Identification of Blackwell Optimal Policies for Deterministic MDPs

Victor Boone **Bruno Gaujal**
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

Abstract

This paper investigates a new learning problem, the identification of Blackwell optimal policies on deterministic MDPs (DMDPs): A learner has to return a Blackwell optimal policy with fixed confidence using a minimal number of queries. First, we characterize the maximal set of DMDPs for which the identification is possible. Then, we focus on the analysis of algorithms based on product-form confidence regions. We minimize the number of queries by efficiently visiting the state-action pairs with respect to the shape of confidence sets. Furthermore, these confidence sets are themselves optimized to achieve better performance. The performance of our method compares to the lower bound up to a factor n^2 in the worst case – where n is the number of states, and constant in certain classes of DMDPs.

1 INTRODUCTION

Blackwell optimality Blackwell (1962) is arguably the most refined optimality criterion that exists on Markov Decision Processes (MDPs). Indeed, Blackwell optimal policies optimize the gain, the bias, higher order biases, as well as β-discounted scores for high enough discount factors β, see Puterman (1994). Despite its properties, Blackwell optimality drew very little attention from reinforcement learning communities. There are actually several hurdles that make learning Blackwell optimal policies difficult. First, the computation of Blackwell optimal policies is already tedious, even when the MDP is fully known Puterman (1994), and its complexity is still open.

Also, there is no concept of "near" Blackwell optimality because of the hierarchy involved in its definition so that one cannot gradually approach Blackwell optimal policies. This makes learning even harder.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

To our knowledge, this paper is the first to investigate the Blackwell optimal policy identification problem within the PAC-RL framework Fiechter (1994); Strehl (2007). We focus on the *generative model* setting Kearns and Singh (1998), where the learner is allowed to sample state-action pairs without restriction and halts according to a stopping time. This random time is such that the algorithm is able to produce an optimal policy with an *a priori* fixed confidence. This work focuses on MDPs with *deterministic* transitions where unlike general MDPs, the transition structure is easily determined by a learner. Learning such MDPs is analogue to *learning on graphs* with unknown probabilistic arc-weights; albeit learning Blackwell optimality is nothing alike a routing problem. Together with the generative model assumption, the problem resembles best arm identification in the multi-armed bandit case with exponentially many correlated arms.

Our contributions. We show that learning a Blackwell optimal policy is possible if and only if the maximal mean weight cycle is unique (H1) and the bias optimal policy is unique (H2). Under these assumptions, Blackwell optimality collapses to bias optimality, which means that optimalities of higher orders cannot be identified on MDPs when they strictly refine bias optimality. We provide a learning algorithm scheme based on generalized Bellman coefficients, together with explicit bounds on its asymptotic expected number of transition samples. We further show that when rewards are Gaussian, then up to multiplicative constants, the performances of our methods compare to lower bounds of possible performances.

Related Work. Regret minimization is a popular way to address learning on MDPs. For example, UCRL2 Auer et al. (2009) is an unavoidable milestone in that direction, with its contemporary UCYCLE Ortner (2010) analogue which is specific to deterministic MDPs. For a more recent review of the state-of-the-art, refer to Wei et al. (2020). In spite of this abundant literature, the design of a no-regret learning setting for Blackwell optimality is still missing and arguably conceptually difficult. Instead, another way to learn is to measure how likely it is that a learner may identify an optimal policy after a given number of samples, leading to policy search problems, e.g. instant regret

Bubeck et al. (2009) and probably approximately correct Fiechter (1994) settings.

For the special case of multi-armed bandits, Bubeck et al. (2009) argues that efficient policy search and regret minimization are orthogonal learning tasks.

As far as policy identification with generative model is concerned, the minimax approach is the most common, with a major focus on the discounted settings, see Azar et al. (2011); Gheshlaghi Azar et al. (2013); Li et al. (2021) and references therein. The finite horizon case is dealt in Dann and Brunskill (2015); Wang et al. (2020); Domingues et al. (2021); Li et al. (2022). Worth mentioning is the work of Tarbouriech et al. (2021) dedicated to the stochastic shortest path problem (SSP) on a type of MDP with zero-reward absorbing state, where optimal policies are *de facto* bias optimal ones. Although their work may be considered as a first step toward the learning of higher order optimalities, it doesn't cover the general class of (deterministic) MDPs – and is not motivated as such. As far as the minimax setting is concerned, the performances of the learner are bounded by her performances for the *worst* possible MDP; by *performances*, understand the number of samples the learner asks to guess an optimal policy. For pure identification without ϵ-tolerance on the suboptimality of policies, there exist MDPs for which the lower bounds of achievable runtimes can be made arbitrarily big. In this case, the minimax approach is irrelevant. This is especially the case for Blackwell optimality, as, to our knowledge, there is no notion of ϵ-Blackwell optimality. In the present paper, the expected sample complexity is compared to *instance specific* lower bounds of the possible performances in the line of the works of Kaufmann et al. (2016) and Garivier and Kaufmann (2016) who provide information-theoretical lower bounds of the expected sample complexity in the best arm identification problem for multi-armed bandits. Recently, this approach has been extended to discounted MDPs with generative model by Marjani and Proutiere (2021) and with navigation constraints by Marjani et al. (2021). Our methods were inspired by this line of work Kaufmann et al. (2016); Marjani and Proutiere (2021); Marjani et al. (2021), especially for the derivation of lower bounds and the exploration methods (D-tracking rule).

2 BLACKWELL OPTIMALITY AND DETERMINISTIC MDPS

Throughout the paper vectors are column vectors by default. We use the infinity-norm on matrices $\|(A_{ij})_{ij}\| = \max_i \sum_j |A_{ij}|$. Accordingly, we use the infinity-norm on column vectors and the one-norm on row vectors. For $p > 0$ and (a_i) a non-negative vector, we denote $\|(a_i)\|^p := (\sum_i a_i^p)^{1/p}$. $\mathrm{KL}(-\|-)$ denotes the Kullback-Leibler divergence and $\mathrm{kl}(u, v)$ is the divergence between Bernoulli distributions of parameters u and v. The simplex of $\mathbb{R}^m$ is denoted Δ^m.

Let us consider a finite Markov decision process (MDP) M with n states and A actions $M = (\mathcal{S}, \mathcal{A}, P, q)$, where $\mathcal{S}$ is the state space (of size n), $\mathcal{A} := \coprod_{x\in\mathcal{S}} \mathcal{A}_x$ is the action space, P is the transition kernel and q the reward distribution (each time action a is taken in state x, it gets a random real reward with distribution $q(x, a)$). Reward distributions are assumed to be sub-Gaussian[1] with means $r(x, a)$ and standard deviation σ. A *policy* is a stationary deterministic decision rule $\pi : \mathcal{S} \to \mathcal{A}$ that, for each state $x \in \mathcal{S}$, selects a legal action $a \in \mathcal{A}_x$. We write $\pi \in \Pi$. Also, a policy π defines *a*) a Markov chain on $\mathcal{S}$ with transition kernel P_π given by $P_\pi(x, y) := P(y|x, \pi(x))$; and *b*) a reward vector r_π of coordinates $r_\pi(x) := r(x, \pi(x))$.

2.1 Blackwell optimality

The performance of a policy can be measured in various ways. Some performance measures only depend on the transient behavior of the MDP, such as the sum of the rewards over a finite horizon, or the discounted infinite sum of the rewards. For instance, the β-discounted value of a given policy π, starting from state x_0, is

$$v_\beta^\pi(x_0) := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \beta^t r(X_t, \pi(X_t)) \right] \tag{1}$$

where X_t denotes the visited (random) state at time t under the iterations of π. Other performance measures only depend the stationary behavior of the MDP, such as the long run average gain. The long run average gain in state x_0 of a policy π is

$$g_\pi(x_0) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} r(X_t, \pi(X_t)). \tag{2}$$

As for Blackwell optimality, it depends on both the transient and the stationary behaviors of the MDP and therefore may supersede both discounted optimality and average gain optimality. Blackwell optimality was historically defined using that transient approach Blackwell (1962):

Definition 1. A policy π^* is *Blackwell optimal* if it satisfies $v_\beta^{\pi^*} \geq v_\beta^\pi$ for all policies π and all $1 > \beta > \beta^\infty$, for some threshold discount factor $\beta^\infty < 1$. The set of Blackwell optimal policies will be denoted $\Pi_\infty^*(M)$.

Another possible definition of Blackwell optimal policies uses the average gain and biases. The set of gain-optimal policies (maximizing (2)) is denoted $\Pi_{-1}^*(M)$. Now, two policies with the same gain may be distinguished by a second order quantity, namely their 0-bias (or simply bias). If

[1] A random variable R is sub-Gaussian with mean $r \in \mathbb{R}$ and standard deviation $\sigma > 0$ if $\mathbb{P}\{|R - r| > t\} \leq \exp\left(\frac{-t^2}{2\sigma^2}\right)$.

P_π^* is the Cesàro limit of the sequence $(P_\pi^n)_{n\in\mathbb{N}}$, then the bias is defined as

$$h_\pi := \sum_{t=0}^{\infty}[P_\pi^t - P_\pi^*]r_\pi. \tag{3}$$

A policy π is *bias*-optimal if it achieves maximal bias in addition to maximal gain, i.e. for all $\pi' \in \Pi$ and $x \in \mathcal{S}$, either $g_{\pi'}(x) \le g_\pi(x)$ or $h_{\pi'}(x) \le h_\pi(x)$ with $g_\pi(x) = g_{\pi'}(x)$. Their set is denoted $\Pi_0^*(M)$.

This refinement can be pushed further. The k-bias of π, for $n \ge 1$ is

$$h_\pi^{(k)} := -\sum_{t=0}^{\infty}[P_\pi^t - P_\pi^*]h_\pi^{(k-1)}, \tag{4}$$

where $h_\pi^{(0)}$ stands for h_π. A policy that optimizes the gain and all biases up to the k-bias is said k-bias optimal. Their set is denoted $\Pi_k^*(M)$.

It has been shown (see Puterman (1994), for example) that this refinement stops when $k = S$ (*i.e.* $\Pi_k^*(M) = \Pi_S^*(M)$ for all $k \ge S$) and this set coincides with $\Pi_\infty^*(M)$ in Definition 1.

2.2 Deterministic MDPs

Definition 2. A MDP M is *deterministic* if its transition kernels $P(\cdot|x,a)$ are degenerate; Specifically, if $\forall x, y \in \mathcal{S}$, $\forall a \in \mathcal{A}_x$, $P(y|x,a) \in \{0,1\}$. From a given state x, the choice of an action $a \in \mathcal{A}_x$ uniquely determines a successor $s(x,a)$ such that $P(s(x,a)|x,a) = 1$.

This way, a DMDP is naturally endowed with a multi-directed graph structure whose multi-arc spaces is identifiable by $\mathcal{Z}$ the set of state-action pairs (x,a). In the sequel, m denotes the size of $\mathcal{Z}$. The elements of a family $\mathcal{M}$ of DMDPs that share the same transition structure P can be identified by their mean reward vector $(r(z) : z \in \mathcal{Z})$ that belongs to $\mathbb{R}^m$. $\mathcal{M}$ hence inherits the topology of $\mathbb{R}^m$, making it a topological space.

Actually, many important examples of MDPs are deterministic, for example games such as Chess, Go, grid games and more, see Shah et al. (2020) for more. Already mentioned in the introduction is also the example of routing problems. Trying to learn, on a network with random delays on transitions, how to route a packet to a fixed destination, is a subcase of learning Blackwell optimality (yet on such problems, only the transient costs matter). Overall, the properties of Blackwell optimality in the deterministic setting are strikingly graph friendly, as explained below.

Gain and bias on DMDPs. When transitions are deterministic, the iterates $(X_t : t \ge 0)$ of π starting from $x \in \mathcal{S}$ are no more random. To insist on their deterministic nature, these are lowercased as x_t. Those iterates eventually converge in finite time to a terminal cycle $\mathcal{C}_x^\pi \subseteq \mathcal{S}$ and the gain is the average reward on that cycle. Visually,

$$g_\pi(x) = g(\mathcal{C}_x^\pi) := \frac{1}{|\mathcal{C}_x^\pi|}\sum_{u\in\mathcal{C}_x^\pi} r(u,\pi(u)). \tag{5}$$

Also, the bias $h_\pi(x)$ takes a particular form in DMDPs:

Proposition 1. *Let $\pi : \mathcal{S} \to \mathcal{A}$ a policy and $x \in \mathcal{S}$. Denote x_t the state at time t under the iterations of π (from $x = x_0$), and let $z_t := (x_t, \pi(x_t))$. If $T \ge 0$ is such that $x_T \in \mathcal{C}_x^\pi$, then the bias expands as*

$$h_\pi(x) = \sum_{t=0}^{T-1}[r(z_t) - g(\mathcal{C}_x^\pi)] \qquad (transient)$$
$$+ \frac{1}{|\mathcal{C}_x^\pi|}\sum_{\ell=1}^{|\mathcal{C}|}\sum_{t=0}^{\ell-1}[r(z_{T+t}) - g(\mathcal{C}_x^\pi)]. \quad (recurrent)$$

This formula indicates that the bias is, roughly speaking, the normalized weight of the path to the terminal cycle. It will be useful to assess the performance of the identification algorithm presented later.

3 POLICY IDENTIFICATION IN DMDPS

The very purpose of this paper is the identification of Blackwell optimal policies in the generative model, in a similar fashion to best arm identification for stochastic bandits in Kaufmann et al. (2016). By *generative model*, we mean that at each time step, the algorithm is allowed to sample any edge. An *identification algorithm* $\mathcal{I}$ consists in three components :

- an *allocation rule* that chooses the next sampled state-action pair $Z_t := (X_t, A_t)$, then observes the transition $R_t \sim q(X_t, A_t)$, $Y_t \sim P(\cdot|X_t, A_t)$ and is measurable with respect to the usual filtration $\mathcal{F}_{t-1}$ of the history $(X_1, A_1, R_1, Y_1, \dots, X_{t-1}, A_{t-1}, R_{t-1}, Y_{t-1})$;
- a *stopping rule* τ_δ that must be a stopping time with respect to $(\mathcal{F}_t \mid t \ge 0)$, where $\delta > 0$ is a confidence parameter;
- a *recommendation rule* to return a policy $\pi_{\tau_\delta}^{\mathcal{I}}$ when the algorithm stops, which is $\mathcal{F}_{\tau_\delta}$-measurable.[2]

Definition 3. An identification algorithm $\mathcal{I}$ (see above) is said δ-probably correct (δ-PC) on M if

$$\mathbb{P}^{M,\mathcal{I}}\left\{\tau_\delta < \infty \text{ and } \pi_{\tau_\delta}^{\mathcal{I}} \in \Pi_\infty^*(M)\right\} \ge 1 - \delta.$$

Its *expected sample complexity* is $\mathbb{E}^{M,\mathcal{I}}[\tau_\delta]$.

[2] *i.e.*, for all $T \ge 1$ and $\pi \in \Pi$, $\{\tau_\delta = n\} \cap \{\pi_{\tau_\delta} = \pi\} \in \mathcal{F}_T$

Visits and empirical model. Let $t \geq 1$ a time instant. The visit count of a state-action pair $z \in \mathcal{Z}$ is denoted $N_t(z) := \sum_{i=1}^{t-1} \mathbf{1}_{Z_i=z}$ and by $\mathbf{N}_t$ is meant their vector. The *DMDP of empirical observations* up to time t is the DMDP of maximum likelihood, i.e., the DMDP $\hat{M}_t$ with reward distributions $\hat{q}_t(z) := N_t(z)^{-1} \sum_{i=1}^{t-1} \mathbf{1}_{Z_i=z} \text{Dirac}(R_i)$. The associated mean reward vector $\hat{\mathbf{r}}_t$ is called the *empirical mean reward vector*.

3.1 A mandatory set of assumptions

We make the following assumptions.

Assumptions. First, it is reasonable to assume that the transition structure P is fixed and known to the learner because can be learned in $O(m)$. The underlying multigraph is refered to as $\mathcal{G}$. We assume that $\mathcal{G}$ has a unique final strongly connected component. This corresponds to the usual weakly communicating assumption made on MDPs. The two following hypothesis replace the uniqueness of the optimal policy used in Marjani and Proutiere (2021) for the discounted case.

H1 *M has a unique optimal cycle: the cycle $\mathcal{C}_*$ maximizing the average expected reward $\frac{1}{|\mathcal{C}_*|} \sum_{(x,a)\in\mathcal{C}_*} r(x,a)$ is unique.*

H2 *M has a unique bias-optimal policy π^*_M.*

These assumptions are independent: there exists DMDPs where H2 is satisfied but H1 is not (and *vice-versa*), see Figure 1.

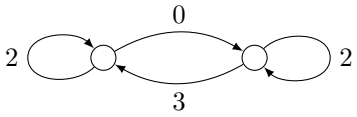

Figure 1: A DMDP where H2 is satisfied but H1 is not. The identification of Blackwell optimal policy is impossible by Theorem 1.

Optimal policy, gain and bias. The uniqueness of the bias-optimal policy implies that Blackwell optimality collapses to bias optimality, that is $\Pi^*_\infty(M) = \Pi^*_0(M)$, and we denote by π^*_M (or π^*) the unique element of $\Pi^*_0(M)$. Also, the optimal gain and bias are respectively $g^* = g^{\pi^*_M}$ and $h^* = h^{\pi^*_M}$. Outside H1 and H2, the identification of Blackwell optimal policies may be impossible. Here is why.

Let $\mathcal{H}$ denote the class of DMDPs with graph $\mathcal{G}$ that satisfies H1 and H2.

Outside the domain spawned by H1-H2, the PAC-RL setting isn't sound: no algorithm can correctly identify Blackwell optimal policies with arbitrarily high confidence *and* stop almost surely on every entry, apart from $\mathcal{H}$. The space $\mathcal{H}$ is thus the largest set on which Blackwell identification is possible. The precise statement is given below:

Theorem 1. *Let $\mathcal{I}$ be an identification algorithm and let $\delta \in (0, \frac{1}{4}n^{-n})$. Let $\mathcal{D}_\mathcal{I}$ be the set of DMDPs on which $\mathcal{I}$ is δ-probably correct. The two following assertions hold:*

(i) The interior of $\mathcal{D}_\mathcal{I}$ is in $\mathcal{H}$: $\mathring{\mathcal{D}}_\mathcal{I} \subseteq \mathcal{H}$;
(ii) If $\mathcal{D}_\mathcal{I}$ contains $\mathcal{H}$, $\mathcal{D}_\mathcal{I} = \mathcal{H}$.

Sketch of the proof. (*i*) Assume on the contrary that $\mathring{\mathcal{D}}_\mathcal{I} \cap \mathcal{H}^\complement \neq \varnothing$ and choose M a member of the intersection. By definition, there exists $\mathcal{B} \subseteq \mathcal{D}_\mathcal{I}$ open that contains M. For M, there exists a policy π that the algorithm returns with probability greater than n^{-n}. If $M \notin \mathcal{H}$, one can find arterial close copies of M such that $\pi \notin \Pi^*(M')$. Then, the algorithm can be fooled by choosing M' arbitrary close to M, so that when running on M', the algorithm returns π with probability approximately n^{-n}. In particular, we can choose $M' \in \mathcal{B}$; Yet $\mathcal{I}$ isn't δ-PAC on M'. A contradiction.

(*ii*) With a similar proof, one can show that if $\mathcal{D}_\mathcal{I}$ contains $\mathcal{H}$, then $\mathcal{D}_\mathcal{I} = \mathcal{H}$, whence the maximality of $\mathcal{H}$ regarding identification. □

Remark 1. Theorem 1 implies that there is no universal δ-PC algorithm, that is, no $\mathcal{I}$ such that $\mathcal{D}_\mathcal{I} = \mathcal{M}$.

Remark 2. The assertion (*i*) cannot be improved to $\mathcal{D}_\mathcal{I} \subseteq \mathcal{H}$ for all $\mathcal{I}$, because for every M and π any of its Blackwell optimal policy, the algorithm $\mathcal{I}(M)$ that stops at time 1 and returns π is correct on M.

We'd like to insist on the fact that under H1-H2, $\Pi^*_0(M)$ is a singleton, so that $\Pi^*_\infty(M) = \Pi^*_0(M)$. In particular, the identification of k-discounted optimalities Puterman (1994) of order $k \geq 1$ is impossible unless they collapse to bias optimality. We end up with the following principle:

> *Probably correct identification cannot go beyond bias optimality.*

As a final remark, while H1-H2 certainly restrict the class of DMDPs where Blackwell optimality can be identified, they are also generic in Baire categories sense.

