A Fully Problem-Dependent Regret Lower Bound for Finite-Horizon MDPs

Andrea Tirinzoni [1] **Matteo Pirotta** [2] **Alessandro Lazaric** [2]

Abstract

[1] We derive a novel asymptotic problem-dependent lower-bound for regret minimization in finite-horizon tabular Markov Decision Processes (MDPs). While, similar to prior work (e.g., for ergodic MDPs), the lower-bound is the solution to an optimization problem, our derivation reveals the need for an additional constraint on the visitation distribution over state-action pairs that explicitly accounts for the dynamics of the MDP. We provide a characterization of our lower-bound through a series of examples illustrating how different MDPs may have significantly different complexity. 1) We first consider a "difficult" MDP instance, where the novel constraint based on the dynamics leads to a larger lower-bound (i.e., a larger regret) compared to the classical analysis. 2) We then show that our lower-bound recovers results previously derived for specific MDP instances. 3) Finally, we show that, in certain "simple" MDPs, the lower bound is considerably smaller than in the general case and it does not scale with the minimum action gap at all. We show that this last result is attainable (up to $poly(H)$ terms, where H is the horizon) by providing a regret upper-bound based on policy gaps for an optimistic algorithm.

1. Introduction

There has been a recent surge of interest for problem-dependent analyses of reinforcement learning (RL) algorithms, both in the context of best policy identification (e.g., Zanette et al., 2019; Marjani & Proutiere, 2021) and regret minimization (e.g., Simchowitz & Jamieson, 2019; He et al., 2020; Yang et al., 2021; Xu et al., 2021). Before this recent trend, problem-dependent bounds were limited to regret minimization in average-reward Markov decision processes (MDPs) (e.g. Burnetas & Katehakis, 1997; Tewari & Bartlett, 2007; Jaksch et al., 2010; Ok et al., 2018). Notably, Burnetas & Katehakis (1997) derived the first problem-dependent asymptotic lower bound for regret minimization in *ergodic* average-reward MDPs and designed an algorithm matching this fundamental limit. Their lower bound was successively extended by Ok et al. (2018) to structured MDPs. However, these results remain restricted to ergodic MDPs, where the need of exploration is limited to the action space, since states are repeatedly visited under any policy.

In finite-horizon MDPs, the literature has focused on deriving problem-dependent "worst-case" lower bounds for regret minimization (Simchowitz & Jamieson, 2019; Xu et al., 2021) with no state reachability assumption (i.e., the counterpart of ergodicity for finite-horizon MDPs). These results are simultaneously *i)* problem-dependent since they scale with instance-specific quantities (e.g., the action-gaps); *ii)* "worst-case" since they are derived only for "hard" instances. Notably, Xu et al. (2021) proved that there exists a specific MDP such that any consistent algorithm must suffer a regret depending on the inverse of the minimum gap and derived an algorithm with matching regret upper bound.

Despite these results, "fully" problem-dependent lower bounds are still missing, i.e., bounds that depend on the properties of any given MDP, instead of relying on specific worst-case instances. In this paper, we a take step in this direction by deriving the first "fully" problem-dependent asymptotic regret lower bound for finite-horizon MDPs. Our lower bound generalizes existing results and provides new insights on the "true" complexity of exploration in this setting. Similarly to average-reward MDPs, our lower-bound is the solution to an optimization problem, but it does not require any assumption on state reachability. Our derivation reveals the need for a constraint on the visitation distribution over state-action pairs that explicitly accounts for the dynamics of the MDP. Interestingly, we show examples where this constraint is crucial to obtain tight lower-bounds and to match existing results derived for specific MDP instances. Finally, we show that, in certain "simple" MDPs, the lower bound is considerably smaller than in the general case and it does not scale with the minimum action-gap, and we show that this is attainable (up to $poly(H)$ terms) by providing a novel regret upper-bound for an optimistic algorithm.

