Adaptive Sampling for Best Policy Identification in Markov Decision Processes

Aymen Al Marjani [1] **Alexandre Proutiere** [2]

Abstract

We investigate the problem of best-policy identification in discounted Markov Decision Processes (MDPs) when the learner has access to a generative model. The objective is to devise a learning algorithm returning the best policy as early as possible. We first derive a problem-specific lower bound of the sample complexity satisfied by any learning algorithm. This lower bound corresponds to an optimal sample allocation that solves a non-convex program, and hence, is hard to exploit in the design of efficient algorithms. We then provide a simple and tight upper bound of the sample complexity lower bound, whose corresponding nearly-optimal sample allocation becomes explicit. The upper bound depends on specific functionals of the MDP such as the sub-optimality gaps and the variance of the next-state value function, and thus really captures the hardness of the MDP. Finally, we devise KLB-TS (KL Ball Track-and-Stop), an algorithm tracking this nearly-optimal allocation, and provide asymptotic guarantees for its sample complexity (both almost surely and in expectation). The advantages of KLB-TS against state-of-the-art algorithms are discussed and illustrated numerically.

1. Introduction

Reinforcement Learning (RL) algorithms are designed to interact with an unknown stochastic dynamical system, and through this interaction, to identify, as fast as possible, an optimal control policy. The efficiency of these algorithms is usually measured through their *sample complexity*, defined as the number of samples (the number of times the algorithm interacts with the system) required to identify an optimal policy with some prescribed levels of accuracy and certainty. This paper, as most related work in this field, focuses on systems and control objectives that are modelled as a standard discounted Markov Decision Processes (MDPs) with finite state and action spaces. Various interaction models have been investigated, but sample complexity analyses have been mainly conducted under the so-called *generative model*, where in each step, the algorithm may sample a transition and a reward from any given (state, action) pair. We also restrict our attention to this model.

We investigate the design of RL algorithms with minimal sample complexity. This problem has attracted a lot of attention over the last two decades. Most studies follow a minimax approach. For example, it is known (Azar et al., 2013) that for the worst possible MDP, identifying an ε-optimal policy with probability $1-\delta$ requires at least $\frac{SA}{\varepsilon^2(1-\gamma)^3}\log(\frac{SA}{\delta})$ samples, where S and A are the number of states and actions, respectively, and γ is the discount factor. Note that to obtain this sample complexity lower bound, one needs to design a very specific worst-case MDP (in particular, its transition probabilities must depend on ε and γ). Since the aforementioned minimax lower bound appeared, most researchers have been aiming at devising algorithms matching this bound. In contrast, we are interested in analyzing the minimal *problem-specific* sample complexity. Specifically, we seek to understand the dependence of the sample complexity on the MDP that has to be learnt. Problem-specific performance metrics are much more informative than their minimax counterparts, because they encode and express the inherent hardness of the MDP. Minimax metrics just represent the hardness of the worst MDP. In particular, establishing that the sample complexity of an algorithm does not exceed the minimax lower bound just reveals that the algorithm performs well for this worst MDP. However, it does not indicate whether the algorithm *adapts* to the hardness of the MDP, i.e., whether the optimal policy of a very easy MDP would be learnt very quickly. As a matter of fact, an algorithm with sample complexity matching the minimax lower bound just consists in sampling (state, action) pairs uniformly at random, and is not adapting to the MDP.

The problem-specific sample complexity of identifying the best arm in stochastic Multi-Armed Bandit (MAB) prob-

[1]UMPA, ENS Lyon. This work was done while Aymen Al Marjani was at KTH. [2]KTH Royal institute of Technology. Correspondence to: Aymen Al Marjani <aymen.al_marjani@ens-lyon.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

lems is now well understood (Garivier & Kaufmann, 2016). In this work, we explore whether the methodology used in (Garivier & Kaufmann, 2016) for MAB problems can be extended to RL problems. This methodology consists in first deriving a problem-specific sample complexity lower bound which should reveal the sample allocation leading to the minimal sample complexity. One may then devise a *track-and-stop* algorithm that (i) tracks the optimal sample allocation identified in the lower bound, and (ii) stops when the information gathered is judged sufficient to get the desired PAC guarantees. As it turns out, extending this methodology to RL problems raises fundamental issues, mainly due to the difficulty of computing the sample allocation leading to the minimal problem-specific sample complexity. We propose a set of tools to solve these issues. Our contributions are as follows:

1. We derive a problem-specific sample complexity lower bound for identifying an optimal policy in a given MDP ϕ. This bound is expressed as $T^*(\phi)\log(1/\delta)$, where the *characteristic time* $T^*(\phi)$ encodes the hardness of the MDP ϕ. $T^*(\phi)$ is the value of a complex non-convex optimization problem. This complexity makes the design of a track-and-stop algorithm similar to that proposed in (Garivier & Kaufmann, 2016) and achieving the sample complexity lower bound elusive. To circumvent this difficulty, we derive an explicit upper bound $U(\phi)$ of $T^*(\phi)$. The advantage of $U(\phi)$ is two-fold: (i) $U(\phi)$ remains problem-specific, and explicitly depends on functionals of the MDP characterizing its hardness. (ii) $U(\phi)$ corresponds to an explicit and simple sample allocation. This allows us to devise a procedure that tracks this allocation.

2. Based on our upper bound analysis, we devise KLB-TS (KL Ball Track-and-Stop), an algorithm whose sample complexity is at most $U(\phi)\log(1/\delta)$. Our algorithm relies on a procedure tracking the sample allocation leading to $U(\phi)$, and a stopping rule that we refer to as KL Ball Stopping rule because of its analogy to the way we derive the upper bound $U(\phi)$.

3. We highlight the differences of our design approach compared to that leading to BESPOKE (Zanette et al., 2019), a recently proposed adaptive algorithm. As it turns out, the adaptive part of BESPOKE is very limited in practice (see related work and Appendix H for details), and KLB-TS exhibits a much better performance numerically.

2. Related Work

Most work on the best policy identification in MDPs adopt a minimax approach (Kearns & Singh, 1999), (Kakade, 2003), (Even-Dar et al., 2006), (Azar et al., 2013), (Sidford et al., 2018), (Agarwal et al., 2020), (Li et al., 2020). In the most recent of these papers (Li et al., 2020), the authors propose an algorithm whose sample complexity achieves the minimax lower bound of (Azar et al., 2013) for a wide range of values of ε, namely for $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Refer to the appendix for a detailed account on the minimax framework.

As far as we are aware, the only paper attempting to propose a problem-specific analysis of the best policy identification in MDPs with a generative model is (Zanette et al., 2019). There, the authors proposed BESPOKE, an adaptive algorithm designed to find ε-optimal policies. BESPOKE starts by allocating an extremely large number of samples $n_{\min} = \frac{2\times 625^2\times\gamma^2\times S\times\log(1/\delta)}{(1-\gamma)^2}$ to each (state, action) pair. Then, at each iteration, BESPOKE solves a convex program whose objective is an upper-bound of the sub-optimality gap (in terms of the ℓ_∞-norm of the value function) of the empirical optimal policy. The solution of this program corresponds to the sampling strategy that the algorithm uses to halve the sub-optimality gap of the empirical policy in the next iteration. Interestingly, BESPOKE is the first algorithm with a problem-dependent sample complexity upper-bound. Note however that BESPOKE has not been tested numerically in (Zanette et al., 2019); we fill this gap in this paper. Because of its very long initialization phase, it turns out that the part where BESPOKE actually adapts its sample allocation is negligible in comparison of its total sample complexity. In Appendix H, we provide a more detailed discussion on BESPOKE, and further compare the sample complexity upper bounds of KLB-TS and BESPOKE. Experiments in Section 7 show that KLB-TS significantly outperforms BESPOKE numerically.

3. Preliminaries and Notation

3.1. Discounted MDPs

We investigate the optimal control of dynamical systems modelled as an infinite time-horizon MDP with finite state space $\mathcal{S}$ and finite action spaces $\mathcal{A}_s$ for any $s \in \mathcal{S}$. Let $\mathcal{A} = \cup_{s\in\mathcal{S}}\mathcal{A}_s$. The MDP is defined by its kernels: $\phi = (p_\phi, q_\phi)$, where p_ϕ captures the system dynamics and q_ϕ the random collected rewards. Specifically, $p_\phi(s'|s,a)$ denotes the probability of the system to be in state s' after taking the action $a \in \mathcal{A}_s$ in state s. Let $p_\phi(s,a) = (p_\phi(s'|s,a))_{s'}$. $q_\phi(\cdot|s,a)$ or simply $q_\phi(s,a)$ is the density of the distribution of the reward collected in state s when action a is selected, w.r.t. some positive measure λ with support included in $[0,1]$. Let $r_\phi(s,a)$ denote the expected reward collected in state s when action a is selected, $r_\phi(s,a) = \int_0^1 Rq_\phi(R|s,a)\lambda(dR)$.

The objective is to identify a control policy $\pi : \mathcal{S} \to \mathcal{A}$ maximizing the long-term discounted reward $\mathbb{E}_\phi[\sum_{t=0}^\infty \gamma^t r_\phi(s^\pi(t), \pi(s^\pi(t))]$, where $s^\pi(t)$ is the state of the system at time t under the policy π and $\mathbb{E}_\phi[\cdot]$ represents the expectation taken w.r.t. to the randomness induced by

(p_ϕ, q_ϕ).

We denote by V_ϕ^π the value function of the MDP ϕ when the control policy is π: for any s, $V_\phi^\pi(s) = \mathbb{E}_\phi[\sum_{t=0}^\infty \gamma^t r_\phi(s^\pi(t), \pi(s^\pi(t)))|s^\pi(0) = s]$. $V_\phi^\star$ corresponds to the value function when the policy π is optimal. Note that since the rewards are lower and upper bounded by 0 and 1, respectively, we have for any s, $V_\phi^\star(s) \in [0, \frac{1}{1-\gamma}]$. Similarly, the Q-function is denoted by Q_ϕ^π, and $Q_\phi^\star$ when π is optimal. The sub-optimality gap of action a in state s is defined as $\Delta_{sa} = V_\phi^\star(s) - Q_\phi^\star(s,a)$. Finally, denote by $\Pi_\phi^\star$ the set of optimal policies for ϕ.

Assumption 1. To simplify notation and the analysis, we assume that ϕ admits a unique optimal control policy denoted by $\pi_\phi^\star$. This means that $\phi \in \Phi$, where Φ is defined as $\Phi = \{\phi : |\Pi_\phi^\star| = 1\}$.

3.2. Best-policy identification

We aim at devising an algorithm identifying $\pi_\phi^\star$ as quickly as possible in the fixed-confidence setting: when the algorithm stops and returns an estimated optimal policy $\hat{\pi}$, we should have $\mathbb{P}_\phi[\hat{\pi} \neq \pi_\phi^\star] \leq \delta$, for some pre-defined confidence parameter $\delta > 0$. Such an algorithm consists of a sampling rule, a stopping rule, and a decision rule. An algorithm χ gathers information sequentially, and we denote by $\mathcal{F}_t^\chi$ the σ-algebra generated by all observations made under χ up to and including round t.

Sampling rule. In round t, the algorithm χ selects a (state, action) pair (s_t, a_t) to explore, depending on past observations. (s_t, a_t) is $\mathcal{F}_{t-1}^\chi$-measurable. χ observes the next state denoted by s_t' and a random reward R_t. Note that any admissible (state, action) pair may be selected (we consider a generative model).

Stopping and decision rules. After gathering enough information, χ may decide to stop sampling and to return an estimated best policy. The algorithm stops after collecting τ samples, and τ is a stopping time w.r.t. the filtration $(\mathcal{F}_t^\chi)_{t\geq 1}$. The estimated best policy $\hat{\pi}$ is then $\mathcal{F}_\tau^\chi$-measurable. τ is referred to as the sample complexity of χ.

δ-PC algorithms. An algorithm is δ-Probably Correct (δ-PC) if it satisfies the two following conditions: for any MDP $\phi \in \Phi$, (i) it stops in finite time almost surely, $\mathbb{P}_\phi[\tau < \infty] = 1$, and (ii) $\mathbb{P}_\phi[\hat{\pi} \neq \pi_\phi^\star] \leq \delta$.

3.3. Additional notation

$\mathbb{1}(s)$ denotes the canonical base vector in $\mathbb{R}^\mathcal{S}$ whose only non-zero entry is at index s. $\Sigma = \{\omega \in [0,1]^{S\times A} : \sum_{s,a} w_{sa} = 1\}$ denotes the simplex in $\mathbb{R}^{S\times A}$. The Kullback-Leibler divergence between two probability distributions P and Q on some discrete space $\mathcal{S}$ is defined as: $KL(P\|Q) = \sum_{s\in\mathcal{S}} P(s)\log(\frac{P(s)}{Q(s)})$. For Bernoulli distributions of respective means p and q, the KL divergence is denoted by $\mathrm{kl}(p,q)$. For distributions over $\mathbb{R}$ defined through their densities p and q w.r.t. some positive measure λ, the KL divergence is: $KL(p\|q) = \int_{-\infty}^\infty p(x)\log\left(\frac{p(x)}{q(x)}\right)\lambda(dx)$. For two MDPs ϕ and ψ, we say that $\phi \ll \psi$ if for all (s,a), $p_\phi(\cdot|s,a) \ll p_\psi(\cdot|s,a)$ and $q_\phi(\cdot|s,a) \ll q_\psi(\cdot|s,a)$. In that case, we define $\mathrm{KL}_{\phi|\psi}(s,a)$ as the KL divergence between the distributions of the random observations made for the (state, action) pair (s,a) under ϕ and ψ:

$$\mathrm{KL}_{\phi|\psi}(s,a) = KL(p_\phi(s,a)\|p_\psi(s,a)) + KL(q_\phi(s,a)\|q_\psi(s,a)).$$

4. Problem-Specific Sample Complexity Lower Bound

To derive a problem-specific sample complexity lower bound, we use classical change-of-measure arguments as those leveraged towards regret and sample complexity lower bounds (Lai & Robbins, 1985; Garivier & Kaufmann, 2016) in bandit problems. These arguments lead to constraints on the expected numbers of times each (state, action) pair should be explored under any δ-PC algorithm.