Proposition 2. *$\mathcal{H}$ is generic in Baire categories sense. In particular, $\mathcal{M} \setminus \mathcal{H}$ has null Lebesgue measure.*

This says that a DMDP whose expected rewards are chosen randomly (continuously w.r.t the Lebesgue measure) will be in $\mathcal{H}$ with probability one.

3.2 Characterization of bias optimal policies on $\mathcal{H}$

Bias optimality is unique under H1 and H2. There is more to it; It is pretty easy to tell whether or not $\pi \in \Pi^*_0(M)$

when $M \in \mathcal{H}$. It is well-known, see Puterman (1994), that a policy is bias optimal if it satisfies three nested Bellman optimality equations. In the deterministic setting, these equations take the simpler form: for all $x \in \mathcal{S}$ and $a \in \mathcal{A}_x$:

$$\begin{aligned} P(x,a) \cdot g_\pi &\leq g_\pi(x) \\ r(x,a) - g_\pi(x) + P(x,a) \cdot h_\pi &\leq h_\pi(x) \\ -h_\pi(x) + P(x,a) \cdot h_\pi^{(1)} &\leq h_\pi^{(1)}(x), \end{aligned}$$

with equality when $a = \pi(x)$. Besides, a policy satisfying these equations achieves optimal gain and bias. The converse is false in general, so what about policies that only satisfy the two first conditions? In general, they may fail to achieve optimal bias. On $\mathcal{H}$, the story is different.

Proposition 3. *Let $M \in \mathcal{H}$ a DMDP. A policy $\pi \in \Pi$ is bias optimal if, and only if $g_\pi(x)$ does not depend on $x \in \mathcal{S}$ and*

$$\forall (x,a) \notin \pi, \quad r(x,a) - g_\pi(x) + P(x,a)h_\pi < h_\pi(x).$$

In particular, π_M^* is the unique policy with unique terminal cycle such that this equation holds. Moreover, a DMDP satisfies H1 and H2 if, and only if there is a policy satisfying this equation with unique terminal cycle.

This provides a characterization of suboptimal policies: if g_π is a constant vector, then $\pi \notin \Pi_0^*(M)$ iff it contains a sub-optimal transition:

$$\exists x \in \mathcal{S}, \quad r(x,\pi(x)) - g_*(x) + P_\pi(\cdot|x)h_* < h_*(x). \quad (6)$$

For $(x,a) \in \mathcal{Z}$, the quantity

$$\Delta(x,a) = h_*(x) - [r(x,a) - g_*(x) + P(x,a)h_*] > 0 \quad (7)$$

measures by how much (x,a) is suboptimal. These quantities, called *Bellman gaps*, do play an important part in the design of identification algorithms. Equation (6) is also of interest for effective computations of Blackwell optimal policies, providing a node-to-node criterion to determine whether or not a policy is bias optimal. Under H1, this leads to the computation of bias optimal policies in $O(nm)$ execution time, see Appendix E.1 for more details.

4 THE LSTS METHOD

In this section, we present a general method to design identification algoritms that are δ-probably correct on $\mathcal{H}$ for all $\delta > 0$. The performance of our method, if well-tuned, is comparable to lower bounds up to explicit multiplicative factor in the asymptotic regime.

4.1 D-tracking Scheme

The standard way to estimate the plausible parameters of the MDP is to construct individual confidence intervals for the estimates of the mean rewards' values, leading to confidence sets in product form, see for e.g. the literature on regret minimization Auer et al. (2009); Filippi et al. (2010); Fruit et al. (2018); Bourel et al. (2020) or PAC-RL on discounted MDPs Azar et al. (2011); Gheshlaghi Azar et al. (2013); Li et al. (2021).

From a frequentist viewpoint, that confidence set is centered at the model of maximum likelihood $\hat{M}_t = (\hat{\mathbf{r}}_t)$ where $\hat{\mathbf{r}}_t$ is the vector of empirical mean rewards. The confidence set is hence of the form

$$\widetilde{\mathcal{M}}_t^\delta = \hat{M}_t + \Lambda_t^\delta. \quad (8)$$

Here, Λ_t^δ is a product of confidence intervals. Specifically, $\Lambda_t^\delta := (-\epsilon_\delta(N_t(x,a),t), \epsilon_\delta(N_t(x,a),t))$ with

$$\epsilon_\delta(s,t) := \left(\frac{2\sigma^2 \log(\frac{4mt^3}{\delta})}{\max(1,s)} \right)^{1/2}. \quad (9)$$

The confidence value ϵ_δ is tuned in order to provide time-uniform confidence sets, see the result below.

Proposition 4. *For all $\mathcal{I}$, $\sum_t \mathbb{P}^{M,\mathcal{I}}\{M \notin \widetilde{\mathcal{M}}_t^\delta\} \leq \delta$.*

The set $\widetilde{\mathcal{M}}_t^\delta$ is refered to as the set of *plausible* MDPs.

It appears natural to stop the learning phase whenever all plausible MDPs share the same optimal policy. Regarding policy identification, boxes of special interest are those centered at M in the style of (8) such that every alternative MDP inside that box has the same optimal policy than M. This motivates the following definition.

Definition 4. A *box* is any element of the family generated by $\Gamma : \mathcal{H} \to \mathbb{R}_+^m \to 2^{\mathcal{H}}$, defined as follows:

$$\Gamma(M,\ell) := M + \prod\nolimits_{x,a} (-\ell(x,a), \ell(x,a)).$$

A box $\mathcal{B}$ is said Π_∞^*-*constant* if $\forall M, M' \in \mathcal{B}$, $\Pi_\infty^*(M) = \Pi_\infty^*(M')$. A family $\mathcal{B} = \{\mathcal{B}(M) : M \in \mathcal{H}\}$ is (1) Π_∞^*-constant if all its boxes are; and (2) is *continuous* if it is continuous with respect to the topology on boxes given by the Hausdorff distance between closed sets in $\mathbb{R}^m$.

Also, write $\ell_{\mathcal{B}} = \ell$ the half-width vector of $\mathcal{B} = \Gamma(M,\ell)$.

Let us consider an identification algorithm whose estimate at time t is $\hat{M}_t$. If $\mathcal{B}$ is a Π_∞^*-constant family, then whenever "$\widetilde{\mathcal{M}}_t^\delta \subseteq \mathcal{B}(\hat{M}_t)$" holds, the set of plausible MDPs (including M with probability $1-\delta$) have the same optimal policy than $\hat{M}_t$. This provides a natural stopping time. Of course, the algorithm should explore efficiently to guarantee that $\widetilde{\mathcal{M}}_t^\delta \subseteq \mathcal{B}(\hat{M}_t)$ holds for t as small as possible. The thing is that if $\hat{M}_t$ is guaranteed to converge to M almost surely, and if $\mathcal{B}$ is continuous, then $\mathcal{B}(\hat{M}_t)$ will resemble $\mathcal{B}(M)$ in the long run. In that scenario, $\mathcal{B}(\hat{M}_t)$ will stabilize eventually. This suggests that exploration can be done with respect to one target box. In the following algorithmic

scheme, the exploration is driven by a frequency parameter $\omega : \mathcal{H} \to \Delta^m$, so that $t \cdot \omega_t(z)$ is the target value of the number of visits $N_t(z)$. Some extra forced exploration is added to keep good convergence properties. It is fixed to the D-tracking rule of Marjani et al. (2021) for simplicity.

Algorithm 1 D-tracking scheme with boxes.

Require: confidence parameter $\delta \in (0,1)$, continuous exploration coefficients $\omega : \mathcal{H} \to \Delta^m$, never-empty continuous Π^*_∞-constant family $\mathcal{B} : \mathcal{H} \to \{\text{boxes}\}$.

Ensure: returns a policy $\pi^\delta \in \Pi^*_\infty(M)$ with probability $1-\delta$.

Sample independent $\mathbf{u}_t \sim \mathcal{U}([-1,1]^m)$ $(t \geq 1)$;
for $t = 1, 2, \dots,$ **do**
  if $\widetilde{\mathcal{M}}^\delta_t \subseteq \mathcal{B}(\hat{M}_t)$ **then**
    return $\pi^*(\hat{M}_t)$
  end if
  compute sampling rates $\omega_t \leftarrow \omega((1-\frac{1}{t})\hat{\mathbf{r}}_t + \frac{1}{t}\mathbf{u}_t)$;
  set $U_t \leftarrow \left\{(x,a) \in \mathcal{Z} \mid N_t(x,a) < \sqrt{t} - m/2\right\}$;
  the D-tracking rule Marjani et al. (2021) selects (X_t, A_t) among

$$\begin{cases} \operatorname{argmin}_{x,a}\left[N_t(x,a)\right] & \text{if } U_t \neq \varnothing, \\ \operatorname{argmin}_{x,a}\left[N_t(x,a) - t \cdot \omega_t(x,a)\right] & \text{if } U_t = \varnothing; \end{cases}$$

  observe reward $R_t \sim q(X_t, A_t)$;
end for

Proposition 5 (Correctedness)**.** *Every D-tracking scheme is δ-probably correct on DMDPs in $\mathcal{H}$.*

Sketch of the proof. We show in the appendix that following $\mathcal{B}(M) \neq \varnothing$, we have $\mathbb{P}^M\{\tau_\delta < \infty\} \geq 1-\delta$. Then, by Proposition 4, the probability that for all t, $\widetilde{\mathcal{M}}^\delta_t$ contains M, is at least $1-\delta$. Also, if $\widetilde{\mathcal{M}}^\delta_t \subseteq \mathcal{B}(\hat{M}_t)$, then for all $\widetilde{M} \in \widetilde{\mathcal{M}}^\delta_t$, we have $\Pi^*_\infty(\widetilde{M}) = \Pi^*_\infty(\hat{M}_t)$. All together,

$$\begin{aligned} \mathbb{P}\left\{\tau_\delta < \infty, \pi^*(\hat{M}_{\tau_\delta}) \notin \Pi^*_\infty(M)\right\} & \\ &\leq \mathbb{P}\left\{\tau_\delta < \infty, M \notin \mathcal{B}(\hat{M}_{\tau_\delta})\right\} \\ &\leq \sum\nolimits_{t\geq 1} \mathbb{P}\left\{\tau_\delta = t, M \notin \widetilde{\mathcal{M}}^\delta_t\right\} \\ &\leq \sum\nolimits_{t\geq 1} \mathbb{P}\left\{M \notin \widetilde{\mathcal{M}}^\delta_t\right\} \leq \delta. \end{aligned}$$ □

Proposition 6 (Performances)**.** *A D-tracking scheme stops on $M \in \mathcal{H}$ a.s.; Also,*

$$\liminf_{\delta\to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} \geq \sum_z \frac{2\sigma^2}{\ell_{\mathcal{B}(M)}(z)^2}.$$

Moreover, if exploration parameters are $\omega \propto \ell^{-2}_{\mathcal{B}(\cdot)}$, then

$$\lim_{\delta\to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} = \sum_z \frac{2\sigma^2}{\ell_{\mathcal{B}(M)}(z)^2}.$$

Sketch of the proof. Termination is guaranteed because the D-tracking rule ensures that every edge is visited $\Omega(\sqrt{t})$ times hence infinitely often. Therefore, $\hat{M}_t \to M$ a.s., so $\mathcal{B}(\hat{M}_t) \to \mathcal{B}(M)$ by continuity. Finally, we show that $\Omega(\sqrt{t})$ visits are enough to guarantee that the set of plausible MDPs gets within the box $\mathcal{B}(M)$ eventually, and regardless of the confidence parameter δ.

Ignoring logarithmic factors, we rewrite the stopping criterion as

$$\forall (x,a) \in \mathcal{Z}, \quad \sqrt{\frac{2\sigma^2 \log(4m/\delta)}{N_{\tau_\delta}(x,a)}} \leq \ell_{\mathcal{B}(\hat{M}_t)}(x,a).$$

For the performance bounds, by property of the D-tracking rule and continuity of $\omega_\cdot$, visit counts satisfy $N_t(x,a) = \omega_M(x,a)t + o(t)$ when t goes to infinity. Similarly, we approximate $\mathcal{B}(\hat{M}_t)$ with $\mathcal{B}(M)$ for large values of t. Accordingly, the previous inequality is asymptotically equivalent to:

$$\forall (x,a) \in \mathcal{Z}, \quad \sqrt{\frac{2\sigma^2 \log(4m/\delta)}{\tau_\delta \omega_M(x,a)}} \leq \ell_{\mathcal{B}(M)}(x,a).$$

Reorganizing terms provides a the lower bound on τ_δ:

$$\forall (x,a) \in \mathcal{Z}, \quad \frac{2\sigma^2 \log(4m/\delta)}{\omega_M(x,a)\ell_{\mathcal{B}(M)}(x,a)^2} \leq \tau_\delta.$$

Taking the limit when $\delta \to 0$, we obtain

$$\max_{(x,a)\in\mathcal{Z}} \frac{2\sigma^2}{\omega_M(x,a)\ell_{\mathcal{B}(M)}(x,a)^2} \leq \lim_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)}.$$

The left-hand term is a convex function of $\omega_M \in \Delta^m$ that can be minimized using KKT-conditions. We retrieve the lower bound of the main statement together with the optimal choice of parameters $\omega_M \propto \ell_{\mathcal{B}(M)}(\cdot)^{-2}$. □

Construction of good familys. Lemma 6 answers one fundamental question: how does a learner choose exploration coefficients $\omega_\cdot$ with respect to a continuous Π^*_∞-constant family? Namely, this solves the problem of (asymptotically) efficient exploration. How familys shall be chosen is an orthogonal task.

4.2 Lower Bounds

We deviate the from main story to deal with the computation of a general lower ground of a learner's possible performances. This will serve as a reference ground for our next algorithms to come. This bound generalizes exiting results and techniques from the multi-armed bandit settings (Kaufmann et al., 2016, Theorem 4). The transport of these information-theoretical results to MDPs is not new and already exist in the context of discounted optimality identification (Marjani and Proutiere, 2021, Proposition 1),

(Marjani et al., 2021, Lemma 1) and regret minimization on deterministic MDPs (Tranos and Proutiere, 2021, Proposition 3).

Proposition 7. *Let $M, M' \in \mathcal{H}$ two DMDPs with different bias-optimal policies such that $q(x,a)$ and $q'(x,a)$ are mutually absolutely continuous for all edges. Every δ-PC algorithm satisfies*

$$\sum\nolimits_{z\in\mathcal{Z}} \mathbb{E}^M[N_{\tau_\delta}(z)]\,\mathrm{KL}(q(z)\|q'(z)) \geq \mathrm{kl}(\delta, 1-\delta).$$

For the sake of self-containedness, we explain how to shift the proof of (Kaufmann et al., 2016, Theorem 4) to fit Proposition 7 in the appendix.

The next section address the issue of efficient stopping times. In the light of Proposition 6, this is about the design of continuous Π^*_∞-constant familys such that $\sum_{x,a} \ell_{\mathcal{B}(M)}(x,a)^{-2}$ is small. As the learning effort should not be uniform of transitions, the question is how the difficulty of a specific transition should be quantified.

4.3 A local suboptimality gap method

The local suboptimality gap method presented in the sequel attaches to each transition a difficulty related to how easily the optimal policy is subject to change under unilateral perturbations of the associated mean reward value. Because these quantities are defined by looking at edges in isolation, we refer to them as *local suboptimality gaps*.

Definition 5 (Local suboptimality gap)**.** Let M a DMDP and $\mathbf{r} \in \mathbb{R}^m$ its mean reward vector. The *local suboptimality gap* at $(x,a) \in \mathcal{Z}$ is given by

$$L_M(x,a) := \sup \left\{ \epsilon > 0 : \begin{array}{c} \Pi^*_\infty(\mathbf{r} + \epsilon \mathbf{e}_{xa}) = \Pi^*_\infty(\mathbf{r}) \\ \text{and} \\ \Pi^*_\infty(\mathbf{r} - \epsilon \mathbf{e}_{xa}) = \Pi^*_\infty(\mathbf{r}) \end{array} \right\}.$$

These local suboptimality gaps are continuous on $\mathcal{H}$. Roughly speaking, these are "generalized" Bellman coefficients. Indeed, $L_M(x,a) = \Delta_M(x,a)$ when (x,a) is not a transition of the optimal policy; Otherwise, $\Delta_M(x,a) = 0$ while $L_M(x,a) > 0$ can be expressed with respect to other $\Delta_M(y,b)$ and traveling times in M, see Lemma 11 in the Appendix. Also, the collection $(L_M(x,a) : (x,a) \in \mathcal{Z})$ can be computed in time $O(nm)$, see Lemma 14 in the Appendix.

Notice that for multi-armed bandits, this definition readily converts to mean reward differences between optimal and suboptimal arms. That last observation, together with the ideas from Kaufmann et al. (2016), leads to a simplification of the lower bound given by Lemma 7 that relies on local suboptimality gaps – thus a tractable lower bound. This is especially striking when rewards are Gaussian.

Proposition 8 (Edgewise Lower Bound)**.** *Let $M \in \mathcal{H}$ with Gaussian rewards of standard deviation $\sigma > 0$. For all δ-PC identification algorithm,*

$$\frac{\mathbb{E}^M[\tau_\delta]}{\mathrm{kl}(\delta, 1-\delta)} \geq \sigma^2 \sum_{z\in\mathcal{Z}} \frac{1}{L_M(z)^2}.$$

The lower bound is basically $\sigma^2 \|(L^{-1}_{x,a})\|_2^2$. By definition, local suboptimality gaps are about the *local* deviations of mean rewards that leave the optimal policy invariant. The next result provides a condition to shift these local deviations to *global* ones, hence explains how to build Π^*_∞-constant familys from local suboptimality gaps.

Proposition 9. *Let $\rho : \mathcal{H} \to \mathbb{R}^m_+$ a continuous function. The family $\mathcal{B}_\rho : \mathcal{H} \to \Gamma$ given by*

$$M + \prod\nolimits_{z\in\mathcal{Z}} (-\rho_M(z)L_M(z), \rho_M(z)L_M(z))$$

is a continuous family. If in addition, a) $\rho_M(z) > 0$ for all $z \in \mathcal{Z}$; and b) $\forall M$, $\forall \pi \in \Pi$, $\sum_{z\in\pi\cup\pi^*_M} \rho_M(z) \leq 1$, then $\mathcal{B}_\rho$ is Π^*_∞-constant and is never empty.

Namely, set $\mathcal{B}_\rho(M) = \Gamma(M, \ell)$ with $\ell(z) = \rho_M(z)L_M(z)$. In the continuity of Lemmas 5 and 6, we already know how to choose sampling parameters in order to accelerate the stopping time with respect to a family $\mathcal{B}_\rho$. We get the result below.

Theorem 2 (LSTS methods)**.** *Algorithms following the scheme 1 with familys of the form $\mathcal{B}_\rho$ and exploration parameters $\omega \propto (\rho L)^{-2}$, where ρ satisfies the conditions a) and b from Lemma 9, are called* LSTS algorithms *(Local Suboptimality based Track&Stop). These algorithms are δ-probably correct and have asymptotic performances:*

$$\lim_{\delta\to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} = 2\sigma^2 \sum_{z\in\mathcal{Z}} \frac{1}{\rho_M(z)^2 L_M(z)^2}.$$

for $M \in \mathcal{H}$ with sub-Gaussian rewards of standard deviations $\sigma > 0$.

What is left is to optimize ρ.

Example 1 (LSTS-cst)**.** Choosing constant coefficients $\rho_M(z) := \frac{1}{2n}$. The associated exploration parameters are $\omega \propto L^{-2}$, with a performance bound

$$\lim_{\delta\to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} = 8n^2 \cdot \sigma^2 \sum_{z\in\mathcal{Z}} \frac{1}{L_M(z)^2}.$$

In comparison with the lower bound from Lemma 8, this method is $8n^2$-asymptotically optimal when rewards follow Gaussian distributions of standard deviations σ.

