s, a)} \\
& + \sum_{(s,a) \notin \pi^*} SH^3 \log\frac{MK}{\delta} \min\left\{\log\frac{MK}{\delta}, \log\frac{MH}{\overline{\operatorname{gap}}(s, a)}\right\} \\
& + SAH^3(S \vee H) \log\frac{M}{\delta}\Bigg).
\end{aligned}$$

Comparing terms which depend polynomially on $1/\overline{\operatorname{gap}}$ to the information theoretic lower bound in Theorem 4.5 we observe only a multiplicative difference of H^2.

F.8 Alternative to integration lemmas

The following lemma is an alternative to the integration lemmas when bounding the sum of the clipped surpluses and in some cases allows us to save additional factors of H.

Lemma F.16. *Consider the following optimization problem*

$$\begin{aligned} \underset{x_1,\dots,x_K}{\textit{maximize}} \quad & \sum_{k=1}^{K} \frac{v_k x_k \sqrt{\log(\sum_{j=1}^{k} x_j)}}{\sqrt{\sum_{j=1}^{k} x_j}} \\ \textit{s.t.} \quad & 1 \le x_1, \quad 0 \le x_k \le 1, \quad \frac{\sqrt{\log(\sum_{j=1}^{k} x_j)}}{\sqrt{\sum_{j=1}^{k} x_j}} \ge \epsilon_k \qquad \forall\, k \in [K], \end{aligned} \tag{32}$$

with $(v_i)_{i\in[K]} \in \mathbb{R}_+^K$ *and* $(\epsilon_i)_{i\in[K]} \in \mathbb{R}_+^K$. *Then the optimal value of Problem 32 is bounded for any* $t \in [K]$ *as*

$$4\frac{\bar{v}_t}{\epsilon_t}\log\left(t \wedge 1 + \frac{1}{\epsilon_t^2}\right) + 4v_t^*\sqrt{\log\left(K \wedge 1 + \frac{1}{\epsilon_K^2}\right)(K-t)}, \tag{33}$$

where $\bar{v}_t = \max_{k\in[t]} v_k$ *and* $v_t^* = \max_{K\ge k\ge t} v_k$.

Proof. Denote by $X_k = \sum_{t=1}^{k} x_t$ the cumulative sum of x_t. The proof consists of splitting the objective of (32) into two terms:

$$\sum_{k=1}^{t} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} + \sum_{k=t+1}^{K} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} \tag{34}$$

and bounding each by the corresponding one in (33) respectively.

Before doing so, we derive the following bound on the sum of $\frac{x_k}{\sqrt{X_k}}$ terms:

$$\sum_{k=m+1}^{M} \frac{x_k}{\sqrt{X_k}} = \sum_{k=m+1}^{M} \frac{X_k - X_{k-1}}{\sqrt{X_k}} \le \int_{X_m}^{X_M} \frac{1}{\sqrt{x}}\,dx = 2(\sqrt{X_M} - \sqrt{X_m}), \tag{35}$$

where the inequality is due to X_k being non-decreasing.

Consider now each term in the objective in (34) separately.

Summands up to t: Since X_k is non-decreasing, we can bound

$$\begin{aligned} \sum_{k=1}^{t} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} \le \bar{v}_t \sqrt{\log(X_t)} \sum_{k=1}^{t} \frac{x_k}{\sqrt{X_k}} &\overset{(i)}{\le} 2\bar{v}_t \sqrt{\log(X_t)}\sqrt{X_t} \\ &\overset{(ii)}{\le} 2\frac{\bar{v}_t}{\epsilon_t}\log(X_t), \end{aligned}$$

where (i) follows from (35) using the convention $X_0 = 0$ and (ii) from the optimization constraint $\sqrt{\log(X_t)} \ge \epsilon_t\sqrt{X_t}$. It remains to bound $\log(X_t)$ by $2\log\left(t \wedge 1 + \frac{1}{\epsilon_t^2}\right)$. Since all increments x_j are at most 1, the bound $\log(X_t) \le \log(t)$ holds.

We claim the following:

Claim F.17. *For any* x *s.t.* $\log(x) \le \log(\log(x)/a)$ *it holds that* $\log(x) \le 2\log(1 + 1/a)$.

Proof. First, we note that if $0 < x \le e$, then $\log(\log(x)) < 0$ and thus the assumption of the claim implies $\log(x) \le \log(1/a)$. Next, assume that $x > e$. Then we have $\frac{\log(\log(x))}{\log(x)} \le 1/e$, which together with the assumption of the claim implies $\log(x) \le 1/e\log(x) + \log(1/a)$ or equivalently $\log(x) \le \frac{e}{e-1}\log(1/a)$. Noting that $e/(e-1) \le 2$ completes the proof. □

The constraints of the problem enforce $\sqrt{X_k} \le \frac{\sqrt{\log(X_k)}}{\epsilon_k}$, which implies after squaring and taking the log: $\log(X_k) \le \log(\log(X_k)/\epsilon_k^2)$. Thus, using Claim F.17 yields:

$$\log(X_k) \le 2\log(k \wedge 1 + 1/\epsilon_k^2). \tag{36}$$

Summands larger than t: Let $v_t^* = \max_{k:\, t<k\le K} v_k$. For this term, we have

$$\begin{aligned}
\sum_{k=t+1}^{K} \frac{v_k x_k \sqrt{\log(X_k)}}{\sqrt{X_k}} &\overset{(36)}{\le} 2v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)} \sum_{k=t+1}^{K} \frac{x_k}{\sqrt{X_k}} \\
&\overset{(35)}{\le} 4v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)}(\sqrt{X_K} - \sqrt{X_t}) \\
&\le 4v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)}(\sqrt{X_K - X_t}) \\
&\le 4v_t^* \sqrt{\log(K \wedge 1 + 1/\epsilon_K^2)}(\sqrt{K - t}),
\end{aligned}$$

where we first bounded $\log(X_k) \le \log(X_K)$, because X_k is non-decreasing, and used the upper bound on $\log(X_K)$. Then we applied (35) and finally used $0 \le x_k \le 1$. □