Definition 1. *The set of alternative MDPs is defined as:* $\mathrm{Alt}(\phi) = \{\psi \text{ MDP} : \phi \ll \psi \text{ and } \Pi_\phi^\star \cap \Pi_\psi^\star = \emptyset\}$.

Let $\psi \in \mathrm{Alt}(\phi)$ be an alternative MDP and consider a δ-PC algorithm. We denote by O_τ the set of observations made under the algorithm until it stops. Further consider L_τ the log-likelihood ratio of O_τ under the MDPs ϕ and ψ. Using similar techniques as those used in the proof of Wald's first lemma, we get (all proofs are detailed in the appendix):

Lemma 1. *Let $n_t(s,a)$ be the number of times (s,a) has been explored up to and including step t. For any $\phi \in \Phi$,* $\mathbb{E}_\phi[L_\tau] = \sum_{s,a} \mathbb{E}_\phi[n_\tau(s,a)]\,\mathrm{KL}_{\phi|\psi}(s,a)$.

From the above lemma, and using the same arguments as in (Kaufmann et al., 2016), one may derive the following data processing inequality, valid for any $\mathcal{F}_\tau$-measurable event E:

$$\sum_{s,a} \mathbb{E}_\phi[n_\tau(s,a)]\,\mathrm{KL}_{\phi|\psi}(s,a) \geq \mathrm{kl}(\mathbb{P}_\phi[E], \mathbb{P}_\psi[E]).$$

Next, we select the event E as $\{\hat{\pi} \notin \Pi^\star(\phi)\}$. Since the algorithm is δ-PC, and since $\psi \in \mathrm{Alt}(\phi)$, we have: $\mathbb{P}_\phi[E] \leq \delta$ and $\mathbb{P}_\psi[E] \geq \mathbb{P}_\psi[\hat{\pi} \in \Pi^\star(\psi)] \geq 1-\delta$. Using the monotonicity of the KL divergence, we deduce that $\mathrm{kl}(\mathbb{P}_\phi[E], \mathbb{P}_\psi[E]) \geq \mathrm{kl}(\delta, 1-\delta)$. We have established that under any δ-PC algorithm, the numbers of times

$(n_\tau(s,a))_{s,a}$ the different (state, action) pairs are explored satisfy: for any MDP $\psi \in \mathrm{Alt}(\phi)$,

$$\sum_{s,a} \mathbb{E}_\phi[n_\tau(s,a)]\, \mathrm{KL}_{\phi|\psi}(s,a) \geq \mathrm{kl}(\delta, 1-\delta). \quad (1)$$

Combining the above constraints with the fact that $\tau = \sum_{s,a} n_\tau(s,a)$, we obtain the following sample complexity lower bound.

Proposition 1. *The sample complexity of any δ-PC algorithm satisfies: for any $\phi \in \Phi$,*

$$\mathbb{E}_\phi[\tau] \geq T^*(\phi)\mathrm{kl}(\delta, 1-\delta), \quad (2)$$

$$\textit{where } T^*(\phi)^{-1} = \sup_{\omega \in \Sigma} \inf_{\psi \in \mathrm{Alt}(\phi)} \sum_{s,a} \omega_{sa} KL_{\phi|\psi}(s,a). \quad (3)$$

In the above proposition, $\omega_{sa}\mathrm{kl}(\delta, 1-\delta)$ can be interpreted as the expected proportion of times the pair (s,a) is explored under the algorithm. Taking the supremum over ω then corresponds to selecting an optimal sampling rule. In the following, ω is referred to as the allocation vector.

Remark. *In (Ok et al., 2018), the authors identify a similar optimization problem as (3) leading to a problem-specific regret lower bound satisfied by online learning algorithms in generic structured MDP. Interestingly, we note that this optimization problem is simpler than (3). This is due to fact that when minimizing regret, confusing MDPs have the same transitions and rewards as in the original MDP at optimal (state, action) pairs. This considerably simplifies the analysis, and explains why, in general, deriving problem-specific sample complexity lower bound is much harder than obtaining regret lower bounds.*

4.1. Properties of the problem (3)

We now provide useful properties of the optimization problem (3). Additional properties of the problem are presented in Appendix B.

(i) The set of alternative MDPs. To simplify the notation we use $\pi^\star$ instead of $\pi^\star_\phi$. Our first result concerns the set $\mathrm{Alt}(\phi)$ of *alternative* MDPs:

Lemma 2. $\mathrm{Alt}(\phi) = \bigcup_{s,a \neq \pi^\star(s)} \mathrm{Alt}_{sa}(\phi)$ *where*

$$\mathrm{Alt}_{sa}(\phi) = \{\psi : Q^{\pi^\star}_\psi(s,a) > V^{\pi^\star}_\psi(s)\}.$$

The above lemma states that an alternative MDP ψ is such that $\pi^\star$, the optimal policy of ϕ, can be improved under ψ locally at some state s, by selecting in s some previously sub-optimal action a, instead of $\pi^\star(s)$. Using this lemma, we can simplify the expression of the characteristic time appearing in Proposition 1. Indeed, (3) is equivalent to:

$$\sup_{\omega \in \Sigma} \min_{s, a \neq \pi^\star(s)} \inf_{\psi \in \mathrm{Alt}_{sa}(\phi)} \sum_{s',a'} \omega_{s',a'} \mathrm{KL}_{\phi|\psi}(s',a'). \quad (4)$$

Next, we rewrite the problem in an analytic manner. To this aim, we parametrize ψ by its transition probabilities and rewards $u = (q_\psi(s,a), p_\psi(s,a))_{s,a \in \mathcal{S}\times\mathcal{A}}$ and introduce the following notations: for all (s,a), $dr(s,a) = (r_\psi - r_\phi)(s,a)$ and $dp(s,a) = (p_\psi - p_\phi)(s,a)$. Further define $dV^{\pi^\star} = \left([V^{\pi^\star}_\psi - V^{\pi^\star}_\phi](s)\right)_{s\in\mathcal{S}}$.

Combining the condition : $Q^{\pi^\star}_\psi(s,a) > V^{\pi^\star}_\psi(s)$ with the fact that $Q^{\pi^\star}_\phi(s,a) + \Delta_{sa} = V^{\pi^\star}_\phi(s)$ we obtain that $\psi \in \mathrm{Alt}_{sa}(\phi)$ if and only if:

$$\begin{aligned}\Delta_{sa} < dr(s,a) + \gamma dp(s,a)^\top V^{\pi^\star}_\phi \\ + [\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top dV^{\pi^\star}. \end{aligned} \quad (5)$$

The above inequality states that for ψ to be in $\mathrm{Alt}_{sa}(\phi)$, the changes in the rewards and transitions between ϕ and ψ should be greater than the sub-optimality gap of action a in state s. Defining $\mathcal{U}_{sa} = \{u : (5) \text{ holds}\}$, we conclude that both the optimization problems (3) and (4) are equivalent to:

$$\sup_{\omega \in \Sigma} \min_{s, a \neq \pi^\star(s)} \inf_{u \in \mathcal{U}_{sa}} \sum_{s',a'} \omega_{s',a'} \mathrm{KL}_{\phi|\psi}(s',a'). \quad (6)$$

(ii) Non-convexity of the problem (3). The characteristic time $T^*(\phi)$, as well as the optimal sampling rule are characterized by the solution of (3) or that of (4). If we think of a track-and-stop algorithm to identify the best policy (as proposed in (Garivier & Kaufmann, 2016) for the simple MAB problem), one would need to repeatedly solve these optimization problems. It is then important to be able to do it in a computationally efficient way. Unfortunately, these problems are probably very hard to solve. This is well illustrated by the fact that the following sub-problem is not convex:

$$T(\phi,\omega)^{-1} = \inf_{\psi \in \mathrm{Alt}(\phi)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{\phi|\psi}(s,a). \quad (7)$$

Actually, in the example presented in Fig. 1, we can specify ϕ such that the sets $\mathrm{Alt}(\phi)$ and $\mathrm{Alt}_{sa}(\phi)$ are not convex.

Consider $\phi, \psi, \overline{\psi}$ belonging to the class of MDPs specified in Fig. 1, each defined by the vector (r_2, r_1, p_1) (all other parameters values are fixed as in the figure):

$$\begin{cases} \psi = (r_2 = 0.25,\ r_1 = 0.93,\ p_1 = 0.7) \\ \overline{\psi} = (r_2 = 0.1,\ r_1 = 0.47,\ p_1 = 0.6) \\ \phi = \frac{\psi + \overline{\psi}}{2} = (r_2 = 0.175,\ r_1 = 0.6925,\ p_1 = 0.65) \end{cases}$$

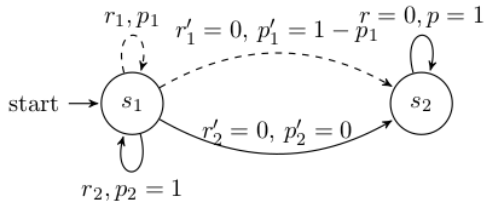

Figure 1. A class of two-state MDPs, with $\gamma = 0.9$. Actions a_1 and a_2 are available in state s_1. State s_2 is absorbing. Dashed (resp. full) arrows indicate the transitions when action a_1 (resp. a_2) is chosen. Numbers above each arrow indicate the transition probability and the average reward, e.g. $p'_2 = \mathbb{P}[s_2|s_1, a_2]$.

Then a simple calculation shows that the pair (s_1, a_1) is optimal : $\frac{r_1}{1-\gamma p_1} > \frac{r_2}{1-\gamma p_2}$ for both ψ and $\overline{\psi}$, while it is sub-optimal : $\frac{r_1}{1-\gamma p_1} < \frac{r_2}{1-\gamma p_2}$ for ϕ. In other words, both ψ and $\overline{\psi}$ are in $\mathrm{Alt}(\phi)$ and $\mathrm{Alt}_{s_1 a_1}(\phi)$ but their average is not: $\frac{\psi+\overline{\psi}}{2} = \phi \notin \mathrm{Alt}(\phi)$. Therefore the sets $\mathrm{Alt}(\phi)$ and $\mathrm{Alt}_{s_1 a_1}(\phi)$ are not convex. Observe that this non-convexity does not arise in simple MAB problems. Indeed, there, the set of parameters (e.g., the average reward vectors $\mu = (\mu_1, \ldots, \mu_K)$) such that a given arm is optimal is always convex, i.e., $\{\mu : \mu_k > \max_{j \neq k} \mu_j\}$ is convex.

Remark. *Note that if we have access to an optimization oracle that solves the problem (3) then we can simply apply the classical Track-and-Stop algorithm and achieve asymptotically optimal sample complexity. In the absence of such an oracle, we will devise an upper bound of $T^*(\phi)$, which we will use in our sampling rule as a proxy for the characteristic time.*

4.2. Upper bound of $T^*(\phi)$

We use the analytic version (6) of the optimization problem that defines the sample complexity lower bound to derive a simple (but still problem-specific) upper bound of the characteristic time $T^*(\phi)$. The upper bound actually corresponds to a sampling rule that is explicit, i.e., we do not need to solve any optimization problem to get it. Using this upper bound and the corresponding sampling rule, we will be able to devise a simple track-and-stop algorithm with provable performance guarantees. In addition, the upper bound has the right dependence in the sub-optimality gaps, and we also prove that it remains smaller than existing minimax sample complexity lower bounds.

Before we state the main result leading to our upper bound, we introduce additional notations.

• $\Delta_{\min} = \min_{s, a \neq \pi^\star(s)} \Delta_{sa}$ denotes the minimum sub-optimality gap in ϕ.

• $\mathrm{Var}_{p_\phi(s,a)}[V^\star_\phi] = \mathrm{Var}_{s' \sim p_\phi(.|s,a)}[V^\star_\phi(s')]$ is the variance of the next-state value after taking state-action pair (s, a). Similarly $\mathrm{Var}^\star_{\max}[V^\star_\phi] = \max_s \mathrm{Var}_{p_\phi(s,\pi^\star(s))}[V^\star_\phi]$ is the maximum variance of the next-state value after taking an optimal action.

• $\mathrm{sp}[V^\star_\phi] = \max_{s,s'} V^\star_\phi(s') - V^\star_\phi(s)$ is the span of the value function.

Theorem 1. *We have for all vectors ω in the simplex $T(\phi, \omega) \leq U(\phi, \omega)$ where,*

$$U(\phi, \omega) \triangleq \max_{s, a \neq \pi^\star(s)} \frac{T_1(s, a; \phi) + T_2(s, a; \phi)}{\omega_{sa}} + \frac{T_3(\phi) + T_4(\phi)}{\min_s \omega_{s, \pi^\star(s)}}, \tag{8}$$

and

$$T_1(s, a; \phi) \triangleq \frac{2}{\Delta^2_{sa}},$$

$$T_2(s, a; \phi) \triangleq \max\left(\frac{16 \mathrm{Var}_{p_\phi(s,a)}[V^\star_\phi]}{\Delta^2_{sa}}, \frac{6 \mathrm{sp}[V^\star_\phi]^{4/3}}{\Delta^{4/3}_{sa}}\right),$$

$$T_3(\phi) \triangleq \frac{2}{[\Delta_{\min}(1-\gamma)]^2},$$

and

$$T_4(\phi) \triangleq \min\left(\frac{27}{\Delta^2_{\min}(1-\gamma)^3}, \max\left(\frac{16 \mathrm{Var}^\star_{\max}[V^\star_\phi]}{\Delta^2_{\min}(1-\gamma)^2}, \frac{6 \mathrm{sp}[V^\star_\phi]^{4/3}}{\Delta^{4/3}_{\min}(1-\gamma)^{4/3}}\right)\right).$$

The proof of the theorem relies on writing each of the difference terms $dr(s, a)$, $dp(s, a)$, $dr^{\pi^\star}$ and $dp^{\pi^\star}$ involved in the constraint (5) as a proportion of the sub-optimality gap Δ_{sa}. Then, using classical f-divergences inequalities, as well as a variance inequality from (Azar et al., 2013), we relate each difference term to the KL divergences appearing in the objective function of the problem (6). With this perspective in mind, the terms $T_1(s, a; \phi)$ and $T_2(s, a; \phi)$ can be interpreted as the sample complexity costs to learn the reward of (state,action) pair (s, a) and the corresponding transition probabilities, respectively. Similarly, the terms $T_3(\phi)$ and $T_4(\phi)$ are interpreted as the sample complexity costs to estimate the future rewards collected from the next state and the transitions from the next state.