Example 2 (LSTS-imp)**.** In general, the computation of the optimal coefficients $\rho_M(x,a)$ is difficult. Regarding condition *b)* from Lemma 9 and the performance bound from Theorem 2, the optimal choices of ρ are solution of the convex optimization problem below:

$$\min_\rho \sum_{z\in\mathcal{Z}} \frac{1}{\rho_M(z)^2 L_M(z)^2} \quad \text{s.t.} \quad \forall \pi, \sum_{z\in\pi\cup\pi^*} \rho_M(z) \leq 1. \tag{10}$$

Although the optimization problem is convex, the number of constraints is exponential and from it arises combinatorial difficulties related to the graph structure $\mathcal{G}$. To make the problem tractable, the constraints (10) are strengthened to $\sum_z \rho_M(z) \leq 1$ so the simplified problem becomes:

$$\min_{\rho} \sum_{z \in \mathcal{Z}} \frac{1}{\rho_M(z)^2 L_M(z)^2} \quad \text{s.t.} \quad \sum_{z \in \mathcal{Z}} \rho_M(z) \leq 1. \quad (11)$$

Proposition 10. *The solution of Eq.* (11) *is* $\rho_M \propto L^{-2/3}$.

The induced family $\mathcal{B}_\rho$ is Π^*_∞-constant and can be improved easily. Indeed, since ρ_M satisfies the condition $\sum_z \rho_M(z) \leq 1$, there is some slackness left with respect to the weaker conditions from (10) that can be removed without changing the ratios $\rho_M(z)/\rho_M(z')$. It means that this improvement won't change the exploration parameters of the LSTS-method but will induce wider boxes, thus will terminate quicker.

This is done like this: Denote ρ the element of the simplex such that $\rho \propto L^{-2/3}$, then pick the largest $\alpha \geq 1$ such that $\alpha\rho$ satisfies:

$$\forall \pi \in \Pi, \sum_{z \in \pi \cup \pi^*} \alpha \rho_M(z) \leq 1.$$

Then solve in α:

$$\alpha = \left(\max_{\pi \neq \pi^*} \sum_{z \in \pi \cup \pi^*} \rho_M(z) \right)^{-1}. \quad (12)$$

This value is can be shown to be tractable, inducing the following result.

Theorem 3 (LSTS-imp)**.** *Let $\rho \propto L^{-2/3}$. Let α be given by Equation* (12)*. The coefficients $\rho^\star := \alpha\rho$ satisfy the conditions a) and b) from Proposition 9. The LSTS method with such $\rho^\star$ is denoted LSTS-imp (standing for* improved *LSTS) and achieves:*

$$\lim_{\delta \to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} \leq 8\sigma^2 n^{4/3} m^{2/3} \sum_z \frac{1}{L_M(z)^2}.$$

When rewards are Gaussian with standard deviations σ, LSTS-imp is $8n^{4/3}m^{2/3}$-asymptotically optimal.

Sketch of the proof. Check that

$$\alpha \geq \left(2 \sum_x \max_a \rho_M(z) \right)^{-1}.$$

Recall that $\rho \propto L^{-2/3}$ then apply Theorem 2. Denote $K := \lim_{\delta \to 0} \frac{E^M[\tau_\delta]}{\log(1/\delta)}$ for short. Simple algebra leads to:

$$K \leq 8\sigma^2 \left(\sum_z L_M(z)^{-2/3} \right) \left(\sum_x \max_a L_M(x,a)^{-2/3} \right)^2.$$

The function $t \geq 0 \mapsto t^{1/3}$ is increasing, so $\max_a L_M(x,a)^{-2/3} = (\max_a L_M(x,a)^{-2})^{1/3}$. Bound both terms with Hölder's inequality with parameters $(3, \frac{3}{2})$, we bound K by

$$8\sigma^2 m^{2/3} \|L^{-1}\|_2^{2/3} \left(n^{2/3} \left(\sum_x \max_a L(x,a)^{-2} \right)^{1/3} \right)^2.$$

Bounding $\max_a$ by $\sum_a$ yields the result. □

4.4 Benefits of LSTS-imp

When all local suboptimality gaps are equal, LSTS-imp and LSTS-cst perform the same, as they explore according to the same exploration coefficients. In general, Theorem 3 only guarantees $8n^{4/3}m^{2/3}$ optimality, which is worse than $8n^2$ of LSTS-cst. So, was the optimization attempt of ρ worth the shot? One observation is that the $8n^2$ factor in LSTS-cst's performances is model agnostic. Specifically, it is always $8n^2$ away from the lower bound given by Proposition 7, while LSTS-imp's performance bounds relies on Hölder's inequality which is pessimistic in many scenarios. In opposition to LSTS-cst, LSTS-imp can take advantage of instances where the number of *critical* edges is small, as in the following example.

Example 3. Consider the family of DMDPs (M_n) given by Figure 2.

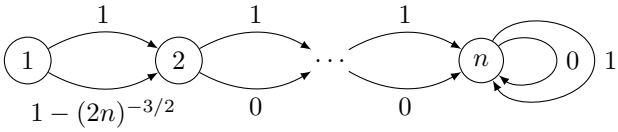

Figure 2: Family (M_n), $n \geq 1$. Rewards are Gaussian with standard deviation σ. Check that $g^{M_n}_*(i) = 1$ and $h^{M_n}_*(i) = 0$ for all $i \leq n$. Also, for all action a, $L_{M_n}(i,a) = 1$ for $i \geq 2$ and $L_{M_n}(1,a) = (2n)^{-3/2}$.

The graph is 2-regular, so we know that LSTS-imp is $12.7n^2$-optimal, so performs *a priori* worse than LSTS-cst. In fact, LSTS-imp is much better. Local suboptimality gaps are all equal to 1, excepted for the edges from 1 to 2, where gaps are $(2n)^{-3/2}$. The lower bound given by Lemma 7 is

$$\sum_z \frac{\sigma^2}{L_{M_n}(z)^2} = 2\sigma^2((2n)^3 + n - 1) \geq 2^4 \cdot \sigma^2 n^3.$$

Ignoring the improvement of LSTS-imp due to the scaling in α, we know by the conjugation of Proposition 10 and Theorem 2 that LSTS-imp's performances are bounded by

$$\sum_z \frac{2\sigma^2}{\rho_{M_n}(z)^2 L_{M_n}(z)^2} = 2\sigma^2 \left(\sum_z \frac{1}{L_{M_n}(z)^{2/3}} \right)^3.$$

Simple algebra leads to

$$\lim_{\delta\to 0}\frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} = 2\sigma^2(3n-1)^3 \leq 2^7\cdot\sigma^2 n^3.$$

This compares to the lower bound up to a constant factor equal to 8. Namely, LSTS-imp is 8-asymptotically optimal on the family (M_n), thus completely outperforms the $8n^2$-asymptotical optimality of LSTS-cst.

The last result below generalizes the method used in the previous example by providing a general template to bound the performances of LSTS-imp when *few state-action pairs are critical*.

Proposition 11. *Assume $M \in \mathcal{H}$ have sub-Gaussian rewards with standard deviation $\sigma > 0$. Assume that there exist $\epsilon > 0$ together with a family of distinct state-action pairs $(z_i)_{i=1}^k$ such that*

$$\sum\nolimits_{z\in\mathcal{Z}} L_M(z)^{-2} \leq (1+\epsilon)\sum\nolimits_{i=1}^{k} L_M(z_i)^{-2}.$$

Thus, if rewards are Gaussian with standard deviations σ, LSTS-imp is $8(k^2+\epsilon m^2)$-asymptotically optimal.

5 PERSPECTIVES

Among possible future research directions, in addition to the general non-deterministic case, the search for improvements of the LSTS scheme is interesting. A first promising direction is the investigation of algorithms relying on more refined confidence sets, e.g. ellipsoid-shaped rather than square-shaped. The latter is seemingly critical and may lead to a significant improvement in performances. See Kaufmann and Koolen (2021) for a discussion on the disadvantage of square-shaped confidence sets in the case of multi-armed bandits. Another promising direction is to find a convincing way to *quantify* near-Blackwell optimality. This would open new directions, with new learning problems such as higher order undiscounted regret minimization and minimax PAC Blackwell optimality learning.

References

Auer, P., Jaksch, T., and Ortner, R. (2009). Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Azar, M. G., Munos, R., Ghavamzadaeh, M., and Kappen, H. J. (2011). Speedy Q-learning. Publisher: Spain, Granada: NIPS.

Blackwell, D. (1962). Discrete dynamic programming. *The Annals of Mathematical Statistics*, pages 719–726. Publisher: JSTOR.

Bourel, H., Maillard, O., and Talebi, M. S. (2020). Tightening Exploration in Upper Confidence Reinforcement Learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1056–1066. PMLR.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure Exploration in Multi-armed Bandits Problems. In Gavaldà, R., Lugosi, G., Zeugmann, T., and Zilles, S., editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 23–37, Berlin, Heidelberg. Springer.

Cochet-Terrasson, J., Cohen, G., Gaubert, S., McGettrick, M., and Quadrat, J.-P. (1998). Numerical Computation of Spectral Elements in Max-Plus Algebra *. *IFAC Proceedings Volumes*, 31(18):667–674.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021). Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR.

Fiechter, C.-N. (1994). Efficient reinforcement learning. In *Proceedings of the seventh annual conference on Computational learning theory*, COLT '94, pages 88–97, New York, NY, USA. Association for Computing Machinery.

Filippi, S., Cappé, O., and Garivier, A. (2010). Optimism in Reinforcement Learning and Kullback-Leibler Divergence. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. arXiv: 1004.5229.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning. *Proceedings of the 35 th International Conference on Machine Learning*.

Garivier, A. and Kaufmann, E. (2016). Optimal Best Arm Identification with Fixed Confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR. ISSN: 1938-7228.

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349.

Karp, R. M. (1978). A characterization of the minimum cycle mean in a digraph. *Discrete Mathematics*, 23(3):309–311.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42.

Kaufmann, E. and Koolen, W. M. (2021). Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. *Journal of Machine Learning Research*. Publisher: Microtome Publishing.

Kearns, M. and Singh, S. (1998). Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, 11.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021). Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. *arXiv:2005.12900 [cs, math, stat]*. arXiv: 2005.12900.

Li, Y., Wang, R., and Yang, L. F. (2022). Settling the horizon-dependence of sample complexity in reinforcement learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 965–976. IEEE.

Marjani, A. A., Garivier, A., and Proutiere, A. (2021). Navigating to the Best Policy in Markov Decision Processes. *arXiv:2106.02847 [cs, stat]*. arXiv: 2106.02847.

Marjani, A. A. and Proutiere, A. (2021). Adaptive Sampling for Best Policy Identification in Markov Decision Processes. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7459–7468. PMLR. ISSN: 2640-3498.

Ortner, R. (2010). Online regret bounds for Markov decision processes with deterministic transitions. *Theoretical Computer Science*, 411(29):2684–2695.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Shah, D., Xie, Q., and Xu, Z. (2020). Non-Asymptotic Analysis of Monte Carlo Tree Search. *ACM SIGMETRICS Performance Evaluation Review*, 48(1):31–32.

Strehl, A. L. (2007). *Probably approximately correct (PAC) exploration in reinforcement learning*. PhD Thesis, Rutgers University-Graduate School-New Brunswick.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. (2021). Sample complexity bounds for stochastic shortest path with a generative model. In *Algorithmic Learning Theory*, pages 1157–1178. PMLR.

Tranos, D. and Proutiere, A. (2021). Regret Analysis in Deterministic Reinforcement Learning. *arXiv:2106.14338 [cs, stat]*. arXiv: 2106.14338.

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. (2020). Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10170–10180. PMLR.

This appendix contains the proofs of all the propositions and theorems of the paper (they are restated in the appendix) as well as the proofs of additional lemmas needed in the main proofs or with an interest of their own.

General Notations.

SYMBOL	MEANING	COMMENT
$\mathcal{S}, n$	set of vertices or states, number of vertices	
$\mathcal{A}, \mathcal{A}_x$	set of actions, set of legal actions from $x \in \mathcal{S}$	
$\mathcal{Z}, m$	state-action space $\coprod_x \mathcal{A}_x$, number of transitions or also $\lvert\mathcal{Z}\rvert$	Section 2.2
$\mathcal{M}$	set of all (deterministic) MDPs	Section 2.2
$\mathcal{H}, \mathcal{M}_{\mathrm{BW}}$	set of DMDPs under H1 and H2	Section 3.1
$\mathcal{R}(\mathcal{H})$	set of mean reward vectors of elements of $\mathcal{H}$	
x, y	vertices, states	
a, b	actions	
z, z'	state-action pairs	Section 2.2
$P(y\lvert x, a)$	transition probability from x to y using a	
$s(x, a)$	successor of x by using a	Definition 2
$q(x, a)$	reward distribution on $(x, a) \in \mathcal{Z}$	Section 2
$r(x, a)$	mean reward on $(x, a) \in \mathcal{Z}$	Section 2
π, π', π''	policy	Sections 2 and 2.2
Π, $\Pi(M)$	set of policies	Section 2
Π_∞, $\Pi_\infty(M)$	set of Blackwell optimal policies	Definition 1
Π_{-1}, $\Pi_{-1}(M)$	set of gain optimal policies	Section 2.1
Π_0, $\Pi_0(M)$	set of bias optimal policies	Section 2.1
Π_k, $\Pi_k(M)$	set of k-discounted optimal policies, $k \geq -1$	Section 2.1
r_π, $r^\pi(x)$, $r_x^\pi(M)$	mean reward vector of π, mean reward of π from x in M	Section 2
$v_\beta^\pi(x)$, $v_\beta^\pi(x, M)$	discounted score of π from x in M	Equation (1)
C_x^π	terminal cycle of the iterates of π from x	Equation (5)
g_π, $g_\pi(x)$, $g_x^\pi(\mathbf{r})$	gain vector of π, gain of π from x under rewards $\mathbf{r}$	Equations (2) and (5)
h_π, $h_\pi(x)$, $h_x^\pi(\mathbf{r})$	bias vector of π, bias of π from x under rewards $\mathbf{r}$	Equation (3),Proposition 1
$h_\pi^{(k)}$, $h_\pi^{(k)}(x)$	k-bias vector of π, bias of π from x in M	Equation (4)
$N_t(z), \mathbf{N}_t$	number of visits of z from time 1 to t (exclusive on t), vector of	Section 3
$\hat{M}_t$	MDP of maximum likelihood or *empirical* MDP at time t	Section 3
$\widetilde{\mathcal{M}}_t^\delta$	confidence region with confidence δ at time t	Equation (8)
$\hat{\mathbf{r}}_t, \hat{r}_t(x, y)$	empirical mean reward at time t	Section 3
$\widetilde{\mathbf{r}}_t, \widetilde{r}_t(x, y)$	noisy mean reward at time t	Equation (15)
$\epsilon_\delta(s, t)$	Hoeffding bonus at time t for an edge sampled s times, confidence δ	Equation (9)
X_t	random picked state at time t	
A_t	random picked action at time t	
$Z_t := (X_t, A_t)$	sampled edge at time t	
R_t	random reward at time t	
$r(Z_t)$	pseudo reward at time t	
δ	confidence threshold	
τ_δ	stopping time under confidence δ	Section 3
$\mathcal{I}$	identification algorithm	Section 3
$\mathbb{P}_\mu^{M,I}\{\cdot\}$	probability measure under M when iterating $\mathcal{I}$	
$\mathbb{E}_\mu^{M,I}[\cdot]$	expectation under M when iterating $\mathcal{I}$	
$\Delta_M(x, a)$	Bellman gap of (x, a) for M	Equation (7)
$L_M(x, a)$	local suboptimality of (x, a) for M	Definition 5
$\mathcal{B}(M)$	box at $M \in \mathcal{H}$	Definition 4
$\ell_{\mathcal{B}(M)}(x, a)$	half-width of box $\mathcal{B}(M)$ at (x, a)	Definition 4
$\rho_M(x, a)$	family extra parameters	Proposition 9
$\omega_M(x, a)$	exploration coefficients	Algorithm 1

A APPENDIX: GAIN AND BIAS

Proposition 1. *Let $\pi : \mathcal{S} \to \mathcal{A}$ a policy and $x \in \mathcal{S}$. Denote x_t the state at time t under the iterations of π (from $x = x_0$), and let $z_t := (x_t, \pi(x_t))$. If $T \geq 0$ is such that $x_T \in \mathcal{C}_x^\pi$, then the bias expands as*

$$h_\pi(x) = \sum_{t=0}^{T-1} [r(z_t) - g(\mathcal{C}_x^\pi)] \qquad (transient)$$
$$+ \frac{1}{|\mathcal{C}_x^\pi|} \sum_{\ell=1}^{|\mathcal{C}|} \sum_{t=0}^{\ell-1} [r(z_{T+t}) - g(\mathcal{C}_x^\pi)]. \qquad (recurrent)$$

Proof. From the definition of the bias, that is $h_\pi(x) := C\text{-lim}\, \mathbb{E}_x^{M,\pi}[\sum_{t=0}^{T-1}[r_t - g(\mathcal{C}_x^\pi)]]$, follows that

$$h_\pi(x_0) = \sum_{t=0}^{T-1} [r(z_t) - g(\mathcal{C}_x^\pi)] + h_\pi(x_T).$$

Let $u_0 = x_T$ and denote $u_t := \pi^t(u_0)$ the t-th state reached by π from u_0 and let $z'_t := (u_t, \pi(u_t))$. Introduce the partial sums $S_t = \sum_{\tau=0}^{t-1}[r(z'_\tau) - g(\mathcal{C}_\pi^*)]$. Remark that for $c := |\mathcal{C}_x^\pi|$ and all $t \geq 0$, $z'_t = z'_{t-\lfloor t/c \rfloor c}$. Together with the observation $S_c = 0$, we see that

$$S_t = S_{t-\lfloor t/c \rfloor c}.$$

Moreover, $h_\pi(u_0) = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} S_\tau$ by definition, hence

$$h_\pi(u_0) = \lim_{t\to\infty} \left(\frac{1}{t} \sum_{\ell=1}^{c} S_\ell \sum_{\tau=0}^{t-1} \mathbf{1}_{\tau \equiv \ell[c]} \right) = \sum_{\ell=1}^{c} S_\ell \lim_{t\to\infty} \left(\frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{1}_{\tau \equiv \ell[c]} \right).$$

Thus $h_\pi(u_0) = \frac{1}{c} \sum_{\ell=1}^{c} S_\ell$. Combine the previous identity with $h_\pi(x_0) = \sum_{t=0}^{T-1}[r(z_t) - g(\mathcal{C}_x^\pi)] + h_\pi(u_0)$ to get the result. □

From H1, all gain optimal policies have the same terminal cycle $\mathcal{C}_*$ over which they coincide. In particular, we see from Lemma 1 that they share the same bias on $\mathcal{C}_*$. Specifically, for $(u_0, u_1, \dots, u_{c-1})$ the decomposition of $\mathcal{C}_*$ introduced in H2, for every gain optimal policy π and $i \leq c-1$, $h_\pi(u_i) = h_*(u_i) := \frac{1}{c} \sum_{\ell=1}^{c} \sum_{k=0}^{l-1} [r(u_{i+k}, \pi(u_{i+k})) - g(\mathcal{C}_*)]$. Therefore, if $\pi \in \Pi_{-1}^*(M)$ and $x_0 \in \mathcal{S}$, for k the smallest integer such that $x_k \in \mathcal{C}_*$, we have

$$h_\pi(x_0) = \sum_{i=0}^{k-1} [r(x_i, \pi(x_i)) - g(\mathcal{C}_*)] + h_*(x_k).$$

Accordingly, from H2 follows the uniqueness of $\pi^* \in \Pi_0^*(M)$ thus bias optimal policies are Blackwell optimal on $\mathcal{H}$.

Lemma 1. *Let $\mathbf{r} \in \mathbb{R}^m$ and $\pi \in \Pi$. Recall that $\mathcal{C}_x^\pi$ denotes the terminal cycle reached by iterating π from x. For all x,*

$$g_x^\pi(\mathbf{r}) = |\mathcal{C}_x^\pi|^{-1} \sum_{z \in \mathcal{C}_x^\pi} r(z).$$

Proof. This is a direct consequence of the fact that under any policy π, a DMDP is a weighted directed graph. All infinite trajectories end up cycling over one simple cycle[3] and their average weight is the average weight of this terminal cycle. □

Lemma 2. *Let $\mathbf{r} \in \mathbb{R}^m$ and $\pi \in \Pi$. Recall that $\mathcal{C}_x^\pi$ denotes the terminal cycle reached by iterating π from x. Denote x_t the state visited at time t by iterating π from the initial state $x = x_0 \in \mathcal{S}$. Introduce the reaching times:*

$$\tau_x^\pi(y) := \inf \{t \geq 0 : x_t = y\} \in \mathbb{N} \cup \{\infty\}.$$

Write $x \rightsquigarrow_\pi^* y$ when $\tau_x^\pi(y) < \infty$. For all $x \in \mathcal{Z}$,

$$h_x^\pi(\mathbf{r}) = \sum_{(y,b) \in \pi \setminus \mathcal{C}_x^\pi} \mathbf{1}_{x \rightsquigarrow_\pi^* y} r(y,b) + \sum_{(y,b) \in \mathcal{C}_x^\pi} \left(1 - \frac{|\mathcal{C}_x^\pi| + 1}{2|\mathcal{C}_x^\pi|} - \frac{\tau_x^\pi(y)}{|\mathcal{C}_x^\pi|} \right) r(y,b).$$

[3] A cycle is *simple* if it cannot be written as a union of cycles.