[1]Inria, Lille [2]Facebook AI Research, Paris. Correspondence to: Andrea Tirinzoni <andrea.tirinzoni@inria.fr>, Matteo Pirotta <pirotta@fb.com>, Alessandro Lazaric <lazaric@fb.com>.

[1]Extended abstract. Full version available at https://arxiv.org/abs/2106.13013.

2. Preliminaries

We consider a time-inhomogeneous finite-horizon MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \{p_h, q_h\}_{h\in[H]}, p_0, H)$ (Puterman, 1994), where $\mathcal{S}$ is a finite set of S states, $\mathcal{A}$ is a finite set of A actions, $p_h : \mathcal{S}\times\mathcal{A} \to \mathbb{P}(\mathcal{S})$ and $q_h : \mathcal{S}\times\mathcal{A} \to \mathbb{P}(\mathbb{R})$ are the transition probabilities and the reward distribution at stage $h \in [H] := \{1, \dots, H\}$, $p_0 \in \mathbb{P}(\mathcal{S})$ is the initial state distribution, and H is the horizon.[2] We denote by $r_h(s, a)$ the expected reward after taking action a in state s. A (deterministic) Markov policy $\pi = \{\pi_h\}_{h\in[H]} \in \Pi$ is a sequence of mappings $\pi_h : \mathcal{S} \to \mathcal{A}$. Let Π be the set of such policies. Executing a policy π on $\mathcal{M}$ yields random trajectories $(s_1, a_1, y_1, \dots, s_H, a_H, y_H)$, where $s_1 \sim p_0$, $a_h = \pi_h(s_h)$, $s_{h+1} \sim p_h(s_h, a_h)$, and $y_h \sim q_h(s_h, a_h)$. We denote by $\mathbb{P}^\pi_\mathcal{M}, \mathbb{E}^\pi_\mathcal{M}$ the corresponding probability and expectation operators, and let $\rho^\pi_{\mathcal{M},h}(s, a) := \mathbb{P}^\pi_\mathcal{M}\{s_h = s, a_h = a\}$ and $\rho^\pi_{\mathcal{M},h}(s) := \rho^\pi_{\mathcal{M},h}(s, \pi_h(s))$ be the state-action and state occupancy measures at stage h. For each $s \in \mathcal{S}$ and $h \in [H]$, we define the action-value function of a policy π in $\mathcal{M}$ as

$$Q^\pi_{\mathcal{M},h}(s, a) := \mathbb{E}^\pi_\mathcal{M}\left[\sum_{h'=h}^{H} r_h(s_{h'}, a_{h'})|s_h = s, a_h = a\right],$$

and the value function is $V^\pi_{\mathcal{M},h}(s) := Q^\pi_{\mathcal{M},h}(s, \pi_h(s))$. Let $V^\pi_{\mathcal{M},0} := \mathbb{E}_{s_1\sim p_0}[V^\pi_{\mathcal{M},1}(s_1)]$ and $V^\star_{\mathcal{M},0} = \sup_\pi V^\pi_{\mathcal{M},0}$. We define the set of *return-optimal* policies as

$$\Pi^\star(\mathcal{M}) := \{\pi \in \Pi \mid V^\pi_{\mathcal{M},0} = V^\star_{\mathcal{M},0}\}. \tag{1}$$

