Planning in Markov Decision Processes with Gap-Dependent Sample Complexity

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar D Domingues, Edouard Leurent, Michal Valko

Planning in Markov Decision Processes with Gap-Dependent Sample Complexity

Anders Jonsson
Universitat Pompeu Fabra
anders.jonsson@upf.edu

Emilie Kaufmann
CNRS & ULille (CRIStAL), Inria Scool
emilie.kaufmann@univ-lille.fr

Pierre Ménard
Inria Lille, Scool team
pierre.menard@inria.fr

Omar Darwiche Domingues
Inria Lille, Scool team
omar.darwiche-domingues@inria.fr

Edouard Leurent
Renault & Inria Lille, Scool team
edouard.leurent@inria.fr

Michal Valko
DeepMind Paris
valkom@deepmind.com

Abstract

We propose MDP-GapE, a new trajectory-based Monte-Carlo Tree Search algorithm for planning in a Markov Decision Process in which transitions have a finite support. We prove an upper bound on the number of calls to the generative model needed for MDP-GapE to identify a near-optimal action with high probability. This problem-dependent *sample complexity* result is expressed in terms of the *sub-optimality gaps* of the state-action pairs that are visited during exploration. Our experiments reveal that MDP-GapE is also effective in practice, in contrast with other algorithms with sample complexity guarantees in the fixed-confidence setting, that are mostly theoretical.

1 Introduction

In reinforcement learning (RL), an agent repeatedly takes *actions* and observes *rewards* in an unknown environment described by a *state*. Formally, the environment is a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $p = \{p_h\}_{h \geq 1}$ a set of transition kernels and $r = \{r_h\}_{h \geq 1}$ a set of reward functions. By taking action a in state s at step h, the agent reaches a state s' with probability $p_h(s'|s,a)$ and receives a random reward with mean $r_h(s,a)$. A common goal is to learn a policy $\pi = (\pi_h)_{h \geq 1}$ that maximizes cumulative reward by taking action $\pi_h(s)$ in state s at step h. If the agent has access to a generative model, it may *plan* before acting by generating additional samples in order to improve its estimate of the best action to take next.

In this work, we consider *Monte-Carlo planning* as the task of recommending a good action to be taken by the agent in a given state s_1, by using samples gathered from a generative model. Let $Q^\star(s_1, a)$ be the maximum cumulative reward, in expectation, that can be obtained from state s_1 by first taking action a, and let $\hat{a}_n$ be the recommended action after n calls to the generative model. The quality of the action recommendation is measured by its *simple regret*, defined as $\bar{r}_n(\hat{a}_n) := V^\star(s_1) - Q^\star(s, \hat{a}_n)$, where $V^\star(s_1) := \max_a Q^\star(s_1, a)$.

We propose an algorithm in the *fixed confidence* setting (ε, δ): after n calls to the generative model, the algorithm should return an action $\hat{a}_n$ such that $\bar{r}_n(\hat{a}_n) \leq \varepsilon$ with probability at least $1 - \delta$. We prove that its *sample complexity* n is bounded in high probability by a quantity that depends on

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Table 1: Different settings of planning algorithms in the literature

Setting	Input	Output	Optimality criterion
(1) Fixed confidence (action-based)	ε, δ	$\widehat{a}_n$	$\mathbb{P}\left(\bar{r}_n(\widehat{a}_n) \leq \varepsilon\right) \geq 1-\delta$
(2) Fixed confidence (value-based)	ε, δ	$\widehat{V}(s_1)$	$\mathbb{P}\left(\lvert\widehat{V}(s_1) - V^\star(s_1)\rvert \leq \varepsilon\right) \geq 1-\delta$
(3) Fixed budget	n (budget)	$\widehat{a}_n$	$\mathbb{E}\left[\bar{r}_n(\widehat{a}_n)\right]$ decreasing in n
(4) Anytime	-	$\widehat{a}_n$	$\mathbb{E}\left[\bar{r}_n(\widehat{a}_n)\right]$ decreasing in n

Table 2: Algorithms with sample complexity guarantees

Algorithm	Setting	Sample complexity	Remarks
Sparse Sampling [19]	(1)-(2)	$H^5(BK)^H/\varepsilon^2$ or $\varepsilon^{-\left(2+\frac{\log(BK)}{\log(1/\gamma)}\right)}$	proved in Lemma 1
OLOP [2]	(3)	$\varepsilon^{-\max\left(2, \frac{\log\kappa}{\log(1/\gamma)}\right)}$	open loop, $\kappa \in [1, K]$
OP [3]	(4)	$\varepsilon^{-\frac{\log\kappa}{\log(1/\gamma)}}$	known MDP, $\kappa \in [1, BK]$
BRUE [8]	(4)	$H^4(BK)^H/\Delta^2$	minimal gap Δ
StOP [28]	(1)	$\varepsilon^{-\left(2+\frac{\log\kappa}{\log(1/\gamma)}+o(1)\right)}$	$\kappa \in [1, BK]$
TrailBlazer [13]	(2)	$\varepsilon^{-\max\left(2, \frac{\log(B\kappa)}{\log(1/\gamma)}+o(1)\right)}$	$\kappa \in [1, K]$
SmoothCruiser [14]	(2)	ε^{-4}	only regularized MDPs
MDP-GapE (ours)	(1)	$\sum_{a_1 \in \mathcal{A}} \frac{H^2(BK)^{H-1}B}{(\Delta_1(s_1,a_1)\vee\Delta\vee\varepsilon)^2}$	see Corollary 1

the sub-optimality gaps of the actions that are applicable in state s_1. We also provide experiments showing its effectiveness. The only assumption that we make on the MDP is that the support of the transition probabilities $p_h(\cdot|s,a)$ should have cardinality bounded by $B < \infty$, for all s, a and h.

Monte-Carlo Tree Search (MCTS) is a form of Monte-Carlo planning that uses a *forward model* to sample transitions from the current state, as opposed to a full generative model that can sample anywhere. Most MCTS algorithms sample *trajectories* from the current state [1], and are widely used in *deterministic* games such as Go. The AlphaZero algorithm [26] guides planning using value and policy estimates to generate trajectories that improve these estimates. The MuZero algorithm [25] combines MCTS with a model-based method which has proven useful for *stochastic* environments. Hence *efficient Monte-Carlo planning* may be instrumental for learning better policies. Despite their empirical success, little is known about the sample complexity of state-of-the-art MCTS algorithms.

Related work The earliest MCTS algorithm with theoretical guarantees is Sparse Sampling [19], whose sample complexity is polynomial in $1/\varepsilon$ in the case $B < \infty$ (see Lemma 1). However, it is not trajectory-based and does not select actions adaptively, making it very inefficient in practice.

Since then, adaptive planning algorithms with small sample complexities have been proposed in different settings with different optimality criteria. In Table 1, we summarize the most common settings, and in Table 2, we show the sample complexity of related algorithms (omitting logarithmic terms and constants) when $B < \infty$. Algorithms are either designed for a discounted setting with $\gamma < 1$ or an episodic setting with horizon H. Sample complexities are stated in terms of the accuracy ε (for algorithms with fixed-budget guarantees we solve $\mathbb{E}\left[\bar{r}_n\right] = \varepsilon$ for n), the number of actions K, the horizon H or the discount factor γ and a problem-dependent quantity κ which is a notion of branching factor of near-optimal nodes whose exact definition varies.

A first category of algorithms rely on optimistic planning [23], and require additional assumptions: a deterministic MDP [15], the *open loop* setting [2, 21] in which policies are sequences of actions instead of state-action mappings (the two are equivalent in MDPs with deterministic transitions), or an MDP with known parameters [3]. For MDPs with stochastic and unknown transitions, polynomial sample complexities have been obtained for StOP [28], TrailBlazer [13] and SmoothCruiser [14] but the three algorithms suffer from numerical inefficiency, even for $B < \infty$. Indeed, StOP explicitly reasons about policies and storing them is very costly, while TrailBlazer and SmoothCruiser require a very large amount of recursive calls even for small MDPs. We remark that popular MCTS algorithms such as UCT [20] are not (ε, δ)-correct and do not have provably small sample complexities.

In the setting $B < \infty$, BRUE [8] is a trajectory-based algorithm that is anytime and whose sample complexity depends on the smallest sub-optimality gap $\Delta := \min_{a \neq a^\star}\left(V^\star(s_1) - Q^\star(s_1, a)\right)$. For

planning in deterministic games, gap-dependent sample complexity bounds were previously provided in a fixed-confidence setting [16, 18]. Our proposal, MDP-GapE, can be viewed as a non-trivial adaptation of the UGapE-MCTS algorithm [18] to planning in MDPs. The defining property of MDP-GapE is that it uses a best arm identification algorithm, UGapE [10], to select the first action in a trajectory, and performs optimistic planning thereafter, which helps refining confidence intervals on the intermediate Q-values. Best arm identification tools have been previously used for planning in MDPs [24, 29] and UGapE also served as a building block for StOP [28].

Finally, going beyond worst-case guarantees for RL is an active research direction, and in a different context gap-dependent bounds on the regret have recently been established for tabular MDPs [27, 30].

Contributions We present MDP-GapE, a new MCTS algorithm for planning in the setting $B < \infty$. MDP-GapE performs efficient Monte-Carlo planning in the following sense: First, it is a simple trajectory-based algorithm which performs well in practice and only relies on a forward model. Second, while most practical MCTS algorithms are not well understood theoretically, we prove upper bounds on the sample complexity of MDP-GapE. Our bounds depend on the *sub-optimality gaps* associated to the state-action pairs encountered during exploration. This is in contrast to StOP and TrailBlazer, two algorithms for the same setting, whose guarantees depend on a notion of near-optimal nodes which can be harder to interpret, and that can be inefficient in practice. In the anytime setting, BRUE also features a gap-dependent sample complexity, but only through the worst-case gap Δ defined above. As can be seen in Table 1, the upper bound for MDP-GapE given in Corollary 1 improves over that of BRUE as it features the gap of each possible first action a_1, $\Delta_1(s_1, a_1) = V^\star(s_1) - Q_1^\star(s_1, a_1)$, and scales better with the planning horizon H. Furthermore, our proof technique relates the *pseudo-counts* of any trajectory prefix to the gaps of state-action pairs on this trajectory, which evidences the fact that MDP-GapE does not explore trajectories uniformly.

2 Learning Framework and Notation

We consider a *discounted episodic setting* where $H \in \mathbb{N}^\star$ is a horizon and $\gamma \in (0, 1]$ a discount parameter. The transition kernels $p = (p_1, \dots, p_H)$ and reward functions $r = (r_1, \dots, r_H)$ can have distinct definitions in each step of the episode. The optimal value of selecting action a in state s_1 is

$$Q^\star(s_1, a) = \max_\pi \mathbb{E}^\pi \left[\sum_{h=1}^{H} \gamma^{h-1} r_h(s_h, a_h) \middle| a_1 = a \right],$$

where the supremum is taken over (deterministic) policies $\pi = (\pi_1, \dots, \pi_H)$, and the expectation is on a trajectory $s_1, a_1, \dots, s_h, a_h$ where $s_h \sim p_{h-1}(\cdot|s_{h-1}, a_{h-1})$ and $a_h = \pi_h(s_h)$ for $h \in [2, H]$. With this definition, an optimal action in state s_1 is $a^\star \in \operatorname{argmax}_{a \in \mathcal{A}(s_1)} Q^\star(s_1, a)$.

We assume that there is a maximal number K of actions available in each state, and that, for each (s, a), the support of $p_h(\cdot|s, a)$ is bounded by B: that is, B is the maximum number of possible next states when applying any action. We further assume that the rewards are bounded in $[0, 1]$. For each pair of integers i, h such that $i \leq h$, we introduce the notation $[i, h] = \{i, \dots, h\}$ and $[h] = [1, h]$.

(ε, δ)-correct planning A sequential planning algorithm proceeds as follows. In each episode t, the agent uses a deterministic policy on the form $\pi^t = (\pi_1^t, \dots, \pi_H^t)$ to generate a trajectory $(s_1, a_1^t, r_1^t, \dots, s_H^t, a_H^t, r_H^t)$, where $a_h^t = \pi_h^t(s_h^t)$, r_h^t is a reward with expectation $r_h(s_h^t, a_h^t)$ and $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$. After each episode the agent decides whether it should perform a new episode to refine its guess for a near-optimal action, or whether it can stop and make a guess. We denote by τ the stopping rule of the agent, that is the number of episodes performed, and $\hat{a}_\tau$ the guess.

We aim to build an (ε, δ)-correct algorithm, that is an algorithm that outputs a guess $\hat{a}_\tau$ satisfying

$$\mathbb{P}\left(Q^\star(s_1, \hat{a}_\tau) > Q^\star(s_1, a^\star) - \varepsilon\right) \geq 1 - \delta \quad \Leftrightarrow \quad \mathbb{P}\left(\bar{r}_{(H\tau)}(\hat{a}_\tau) \leq \varepsilon\right) \geq 1 - \delta \tag{1}$$

while using as few calls to the generative model $n = H\tau$ (i.e. as few episodes τ) as possible.