Proof. Assume that $r(y,b)$ is null everywhere excepted at some coordinate $(y,b) \in \mathcal{Z}$, i.e., $\mathbf{r} = \alpha \cdot \mathbf{e}_{y,b}$. Let $x \in \mathcal{S}$.

If $(y,b) \notin \pi$, then $h_x^\pi(\alpha \mathbf{e}_{y,b}) = 0$ since π never uses the transition.

If $(y,b) \in \pi \setminus \mathcal{C}_x^\pi$, expand the bias $h_x^\pi(\mathbf{r})$ according to Lemma 1 with T such that $x_T \in \mathcal{C}_x^\pi$. Then

$$\begin{aligned} h_x^\pi(\alpha \mathbf{e}_{y,b}) &= \sum_{t=0}^{T-1} [\alpha \mathbf{e}_{y,b}(x_t, \pi(x_t)) - g_x^\pi(\alpha \mathbf{e}_{y,b})] + h_{x_T}^\pi(\alpha \mathbf{e}_{y,b}) \\ &= \sum_{t=0}^{T-1} \alpha \mathbf{e}_{y,b}(x_t, \pi(x_t)) = \alpha \mathbf{1}_{x \rightsquigarrow_\pi^* y}. \end{aligned}$$

If $(y,b) \in \mathcal{C}_x^\pi$, choose $T = \tau_x^\pi(y)$ i.e. the smallest t such that $x_t = y$. Using Lemma 1 again,

$$\begin{aligned} h_x^\pi(\alpha \mathbf{e}_{y,b}) &= -\sum_{t=1}^{\tau_x^\pi(u)} g_x^\pi(\alpha \mathbf{e}_{y,b}) + \frac{1}{|\mathcal{C}_x^\pi|} \sum_{\ell=1}^{|\mathcal{C}_x^\pi|} \sum_{t=0}^{\ell-1} [\alpha \mathbf{e}_{y,b}(x_t, \pi(x_t)) - g_x^\pi(\alpha \mathbf{e}_{y,b})] \\ &= -\frac{\tau_x^\pi(y)}{|\mathcal{C}_x^\pi|}\alpha + \frac{1}{|\mathcal{C}_x^\pi|} \sum_{\ell=1}^{|\mathcal{C}_x^\pi|} \left(1 - \frac{\ell}{|\mathcal{C}_x^\pi|}\right) \alpha \\ &= \left(1 - \frac{|\mathcal{C}_x^\pi| + 1}{2|\mathcal{C}_x^\pi|} - \frac{\tau_x^\pi(y)}{|\mathcal{C}_x^\pi|}\right) \alpha. \end{aligned}$$

We conclude by linearity of $\mathbf{r} \mapsto h_x^\pi(\mathbf{r})$. □

B APPENDIX: ASSUMPTIONS H1 AND H2

Theorem 1. *Let $\mathcal{I}$ be an identification algorithm and let $\delta \in (0, \frac{1}{4}n^{-n})$. Let $\mathcal{D}_\mathcal{I}$ be the set of DMDPs on which $\mathcal{I}$ is δ-probably correct. The two following assertions hold:*

(*i*) *The interior of $\mathcal{D}_\mathcal{I}$ is in $\mathcal{H}$: $\mathring{\mathcal{D}}_\mathcal{I} \subseteq \mathcal{H}$;*
(*ii*) *If $\mathcal{D}_\mathcal{I}$ contains $\mathcal{H}$, $\mathcal{D}_\mathcal{I} = \mathcal{H}$.*

Proof. Let $\mathcal{I}$ be an ϵ-PC identification algorithm that stops almost surely on $M \in \mathring{\mathcal{D}}_I \setminus \mathcal{H}$ with $\epsilon \in (0, \frac{1}{4}n^{-n})$. There exists $T > 0$ such that $\mathbb{P}^{M,\mathcal{I}}\{\tau_\epsilon < T\} > \frac{1}{2}$. The total number of policies is bounded by n^n, so there must exist a policy π such that $\mathbb{P}^{M,\mathcal{I}}\left\{\tau_\epsilon < T, \pi_{\tau_\epsilon}^\mathcal{I} = \pi\right\} > \frac{1}{2}n^{-n}$. Since $\mathcal{I}$ is ϵ-PC with $\epsilon < \frac{1}{2}n^{-n}$, $\pi \in \Pi_\infty^*(M)$. Now, we construct a MDP M' such that (1) $\pi \notin \Pi_\infty^*(M')$ and (2) for all $\mathcal{D} \in \mathcal{F}_T$, we have

$$\mathbb{P}^{M',\mathcal{I}}\left\{H_{\min(T,\tau_\epsilon)} \in \mathcal{D}\right\} \geq \tfrac{1}{2}\mathbb{P}^{M,\mathcal{I}}\left\{H_{\min(T,\tau_\epsilon)} \in \mathcal{D}\right\},$$

where H_T denotes the history up to time T, that is $H_T := (X_0, Y_0, R_0, \ldots, X_{t-1}, Y_{t-1}, R_{t-1})$. Let

$$A = 1 + \max_{(x,a)\in\mathcal{Z}} r(x,a).$$

H1 Assume that M does not satisfy H1 and write $\pi_1 := \pi \in \Pi_{-1}^*(M)$. Because H1 does not hold, there must exists $x \in \mathcal{S}$ and $\pi_2 \in \Pi_{-1}^*(M)$ such that the terminal cycle $\mathcal{C}_1$ from x by using π_1 and the terminal cycle $\mathcal{C}_2$ from x by using π_2 are different. Let $(u,v) \in \mathcal{C}_2 \setminus \mathcal{C}_1$ and define M' the copy of M such that $q'(u,a) = (1-(\frac{1}{2})^{1/T})\delta_A + (\frac{1}{2})^{1/T} q(u,a))$, meaning that for any Borelian set $\mathcal{U} \in \mathcal{B}([0,1])$,

$$q'(u,v)(\mathcal{U}) = \left(1 - \left(\tfrac{1}{2}\right)^{\frac{1}{T}}\right) \delta_A(\mathcal{U}) + \left(\tfrac{1}{2}\right)^{\frac{1}{T}} q(u,v)(\mathcal{U})$$

as probability measures (δ_A is the Dirac at A). Then by construction $r'(u,a) > r(u,a)$ so $g'_{\pi_2}(x) > g'_{\pi_1}(x)$. Therefore, $\pi_1 \notin \Pi_{-1}^*(M')$. In addition, if $Z_1, \ldots, Z_T$ are i.i.d. random variables with laws $q'(u,v)$, then for all $t \leq T$ and

$(\mathcal{U}_1, \dots, \mathcal{U}_T) \in \prod_{i=1}^T \mathcal{B}([0,1])$,

$$\mathbb{P}\left\{(Z_1, \dots, Z_t) \in \prod_{i=1}^t \mathcal{U}_i\right\} = \prod_{i=1}^t \mathbb{P}\left\{Z_i \in \mathcal{U}_i\right\} \geq \tfrac{1}{2} \prod_{i=1}^t q(u,v)(\mathcal{U}_i).$$

It follows that for any event $\mathcal{D} \in \mathcal{F}_T$,

$$\mathbb{P}^{M',\mathcal{I}}\left\{H_T \in \mathcal{D}\right\} \geq \tfrac{1}{2}\mathbb{P}^{M,\mathcal{I}}\left\{H_T \in \mathcal{D}\right\}.$$

This is for the construction when M does not satisfy H1.

H2 The construction is similar when M does not satisfy H2. Find $x \in \mathcal{S}$ such that $\pi_1(x) \neq \pi_2(x)$ and change $q(x, \pi_2(x))$ to $q'(x, \pi_2(x)) := (1 - (\frac{1}{2})^{1/T})\delta_A + (\frac{1}{2})^{1/T} q(x, \pi_2(x))$. If x belongs to a terminal cycle of π_2, then $g'_{\pi_2}(x) > g'_{\pi_1}(x)$ so $\pi_1 \notin \Pi^*_{-1}(M)$; otherwise x is transient for π_2 and we get $h'_{\pi_2}(x) > h_{\pi_2}(x)$ with $h'_{\pi_1}(x) = h_{\pi_1}(x)$, so $\pi_1 \notin \Pi^*_0(M)$.

Then, by taking $\mathcal{D} := \{H_T \text{ such that } \tau_\epsilon < T \text{ and } \pi^{\mathcal{I}}_{\tau_\epsilon} = \pi\}$, we obtain

$$\mathbb{P}^{M',\mathcal{I}}\left\{\pi^{\mathcal{I}}_{\tau_\epsilon} = \pi\right\} > \tfrac{1}{4} n^{-n}.$$

As $\pi \notin \Pi^*_\infty(M')$, follows $\mathbb{P}^{M',\mathcal{I}}\left\{\pi^{\mathcal{I}}_{\tau_\epsilon} \notin \Pi^*_\infty(M')\right\} > \frac{1}{4} n^{-n}$. This proves the Theorem. □

B.1 Topological properties of $\mathcal{H}$

To be accurate, the property $M \in \mathcal{H}$ does not depend on the reward distributions $q(x,a)$ but rather on their means $r(x,a)$. Denote $\mathcal{R}(\mathcal{H})$ the set of mean reward vectors that correspond to a DMDP of $\mathcal{H}$. As a subset of $\mathbb{R}^{\mathcal{Z}}$, $\mathcal{R}(\mathcal{H})$ inherits its natural topology, making it a Baire space.

Lemma 3. *$\mathcal{R}(\mathcal{H})$ is open.*

Proof. Let $M \in \mathcal{H}$ with mean reward vector r and let π its bias-optimal policy. To emphasize that the gain and the bias of π from a given state $x \in \mathcal{S}$ are functions of $\mathbf{r} \in \mathbb{R}^{\mathcal{Z}}$, we write $g^\pi_x(\mathbf{r})$ and $h^\pi_x(\mathbf{r})$. From Theorem 3, π remains optimal under reward profile $\mathbf{r}'$, and its terminal is the unique maximal mean weight cycle if and only if

$$\forall (x,a) \notin \pi, \quad r'(x,a) - g^\pi_x(\mathbf{r}') + h^\pi_y(\mathbf{r}') < h^\pi_x(\mathbf{r}').$$

Recall that $\Delta_{x,a}(\mathbf{r}) = h^\pi_x(\mathbf{r}) - [r(x,a) - g^\pi_x(\mathbf{r}) + h^\pi_y(\mathbf{r})]$. Denote $\mathrm{d}\mathbf{r} := \mathbf{r}' - \mathbf{r}$. Using the linearity of the gain and the bias with respect to r, we obtain that π is the unique bias optimal policy under reward profile r' if, and only if for all $(x,a) \notin \pi^*$,

$$\mathrm{d}\mathbf{r}(x,a) - g^\pi_x(\mathrm{d}\mathbf{r}) + h^\pi_y(\mathrm{d}\mathbf{r}) - h^\pi_x(\mathrm{d}\mathbf{r}) < \Delta_{x,a}(\mathbf{r}). \tag{13}$$

The function $\mathrm{d}\mathbf{r} \mapsto dr(x,a) - g^\pi_x(\mathrm{d}\mathbf{r}) + h^\pi_y(\mathrm{d}\mathbf{r}) - h^\pi_x(\mathrm{d}\mathbf{r})$ is a continuous function of $\mathrm{d}\mathbf{r}$ and $\Delta_{x,a}(\mathbf{r}) > 0$, thus the set of $\mathrm{d}\mathbf{r}$ satisfying the equation (13) is an open set $U_{x,a} \subseteq \mathbb{R}^{\mathcal{Z}}$. One obtains the set of $\mathbf{r}'$ such that π is the bias unique optimal policy and its terminal cycle the unique maximal mean weight cycle as $\mathbf{r} + \bigcap_{(x,a)\notin\pi} U_{x,a}$, which is open. □

Proposition 2. *$\mathcal{H}$ is generic in Baire categories sense. In particular, $\mathcal{M} \setminus \mathcal{H}$ has null Lebesgue measure.*

Proof. We show first that $\mathbb{R}^{\mathcal{Z}} \setminus \mathcal{R}(\mathcal{H})$ is *meagre* in Baire category sense, i.e., is a countable union of closed set with empty interior. In fact, we show the stronger statement: *$\mathbb{R}^{\mathcal{Z}} \setminus \mathcal{R}(\mathcal{H})$ is a finite union of hyperplanes.*

Let $\mathcal{C}$ and $\mathcal{C}'$ two (different) simple cycles of $\mathcal{G}$. The difference between their mean weight under reward profile $\mathbf{r} \in \mathbb{R}^{\mathcal{Z}}$ is given by

$$f(\mathbf{r}) := \frac{1}{|\mathcal{C}|} \sum_{(x,a)\in\mathcal{C}} r(x,a) - \frac{1}{|\mathcal{C}'|} \sum_{(x,a)\in\mathcal{C}'} r(x,a)$$

which is a non-trivial linear function of $\mathbf{r}$. Thus, $\mathcal{C}$ and $\mathcal{C}'$ have the same mean weights if, and only if $\mathbf{r} \in \ker f$ which is an hyperplane of $\mathbb{R}^{\mathcal{Z}}$. The set of simple cycles being finite, the set Γ of mean rewards such that at least two simple cycles have the same mean weights is hence a finite union of hyperplanes of $\mathbb{R}^{\mathcal{Z}}$.

Then, we show that two distinct policies that share the same terminal cycle only have the same bias for reward vectors on a finite union of hyperplanes of $\mathbb{R}^{\mathcal{Z}}$. Let $\mathbf{r} \notin \Gamma$ and for $\mathcal{C}$ a simple cycle, let $\Pi_{\mathcal{C}}$ the set of policy with terminal cycle $\mathcal{C}$. Let $\pi \neq \pi' \in \Pi_{\mathcal{C}}$. There is some $x \in \mathcal{S}$ such that $\pi(x) \neq \pi'(x)$. Then with the same notations as in the proof of Lemma 3, again, $h_x^{\pi}(\mathbf{r}) - h_x^{\pi'}(\mathbf{r})$ is a non-trivial linear function of $\mathbf{r}$. Hence $h_x^{\pi}(\mathbf{r}) = h_x^{\pi'}(\mathbf{r})$ if, and only if $\mathbf{r}$ belongs to some hyperplane $H_x^{\pi,\pi'}$ of $\mathbb{R}^{\mathcal{Z}}$. Finally,

$$\mathcal{R}(\mathcal{H})^{\complement} = \Gamma \cup \bigcup_{\mathcal{C} \text{ simple}} \bigcup_{x \in \mathcal{S}} \bigcup_{\substack{\pi,\pi' \in \Pi_{\mathcal{C}} \\ \pi(x) \neq \pi'(x)}} H_x^{\pi,\pi'}$$

is a finite union of hyperplanes.

Because a hyperplane has null Lebesgue measure (denoted λ), $\mathcal{R}(\mathcal{H})^{\complement}$ has null Lebesgue measure by union bound. □

This means that if $\mathbf{r}$ is chosen randomly according to a probability measure $\mu \ll \lambda$, then $\mathbf{r} \in \mathcal{R}(\mathcal{H})$ with probability one.

B.2 Characterization of the bias optimal policy

Proposition 3. *Let $M \in \mathcal{H}$ a DMDP. A policy $\pi \in \Pi$ is bias optimal if, and only if $g_\pi(x)$ does not depend on $x \in \mathcal{S}$ and*

$$\forall (x,a) \notin \pi, \quad r(x,a) - g_\pi(x) + P(x,a)h_\pi < h_\pi(x).$$

In particular, π_M^* is the unique policy with unique terminal cycle such that this equation holds. Moreover, a DMDP satisfies H1 and H2 if, and only if there is a policy satisfying this equation with unique terminal cycle.

Proof. The proof is done in several steps.

Characterization of bias optimality on $\mathcal{H}$. Let us assume $M \in \mathcal{H}$ and let $\pi \in \Pi$ with $g_\pi(x)$ independent of $x \in \mathcal{S}$ and such that

$$\forall (x,a) \in \pi^*, \qquad r(x,a) - g_\pi(x) + h_\pi(y) < h_\pi(x) \tag{14}$$

where $y := s(x,a)$. For $(x,a) \in \mathcal{Z}$, define $\Delta^\pi(x,a) := h_\pi(x) - [r(x,a) - g_\pi(x) + h_\pi(y)]$. By definition of $h_\pi(x)$, we have $\Delta^\pi(x,a) = 0$ for all $(x,a) \in \pi$ and $\Delta^\pi(x,a) > 0$ everywhere else. By assumption, we may write g_π instead of $g_\pi(x)$. Let $(u_0, a_1, u_1, a_2, u_2, \ldots, a_k, u_k)$ be a path in $\mathcal{G}$. One checks by induction on $k \geq 1$ that

$$\sum_{i=0}^{k-1} [r(u_i,a_i) - g_\pi] = h_\pi(u_0) - h_\pi(u_k) - \sum_{i=0}^{k-1} \Delta^\pi(u_i,a_i).$$

Choose $(u_0, a_0, \ldots, u_{c-1}, a_{c-1}, u_c)$ any cycle $\mathcal{C}$ of $\mathcal{G}$ i.e. $u_0 = u_c$. From the formula above follows

$$-\sum_{i=0}^{c-1} \Delta^\pi(u_i,u_{i+1}) = \sum_{i=0}^{c-1} [r(u_i,u_{i+1}) - g_\pi] = |\mathcal{C}|[g(\mathcal{C}) - g_\pi].$$

As all $\Delta^\pi(x,a)$ are non-negative, applying the formula with $\mathcal{C} = \mathcal{C}_*$ the optimal cycle of M, we get that $g_\pi = g_*$. In particular, π has $\mathcal{C}_*$ for unique terminal cycle so $\pi \in \Pi_{-1}^*(M)$. To show that $\pi \in \Pi_0^*(M)$, compare the bias of π to the one of the optimal policy π^*. Let $x \in \mathcal{S}$ and k minimal such that $x_k := \pi^{*k}(x) \in \mathcal{C}_*$. We have

$$\begin{aligned} h_*(x_0) - h_*(x_k) &= \sum_{i=0}^{k-1} [r(x_i,a_i) - g_*] \\ &= h_\pi(x_0) - h_\pi(x_k) - \sum_{i=0}^{k-1} \Delta^\pi(x_i,a_i) \\ &\leq h_\pi(x_0) - h_\pi(x_k). \end{aligned}$$

But x_k belongs to $\mathcal{C}_*$, thus starting from x_k, the iterates of π and π^* coincide. Thus $h_\pi(x_k) = h_*(x_k)$. Whence $h_*(x_0) \leq h_\pi(x_0)$. Accordingly, $\pi \in \Pi_0^*(M)$.