By standard MDP theory (e.g., Puterman, 1994), there exists a unique optimal action-value function $Q^\star_{\mathcal{M},h}$ that satisfies the Bellman optimality equations *for any* $h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$, $Q^\star_{\mathcal{M},h}(s, a) = r_h(s, a) + p_h(s, a)^\mathsf{T} V^\star_{\mathcal{M},h+1}$, where $V^\star_{\mathcal{M},h}(s) := \max_{a\in\mathcal{A}} Q^\star_{\mathcal{M},h}(s, a)$. We define the set of Bellman-optimal actions at state-stage (s, h) as $\mathcal{O}_{\mathcal{M},h}(s) := \{a \in \mathcal{A} : Q^\star_{\mathcal{M},h}(s, a) = V^\star_{\mathcal{M},h}(s)\}$, then the set of *Bellman-optimal* policies is $\Pi^\star_\mathcal{O}(\mathcal{M}) := \{\pi \in \Pi \mid \forall s, h : \pi_h(s) \in \mathcal{O}_{\mathcal{M},h}(s)\}$. A Bellman-optimal policy is always return optimal, i.e., $\Pi^\star_\mathcal{O}(\mathcal{M}) \subseteq \Pi^\star(\mathcal{M})$, while it easy to construct examples where the reverse is not true (i.e., a return-optimal policy is not Bellman optimal). Finally, we introduce the *policy gap* $\Gamma_\mathcal{M}(\pi) := V^\star_{\mathcal{M},0} - V^\pi_{\mathcal{M},0}$ and the sub-optimality gap of action a in state s at stage h as

$$\Delta_{\mathcal{M},h}(s, a) := V^\star_{\mathcal{M},h}(s) - Q^\star_{\mathcal{M},h}(s, a). \tag{2}$$

These two notions of sub-optimality are related by the following equation (proof in the full paper[1]):

$$\Gamma_\mathcal{M}(\pi) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h\in[H]} \rho^\pi_{\mathcal{M},h}(s, a)\Delta_{\mathcal{M},h}(s, a). \tag{3}$$

We consider the standard online learning protocol for finite-horizon MDPs. At each *episode* $k \in [K]$, the learner plays a policy π_k and observes a random trajectory $(s_{k,h}, a_{k,h}, y_{k,h}, \dots, s_{k,H}, a_{k,H}, y_{k,H}) \sim \mathbb{P}^{\pi_k}_\mathcal{M}$. The choice of π_k is made by a *learning algorithm* $\mathfrak{A}$, i.e., a measurable function that maps the observations up to episode $k-1$ to policies. The goal is to minimize the cumulative regret,

$$\mathrm{Regret}_K(\mathcal{M}) := \sum_{k=1}^{K} \Gamma_\mathcal{M}(\pi_k). \tag{4}$$

This definition together with (3) reveals that a policy π can have zero regret (i.e., being return-optimal) while selecting actions with $\Delta_{\mathcal{M},h}(s, a) > 0$ (i.e., not Bellman-optimal) at states that have zero occupancy measure $\rho^\pi_{\mathcal{M},h}(s, a)$.

All the proofs can be found in the extended version of the paper[1].

3. Problem-dependent lower bound

As customary in problem-dependent lower bounds, we derive our result for any MDP $\mathcal{M}$ in a given set $\mathfrak{M}$ of MDPs with the same state-action space but different transition probabilities and reward distributions. Formally, we derive an *asymptotic problem-dependent* lower bound on the expected regret of any "provably-efficient" learning algorithm on the set of MDPs $\mathfrak{M}$.

Definition 1 (α-uniformly good algorithm)**.** *Let $\alpha \in (0, 1)$, then a learning algorithm $\mathfrak{A}$ is α-uniformly good on $\mathfrak{M}$ if, for each $K \in \mathbb{N}_{>0}$ and $\mathcal{M} \in \mathfrak{M}$, there exists a constant $c(\mathcal{M})$ such that $\mathbb{E}^\mathfrak{A}_\mathcal{M}[\mathrm{Regret}_K(\mathcal{M})] \le c(\mathcal{M})K^\alpha$.*

Note that existing algorithms with $\mathcal{O}(\sqrt{K})$ worst-case regret (e.g., Azar et al., 2017; Zanette & Brunskill, 2019) are $1/2$-uniformly good, while those with logarithmic regret (e.g., Simchowitz & Jamieson, 2019; Xu et al., 2021) are α-uniformly good for all $\alpha \in (0, 1)$. We make the following assumption on the MDP $\mathcal{M}$.