Our setup permits to propose algorithms for planning in the undiscounted episodic case (in which our bounds will not blow up when $\gamma = 1$) and in discounted MDPs with infinite horizon. Indeed, choosing H such that $2\gamma^H/(1-\gamma) \leq \varepsilon$, an (ε, δ)-correct algorithm for the discounted episodic setting recommends an action that is 2ε-optimal for the discounted infinite horizon setting.

A (recursive) baseline Sparse Sampling [19] can be tuned to output a guess $\hat{a}$ that satisfies (1), as specified in the following lemma, which provides a baseline for our undiscounted episodic setting (see Appendix F). Note that Sparse Sampling is not strictly sequential as it does not repeatedly select trajectories. However, it can still be implemented using a forward model by storing states on a stack.

Lemma 1. *If $B < \infty$, Sparse Sampling using horizon H and performing $\mathcal{O}\left((H^5/\varepsilon^2)\log(BK/\delta)\right)$ transitions in each node is (ε, δ)-correct with sample complexity $O(n_{SS})$ for $n_{SS} := H^5(BK)^H/\varepsilon^2$.*

Structure of the optimal Q-value function In our algorithm, we will build estimates of the intermediate Q-values, that are useful to compute the optimal Q-value function $Q^\star(s_1, a)$. Defining

$$Q_h(s_h, a_h) = \max_\pi \mathbb{E}^\pi \left[\sum_{i=h}^{H} \gamma^{i-h} r(s_i, a_i) \middle| s_h, a_h \right],$$

$Q^\star(s_1, a) = Q_1(s_1, a)$ and the optimal action-values $Q = (Q_1, \dots, Q_H)$ can be computed recursively using the Bellman equations, where we use the convention $Q_{H+1}(\cdot, \cdot) = 0$:

$$Q_h(s_h, a_h) = r_h(s_h, a_h) + \gamma \sum_{s'} p_h(s'|s_h, a_h) \max_{a'} Q_{h+1}(s', a'), \quad h \in [H].$$

Let $\pi^\star = (\pi_1^\star, \dots, \pi_H^\star)$ denote a deterministic optimal policy where, for $h \in [H]$, $\pi_h^\star(s_h) = \arg\max_a Q_h(s_h, a)$, with ties arbitrarily broken. Hence the optimal value in s_h is $Q_h(s_h, \pi_h^\star(s_h))$.

3 The MDP-GapE Algorithm

In this section we present MDP-GapE, a generalization of UGapE [10] to Monte-Carlo planning. Like BAI-MCTS for games [18] a core component is the construction of confidence intervals on $Q_1(s_1, a)$. The construction below generalizes that of OP-MDP [3] for known transition probabilities.

Confidence bounds on the Q-values Our algorithm maintains empirical estimates, superscripted with the episode t, of the transition kernels p and expected rewards r, which are assumed unknown.

Let $n_h^t(s_h, a_h, s_{h+1}) := \sum_{s=1}^t \mathbb{1}\left((s_h^s, a_h^s, s_{h+1}^s) = (s_h, a_h, s_{h+1})\right)$ be the number of observations of transition (s_h, a_h, s_{h+1}), and $R_h^t(s_h, a_h) := \sum_{s=1}^t r_h^s(s_h, a_h)\mathbb{1}\left((s_h^s, a_h^s) = (s_h, a_h)\right)$ the sum of rewards obtained when selecting a_h in s_h. We define the empirical transition probabilities $\hat{p}^t$ and expected rewards $\hat{r}^t$ as follows, for state-action pairs such that $n_h^t(s_h, a_h) := \sum_s n_h^t(s_h, a_h, s) > 0$:

$$\hat{p}_h^t(s_{h+1}|s_h, a_h) := \frac{n_h^t(s_h, a_h, s_{h+1})}{n_h^t(s_h, a_h)}, \quad \text{and} \quad \hat{r}_h^t(s_h, a_h) := \frac{R_h^t(s_h, a_h)}{n_h^t(s_h, a_h)}.$$

As rewards are bounded in $[0, 1]$, we define the following Kullback-Leibler upper and lower confidence bounds on the mean rewards $r_h(s_h, a_h)$ [4]:

$$u_h^t(s_h, a_h) := \max\left\{ v : \mathrm{kl}\big(\hat{r}_h^t(s_h, a_h), v\big) \le \frac{\beta^r(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h)} \right\},$$
$$\ell_h^t(s_h, a_h) := \min\left\{ v : \mathrm{kl}\big(\hat{r}_h^t(s_h, a_h), v\big) \le \frac{\beta^r(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h)} \right\},$$

where β^r is an exploration function and $\mathrm{kl}(u, v)$ is the binary Kullback-Leibler divergence between two Bernoulli distributions $\mathcal{B}\mathrm{er}(u)$ and $\mathcal{B}\mathrm{er}(v)$: $\mathrm{kl}(u, v) = u \log \frac{u}{v} + (1-u)\log\frac{1-u}{1-v}$. We adopt the convention that $u_h^t(s_h, a_h) = 1, \ell_h^t(s_h, a_h) = 0$ when $n_h^t(s_h, a_h) = 0$.

In order to define confidence bounds on the values Q_h, we introduce a confidence set on the probability vector $p_h(\cdot|s_h, a_h)$. We define $\mathcal{C}_h^t(s_h, a_h) = \Sigma_B$ if $n_h^t(s_h, a_h) = 0$ and otherwise

$$\mathcal{C}_h^t(s_h, a_h) := \left\{ p \in \Sigma_B : \mathrm{KL}\big(\widehat{p}_h^t(\cdot|s_h, a_h), p\big) \le \frac{\beta^p(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h)} \right\},$$

where Σ_B is the set of probability distribution over B elements, β^p is an exploration function and $\mathrm{KL}(p, q) = \sum_{s \in \mathrm{Supp}(p)} p(s) \log \frac{p(s)}{q(s)}$ is the Kullback-Leibler divergence between two categorical distributions p and q with supports satisfying $\mathrm{Supp}(p) \subseteq \mathrm{Supp}(q)$.

We now define our confidence bounds on the action values inductively. We use the convention $U^t_{H+1}(\cdot,\cdot) = L^t_{H+1}(\cdot,\cdot) = 0$, and for all $h \in [H]$,

$$U^t_h(s_h,a_h) = u^t_h(s_h,a_h) + \gamma \max_{p \in \mathcal{C}^t_h(s_h,a_h)} \sum_{s'} p(s'|s_h,a_h) \max_{a'} U^t_{h+1}(s',a'),$$

$$L^t_h(s_h,a_h) = \ell^t_h(s_h,a_h) + \gamma \min_{p \in \mathcal{C}^t_h(s_h,a_h)} \sum_{s'} p(s'|s_h,a_h) \max_{a'} L^t_{h+1}(s',a').$$

As explained in Appendix A of [9], optimizing over these KL confidence sets can be reduced to a linear program with convex constraints, that can be solved efficiently using Newton's method, which has complexity $O(B\log(d))$ where d is the desired digit precision.

We provide in Section 4.1 an explicit choice for the exploration functions $\beta^r(n,\delta)$ and $\beta^p(n,\delta)$ that govern the size of the confidence intervals. Note that if the rewards or transitions are deterministic, or if we know p, we can adapt our confidence bounds by setting $\beta^p = 0$ or $\beta^r = 0$.

MDP-GapE As any fixed-confidence algorithm, MDP-GapE depends on the tolerance parameter ε and the risk parameter δ. The dependency in ε is explicit in the stopping rule (4), while the dependency in δ is in the tuning of the confidence bounds, that depend on δ.

After t trajectories observed, MDP-GapE selects the $(t+1)$-st trajectory using the policy $\pi^{t+1} = (\pi^{t+1}_1, \dots, \pi^{t+1}_H)$ where the first action choice is made according to UGapE:

$$\pi^{t+1}_1(s_1) = \operatorname*{argmax}_{b \in \{b^t, c^t\}} \left[U^t_1(s_1,b) - L^t_1(s_1,b)\right],$$

where b^t is the current guess for the best action, which is the action b with the smallest upper confidence bound on its gap $Q^\star_1(s_1,a^\star) - Q_1(s_1,b)$, and c^t is some challenger:

$$b^t = \operatorname*{argmin}_b \left[\max_{a \neq b} U^t_1(s_1,a) - L^t_1(s_1,b)\right], \tag{2}$$

$$c^t = \operatorname*{argmax}_{c \neq b^t} U^t_1(s_1,c). \tag{3}$$

Then for all remaining steps we follow an optimistic policy, for all $h \in [2,H]$,

$$\pi^{t+1}_h(s_h) = \operatorname*{argmax}_a U^t_h(s_h,a).$$

The stopping rule of MDP-GapE is

$$\tau = \inf\{t \in \mathbb{N} : U^t_1(s_1,c^t) - L^t_1(s_1,b^t) \leq \varepsilon\}, \tag{4}$$

and the guess output when stopping is $\hat{a}_\tau = b^\tau$. A generic implementation of MDP-GapE is given in Algorithm 1 in Appendix A, where we also discuss some implementation details. Note that, in sharp contrast with the deterministic stopping rule proposed for Sparse Sampling in Lemma 1, MDP-GapEuses an adaptive stopping rule.

The high-level intuition behind MDP-GapE is that unlike a greedy optimistic policy, MDP-GapE does not attempt to minimize regret while learning the best action to take in step 1. The UGapE policy followed at depth 1 indeed explores much more than a purely optimistic algorithm. At depths larger than 1, however, MDP-GapE does follow a greedy optimistic policy. This combination of policy choices is crucial for the theoretical analysis of the algorithm, and for quickly achieving the proposed stopping condition (4): stop when one of the confidence intervals on the value at depth 1 is larger than and separated from the others.

4 Analysis of MDP-GapE

Recall that MDP-GapE uses policy $\pi^{t+1} = (\pi^{t+1}_1, \dots, \pi^{t+1}_H)$ to select the $(t+1)$-st trajectory, $s_1, a^{t+1}_1, s^{t+1}_2, a^{t+1}_2, \dots, s^{t+1}_H, a^{t+1}_H$, satisfying $a^{t+1}_h = \pi^{t+1}_h(s^{t+1}_h)$ and $s^{t+1}_{h+1} \sim p_h\left(\cdot \middle| s^{t+1}_h, a^{t+1}_h\right)$.

High probability event To define an event $\mathcal{E}$ that holds with high probability, let $\mathcal{E}^r$ (resp. $\mathcal{E}^p$) be the event that the confidence regions for the mean rewards (resp. transition kernels) are correct:

$$\mathcal{E}^r := \left\{\forall t \in \mathbb{N}^*, \forall h \in [H], \forall (s_h, a_h) \in \mathcal{S} \times \mathcal{A} :\ r_h(s_h, a_h) \in \left[\ell_h^t(s_h, a_h), u_h^t(s_h, a_h)\right]\right\},$$
$$\mathcal{E}^p := \left\{\forall t \in \mathbb{N}^*, \forall h \in [H], \forall (s_h, a_h) \in \mathcal{S} \times \mathcal{A} :\ p_h(\cdot|s_h, a_h) \in \mathcal{C}_h^t(s_h, a_h)\right\}.$$

For a state-action pair (s_h, a_h), let $p_h^\pi(s_h, a_h)$ be the probability of reaching it at step h under policy π, and let $p_h^t(s_h, a_h) = p_h^{\pi^t}(s_h, a_h)$. We define the *pseudo-counts* of the number of visits of (s_h, a_h) as $\bar{n}_h^t(s_h, a_h) := \sum_{s=1}^t p_h^s(s_h, a_h)$. As $n_h^t(s_h, a_h) - \bar{n}_h^t(s_h, a_h)$ is a martingale, the counts should not be too far from the pseudo-counts. Given a rate function β^{cnt}, we define the event

$$\mathcal{E}^{\text{cnt}} := \left\{\forall t \in \mathbb{N}^\star, \forall h \in [H], \forall (s_h, a_h) \in \mathcal{S} \times \mathcal{A} :\ n_h^t(s_h, a_h) \geq \frac{1}{2}\bar{n}_h^t(s_h, a_h) - \beta^{\text{cnt}}(\delta)\right\}.$$

Finally, we define $\mathcal{E}$ to be the intersection of these three events: $\mathcal{E} = \mathcal{E}^r \cap \mathcal{E}^p \cap \mathcal{E}^{\text{cnt}}$.

4.1 Correctness

One can easily prove by induction (see Appendix B) that

$$\mathcal{E}^r \cap \mathcal{E}^p \subseteq \bigcap_{t \in \mathbb{N}^\star} \bigcap_{h=1}^H \Big[\bigcap_{s_h, a_h} \Big(Q_h(s_h, a_h) \in \left[L_h^t(s_h, a_h), U_h^t(s_h, a_h)\right]\Big)\Big].$$

As the arm $\hat{a}$ output by MDP-GapE satisfies $L_1(s_1, \hat{a}) > \max_{c \neq \hat{a}} U_1(s_1, c) - \varepsilon$, on the event $\mathcal{E} \subseteq \mathcal{E}^r \cap \mathcal{E}^p$ it holds that $Q_1(s_1, \hat{a}) > \max_{c \neq \hat{a}} Q_1(s_1, c) - \varepsilon$. Thus MDP-GapE can only output an ε-optimal action. Hence a sufficient condition for MDP-GapE to be (ε, δ)-correct is $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

In Lemma 2 below, we provide a calibration of the thresholds functions β^r, β^p and β^{cnt} such that this sufficient condition holds. This result, proved in Appendix C, relies on new time-uniform concentration inequalities that follow from the method of mixtures [7].