Corollary 1. *Let $H_{sa} \triangleq T_1(s, a; \phi) + T_2(s, a; \phi)$ and $H^\star \triangleq S(T_3(\phi) + T_4(\phi))$. Then the solution of the problem $\inf_{\omega \in \Sigma} U(\phi, \omega)$ is given by the unique allocation vector $\overline{\omega} \in \Sigma$ defined by ($\sim$ means proportional to): for all $s \in \mathcal{S}$,*

$$\begin{cases} \overline{\omega}_{s, \pi^\star(s)} \sim \frac{1}{S} \sqrt{H^\star (\sum_{s, a \neq \pi^\star(s)} H_{sa})}, \\ \overline{\omega}_{sa} \sim H_{sa}, \quad \text{for } s, a \neq \pi^\star(s). \end{cases} \tag{9}$$

This allocation yields the following upper bound:

$$T^*(\phi) \leq U(\phi) \triangleq 2(H^\star + \sum_{s,a \neq \pi^\star(s)} H_{sa}). \qquad (10)$$

In the previous corollary, $\overline{\omega}_{sa}$ is the optimal proportion of times (s,a) should be sampled, and hence for $s, a \neq \pi^\star(s)$, H_{sa} corresponds to the *hardness* of learning that (s,a) is sub-optimal. It scales as the inverse of the square of the gap Δ_{sa} and is proportional to the variance of future rewards after taking (s,a).

Further observe that since the rewards are normalized, we always have: for all (s,a), $\mathrm{Var}_{p_\phi(s,a)}[V_\phi^\star] \leq \frac{1}{(1-\gamma)^2}$ and $\mathrm{sp}[V_\phi^\star] \leq \frac{1}{(1-\gamma)}$. In addition, we show in Lemma 7 (see Appendix E) that $\Delta_{\min}$ is always smaller than 1. These observations allow us to upper bound $T_1(s,a;\phi)$, $T_2(s,a;\phi)$, $T_3(\phi)$ and $T_4(\phi)$, and to prove the following corollary.

Corollary 2. *We have:* $U(\phi) = \mathcal{O}\left(\frac{SA}{\Delta_{\min}^2(1-\gamma)^3}\right)$.

The above result is obtained by plugging the uniform allocation $\omega_{sa} = 1/SA$ in (8). Hence this naive uniform allocation yields an upper bound scaling as the known minimax sample complexity lower bound $\frac{SA}{\Delta_{\min}^2(1-\gamma)^3}$. This result also implies that a track-and-stop algorithm sampling the pairs (s,a) according to $\overline{\omega}$ will perform better than the minimax bound. This algorithm will become strictly better when $\mathrm{Var}^\star_{\max}[V_\phi^\star] = o(1/(1-\gamma))$, i.e., when the variance of the next-state value after taking the optimal action is small.

5. Algorithm

In this section, we present KLB-TS (KL-Ball Track-and-Stop), an algorithm that selects the successive (state, action) pairs so as to track the allocation $\overline{\omega}$, the problem-specific allocation (9) that leads to the upper bound (10). The algorithm is a track-and-stop, whose stopping rule does not follow a generic Generalized Likelihood Ratio Test as that used (Garivier & Kaufmann, 2016) for MAB problems (refer to Subsection 5.2 for detail).

The algorithm takes as input the confidence parameter δ and any black-box planner MDP-SOLVER. The latter takes as input an MDP ϕ, and returns an optimal policy $\pi_\phi^\star \in \Pi_\phi^\star$. For practical implementations, we use the Policy Iteration algorithm.

KLB-TS starts exploring each (state, action) pair once, to construct an initial estimate $\widehat{\phi}$ of the true MDP ϕ. The algorithm maintains, after t collected observations, an estimate $\widehat{\phi}_t$ of the true MDP. Based on this estimate, KLB-TS computes an estimate of the allocation $\overline{\omega}$, and selects the next (state, action) pair to track it. After each observation, the estimated MDP $\widehat{\phi}_t$ is updated. Finally, the algorithm checks if a stopping condition is satisfied, in which case the algorithm stops and returns the empirical optimal policy $\widehat{\pi}_\tau^\star$. The stopping condition is referred to as the *KL-Ball stopping rule* since it is inspired by the derivation of the upper bound of $T^*(\phi)$. There, the various terms involved in the exploration constraints are upper bounded by KL divergences, i.e., are in a KL ball.

The pseudo-code of KLB-TS is presented in Algorithm 1. Its sampling and stopping rule are described in detail in the next two sub-sections.

Algorithm 1 KLB-TS

input Black-box planner MDP-SOLVER(), Confidence parameter δ.
Collect one sample from each (s,a) in $\mathcal{S} \times \mathcal{A}$.
Set $t \leftarrow SA$ and $n_t(s,a) \leftarrow 1$, for all (s,a).
Initialize empirical estimate $\widehat{\phi}_t$ of ϕ.
$\widehat{\pi}_t^\star \leftarrow$ MDP-SOLVER$(\widehat{\phi}_t)$.
while Stopping condition (14) is not satisfied **do**
  Compute allocation vector $\overline{\omega}(\widehat{\phi}_t)$ of equation (9).
  Sample from (s_{t+1}, a_{t+1}) determined by equation (11).
  For all (s,a) set:

$$n_{t+1}(s,a) \leftarrow \begin{cases} n_t(s,a) + 1 \text{ if } (s,a) = (s_{t+1}, a_{t+1}) \\ n_t(s,a) \text{ Otherwise} \end{cases}$$

  $t \leftarrow t+1$.
  Update empirical estimate $\widehat{\phi}_t$ of ϕ.
  $\widehat{\pi}_t^\star \leftarrow$ MDP-SOLVER$(\widehat{\phi}_t)$.
end while
output Empirical optimal policy $\widehat{\pi}_\tau^\star$

5.1. Sampling rule

To build an algorithm with sample complexity matching the upper-bound of Corollary 1, the sampling proportions of (state,action) pairs should be as close as possible to the near-optimal weights defined in (9). To this aim, we simply use the C-tracking rule defined in (Garivier & Kaufmann, 2016), which we recall below.

Define $\overline{\omega}^\varepsilon(\phi)$ as the L^∞ projection of $\overline{\omega}(\phi)$ onto $\Sigma^\varepsilon = \{\omega \in [\varepsilon, 1]^{SA} : \sum_{s,a} \omega_{sa} = 1\}$. Further define $\varepsilon_t = (S^2A^2 + t)^{-1/2}/2$. Then the (state, action) pair to be sampled in round $t+1$ is defined as:

$$(s_{t+1}, a_{t+1}) \in \underset{(s,a)\in\mathcal{S}\times\mathcal{A}}{\arg\max} \sum_{k=1}^{t} \overline{\omega}_{sa}^{\varepsilon_k}(\widehat{\phi}_k) - n_t(s,a) \qquad (11)$$

with ties broken arbitrarily. The projection onto Σ^ε forces a minimal amount of exploration so that no pair is left under-

explored because of bad initial estimates. The same analysis of the sampling rule given in (Garivier & Kaufmann, 2016) holds in the MDP case and guarantees that:

$$\mathbb{P}_\phi\left(\forall (s,a)\in\mathcal{S}\times\mathcal{A},\quad \lim_{t\to\infty}\frac{n_t(s,a)}{t}=\overline{\omega}_{sa}(\phi)\right)=1.$$

5.2. Stopping rule

It is first worth noting that the proposed stopping condition constitutes the first stopping rule for best-policy identification in the MDP setting. Previous stopping rules in the literature are designed to identify ε-optimal policies. Unless we have access to an oracle that reveals the minimal gap between the best policy and a sub-optimal policy (in which case we can set ε smaller than this gap), we cannot identify the best-policy using these rules.

A *good* stopping rule determines when the set of samples collected so far is *just* enough to declare that $\widehat{\pi}_t^\star=\pi^\star$ with probability $1-\delta$. The design of our stopping rule is inspired by the proof of the upper-bound $U(\phi)$, which uses the following fact (refer to the inequalities (19)-(20)-(21)-(24)-(23) in the appendix): For all $\psi\in\mathrm{Alt}(\phi)$, there exists $s,a\neq\pi^\star(s)$ and a vector α in the simplex of $\mathbb{R}^4$ (which we denote Σ_4) such that the four following conditions are verified:

$$\begin{cases}\frac{\alpha_1^2}{T_1(s,a;\phi)}\leq \mathrm{kl}\left(r_\phi(s,a),r_\psi(s,a)\right),\\ \frac{\alpha_2^2}{T_2(s,a;\phi)}\leq KL\left(p_\phi(s,a)\|p_\psi(s,a)\right),\\ \frac{\alpha_3^2}{T_3(\phi)}\leq \max\limits_{s\in\mathcal{S}}\mathrm{kl}\left(r_\phi(s,\pi_\phi^\star(s)),r_\psi(s,\pi_\phi^\star(s))\right),\\ \frac{\alpha_4^2}{T_4(\phi)}\leq \max\limits_{s\in\mathcal{S}}KL\left(p_\phi(s,\pi_\phi^\star(s))\|p_\psi(s,\pi_\phi^\star(s))\right).\end{cases}\tag{12}$$

Then defining the quantities

$$\begin{cases}\rho_1(\phi,\psi)(s,a)=T_1(s,a;\phi)\mathrm{kl}\left(r_\phi(s,a),r_\psi(s,a)\right),\\ \rho_2(\phi,\psi)(s,a)=T_2(s,a;\phi)KL\left(p_\phi(s,a)\|p_\psi(s,a)\right),\\ \rho_3(\phi,\psi)=\max\limits_{s\in\mathcal{S}}T_3(\phi)\mathrm{kl}(r_\phi(s,\pi_\phi^\star(s)),r_\psi(s,\pi_\phi^\star(s))),\\ \rho_4(\phi,\psi)=\max\limits_{s\in\mathcal{S}}T_4(\phi)KL(p_\phi(s,\pi_\phi^\star(s))\|p_\psi(s,\pi_\phi^\star(s))),\end{cases}\tag{13}$$

(12) suggests that to design a PAC stopping condition, it is sufficient to check that the event

$$\mathcal{E}=\Big(\forall\alpha\in\Sigma_4\ \forall s,a\neq\widehat{\pi}_t^\star(s),\ \rho_1(\widehat{\phi}_t,\phi)(s,a)<\alpha_1^2\text{ or }\rho_2(\widehat{\phi}_t,\phi)(s,a)<\alpha_2^2\text{ or }\rho_3(\widehat{\phi}_t,\phi)<\alpha_3^2\text{ or }\rho_4(\widehat{\phi}_t,\phi)<\alpha_4^2\Big)$$

or equivalently[1]:

$$\mathcal{E}=\Big(\forall s,a\neq\widehat{\pi}_t^\star(s),\ \sqrt{\rho_1(\widehat{\phi}_t,\phi)(s,a)}+\sqrt{\rho_2(\widehat{\phi}_t,\phi)(s,a)}+\sqrt{\rho_3(\widehat{\phi}_t,\phi)}+\sqrt{\rho_4(\widehat{\phi}_t,\phi)}<1\Big)$$

[1]Hence the name KL-Ball stopping rule.

holds with probability $1-\delta$. Indeed, if $\mathcal{E}$ holds, then by contraposition of (12), we have $\phi\notin\mathrm{Alt}(\widehat{\phi}_t)$, which means that $\widehat{\pi}_t^\star=\pi^\star$. To define our stopping rule, we further introduce the threshold function:

$$x(\delta,n,m)=\log(1/\delta)+(m-1)[1+\log\left(1+n/(m-1)\right)].$$

We finally define $\widehat{T}_1(s,a)=T_1(s,a;\widehat{\phi}_t)$, $\widehat{T}_2(s,a)=T_2(s,a;\widehat{\phi}_t)$, $\widehat{T}_3=T_3(\widehat{\phi}_t)$, $\widehat{T}_4=T_4(\widehat{\phi}_t)$ and $\delta'=\frac{\delta}{4S^3A}$. The KL-Ball stopping condition, which guarantees that the event $\mathcal{E}$ above holds with probability $1-\delta$, is:

$$\max_{s,a\neq\widehat{\pi}_t^\star(s)}\frac{\sqrt{\widehat{T}_1(s,a)x(\delta',n_t(s,a),2)}+\sqrt{\widehat{T}_2(s,a)x(\delta',n_t(s,a),S)}}{\sqrt{n_t(s,a)}}+\max_{s\in\mathcal{S}}\frac{\sqrt{\widehat{T}_3x(\delta',n_t(s,\widehat{\pi}_t^\star(s)),2)}+\sqrt{\widehat{T}_4x(\delta',n_t(s,\widehat{\pi}_t^\star(s)),S)}}{\sqrt{n_t(s,\widehat{\pi}_t^\star(s))}}\leq 1\tag{14}$$

More precisely: $\tau_\delta=\inf\{t\in\mathbb{N}:(14)\text{ holds}\}$.

Theorem 2. *Under the KL-Ball stopping rule, we have:* $\mathbb{P}_\phi(\tau_\delta<\infty,\widehat{\pi}_{\tau_\delta}^\star\neq\pi_\phi^\star)\leq\delta$.

6. Sample Complexity Analysis

6.1. Main Results

Our main results take the form of asymptotic (when δ goes to 0) upper bounds on the sample complexity of KLB-TS. These bounds are proved as follows. First, the use of the C-tracking rule makes it possible to establish the convergence of the vector $(n_t(s,a))_{s,a}/t$ (the (state, action) pair visit frequencies) to the nearly-optimal allocation vector $\overline{\omega}$, as well as the convergence of the empirical MDP $\widehat{\phi}_t$ to the true MDP ϕ. Then, plugging these convergence results in the definition of the stopping rule (14), and combining the obtained results with the asymptotic shape of the threshold function $x(\delta',n,m)\underset{\delta\to0}{\sim}\log(1/\delta)$, we obtain (refer to Appendix G for a detailed description of these arguments):

$$\tau_\delta\underset{\delta\to0}{\sim}\inf\left\{t\in\mathbb{N}:\sqrt{\log(1/\delta)}\left(\max_{s,a\neq\pi^\star(s)}\frac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{t\times\overline{\omega}_{sa}}}+\max_{s\in\mathcal{S}}\frac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{t\times\overline{\omega}_{s,\pi^\star(s)}}}\right)\leq 1\right\}.$$

Finally, we show that the condition in the 'inf' above holds as soon as $t\geq 4U(\phi)\log(1/\delta)$ (see Lemma 11). The above arguments lead to an upper bound of the sample complexity of KLB-TS, valid almost surely (Proposition 2) and in expectation (Theorem 3).