Assumption 2 (Unique optimal state distribution)**.** *There exists $\rho^\star_{\mathcal{M},h} \in \mathbb{P}(\mathcal{S})$ such that, for any optimal policy $\pi \in \Pi^\star(\mathcal{M})$ and for any $s \in \mathcal{S}, h \in [H]$, $\rho^\star_{\mathcal{M},h}(s) = \rho^\pi_{\mathcal{M},h}(s)$.*

This assumption requires all return-optimal policies of $\mathcal{M}$ to induce the same distribution over the state space. This is strictly weaker than assuming a unique optimal action at each state (see full paper [1]), as commonly done in the contextual bandit setting (Hao et al., 2020; Tirinzoni et al., 2020) and in MDPs (Marjani & Proutiere, 2021).

Let $\mathcal{O}^\star_\mathcal{M} := \{s, a, h : s \in \mathrm{supp}(\rho^\star_{\mathcal{M},h}), a \in \mathcal{O}_{\mathcal{M},h}(s)\}$ be the set of state-action-stage triplets containing all optimal actions in states that are visited by optimal policies. We introduce the following set of *alternative* MDPs to $\mathcal{M}$:

$$\Lambda(\mathcal{M}) := \Lambda^{\mathrm{wa}}(\mathcal{M}) \cap \Lambda^{\mathrm{wc}}(\mathcal{M}),$$

[2] $\mathbb{P}(\Omega)$ denotes the set of probability measures over a set Ω.

where $\Lambda^{\text{wa}}(\mathcal{M}) := \{\mathcal{M}' \in \mathfrak{M} \mid \Pi^\star(\mathcal{M}) \cap \Pi^\star(\mathcal{M}') = \emptyset\}$ and[3]

$$\Lambda^{\text{wc}}(\mathcal{M}) := \{\mathcal{M}' \in \mathfrak{M} \mid \forall z \in \mathcal{O}^\star_{\mathcal{M}} : \text{KL}_z(\mathcal{M}, \mathcal{M}') = 0\}.$$

The set of alternatives is a key component in the derivation of information-theoretic problem-dependent lower bounds (e.g., Lai & Robbins, 1985). Similar to (Burnetas & Katehakis, 1997; Ok et al., 2018), the set of alternatives $\Lambda(\mathcal{M})$ is the intersection of two sets: (1) the set of *weak alternatives* $\Lambda^{\text{wa}}(\mathcal{M})$, i.e., MDPs that have no return-optimal policy in common with $\mathcal{M}$; and (2) the set of *weakly confusing* $\Lambda^{\text{wc}}(\mathcal{M})$, i.e., MDPs whose dynamics and rewards are indistinguishable from $\mathcal{M}$ on the state-action pairs observed while executing any return-optimal policy for $\mathcal{M}$. Notice that the set $\Lambda^{\text{wc}}(\mathcal{M})$ differs from the set of *confusing* MDPs considered in (Burnetas & Katehakis, 1997; Ok et al., 2018). In their case, the zero-KL condition is imposed over all states since the MDP $\mathcal{M}$ is assumed ergodic, which implies that any optimal policy visits all the states with positive probability. In our case, since we do not make any ergodicity assumption, optimal policies may not visit some states at some stages. Therefore, even if the kernels of $\mathcal{M}$ and $\mathcal{M}'$ differ at some optimal action in any such state, the two MDPs remain indistinguishable by playing return-optimal policies. With these notions in mind, we are now ready to state our problem-dependent lower bound.