Characterization of $\mathcal{H}$. If $M \in \mathcal{H}$, we know that the bias optimal policy has constant gain vector and is such that (14) holds. The proof of the converse statement is very similar to the previous argument. So, conversely, assume that there exists π with unique terminal cycle satisfying (14). The formula

$$\forall (x,a) \in \pi^*, \qquad r(x,a) - g_\pi(x) + h_\pi(y) < h_\pi(x)$$

is still valid. Thus for $(u_0, a_0, \dots, u_{c-1}, a_{c-1}, u_c)$ any cycle $\mathcal{C}$, we have

$$|\mathcal{C}|[g(\mathcal{C}) - g_\pi] = -\sum_{i=0}^{c-1} \Delta^\pi(u_i, a_i).$$

Any cycle $\mathcal{C} \neq \mathcal{C}^\pi$ is such that $\sum_{i=0}^{c-1} \Delta^\pi(u_i,a_i) > 0$, i.e., $g(\mathcal{C}) < g_\pi = g(\mathcal{C}^\pi)$. This means that the optimal cycle is unique, so M satisfies H1. Finally, to show that π is the unique bias optimal policy, consider $\pi' \in \Pi^*_{-1}(M)$ different from π and show that there exists $x \in \mathcal{S}$ such that $h_{\pi'}(x) < h_\pi(x)$. There exists $(x,a) \in \pi' \setminus \pi$, so that $\Delta^\pi(x,a) > 0$. Let k minimal such that $x_k := \pi'^k(x) \in \mathcal{C}_*$. Then,

$$\begin{aligned} h_{\pi'}(x_0) - h_{\pi'}(x_k) &= \sum_{i=0}^{k-1} [r(x_i,a_i) - g_\pi] \\ &= h_\pi(x_0) - h_\pi(x_k) - \sum_{i=0}^{k-1} \Delta^\pi(x_i,a_i) \\ &< h_\pi(x_0) - h_\pi(x_k). \end{aligned}$$

But since $x_k \in \mathcal{C}_* = \mathcal{C}^\pi = \mathcal{C}^{\pi'}$, $h_\pi(x_k) = h_{\pi'}(x_k)$. Thus $h_{\pi'}(x_0) < h_\pi(x_0)$. □

C APPENDIX: ANALYSIS OF LSTS

To settle notations, the noisy perturbation vector of $\hat{\mathbf{r}}_t$ is denoted

$$\widetilde{\mathbf{r}}_t := (1 - \tfrac{1}{t})\hat{\mathbf{r}}_t + \tfrac{1}{t}\mathbf{u}_t \tag{15}$$

where $\mathbf{u}_t \sim \mathcal{U}([-1,1]^{\mathcal{Z}})$. A comment first: in order to have a well-defined scheme, one have to check that $\omega_\cdot$ (the sampling parameters) are well-defined other time.[4] Let $\mathcal{R}(\mathcal{H}) := \bigcup_{M \in \mathcal{H}} \mathbf{r}_M$ the set of mean reward vectors of elements of $\mathcal{H}$. Proposition 2 shows that for all $\mathbf{r}_M$ and $\epsilon > 0$,

$$\mathbb{P}\left\{\mathbf{r}_M + \epsilon \mathbf{u}_t \in \mathcal{R}(\mathcal{H})\right\} = 1.$$

Now, $\mathbf{u}_t$ and $\hat{\mathbf{r}}_t$ are independent, thus for all t,

$$\mathbb{P}\left\{(1 - \tfrac{1}{t})\hat{\mathbf{r}}_t + \tfrac{1}{t}\mathbf{u}_t \in \mathcal{R}(\mathcal{H})\right\} = 1.$$

Thanks to that perturbation, exploration coefficients are almost surely well-defined, since:

$$\mathbb{P}\left\{\exists t, \widetilde{\mathbf{r}}_t \notin \mathcal{R}(\mathcal{H})\right\} \leq \sum\nolimits_t \mathbb{P}\left\{\widetilde{\mathbf{r}}_t \notin \mathcal{R}(\mathcal{H})\right\} = 0.$$

C.1 Time uniform confidence sets

Recall that the confidence sets are of the form

$$\widetilde{\mathcal{M}}_t^\delta := \hat{M}_t + \Lambda_t^\delta$$

where Λ_t^δ is the product of confidence intervals

$$\Lambda_t^\delta := \prod_{z \in \mathcal{Z}} (-\epsilon_\delta(N_t(z),t), \epsilon_\delta(N_t(z),t))$$

[4] The test "$\widetilde{\mathcal{M}}_t^\delta \subseteq \mathcal{B}(\hat{M}_t)$" isn't defined, hence is evaluated to false, when $\hat{M}_t \notin \mathcal{H}$.

where $\epsilon_\delta(N_t(z), t)$ is given as per Equation (9), namely

$$\epsilon_\delta(s,t) := \left(\frac{2\sigma^2 \log(\frac{4mt^3}{\delta})}{s}\right)^{1/2}$$

As claimed in the main body of this paper, this construction provides time uniform confidence sets.

Proposition 4. *For all $\mathcal{I}$, $\sum_t \mathbb{P}^{M,\mathcal{I}}\{M \notin \widetilde{\mathcal{M}}_t^\delta\} \leq \delta$.*

Proof of Proposition 4. Fix $z \in \mathcal{Z}$. Let $V_1, V_2, \ldots$ i.i.d. random variables of distribution $q(z)$. By Hoeffding's Lemma, for all $s, t \geq 1$, we have

$$\mathbb{P}\left\{\left|\frac{1}{s}\sum_{i=1}^{s} V_i - r(z)\right| > \epsilon_\delta(s,t)\right\} \leq 2\exp\left(-\tfrac{s}{2\sigma^2}\tfrac{2\sigma^2\log(4mt^3/\delta)}{s}\right) = \frac{\delta}{2mt^3}.$$

Thus for all $t \geq 1$, by union bound,

$$\begin{aligned}\mathbb{P}^{M,\mathcal{I}}\{|\hat{r}_t(z) - r(z)| > \epsilon_\delta(N_t(z),t)\} &\leq \sum_{s=0}^{t-1} \mathbb{P}^{M,\mathcal{I}}\left\{\begin{array}{c}|\hat{r}_t(z) - r(z)| > \epsilon_\delta(N_t(z),t) \\ \text{and } N_t(z) = s\end{array}\right\} \\ &\leq t \cdot \tfrac{\delta}{2mt^3} = \tfrac{\delta}{2mt^2}.\end{aligned}$$

Finally, we get that $M \in \widetilde{\mathcal{M}}_t^\delta$ uniformly in time in the following way:

$$\mathbb{P}^M\left\{\exists t, M \notin \widetilde{\mathcal{M}}_t^\delta\right\} \leq \sum_t \mathbb{P}^M\left\{M \notin \widetilde{\mathcal{M}}_t^\delta\right\} \leq \frac{\delta}{2}\sum_{t=1}^{\infty}\frac{1}{t^2} < \delta. \qquad \square$$

C.2 The D-tracking rule

The D-tracking scheme heavily relies on the D-tracking rule in the style of Garivier and Kaufmann (2016); Marjani et al. (2021) – just like the name suggested.

Lemma 4 (D-tracking rule)**.** *The D-tracking rule ensures:*

1. *For all $z \in \mathcal{Z}$, $N_t(z) \geq (\sqrt{t} - m/2)_+ - 1$;*
2. *For all $\epsilon > 0$, $\exists t_\epsilon > 0$, s.t. $\forall t_0 > 0$:*

$$\sup_{t \geq t_0} \|\omega_{\widetilde{\mathbf{r}}_t} - \omega_{\mathbf{r}}\|_\infty \leq \epsilon \implies \sup_{t \geq \max(t_0, t_\epsilon)} \left\|\frac{\mathbf{N}(t)}{t} - \omega_{\mathbf{r}}\right\|_\infty \leq 2(m-1)\epsilon$$

About the proof. This result has nothing to do with DMDPs, and is rather a result about "sequences that concentrates" (Garivier and Kaufmann, 2016, p.22). Apply Lemma 17 from Garivier and Kaufmann (2016) with $g(k) := (\sqrt{k} - m/2)_+$. $\square$

The choice to rely on D-tracking rule is arbitrary. Many other tracking methods exist in the literature; That many that would also work here. As these are not the main focus, such discussions are excluded from this paper for the sake of conciseness.

C.3 Proof of Proposition 5: The D-tracking scheme is δ-PC

The δ-probably correctedness is proved using the continuity of $\mathcal{B}(M)$ together with the Π_∞^*-constant property. In more details, $\mathcal{B}(M) \neq \varnothing$ and continuity ensures finite stopping time with high probability, while the Π_∞^*-constant property guarantees that the recommended policy is correct.

Proposition 5 (Correctedness)**.** *Every D-tracking scheme is δ-probably correct on DMDPs in $\mathcal{H}$.*

Proof of Proposition 5. It follows from Lemma 4 that for all z and t, $N_t(z) \geq \sqrt{t} - m$. Therefore, for a fixed $\delta > 0$, the confidence width $\epsilon_\delta(N_t(z), t) \to 0$ a.s. when $t \to \infty$. In particular, for all $\eta > 0$, there exists $t_\eta > 0$ a constant such that for all $t > t_\eta$, $\Lambda_t^\delta \subseteq \prod_z [-\eta, \eta]$. Accordingly, conditioned on the event

$$\mathcal{E} := \left\{ \forall t, M \in \widetilde{\mathcal{M}}_t^\delta \right\},$$

we have

$$\forall \eta > 0, \forall t > t_\eta, \quad d(M, \hat{M}_t) \leq 2\eta.$$

So, by continuity of $\mathcal{B}(\cdot)$, there exists $t_{\mathcal{B}}$ such that on $\mathcal{E}$, for $t > t_{\mathcal{B}}$, $\mathcal{B}(M)$ and $\mathcal{B}(\hat{M}_t)$ look alike in that:

$$\ell_{\mathcal{B}(\hat{M}_t)} \geq \frac{1}{2} \ell_{\mathcal{B}(M)} > 0.$$

Set $\eta := \frac{1}{8} \min_z \ell_{\mathcal{B}(M)}(z) > 0$. Let $t' = \max(t_{\mathcal{B}}, t_\eta)$. On $\mathcal{E}$, for all $t > t'$, we have:

$$d(\hat{M}_t, M) < \frac{1}{2} \ell_{\mathcal{B}(\hat{M}_t)}(z) \tag{$*$}$$

$$\Lambda_t^\delta \subseteq \prod_z [-\eta, \eta] \subseteq \prod_z [-\tfrac{1}{2}\ell_{\mathcal{B}(\hat{M}_t)}(z), \tfrac{1}{2}\ell_{\mathcal{B}(\hat{M}_t)}(z)] \tag{$**$}$$

The combination of $(*)$ and $(**)$ yields that on $\mathcal{E}$, for $t > t'$, we have $\widetilde{\mathcal{M}}_t^\delta \subseteq \mathcal{B}(\hat{M}_t)$. So τ_δ is upper bounded by t' on $\mathcal{E}$; hence finite. By Proposition 4,

$$\mathbb{P}\{\tau_\delta < \infty\} \geq \mathbb{P}\left(\mathcal{E} \cap \{\tau_\delta < \infty\}\right) = \mathbb{P}(\mathcal{E}) \geq 1 - \delta.$$

Finally, we show that under $\{\tau_\delta < \infty\}$ the algorithm returns an optimal policy with probability δ. By construction of the set of plausible MDPs, the probability that for all t, $\widetilde{\mathcal{M}}_t^\delta$ contains M, is at least $1 - \delta$. Also, if $\widetilde{\mathcal{M}}_t^\delta \in \mathcal{B}(\hat{M}_t)$, then for all $\widetilde{M} \in \widetilde{\mathcal{M}}_t^\delta$, we have $\Pi_\infty^*(\widetilde{M}) = \Pi_\infty^*(\hat{M}_t)$. All together,

$$\begin{aligned}
\mathbb{P}\left\{\tau_\delta < \infty, \pi^*(\hat{M}_{\tau_\delta}) \notin \Pi_\infty^*(M)\right\} &\leq \mathbb{P}\left\{\tau_\delta < \infty, M \notin \mathcal{B}(\hat{M}_{\tau_\delta})\right\} \\
&\leq \sum\nolimits_{t \geq 1} \mathbb{P}\left\{\tau_\delta = t, M \notin \widetilde{\mathcal{M}}_t^\delta\right\} \\
&\leq \sum\nolimits_{t \geq 1} \mathbb{P}\left\{M \notin \widetilde{\mathcal{M}}_t^\delta\right\} \leq \delta.
\end{aligned}$$

□

C.4 Proof of Proposition 6: Performances of the D-tracking scheme

We switch the focus to the proof of:

Proposition 6 (Performances)**.** *A D-tracking scheme stops on $M \in \mathcal{H}$ a.s.; Also,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} \geq \sum_z \frac{2\sigma^2}{\ell_{\mathcal{B}(M)}(z)^2}.$$

Moreover, if exploration parameters are $\omega \propto \ell_{\mathcal{B}(\cdot)}^{-2}$, then

$$\lim_{\delta \to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} = \sum_z \frac{2\sigma^2}{\ell_{\mathcal{B}(M)}(z)^2}.$$

The performance of the D-tracking schemes described by Proposition 6 can, in fact, be described in much more precision as given by the result below.

Lemma 5 (Performances of D-tracking schemes)**.** *Let $M \in \mathcal{H}$ with mean reward vector $\mathbf{r}$. Then*

$$\lim_{\delta \to 0} \frac{\mathbb{E}^{M,\mathcal{I}}[\tau_\delta]}{\log(1/\delta)} = \max_z \frac{2\sigma^2}{\omega_{\mathbf{r}}(z) \ell_{\mathcal{B}(\mathbf{r})}(z)^2 L_{\mathbf{r}}(z)^2}.$$

The right-hand term is proved to be an upper bound and a lower bound of the asymptotic expected sampling complexity growth rate separately.

C.4.1 The upper bound

The main steps of the upper-bound part of the proof of Lemma 5 are summarized below.

- Step 1: Properties of the D-tracking rule, see Lemma 4;
- Step 2: Quantify on how much local gaps $L_z(\cdot)$, boxes $\mathcal{B}(\cdot)$ and exploration coefficients $\omega_\cdot$ w.r.t. $\mathbf{r}$, $\hat{\mathbf{r}}_t$, $\widetilde{\mathbf{r}}_t$ all are close to each over in the run;
- Step 3: Relate the actual visit counts $\mathbf{N}_t$ to $t \cdot \omega_{\mathbf{r}}$;
- Step 4: Pseudo-inverse of $\epsilon_\delta(\kappa \cdot t, t)$ in t, in order to quantify how small time needs to be for the confidence set to collapse;
- Step 5: Steps 2-3 define a high probability *good event*, on which τ_δ can be accurately approximated thanks to Step 4.

The conclusion follows quickly from step 5. As step 1 is done already (see Lemma 4), we skip to step 2.

By continuity of $L_z(-)$ at $\mathbf{r} \in \mathcal{R}(\mathcal{H})$, for all $\epsilon > 0$, there exists $\alpha(\epsilon) > 0$ such that for $\|\mathbf{r} - \mathbf{r}'\|_\infty < \alpha(\epsilon)$, we have

$$\max_z |L_{\mathbf{r}'}(z) - L_{\mathbf{r}}(z)| < \epsilon \tag{16}$$

$$\max_z |\omega_{\mathbf{r}'}(z) - \omega_{\mathbf{r}}(z)| < \tfrac{\min_{z' \in \mathcal{Z}} \omega_{z'}(\mathbf{r})}{2(m-1)} \epsilon \tag{17}$$

$$\max_z |\ell_{\mathcal{B}(\mathbf{r}')}(z) - \ell_{\mathcal{B}(\mathbf{r})}(z)| < \epsilon. \tag{18}$$

The set of such $\mathbf{r}'$ is denoted $\mathcal{B}(\mathbf{r}, \alpha(\epsilon))$. Introduce

$$\mathcal{F}_\epsilon(T) := \bigcap_{t=T^{1/4}}^{T} \left\{ \hat{\mathbf{r}}_t, \widetilde{\mathbf{r}}_t \in \mathcal{B}(\mathbf{r}, \alpha(\epsilon)) \right\}.$$

We show that this event hold with high probability.

Lemma 6 (Step 2)**.** *There exist constants $B_\epsilon, C_\epsilon > 0$ such that*

$$\forall T > 0, \qquad \mathbb{P}^{M,\mathcal{I}} \left\{ \mathcal{F}_\epsilon(T)^{\complement} \right\} \leq B_\epsilon T \exp(-C_\epsilon T^{1/8}).$$

Proof of Lemma 6. By definition of $\widetilde{r}_t(z)$, for all $z \in \mathcal{Z}$, $\widetilde{r}_t(z) - r(z)$ unfolds as

$$\left(1 - \tfrac{1}{t}\right) \left[\hat{r}_t(z) - r(z)\right] + \tfrac{1}{t} \left[\mathbf{u}_t(z) - r(z)\right].$$

There exists $A_\epsilon > 0$ such that if $t \geq A_\epsilon$ then for all $z \in \mathcal{Z}$ and $z \in [0,1]$, $|\frac{1}{t}(\mathbf{u}_t(z) - r(z))| \leq \frac{1}{2}\alpha(\epsilon)$. In addition, we know that $t \geq N_t(z) \geq \sqrt{t} - m$, so for $t \geq A_\epsilon$ and $z \in \mathcal{Z}$,

$$\begin{aligned}
\mathbb{P}\left\{|\hat{r}_t(z) - r(z)| > \alpha(\epsilon) \text{ and } |\widetilde{r}_t(z) - r(z)| > \alpha(\epsilon)\right\} &\leq \mathbb{P}\left\{|\hat{r}_t(z) - r(z)| > \tfrac{1}{2}\alpha(\epsilon), N_z(t) \geq \sqrt{t} - m\right\} \\
&\leq 2 \sum_{s=\sqrt{t}-m}^{t} \exp\left(-\frac{s\alpha(\epsilon)^2}{8\sigma^2}\right) \\
&\leq \frac{2\exp\left(-\alpha(\epsilon)^2(\sqrt{t} - m)/8\sigma^2\right)}{1 - \exp(-\alpha(\epsilon)^2/8\sigma^2)}.
\end{aligned}$$

So by union bound and Hoeffding's Lemma,

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{F}_\epsilon(T)^{\complement}\right\} &= \mathbb{P}\left\{\bigcup_{z \in \mathcal{Z}} \bigcup_{t \in [T^{1/4}, T]} |\hat{r}_t(z) - r(z)| > \alpha(\epsilon)\right\} \\
&\leq m\left(|T^{1/4} - A_\epsilon|_+ + 2\sum_{t=T^{1/4}}^{T} \frac{e^{-\alpha(\epsilon)^2(\sqrt{t}-m)/8\sigma^2}}{1 - e^{-\alpha(\epsilon)^2/8\sigma^2}}\right) \\
&\leq m|T^{1/4} - A_\epsilon|_+ \frac{2me^{m\alpha(\epsilon)^2/8\sigma^2}}{1 - e^{-\alpha(\epsilon)^2/8\sigma^2}} \cdot T \cdot e^{-\alpha(\epsilon)^2 T^{1/8}/8\sigma^2}.
\end{aligned}$$

Last but not least, $|T^{1/4} - A_\epsilon|_+ = 0$ for $T \geq A_\epsilon^4$, so $m|T^{1/4} - A_\epsilon|$ is negligible in comparison to the right term when T goes to infinity. □

On $\mathcal{F}_\epsilon(T)$, the actual ratio of visits is closed to $\omega_{\mathbf{r}}$ up to $\epsilon \min_z \omega_{\mathbf{r}}(z)$ when T is high enough. It puts the informal statement $\mathbf{N}_t = t\omega_{\mathbf{r}} + o(t)$ in a formal way.

Lemma 7 (Step 3)**.** *For all $\epsilon > 0$, there exists $T_\epsilon > 0$ such that for all $T \geq 0$, on $\mathcal{F}_\epsilon(T)$,*

$$\forall t \in [\max(T^{1/4}, T_\epsilon), T], \quad \left\| \frac{\mathbf{N}_t}{t} - \omega_{\mathbf{r}} \right\|_\infty \leq \epsilon \min_{z \in \mathcal{Z}} \omega_{\mathbf{r}}(z).$$

Proof. On $\mathcal{F}_\epsilon(T)$, for all $t \geq T^{1/4}$, we have $\|\omega_{\widetilde{\mathbf{r}}_t} - \omega_{\mathbf{r}}\|_\infty \leq \epsilon \min \omega_{\mathbf{r}}(x,a)/2(m-1)$ by construction, see Equation (17). So, from Lemma 4 (2), for $t \geq \max(T^{1/4}, t_\epsilon)$,

$$\left\| \frac{N(t)}{t} - \omega_{\mathbf{r}} \right\|_\infty \leq \epsilon \min_{z \in \mathcal{Z}} \omega_{\mathbf{r}}(z).$$

Set $T_\epsilon := t_\epsilon$. □

The next result provides a way to *invert* $\epsilon_\delta(\cdot)$. This will give information on how fast the confidence set shrink. Recall that $\epsilon_\delta(s,t) := (\frac{2\sigma^2}{s} \log(\frac{4mt^3}{\delta}))^{1/2}$ is the precision that the algorithm assumes to have on a reward with s samples at time t.