Proposition 2. *The KL-Ball stopping rule, coupled with any sampling rule ensuring that for every state-action pair* (s,a), $n_t(s,a)/t$ *converges almost surely to the nearly-optimal allocations* $\overline{\omega}_{sa}$ *of Corollary 1, yields a sample complexity* τ_δ *satisfying for all* $\delta\in(0,1)$: $\mathbb{P}_\phi(\tau_\delta<\infty)=1$ *and* $\mathbb{P}_\phi\left(\limsup_{\delta\to0}\frac{\tau_\delta}{\log(1/\delta)}\leq 4U(\phi)\right)=1$.

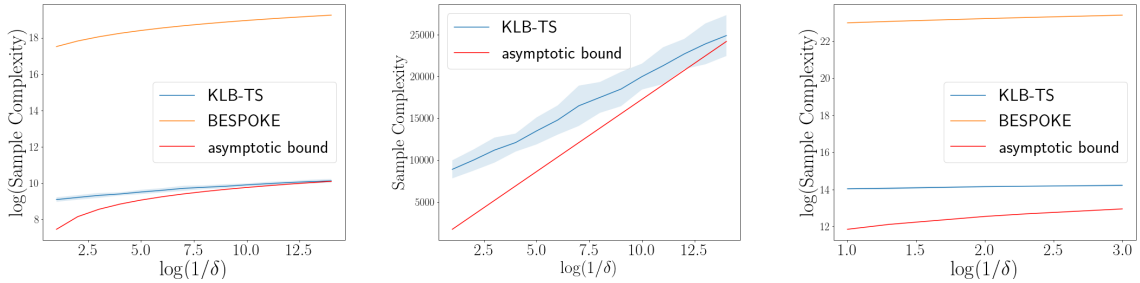

Figure 2. KLB-TS vs. BESPOKE. Left and center: S=A=2, $\gamma = 0.5$, right: $S = 5, A = 10, \gamma = 0.7$.

Theorem 3. *The KL-Ball stopping rule, coupled with the C-tracking rule defined in (11), yields a sample complexity τ_δ satisfying: for all $\delta \in (0,1)$, $\mathbb{E}_\phi[\tau_\delta]$ is finite and* $\limsup_{\delta \to 0} \frac{\mathbb{E}_\phi[\tau_\delta]}{\log(1/\delta)} \leq 4U(\phi)$.

The proof of the theorem above is similar to that of Theorem 14 in (Garivier & Kaufmann, 2016) with a few notable differences. First, we defined a distance on MDPs through the L^∞-norm of their reward and transition kernels. Then, we adapted Lemma 19 from (Garivier & Kaufmann, 2016), which gives a concentration inequality of the empirical average-rewards in the MAB setting, to include the concentration of transition probabilities of the empirical MDP.

6.2. Interpretation of the bound $U(\phi)$

The bound $U(\phi)$ is the sum of two interpretable components. The first term $U_1 = 2\sum_{s,a \neq \pi^\star(s)} H_{sa}$ represents the number of samples needed to estimate and identify sub-optimal pairs (s,a). It is proportional to the variance of the value function $\mathrm{Var}_{p(s,a)}[V_\phi^\star]$ and inversely proportional to the squared gap Δ_{sa}^2:

$$U_1 = \mathcal{O}\Big(\sum_{s,a \neq \pi^\star(s)} \max\Big\{ \frac{1 + \mathrm{Var}_{p(s,a)}[V_\phi^\star]}{\Delta_{sa}^2}, \frac{\mathrm{sp}[V_\phi^\star]^{4/3}}{\Delta_{sa}^{4/3}} \Big\} \Big).$$

The second term $2H^\star$ represents the samples needed to estimate and identify the optimal (state, action) pairs. It is proportional to $\mathrm{Var}^*_{max}[V_\phi^\star]$ the maximum variance of the value function across the trajectory of the optimal policy and inversely proportional to the squared minimum gap $\Delta_{\min}^2$ and to $(1-\gamma)^3$:

$$H^\star = \mathcal{O}\Big(\frac{S}{(1-\gamma)^2 \Delta_{\min}^2} + \min\Big\{ \frac{S}{(1-\gamma)^3 \Delta_{\min}^2}, \max\Big\{ \frac{S \mathrm{Var}^*_{max}[V_\phi^\star]}{(1-\gamma)^2 \Delta_{\min}^2}, \frac{S \mathrm{sp}[V_\phi^\star]^{4/3}}{(1-\gamma)^{4/3} \Delta_{\min}^{4/3}} \Big\} \Big\} \Big).$$

These dependencies were intuitively expected: The larger the variances are, the more samples we need to accurately estimate the value function. Similarly, the smaller the gaps are, the harder it is to distinguish optimal (state, action) pairs from sub-optimal ones. Finally, as γ gets closer to one, it is natural to require more samples, as a small error in estimating rewards or transitions can induce a large change in the total discounted reward, thereby modifying the optimal policy.

7. Experiments

In this section, we run numerical experiments to compare the performances of KLB-TS and BESPOKE (these are so far the two algorithms with problem-specific sample complexity guarantees). We refer the reader to Appendix H for a detailed description of the differences between KLB-TS and BESPOKE, as well as a comparison of their theoretical guarantees. To compare the two algorithms, we generated two MDPs randomly: a first small MDP with two states and two actions, and a second larger and more realistic MDP with five states and ten actions per state. We used BESPOKE with an accuracy parameter $\epsilon = 0.9\Delta_{\min}$ (note that $\Delta_{\min}$ is revealed to BESPOKE). For each value of the confidence level δ, we run 10 simulations for the first MDP under both algorithms. To save computation time in the case of the second MDP, we run 5 simulations for each δ and only compare KLB-TS's sample complexity with BESPOKE's initial number of samples $n_{\min}$ which, as noted in Appendix H, contributed for more than 99% of its sample complexity.

Figure 2 shows the mean sample complexity along with its 2-standard-deviations interval (which seems very small due to the use of a log-scale). The red curve (referred to as 'asymptotic bound') shows the upper bound $4U(\phi)\log(1/\delta)$ guaranteed by Theorem 3. Note that KLB-TS sample complexity is greater than $4U(\phi)\log(1/\delta)$ for moderate values of δ and only matches it for $\delta = 10^{-14}$. For both MDPs, KLB-TS clearly outperforms BESPOKE.

8. Conclusion

In this work, we have investigated the design of RL algorithms with *minimal problem-specific* sample complexity. To this aim, we first derived the information-theoretical sample complexity limit (a lower bound on the sample complexity satisfied by any algorithm) and the corresponding optimal sample allocation. Our hope was that, as for the MAB problem, this allocation would be easy to compute and could then lead to a simple and optimal track-and-stop algorithm. Unfortunately, for RL problems, it turns out that the optimal allocation solves an involved non-convex program. Approaching the fundamental sample complexity limit seems possible only if one could solve this program. To circumvent this issue, we derived a tight upper bound of the characteristic time. Remarkably, this bound corresponds to a sample allocation that is explicit, and hence can be easily plugged in into a track-and-stop algorithm. Based on this upper bound, we proposed KLB-TS, an algorithm whose sample complexity matches this upper bound.

This work opens up interesting research directions. First, the computational complexity of the sample complexity lower bound strongly suggests the existence of a fundamental trade-off between sample and computational complexities. Investigating this trade-off is intriguing. Then, we restricted our attention to the generative model, where one can sample any (state, action) pair at any step. In most practical cases however, one needs to learn an optimal policy by observing a single trajectory of the system. Hence, the numbers of times one observes the various (state, action) pairs are correlated, inducing some additional constraints in the optimization problem leading to the sample complexity lower bound. It is worth studying the impact of these navigation constraints on the sample complexity. Finally, we plan to extend our results to the framework of RL with function approximation.

Acknowledgments: The authors would like to thank the reviewers whose comments and questions helped improve the clarity of the paper.

References

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. volume 125 of *Proceedings of Machine Learning Research*, pp. 67–83. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/agarwal20b.html.

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

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In Feldman, V., Rakhlin, A., and Shamir, O. (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL http://proceedings.mlr.press/v49/garivier16a.html.

Jonsson, A., Kaufmann, E., Ménard, P., Domingues, O. D., Leurent, E., and Valko, M. Planning in markov decision processes with gap-dependent sample complexity. *arXiv preprint*, arXiv:2006.05879, 2020.

Kakade, S. M. *On the sample complexity of reinforcement learning*. PhD thesis, University of London, England, 2003.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Kearns, M. and Singh, S. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in Neural Information Processing*, 11, 04 1999.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–2, 1985.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint*, arXiv:2005.12900, 2020.

Ok, J., Proutiere, A., and Tranos, D. Exploration in structured reinforcement learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d693d554e0ede0d75f7d2873b015f228-Paper.pdf.

Reiss, R.-D. *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics.*, pp. 98–99. 1st edition, 1989. ISBN 9781461396208.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 5186–5196. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7765-near-optimal-time-and-sample-complexities-for-solving-markov-decision-processes-with-a-generative-model.pdf.

Zanette, A., Kochenderfer, M. J., and Brunskill, E. Almost horizon-free structure-aware best policy identification with a generative model. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 5625–5634. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8800-almost-horizon-free-structure-aware-best-policy-identification-with-a-generative-model.pdf.

A. Related work: The minimax approach

One of the first works on best-policy identification in discounted MDPs is (Kearns & Singh, 1999). There, the authors introduce a model referred to as *parallel sampling*, where the agent can sample transitions from all (state,action) pairs simultaneously (instead of following a trajectory determined by the MDP dynamics). They proposed Phased Q-Learning and showed that it requires $\widetilde{\mathcal{O}}\left(\frac{SA\log(SA/\delta)}{\varepsilon^2}\right)$ samples[2] to find an ε-optimal value function. Later on, (Kakade, 2003)(Chapter 2.5) proposed the generative model as a variant of the parallel sampling model. Both (Kearns & Singh, 1999) and (Kakade, 2003) proved upper-bounds on the sample complexity of model-based Q-Value-Iteration (QVI) by $\mathcal{O}\left(\frac{SA\log(SA/\delta)}{\varepsilon^2(1-\gamma)^4}\right)$. Using a variance trick, (Azar et al., 2013) improved their analysis and showed that when $\varepsilon \in (0, \frac{1}{\sqrt{(1-\gamma)S}}]$, both model-based QVI along and Policy Iteration (PI) can find an ε-optimal policy after collecting $\mathcal{O}\left(\frac{SA\log(SA/\delta)}{\varepsilon^2(1-\gamma)^3}\right)$ samples. They also proved that the latter quantity is the minimax lower bound of sample complexity required to find an ε-optimal policy. (Even-Dar et al., 2006) used Action-Elimination techniques from the Multi-Armed Bandit setting(MAB) to devise MAB-Phased-Q-Learning, an algorithm for MDPs with a generative model which finds an ε-optimal policy using $\widetilde{\mathcal{O}}(\frac{SAV_{\max}^2}{(1-\gamma)^5\varepsilon^2})$ samples, where $V_{\max}$ is the maximum range of the value function. (Sidford et al., 2018) proposed Variance-Reduced-Q-Value-Iteration (vQVI) which matches the minimax bound for a wider range of $\varepsilon \in (0, 1]$. The same bound was derived by (Agarwal et al., 2020) for $\varepsilon \in (0, \frac{1}{\sqrt{1-\gamma}}]$ using a model-based approach. Finally, (Li et al., 2020) used a reward perturbation technique to widen the set of ε where their algorithm is minimax optimal to the full range of accuracy levels: $(0, \frac{1}{1-\gamma}]$. It is worth noting that, except for (Even-Dar et al., 2006), the aforementioned papers only sample transitions and assume a reward function known in advance by the agent.

[2]Their analysis ignored the dependency on the horizon $H = \frac{1}{1-\gamma}$, treating γ as a constant.

B. Additional Proprerties of the lower bound program

Most alternative MDPs. We refer to an MDP $\psi \in \overline{\text{Alt}(\phi)}$[3] solving the problem (7) as *most alternative*, since for a given allocation ω, the sample complexity lower bound is determined by the number of samples needed to distinguish ϕ from ψ.

Observe that the condition (5) involves transition probabilities and rewards of the (state, action) pairs (s,a) and $(s', \pi^\star(s'))$ for all s', only. Hence $\psi \in \text{Alt}_{sa}(\phi)$ can be obtained from ϕ by changing at most the transition probabilities and rewards of these (state, action) pairs. Next, let $\psi \in \overline{\text{Alt}_{sa}(\phi)}$ solve (7). Then we can verify that the constraint (5) is active and that we have:

$$\Delta_{sa} = dr(s,a) + \gamma dp(s,a)^\top V_\phi^{\pi^\star} + [\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top dV^{\pi^\star}.$$

This means that to design a most alternative MDP, one should change the rewards and transitions of optimal (state, action) pairs and only one sub-optimal pair (s,a) and those changes should be just enough to fill sub-optimality gap Δ_{sa}. The next lemma formalizes these findings.

Lemma 3. *Denote by $\mathcal{O}(\phi) = \{(s,a) :\ Q_\phi^\star(s,a) = V_\phi^\star(s)\}$ the set of optimal (state,action) pairs in the MDP ϕ and let $\psi \in \overline{\text{Alt}(\phi)}$ solve (7). Then:*
(i) For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $(p_\psi(.|s,a), q_\psi(.|s,a)) \neq (p_\phi(.|s,a), q_\phi(.|s,a)) \implies (s,a) \in \mathcal{O}(\psi) \setminus \mathcal{O}(\phi)$ or $a = \pi^\star(s)$;
(ii) $\mathcal{O}(\phi) \subset \mathcal{O}(\psi)$.

