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

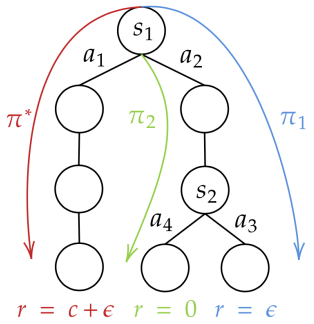
Reinforcement Learning (RL) is a general scenario where agents interact with the environment to achieve some goal. The environment and an agent's interactions are typically modeled as a Markov decision process (MDP) [29], which can represent a rich variety of tasks. But, for which MDPs can an agent or 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 $\mathcal{Z}_\theta = \{(s,a)\in\mathcal{S}\times\mathcal{A}\colon \forall \pi^* \in \Pi^*_\theta \quad w^{\pi^*}_\theta(s,a) = 0\}$ 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.