Lemma 2. *For all $\delta \in [0, 1]$, it holds that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ for the choices*

$$\beta^r(n, \delta) = \log(3(BK)^H/\delta) + \log\left(e(1 + n)\right), \quad \beta^{\text{cnt}}(\delta) = \log\left(3(BK)^H/\delta\right),$$
$$\text{and} \quad \beta^p(n, \delta) = \log\left(3(BK)^H/\delta\right) + (B - 1)\log\left(e(1 + n/(B - 1))\right).$$

Moreover, the maximum of these three thresholds defined (by continuity when $B = 1$) as

$$\beta(n, \delta) \quad := \max_{c \in \{r, p, \text{cnt}\}} \beta^c(n, \delta) = \log\left(3(BK)^H/\delta\right) + (B - 1)\log\left(e(1 + n/(B - 1))\right),$$

is such that $n \mapsto \beta(n, \delta)$ is non-decreasing and $n \mapsto \beta(n, \delta)/n$ is non-increasing.

4.2 Sample Complexity

In order to state our results, we define the following sub-optimality gaps. $\Delta_h(s_h, a_h)$ measures the gap in future discounted reward between the optimal action $\pi_h^\star(s_h)$ and the action a_h, whereas $\Delta_1^\star(s_1, a_1)$ also takes into account the gap of the second best action and the tolerance level ε.

Definition 1. *Recall that $\Delta = \min_{a \neq a^\star}\left[Q_1(s_1, a^\star) - Q_1(s_1, a)\right]$. For all $h \in [H]$, we let*

$$\Delta_h(s_h, a_h) \quad = \quad Q_h(s_h, \pi_h^\star(s_h)) - Q_h(s_h, a_h),$$
$$\Delta_1^\star(s_1, a_1) \quad = \quad \max\left(\Delta_1(s_1, a_1); \Delta; \varepsilon\right),$$

and we denote $\tilde{\Delta}_h(s_h, a_h) = \begin{cases} \Delta_1^\star(s_h, a_h), & \text{if } h = 1, \\ \Delta_h(s_h, a_h), & \text{if } h \geq 2. \end{cases}$

Our sample complexity bounds follow from the following crucial theorem, which we prove in Appendix D, that relates the pseudo-counts of state-action pairs at time τ to the corresponding gap.

Theorem 1. *If $\mathcal{E}$ holds, every (s_h, a_h) is such that*

$$\bar{n}_h^\tau(s_h, a_h)\tilde{\Delta}_h(s_h, a_h) \leq 64\sqrt{2}(1 + \sqrt{2})\left(\sqrt{BK}\right)^{H-h}\sqrt{\bar{n}_h^\tau(s_h, a_h)\beta(\bar{n}_h^\tau(s_h, a_h), \delta)}.$$

Introducing the constant $C_0 = (64\sqrt{2}(1+\sqrt{2}))^2$ and letting $c_\delta = \log\left(\frac{3(BK)^H}{\delta}\right)$, Lemma 12 stated in Appendix G permits to prove that, on the event $\mathcal{E}$, any (s_h, a_h) for which $\tilde{\Delta}_h(s_h, a_h) > 0$ satisfies

$$\bar{n}_h^\tau(s_h, a_h) \leq \frac{C_0(BK)^{H-h}}{\tilde{\Delta}_h^2(s_h, a_h)}\left[c_\delta + 2(B-1)\log\left(\frac{C_0(BK)^{H-h}}{\tilde{\Delta}_h^2(s_h, a_h)}\left[\frac{c_\delta}{\sqrt{B-1}} + 2\sqrt{e(B-1)}\right]\right) + (B-1)\right] \quad (5)$$

As $\tilde{\Delta}_1(s_1, a_1) = \max\left(\Delta_1(s_1, a_1); \Delta; \varepsilon\right)$ is positive, the following corollary follows from summing the inequality over a_1, as $\bar{n}_1^\tau(s_1, a_1) = n_1^\tau(s_1, a_1)$ and $\tau = \sum_{a_1} n_1^\tau(s_1, a_1)$.

Corollary 1. *The number of episodes used by MDP-GapE satisfies*

$$\mathbb{P}\left(\tau = \mathcal{O}\left(\sum_{a_1} \frac{(BK)^{H-1}}{\left(\Delta_1(s_1, a_1) \vee \Delta \vee \varepsilon\right)^2}\left[\log\left(\frac{1}{\delta}\right) + BH\log(BK)\right]\right)\right) \geq 1 - \delta\,.$$

The upper bound on the sample complexity $n = H\tau$ of `MDP-GapE` that follows from Corollary 1 improves over the $\mathcal{O}(H^5(BK)^H/\varepsilon^2)$ sample complexity of `Sparse Sampling`. It is also smaller than the $\mathcal{O}(H^4(BK)^H/\Delta^2)$ samples needed for `BRUE` to have a reasonable upper bound on its simple regret. The improvement is twofold: first, this new bound features the problem dependent gap $\Delta(s_1, a_1) \vee \Delta \vee \varepsilon$ for each action a_1 in state s_1, whereas previous bounds were only expressed with ε or Δ. Second, it features an improved scaling in H^2.

It is also possible to provide bounds that features the gaps $\tilde{\Delta}_h(s_h, a_h)$ in the *whole* tree, beyond depth one. To do so, we shall consider *trajectories* $t_{1:H} = (s_1, a_1, \ldots, s_H, a_H)$ or trajectory *prefixes* $t_{1:h} = (s_1, a_1, \ldots, s_h, a_h)$ for $h \in [H]$. Introducing the probability $p_h^\pi(t_{1:h})$ that the prefix $t_{1:h}$ is visited under policy π, we can further define the pseudo-counts $\bar{n}_h^t(t_{1:h}) = \sum_{s=1}^t p_h^{\pi^s}(t_{1:h})$. One can easily show that for all $h \in [H]$, $\bar{n}_H^\tau(t_{1:H}) \leq \bar{n}_h^\tau(t_{1:h}) \leq \bar{n}_h^\tau(s_h, a_h)$, if (s_h, a_h) is the state-action pair visited in step h in the trajectory $t_{1:H}$, and (5) leads to the following upper bound.

Corollary 2. *On the event $\mathcal{E}$, $\bar{n}_h^\tau(t_{1:h}) = \mathcal{O}\left(\left[\min_{\ell=1}^h \frac{(BK)^{H-\ell}}{\left(\tilde{\Delta}_\ell(s_\ell, a_\ell)\right)^2}\right]\log\left(\frac{3(BK)^H}{\delta}\right)\right)$.*

In particular, using that $\tau = \sum_{t_{1:H}\in\mathcal{T}} \bar{n}_h^\tau(t_{1:H})$ where $\mathcal{T}$ is the set of $(BK)^H$ complete trajectories leads to a sample complexity bound featuring all gaps. However, its improvement over the bound of Corollary 1 is not obvious in the general case. For $B = 1$, that is for planning in a deterministic MDP with possibly random rewards, a slightly different proof technique leads to the following improved gap-dependent sample complexity bound (see the proof in Appendix E).

Theorem 2 (deterministic case)**.** *When $B = 1$, `MDP-GapE` satisfies*

$$\mathbb{P}\left(\tau = \mathcal{O}\left(\sum_{t_{1:H}\in\mathcal{T}}\left[\min_{h=1}^H \frac{\left(\sum_{\ell=h}^H \gamma^\ell\right)^2}{\left(\tilde{\Delta}_h^2(s_h, a_h)\right)^2}\right]\left(\log\left(\frac{1}{\delta}\right) + H\log(K)\right)\right)\right) \geq 1 - \delta.$$

Scaling in ε A majority of prior work on planning in MDPs has obtained sample complexity bounds that scale with ε only, in the discounted setting. Neglecting the gaps, Corollary 1 gives a $\mathcal{O}(H^2(BK)^H/\varepsilon^2)$ upper bound that yields a crude $\tilde{\mathcal{O}}\left(\varepsilon^{-[2+\log(BK)/\log(1/\gamma)]}\right)$ sample complexity in the discounted setting in which $H \sim \log(1/\varepsilon)/\log(1/\gamma)$. This exponent is larger than that in previous work, which features some notion of near-optimality dimension κ (see Table 1). However, our analysis was not tailored to optimizing this exponent, and we show in Section 5 that the empirical scaling of `MDP-GapE` in ε can be much smaller than the one prescribed by the above crude bound.

Lower bounds To the best of our knowledge, the only available lower bound on the sample complexity of MCTS planning in general MDPs is the $(1/\varepsilon)^{1/\log(1/\gamma)}$ worst-case bound given by Kearns et al. [19], which is proved using an MDP that is a binary tree ($B < \infty$). In the open-loop setting, Bubeck and Munos [2] prove a minimax lower bound that is $\Omega\left(\varepsilon^{-\log K/\log(1/\gamma)}\right)$ if $\gamma\sqrt{K} > 1$ and $\Omega\left(\varepsilon^{-2}\right)$ if $\gamma\sqrt{K} \leq 1$. As for problem-dependent results, the only available results hold for $H = 1$, for which MCTS planning boils down to finding an arm with mean that is within ε of the best mean in a bandit model. In that case, the lower bound of Mannor and Tsitsiklis [22] indeed features the gaps at depth-one that appear in Corollary 1. Deriving problem-dependent lower bound for $H \geq 2$ is left as an important future work.

5 Numerical Experiments[1]

We consider random discounted MDPs with infinite horizon in which the maximal number B of successor states and the sparsity of rewards are controlled. The transition kernel is generated as follows: for each transition in $\mathcal{S} \times \mathcal{A}$, we uniformly pick B next states in $\mathcal{S}$. The cumulative transition probabilities to these states are computed by sorting $B-1$ numbers uniformly sampled in $(0,1)$. The reward kernel is computed by selecting a proportion of the transitions to have non-zero rewards with means sampled uniformly in $(0,1)$. The values for these parameters are shown in Table 3a.

Table 3: Experimental setting.

(a) Environment parameters

States $\mathcal{S}$	10^5
Actions $\mathcal{A}$	5
Number B of successors	2
Reward sparsity	0.5

(b) MDP-GapE parameters

Discount factor γ	0.7
Confidence level δ	0.1
Exploration function $\beta_r(n_h^t, \delta)$	$\log \frac{1}{\delta} + \log n_h^t$
Exploration function $\beta_p(n_h^t, \delta)$	$\log \frac{1}{\delta} + \log n_h^t$

The main objective of our numerical experiments is to empirically verify several properties of MDP-GapE, but we acknowledge that these experiments have some limitations. Planning algorithms such as MDP-GapE are usually intended for the case $(BK)^{H-1} \ll SA$, which does not hold in our experiments (despite the large state space). Moreover, we use tighter threshold functions than those prescribed by theory, as is sometimes done in the bandit literature. These choices of thresholds are still inspired by our theoretical results, for their scaling in $n_h^t(s,a)$, un-doing a few union bounds that were found to be conservative in practice.

Fixed-confidence: Correction and sample complexity We verify empirically that MDP-GapE is (ε, δ)-correct while stopping with a reasonable number of oracle calls. Table 3b shows the choice of parameters for the algorithm. For various values of the desired accuracy ε and of the corresponding planning horizon $H = \lceil \log_\gamma(\varepsilon(1-\gamma)/2) \rceil$ (see Section 2), we run simulations on 200 random MDPs. We report in Table 4 the distribution of the number $n = \tau H$ of oracle calls and the simple regret $\bar{r}_n(\hat{a}_n)$ of MDP-GapE over these 200 runs. We first observe that MDP-GapE satisfies $\bar{r}_n(\hat{a}_n) < \varepsilon$ in all simulations, despite the use of smaller exploration functions compared to those prescribed in Lemma 2. We then compare the empirical sample complexity of MDP-GapE to the number of samples that Sparse Sampling would use. The sample complexity of Sparse Sampling with parameter C (number of calls to the generative model in each node) is $(K^{H+1} - K)/(K-1)$ for $C = 1$ and of order $\sum_{h=0}^{H-1}[(KC) \times (K(\min(B,C)))^h]$ for larger values of C. Thus, beyond very small C, the runtime of SS is prohibitively too large to try the algorithm in our setting (larger than $(BK)^H = 10^H$). For $C = 1$, the sample complexity of SS is 2.0×10^4, 4.9×10^5 and 1.2×10^7 in the 3 experiments in Table 4, which is larger than the maximal sample complexity observed for MDP-GapE.