Proof. First we recall the following facts which we will make use of.

Fact 1. $Q^\star$ is Liptschitz w.r.t rewards and transitions (by simple bounds on Bellman operator):

$$\left\|Q_\phi^\star - Q_\psi^\star\right\|_\infty \leq \left(1 + \frac{1}{1-\gamma}\right)\left(\|r_\phi - r_\psi\|_\infty + \frac{\gamma}{(1-\gamma)}\|p_\phi - p_\psi\|_{1,\infty}\right).$$

Fact 2. If we change only the kernels $(p_\phi(s,a), q_\phi(s,a)) \to (p_\psi(s,a), q_\psi(s,a))$ of some sub-optimal (state, action) pair $s, a \neq \pi^\star(s)$ and the action a doesn't become strictly optimal $(s,a) \notin \mathcal{O}(\psi)$, then the value function remains unchanged $V_\psi^\star = V_\phi^\star$.

This is because there exists $(\pi_1, \pi_2) \in \Pi_\phi^\star \times \Pi_\psi^\star$ such that $\pi_2(a|s) = \pi_1(a|s) = 0$ (where we recall that $\pi(a|s)$ denotes the probability that π selects a in state s) which implies:

$$\begin{cases} \left(P_\psi^{\pi_1}, r_\psi^{\pi_1}\right) = \left(P_\phi^{\pi_1}, r_\phi^{\pi_1}\right) \\ \left(P_\psi^{\pi_2}, r_\psi^{\pi_2}\right) = \left(P_\phi^{\pi_2}, r_\phi^{\pi_2}\right) \end{cases} \implies \begin{cases} V_\psi^\star \geq V_\psi^{\pi_1} = \left(I - \gamma P_\psi^{\pi_1}\right)^{-1} r_\psi^{\pi_1} = \left(I - \gamma P_\phi^{\pi_1}\right)^{-1} r_\phi^{\pi_1} = V_\phi^\star \\ V_\phi^\star \geq V_\phi^{\pi_2} = \left(I - \gamma P_\phi^{\pi_2}\right)^{-1} r_\phi^{\pi_2} = \left(I - \gamma P_\psi^{\pi_2}\right)^{-1} r_\psi^{\pi_1} = V_\psi^\star \end{cases}$$

Fact 3: We can restrict our attention to allocation vectors ω with zero-null entries: $\forall (s,a) \in \mathcal{S} \times \mathcal{A} :\ \omega_{sa} > 0$.

In fact, any allocation vector ω such that $\omega_{sa} = 0$ is suboptimal. Indeed, consider ψ obtained from ϕ by changing the kernels in (s,a) so that they become equal to the kernels in $(s, \pi^\star(s))$, while keeping everything else unchanged. Then by definition of ψ: $\sum_{s',a'} \omega_{s',a'} KL_{\phi|\psi}(s',a') = 0$. Furthermore one can easily show that $\psi \in \overline{\text{Alt}(\phi)}$ which implies that $K(\phi,\omega)^{-1} = 0$.

We are now ready to prove the lemma. Let $\psi \in \overline{\text{Alt}(\phi)}$ solving (7). We can write: $\psi = \lim_{n\to\infty} \psi_n$, where $(\psi_n)_{n\geq 1} \in \text{Alt}(\phi)^{\mathbb{N}}$ and $\lim_{n\to\infty} \sum_{s,a} \omega_{sa} KL_{\phi|\psi_n}(s,a) = \inf_{\psi \in \text{Alt}(\phi)} \sum_{s,a} \omega_{sa} KL_{\phi|\psi}(s,a)$. Therefore, by continuity of the KL function:

$$\sum_{s,a} \omega_{sa} KL_{\phi|\psi}(s,a) = \inf_{\psi \in \text{Alt}(\phi)} \sum_{s,a} \omega_{sa} KL_{\phi|\psi}(s,a) \tag{15}$$

[3] We use $\overline{E}$ to denote the closure of a set E.

Proof of (i): $(p_\psi(.|s,a), q_\psi(.|s,a)) \neq (p_\phi(.|s,a), q_\phi(.|s,a)) \implies (s,a) \in \mathcal{O}(\psi) \setminus \mathcal{O}(\phi)$ **or** $a = \pi^\star(s)$

By contradiction: Suppose there exists (s,a) such that: $(p_\psi(s,a), q_\psi(s,a)) \neq (p_\phi(s,a), q_\phi(s,a))$ and $(s,a) \in \mathcal{O}(\psi)^c \cup \mathcal{O}(\phi)$ and $a \neq \pi^\star(s)$. Combined together, the latter two conditions imply that:

$$(s,a) \in \mathcal{O}(\psi)^c. \tag{16}$$

We will use the following operator (ε-transform) where we move the rewards and transitions of ψ at (s,a) in the direction of ϕ by $\varepsilon \geq 0$: $T^{s,a}_{\phi,\varepsilon}(\psi) \triangleq \psi_\varepsilon$ where

$$(p_{\psi_\varepsilon}(s',a'), q_{\psi_\varepsilon}(s',a')) = \begin{cases} (1-\varepsilon)\,(p_\psi(s,a), q_\psi(s,a)) + \varepsilon\,(p_\phi(s,a), q_\phi(s,a))\,, & \text{if } (s',a') = (s,a), \\ (p_\psi(s',a'), q_\psi(s',a')) & \text{otherwise.} \end{cases} \tag{17}$$

Note that the objective function of the infimum problem takes a smaller value at ψ_ε than at ψ:

$$\begin{aligned} \sum_{s',a'} \omega_{s',a'} KL_{\phi|\psi_\varepsilon}(s',a') &\leq \left[(1-\varepsilon)\,\omega_{sa} KL_{\phi|\psi}(s,a) + \varepsilon\,\omega_{sa} KL_{\phi|\phi}(s,a)\right] + \sum_{(s',a')\neq(s,a)} \omega_{s',a'} KL_{\phi|\psi}(s',a') \\ &< \sum_{s',a'} \omega_{s',a'} KL_{\phi|\psi}(s',a') \end{aligned}$$

where the first inequality stems from the convexity of KL-function and the second from the property $p \neq q \implies KL(p\|q) > 0$. We will prove that there exists $\varepsilon > 0$ such that ψ_ε is the limit of a sequence of elements in $\mathrm{Alt}(\phi)$, which clearly contradicts the optimality of ψ (see equation 15).

Consider $a^\star$ an optimal action at state s in ψ, ie such $(s,a^\star) \in \mathcal{O}(\psi)$. Since $(s,a) \notin \mathcal{O}(\psi)$ (16), then for $\varepsilon = 0$, we have: $\psi_0 = \psi$ and $\delta \triangleq \delta_\psi(s,a) = Q^\star_\psi(s,a^\star) - Q^\star_\psi(s,a) > 0$. By continuity of $Q^\star$ w.r.t the rewards and transitions (**Fact 1**), there exists $\varepsilon > 0$ small enough such that:

$$Q^\star_{\psi_\varepsilon}(s,a^\star) - Q^\star_{\psi_\varepsilon}(s,a) > \delta/2 > 0.$$

Fix such ε and define $(\theta_n)_{n\geq 1} = \left(T^{s,a}_{\phi,\varepsilon}(\psi_n)\right)_{n\geq 1}$ where $(\psi_n)_{n\geq 1}$ is any sequence converging to ψ. By continuity of the operator $T^{s,a}_{\phi,\varepsilon}$, we have: $\lim_{n\to\infty} \theta_n = \psi_\varepsilon$. It remains to show that $(\theta_n)_{n\geq 1} \in \mathrm{Alt}(\phi)^{\mathbb{N}}$. Using the continuity of $Q^\star$ another time, we get:

$$\begin{aligned} \begin{cases} \lim_{n\to\infty} \psi_n = \psi \\ \lim_{n\to\infty} \theta_n = \psi_\varepsilon \end{cases} &\implies \begin{cases} \lim_{n\to\infty} Q^\star_{\psi_n}(s,a^\star) - Q^\star_{\psi_n}(s,a) = Q^\star_\psi(s,a^\star) - Q^\star_\psi(s,a) > \delta/2 \\ \lim_{n\to\infty} Q^\star_{\theta_n}(s,a^\star) - Q^\star_{\theta_n}(s,a) = Q^\star_{\psi_\varepsilon}(s,a^\star) - Q^\star_{\psi_\varepsilon}(s,a) > \delta/2 \end{cases} \\ &\implies \exists N_0 \in \mathbb{N} \quad \forall n \geq N_0 \quad \begin{cases} Q^\star_{\psi_n}(s,a^\star) - Q^\star_{\psi_n}(s,a) > \delta/2 \\ Q^\star_{\theta_n}(s,a^\star) - Q^\star_{\theta_n}(s,a) > \delta/2 \end{cases} \\ &\implies \forall n \geq N_0 \text{ (s,a) is sub-optimal in both } \psi_n \text{ and } \theta_n. \end{aligned}$$

This implies, by Fact 2 on ψ_n and θ_n, that: $\forall n \geq N_0\ V^\star_{\theta_n} = V^\star_{\psi_n}$. Since, we only changed kernels of ψ_n at (s,a) to obtain θ_n, then this also implies that for all $n \geq N_0$:

$$\begin{cases} \forall (s',a') \neq (s,a),\ Q^\star_{\psi_n}(s',a') = r_{\psi_n}(s',a') + \gamma p_{\psi_n}(s',a')^T V^\star_{\psi_n} = r_{\theta_n}(s',a') + \gamma p_{\theta_n}(s',a')^T V^\star_{\theta_n} = Q^\star_{\theta_n}(s',a') \\ \text{(s,a) is sub-optimal in both } \psi_n \text{ and } \theta_n \end{cases}$$

Therefore, $\forall n \geq N_0,\ \Pi^*_{\theta_n} = \Pi^*_{\psi_n}$, and consequently $\theta_n \in \mathrm{Alt}(\phi)$.
To sum up, modulo a reindexing of the sequence: $\exists (\theta_n)_{n\geq 1} \in \mathrm{Alt}(\phi)^{\mathbb{N}} : \lim_{n\to\infty} \theta_n = \psi_\varepsilon$. This is a contradiction.

Proof of (ii): $\mathcal{O}(\phi) \subset \mathcal{O}(\psi)$

We proceed in the same way, i.e., we suppose that there exists $(s,a) \in \mathcal{O}(\phi) \setminus \mathcal{O}(\psi)$. Only this time, we consider $\psi_\varepsilon \triangleq \prod_{s',a'} T^{s',a'}_{\phi,\varepsilon}(\psi)$ where the product sign stands for composition of operators. It's straightforward to show, using continuity of $Q^\star$ w.r.t rewards and transitions, that there exists $\varepsilon > 0$ such that (s,a) is still not optimal: $a \notin \mathcal{O}(\psi_\varepsilon)$. Hence $\psi_\varepsilon \in \mathrm{Alt}(\phi)$, which contradicts the optimality of ψ. $\square$

C. Lower Bound $T^*(\phi)$

C.1. Proof of Lemma 1

Proof. Let τ be a stopping time w.r.t. the filtration $(\mathcal{F}_t)_{t\geq 1}$. The observations made up to the beginning of round t are $\mathcal{O}_t = (s_1, a_1, R_1, s'_1 \ldots, s_t, a_t, R_t, s'_t)$. Let $p(\cdot)$ denote the distribution of the first state. We have:

$$\mathbb{P}_\phi(\mathcal{O}_t) = p(s_1) \prod_{k=1}^{t} p_\phi(s'_k|s_k, a_k) \times \prod_{k=1}^{t} q_\phi(R_k|s_k, a_k).$$

The log-likelihood ratio of the observations up to the end of round t under ϕ and ψ is then:

$$\begin{aligned} L_t &= \sum_{k=1}^{t} \left(\log \frac{p_\phi(s'_k|s_k, a_k)}{p_\psi(s'_k|s_k, a_k)} + \log \frac{q_\phi(R_k|s_s, a_k)}{q_\psi(R_k|s_k, a_k)} \right) \\ &= \sum_{s,a} L_t^{s,a}, \end{aligned}$$

where

$$L_t^{s,a} = \sum_{k=1}^{t} \mathbb{1}_{\{s_k=s, a_k=a\}} \left(\log \frac{p_\phi(s'_k|s, a)}{p_\psi(s'_k|s, a)} + \log \frac{q_\phi(R_k|s, a)}{q_\psi(R_k|s, a)} \right).$$

Next we study $L_t^{s,a}$ for a given pair (s, a). Introduce the following random variables: Y_k and Z_k denote the next state and the collected reward after the k-th time (s, a) has been visited. We can re-write $L_t^{s,a}$ as:

$$L_t^{s,a} = \sum_{k=1}^{N_t(s,a)} \left(\log \frac{p_\phi(Y_k|s, a)}{p_\psi(Y_k|s, a)} + \log \frac{q_\phi(Z_k|s, a)}{q_\psi(Z_k|s, a)} \right)$$

Observe that $\xi_k := \log \frac{p_\phi(Y_k|s,a)}{p_\psi(Y_k|s,a)} + \log \frac{q_\phi(Z_k|s,a)}{q_\psi(Z_k|s,a)}$ and $\mathbb{1}_{\{N_\tau(s,a)>k-1\}}$ are independent, because under the event $\{N_\tau(s, a) \leq k - 1\}$, Y_s and Z_s have not been observed yet. Further notice that $\mathbb{E}_\phi[\xi_k] = \mathrm{KL}_{\psi|\phi}(s, a)$. We deduce that:

$$\begin{aligned} \mathbb{E}_\phi[L_\tau^{s,a}] &= \mathbb{E}_\phi \left[\sum_{k=1}^{\infty} \xi_k \mathbb{1}_{\{N_\tau(s,a)>k-1\}} \right] \\ &= \sum_{k=1}^{\infty} \mathbb{P}_\phi[N_\tau(s, a) > k - 1] \, \mathrm{KL}_{\psi|\phi}(s, a) \\ &= \mathbb{E}_\phi[N_\tau(s, a)] \, \mathrm{KL}_{\psi|\phi}(s, a). \end{aligned}$$

Summing over all pairs (s, a) completes the proof. □

D. Main properties of the problem (3)

D.1. Proof of Lemma 2

Proof. To simplify the notation, we denote $\pi = \pi^\star_\phi$.