Lemma 8 (Step 4)**.** *Let $\alpha, \kappa > 0$. For all $\beta < \frac{1}{2}$, there exists a constant $D_{\alpha,\beta,\kappa} > 0$, independent of δ, such that if*

$$t \geq \frac{2\sigma^2 \log(1/\delta)}{\kappa(1-2\beta)\alpha^2} + D_{\alpha,\beta,\kappa} \tag{19}$$

then for all $s \geq \kappa \cdot t$, $\epsilon_\delta(s,t) \leq \alpha$.

Proof of Lemma 8. Assume that $\kappa \cdot t \leq s$. We are searching for a sufficient condition on t such that $\epsilon_\delta(s,t) \leq \alpha$. Observe that it boils down to

$$\log(1/\delta) + \log(4m) + 3\log t \leq \frac{\kappa\alpha^2}{2\sigma^2} \cdot t$$

Using the global upper bound $\log t \leq \sqrt{t}$, we get the three sufficient conditions

$$\begin{cases} \log(1/\delta) & \leq (1-2\beta)\frac{\kappa\alpha^2}{2\sigma^2} \cdot t \\ \log(4m) & \leq \beta\frac{\kappa\alpha^2}{2\sigma^2} \cdot t \\ 3\sqrt{t} & \leq \beta\frac{\kappa\alpha^2}{2\sigma^2} \cdot t. \end{cases}$$

Solving in t and using $\max\{a,b,c\} \leq a+b+c$ for $a,b,c \geq 0$, the condition above reduces to

$$\frac{2\sigma^2 \log(1/\delta)}{\kappa(1-2\beta)\alpha^2} + \frac{2\sigma^2 \log(4m)}{\beta\kappa\alpha^2} + \frac{36\sigma^4}{\beta^2\kappa^2\alpha^4} \leq t$$

where we identify $D_{\alpha,\beta,\kappa} := (\beta\kappa\alpha^2)^{-1} 2\sigma^2 \log(4m) + 36\sigma^4(\beta^2\kappa^2\alpha^4)^{-1}$. □

The last lemma basically bounds τ_δ on $\mathcal{F}_\epsilon(T)$ provided by T is large enough. It is the last major step of the proof.

Lemma 9 (Step 5)**.** *For all $\epsilon > 0$, there exist constants $H_\epsilon, D_\epsilon \geq 0$ such that denoting*

$$T_0(\epsilon,\delta) := D_\epsilon + \max_{z \in \mathcal{Z}} \frac{2\sigma^2 \log(1/\delta)}{(1-\epsilon)^2 \omega_{\mathbf{r}}(z)(\ell_{\mathcal{B}(\mathbf{r})}(z) - \epsilon)^2 (L_{\mathbf{r}}(z) - \epsilon)^2},$$

then for all $T \geq \max(H_\epsilon, (1+\epsilon)T_0(\delta,\epsilon))$, then on $\mathcal{F}_\epsilon(T)$:

$$\tau_\delta \leq T^{1/4} + T_0(\epsilon,\delta) \leq T$$

Proof of Lemma 9. By definition of the stopping time, on $\mathcal{F}_\epsilon(T)$, if for all $z \in \mathcal{Z}$, we have

$$\epsilon_\delta(N_{x,a}(t), t) \leq (\ell_{\mathcal{B}(\mathbf{r})}(z) - \epsilon)(L_{\mathbf{r}}(z) - \epsilon),$$

then $\epsilon_\delta(N_t(z), t) \leq \ell_{\hat{\mathbf{r}}_t}(z) L_z(\hat{\mathbf{r}}_t)$, so the scheme stops. Introduce the following random time:

$$T_{\text{upper}}(\delta, T) := \inf\left\{t \in [\max(T^{1/4}, T_\epsilon), T] : \forall z \in \mathcal{Z}, \epsilon_\delta(N_t(z), t) \leq (\ell_{\mathcal{B}(\mathbf{r})}(z) - \epsilon)(L_{\mathbf{r}}(z) - \epsilon)\right\}$$

with the convention that $T_{\text{upper}}(\delta, T) = +\infty$ if the infimum goes over an empty set. From the previous discussion follows $\mathcal{F}_\epsilon(T) \subseteq \{\tau_\delta \leq T_{\text{upper}}(\delta, T)\}$. Namely, if $\mathcal{F}_\epsilon(T)$ holds then $T_{\text{upper}}(\delta, T)$ is an upper bound of τ_δ. From Lemma 7, on $\mathcal{F}_\epsilon(T)$ and for $t \in [\max(T^{1/4}, T_\epsilon), T]$, we have

$$\forall z \in \mathcal{Z}, \quad \frac{N_t(z)}{t} - \omega_{\mathbf{r}}(z) \geq -\omega_{\mathbf{r}}(z)\epsilon$$

Thus $N_t(z) \geq (1-\epsilon)\omega_{\mathbf{r}}(z)t$. We provide an upper bound of $T_{\text{upper}}(\delta, T)$ by applying Lemma 8 with $\kappa_z := (1-\epsilon)\omega_{\mathbf{r}}(z)$, $\alpha_z := (\ell_{\mathcal{B}(\mathbf{r})}(z) - \epsilon)(L_{\mathbf{r}}(z) - \epsilon)$ and $\beta = \epsilon/2$. Define

$$T_0(\epsilon, \delta) := T_\epsilon + \max_{z \in \mathcal{Z}} \left(\frac{2\sigma^2(1-\epsilon)^{-2}\log(1/\delta)}{\omega_{\mathbf{r}}(z)(\ell_{\mathcal{B}(\mathbf{r})}(z) - \epsilon)^2 (L_{\mathbf{r}}(z) - \epsilon)^2} + D_{\alpha_z, \frac{\epsilon}{2}, \kappa_z} \right).$$

When $T \geq T^{1/4} + T_0(\epsilon, \delta)$, we get $T_{\text{upper}}(\delta, T) \leq T_0(\epsilon, \delta) + T^{1/4}$ on $\mathcal{F}_\epsilon(T)$.

To conclude the proof, observe that the condition

$$\text{“}T \geq T^{1/4} + T_0(\epsilon, \delta)''$$

is satisfied as soon as $T \geq \max((1 + \frac{1}{\epsilon})^{4/3}, (1+\epsilon)T_0(\epsilon, \delta))$. Indeed, if this is the case, then $T(1 - \frac{1}{1+\epsilon}) \geq T^{1/4}$, so

$$T_0(\epsilon, \delta) + T^{1/4} \leq \tfrac{1}{1+\epsilon} T + T^{1/4} \leq T(\tfrac{1}{1+\epsilon} + 1 - \tfrac{1}{1+\epsilon}) = T.$$

Set $D_\epsilon := T_\epsilon + \max_{z \in \mathcal{Z}} D_{\alpha_z, \epsilon/2, \kappa_z}$ and $H_\epsilon := (1 + \frac{1}{\epsilon})^{4/3}$. This proves Lemma 9. □

The final step. Following Lemma 9 (step 5), if $T \geq \max(H_\epsilon, (1+\epsilon)T_0(\delta, \epsilon))$, then $\mathcal{F}_\epsilon(T) \subseteq \{\tau_\delta \leq T\}$. So,

$$\mathbb{P}^{M,\mathcal{I}}\{\tau_\delta > T\} \leq \mathbb{P}^{M,\mathcal{I}}\left\{\mathcal{F}_\epsilon(T)^{\complement}\right\} \leq B_\epsilon T \exp(-C_\epsilon T^{1/8}).$$

Then, we compute the expected sampling complexity has:

$$\begin{aligned}
\mathbb{E}^{M,\mathcal{I}}[\tau_\delta] &= \sum_{T=1}^{\infty} \mathbb{P}^{M,\mathcal{I}}\{\tau_\delta \geq T\} \\
&\leq \left[\sum_{T=1}^{H_\epsilon + (1+\epsilon)T_0(\epsilon,\delta)} \mathbb{P}^{M,\mathcal{I}}\{\tau_\delta \geq T, \mathcal{F}_\epsilon(T)\}\right] + \left[\sum_{T=1}^{\infty} \mathbb{P}^{M,\mathcal{I}}\left\{\mathcal{F}_\epsilon(t)^{\complement}\right\}\right] \\
&\leq H_\epsilon + (1+\epsilon)T_0(\epsilon, \delta) + \sum_{T=1}^{\infty} B_\epsilon T \exp(-C_\epsilon T^{1/8}).
\end{aligned}$$

Divide by $\log(1/\delta)$ and take the supremum limit. Expanding $T_0(\epsilon, \delta)$, we obtain:

$$\limsup_{\delta \to 0} \frac{\mathbb{E}^{M,\mathcal{I}}[\tau_\delta]}{\log(1/\delta)} \leq \max_{z \in \mathcal{Z}} \frac{(1+\epsilon)2\sigma^2}{(1-\epsilon)^2 \omega_{\mathbf{r}}(z)(\ell_{\mathcal{B}(\mathbf{r})}(z) - \epsilon)^2 (L_{\mathbf{r}}(z) - \epsilon)^2}.$$

This holds for all $\epsilon > 0$ sufficiently small, so letting ϵ goes to 0 yields

$$\limsup_{\delta \to 0} \frac{\mathbb{E}^{M,\mathcal{I}}[\tau_\delta]}{\log(1/\delta)} \leq \max_{z \in \mathcal{Z}} \frac{2\sigma^2}{\omega_{\mathbf{r}}(z)\ell_{\mathcal{B}(\mathbf{r})}(z)^2 L_{\mathbf{r}}(z)^2}. \tag{20}$$

Finally, as $\mathrm{kl}(\delta, 1-\delta) \sim \log(1/\delta)$ when $\delta \to 0$, $\log(1/\delta)$ can be changed to $\mathrm{kl}(\delta, 1-\delta)$ in Equation (20).

C.4.2 The lower bound

Equation (20) is seemingly an upper bound. This is in fact an equality. The proof of the reverse equality is mainly similar and in particular, it shares the technical aspect of the upper bound's proof. Details are spared; we focus on the main idea. We rely an additional crucial property: specifically that when the targeted confidence is high, the *probability* that the algorithm stops early is small. More precisely:

Lemma 10. *Consider a D-tracking scheme $\mathcal{I}$ and DMDP $M \in \mathcal{H}$. For all $T, \eta > 0$, there exists $\delta_{T,\eta} > 0$ such that if $\delta < \delta_{T,\eta}$, then*

$$\mathbb{P}^{M,\mathcal{I}}\{\tau_\delta < T\} < \eta.$$

Proof. Let $T, \eta > 0$.

Denote

$$\epsilon_t := \epsilon_\delta(\sqrt{t} - m, t)$$

For $\delta < \eta$, the event

$$\mathcal{E} := \{\forall z, \forall t \geq m,\ \epsilon_\delta(N_t(z), t) < \epsilon_t\}$$

is almost sure by Lemma 4. Let $\overline{\mathcal{U}}(M, \epsilon_t)$ the closed ball of $\mathcal{M}$ centered at M of radius ϵ_t. By Hoeffding's Lemma (refer to the proof of Proposition 4), if $\delta < \eta$, we get

$$\begin{aligned}\mathbb{P}\left\{\forall t, \hat{M}_t \in \overline{\mathcal{U}}(M, \epsilon_t)\right\} &\geq 1 - \mathbb{P}\left\{\exists t, \hat{M}_t \notin \overline{\mathcal{U}}(M, \epsilon_t)\right\} \\ &\geq 1 - \mathbb{P}\left\{\exists t, \exists z, |\hat{\mathbf{r}}_t(z) - \mathbf{r}(z)| > \epsilon_t\right\} \\ &\geq 1 - \delta \geq 1 - \eta.\end{aligned}$$

The sequence of sets given by $\overline{\mathcal{U}}(M, \epsilon_t)$ is compact and non-increasing. The quantity

$$\ell_\eta := \sup_{M' \in \overline{\mathcal{U}}(M, \epsilon_t)} \sup_{z \in \mathcal{Z}} \ell_{\mathcal{B}(M')}(z)$$

is, by continuity of the family and compacity of $\overline{\mathcal{U}}(M, \epsilon_1)$, finite. Let $\delta_{\eta,T} \in (0, \eta)$ such that for all $\delta < \delta_{\eta,T}$,

$$\epsilon_\delta(T, 1) > \ell_\eta.$$

On the event $\mathcal{E}' := \{\forall t, \hat{M}_t \in \bar{\mathcal{U}}(M, \epsilon_t)\}$ and for $\delta < \delta_{\eta,T}$, we are guaranteed that for $t < T$, $\widetilde{\mathcal{M}}_t^\delta \not\subseteq \mathcal{B}(\hat{M}_t)$; Indeed, for $t < T$,

$$\begin{aligned}\mathbb{P}\left\{\widetilde{\mathcal{M}}_t^\delta \subseteq \mathcal{B}(\hat{M}_t) \mid \mathcal{E}'\right\} &\leq \mathbb{P}\left\{\exists z, \epsilon_\delta(N_t(z), t) \leq \ell_{\mathcal{B}(\hat{M}_t)}(z) \mid \mathcal{E}'\right\} \\ &\leq \mathbb{P}\{\exists z, \epsilon_\delta(N_t(z), t) \leq \ell_\eta\} \\ &\leq \mathbb{P}\{\exists z, \epsilon_\delta(T, 1) \leq \ell_\eta\} = 0.\end{aligned}$$

Accordingly, $\{\tau_\delta > T\} \supseteq \mathcal{E}'$ for $\delta < \delta_{\eta,T}$. So, for such δ,

$$\mathbb{P}\{\tau_\delta > T\} \geq \mathbb{P}(\mathcal{E}') \geq 1 - \eta. \qquad \square$$

With that in mind, the general idea to prove the lower bound

$$\liminf_{\delta \to 0} \frac{\mathbb{E}^{M,\mathcal{I}}[\tau_\delta]}{\log(1/\delta)} \geq \max_{z \in \mathcal{Z}} \frac{2\sigma^2}{\omega_{\mathbf{r}}(z)\ell_{\mathcal{B}(\mathbf{r})}(z)^2 L_{\mathbf{r}}(z)^2} \tag{21}$$

is the following: choose T such that once $t > T$, $\mathcal{B}(\hat{M}_t)$ and $\mathcal{B}(M)$ are very similar (in the same fashion as $\mathcal{F}_\epsilon(T)$ in the proof of the upper bound). Choose $\eta > 0$. Then with confidence parameter $\delta < \delta_{T,\eta}$, the event $\mathcal{E}' := \{\tau_\delta > T\}$ holds with probability at least $1 - \eta$. On that event, with a similar analysis than the one of the upper bound, show that most of the expected sample complexity is *waiting* for Λ_t^δ to shrink, so that on $\mathcal{E}'$,

$$\tau_\delta < T' \implies \widetilde{\mathcal{M}}_t^\delta \subseteq \mathcal{B}'$$

where $\mathcal{B}'$ is a box that very much resembles $\mathcal{B}(M)$. Specifically, with the same computations as in Appendix C.4, we find that under $\{\tau_\delta > T\}$, τ_δ is of order

$$\max_{z\in\mathcal{Z}} \frac{2\sigma^2}{\omega_{\mathbf{r}}(z)\ell_{\mathcal{B}(\mathbf{r})}(z)^2 L_{\mathbf{r}}(z)^2} \cdot \log(1/\delta).$$

Since for $\delta < \delta_{T,\eta}$, $\{\tau_\delta > T\}$ holds with probability at least $1-\eta$, we get:

$$\liminf_{\delta\to 0} \frac{\mathbb{E}^{M,\mathcal{I}}[\tau_\delta]}{\log(1/\delta)} \geq (1-\eta) \max_{z\in\mathcal{Z}} \frac{2\sigma^2}{\omega_{\mathbf{r}}(z)\ell_{\mathcal{B}(\mathbf{r})}(z)^2 L_{\mathbf{r}}(z)^2}.$$

Make η go to 0 and conclude that

$$\liminf_{\delta\to 0} \frac{\mathbb{E}^{M,\mathcal{I}}[\tau_\delta]}{\log(1/\delta)} \geq \max_{z\in\mathcal{Z}} \frac{2\sigma^2}{\omega_{\mathbf{r}}(z)\ell_{\mathcal{B}(\mathbf{r})}(z)^2 L_{\mathbf{r}}(z)^2}. \tag{22}$$

C.5 Proof of Proposition 6: Optimal exploration coefficients

On our way to prove Proposition 6, what is left is to compute the optimal ω for a given family $\mathcal{B}(\cdot)$. Fix $\mathbf{r} \in \mathbb{R}^m$. We wish to solve the following minimization problem:

$$\min_{\omega} \max_{z\in\mathcal{Z}} \frac{1}{\omega_z \alpha_z} \quad \text{s.t.} \quad \sum_z \omega_z = 1.$$

where $\alpha_z := \ell_{\mathcal{B}(\mathbf{r})}(z)^2 L_{\mathbf{r}}(z)^2 > 0$. This problem is equivalent to

$$\min_{\omega,t} t \quad \text{s.t.} \quad \begin{cases} \sum_z \omega_z - 1 = 0 \\ \forall z, \quad \frac{1}{\omega_z\alpha_z} - t \leq 0 \end{cases}$$

The Lagrangian of this convex optimization problem is:

$$\mathcal{L}(\omega, t; \mu, \lambda) := t + \sum_z \mu_z \left(\frac{1}{\omega_z\alpha_z} - t\right) + \lambda\left(\sum_z \omega_z - 1\right).$$

KKT-conditions provide:

$$\partial_t : \quad \sum_z \mu_z = 1 \tag{$*$}$$

$$\partial_{\omega_z} : \quad \lambda = \frac{\mu_z}{\omega_z^2 \alpha_z} \tag{$**$}$$

$$\text{CS} : \quad \mu_z \left(\frac{1}{\omega_z\alpha_z} - t\right) = 0 \tag{$***$}$$

$$\&\text{c}$$

Clearly, $\mu_z \neq 0$ for all z (or $\lambda = 0$ and we quickly derive a contradiction), so the complementary slackness condition $(***)$ implies that $t = \omega_z\alpha_z$. Injecting that into $(**)$, we get that μ_z and ω_z are proportional, hence equal, as they both belong to the simplex of $\mathbb{R}^m$. Hence $(**)$ produces again:

$$\lambda = \frac{1}{\omega_z\alpha_z}$$

so $\omega_z \propto \alpha_z^{-1}$. Specifically, the optimal $\omega \in \Delta^m$ satisfies $\omega_z \propto \ell_{\mathcal{B}(\mathbf{r})}(z)^{-2} L_{\mathbf{r}}(z)^{-2}$. Write this element $\omega_{\mathbf{r}}^{\mathcal{B}*}$. The D-tracking scheme with exploration coefficients $\omega_{\mathbf{r}}^{\mathcal{B}*}$ has asymptotic expected sample complexity:

$$\lim_{\delta\to 0} \frac{\mathbb{E}^{M,\mathcal{I}}}{\log(1/\delta)} = \sum_z \frac{2\sigma^2}{\ell_{\mathcal{B}(M)}(z)^2 L_M(z)^2}.$$

This ends the proof of the Proposition 6.

C.6 Properties of local suboptimality gaps

This section is dedicated to a review of the most important properties of local suboptimality gaps $L_M(z)$ (see Definition 5). We start with the intimate relationship between local suboptimality gaps and Bellman gaps that motivates the denomination "generalized Bellman coefficients".