Table 4: Simple regret and number of oracle calls, collected on 200 simulations

ε	H	MDP-GapE		
		max r_n	median n	max n
1	6	3.6×10^{-2}	8.6×10^3	1.8×10^4
0.5	8	5.2×10^{-3}	7.3×10^4	2.0×10^5
0.2	10	0	5.0×10^5	2.3×10^6

Scaling in ε As discussed above, Corollary 1 with the aforementioned choice of the planning horizon, yields a crude sample complexity bound of order $\tilde{\mathcal{O}}\left(\varepsilon^{-[2+\log(BK)/\log(1/\gamma)]}\right) = \tilde{\mathcal{O}}\left((1/\varepsilon)^{8.4}\right)$ in our experimental setting. However, we observe that the empirical exponent can be much smaller in practice. To see that, we plot in in log-log scale in Figure 1 the sample complexity n as a function of $1/(\Delta \vee \varepsilon)$ when running MDP-GapE for 5 different values of ε and 200 random MDPs for each

[1]The source code of our experiments is available at https://eleurent.github.io/planning-gap-complexity/

value (each dot correponds to one value of ε and one MDP). The 5 vertical groups of dots correspond to the 5 values of ε and to MDPs for which $\Delta \vee \varepsilon = \varepsilon$. In particular, by measuring the slope of the curve we obtain that the maximal sample complexity among those MDPs scales in $n \simeq \mathcal{O}\big((1/\varepsilon)^{3.0} \big)$. Dots that are between the vertical groups correspond to MDPs for which the smallest gap Δ was larger than ε, and for which the sample complexity is typically smaller than this worst-case value.

Comparison to the state of the art In the fixed-confidence setting, most existing algorithms are considered theoretical and cannot be applied to practical cases. For instance, for our problem with $K = 5$ and $\varepsilon = 1$, `Sparse Sampling` [19] and `SmoothCruiser` [14] both require a fixed budget[2] of at least $n_{\text{SS}} = 8 \times 10^9$. Likewise, `Trailblazer` [13] is a recursive algorithm which did not terminate in our setting. We did not implement `StOP` [28] as it requires to store a tree of policies, which is very costly even for moderate horizons. In comparison, Table 4 shows that `MDP-GapE` stopped after $n = 1.8 \times 10^4$ oracle calls in the worst case. To the best of our knowledge, `MDP-GapE` is the first (ε, δ)-correct algorithm for general MDPs with an easy implementation and a reasonable running time in practice. The only planning algorithms that can be run in practice are in the fixed-budget setting, which we now consider.

Fixed-budget evaluation We compare `MDP-GapE` to three existing baselines: first, the `KL-OLOP` algorithm [21], which uses the same upper-confidence bounds on the rewards u_h^t and states values U_h^t as `MDP-GapE`, but is restricted to *open-loop* policies, i.e. sequences of actions only. Second, the `BRUE` algorithm [8] which explores uniformly and handles closed-loop policies. Third, the popular `UCT` algorithm [20], which is also closed-loop and performs optimistic exploration at all depths. `UCT` and its variants lack theoretical guarantees, but they have been shown successful empirically in many applications. For each algorithm, we tune the planning horizon H similarly to `KL-OLOP`, by dividing the available budget n into τ episodes, where τ is the largest integer such that $\tau \log \tau / (2 \log 1/\gamma) \leq n$, and choose $H = \log \tau / (2 \log 1/\gamma)$. The exploration functions are those of `KL-OLOP` and depend on τ: $\beta_r(n_h^t, \delta) = \beta_p(n_h^t, \delta) = \log(\tau)$. Again, we perform 200 simulations and report in Figure 2 the mean simple regret, along with its 95% confidence interval. We observe that `MDP-GapE` compares favourably with these baselines in the high-budget regime.

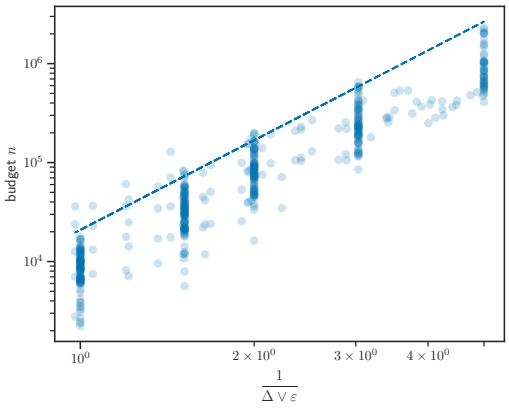

Figure 1: Dependency of the maximum number n of oracle calls with respect to $1/(\Delta \vee \varepsilon)$.

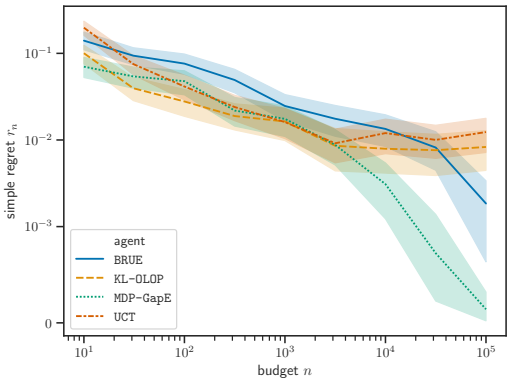

Figure 2: Comparison to other planning algorithms in a fixed-budget setting.

6 Conclusion

We proposed a new, efficient algorithm for Monte-Carlo planning in Markov Decision Processes, that combines tools from best arm identification and optimistic planning and exploits tight confidence regions on mean rewards and transitions probabilities. We proved that `MDP-GapE` attains the smallest existing gap-dependent sample complexity bound for general MDPs with stochastic rewards and transitions, when the branching factor B is finite. In future work, we will investigate the worst-case complexity of `MDP-GapE`, that is try to derive an upper bound on its sample complexity that only features ε and some appropriate notion of near-optimality dimension.

[2] In non-regularized MDPs, `SmoothCruiser` has the same sample complexity as `Sparse Sampling`.

Acknowledgments and Disclosure of Funding

We acknowledge the support of the European CHIST-ERA project DELTA and the French ANR project BOLD (ANR-19-CE23-0026-04). Anders Jonsson is partially supported by the Spanish grants TIN2015-67959 and PCIN-2017-082.

References

[1] C. Browne, E. Powley, D. Whitehouse, S. Lucas, P. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in games,*, 4(1):1–49, 2012.

[2] S Bubeck and R Munos. Open loop optimistic planning. In *Conference on Learning Theory*, 2010.

[3] Lucian Busoniu and Rémi Munos. Optimistic planning for Markov decision processes. In *Artificial Intelligence and Statistics*, pages 182–189, 2012.

[4] O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

[5] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[6] 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.

[7] Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933, 2004.

[8] Zohar Feldman and Carmel Domshlak. Simple Regret Optimization in Online Planning for Markov Decision Processes. *Journal of Artifial Intelligence Research*, 51:165–205, 2014.

[9] S. Filippi, O. Cappé, and A. Garivier. Optimism in Reinforcement Learning and Kullback-Leibler Divergence. In *Allerton Conference on Communication, Control, and Computing*, 2010.

[10] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

[11] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

[12] Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv preprint arXiv:1805.05071*, 2018.

[13] J.-B. Grill, M. Valko, and R. Munos. Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning. In *Neural Information Processing Systems (NIPS)*, 2016.

[14] Jean-Bastien Grill, Omar Darwiche Domingues, Pierre Ménard, Rémi Munos, and Michal Valko. Planning in entropy-regularized Markov decision processes and games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[15] Jean-Francois Hren and Rémi Munos. Optimistic planning of deterministic systems. In *European Workshop on Reinforcement Learning*, 2008.

[16] Ruitong Huang, Mohammad M. Ajallooeian, Csaba Szepesvári, and Martin Müller. Structured Best Arm Identification with Fixed Confidence. In *International Conference on Algorithmic Learning Theory (ALT)*, 2017.

[17] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[18] Emilie Kaufmann and Wouter M. Koolen. Monte-Carlo tree search by best arm identification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[19] Michael J. Kearns, Yishay Mansour, and Andrew Y. Ng. A Sparse Sampling Algorithm for Near-Optimal Planning in Large Markov Decision Processes. *Machine Learning*, 49(2-3):193–208, 2002.

[20] Levente Kocsis and Csaba Szepesvári. Bandit Based Monte-Carlo Planning. In *Proceedings of the 17th European Conference on Machine Learning (ECML)*, 2006.

[21] Edouard Leurent and Odalric-Ambrym Maillard. Practical open-loop optimistic planning. In *Proceedings of the 19th European Conference on Machine Learning and Principles and Practice (ECML-PKDD)*, 2019.

[22] S. Mannor and J. Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, pages 623–648, 2004.

[23] R. Munos. *From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning*, volume 7. Foundations and Trends in Machine Learning, 2014.

[24] Tom Pepels, Tristan Cazenave, Mark H. M. Winands, and Marc Lanctot. Minimizing Simple and Cumulative Regret in Monte-Carlo Tree Search. In *Third Workshop on Computer Games (CGW)*, pages 1–15, 2014.

[25] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *arXiv:1911.08265*, 2019.

[26] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362, 2018.

[27] Max Simchowitz and Kevin G Jamieson. Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs. In *Advances in Neural Information Processing Systems 32*, pages 1153–1162, 2019.

[28] B. Szorenyi, G. Kedenburg, and R. Munos. Optimistic Planning in Markov Decision Processes using a generative model. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[29] David Tolpin and Solomon Eyal Shimony. MCTS based on simple regret. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.*, 2012.

[30] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7304–7312, 2019.

A Detailed Algorithm

In this section we provide a detailed algorithm for MDP-GapE, namely Algorithm 1.

Algorithm 1 MDP-GapE

Input: confidence level δ, tolerance ε
initialize data lists $\mathcal{D}_h \leftarrow [\,]$ for all $h \in [H]$
for $t = 1 \dots$ **do**
  //Update confidence bounds
  $U_h^{t-1}, L_h^{t-1} \leftarrow \texttt{UpdateBounds}(t, \delta, \mathcal{D}_h)$
  if $U_1^{t-1}(s_1, c^t) - L_1^{t-1}(s_1, b^t) \leq \varepsilon$ **then**
    return b_{t-1}, **break**
  end if
  // Best
  $b^{t-1} \leftarrow \operatorname{argmin}_b \left[\max_{a \neq b} U_1^{t-1}(s_1, a) - L_1^{t-1}(s_1, b)\right]$
  //Challenger
  $c^{t-1} \leftarrow \operatorname{argmax}_{c \neq b^t} U_1^{t-1}(s_1, c)$
  //Exploration
  $a_1^t \leftarrow \operatorname{argmax}_{a \in \{b^{t-1}, c^{t-1}\}} \left[U_1^{t-1}(s_1, a) - L_1^{t-1}(s_1, a)\right]$
  observe reward r_1^t, next state s_2^t, save $\mathcal{D}_1.\text{append}(s_1^t, a_1^t, s_2^t, r_1^t)$
  for step $h = 2, \dots, H$ **do**
    $a_h^t \leftarrow \operatorname{argmax}_a U_h^{t-1}(s_h^t, a)$
    observe reward r_{h-1}^t, next state s_h^t,save $\mathcal{D}_h.\text{append}(s_h^t, a_h^t, s_{h+1}^t, r_h^t)$
  end for
end for

Implementation details There are different ways to store and update the confidence bounds on the Q-value (that is, to specify the UpdateBounds subroutine) according to how we merge information across states.

The most obvious one, suggested by previous work [2, 21, 3] (and also implemented for our experiments) does not merge information at all and builds a search *tree* in which a node (s_h, a_h) at depth h is identified with the sequence of h states and actions that leads to it. It leads to a very simple update: after each trajectory, one only needs to update the confidence bounds, $U_h(s_h, a_h)$ and $L_h(s_h, a_h)$, of the visited action-state pairs. Another option is to merge information for the same states and a fixed depth. But in this case the search tree becomes a *graph* and after each trajectory we need to re-compute the values $U_h(s_h, a_h)$ for all stored state action pairs (s_h, a_h) at each depth.

B Correctness of MDP-GapE

In this section we prove the correctness of MDP-GapEunder the assumption that the event $\mathcal{E}^r \cap \mathcal{E}^p$ holds. Concretely, we prove by induction that

$$\mathcal{E}^r \cap \mathcal{E}^p \subseteq \bigcap_{t \in \mathbb{N}^\star} \bigcap_{h=1}^{H} \Big[\bigcap_{s_h, a_h} \Big(Q_h(s_h, a_h) \in \big[L_h^t(s_h, a_h), U_h^t(s_h, a_h)\big] \Big) \Big].$$

The base case is given by $h = H + 1$, in which case by our previous convention,

$$L_{H+1}^t(\cdot, \cdot) = Q_{H+1}(\cdot, \cdot) = U_{H+1}^t(\cdot, \cdot) = 0.$$

For the inductive case, assume that the inclusion holds at depth $h+1$. Then we have

$$\begin{aligned}
L_h^t(s_h,a_h) &= \ell_h^t(s_h,a_h) + \gamma \min_{p\in\mathcal{C}_h^t(s_h,a_h)} \sum_{s'} p(s'|s_h,a_h) \max_{a'} L_{h+1}^t(s',a') \\
&\leq \ell_h^t(s_h,a_h) + \gamma \sum_{s'} p_h(s'|s_h,a_h) \max_{a'} L_{h+1}^t(s',a') \\
&\leq r_h(s_h,a_h) + \gamma \sum_{s'} p_h(s'|s_h,a_h) Q_{h+1}(s', \arg\max_{a'} L_{h+1}^t(s',a')) \\
&\leq r_h(s_h,a_h) + \gamma \sum_{s'} p_h(s'|s_h,a_h) Q_{h+1}(s', \pi_{h+1}^*(s')) = Q_h(s_h,a_h) \\
&\leq u_h^t(s_h,a_h) + \gamma \sum_{s'} p_h(s'|s_h,a_h) \max_{a'} U_{h+1}^t(s',a') \\
&\leq u_h^t(s_h,a_h) + \gamma \max_{p\in\mathcal{C}_h^t(s_h,a_h)} \sum_{s'} p(s'|s_h,a_h) \max_{a'} U_{h+1}^t(s',a') = U_h^t(s_h,a_h),
\end{aligned}$$

where we have used $r_h(s_h,a_h) \in \left[\ell_h^t(s_h,a_h), u_h^t(s_h,a_h)\right]$ and $p_h(\cdot|s_h,a_h) \in \mathcal{C}_h^t(s_h,a_h)$.