First part: $\mathrm{Alt}(\phi) \subset \bigcup_{s,a \neq \pi^\star(s)} \{\psi : Q^\pi_\psi(s,a) > V^\pi_\psi(s)\}$

By contradiction: Suppose there exists $\psi \in \mathrm{Alt}(\phi)$ such that $\forall s, a \neq \pi^\star(s),\ Q^\pi_\psi(s,a) \leq V^\pi_\psi(s)$. Since $Q^\pi_\psi(s,\pi(s)) = V^\pi_\psi(s)$ then the inequality is valid for all pairs:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ Q^\pi_\psi(s,a) \leq V^\pi_\psi(s)$$

Let $\pi^\star_\psi$ be an optimal policy under ψ. Then:

$$\forall s \in \mathcal{S},\ Q^\pi_\psi(s,,\pi^\star_\psi(s)) \leq V^\pi_\psi(s)$$

Define the Bellman operator of π under ψ as $\mathcal{B}^\pi_\psi : \mathbb{R}^S \to \mathbb{R}^S$ and for all $s \in \mathcal{S}$,

$$(\mathcal{B}^\pi_\psi V)(s) = r_\psi(s,\pi(s)) + \gamma p_\psi(s,\pi(s))^\top V.$$

Using the Bellman operator of the policy $\pi^\star_\psi$ under ψ, we rewrite the inequalities above:

$$\mathcal{B}^{\pi^\star_\psi}_\psi V^\pi_\psi \leq V^\pi_\psi.$$

By monotonicity of Bellman operator, this implies that: $\forall n \geq 1,\ \left(\mathcal{B}^{\pi^\star_\psi}_\psi\right)^n V^\pi_\psi \leq V^\pi_\psi$. Hence:

$$V^\star_\psi = \lim_{n\to\infty} \left(\mathcal{B}^{\pi^\star_\psi}_\psi\right)^n V^\pi_\psi \leq V^\pi_\psi,$$

i.e., the policy π is optimal under ψ. This is a contradiction.

Second part: $\bigcup_{s,a \neq \pi^\star(s)} \{\psi : Q^\pi_\psi(s,a) > V^\pi_\psi(s)\} \subset \mathrm{Alt}(\phi)$

By contradiction: Let $s, a \neq \pi^\star(s)$ and suppose there exists $\psi \in \{\psi : Q^\pi_\psi(s,a) > V^\pi_\psi(s)\}$ such that $\pi = \pi^\star_\phi$ is optimal under ψ. Define the modified policy π_1 as:

$$\pi_1(s') = \begin{cases} a & \text{if } s' = s, \\ \pi(s') & \text{otherwise.} \end{cases}$$

Then the fact that $Q^\pi_\psi(s,a) > V^\pi_\psi(s)$ translates to:

$$\mathcal{B}^{\pi_1}_\psi V^\star_\psi = \mathcal{B}^{\pi_1}_\psi V^\pi_\psi > V^\pi_\psi = V^\star_\psi$$

where the equality comes from the assumption that π is an optimal policy in ψ. Therefore, by monotonicity of Bellman operator, we have:

$$V^{\pi_1}_\psi = \lim_{n\to\infty} \left(\mathcal{B}^{\pi_1}_\psi\right)^n V^\star_\psi > V^\star_\psi.$$

We got a a contradiction. □

E. Upper bound $U(\phi)$ and the near-optimal sampling allocation $\overline{\omega}$

E.1. First technical lemma

We will need the following technical lemma which relates the change in the future discounted rewards between ϕ and ψ due to different transitions $dp(s,a)^\top V_\phi^\star$ to the Kullback-Leibler divergence of the transition kernels as well as the variance and maximum-deviation of the next-state value.

Lemma 4. *Using the notations of Sections 4.1 and 4.2, we have:*

$$|dp(s,a)^\top V_\phi^\star|^2 \leq 8KL(p_\phi(s,a)\|p_\psi(s,a))Var_{p_\phi(s,a)}[V_\phi^\star] + 4\sqrt{2}KL(p_\phi(s,a)\|p_\psi(s,a))^{3/2}sp[V_\phi^\star]^2.$$

Proof. We have:

$$\begin{aligned}
dp(s,a)^\top V_\phi^\star &= \sum_{s'} \left(p_\psi(s'|s,a) - p_\phi(s'|s,a)\right)\left[V_\phi^\star(s') - \mathbb{E}_{\widetilde{s}\sim p_\phi(.|s,a)}[V_\phi^\star(\widetilde{s})]\right] \\
&= \sum_{s'} \left(\sqrt{p_\psi(s'|s,a)} - \sqrt{p_\phi(s'|s,a)}\right) \\
&\qquad \times \left[\left(\sqrt{p_\psi(s'|s,a)} + \sqrt{p_\phi(s'|s,a)}\right)\left(V_\phi^\star(s') - \mathbb{E}_{\widetilde{s}\sim p_\phi(.|s,a)}[V_\phi^\star(\widetilde{s})]\right)\right].
\end{aligned}$$

Thus, by Cauchy-Schwartz inequality:

$$\begin{aligned}
|dp(s,a)^\top V_\phi^\star|^2 \leq& 2d_H(p_\phi(s'|s,a), p_\psi(s'|s,a))^2 \times \\
&\left[\sum_{s'}\left(\sqrt{p_\psi(s'|s,a)} + \sqrt{p_\phi(s'|s,a)}\right)^2\left(V_\phi^\star(s') - \mathbb{E}_{p_\phi(.|s,a)}[V_\phi^\star(\widetilde{s})]\right)^2\right] \\
&\leq 4d_H(p_\phi(s'|s,a), p_\psi(s'|s,a))^2\left[\sum_{s'}\left(p_\psi(s'|s,a) + p_\phi(s'|s,a)\right)\left(V_\phi^\star(s') - \mathbb{E}_{p_\phi(.|s,a)}[V_\phi^\star(\widetilde{s})]\right)^2\right],
\end{aligned}$$

where we have used $(a+b)^2 \leq 2(a^2+b^2)$ and $d_H(p,q) = \left[\frac{1}{2}\sum_i(\sqrt{p_i} - \sqrt{q_i})^2\right]^{1/2}$ is the Hellinger distance between two probability distributions. Therefore:

$$\begin{aligned}
|dp(s,a)^\top V_\phi^\star|^2 \leq& 4d_H(p_\phi(s'|s,a), p_\psi(s'|s,a))^2 \\
&\times \left[2\mathrm{Var}_{s'\sim p_\phi(.|s,a)}[V_\phi^\star(s')] + \|p_\phi(s'|s,a) - p_\psi(s'|s,a)\|_1 \,\mathrm{sp}[V_\phi^\star]^2\right].
\end{aligned}$$

We conclude the proof using Pinsker's inequality $\|p-q\|_1 \leq \sqrt{2KL(p\|q)}$ along with the inequality $d_H(p,q)^2 \leq KL(p\|q)$ (see (Reiss, 1989)). □

E.2. Proof of Theorem 1

Proof. Consider the simplified problem (6). Note that the constraint (5) doesn't involve the pairs $(\widetilde{s},\widetilde{a}) \in \mathcal{S}\times\mathcal{A} \setminus \{(s,a),(s',\pi^\star(s'))_{s'\in\mathcal{S}}\}$. One can easily show that any solution of the $\inf_{u\in\mathcal{U}_{sa}}$ part of (6) must satisfy $KL_{\phi|\psi}(\widetilde{s},\widetilde{a}) = 0$ for these unconstrained pairs $(\widetilde{s},\widetilde{a}) \in \mathcal{S}\times\mathcal{A} \setminus \{(s,a),(\widetilde{s},\pi^\star(\widetilde{s}))_{\widetilde{s}\in\mathcal{S}}\}$ (a trivial way to do it is by setting $\left(p_\psi(.|\widetilde{s},\widetilde{a}), q_\psi(.|\widetilde{s},\widetilde{a})\right) = \left(p_\phi(.|\widetilde{s},\widetilde{a}), q_\phi(.|\widetilde{s},\widetilde{a})\right)$). Therefore:

$$T(\phi,\omega)^{-1} = \min_{s,a\neq\pi^\star(s)} \inf_{u\in\mathcal{U}_{sa}} \omega_{sa}\mathrm{KL}_{\phi|\psi}(s,a) + \sum_{s'}\omega_{s',\pi_\phi^\star(s')}\mathrm{KL}_{\phi|\psi}(s',\pi_\phi^\star(s')). \tag{18}$$

We fix $s, a \neq \pi^\star(s)$ and derive a lower bound of $\inf_{u \in \mathcal{U}_{sa}} \omega_{sa} \mathrm{KL}_{\phi|\psi}(s,a) + \sum_{s'} \omega_{s',\pi^\star_\phi(s')} \mathrm{KL}_{\phi|\psi}(s', \pi^\star_\phi(s'))$. To do so, we rewrite the condition (5) by expanding the expression of $dV^{\pi^\star}$ as follows:

$$dr(s,a) + \gamma dp(s,a)^\top V^\star_\phi + [\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top \left(I - \gamma P^{\pi^\star}_\psi\right)^{-1} \left[r^{\pi^\star}_\psi - r^{\pi^\star}_\phi\right]$$
$$+ [\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top \left[\left(I - \gamma P^{\pi^\star}_\psi\right)^{-1} - \left(I - \gamma P^{\pi^\star}_\phi\right)^{-1}\right] r^{\pi^\star}_\phi > \Delta_{sa}.$$

We then write each of the four terms on the left-hand side as a "fraction" of Δ_{sa}:

$$\begin{cases} dr(s,a) = \alpha_1 \Delta_{sa} \\ dp(s,a)^\top V^\star_\phi = \alpha_2 \Delta_{sa} \\ [\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top \left(I - \gamma P^{\pi^\star}_\psi\right)^{-1} \left[r^{\pi^\star}_\psi - r^{\pi^\star}_\phi\right] = \alpha_3 \Delta_{sa} \\ [\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top \left[\left(I - \gamma P^{\pi^\star}_\psi\right)^{-1} - \left(I - \gamma P^{\pi^\star}_\phi\right)^{-1}\right] r^{\pi^\star}_\phi = \alpha_4 \Delta_{sa} \\ \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 > 1 \end{cases}$$

We use Pinsker's inequality and Lemma 4 to lower bound each term.

<u>1st **term.**</u> By Pinsker's inequality:

$$|dr(s,a)| = \left| \int_0^1 u[q_\psi(u|s,a) - q_\phi(u|s,a)] \lambda(du) \right| \leq \int_0^1 |q_\psi(u|s,a) - q_\phi(u|s,a)| \, \lambda(du)$$
$$\leq \sqrt{2KL(q_\phi(.|s,a) \| q_\psi(.|s,a))}.$$

Thus:

$$\boxed{\frac{1}{2}(\alpha_1 \Delta_{sa})^2 \leq KL(q_\phi(.|s,a) \| q_\psi(.|s,a))} \tag{19}$$

<u>2nd **term.**</u> By Lemma 4, we have:

$$(\alpha_2 \Delta_{sa})^2 \leq 8KL(p_\phi(s,a) \| p_\psi(s,a)) \mathrm{Var}_{s' \sim p_\phi(.|s,a)}[V^\star_\phi(s')] + 4\sqrt{2} KL(p_\phi(s,a) \| p_\psi(s,a))^{3/2} \mathrm{sp}[V^\star_\phi]^2.$$

Thus either:

$$\frac{1}{2}(\alpha_2 \Delta_{sa})^2 \leq 8KL(p_\phi(s,a) \| p_\psi(s,a)) \mathrm{Var}_{s' \sim p_\phi(.|s,a)}[V^\star_\phi(s')]$$

or

$$\frac{1}{2}(\alpha_2 \Delta_{sa})^2 \leq 4\sqrt{2} KL(p_\phi(s,a) \| p_\psi(s,a))^{3/2} \mathrm{sp}[V^\star_\phi]^2.$$

Therefore, we obtain:

$$\boxed{\min\left(\frac{\alpha_2^2 \Delta_{sa}^2}{16 \mathrm{Var}_{p_\phi(s,a)}[V^\star_\phi]}, \frac{\alpha_2^{4/3} \Delta_{sa}^{4/3}}{2^{7/3} \mathrm{sp}[V^\star_\phi]^{4/3}}\right) \leq KL\left(p_\phi(s,a) \| p_\psi(s,a)\right)} \tag{20}$$

<u>3rd **term.**</u> We have:

$$|\alpha_3| \Delta_{sa} = \left\| [\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top \left(I - \gamma P^{\pi^\star}_\psi\right)^{-1} \left[r^{\pi^\star}_\psi - r^{\pi^\star}_\phi\right] \right\|$$
$$\leq \|\gamma p_\psi(s,a) - \mathbb{1}(s)\|_\infty \times \left\| \left(I - \gamma P^{\pi^\star}_\psi\right)^{-1} \right\|_\infty \times \left\| r^{\pi^\star}_\psi - r^{\pi^\star}_\phi \right\|_\infty$$
$$\leq \frac{1}{1-\gamma} \left\| r^{\pi^\star}_\psi - r^{\pi^\star}_\phi \right\|_\infty,$$

which, following the same reasoning as the first term, implies:

$$\boxed{\frac{(\alpha_3\Delta_{sa}(1-\gamma))^2}{2} \leq \max_{s\in\mathcal{S}} KL\left(q_\phi(.|s,\pi^\star(s))\|q_\psi(.|s,\pi^\star(s))\right)} \tag{21}$$