Theorem 3. *Let $\mathfrak{A}$ be any α-uniformly good learning algorithm on $\mathfrak{M}$ with $\alpha \in (0,1)$. Then, for any $\mathcal{M} \in \mathfrak{M}$ that satisfies Assumption 2,*

$$\liminf_{K\to\infty} \frac{\mathbb{E}^{\mathfrak{A}}_{\mathcal{M}}\left[\text{Regret}_K(\mathcal{M})\right]}{\log(K)} \geq v^\star(\mathcal{M}),$$

where $v^\star(\mathcal{M})$ is the value of the optimization problem

$$\begin{aligned}
&\inf_{\eta \in \mathbb{R}^{SAH}_{\geq 0}} \sum_{h,s,a} \eta_h(s,a)\Delta_{\mathcal{M},h}(s,a) \quad \textit{subject to} \\
&\quad \inf_{\mathcal{M}' \in \Lambda(\mathcal{M})} \sum_{h,s,a} \eta_h(s,a)\text{KL}_{s,a,h}(\mathcal{M},\mathcal{M}') \geq 1-\alpha, \\
&\quad \sum_{a\in\mathcal{A}} \eta_h(s,a) = \sum_{s',a'} p_h(s|s',a')\eta_{h-1}(s',a') \;\forall s, h>1, \\
&\quad \sum_{a\in\mathcal{A}} \eta_1(s,a) = 0 \quad \forall s \notin \text{supp}(p_0).
\end{aligned}$$

The lower bound is the solution to a constrained optimization problem that defines an optimal *exploration strategy* $\eta \in \mathbb{R}^{SAH}$, where $\eta_h(s,a)$ is proportional to the number of visits allocated to each state s and action a at stage h. Such optimal exploration strategy must minimize the resulting regret (written as a weighted sum of local sub-optimality gaps), while satisfying three constraints. First, the *KL constraint*, which is common in this type of information-theoretic lower bounds, requires that the exploration strategy allocates sufficient visits to relevant state-action-stage triplets so as to discriminate $\mathcal{M}$ from all its alternatives $\mathcal{M}' \in \Lambda(\mathcal{M})$. The last two constraints, taken as a whole, form what we refer to as the *dynamics constraint*. This requires the optimal exploration strategy to be *realizable* according to (i.e., compatible with) the MDP dynamics. As we shall see in our examples later, the dynamics constraint is a crucial component to introduce MDP structure into the optimization problem. Without it, an exploration strategy would be allowed to allocate visits to certain state-action pairs regardless of the probability to reach them (i.e., as if a generative model were available), thus resulting in a *non-realizable* allocation in most cases and loose lower bounds.

The policy-based perspective. Note that, by definition, we can realize any allocation η that satisfies the dynamics constraint by playing some *stochastic* policy. Moreover, we can always express the occupancy measure of any stochastic Markov policy as a mixture of *deterministic* Markov policies (e.g., Altman, 1999, Remark 6.1, page 64). This implies that an allocation η satisfies the dynamics constraint in the optimization problem of Thm. 3 if, and only if, there exists a vector $\omega \in \mathbb{R}^{|\Pi|}_{\geq 0}$ of "mixing coefficients" such that $\eta_h(s,a) = \sum_{\pi\in\Pi} \omega_\pi \rho^\pi_h(s,a)$ for all s,a,h. This allows us to rewrite the optimization problem in a simpler form.

Proposition 4. *The optimization problem of Thm. 3 can be rewritten in the following equivalent form*

$$\begin{aligned}
&\inf_{\omega \in \mathbb{R}^{|\Pi|}_{\geq 0}} \sum_{\pi\in\Pi} \omega_\pi \Gamma_{\mathcal{M}}(\pi) \quad \textit{subject to} \\
&\quad \inf_{\mathcal{M}' \in \Lambda(\mathcal{M})} \sum_{\pi\in\Pi} \omega_\pi \text{KL}_\pi(\mathcal{M},\mathcal{M}') \geq 1-\alpha,
\end{aligned}$$

where $\text{KL}_\pi(\mathcal{M},\mathcal{M}') := \sum_{h,s,a} \rho^\pi_{\mathcal{M},h}(s,a)\text{KL}_{s,a,h}(\mathcal{M},\mathcal{M}')$.