C Concentration Events

In this section we prove that the event $\mathcal{E}$ holds with high probability. But before we need several concentration inequalities.

C.1 Deviation Inequality for Categorical Distributions

Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. samples from a distribution supported over $\{1,\ldots,m\}$, of probabilities given by $p \in \Sigma_m$, where Σ_m is the probability simplex of dimension $m-1$. We denote by $\widehat{p}_n$ the empirical vector of probabilities, i.e. for all $k \in \{1,\ldots,m\}$

$$\widehat{p}_{n,k} = \frac{1}{n} \sum_{\ell=1}^{n} \mathbb{1}(X_\ell = k)\,.$$

Note that an element $p \in \Sigma_m$ will sometimes be seen as an element of $\mathbb{R}^{m-1}$ since $p_m = 1 - \sum_{k=1}^{m-1} p_k$. This should be clear from the context. We denote by $H(p)$ the (Shannon) entropy of $p \in \Sigma_m$,

$$H(p) = \sum_{k=1}^{m} p_k \log(1/p_k)\,.$$

Proposition 1. *For all $p \in \Sigma_m$, for all $\delta \in [0,1]$,*

$$\mathbb{P}\left(\exists n \in \mathbb{N}^*,\, n\,\mathrm{KL}(\widehat{p}_n, p) > \log(1/\delta) + (m-1)\log\left(e(1+n/(m-1))\right)\right) \leq \delta\,.$$

Proof. We apply the method of mixture with a Dirichlet prior on the mean parameter of the exponential family formed by the set of categorical distribution on $\{1,\ldots,m\}$. Letting

$$\varphi_p(\lambda) = \log \mathbb{E}_{X\sim p}\left[e^{\lambda X}\right] = \log(p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k}),$$

be the log-partition function, the following quantity is a martingale:

$$M_n^\lambda = e^{n\langle \lambda, \widehat{p}_n\rangle - n\varphi_p(\lambda)}.$$

We set a Dirichlet prior $q \sim \mathcal{D}\mathrm{ir}(\alpha)$ with $\alpha \in {\mathbb{R}_+^*}^m$ and for $\lambda_q = (\nabla\varphi_p)^{-1}(q)$ and consider the integrated martingale

$$\begin{aligned}
M_n &= \int M_n^{\lambda_q} \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} q_k^{\alpha_k - 1} \,\mathrm{d}q \\
&= \int e^{n\left(\mathrm{KL}(\widehat{p}_n, p) - \mathrm{KL}(\widehat{p}_n, q)\right)} \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} q_k^{\alpha_k - 1} \,\mathrm{d}q \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} \int \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} q_k^{n\widehat{p}_{n,k} + \alpha_k - 1} \,\mathrm{d}q \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} \frac{\prod_{k=1}^m \Gamma(\alpha_k + n\widehat{p}_{n,k})}{\Gamma\Big(\sum_{k=1}^m \alpha_k + n\Big)} \,,
\end{aligned}$$

where in the second inequality we used Lemma 3. Now we choose the uniform prior $\alpha = (1, \dots, 1)$. Hence we get

$$\begin{aligned}
M_n &= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} (m-1)! \frac{\prod_{k=1}^m \Gamma(1 + n\widehat{p}_{n,k})}{\Gamma(m+n)} \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} (m-1)! \frac{\prod_{k=1}^m (n\widehat{p}_{n,k})!}{n!} \frac{n!}{(m+n-1)!} \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} \frac{1}{\binom{n}{n\widehat{p}_n}} \frac{1}{\binom{m+n-1}{m-1}} .
\end{aligned}$$

Thanks to Theorem 11.1.3 by [5] we can upper bound the multinomial coefficient as follows: for $M \in \mathbb{N}^*$ and $x \in \{0, \dots, M\}^m$ such that $\sum_{k=1}^m x_k = M$ it holds

$$\binom{M}{x} = \frac{M!}{\prod_{k=1}^m x_k!} \le e^{MH(x/M)} \,.$$

Using this inequality we obtain

$$\begin{aligned}
M_n &\ge e^{n\,\mathrm{kl}(\widehat{p}_n, p) + nH(\widehat{p}_n) - nH(\widehat{p}_n) - (m+n-1)H\left((m-1)/(m+n-1)\right)} \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) - (m+n-1)H\left((m-1)/(m+n-1)\right)} \,.
\end{aligned}$$

It remains to upper-bound the entropic term

$$\begin{aligned}
(m+n-1)H\big((m-1)/(m+n-1)\big) &= (m-1)\log\frac{m+n-1}{m-1} + n\log\frac{m+n-1}{n} \\
&\le (m-1)\log\big(1 + n/(m-1)\big) + n\log(1 + (m-1)/n) \\
&\le (m-1)\log\big(1 + n/(m-1)\big) + (m-1) \,.
\end{aligned}$$

Thus we can lower bound the martingale as follows

$$M_n \ge e^{n\,\mathrm{KL}(\widehat{p}_n, p)} \big(e(1 + n/(m-1))\big)^{m-1} \,.$$

Using the fact that, for any supermartingale it holds that

$$\mathbb{P}\left(\exists n \in \mathbb{N}^* : M_n > 1/\delta\right) \le \delta \mathbb{E}[M_1], \tag{6}$$

which is a well-known property used in the method of mixtures (see [7]), we conclude that

$$\mathbb{P}\left(\exists n \in \mathbb{N}^*,\, n\,\mathrm{KL}(\widehat{p}_n, p) > (m-1)\log\big(e(1 + n/(m-1))\big) + \log(1/\delta)\right) \le \delta \,.$$

□

Lemma 3. *For $q, p \in \Sigma_m$ and $\lambda \in \mathbb{R}^{m-1}$,*

$$\langle \lambda, q \rangle - \varphi_p(\lambda) = \mathrm{KL}(q, p) - \mathrm{KL}(q, p^\lambda) \,,$$

where $\varphi_p(\lambda) = \log(p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k})$ and $p^\lambda = \nabla\varphi_{p_0}(\lambda)$.

Proof. There is a more general way than the ad hoc one below to prove the result. First note that

$$p_k^\lambda = \frac{p_k e^{\lambda_k}}{p_m + \sum_{\ell=1}^{m-1} p_\ell e^{\lambda_\ell}},$$

which implies that

$$p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k} = \frac{p_m}{p_m^\lambda}, \qquad \lambda_k = \log \frac{p_k^\lambda}{p_k} + \log \frac{p_m}{p_m^\lambda}.$$

Therefore we get

$$\begin{aligned}
\langle \lambda, q \rangle - \varphi_p(\lambda) &= \sum_{k=1}^{m-1} q_k \log \left(\frac{p_k^\lambda}{p_k} \frac{p_m}{p_m^\lambda} \right) - \log \left(p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k} \right) \\
&= \sum_{k=1}^{m-1} q_k \log \frac{p_k^\lambda}{p_k} + (1 - q_m) \log \frac{p_m}{p_m^\lambda} - \log \frac{p_m}{p_m^\lambda} \\
&= \sum_{k=1}^{m} q_k \log \frac{p_k^\lambda}{p_k} = \mathrm{KL}(q, p) - \mathrm{KL}(q, p^\lambda).
\end{aligned}$$

□

C.2 Deviation Inequality for Bounded Distribution

Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. samples from a distribution ν of mean μ supported on $[0, 1]$. We denote by $\widehat{\mu}_n$ the empirical mean

$$\widehat{\mu}_n = \frac{1}{n} \sum_{\ell=1}^{n} X_\ell .$$

It is well known, see [11], that we can "project" the distribution ν on a Bernoulli distribution with the same mean and then use deviation inequality for Bernoulli to concentrate the empirical mean. This method dos not lead to the sharpest confidence intervals but it provides a good trade-off between complexity computation and accuracy.

Proposition 2. *For all distribution ν of mean μ supported on the unit interval, for all $\delta \in [0, 1]$,*

$$\mathbb{P}\left(\exists n \in \mathbb{N}^*,\, n\, \mathrm{kl}(\widehat{\mu}_n, \mu) > \log(1/\delta) + \log\big(e(1+n)\big) \right) \leq \delta .$$

Proof. First note that we can upper bound the log-partition function of ν by the one of a Bernoulli $\mathcal{B}\mathrm{er}(\mu)$, for all $\lambda \in \mathbb{R}$,

$$\log\left(\mathbb{E}[e^{\lambda X_n}]\right) \leq \log\left(\mathbb{E}[X_n e^\lambda + 1 - X_n]\right) = \log\left(1 - \mu + \mu e^\lambda\right) = \varphi_\mu(\lambda).$$

Then we can follow the proof of Proportion 1 with $m = 2$ and where M_n^λ is only a supermartingale but this does not change the result as the property (6) still holds. Thus the proposition follows by specifying Proposition 1 to the case $m = 2$. □

C.3 Deviation Inequality for sequence of Bernoulli Random Variables

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of Bernoulli random variables adapted to the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$. We restate here Lemma F.4. of [6].

Proposition 3. *If we denote $p_n = \mathbb{P}(X_n = 1 | \mathcal{F}_{n-1})$, then for all $\delta \in (0, 1]$*

$$\mathbb{P}\left(\exists n \in \mathbb{N}^* : \sum_{\ell=1}^{n} X_\ell < \sum_{\ell=1}^{n} p_\ell/2 - \log(1/\delta) \right) \leq \delta .$$

C.4 Proof of Lemma 2

We just prove that each event forming $\mathcal{E} = \mathcal{E}^r \cap \mathcal{E}^p \cap \mathcal{E}^n$ holds with high probability. For the first one using Proposition 2, since the reward are bounded in the unit interval we have

$$\begin{aligned}\mathbb{P}\big((\mathcal{E}^r)^c\big) &\leq \sum_{h\in[H]}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\big(\exists t\in\mathbb{N}^* :\ n_h^t(s_h,a_h)\,\mathrm{kl}\big(\hat{r}_h^t(s_h,a_h), r_h(s_h,a_h)\big) > \beta_r(n_h^t(s_h,a_h),\delta)\big)\\ &\leq \sum_{h\in[H]}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \frac{\delta}{3AS^H} \leq \frac{\delta}{3}\,.\end{aligned}$$

where we used Doob's optional skipping in the second inequality in order to apply Proposition 2, see Section 4.1 of [12]. Similarly for the confidence regions for the probabilities transitions, using Proposition 1 we obtain

$$\begin{aligned}\mathbb{P}\big((\mathcal{E}^p)^c\big) &\leq \sum_{h\in[H]}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\big(\exists t\in\mathbb{N}^* :\ n_h^t(s_h,a_h)\,\mathrm{KL}\big(\widehat{p}_h^t(\cdot|s_h,a_h), p_h(\cdot|s_h,a_h)\big) > \beta_p(n_h^t(s_h,a_h),\delta)\big)\\ &\leq \sum_{h\in[H]}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \frac{\delta}{3AS^H} \leq \frac{\delta}{3}\,.\end{aligned}$$

It remains to control the counts, using Proposition 3,

$$\begin{aligned}\mathbb{P}\big((\mathcal{E}^{\mathrm{cnt}})^c\big) &\leq \sum_{h\in[H]}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists t\in\mathbb{N}^* :\ n_h^t(s_h,a_h) < \frac{1}{2}\bar{n}_h^t(s_h,a_h) - \beta^{\mathrm{cnt}}(\delta)\right)\\ &\leq \sum_{h\in[H]}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \frac{\delta}{3AS^H} \leq \frac{\delta}{3}\,,\end{aligned}$$

where we used that by definition of the pseudo-counts

$$\bar{n}_h^t(s_h,a_h) = \sum_{\ell=1}^{t} \mathbb{P}\left((s_h^\ell, a_h^\ell) = (s_h,a_h)|\mathcal{F}_{\ell-1}\right),$$

and $\mathcal{F}_{\ell-1}$ is the information available to the agent at step ℓ. An union bound allows us to conclude

$$\mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}\big((\mathcal{E}^r)^c\big) + \mathbb{P}\big((\mathcal{E}^p)^c\big) + \mathbb{P}\big((\mathcal{E}^{\mathrm{cnt}})^c\big) \leq \delta\,.$$

D Proof of Theorem 1

In this section we present the proof of Theorem 1, which relies on three important ingredients. The first ingredient is Lemma 5 in Appendix D.1, which provides a relationship between the state-action gaps and the diameter $D_h^t(s_h,a_h) := U_h^t(s_h,a_h) - L_h^t(s_h,a_h)$ of the confidence intervals. The second ingredient is Lemma 8 in Appendix D.2, which provides an upper bound on the diameter $D_h^t(s_h,a_h)$. The third ingredient is Lemma 9 in Appendix D.3, which relates the actual counts of state-action pairs to the corresponding pseudo-counts. After providing these ingredients, we present the detailed proof of Theorem 1 in Appendix D.4.