Lemma 11. *Let $M \in \mathcal{H}$ with mean reward vector $\mathbf{r}$. Let $\pi \in \Pi_0^*(M)$ and $\mathcal{C}$ its unique terminal cycle seen as a set of transitions $\{(x,a)\}$ with x recurrent and $a = \pi(x)$. Denote x_t the state visited at time $t \geq 0$ under the iterations of π starting from $x = x_0 \in \mathcal{S}$. For $x, y \in \mathcal{S}$, introduce the reaching time of y from x:*

$$\tau_x(y) := \inf\{t \geq 0 \mid x_t = y\} \in \mathbb{N} \cup \{\infty\}$$

and write $x \rightsquigarrow^* y$ if $\tau_x(y) < \infty$. With the convention that $\Delta_{x,a}(M)/0 = +\infty$, for all $z = (x,a) \in \mathcal{Z}$:

(*i*) *if $z \notin \pi$, then $L_M(z) = \Delta_M(z)$;*

(*ii*) *if $z \in \pi \setminus \mathcal{C}$, then*

$$L_M(z) = \min_{(y,b)\notin\pi} \frac{\Delta_M(y,b)}{\left|\mathbf{1}_{s(y,b)\rightsquigarrow^* x} - \mathbf{1}_{y\rightsquigarrow^* x}\right|};$$

(*iii*) *if $z \in \mathcal{C}$, then*

$$L_M(z) = \min_{(y,b)\notin\pi} \frac{|\mathcal{C}|\Delta_M(y,b)}{\left|\tau_y(x) - \tau_{s(y,b)}(x) - 1\right|}.$$

Proof. So π is the only optimal policy of M. From Theorem 3, M' has the same (unique) bias optimal policy if, and only if

$$\forall (y,b) \notin \pi, \quad \mathbf{r}'(y,b) - g_y^\pi(\mathbf{r}') + h_{s(y,b)}^\pi(\mathbf{r}') < h_y^\pi(\mathbf{r}').$$

Writing $\mathbf{r}' = \mathbf{r} + \mathrm{d}\mathbf{r}$ and using linearity, the condition above translates to: for all $(y,b) \notin \pi$,

$$\mathrm{d}\mathbf{r}(y,b) - g_y^\pi(\mathrm{d}\mathbf{r}) + h_{s(y,b)}^\pi(\mathrm{d}\mathbf{r}) - h_y^\pi(\mathrm{d}\mathbf{r}) < \Delta_M(y,b). \tag{23}$$

Let us assume that $\mathbf{r}'$ is a unilateral deviation of $\mathbf{r}$ at $(x,a) \in \mathcal{Z}$, that is to say $\mathrm{d}\mathbf{r} = \alpha \mathbf{e}_{x,a}$. To express $L_M(x,a)$, we expand the variational quantities of (23) in the light of Lemma 1 and Lemma 2.

- *Case $(x,a) \notin \pi^*$.* Then, both the gain and the bias are unchanged. The only non-trivial condition from (23) is given for $(y,b) = (x,a)$. Subsequently,

  $$\Pi_0^*(\mathbf{r} + \mathrm{d}\mathbf{r}) = \Pi_0^*(\mathbf{r}) \iff \alpha < \Delta_M(x,a)$$

  Accordingly, $L_M(x,a) = \Delta_M(x,a)$.

- *Case $(x,a) \in \pi \setminus \mathcal{C}$.* For the same reason, $g_u^\pi(\mathbf{r}') = g_u^\pi(\mathbf{r})$ for all $u \in \mathcal{S}$. For the bias, from Lemma 2, we have

  $$h_u^\pi(\mathrm{d}\mathbf{r}) = \mathbf{1}_{u\rightsquigarrow^* x}\alpha.$$

  From (23) follows that $\Pi_0^*(\mathbf{r} + \mathrm{d}\mathbf{r}) = \Pi_0^*(\mathbf{r})$ iff for all $(y,b) \notin \pi$,

  $$\alpha\left(\mathbf{1}_{(x,a)=(y,b)} + \mathbf{1}_{s(y,b)\rightsquigarrow^* x} - \mathbf{1}_{y\rightsquigarrow^* x}\right) < \Delta_M(y,b).$$

  As $(x,a) \in \pi$, we have $\mathbf{1}_{(y,b)=(x,a)} = 0$. With the convention $\Delta_{y,b}/0 = +\infty$, we get

  $$L_M(z) = \min_{(y,b)\notin\pi} \frac{\Delta_M(y,b)}{\left|\mathbf{1}_{s(y,b)\rightsquigarrow^* x} - \mathbf{1}_{y\rightsquigarrow^* x}\right|}.$$

- *Case $(u,v) \in \mathcal{C}$.* This time, both the gain and the bias will be subject to variations. By Lemma 1, for all $u \in \mathcal{S}$,

  $$g_u^\pi(\mathrm{d}\mathbf{r}) = \alpha|\mathcal{C}|^{-1}.$$

For the bias, Lemma 2 provides

$$h_u^\pi(\mathrm{d}\mathbf{r}) = \alpha\left(1 - \tfrac{|\mathcal{C}|+1}{2|\mathcal{C}|} - \tfrac{\tau_u(x)}{|\mathcal{C}|}\right)$$

Injecting those expressions into (23), we get that $\Pi_0^*(\mathbf{r}+\mathrm{d}\mathbf{r}) = \Pi_0^*(\mathbf{r})$ if, and only if for all $(y,b) \notin \pi$,

$$\alpha\left(\mathbf{1}_{(y,b)=(x,a)} + \tfrac{\tau_y(x) - \tau_{s(y,b)}(x) - 1}{|\mathcal{C}|}\right) < \Delta_M(y,b).$$

Again, $(x,a) \in \pi$ so $\mathbf{1}_{(y,b)=(x,a)} = 0$. Therefore,

$$L_M(z) = \min_{(y,b)\notin\pi} \frac{|\mathcal{C}|\Delta_M(y,b)}{\left|\tau_y(x) - \tau_{s(y,b)}(x) - 1\right|}. \qquad \square$$

This settles the correspondence between $L_M(x,a)$ and $\Delta_M(x,a)$ on $\mathcal{H}$. These quantities are also used to design Π_∞-constant familys, see Proposition 9 that we restate and prove below.

Proposition 9. *Let $\rho : \mathcal{H} \to \mathbb{R}_+^m$ a continuous function. The family $\mathcal{B}_\rho : \mathcal{H} \to \Gamma$ given by*

$$M + \prod\nolimits_{z\in\mathcal{Z}} \left(-\rho_M(z)L_M(z), \rho_M(z)L_M(z)\right)$$

is a continuous family. If in addition, a) $\rho_M(z) > 0$ for all $z \in \mathcal{Z}$; and b) $\forall M$, $\forall \pi \in \Pi$, $\sum_{z\in\pi\cup\pi_M^*} \rho_M(z) \leq 1$, then $\mathcal{B}_\rho$ is Π_∞^*-constant and is never empty.

Proof. We show the equivalent following statement: *Assume that $\rho : \mathcal{H} \to \mathbb{R}_+^m$ satisfies the assumptions a) and b) of Proposition 9. Let $M \in \mathcal{H}$ with mean reward vector $\mathbf{r}$. Let $M' \in \mathcal{M}$ with mean reward vector $\mathbf{r}'$ such that:*

$$\forall z \in \mathcal{Z}, \quad |r'(z) - r(z)| < \rho_M(z)L_M(z).$$

Then $M' \in \mathcal{H}$ and $\Pi_\infty^*(M) = \Pi_\infty^*(M')$.

For $\pi \in \Pi$ and $x \in \mathcal{S}$, the gain $g_\pi(x)$ and the bias $h_\pi(x)$ are linear functions of the mean reward vector $\mathbf{r}$ that we shall write $g_x^\pi(\mathbf{r})$ and $h_x^\pi(\mathbf{r})$ respectively. Let $\pi^* \in \Pi_0^*(\mathbf{r})$ and $\pi \neq \pi^*$. We show that for all $x \in \mathcal{S}$, $g_x^{\pi^*}(\mathbf{r}') \geq g_x^\pi(\mathbf{r}')$; and, if there is equality for all x in the preceding equations, that $h_x^{\pi^*}(\mathbf{r}') > h_x^\pi(\mathbf{r}')$ for at least one x. Then $\pi^* \in \Pi_0^*(\mathbf{r}')$ will follow by definition of bias optimality and by cardinality, $\Pi_0^*(\mathbf{r}') = \{\pi^*\}$ will hold. Accordingly, by Proposition 3, we will get $M' \in \mathcal{H}$ with $\Pi_\infty^*(M') = \Pi_\infty^*(M)$.

Let $x \in \mathcal{S}$. Denote $\mathrm{d}\mathbf{r} = \mathbf{r}' - \mathbf{r}$. We know that for all $z \in \mathcal{Z}$, $|\,\mathrm{d}\mathbf{r}(z)| < \rho_M(z)L_M(z)$, so there exists $\alpha(z) \in (-1,1)$ such that

$$\mathrm{d}\mathbf{r}(z) = \alpha(z)\rho_M(z)L_M(z).$$

We have $\mathrm{d}\mathbf{r} = \sum_{z\in\mathcal{Z}} \mathrm{d}\mathbf{r}(z)\mathbf{e}_z$. Let $C \neq C^*$ a cycle which is different from the terminal cycle of the optimal policy. Because M satisfies H1, $g_C(\mathbf{r}) < g_{C_*}(\mathbf{r})$. Both g_C and g_{C_*} are linear functions of $\mathbf{r}$, so

$$g_{C_*}(\mathbf{r}') - g_C(\mathbf{r}') = [g_{C_*}(\mathbf{r}) - g_C(\mathbf{r})] + [g_{C_*}(\mathrm{d}\mathbf{r}) - g_C(\mathrm{d}\mathbf{r})]$$

The quantities $g_{C_*}(\cdot)$ and $g_C(\cdot)$ only depend on the coordinates $z \in C \cup C_*$. Because $g_{C_*}(\mathbf{r}) - g_C(\mathbf{r}) \geq 0$ and by assumption *b*), $\sum_{z\in\pi\cup\pi^*} \rho_M(z) \leq 1$, one can factorize the following way (L_z and ρ_z are shorthands for $L_M(z)$ and $\rho_M(z)$):

$$\begin{aligned}
&g_{C_*}(\mathbf{r}') - g_C(\mathbf{r}') \\
&\overset{b)}{\geq} \sum_{z\in C\cup C_*} \rho_z \left(\left[g_{C_*}(\mathbf{r}) - g_C(\mathbf{r})\right] + \left[g_{C_*}(\alpha_z L_z \mathbf{e}_z) - g_C(\alpha_z L_z \mathbf{e}_z)\right]\right) \\
&= \sum_{z\in C\cup C_*} \rho_z \left(g_{C_*}(\mathbf{r}_z + \alpha_z L_z \mathbf{e}_z) - g_C(\mathbf{r}_z + \alpha_z L_z \mathbf{e}_z)\right).
\end{aligned}$$

As $|\alpha_z L_z| < L_z$, by definition of $L_z = L_M(z)$, all terms in the sum above are non-negative – in fact, these are positive. It means that $g_{C_*}(\mathbf{r}') > g_C(\mathbf{r}')$. Accordingly, a policy π that doesn't have terminal cycle C_* is not gain-optimal for M'. Therefore, unless $\pi \in \Pi_{-1}^*(M)$, $\pi \notin \Pi_\infty^*(M')*$.

So, what if $\pi \in \Pi^*_{-1}(M)$? It then has the same terminal cycle than π^*, and we can do the same computation for $h_x^{\pi^*}(\mathbf{r}') - h_x^{\pi}(\mathbf{r}')$, and find that it has to be positive again unless the iterate of π and π^* coincide from x. But for at least one x, they do not, since $\pi \neq \pi^*$.

In the end, π^* appears to be the only bias optimal policy of M' and its terminal cycle is unique. From Proposition 3 follows that $M' \in \mathcal{H}$. Overall, $M' \in \mathcal{H}$ and $\Pi^*_\infty(M) = \Pi^*_\infty(M')$. □

C.7 Details for LSTS-imp

This section provides details for the proof of Proposition 10, Proposition 11 and Theorem 3. Recall that LSTS-imp starts by improving on the ρ that are solution of the following optimization problem:

$$\min_{\rho} \sum_{z \in \mathcal{Z}} \frac{1}{\rho_M(z)^2 L_M(z)^2} \quad \text{s.t.} \quad \sum_{z \in \mathcal{Z}} \rho_M(z) \leq 1. \tag{11}$$

Proposition 10. *The solution of Eq.* (11) *is* $\rho_M \propto L^{-2/3}$.

Proof. The DMDP M is fixed, so write ρ_z and L_z instead of $\rho_M(z)$ and $L_M(z)$. The Lagrangian of the optimization problem is

$$\mathcal{L}(\rho; \mu) := \sum_z \frac{1}{\rho_z^2 L_z^2} + \mu \left(\sum_z \rho_z - 1 \right).$$

The KKT conditions give:

$$\begin{aligned} \partial_{\rho_z} : \quad & \frac{2}{\rho_z^3 L_z^2} = \mu \\ \text{C.S.} : \quad & \sum_z \rho_z = 1. \end{aligned}$$

In particular, we see that $\rho_z^3 \propto L_z^{-2}$, hence $\rho_z \propto L_z^{-2/3}$. Because $\rho \in \Delta^m$, this uniquely determines ρ. □

To paraphrase the main document, LSTS-imp is build from $\rho \propto L^{-2/3}$ with an additional scaling. Specifically, denote simply ρ the element of Δ^m such that $\rho \propto L^{-2/3}$. This element is pullbacked to an approximate solution of Equation (10) by picking the largest possible $\alpha \geq 1$ such that

$$\forall \pi \in \Pi, \quad \sum_{z \in \pi \cup \pi^*} \alpha \rho(z) \leq 1.$$

It is equivalent to

$$\forall \pi \in \Pi, \quad \alpha^{-1} \geq \sum_{z \in \pi \cup \pi^*} \rho(z).$$

We deduce that α is

$$\alpha = \left(\max_{\pi \in \Pi} \sum_{z \in \pi \cup \pi^*} \rho(z) \right)^{-1}.$$

Theorem 3 (LSTS-imp)**.** *Let $\rho \propto L^{-2/3}$. Let α be given by Equation* (12)*. The coefficients $\rho^\star := \alpha\rho$ satisfy the conditions a) and b) from Proposition 9. The LSTS method with such $\rho^\star$ is denoted LSTS-imp (standing for* improved *LSTS) and achieves:*

$$\lim_{\delta \to 0} \frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} \leq 8\sigma^2 n^{4/3} m^{2/3} \sum_z \frac{1}{L_M(z)^2}.$$

When rewards are Gaussian with standard deviations σ, LSTS-imp is $8n^{4/3}m^{2/3}$-asymptotically optimal.

There isn't much more to say that the provided sketch of the proof from the main body.

Proof. We have

$$\max_{\pi\in\Pi}\sum_{z\in\pi\cup\pi^*}\rho(z) \le \max_{\pi\in\Pi}\sum_{x\in\mathcal{S}}(\rho(x,\pi(x))+\rho(x,\pi^*(x))) \le \sum_{x\in\mathcal{S}} 2\max_{a\in\mathcal{A}_x}\rho(x,a).$$

We get the claimed:

$$\alpha \ge \left(2\sum_x \max_a \rho_M(z)\right)^{-1}$$

Recall that $\rho \propto L^{-2/3}$ then apply Theorem 2. Simple algebra leads to:

$$\lim_{\delta\to 0}\frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} \le 8\sigma^2\left(\sum_z L_M(z)^{-2/3}\right)\left(\sum_x \max_a L_M(x,a)^{-2/3}\right)^2$$

The function $t \ge 0 \mapsto t^{1/3}$ is increasing, so $\max_a L_M(x,a)^{-2/3} = (\max_a L_M(x,a)^{-2})^{1/3}$. Bound both terms with Hölder's inequality with parameters $(3, \frac{3}{2})$:

$$\lim_{\delta\to 0}\frac{\mathbb{E}^M[\tau_\delta]}{\log(1/\delta)} \le 8\sigma^2 m^{2/3}\|L^{-1}\|_2^{2/3}\left(n^{2/3}\left(\sum_x \max_a L_{x,a}^{-2}\right)^{1/3}\right)^2$$

Bounding $\max_a$ by $\sum_a$ yields the result. □

Finally, we prove the generalization of the method used in Example 3.

Proposition 11. *Assume $M \in \mathcal{H}$ have sub-Gaussian rewards with standard deviation $\sigma > 0$. Assume that there exist $\epsilon > 0$ together with a family of distinct state-action pairs $(z_i)_{i=1}^k$ such that*

$$\sum\nolimits_{z\in\mathcal{Z}} L_M(z)^{-2} \le (1+\epsilon)\sum\nolimits_{i=1}^k L_M(z_i)^{-2}.$$

Thus, if rewards are Gaussian with standard deviations σ, LSTS-imp is $8(k^2+\epsilon m^2)$-asymptotically optimal.

Proof. We show that if $(a_i)_{i\le m}$ is a family of non-negative coefficients such that

$$\sum_{i=1}^m a_i \le (1+\epsilon)\sum_{j=1}^k a_j,$$

then

$$\left(\sum_{j=1}^m a_j^{1/3}\right)^3 \le 4(k^2+\epsilon m^2)\sum_{i=1}^m a_i.$$

Enumerating $\mathcal{Z}$ as $(z_i)_{i=1}^m$ in the right order, with $m := |\mathcal{Z}|$ and $a_i = L_M(z_i)^{-2}$, we will get the result.

By Hölder's inequality,

$$\sum_{i=1}^m a_i^{1/3} = \sum_{i=1}^k a_i^{1/3} + \sum_{j=k+1}^m a_j^{1/3} \le k^{2/3}\left(\sum_{i=1}^k a_i\right)^{1/3} + \sum_{j=k+1}^m a_i^{1/3}.$$

Using that $\sum_{j=k+1}^m a_j \le \epsilon\sum_{i=1}^k a_i$ together with Hölder's inequality gives

$$\sum_{j=k+1}^m a_j^{1/3} \le (m-k)^{2/3}\left(\sum_{j=k+1}^m a_j\right)^{1/3} \le (m-k)^{2/3}\epsilon^{1/3}\left(\sum_{i=1}^k a_i\right)^{1/3}.$$

By merging the two equations together, we get

$$\left(\sum_{i=1}^m a_i^{1/3}\right)^3 \leq \left(k^{2/3} + \epsilon^{1/3}(m-k)^{2/3}\right)^3 \sum_{i=1}^m a_i.$$

Again, for $x, y \geq 0$, by Hölder's inequality, $a^{1/3} + b^{1/3} \leq 4^{1/3}(a+b)^{1/3}$. Applied to $a := k^2$ and $b := \epsilon(m-k)^2$, it provides

$$\begin{aligned}\left(\sum_{i=1}^m a_i^{1/3}\right)^3 &\leq 4\left(k^2 + \epsilon(m-k)^2\right)\sum_{i=1}^m a_i \\ &\leq 4\left(k^2 + \epsilon m^2\right)\sum_{i=1}^m a_i.\end{aligned}$$

□

D APPENDIX: LOWER BOUNDS

Proposition 7. *Let $M, M' \in \mathcal{H}$ two DMDPs with different bias-optimal policies such that $q(x,a)$ and $q'(x,a)$ are mutually absolutely continuous for all edges. Every δ-PC algorithm satisfies*

$$\sum\nolimits_{z\in\mathcal{Z}} \mathbb{E}^M[N_{\tau_\delta}(z)]\,\mathrm{KL}(q(z)\|q'(z)) \geq \mathrm{kl}(\delta, 1-\delta).$$

Proof of Proposition 7. The original result of Kaufmann et al. (2016) relies on a change of distributions argument. Recall the definition of $\mathcal{F}_{t+1} := \sigma(X_1, A_1, R_1, Y_1, \ldots, X_t, A_t, R_t, Y_t)$. Assume that $q(x,a)$ and $q'(x,a)$ are mutually absolutely continuous, so that we can find a common measure $\lambda_{x,a}$ on $\mathbb{R}$ for which $q(x,a)$ and $q'(x,a)$ have the respective densities $f_{x,a}$ and $f'_{x,a}$. Define the log-likelihood of observations up to time t under the execution of an algorithm $\mathcal{I}$ as

$$\begin{aligned}L_t &:= L_t(X_1, A_1, R_1, Y_1, \ldots, X_t, A_t, R_t, Y_t) \\ &:= \sum_{z\in\mathcal{Z}}\sum_{i=1}^t \mathbf{1}_{Z_i=z}\log\left(\frac{f_z(R_i)}{f'_z(R_i)}\right).\end{aligned}$$

Their proof of Lemma 7 is based on the following Lemma, which generalizes the change of measure argument with almost surely constant stopping time[5] to general stopping times.