While computationally harder than its counterpart in Thm. 3 (we moved from optimizing over SAH variables to $|\Pi| = A^{SH}$ variables), this policy-based perspective is convenient to interpret and instantiate the lower bound in specific cases.

4. Discussion

We briefly recall existing lower bounds. In *ergodic* average-reward MDPs, Burnetas & Katehakis (1997); Tewari & Bartlett (2007); Ok et al. (2018) showed that the optimal problem-dependent regret scales as the sum of the inverse sub-optimality gaps[4]. In finite-horizon MDPs, Simchowitz & Jamieson (2019) first showed that the sum of the inverse

[3] The KL divergence is defined as $\text{KL}_{(s,a,h)}(\mathcal{M},\mathcal{M}') = \text{KL}(p_h(s,a), p'_h(s,a)) + \text{KL}(q_h(s,a), q'_h(s,a))$.

[4] More precisely, it scales with the sum of local complexity measures which are related to the gaps (Tewari & Bartlett, 2007).

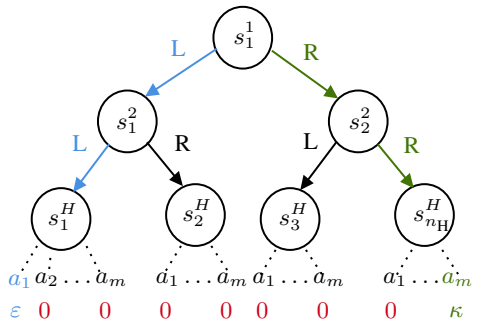

Figure 1. Variant of the example in (Xu et al., 2021). The MDP is binary tree with $S = 2^H - 1$ states, $A = m \geq 2$ actions, and deterministic transitions. The figure shows an instance with $H = 3$. The agent starts from the root state s_1^1 and descends the tree using only two actions (L and R). In the leaf states, all the m actions are available. The rewards follow a Gaussian distribution with unit variance and mean equal to zero everywhere except for at most two leaf state-action pairs (whose values are ε and κ).

gaps is a loose lower bound for a specific family of optimistic algorithms, which in the worst-case may suffer from a regret of at least $S/\Delta_{\min}$. Xu et al. (2021) later refined this result showing that there exists a "hard" MDP where any α-good algorithm (Def. 1) suffers a regret proportional to $SA/\Delta_{\min}$, which is (exponentially) larger than the sum of inverse gaps and it is proportional to $\frac{Z_{\text{mul}}}{\Delta_{\min}}$, where Z_{mul} is the total number of optimal actions in states where the optimal action is not unique. This suggests that the number of non-unique optimal actions may be key to characterize the "worst-case" complexity in finite-horizon MDPs.

On the importance of the dynamics constraint to match existing lower bounds. We consider the MDP $\mathcal{M}$ introduced by Xu et al. (2021) (see Fig. 1 with $\kappa = 0$) and we define $\mathfrak{M}$ as the set of MDPs with exactly the same dynamics as $\mathcal{M}$ but arbitrary Gaussian rewards. In this problem $\Delta_{\min} = \varepsilon > 0$. We instantiate our lower bound in this case with and without the dynamics constraints.

Corollary 5. *Let $\mathcal{M}$ be the MDP of Fig. 1 with $\kappa = 0$. Let $\widetilde{v}(\mathcal{M})$ the value of the optimization problem of Thm. 3* ***without dynamics constraints****, then $\widetilde{v}(\mathcal{M}) = 2(1-\alpha)(\log_2(S+1) + A - 2)/\Delta_{\min}$. On the other hand, the lower bound in Thm. 3 yields $v^\star(\mathcal{M}) \geq (1-\alpha)SA/\Delta_{\min}$.*

This result shows that ignoring the dynamics constraints leads to an exponentially smaller (and thus looser) bound w.r.t. $v^\star(\mathcal{M})$. On the other hand, when computing the lower bound of Thm. 3, we match the lower bound of Xu et al. (2021) for this configuration.