D.1 Relating state-action gaps to diameters

Before stating Lemma 5, we prove an important property of the `UGapE` algorithm. We recall that b^t and c^t are the candidate best action and its challenger, defined as

$$\begin{aligned}b^t &= \operatorname*{argmin}_b \left[\max_{a\neq b} U_1^t(s_1,a) - L_1^t(s_1,b)\right],\\ c^t &= \operatorname*{argmax}_{c\neq b^t} U_1^t(s_1,c).\end{aligned}$$

The policy at the root is then defined as $\pi_1^{t+1}(s_1) = \operatorname*{argmax}_{b\in\{b^t,c^t\}} [U_1^t(s_1,b) - L_1^t(s_1,b)]$.

Lemma 4. *For all $t \in [\tau_\delta - 1]$, the following inequalities hold:*

1. $U_1^t(s_1, c^t) - L_1^t(s_1, b^t) \leq U_1^t\left(s_1, \pi^{t+1}(s_1)\right) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right)$,
2. $U_1^t(s_1, b^t) - L_1^t(s_1, c^t) < 2\left[U_1^t\left(s_1, \pi^{t+1}(s_1)\right) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right)\right]$.

Proof. We show the first part by contradiction. If the inequality does not hold, we obtain

$$\begin{aligned}
U_1^t(s_1, b^t) - L_1^t(s_1, b^t) &\leq U_1^t(s_1, \pi_1^{t+1}(s_1)) - L_1^t(s_1, \pi_1^{t+1}(s_1)) < U_1^t(s_1, c^t) - L_1^t(s_1, b^t), \\
U_1^t(s_1, c^t) - L_1^t(s_1, c^t) &\leq U_1^t(s_1, \pi_1^{t+1}(s_1)) - L_1^t(s_1, \pi_1^{t+1}(s_1)) < U_1^t(s_1, c^t) - L_1^t(s_1, b^t) \\
&= \max_{a \neq b^t} U_1^t(s_1, a) - L_1^t(s_1, b^t) \leq \max_{a \neq c^t} U_1^t(s_1, a) - L_1^t(s_1, c^t),
\end{aligned}$$

where the last inequality follows from the definition of b^t. Combining the two inequalities yields $U_1^t(s_1, b^t) < U_1^t(s_1, c^t) < \max_{a \neq c^t} U_1^t(s_1, a)$, which contradicts the definition of c^t.

For the second part, if $t < \tau_\delta$ then the algorithm has not yet stopped, implying

$$\begin{aligned}
U_1^t(s_1, b^t) - L_1^t(s_1, c^t) &= U_1^t(s_1, b^t) - L_1^t(s_1, b^t) + U_1^t(s_1, c^t) - L_1^t(s_1, c^t) \\
&\quad - \left[U_1^t(s_1, c^t) - L_1^t(s_1, b^t)\right] \\
&< 2\left[U_1^t(s_1, \pi_1^{t+1}(s_1)) - L_1^t(s_1, \pi_1^{t+1}(s_1)\right] - \varepsilon.
\end{aligned}$$

□

As a consequence of Lemma 4, we can upper bound any confidence interval involving b^t and c^t.

Corollary 3. *For each pair of actions $a, a' \in \{b^t, c^t\}$, it holds that*

$$U_1^t(s_1, a) - L_1^t(s_1, a') \leq 2\left[U_1^t\left(s_1, \pi^{t+1}(s_1)\right) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right)\right].$$

We are now ready to state Lemma 5.

Lemma 5. *If $\mathcal{E}$ holds and $t < \tau_\delta$, for all $h \in [H]$ and $s_h \in \mathcal{S}_h(\pi^{t+1})$,*

$$\tilde{\Delta}_h(s_h, \pi_h^{t+1}(s_h)) \leq 2\left[U_h^t(s_h, \pi_h^{t+1}(s_h)) - L_h^t(s_h, \pi_h^{t+1}(s_h))\right].$$

Proof. The proof for $h \in [2, H]$ is immediate from the correctness of the confidence bounds implied by $\mathcal{E}$, and the fact that the selection is optimistic:

$$\begin{aligned}
\Delta_h(s_h, \pi_h^{t+1}(s_h)) &= Q_h(s_h, \pi_h^\star(s_h)) - Q_h(s_h, \pi_h^{t+1}(s_h)) \\
&\leq \max_a U_h^t(s_h, a) - L_h^t(s_h, \pi_h^{t+1}(s_h)) = U_h^t(s_h, \pi_h^{t+1}(s_h)) - L_h^t(s_h, \pi_h^{t+1}(s_h)).
\end{aligned}$$

For $h = 1$, we prove separately that each term in the max is smaller that the right hand side of desired inequality, that is

$$\max\left(\Delta_1\left(s_1, \pi^{t+1}(s_1)\right); \Delta; \varepsilon\right) \leq 2\left[U_h^t(s_h, \pi_h^{t+1}(s_h)) - L_h^t(s_h, \pi_h^{t+1}(s_h))\right].$$

Now, by definition of the stopping rule, if $t < \tau_\delta$, $U_1^t(s_1, c^t) - L_1^t(s_1, b^t) > \varepsilon$. Using the first property in Lemma 4 yields

$$\varepsilon < U_1^t\left(s_1, \pi^{t+1}(s_1)\right) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right). \tag{7}$$

Then, exploiting the fact that the action with largest UCB is either b^t or c^t, it holds on $\mathcal{E}$ that

$$\begin{aligned}
\Delta_1\left(s_1, \pi^{t+1}(s_1)\right) &= Q_1(s_1, a^\star) - Q_1\left(s_1, \pi^{t+1}(s_1)\right) \\
&\leq \max_a U_1^t(s_1, a) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right) \\
&= \max_{a \in \{b^t, c^t\}} U_1^t(s_1, a) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right).
\end{aligned}$$

Using Corollary 3 to further upper bound the right hand side yields

$$\Delta_1\left(s_1, \pi^{t+1}(s_1)\right) < 2\left[U_1^t\left(s_1, \pi^{t+1}(s_1)\right) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right)\right]. \tag{8}$$

Finally, one can also write, on the event $\mathcal{E}$,

$$\begin{aligned}\Delta = \min_{a \neq a^\star} \left[Q_1(s_1, a^\star) - Q_1(s_1, a)\right] &\leq U_1^t(s_1, a^\star) - \max_{a \neq a^\star} Q_1(s_1, a) \\ &\leq \max_{a' \in \{b^t, c^t\}} U_1^t(s_1, a') - \min_{a \in \{b^t, c^t\}} Q_1(s_1, a) \\ &\leq \max_{a' \in \{b^t, c^t\}} U_1^t(s_1, a') - \min_{a \in \{b^t, c^t\}} L_1^t(s_1, a).\end{aligned}$$

In each of the four possible choices of (a, a'), Corollary 3 implies that

$$\Delta \leq 2\left[U_1^t\left(s_1, \pi^{t+1}(s_1)\right) - L_1^t\left(s_1, \pi^{t+1}(s_1)\right)\right]. \tag{9}$$

Lemma 5 follows by combining (7), (8) and (9) with the definition of $\Delta_1^\star\left(s_1, \pi^{t+1}(s_1)\right)$. □

D.2 Upper bounding the diameters

In this section we state and prove Lemma 8. We use the notation $\sigma_h = \sum_{i=0}^{h-1} \gamma^i$ to upper bound the discounted reward in h steps. As a first step, we prove the following auxiliary lemma.

Lemma 6. *If $\mathcal{E}$ holds, for each $h \in [H]$, each (s_h, a_h) and each $q \in \mathcal{C}_h^t(s_h, a_h)$,*

$$\sum_{s'} \left(q(s'|s_h, a_h) - p_h(s'|s_h, a_h)\right) U_{h+1}^t(s', \pi_{h+1}^{t+1}(s')) \leq 2\sqrt{2}\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}}.$$

Proof. First note that for each state s', $U_{h+1}^t(s', \pi_{h+1}^{t+1}(s'))$ can be expressed as an expectation on the form $\mathbb{E}^{\pi^{t+1}}\left\{\sum_{i=h+1}^H \gamma^{i-h-1} u_i^t(s_i, a_i) \mid s_{h+1} = s'\right\}$, which is upper bounded by $\sum_{i=h+1}^H \gamma^{i-h-1} = \sigma_{H-h}$ since $u_i^t(s_i, a_i) \leq 1$ for each (s_i, a_i). Note that for $h = H$, $\sigma_{H-H} = \sigma_0 = 0$. If $n_h^t(s_h, a_h) = 0$ the result trivially holds by the conventions adopted for the confidence bounds and regions. Now, if $n_h^t(s_h, a_h) > 0$, we have

$$\begin{aligned}&\sum_{s'} \left(q(s'|s_h, a_h) - p_h(s'|s_h, a_h)\right) U_{h+1}^t(s', \pi_{h+1}^{t+1}(s')) \\ &\quad\leq \|q(\cdot|s_h, a_h) - p_h(\cdot|s_h, a_h)\|_1 \, \|U_{h+1}^t(\cdot, \pi_{h+1}^{t+1}(\cdot))\|_\infty \\ &\quad\leq \sigma_{H-h}\left(\|q(\cdot|s_h, a_h) - \hat{p}_h^t(\cdot|s_h, a_h)\|_1 + \|p_h(\cdot|s_h, a_h) - \hat{p}_h^t(\cdot|s_h, a_h)\|_1\right) \\ &\quad\leq \sigma_{H-h}\left(\sqrt{2\,\mathrm{KL}(\hat{p}_h^t(\cdot|s_h, a_h), q(\cdot|s_h, a_h))} + \sqrt{2\,\mathrm{KL}(\hat{p}_h^t(\cdot|s_h, a_h), p_h(\cdot|s_h, a_h))}\right) \\ &\quad\leq 2\sqrt{2}\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}},\end{aligned}$$

where we have used Pinsker's inequality to bound the L^1-norm using the KL divergence, combined with the fact that both q and p are close to the empirical transition probabilities $\hat{p}^t$ under $\mathcal{E}$. □

As a consequence, we can express the upper bound U^t in terms of the true transition probabilities p.

Corollary 4. *If $\mathcal{E}$ holds, for each $h \in [H]$ and each (s_h, a_h),*

$$U_h^t(s_h, a_h) \leq u_h^t(s_h, a_h) + \gamma \sum_{s'} p_h(s'|s_h, a_h) U_{h+1}^t(s', \pi_{h+1}^{t+1}(s')) + 2\sqrt{2}\gamma\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}}.$$

We can also express the lower bound L^t in terms of the transition probabilities p and policy π^{t+1}.

Lemma 7. *If $\mathcal{E}$ holds, for each $h \in [H]$ and each (s_h, a_h),*

$$L_h^t(s_h, a_h) \geq \ell_h^t(s_h, a_h) + \gamma \sum_{s'} p_h(s'|s_h, a_h) L_{h+1}^t(s', \pi_{h+1}^{t+1}(s')) - 2\sqrt{2}\gamma\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}}.$$

Proof. We exploit the fact that for each $h \in [H]$, each (s_h, a_h) and each $q \in \mathcal{C}_h^t(s_h, a_h)$,

$$\sum_{s'} (q(s'|s_h, a_h) - p_h(s'|s_h, a_h)) \max_{a'} L_{h+1}^t(s', a') \geq -2\sqrt{2}\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}}.$$

The proof is analogous to the proof of Lemma 6. We can now write

$$\begin{aligned}
L_h^t(s_h, a_h) &= \ell_h^t(s_h, a_h) + \gamma \min_{p \in \mathcal{C}_h^t(s_h, a_h)} \sum_{s'} p(s'|s_h, a_h) \max_{a'} L_{h+1}^t(s', a') \\
&\geq \ell_h^t(s_h, a_h) + \gamma \sum_{s'} p_h(s'|s_h, a_h) \max_{a'} L_{h+1}^t(s', a') - 2\sqrt{2}\gamma\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}} \\
&\geq \ell_h^t(s_h, a_h) + \gamma \sum_{s'} p_h(s'|s_h, a_h) L_{h+1}^t(s', \pi_{h+1}^{t+1}(s')) - 2\sqrt{2}\gamma\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}}.
\end{aligned}$$

□

We are now ready to state Lemma 8.