4th term (first bound). We have:

$$|\alpha_4|\Delta_{sa} = \left\|[\gamma p_\psi(s,a) - \mathbb{1}(s)]^\top \left[\left(I-\gamma P_\psi^{\pi^\star}\right)^{-1} - \left(I-\gamma P_\phi^{\pi^\star}\right)^{-1}\right] r_\phi^{\pi^\star}\right\| \leq \|B\|_\infty, \tag{22}$$

where $B = \left[\left(I-\gamma P_\psi^{\pi^\star}\right)^{-1} - \left(I-\gamma P_\phi^{\pi^\star}\right)^{-1}\right] r_\phi^{\pi^\star}$. Hence:

$$\begin{aligned}|\alpha_4|\Delta_{sa} \leq \|B\|_\infty &= \gamma \left\|\left(I-\gamma P_\psi^{\pi^\star}\right)^{-1}\left[P_\psi^{\pi^\star} - P_\phi^{\pi^\star}\right] V_\phi^\star\right\|_\infty \\ &\leq \frac{\max_{s'\in\mathcal{S}} |dp(s',\pi^\star(s'))^\top V_\phi^\star|}{1-\gamma}.\end{aligned}$$

Therefore, applying Lemma 4, we get:

$$\boxed{\begin{aligned}\min\left(\frac{[\alpha_4\Delta_{sa}(1-\gamma)]^2}{16\text{Var}^\star_{max}[V_\phi^\star]}, \frac{\alpha_4^{4/3}\Delta_{sa}^{4/3}(1-\gamma)^{4/3}}{2^{7/3}\text{sp}[V_\phi^\star]^{4/3}}\right) \\ \leq \max_{s'\in\mathcal{S}} KL\left(p_\phi(s',\pi_\phi^\star(s'))\|p_\psi(s',\pi_\phi^\star(s'))\right)\end{aligned}} \tag{23}$$

4th term (second bound): We will now derive a second bound for the 4th term. Using Lemma 5, we get:

$$|\alpha_4|\Delta_{sa} \leq \|B\|_\infty \leq \frac{2^{5/2}\log(2)\text{KL}^{1/2}}{(1-\gamma)^{3/2}} + \frac{2^3\log(2)\gamma\text{KL}}{(1-\gamma)^{5/2}} + \frac{2^{5/4}\text{KL}^{3/4}\text{sp}[V_\phi^\star]}{1-\gamma}$$

where $\text{KL} = \max_{s\in\mathcal{S}} KL(p_\phi\left(s,\pi_\phi^\star(s))\|p_\psi(s,\pi_\phi^\star(s)\right))$. This means one of the three terms on the right-hand side is greater than $\frac{|\alpha_4|\Delta_{sa}}{3}$, which implies:

$$\boxed{\begin{aligned}\min\left(\frac{\alpha_4^2\Delta_{sa}^2(1-\gamma)^3}{288\log(2)^2}, \frac{|\alpha_4|\Delta_{sa}(1-\gamma)^{5/2}}{24\log(2)}, \frac{\alpha_4^{4/3}\Delta_{sa}^{4/3}(1-\gamma)^{4/3}}{2^{5/3}\times 3^{4/3}\text{sp}[V_\phi^\star]^{4/3}}\right) \\ \leq \max_{s\in\mathcal{S}} KL\left(p_\phi(s,\pi_\phi^\star(s))\|p_\psi(s,\pi_\phi^\star(s))\right)\end{aligned}} \tag{24}$$

Putting the individual lower bounds together: Summing up all inequalities from (19), (20), (21), (24) and (23), we deduce:

$$\inf_{\sum\alpha_i>1} \sum_{i=1}^3 B_i + \max(B_4,B_5) \leq \inf_{u\in\mathcal{U}_{sa}} \omega_{sa}\text{KL}_{\phi|\psi}(s,a) + \sum_{s'} \omega_{s',\pi_\phi^\star(s')}\text{KL}_{\phi|\psi}(s',\pi_\phi^\star(s'))$$

where

$$\begin{cases} B_1 = \frac{1}{2}\omega_{sa}(\alpha_1\Delta_{sa})^2 \\ B_2 = \omega_{sa}\min\left(\frac{\alpha_2^2\Delta_{sa}^2}{16\mathrm{Var}_{p_\phi(s,a)}[V_\phi^\star]}, \frac{\alpha_2^{4/3}\Delta_{sa}^{4/3}}{2^{7/3}\mathrm{sp}[V_\phi^\star]^{4/3}}\right) \\ B_3 = \frac{1}{2}\min_s \omega_{s,\pi^\star(s)}\ (\alpha_3\Delta_{sa}(1-\gamma))^2 \\ B_4 = \min_s \omega_{s,\pi^\star(s)}\ \min\left(\frac{\alpha_4^2\Delta_{sa}^2(1-\gamma)^3}{288\log(2)^2}, \frac{|\alpha_4|\Delta_{sa}(1-\gamma)^{5/2}}{24\log(2)}, \frac{\alpha_4^{4/3}\Delta_{sa}^{4/3}(1-\gamma)^{4/3}}{2^{5/3}\times 3^{4/3}\mathrm{sp}[V_\phi^\star]^{4/3}}\right) \\ B_5 = \min_s \omega_{s,\pi^\star(s)}\ \min\left(\frac{[\alpha_4\Delta_{sa}(1-\gamma)]^2}{16\mathrm{Var}^\star_{max}[V_\phi^\star]}, \frac{\alpha_4^{4/3}\Delta_{sa}^{4/3}(1-\gamma)^{4/3}}{2^{7/3}\mathrm{sp}[V_\phi^\star]^{4/3}}\right) \end{cases}$$

Notice that if α verifies the inequalities above, and $\sum_{i=1}^4 \alpha_i > 1$, then the vector whose entries are $\left(\frac{|\alpha_i|}{\sum_{j=1}^4 |\alpha_j|}\right)_{1\le i\le 4}$ also verifies these inequalities. Therefore we can restrict our attention to vectors α in the simplex Σ_4. In particular, we have $\alpha_i^2 \le \alpha_i^{4/3} \le \alpha_i$. Furthermore, we lower bound Δ_{sa} by $\Delta_{\min}$ in the terms $(B_j)_{3\le j\le 5}$. This simplifies the bound to:

$$\begin{aligned} \min_{s,a\neq\pi^\star(s)} \inf_{\alpha\in\Sigma_4} \sum_{i=1}^3 B_i'\alpha_i^2 + \max(B_4', B_5')\alpha_4^2 \ &\le \sup_{\omega\in\Sigma} \min_{s,a\neq\pi^\star(s)} \inf_{u\in\mathcal{U}_{sa}} \Bigg(\omega_{sa}\mathrm{KL}_{\phi|\psi}(s,a) \\ &\qquad + \sum_{s'} \omega_{s',\pi_\phi^\star(s')}\mathrm{KL}_{\phi|\psi}(s',\pi_\phi^\star(s'))\Bigg) \\ &= T(\phi,\omega)^{-1} \end{aligned} \tag{25}$$

where

$$\begin{cases} B_1' = \frac{1}{2}\omega_{sa}(\Delta_{sa}^2 \\ B_2' = \omega_{sa}\min\left(\frac{\Delta_{sa}^2}{16\mathrm{Var}_{p_\phi(s,a)}[V_\phi^\star]}, \frac{\Delta_{sa}^{4/3}}{2^{7/3}\mathrm{sp}[V_\phi^\star]^{4/3}}\right) \\ B_3' = \frac{1}{2}\min_s \omega_{s,\pi^\star(s)}\ (\Delta_{\min}(1-\gamma))^2 \\ B_4' = \min_s \omega_{s,\pi^\star(s)}\ \min\left(\frac{\Delta_{\min}^2(1-\gamma)^3}{288\log(2)^2}, \frac{\Delta_{\min}(1-\gamma)^{5/2}}{24\log(2)}, \frac{\Delta_{\min}^{4/3}(1-\gamma)^{4/3}}{2^{5/3}\times 3^{4/3}\mathrm{sp}[V_\phi^\star]^{4/3}}\right) \\ B_5' = \min_s \omega_{s,\pi^\star(s)}\ \min\left(\frac{\Delta_{\min}^2(1-\gamma)^2}{16\mathrm{Var}^\star_{max}[V_\phi^\star]}, \frac{\Delta_{\min}^{4/3}(1-\gamma)^{4/3}}{2^{7/3}\mathrm{sp}[V_\phi^\star]^{4/3}}\right) \end{cases}$$

Solving the left-hand side problem above in α, we get:

$$\min_{s,a\neq\pi^\star(s)} \left(\sum_{i=1}^3 \frac{1}{B_i'} + \min(\frac{1}{B_4'}, \frac{1}{B_5'})\right)^{-1} \le T(\phi,\omega)^{-1}.$$

Therefore:

$$T(\phi,\omega) \le \max_{s,a\neq\pi^\star(s)} \frac{T_1(s,a;\phi) + T_2(s,a;\phi)}{\omega_{sa}} + \frac{T_3(\phi) + T_4(\phi)}{\min_s \omega_{s,\pi^\star(s)}},$$

where

$$\begin{cases} T_1(s,a;\phi) = \frac{2}{\Delta_{sa}^2} \\ T_2(s,a;\phi) = \max\left(\frac{16\mathrm{Var}_{p_\phi(s,a)}[V_\phi^\star]}{\Delta_{sa}^2}, \frac{6\mathrm{sp}[V_\phi^\star]^{4/3}}{\Delta_{sa}^{4/3}}\right) \\ T_3(\phi) = \dfrac{2}{\Delta_{\min}^2(1-\gamma)^2} \\ T_4(\phi) = \min\left(V_1(\phi), V_2(\phi)\right), \end{cases}$$

and

$$V_1(\phi) = \max\left(\frac{27}{\Delta_{\min}^2(1-\gamma)^3}, \frac{8}{\Delta_{\min}(1-\gamma)^{5/2}}, \frac{14\mathrm{sp}[V_\phi^\star]^{4/3}}{\Delta_{\min}^{4/3}(1-\gamma)^{4/3}}\right),$$
$$V_2(\phi) = \max\left(\frac{16\mathrm{Var}_{\max}^\star[V_\phi^\star]}{\Delta_{\min}^2(1-\gamma)^2}, \frac{6\mathrm{sp}[V_\phi^\star]^{4/3}}{\Delta_{\min}^{4/3}(1-\gamma)^{4/3}}\right).$$

By Lemma 7, we always have $\Delta_{\min} \leq 1$. In addition $\mathrm{sp}[V_\phi^\star] \leq \frac{1}{1-\gamma}$, hence $V_1(\phi) = \frac{27}{\Delta_{\min}^2(1-\gamma)^3}$, which simplifies the expression of $T_4(\phi)$:

$$T_4(\phi) = \min\left(\frac{27}{\Delta_{\min}^2(1-\gamma)^3}, \max\left(\frac{16\mathrm{Var}_{\max}^\star[V_\phi^\star]}{\Delta_{\min}^2(1-\gamma)^2}, \frac{6\mathrm{sp}[V_\phi^\star]^{4/3}}{\Delta_{\min}^{4/3}(1-\gamma)^{4/3}}\right)\right).$$

□

E.3. Second technical lemma: Contributions of transitions at optimal pairs to the sample complexity

Lemma 5. *Define:*

$$B = \left[\left(I - \gamma P_\psi^{\pi^\star}\right)^{-1} - \left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\right] r_\phi^{\pi^\star} \quad \textit{and} \quad KL = \max_{s\in\mathcal{S}} KL\left(p_\phi(s,\pi^\star(s)) \| p_\psi(s,\pi^\star(s))\right).$$

Then we have:

$$\|B\|_\infty \leq \frac{2^{5/2}\log(2) KL^{1/2}}{(1-\gamma)^{3/2}} + \frac{2^3\log(2)\gamma KL}{(1-\gamma)^{5/2}} + \frac{2^{5/4} KL^{3/4} sp[V_\phi^\star]}{1-\gamma}. \tag{26}$$

Proof. Let us further develop the expression of B:

$$\begin{aligned} B &= \left[\left(I - \gamma P_\psi^{\pi^\star}\right)^{-1} - \left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\right] r_\phi^{\pi^\star} \\ &= \left(I - \gamma P_\psi^{\pi^\star}\right)^{-1}\left[\gamma P_\psi^{\pi^\star} - \gamma P_\phi^{\pi^\star}\right]\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1} r_\phi^{\pi^\star} \\ &= \gamma\left(I - \gamma P_\psi^{\pi^\star}\right)^{-1}\left[P_\psi^{\pi^\star} - P_\phi^{\pi^\star}\right] V_\phi^\star \\ &= \gamma\left[\left(I - \gamma P_\psi^{\pi^\star}\right)^{-1}\left(I - \gamma P_\phi^{\pi^\star}\right)\right]\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\left[P_\psi^{\pi^\star} - P_\phi^{\pi^\star}\right] V_\phi^\star \\ &\triangleq \gamma\, M_{\psi,\phi}\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\left[P_\psi^{\pi^\star} - P_\phi^{\pi^\star}\right] V_\phi^\star. \end{aligned} \tag{27}$$

Notice that the quantity $\gamma\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\left[P_\psi^{\pi^\star} - P_\phi^{\pi^\star}\right] V_\phi^\star$ is similar to the one that appears in Lemma 3 of (Azar et al., 2013), with ψ playing the role of $\widehat{\phi}$ in this case. We will try to relate it to the variances of the value function in the ϕ. Define:

$$\begin{cases} M_{\psi,\phi} = \left(I - \gamma P_{\psi}^{\pi^\star}\right)^{-1}\left(I - \gamma P_{\phi}^{\pi^\star}\right), \\ \mathrm{KL} = \max\limits_{s\in\mathcal{S}} KL\left(p_\phi(s,\pi^\star(s))\|p_\psi(s,\pi^\star(s))\right), \\ v^\pi(s) = \gamma^2 Var_{s'\sim p_\phi(.|s,\pi(s))}[V_\phi^\pi(s')], \\ \sigma^\pi(s) = \gamma^2 Var_{(s',a')\sim p_\phi(.|s,\pi(s))\otimes\pi(.|s')}[Q_\phi^\pi(s',a')]. \end{cases}$$