On the dependence on the sum of inverse gaps. While the lower bound of Xu et al. (2021) shows that there exists an MDP where the regret is significantly larger than the sum of inverse gaps whenever multiple equivalent optimal actions exist, in the following we derive a result that is somewhat complementary: we show that there exists a large class of MDPs where the lower bound scales as the sum of the inverse gaps, even when $Z_{\text{mul}} > 0$.

Proposition 6. *Let $\mathcal{M}$ be an MDP satisfying Asm. 2 such that $\rho^\star_{\mathcal{M}}$ is* full-support *(i.e., $\rho^\star_{\mathcal{M},h}(s) > 0, \forall s, h$). Then,*

$$v^\star(\mathcal{M}) = (1-\alpha) \sum_{h,s,a} \frac{\Delta_{\mathcal{M},h}(s,a)}{\mathcal{K}_{s,a,h}(\mathcal{M})} \leq \sum_{h,s,a} \frac{2(H-h)^2}{\Delta_{\mathcal{M},h}(s,a)},$$

where $\mathcal{K}_{s,a,h}(\mathcal{M}) := \inf_{\bar{p},\bar{q} \in \Lambda_s(\mathcal{M})} \{\mathrm{KL}(p_h(s,a),\bar{p}) + \mathrm{KL}(q_h(s,a),\bar{q})\}$ *and* $\Lambda_s(\mathcal{M}) := \{\bar{p} \in \mathbb{P}(\mathcal{S}), \bar{q} \in \mathbb{P}([0,1]) : \mathbb{E}_{y\sim\bar{q}}[y] + \bar{p}^T V^\star_{\mathcal{M},h+1} > V^\star_{\mathcal{M},h}(s)\}$.

Note that the full-support condition for $\rho^\star_{\mathcal{M},h}$ is weaker than ergodicity for average-reward MDPs since it is required only for the optimal policy. For MDPs with this property, the lower bound is obtained by a *decoupled* exploration strategy similar to the one for ergodic MDPs, where the optimal allocation focuses on exploring sub-optimal actions regardless of how to reach the corresponding state, while the exploration of the state space comes "for free" from trying to minimize regret w.r.t. the optimal policy itself. Interestingly, this result holds even for $Z_{\text{mul}} > 0$, suggesting that the dependency $\frac{Z_{\text{mul}}}{\Delta_{\min}}$ derived in (Xu et al., 2021) may be relevant only under specific reachability properties (e.g., when the optimal occupancy measure is not full support).

On the dependence on the minimum gap. Let us consider again the MDP of Fig. 1 under the same setting as before except that $\kappa \geq 2\varepsilon > 0$. In this problem, $\Delta_{\min} = \epsilon$ and $\Delta_{\max} = \kappa$ are the minimum and maximum action gap, respectively. Perhaps surprisingly, despite we only added a single positive reward (κ) to the original hard instance of Xu et al. (2021), we now show that the lower bound of Thm. 3 does not scale with the minimum gap at all.

Proposition 7. *Let $\mathcal{M}$ be the MDP of Fig. 1 with $\kappa \geq 2\varepsilon > 0$, then the lower bound of Thm. 3 yields $v^\star(\mathcal{M}) \leq 12(1-\alpha)\frac{SA}{\Delta_{\max}}$. On the other hand, the sum of inverse gaps of $\mathcal{M}$ is at least $(\log_2(S+1) + A - 3)/\Delta_{\min}$.*