Lemma 8. *If $\mathcal{E}$ holds, for all $h \in [H]$, $s_h \in \mathcal{S}_h(\pi^{t+1})$ and a_h,*

$$D_h^t(s_h, a_h) \leq \sigma_{H-h+1}\left[4\sqrt{2}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h)}} \wedge 1\right] + \gamma \sum_{s'} p_h(s'|s_h, a_h) D_{h+1}^t(s', \pi_{h+1}^{t+1}(s')).$$

Proof. The bound on the diameter follows directly from Corollary 4 and Lemma 7:

$$\begin{aligned}
D_h^t(s_h, a_h) &= U_h^t(s_h, a_h) - L_h^t(s_h, a_h) \\
&\leq \left(u_h^t(s_h, a_h) - \ell_h^t(s_h, a_h)\right) + \gamma \sum_{s'} p_h(s'|s_h, a_h) \left(U_{h+1}^t(s', \pi_{h+1}^{t+1}(s')) - L_{h+1}^t(s', \pi_{h+1}^{t+1}(s'))\right) \\
&\quad + 4\sqrt{2}\gamma\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}} \\
&\leq 4\sqrt{2}\sigma_{H-h+1}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}} + \gamma \sum_{s'} p_h(s'|s_h, a_h) D_{h+1}^t(s', \pi_{h+1}^{t+1}(s')),
\end{aligned}$$

where we used $\mathcal{E}^r \supseteq \mathcal{E}$ and Pinsker's inequality to bound

$$u_h^t(s_h, a_h) - \ell_h^t(s_h, a_h) \leq \sqrt{\frac{2\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}} < 4\sqrt{2}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}}.$$

To obtain the final expression in Lemma 8, we observe that it also trivially holds that

$$D_h^t(s_h, a_h) \leq \sigma_{H-h+1} \leq \sigma_{H-h+1} + \gamma \sum_{s'} p_h(s'|s_h, a_h) D_{h+1}^t(s', \pi_{h+1}^{t+1}(s')),$$

hence

$$D_h^t(s_h, a_h) \leq \sigma_{H-h+1} \min\left[4\sqrt{2}\sqrt{\frac{\beta(n_h^t(s_h, a_h), \delta)}{n_h^t(s_h, a_h) \vee 1}}, 1\right] + \gamma \sum_{s'} p_h(s'|s_h, a_h) D_{h+1}^t(s', \pi_{h+1}^{t+1}(s')).$$

The conclusion follows by observing that one can get rid of the maximum with 1 in the denominator by using instead the convention $1/0 = +\infty$. □

D.3 Relating counts to pseudo-counts

We now assume that the event $\mathcal{E}$ holds and fix some $h \in [H]$ and some state-action pair (s_h, a_h). For every $\ell \geq h$, we define $p_{h,\ell}^{\pi}(s, a|s_h, a_h)$ to be the probability that starting from (s_h, a_h) in step h

and following π thereafter, we end up in (s,a) in step ℓ. We use $p^t_{h,\ell}(s,a|s_h,a_h)$ as a shorthand for $p^{\pi^t}_{h,\ell}(s,a|s_h,a_h)$.

Introducing the *conditional pseudo-counts* $\bar{n}^t_{h,\ell}(s,a;s_h,a_h) := \sum_{i=1}^t p^i_h(s_h,a_h)p^i_{h,\ell}(s,a|s_h,a_h)$ and using that on the event $\mathcal{E}^{\mathrm{cnt}} \supseteq \mathcal{E}$ the counts are close to the pseudo-counts, one can prove:

Lemma 9. *If the event $\mathcal{E}^{\mathrm{cnt}}$ holds,* $\left[\sqrt{\frac{\beta(n^t_\ell(s,a),\delta)}{n^t_\ell(s,a)}} \wedge 1\right] \le 2\sqrt{\frac{\beta(\bar{n}^t_{h,\ell}(s,a;s_h,a_h),\delta)}{\bar{n}^t_{h,\ell}(s,a;s_h,a_h)\vee 1}}$.

Proof. As the event $\mathcal{E}^{\mathrm{cnt}}$ holds, we know that for all $t < \tau$,

$$\begin{aligned} n^t_\ell(s,a) &\ge \frac{1}{2}\bar{n}^t_\ell(s,a) - \beta^{\mathrm{cnt}}(\delta) \\ &\ge \frac{1}{2}\bar{n}^t_{h,\ell}(s,a;s_h,a_h) - \beta^{\mathrm{cnt}}(\delta). \end{aligned}$$

We now distinguish two cases. First, if $\beta^{\mathrm{cnt}}(\delta) \le \frac{1}{4}\bar{n}^t_{h,\ell}(s,a;s_h,a_h)$, then

$$\sqrt{\frac{\beta(n^t_\ell(s,a),\delta)}{n^t_\ell(s,a)}} \le \sqrt{\frac{\beta\left(\frac{1}{4}\bar{n}^t_{h,\ell}(s,a;s_h,a_h),\delta\right)}{\frac{1}{4}\bar{n}^t_{h,\ell}(s,a;s_h,a_h)}} \le 2\sqrt{\frac{\beta\left(\bar{n}^t_{h,\ell}(s,a;s_h,a_h),\delta\right)}{\bar{n}^t_{h,\ell}(s,a;s_h,a_h)\vee 1}},$$

where we use that $x \mapsto \sqrt{\beta(x,\delta)/x}$ is non-increasing for $x \ge 1$, $x \mapsto \beta(x,\delta)$ is non-decreasing, and $\beta^{\mathrm{cnt}}(\delta) \ge 1$. If $\beta^{\mathrm{cnt}}(\delta) > \frac{1}{4}\bar{n}^t_{h,\ell}(s,a;s_h,a_h)$, simple algebra shows that

$$1 < 2\sqrt{\frac{\beta^{\mathrm{cnt}}(\delta)}{\bar{n}^t_{h,\ell}(s,a;s_h,a_h)\vee 1}} \le 2\sqrt{\frac{\beta(\bar{n}^t_{h,\ell}(s,a;s_h,a_h),\delta)}{\bar{n}^t_{h,\ell}(s,a;s_h,a_h)\vee 1}},$$

where we use that $\beta^{\mathrm{cnt}}(\delta) \le \beta(0,\delta)$ and $x \mapsto \beta(x,\delta)$ is non-decreasing. If $\bar{n}^t_{h,\ell}(s,a;s_h,a_h) < 1$, the expression uses the trivial bound $\beta^{\mathrm{cnt}}(\delta) > \frac{1}{4}$. In both cases, we have

$$\left[\sqrt{\frac{\beta(n^t_\ell(s,a),\delta)}{n^t_\ell(s,a)}} \wedge 1\right] \le 2\sqrt{\frac{\beta(\bar{n}^t_{h,\ell}(s,a;s_h,a_h),\delta)}{\bar{n}^t_{h,\ell}(s,a;s_h,a_h)\vee 1}}.$$

□

D.4 Detailed proof of Theorem 1

We assume that the event $\mathcal{E}$ holds and fix some $h \in [H]$ and some state-action pair (s_h,a_h). We define some notion of expected diameter in a future step ℓ given that (s_h,a_h) is visited at step h under policy π^{t+1}. For every $(h,\ell) \in [H]^2$ such that $h \le \ell$ we let

$$q^t_{h,\ell}(s_h,a_h) := \sum_{(s,a)} p^{t+1}_h(s_h,a_h)p^{t+1}_{h,\ell}(s,a|s_h,a_h)\, D^t_\ell(s,a).$$

To be more accurate, $q^t_{h,\ell}(s_h,a_h)$ is equal to the probability that (s_h,a_h) is visited by π^{t+1}, multiplied by the expected diameter of the state-action pair (s,a) that is reached at step ℓ if one applies π^{t+1} after choosing a_h in state s_h. In particular, $q^t_{h,\ell}(s_h,a_h) = 0$ if $a_h \ne \pi^{t+1}(s_h)$.

Step 1: lower bounding $q^t_{h,h}(s_h,a_h)$ in terms of the gaps From the above definition,

$$q^t_{h,h}(s_h,a_h) = p^{t+1}_h(s_h,a_h)D^t_h(s_h,a_h).$$

Using Lemma 5 and the fact that $p^{t+1}_h(s_h,a_h) = 0$ if $a_h \ne \pi^{t+1}(s_h)$ yields

$$\text{if } t < \tau, \quad q^t_{h,h}(s_h,a_h) \ge \frac{1}{2}p^{t+1}_h(s_h,a_h)\Delta_h(s_h,a_h). \tag{10}$$

Step 2: upper bounding $q^t_{h,h}(s_h,a_h)$ in terms of the counts Using Lemma 8 and the fact that

$$\sum_{(s,a)} p_h^{t+1}(s_h,a_h)p_{h,\ell}^{t+1}(s,a|s_h,a_h)\left[\sum_{(s',a')} p_\ell(s'|s,a)\mathbb{1}\left(a'=\pi_{\ell+1}^{t+1}(s')\right)D^t_{\ell+1}(s',a')\right]$$

$$= \sum_{(s',a')} p_h^{t+1}(s_h,a_h)\underbrace{\left[\sum_{(s,a)} p_{h,\ell}^{t+1}(s,a|s_h,a_h)p_\ell(s'|s,a)\mathbb{1}\left(a'=\pi_{\ell+1}^{t+1}(s')\right)\right]}_{=p^{t+1}_{h,\ell+1}(s',a'|s_h,a_h)} D^t_{\ell+1}(s',a'),$$

one can establish the following relationship between $q^t_{h,\ell}(s_h,a_h)$ and $q^t_{h,\ell+1}(s_h,a_h)$:

$$q^t_{h,\ell}(s_h,a_h) \le \sum_{(s,a)} p_h^{t+1}(s_h,a_h)p_{h,\ell}^{t+1}(s,a|s_h,a_h)\left[4\sqrt{2}\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}\wedge 1\right] + \gamma q^{t+1}_{h,\ell+1}(s_h,a_h).$$

By induction, one then obtains the following upper bound:

$$q^t_{h,h}(s_h,a_h) \le \sum_{\ell=h}^{H}\gamma^{\ell-h}\sigma_{H-\ell+1}\sum_{(s,a)} p_h^{t+1}(s_h,a_h)p_{h,\ell}^{t+1}(s,a|s_h,a_h)\left[4\sqrt{2}\sqrt{\frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)}}\wedge 1\right]. \quad (11)$$

Step 3: summing the inequalities to get an upper bound on $\overline{n}^t_h(s_h,a_h)$ Summing for $t\in\{0,\dots,\tau-1\}$ the inequalities given by (10) yields

$$\sum_{t=0}^{\tau-1} q^t_{h,h} \ge \frac{\tilde{\Delta}_h(s_h,a_h)}{2}\left(\sum_{t=0}^{\tau-1} p_h^{t+1}(s_h,a_h)\right) = \frac{\tilde{\Delta}_h(s_h,a_h)}{2}\overline{n}^\tau_h(s_h,a_h).$$

Summing the upper bounds in (11) yields that $\tilde{\Delta}_h(s_h,a_h)n^\tau_h(s_h,a_h)$ is upper bounded by

$$B^\tau_h(s_h,a_h) := 2\sum_{t=0}^{\tau-1}\sum_{\ell=h}^{H}\gamma^{\ell-h}\sigma_{H-\ell+1}\sum_{(s,a)} p_h^{t+1}(s_h,a_h)p_{h,\ell}^{t+1}(s,a|s_h,a_h)\left[4\sqrt{2}\sqrt{\frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)}}\wedge 1\right].$$

The rest of the proof consists in upper bounding $B^\tau_h(s_h,a_h)$ in terms of the pseudo counts $\overline{n}^\tau_h(s_h,a_h)$.

Step 4: from counts to pseudo-counts For all $\ell\ge h$, we introduce the set $\mathcal{S}_\ell(s_h,a_h)$ of states-action pairs (s,a) that can be reached at step ℓ from (s,a).

For each $(s,a)\in\mathcal{S}_\ell(s_h,a_h)$, we define

$$C_\ell(s,a;s_h,a_h) = \sum_{t=0}^{\tau-1} p_h^{t+1}(s_h,a_h)p_{h,\ell}^{t+1}(s,a|s_h,a_h)\left[4\sqrt{2}\sqrt{\frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)}}\wedge 1\right].$$

One can observe that $B^\tau_h(s_h,a_h) = 2\sum_{\ell=h}^{H}\sum_{(s,a)\in\mathcal{S}_\ell(s_h,a_h)}\gamma^{\ell-h}\sigma_{H-\ell+1}C_\ell(s,a;s_h,a_h)$. To upper bound $C_\ell(s,a;s_h,a_h)$ we further introduce the *conditional pseudo-counts*

$$\bar{n}^t_{h,\ell}(s,a;s_h,a_h) := \sum_{i=1}^{t} p_h^i(s_h,a_h)p^i_{h,\ell}(s,a|s_h,a_h),$$

for which one can write

$$C_\ell(s,a;s_h,a_h) = \sum_{t=0}^{\tau-1}[\bar{n}^{t+1}_{h,\ell}(s,a;s_h,a_h) - \bar{n}^t_{h,\ell}(s,a;s_h,a_h)]\left[4\sqrt{2}\sqrt{\frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)}}\wedge 1\right].$$

Using Lemma 9 to relate the counts to the conditional pseudo-counts, one can write

$$\begin{aligned}
C_\ell(s,a;s_h,a_h) &\le 8\sqrt{2}\sum_{t=0}^{\tau-1}[\bar{n}^{t+1}_{h,\ell}(s,a;s_h,a_h) - \bar{n}^t_{h,\ell}(s,a;s_h,a_h)]\sqrt{\frac{\beta(\overline{n}^t_{h,\ell}(s,a;s_h,a_h),\delta)}{\overline{n}^t_{h,\ell}(s,a;s_h,a_h)\vee 1}}\\
&\le 8\sqrt{2}\sqrt{\beta(\overline{n}^\tau_{h,\ell}(s,a;s_h,a_h),\delta)}\sum_{t=0}^{\tau-1}\frac{\bar{n}^{t+1}_{h,\ell}(s,a;s_h,a_h) - \bar{n}^t_{h,\ell}(s,a;s_h,a_h)}{\sqrt{\overline{n}^t_{h,\ell}(s,a;s_h,a_h)\vee 1}}\\
&\le 8\sqrt{2}(1+\sqrt{2})\sqrt{\beta(\overline{n}^\tau_{h,\ell}(s,a;s_h,a_h),\delta)\times\bar{n}^\tau_{h,\ell}(s,a;s_h,a_h)},
\end{aligned}$$

where the last step uses Lemma 19 in [17].