Lemma 12. *Let τ be any almost surely finite stopping time w.r.t. $\mathcal{F}_t$. For every event $\mathcal{U} \in \mathcal{F}_\tau$ (i.e. $\{\tau = t\} \cap \mathcal{U} \in \mathcal{F}_t$),*

$$\mathbb{E}^{M',\mathcal{I}}[L_\tau] \geq \mathrm{kl}(\mathbb{P}^{M,\mathcal{I}}(\mathcal{U}), \mathbb{P}^{M',\mathcal{I}}(\mathcal{U})).$$

This Lemma is a mere restatement of the Lemma 19 from Kaufmann et al. (2016). Although the later is stated for identification algorithms for multi-armed bandits, the stochastic process underlying to an identification algorithm for a DMDP M with edge space $\mathcal{Z}$, rewards $q(x,a)$ and generative model is the same as to the stochastic process corresponding to a $\mathcal{Z}$-multi-armed bandits with the same rewards. This is thanks to the generative model assumption; The DMDP setting *is* actually an instance of multi-armed bandits with arms $\mathcal{Z}$ – but with a much subtler notion of optimality.[6]

An application of Wald's Lemma to L_τ produces

$$\mathbb{E}^{M,\mathcal{I}}[L_\tau] := \sum_{z\in\mathcal{Z}} \mathbb{E}^{M,\mathcal{I}}[N_z(\tau)]\,\mathrm{KL}(q(z)\|q'(z)).$$

To get Lemma 7, apply Lemma 12 with $\mathcal{U} := \left\{\tau_\delta < \infty, \pi^{\mathcal{I}}_{\tau_\delta} \neq \pi^*_M\right\}$. We have $\pi^*_M \in \Pi^*_0(M)$, $\pi^*_M \notin \Pi^*_0(M')$ and $\mathcal{I}$ is δ-PC, so

$$\begin{aligned}&\mathbb{P}^{M,\mathcal{I}}(\mathcal{U}) \leq \delta, \qquad \text{and} \\ &\mathbb{P}^{M',\mathcal{I}}(\mathcal{U}) \geq 1-\delta.\end{aligned}$$

So overall, $\mathrm{kl}(\mathbb{P}^{M,\mathcal{I}}(\mathcal{U}), \mathbb{P}^{M',\mathcal{I}}(\mathcal{U})) \geq \mathrm{kl}(\delta, 1-\delta)$. □

[5] Namely, most informational lower bounds come from the famous *change of measure* argument: for a fixed $T \geq 1$, if U is $\mathcal{F}_T$-measurable, then $\mathbb{E}^M[U] = \mathbb{E}^{M'}[U \cdot L_T]$.

[6] Without the generative model assumption, the selection of actions is also subject to constraints (the current state). This case has been covered by the works of Marjani et al. (2021).

Although the lower bound provided by Proposition 7 is very powerful, it isn't tractable in general. Albeit weaker, the edgewise lower bound variant of Proposition 8 is fairly easy to compute in practice.

Proposition 8 (Edgewise Lower Bound)**.** *Let $M \in \mathcal{H}$ with Gaussian rewards of standard deviation $\sigma > 0$. For all δ-PC identification algorithm,*

$$\frac{\mathbb{E}^M[\tau_\delta]}{\mathrm{kl}(\delta, 1-\delta)} \geq \sigma^2 \sum_{z \in \mathcal{Z}} \frac{1}{L_M(z)^2}.$$

Proof. Let $\epsilon > 0$ and $z \in \mathcal{Z}$. There is $\sigma \in \{\pm 1\}$ such that $\Pi_0^*(\mathbf{r} + \sigma(1+\epsilon)L_{\mathbf{r}}(z)\mathbf{e}_z) \neq \Pi_0^*(\mathbf{r})$. Let M' the DMDP with Gaussian rewards of standard deviation σ and mean reward vector $\mathbf{r}' := \mathbf{r} + \sigma(1+\epsilon)L_{\mathbf{r}}(z)\mathbf{e}_z$. Since $\mathcal{H}$ is dense in the whole set of DMDPs and that

$$q' \mapsto \sum_{z' \in \mathcal{Z}} \mathbb{E}^{M,\mathcal{I}}[N_{z'}(\tau_\delta)] \,\mathrm{KL}(q(z') \| q'(z'))$$

is continuous, we may assume that $M' \in \mathcal{H}$ up to an infinitesimal perturbation of $\mathbf{r}'$. Then by Lemma 7, we have

$$\sum_{z' \in \mathcal{Z}} \mathbb{E}^{M,\mathcal{I}}[N_{z'}(\tau_\delta)] \,\mathrm{KL}(q(z') \| q'(z')) \geq \mathrm{kl}(\delta, 1-\delta).$$

By construction, $\sum_{z' \in \mathcal{Z}} \mathbb{E}^{M,\mathcal{I}}[N_{z'}(\tau_\delta)] \,\mathrm{KL}(q(z') \| q'(z')) = \mathbb{E}^{M,\mathcal{I}}[N_z(\tau_\delta)] \,\mathrm{KL}(q(z) \| q'(z))$. Moreover,

$$\mathrm{KL}(q(z) \| q'(z)) = \frac{(1+\epsilon)^2 L_{\mathbf{r}}(z)^2}{\sigma^2}.$$

Therefore,

$$\frac{\mathbb{E}^{M,\mathcal{I}}[N_z(\tau_\delta)]}{\mathrm{kl}(\delta, 1-\delta)} \geq \frac{1}{\mathrm{KL}(q(z) \| q'(z))} = \frac{\sigma^2}{(1+\epsilon)^2 L_M(z)^2}.$$

This holds for all $\epsilon > 0$. Letting ϵ go to 0, we obtain the announced lower bound. □

E APPENDIX: COMPUTATIONS

E.1 The computation of bias optimal policies on $\mathcal{H}$ is $O(nm)$

E.1.1 Howard's Policy Iteration

Proposition 3 is of interest when it comes to the effective computation of Blackwell optimal policies. This section describes two algorithms to compute them in efficient time. The first one below is an adaptation Howard's Policy Iteration algorithm.

This policy iteration algorithm is known to terminate in a finite time, see Cochet-Terrasson et al. (1998). More precisely, in the deterministic case one step costs $O(n+m)$ and the number of steps is bounded above by $|\Pi|$, the number of policies. Although the complexity bound $O(m|\Pi|)$ is not polynomial, the algorithm is known to perform extremely well numerically, see Cochet-Terrasson et al. (1998); this is no surprise, knowning the intimate relationship with the simplex algorithm. For the sake of self-containedness, we provide an ad-hoc proof of the termination of HPI-BO.

Lemma 13. *Let $(\pi_t)_{t \geq 1}$ the sequence of policies generated by HPI-BO. For all $x \in \mathcal{S}$, $g_{\pi_t}(x)$ is non-decreasing with t and on time-steps such that $g_{\pi_t} = g_{\pi_{t+1}}$, we have $\forall x, h_{\pi_t}(x) \leq h_{\pi_{t+1}}(x)$ with strict inequality for at least one $x \in \mathcal{S}$. Made short, using the product order on $\mathbb{R}^{\mathcal{S}}$, the pair of vectors (g_{π_t}, h_{π_t}) is increasing for the lexicographic order.*

Proof. To begin with, because the transition structure of a DMDP is deterministic, we abuse of notations and denote $\pi^k(x)$ the k-th states reached by π starting from x. In particular, $\pi^0(x) = x$ and $\pi^1(x) = s(x, \pi(x))$.

The proof distinguishes two cases : updates of the current policy π_t under the G-rule and under the H-rule. Under the G-rule, terminal cycles do not change, but $\pi_t(x)$ may be modified to another $a \in \mathcal{A}_x$ that leads to a better terminal cycle. Under the H-rule, terminal cycles may change. Specifically, if a new cycle is created, its value is higher than the previous ones ; if no cycle emerges under H-rule, the gain is unchanged and the bias increases at some vertex. The proof of these statements goes by induction on $\mathcal{S}$ with respect to the following quantities. For $t \geq 1$ and $x \in \mathcal{S}$, define

$$k_x^t := \left|\left\{\ell : \pi_{t+1}^1(\pi_{t+1}^\ell(x)) \neq \pi_t^1(\pi_{t+1}^\ell(x))\right\}\right|$$

Algorithm 2 Howard's Policy Iteration for Bias Optimality (HPI-BO)

Require: a DMDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ with multigraph $\mathcal{G}$;
Ensure: return π a bias optimal policy.
$t \leftarrow 0$
initialize $\pi_0 : \mathcal{S} \to \mathcal{A}$ arbitrarily
while $\pi_t \neq \pi_{t-1}$ **do**
  compute $g_{\pi_t}(x)$ and $h_{\pi_t}(x)$ for $x \in \mathcal{S}$
  compute $G_x^t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}_x}[g_{\pi_t}(s,(x,a))]$ for $x \in \mathcal{S}$ *G-sets*
  compute $H_x^t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}_x \cap G_x^t}[r(x,a) - g_{\pi_t}(x) + h_{\pi_t}(s(x,a))]$ for $x \in \mathcal{S}$ *H-sets*
  choose for all $x \in \mathcal{S}$ *G-rule*

$$\pi_t'(x) := \begin{cases} a \in G_x^t & \text{if } \pi_t(x) \notin G_x^t \\ \pi_t(x) & \text{if } \pi_t(x) \in G_x^t \end{cases} \tag{24}$$

  if $\pi_t' = \pi_t$ **then**
    choose for all $x \in \mathcal{S}$ *H-rule*

$$\pi_t'(x) := \begin{cases} a \in H_x^t & \text{if } \pi_t(x) \notin H_x^t \\ \pi_t(x) & \text{if } \pi_t(x) \in H_x^t \end{cases} \tag{25}$$

  end if
  $\pi_{t+1} \leftarrow \pi_t'$
  $t \leftarrow t+1$
end while
return π_t.

the number, possibly infinite, of successors of x by π_{t+1} from which π_t and π_{t+1} take different decisions. Define also d_x^t the distance of x to $\mathcal{C}_x^{\pi_{t+1}}$, that is

$$d_x^t := \inf\left\{\ell \geq 0 \mid \pi_{t+1}^\ell(x) \in \mathcal{C}_x^{\pi_{t+1}}\right\} < \infty.$$

When a non-trivial G-rule is applied, it means there exists $(x,a) \in \mathcal{Z}$ such that $g_{\pi_t}(\pi_t^1(x)) < g_{\pi_t}(s(x,a))$, so that under π_t, $s(x,a)$ leads to a better cycle of π_t than $\pi_t(x)$. By induction on (k_x^t, d_x^t) for the lexicographic order, we show that $\forall x \in \mathcal{S}, g_{\pi_{t+1}}(x) \geq g_{\pi_{t+1}}(x)$. Moreover, for $x \in \mathcal{S}$ such that $k_x^t = 1$, we have

$$g_{\pi_{t+1}}(x) = g_{\pi_t}(\pi_{t+1}^1(x)) > g_{\pi_t}(\pi_t^1(x)) = g_{\pi_t}(x).$$

Therefore, the gain increases for the product order on $\mathbb{R}^{\mathcal{S}}$.

When a non-trivial H-rule is applied, it means that for all $(x,a) \in \mathcal{Z}$, $g_{\pi_t}(x) \geq g_{\pi_t}(s(x,a))$. A quick induction shows that if there exists a path from u to v, then $g_{\pi_t}(u) \geq g_{\pi_t}(v)$. Now, remark that $k_x^t < \infty$ if, and only if $\mathcal{C}_x^{\pi_{t+1}}$ is a cycle of π. We start by investigating the case $k_x^t = \infty$ where under π_{t+1}, x converges to a cycle $\mathcal{C}$ which is not a cycle of π. Expand $\mathcal{C}$ as a sequence of states-actions $(u_0, a_0, \dots, u_{c-1}, a_{c-1})$. For all $i \leq c-1$, we have

$$r(u_i, a_i) - g_{\pi_t}(u_i) + h_{\pi_t}(u_{i+1}) \geq h_{\pi_t}(u_i)$$

with strict inequality for at least one i. Hence, summing over i,

$$\frac{1}{c}\sum_{i=0} r(u_i, a_i) + \frac{1}{c}\sum_{i=1}^{c} h_{\pi_t}(u_i) > \frac{1}{c}\sum_{i=0}^{c-1} g_{\pi_t}(u_i) + \frac{1}{c}\sum_{i=0}^{c-1} h_{\pi_t}(u_i).$$

Since all u_i are connected, $g_{\pi_t}(u_i)$ does not depend on i. Moreover, the terms involving the bias cancel out. It leaves

$$\forall j \leq c-1, \qquad g_{\pi_{t+1}}(u_j) = g(\mathcal{C}) > g_{\pi_t}(u_j).$$

Hence, if $k_x^t = \infty$, then $\mathcal{C}_x^{\pi_{t+1}}$ is not a cycle of π_t and $g_{\pi_{t+1}}(x) > g_{\pi_t}(x)$.

In the case $k_x^t < \infty$, x converges (under π_{t+1}) to a cycle of π_t and by construction of the H-sets, $g_{\pi_{t+1}}(x) = g_{\pi_t}(x)$. We prove by induction over (k_x^t, d_x^t) for the lexicographic order that $h_{\pi_{t+1}}(x) \geq h_{\pi_t}(x)$. For $k_x^t = 0$, this is obvious because π_t and π_{t+1} coincide on their successors from x. Assume that $k_x^t > 0$. If $k_{\pi_{t+1}^1(x)}^t < k_x^t$, it means that $\pi_t^1(x) \neq \pi_{t+1}^1(x)$ and

$$r(x, \pi_{t+1}(x)) - g_{\pi_t}(x) + h_{\pi_t}(\pi_{t+1}^1(x)) > h_{\pi_t}(x).$$

By induction, $h_{\pi_t}(\pi^1_{t+1}(x)) \geq h_{\pi_{t+1}}(\pi^1_{t+1}(x))$. Moreover, $g_{\pi_t}(x) = g_{\pi_{t+1}}(x)$. In the end,

$$h_{\pi_{t+1}}(x) > h_{\pi_t}(x). \tag{26}$$

The other possible case is $k^t_{\pi^1_{t+1}(x)} = k^t_x$. Then, $d^t_{\pi^1_{t+1}(x)} < d^t_x$ and the same argument leads to $h_{\pi_{t+1}}(x) \geq h_{\pi_t}(x)$. Overall, we have $h_{\pi_{t+1}}(x) \geq h_{\pi_t}(x)$ for all $x \in \mathcal{S}$ such that $k^t_x < \infty$. Furthermore, when $k^t_x > 0$, the equation (26) enlighten that there exists a successor of x which increases the bias.

To summary, if there exists $k^t_x = \infty$, the gain increases under H-rule ; otherwise, all k^t_x are finite and because $\pi_t \neq \pi_{t+1}$, there must exists at least one $k^t_x > 0$ so the bias increases for the product order of $\mathbb{R}^{\mathcal{S}}$ under H-rule. □

Corollary 1. *HPI-BO terminates in finite time and returns a bias optimal policy.*

Proof. Let $T \in [0, +\infty]$ number of time-steps before termination. For each $t < T$, either the gain or the bias increases, so $t \in [0, T-1] \mapsto \pi_t \in \Pi$ is an injective map. Because Π is finite, T must be finite. Now, by definition, $\pi_{T-1} = \pi_T$, so that π_T satisfies the assertion 3 of Theorem 3. Therefore, $\pi_T \in \Pi^*_\infty(M)$. □

E.1.2 Bias optimality via maximal mean cycle

Another approach is to compute the maximal mean weight cycle of $\mathcal{G}$ directly, using algorithms such as Karp's maximal mean cycle algorithm (see Karp (1978)). This optimal cycle is obtained in time $O(nm)$ and once $\mathcal{C}_*$ is known, we show next that Howard Policy Iteration computes the bias optimal policy in at most n steps.

Algorithm 3 Karp Maximal Mean Cycle + Howard Policy Iteration (KMMC+HPI)

Require: a DMDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ satisfying (H1);
Ensure: π is bias optimal
1: compute the maximal mean cycle $\mathcal{C}_*$ using Karp's algorithm;
2: compute a policy π_0 with unique terminal cycle $\mathcal{C}_*$;
3: apply HPI-BO starting from π_0 to get π_t;
4: **return** π_t.

Proposition 12. *On DMDPs satisfying H1, the KMMC+HPI algorithm computes a bias-optimal policy in time $O(nm)$.*

Proof. Assume that M satisfies (H1). Karp's algorithm has execution time $O(nm)$, and the computation of a policy π_0 with unique terminal cycle $\mathcal{C}_*$ can be done in $O(n+m)$. Let (π_t) the sequence of policies generated by HPI-BO. We show that HPI-BO runs at most S steps. Let $\mathcal{Z}^* = \{(x,a) \in \mathcal{Z} \mid \exists \pi^* \in \Pi^*_0(M), a = \pi^*(x)\}$. For $x \in \mathcal{S}$, set d^t_x the distance of x to the terminal cycle $\mathcal{C}_*$ under π_t, that is $d^t_x := \inf\left\{\ell \geq 0 \mid \pi^\ell_t(x) \in \mathcal{C}_*\right\}$. We claim that after t steps of HPI, for all $x \in \mathcal{S}$ such that $d^t_x \leq t$, we have $(x, \pi_t(x)) \in \mathcal{Z}^*$ and $h_{\pi_t}(x) = h_*(x)$. The proof goes by induction on t.

Because $\mathcal{C}_*$ is unique, the result is obvious for $t = 0$. Assume that $t > 1$. We know that the application of the G-rule increases the gain, so π_t must be updated according to the H-rule. What is more, the H-rule does not change the terminating cycle. Let x such that $d^t_x < t$. By induction, π_{t-1} achieves optimal bias from x, so from the H-rule, $\pi_{t-1}(x) = \pi_t(x)$. Accordingly, π_t is bias-optimal from all x at distance less than t of $\mathcal{C}_*$. Let x such that $d^t_x = t$. Because $g_{\pi_t}(y) = g(\mathcal{C}_*)$ for all $y \in \mathcal{S}$, we see that

$$\pi_t(x) \in \operatorname*{argmax}_{a \in \mathcal{A}_x} \left[r(x,a) + h_{\pi_{t-1}}(s(x,a))\right]$$

By definition, $d^t_y = d^t_x - 1 < t$, so $h_{\pi_{t-1}}(y) = h_*(y)$ and $h_{\pi_t}(y) = h_*(y)$. Remark that if π^* is a bias-optimal policy, then for all $(x,a) \notin \mathcal{Z}^*$, we have $r(x,a) + h_*(s(x,a)) < r(x,\pi^*(x)) + h_*(\pi^*(x))$ by Theorem 3. Using $h_{\pi_{t-1}}(s(x,a)) \leq h_*(s(x,a))$, we get that for all $z \in \mathcal{S}$ such that $(x,a) \notin \mathcal{Z}^*$,

$$r(x,a) + h_{\pi_{t-1}}(s(x,a)) < r(x,\pi_t(x)) + h_{\pi_{t-1}}(\pi_t(x)).$$

In particular, $(x, \pi_t(x)) \in \mathcal{Z}^*$. Let $\pi_* \in \Pi^*_0(M)$ such that $\pi_t(x) = \pi_*(x)$. Then

$$\begin{aligned} h_*(x) &= r(x,\pi_*(x)) - g_*(x) + h_*(\pi_*(x)) \\ &= r(x,\pi_t(x)) - g_*(x) + h_{\pi_t}(\pi_t(x)) = h_{\pi_t}(x). \end{aligned}$$

Setting $t = n-1$, we obtain that $\forall x \in \mathcal{S}, h_{\pi_t}(x) = h_*(x)$. Hence HPI terminates in at most n steps, each step running in time $O(m)$. We end up with a total time complexity $O(nm)$. □

E.2 The computation of local suboptimality gaps is $O(nm)$

From the identities derived by Lemma 11, we are able to compute the local suboptimality gaps efficiently.

Lemma 14. *Let $M \in \mathcal{H}$ a DMDP. If the optimal policy $\pi \in \Pi^*_\infty(M)$ is given, the family $(L_M(z))_{z\in\mathcal{Z}}$ can be computed in $O(nm)$ time.*

Proof. When $\pi \in \Pi^*_0(M)$ is given, the computation of g_* and h_* are done in time $O(n)$ and the Bellman gaps $\Delta_M(x,a)$ are deduced in $O(1)$ time each, for a total of $O(m)$ computations. We deduce $L_M(x,a)$ for $(x,a) \notin \pi$ immediately and are left to compute the quantities $L_M(x,a)$ with $(x,a) \in \pi$. One starts with a preliminary computation of the relation $\rightsquigarrow^*_\pi$ and the family of reaching times $(\tau^\pi_u(v))_{u,v\in\mathcal{S}}$ for all total time $O(n^2)$ by backpropagating the values from the terminal cycle of π (that can be computed in $O(n)$). Finally, for each $(x,a) \in \pi$, $L_M(x,a)$ is computed in $O(m)$ each and there are n of them. This results in an overall $O(n^2 + nm)$ time complexity. □

Because the computation of bias optimal policies can be done in time $O(nm)$ (see the previous section), the computation of local gaps are $O(nm)$ in total.