This result shows that, for given S, A, H, one can always construct an MDP where the lower bound of Thm. 3 is smaller than the sum of inverse gaps by an arbitrarily large factor. The intuition is as follows. In order to learn an optimal policy, any consistent algorithm must figure out which among actions L and R is optimal at the root state s_1^1. Action L leads to a return of ϵ, while action R yields a (possibly much larger) return κ. Suppose the agent has estimated all the rewards in the MDP up to an error of $\kappa/2$. This is enough for it to "prune" the whole left branch of the tree since its return is certainly smaller than the one in the right branch. This is better illustrated using the *policy view*: each policy in this MDP has a gap $\Gamma(\pi) \geq \kappa/2$. Thus, an estimation error below the minimum policy gap suffices to

discriminate all sub-optimal policies w.r.t. the optimal one. Notably, this means that the left branch need not be explored to gain ϵ-accurate estimates, which would translate into a much larger $O(1/\Delta_{\min})$ regret. In other words, the agent is not required to explore until it learns a *Bellman optimal* policy (i.e., one that correctly chooses action a_1 in state s_1^H); any *return optimal* policy suffices to minimize regret, and this can be obtained by simply learning to take the right path while playing arbitrary actions at all other states.

To confirm that this result is not an artifact of our lower bound, we provide a novel problem-dependent regret bound for the optimistic algorithm UCBVI (Azar et al., 2017) that scales with the minimum *policy gap* $\Gamma_{\min}$.[5]

Theorem 8. *Let $\mathcal{M}$ be any MDP with rewards in $[0,1]$ and $K \geq 1$, then the expected regret of UCBVI with Chernoff-Hoeffding bonus (ignoring low-order terms in $\log(K)$) is*

$$\mathbb{E}_{\mathcal{M}}[\mathrm{Regret}(K)] \lesssim \frac{4H^4SA}{\Gamma_{\min}} \log(4SAHK^2).$$

This shows that *1)* UCBVI attains the result in Prop. 7 (up to poly(H) factors) where $\Gamma_{\min} = \Delta_{\max}$, even when dynamics are unknown; *2)* Prop. 7 is tight w.r.t. the gaps; *3)* it is possible to achieve regret not scaling with $\Delta_{\min}$.

Outlook. While Thm. 3 provides the first "fully" problem-dependent lower bound for finite-horizon MDPs, it opens a number of interesting directions. **1)** As all existing problem-dependent lower bounds for this setting, the result is asymptotic in nature. A more refined finite-time analysis could be obtained following Garivier et al. (2019). **2)** For the case studied in Cor. 5, Xu et al. (2021) provide an algorithm with matching upper bound (up to poly(H) factors), while we provide a matching upper bound (Thm. 8) for Prop. 7. It remains an open question how to design an algorithm to match the bound of Thm. 3. **3)** Most of the ingredients in deriving Thm. 3 could be adapted to the average-reward case to obtain a lower bound with no ergodicity assumption.

References

Altman, E. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Burnetas, A. N. and Katehakis, M. N. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Hao, B., Lattimore, T., and Szepesvári, C. Adaptive exploration in linear contextual bandit. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3536–3545. PMLR, 2020.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. *CoRR*, abs/2011.11566, 2020.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Marjani, A. A. and Proutiere, A. Adaptive sampling for best policy identification in markov decision processes, 2021.

Ok, J., Proutiere, A., and Tranos, D. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 8874–8882, 2018.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. USA, 1st edition, 1994. ISBN 0471619779.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *NeurIPS*, pp. 1151–1160, 2019.

Tewari, A. and Bartlett, P. L. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *NIPS*, pp. 1505–1512. Curran Associates, Inc., 2007.

Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Xu, H., Ma, T., and Du, S. S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

Yang, K., Yang, L. F., and Du, S. S. Q-learning with logarithmic regret. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1576–1584. PMLR, 2021.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7304–7312. PMLR, 2019.

[5] Exisiting problem-dependent bounds (Simchowitz & Jamieson, 2019; Xu et al., 2021) scale with the action gaps $\Delta_{\mathcal{M}}$.

Zanette, A., Kochenderfer, M. J., and Brunskill, E. Almost horizon-free structure-aware best policy identification with a generative model. In *NeurIPS*, pp. 5626–5635, 2019.