Finally, by summing over episodes ℓ and over reachable states $(s,a) \in \mathcal{S}_\ell(s_h,a_h)$, we can upper bound $B_h^\tau(s_h,a_h)$ by

$$
\begin{aligned}
&2\sum_{\ell=h}^{H}\gamma^{\ell-h}\sigma_{H-\ell+1}\left[8\sqrt{2}(1+\sqrt{2})\sqrt{\beta(\overline{n}_h^\tau(s_h,a_h),\delta)}\sum_{(s,a)\in\mathcal{S}_\ell(s_h,a_h)}\sqrt{\bar{n}_{h,\ell}^\tau(s,a;s_h,a_h)}\right]\\
&\le 2\sum_{\ell=h}^{H}\gamma^{\ell-h}\sigma_{H-\ell+1}\left[8\sqrt{2}(1+\sqrt{2})\sqrt{\beta(\overline{n}_h^\tau(s_h,a_h),\delta)}\sqrt{(BK)^{h-\ell}}\sqrt{\sum_{(s,a)\in\mathcal{S}_\ell(s_h,a_h)}\bar{n}_{h,\ell}^\tau(s,a;s_h,a_h)}\right]\\
&= 2\sum_{\ell=h}^{H}\gamma^{\ell-h}\sigma_{H-\ell+1}\left[8\sqrt{2}(1+\sqrt{2})\sqrt{\beta(\overline{n}_h^\tau(s_h,a_h),\delta)}\sqrt{(BK)^{h-\ell}}\sqrt{\bar{n}_h^\tau(s_h,a_h)}\right],
\end{aligned}
$$

where we have used that $\sum_{(s,a)\in\mathcal{S}_\ell(s_h,a_h)}\bar{n}_{h,\ell}^\tau(s,a;s_h,a_h) = \bar{n}_h^\tau(s_h,a_h)$. By using further Lemma 10 to upper bound all the constants, we obtain

$$
B_h^\tau(s_h,a_h) \le 64\sqrt{2}(1+\sqrt{2})\left(\sqrt{BK}\right)^{H-h}\sqrt{\bar{n}_h^\tau(s_h,a_h)\beta(\overline{n}_h^\tau(s_h,a_h),\delta)}\,.
$$

Lemma 10. *For every $x>1$, $\sum_{\ell=h}^{H}(\gamma x)^{\ell-h}\sigma_{H-\ell+1} \le \frac{x^{H-h}}{\left(1-\frac{1}{x}\right)^2}$.*

Proof. Since $\gamma \le 1$ and $x>1$, we can write

$$
\begin{aligned}
\sum_{\ell=h}^{H}(\gamma x)^{\ell-h}\sigma_{H-\ell+1} &\le \sum_{\ell=h}^{H}x^{\ell-h}(H-\ell+1) = \sum_{\ell=0}^{H-h}x^{\ell}(H-h-\ell+1)\\
&= x^{H-h}\sum_{\ell=0}^{H-h}\frac{H-h-\ell+1}{x^{H-h-\ell}} = x^{H-h}\sum_{\ell=0}^{H-h}(\ell+1)r^\ell,
\end{aligned}
$$

where $r=1/x<1$. The latter is an *arithmetico-geometric sum* that can be upper bounded as

$$
\sum_{\ell=0}^{H-h}(\ell+1)r^\ell \le \sum_{\ell=0}^{\infty}(\ell+1)r^\ell = \frac{1}{(1-r)^2} = \frac{1}{\left(1-\frac{1}{x}\right)^2}\,.
$$

□

E Proof of Theorem 2

The proof of Theorem 2 uses the same ingredients as the proof of Theorem 1: Lemma 5 which relates the gaps to the diameters of the confidence intervals $D_h^t(s_h,a_h) = U_h^t(s_h,a_h) - L_h^t(s_h,a_h)$ and a counterpart of Lemma 8 for the deterministic case, stated below.

Lemma 11. *If $\mathcal{E}$ holds, and $t_{1:H} = (s_1,a_1,\dots,s_H,a_H)$ is the $(t+1)$-st trajectory generated by* `MDP-GapE`*, for all $h\in[H]$,*

$$
D_h^t(s_h,a_h) \le \left[\sqrt{\frac{2\beta(n_h^t(s_h,a_h),\delta)}{n_h^t(s_h,a_h)}} \wedge 1\right] + \gamma D_{h+1}^t(s_{h+1},a_{h+1}).
$$

It follows from Lemma 11 that for all $h\in[H]$, along the $(t+1)$-st trajectory $t_{1:H} = (s_1,a_1,\dots,s_H,a_H)$,

$$
D_h^t(s_h,a_h) \le \sum_{\ell=h}^{H}\gamma^{\ell-h}\left[\sqrt{\frac{2\beta(n_\ell^t(s_\ell,a_\ell),\delta)}{n_\ell^t(s_\ell,a_\ell)}} \wedge 1\right].
$$

Letting $n^t(t_{1:H})$ be the number of times the trajectory $t_{1:H}$ has been selected by `MDP-GapE` in the first t episodes, one has $n_\ell^t(s_\ell,a_\ell) \ge n^t(t_{1:H})$. Hence, if $n^t(t_{1:H})>0$, it holds that

$$
D_h^t(s_h,a_h) \le \sum_{\ell=h}^{H}\gamma^{\ell-h}\sqrt{\frac{2\beta(n^t(t_{1:H}),\delta)}{n^t(t_{1:H})}} = \sigma_{H-h+1}\sqrt{\frac{2\beta(n^t(t_{1:H}),\delta)}{n^t(t_{1:H})}}.
$$

Using Lemma 5, if $t < \tau$, if $t_{1:H}$ is the trajectory selected at time $(t+1)$, either $n^t(t_{1:H}) = 0$ or

$$\forall h \in [H], \;\; \tilde{\Delta}_h(s_h, a_h) \leq \sigma_{H-h+1}\sqrt{\frac{2\beta(n^t(t_{1:H}), \delta)}{n^t(t_{1:H})}}$$

It follows that for any trajectory $t_{1:H}$,

$$n^\tau(t_{1:H})\left[\max_{h\in[H]} \frac{\left(\tilde{\Delta}_h(s_h, a_h)\right)^2}{(\sigma_{H-h+1})^2}\right] \leq 2\beta(n^\tau(t_{1:H}), \delta).$$

The conclusion follows from Lemma 12 and from the fact that $\tau = \sum_{t_{1:H}\in\mathcal{T}} n^\tau(t_{1:H})$.

F Sample complexity of Sparse Sampling in the Fixed-Confidence Setting

In this section, we prove Lemma 1.

For simplicity, and without loss of generality, assume that the reward function is known. Let $C > 0$. Sparse Sampling builds, recursively, the estimates $\widehat{V}_h$ and $\widehat{Q}_h$ for $h \in [H+1]$, starting from $\widehat{V}_{H+1}(s) = 0$ and $\widehat{Q}_{H+1}(s,a) = 0$ for all (s,a). Then, from a target state-action pair (s,a), it samples C transitions $Z_i \sim p_h(\cdot|s,a)$ for $i \in [C]$ and computes:

$$\widehat{Q}_h(s,a) = r_h(s,a) + \frac{1}{C}\sum_{i=1}^{C} \widehat{V}_{h+1}(Z_i), \;\; \text{with } \widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$$

For an initial state s, its output is $\widehat{Q}_1(s,a)$ for all $a \in [K]$. For any state s, consider the events

$$\mathcal{G}(s,a,h) = \left\{\left|\widehat{Q}_h(s,a) - Q^\star_h(s,a)\right| \leq \varepsilon_h\right\} \bigcap \left\{\bigcap_{z\in \mathrm{supp}[p_h(\cdot|s,a)]} \mathcal{G}(z, h+1)\right\}.$$

and

$$\mathcal{G}(s,h) = \bigcap_{a\in[K]} \mathcal{G}(s,a,h).$$

defined for $h \in [H+1]$, where $\varepsilon_h := (H-h+1)H\sqrt{(2/C)\log(2/\delta')}$ for some $\delta' > 0$.

Let

$$\delta_h = \frac{2K\delta'}{BK-1}\left((BK)^{H-h+1} - 1\right)$$

We prove that, for all s and all h, $\mathbb{P}\left[\mathcal{G}(s,h)\right] \geq 1 - \delta_h$. We proceed by induction on h. For $h = H+1$, we have $\widehat{Q}_{H+1}(s,a) = Q^\star_{H+1}(s,a) = 0$ for all (s,a) by definition, which gives us $\mathbb{P}\left[\mathcal{G}(s,a,H+1)\right] = 1$ and, consequently, $\mathbb{P}\left[\mathcal{G}(s,H+1)\right] = 1$.

Now, assume that $\mathbb{P}\left[\mathcal{G}(z,h+1)\right] \geq 1 - \delta_h$ for all z. Since

$$\left|\widehat{Q}_h(s,a) - Q^\star_h(s,a)\right| \leq \frac{1}{C}\left|\sum_{i=1}^{C}\left(\widehat{V}_{h+1}(Z_i) - V^\star_{h+1}(Z_i)\right)\right| + \frac{1}{C}\left|\sum_{i=1}^{C}\left(V^\star_{h+1}(Z_i) - \mathbb{E}\left[V^\star_{h+1}(Z_i)\right]\right)\right|$$

We have,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{G}(s,a,h)^\complement\right] &\leq \sum_{z\in\mathrm{supp}[p_h(\cdot|s,a)]} \mathbb{P}\left[\mathcal{G}(z,h+1)^\complement\right] + \mathbb{P}\left[\frac{1}{C}\left|\sum_{i=1}^{C}\left(V^\star_{h+1}(Z_i) - \mathbb{E}\left[V^\star_{h+1}(Z_i)\right]\right)\right| \geq \varepsilon_h - \varepsilon_{h+1}\right] \\
&\leq B\delta_{h+1} + 2\exp\left(-\frac{C(\varepsilon_h - \varepsilon_{h+1})^2}{2H^2}\right) \leq B\delta_{h+1} + 2\delta'
\end{aligned}$$

and, consequently,

$$\mathbb{P}\left[\mathcal{G}(s,h)^{\complement}\right] \leq BK\delta_{h+1} + 2K\delta' = \delta_h.$$

which gives us $\mathbb{P}\left[\mathcal{G}(s,h)\right] \geq 1 - \delta_h$, as claimed above. In particular, taking $h = 1$, we have

$$\left|\widehat{Q}_1(s,a) - Q_1^\star(s,a)\right| \leq H^2\sqrt{(2/C)\log(2/\delta')}$$

with probability at least $1-\delta$, where $\delta = 2K\delta'\left((BK)^H - 1\right)/(BK-1)$. Finally, we let $\varepsilon := H^2\sqrt{(2/C)\log(2/\delta')}/2$ and solve for C, obtaining

$$C = \mathcal{O}\left(\frac{H^5}{\varepsilon^2}\log\left(\frac{BK}{\delta}\right)\right).$$

Thus predicting $\hat{a} = \operatorname*{argmax}_a \widehat{Q}_1(s_1,a)$ after $\mathcal{O}\left(C(BK)^H\right)$ sampled transitions we have

$$\mathbb{P}\left(Q^\star(s_1,\hat{a}_\tau) > Q^\star(s_1,a^\star) - \varepsilon\right) \geq 1-\delta\,.$$

G A Technical Lemma

We state and prove below a technical result that permits to obtain an upper bound on n from a condition of the form $n\Delta^2 \leq \beta(n,\delta)$, like the one which appears in Theorem 1.

Lemma 12. *Let $n \geq 1$ and $a,b,c,d > 0$. If $n\Delta^2 \leq a + b\log(c+dn)$ then*

$$n \leq \frac{1}{\Delta^2}\left[a + b\log\left(c + \frac{d}{\Delta^4}(a + b(\sqrt{c}+\sqrt{d}))^2\right)\right].$$

Proof. Since $\log(x) \leq \sqrt{x}$ and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y > 0$, we have

$$\begin{aligned}
&n\Delta^2 \leq a + b\sqrt{c+dn} \leq a + b\sqrt{c} + b\sqrt{d}\sqrt{n} \\
&\implies \sqrt{n}\Delta^2 \leq \frac{a+b\sqrt{c}}{\sqrt{n}} + b\sqrt{d} \leq a + b(\sqrt{c}+\sqrt{d}) \\
&\implies n \leq \frac{1}{\Delta^4}\left(a + b(\sqrt{c}+\sqrt{d})\right)^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&n\Delta^2 \leq a + b\log(c+dn) \\
&\implies n\Delta^2 \leq a + b\log(c+dn) \quad \text{and} \quad n \leq \frac{1}{\Delta^4}\left(a + b(\sqrt{c}+\sqrt{d})\right)^2 \\
&\implies n\Delta^2 \leq a + b\log\left(c + \frac{d}{\Delta^4}\left(a + b(\sqrt{c}+\sqrt{d})\right)^2\right).
\end{aligned}$$

□