Using Lemma 4 and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we can write: $\forall s \in \mathcal{S}$,

$$\begin{aligned} \left|\gamma\left([P_\psi^{\pi^\star} - P_\phi^{\pi^\star}]V_\phi^\star\right)(s)\right| &= \left|\gamma dp(s,\pi^\star(s))^\top V_\phi^\star\right| \\ &\le \gamma\sqrt{8KL(p_\phi(s,\pi^\star(s))\|p_\psi(s,\pi^\star(s)))\ \mathrm{Var}_{s'\sim p_\phi(.|s,\pi^\star(s))}[V_\phi^\star(s')]} \\ &\quad + \gamma\sqrt{4\sqrt{2}KL(\ p_\phi(s,\pi^\star(s))\ ||\ p_\psi(s,\pi^\star(s))\ )^{3/2}\mathrm{MD}_{p_\phi(s,\pi^\star(s))}[V_\phi^\star]^2} \\ &\le 2^{3/2}\mathrm{KL}^{1/2}\sqrt{v^{\pi^\star}(s)} + 2^{5/4}\mathrm{KL}^{3/4}\mathrm{sp}[V_\phi^\star] \\ &\le 2^{3/2}\mathrm{KL}^{1/2}\sqrt{\sigma^{\pi^\star}(s)} + 2^{5/4}\mathrm{KL}^{3/4}\mathrm{sp}[V_\phi^\star], \end{aligned} \tag{28}$$

where the last inequality comes from Total Variance theorem:

$$\begin{aligned} \sigma^\pi(s) &= \gamma^2 Var_{(s',a')\sim p_\phi(.|s,\pi(s))\otimes\pi(.|s')}[Q_\phi^\pi(s',a')] \\ &= \gamma^2 Var_{s'\sim p_\phi(.|s,\pi(s))}\left[\mathbb{E}_{a'\sim\pi(.|s')}[Q_\phi^\pi(s',a')]\right] + \gamma^2\mathbb{E}_{s'\sim p_\phi(.|s,\pi(s))}\left[Var_{a'\sim\pi(.|s')}[Q_\phi^\pi(s',a')]\right] \\ &= v^\pi(s) + \gamma^2\mathbb{E}_{s'\sim p_\phi(.|s,\pi(s))}\left[Var_{a'\sim\pi(.|s')}[Q_\phi^\pi(s',a')]\right] \\ &\ge v^\pi(s). \end{aligned}$$

Denote $\sqrt{\sigma^{\pi^\star}} \triangleq \left(\sqrt{\sigma^{\pi^\star}(s)}\right)_{s\in\mathcal{S}}$. Then from (27) and (28), we deduce:

$$\begin{aligned} \|B\|_\infty &= \left\|M_{\psi,\phi}\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\gamma[P_\psi^{\pi^\star} - P_\phi^{\pi^\star}]V_\phi^\star\right\|_\infty \\ &\le \left\|M_{\psi,\phi}\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\left[2^{3/2}\mathrm{KL}^{1/2}\sqrt{\sigma^{\pi^\star}} + 2^{5/4}\mathrm{KL}^{3/4}\mathrm{sp}[V_\phi^\star]\mathbb{1}\right]\right\|_\infty \\ &\le 2^{3/2}\mathrm{KL}^{1/2}\left\|M_{\psi,\phi}\right\|_\infty\left\|\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\sqrt{\sigma^{\pi^\star}}\right\|_\infty + 2^{5/4}\mathrm{KL}^{3/4}\mathrm{sp}[V_\phi^\star]\left\|M_{\psi,\phi}\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\mathbb{1}\right\|_\infty \\ &= 2^{3/2}\mathrm{KL}^{1/2}\left\|M_{\psi,\phi}\right\|_\infty\left\|\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\sqrt{\sigma^{\pi^\star}}\right\|_\infty + 2^{5/4}\mathrm{KL}^{3/4}\mathrm{sp}[V_\phi^\star]\left\|\left(I - \gamma P_\psi^{\pi^\star}\right)^{-1}\mathbb{1}\right\|_\infty \\ &\le 2^{3/2}\mathrm{KL}^{1/2}\left\|M_{\psi,\phi}\right\|_\infty\left\|\left(I - \gamma P_\phi^{\pi^\star}\right)^{-1}\sqrt{\sigma^{\pi^\star}}\right\|_\infty + \frac{2^{5/4}}{1-\gamma}\mathrm{KL}^{3/4}\mathrm{sp}[V_\phi^\star]. \end{aligned} \tag{29}$$

Now observe that:

$$\begin{aligned} \left\|M_{\psi,\phi}\right\|_\infty &= \left\|\left(I - \gamma P_\psi^{\pi^\star}\right)^{-1}\left(I - \gamma P_\phi^{\pi^\star}\right)\right\|_\infty \\ &= \left\|I - \gamma\left(I - \gamma P_\psi^{\pi^\star}\right)^{-1}\left(P_\phi^{\pi^\star} - P_\psi^{\pi^\star}\right)\right\|_\infty \\ &\le 1 + \frac{\gamma\left\|P_\phi^{\pi^\star} - P_\psi^{\pi^\star}\right\|_\infty}{1-\gamma} \\ &\le 1 + \frac{\gamma(2\mathrm{KL})^{1/2}}{1-\gamma}, \end{aligned} \tag{30}$$

where the last inequality stems from Pinsker's inequality. Next we recall a variance inequality from (Azar et al., 2013):

Lemma 6. *(Lemma 8, (Azar et al., 2013))*

$$\left\| \left(I - \gamma P_\phi^{\pi^\star} \right)^{-1} \sqrt{\sigma^{\pi^\star}} \right\|_\infty \leq \frac{2\log(2)}{(1-\gamma)^{3/2}}.$$

Summing up equations (29), (30) and Lemma 6, we get:

$$\|B\|_\infty \leq \frac{2^{5/2}\log(2)\mathrm{KL}^{1/2}}{(1-\gamma)^{3/2}} + \frac{2^3\log(2)\gamma\mathrm{KL}}{(1-\gamma)^{5/2}} + \frac{2^{5/4}\mathrm{KL}^{3/4}\mathrm{sp}[V_\phi^\star]}{1-\gamma}. \tag{31}$$

□

E.4. Third technical lemma: The minimum gap is smaller than 1

Lemma 7. $\Delta_{\min} \leq 1$.

Proof. By contradiction, suppose $\Delta_{\min} > 1$, then:

$$\forall s, a \neq \pi^\star(s),\ \Delta_{sa} = V_\phi^\star(s) - Q_\phi^\star(s,a) > 1.$$

This means that for all policies $\pi \in \{\pi\, \forall s \in \mathcal{S},\ \pi(s) \neq \pi^\star(s)\}$, we have:

$$\forall s \in \mathcal{S},\ Q_\phi^\star(s, \pi(s)) < V_\phi^\star(s) - 1.$$

Using Bellman operator, the above inequality becomes:

$$\mathcal{B}_\phi^\pi V_\phi^\star < V_\phi^\star - \mathbb{1}.$$

By induction, using that the monotonicity of Bellman operator:

$$\forall n \geq 1,\ \left(\mathcal{B}_\phi^\pi\right)^n V_\phi^\star < V_\phi^\star - (\sum_{i=0}^{n-1}\gamma^i)\mathbb{1}.$$

Therefore:

$$\begin{aligned}
\forall \pi \in \{\pi\, \forall s \in \mathcal{S},\ \pi(s) \neq \pi^\star(s)\},\ V_\phi^\pi &= \lim_{n\to\infty} \left(\mathcal{B}_\phi^\pi\right)^{n+1} V_\phi^\star \\
&\leq \lim_{n\to\infty} \left(\mathcal{B}_\phi^\pi\right) \left[V_\phi^\star - (\sum_{i=0}^{n-1}\gamma^i)\mathbb{1} \right] \\
&= \left(\mathcal{B}_\phi^\pi\right) V_\phi^\star - \lim_{n\to\infty} (\sum_{i=1}^{n}\gamma^i)\mathbb{1} \\
&= \left(\mathcal{B}_\phi^\pi\right) V_\phi^\star - \frac{\gamma}{1-\gamma}\mathbb{1} \\
&< V_\phi^\star - \frac{1}{1-\gamma} \\
&< 0.
\end{aligned}$$

We obtained a contradiction. Thus, $\Delta_{\min} \leq 1$. □

E.5. Proof of Corollary1

Proof. The ω solving the problem in the right-hand side of (8) clearly verifies:

$$\forall s \in \mathcal{S},\ \ \omega_{s,\pi^\star(s)} = \min_{s'} \omega_{s',\pi^\star(s')} \triangleq \omega_0.$$

The problem of Theorem 1 then rewrites as:

$$\inf_{\omega_0} \ \max_{s,a\neq\pi^\star(s)} \frac{H_{sa}}{\omega_{sa}} + \frac{H^\star}{S\omega_0} \tag{32}$$

$$(\omega_{\widetilde{s},\widetilde{a}})_{\widetilde{s},\widetilde{a}\neq\pi^\star(\widetilde{s})} \tag{33}$$

where $H_{sa} = T_1(s,a;\phi) + T_2(s,a;\phi)$ and $H^\star = S(T_3(\phi) + T_4(\phi))$. We reformulate (33) as a convex program:

$$\begin{aligned}
&\inf_{t,\omega_0} \quad t + \frac{H^\star}{S\omega_0} \\
&(\omega_{sa})_{s,a\neq\pi^\star(s)} \\
&\text{s.t. } \omega^\top \mathbb{1} = 1, \\
&t \geq \frac{H_{sa}}{\omega_{sa}}, \forall s, a \neq \pi^\star(s)
\end{aligned}$$

Using KKT conditions, one can easily derive the expression of the solution:

$$\begin{cases}
\overline{\omega}_{s,a} = \dfrac{H_{sa}}{\sum\limits_{s,a\neq\pi^\star(s)} H_{sa} + \sqrt{H^\star\left(\sum\limits_{s,a\neq\pi^\star(s)} H_{sa}\right)}} & \forall s, a \neq \pi^\star(s), \\
\overline{\omega}_{s,\pi^\star(s)} = \frac{1}{S} \times \dfrac{\sqrt{H^\star\left(\sum\limits_{s,a\neq\pi^\star(s)} H_{sa}\right)}}{\sum\limits_{s,a\neq\pi^\star(s)} H_{sa} + \sqrt{H^\star\left(\sum\limits_{s,a\neq\pi^\star(s)} H_{sa}\right)}} & \forall s \in \mathcal{S}.
\end{cases} \tag{34}$$

The value V_P of the program is:

$$V_P = \sum_{s,a\neq\pi^\star(s)} H_{sa} + H^\star + 2\sqrt{H^\star\left(\sum_{s,a\neq\pi^\star(s)} H_{sa}\right)} \leq 2\left(\sum_{s,a\neq\pi^\star(s)} H_{sa} + H^\star\right) \triangleq U(\phi).$$

□

F. PAC Guarantee:

F.1. Proof of Theorem 2

First we recall two concentration inequalities and a technical lemma that we will be using. The first two lemmas are taken from (Jonsson et al., 2020). The third lemma is immediate.

Define the threshold function $x(n, \delta, m) = \log(1/\delta) + (m-1)\log\Big(e(1 + n/(m-1))\Big)$

Lemma 8. *(Proposition 2, (Jonsson et al., 2020)) For all distributions q of mean r supported on the unit interval, for all $\delta \in [0, 1]$:*

$$\mathbb{P}\Big(\exists n \in \mathbb{N}\ n\mathrm{kl}(\widehat{r}_n, r) > x(\delta, n, 2)\Big) \leq \delta.$$

Lemma 9. *(Proposition 1, (Jonsson et al., 2020)) Let P be a distribution over a finite set $\mathcal{S}$, and $(X_i)_{i\in\mathbb{N}}$ be i.i.d. variables with distribution P. For $s \in \mathcal{S}$, denote by $\widehat{P}_n = (\widehat{p}_n(s))_{s\in\mathcal{S}}$ the empirical estimate of P from the first n samples. Then for all $\delta \in [0, 1]$:*

$$\mathbb{P}\left(\exists n \in \mathbb{N}\ nKL(\widehat{P}_n \,||\, P) > x(\delta, n, S)\right) \leq \delta,$$

where we used S as a shorthand for $|\mathcal{S}|$.

Lemma 10. *Let $(\rho_i)_{1\leq i\leq 4} \in \mathbb{R}_+^4$. Then:*

$$\forall \alpha \in \Sigma_4\ \exists i \in [|0, 4|],\ \rho_i < \alpha_i^2 \iff \sum_{i=0}^{4} \sqrt{\rho_i} < 1.$$

We are now ready to prove Theorem 2 :

Proof. Recall the definition of the "correctness" event:

$$\mathcal{E}_t = \Big(\forall \alpha \in \Sigma_4\ \forall s, a \neq \widehat{\pi}_t^\star(s),\ \rho_1(\widehat{\phi}_t, \phi)(s, a) < \alpha_1^2 \text{ or } \rho_2(\widehat{\phi}_t, \phi)(s, a) < \alpha_2^2 \text{ or } \rho_3(\widehat{\phi}_t, \phi) < \alpha_3^2 \text{ or } \rho_4(\widehat{\phi}_t, \phi) < \alpha_4^2\Big)$$

where:

$$\begin{cases} \rho_1(\phi, \psi)(s, a) \triangleq T_1(s, a; \phi) KL(r_\phi(s, a)||r_\psi(s, a)), \\ \rho_2(\phi, \psi)(s, a) \triangleq T_2(s, a; \phi) KL(p_\phi(s, a)||p_\psi(s, a)), \\ \rho_3(\phi, \psi)(s) \triangleq T_3(\phi) KL\Big(r_\phi(s, \pi_\phi^\star(s)) \,||\, r_\psi(s, \pi_\phi^\star(s))\Big), \\ \rho_4(\phi, \psi)(s) \triangleq T_4(\phi) KL\Big(p_\phi(s, \pi_\phi^\star(s)) \,||\, p_\psi(s, \pi_\phi^\star(s))\Big), \\ \rho_3(\phi, \psi) \triangleq \max\limits_{s\in\mathcal{S}} \rho_3(\phi, \psi)(s), \\ \rho_4(\phi, \psi) \triangleq \max\limits_{s\in\mathcal{S}} \rho_4(\phi, \psi)(s). \end{cases}$$

Applying Lemma 10, we can simplify the event $\mathcal{E}_t$:

$$\mathcal{E}_t = \bigcap_{s, a\neq\widehat{\pi}_t^\star(s)} \left(\sqrt{\rho_1(\widehat{\phi}_t, \phi)(s, a)} + \sqrt{\rho_2(\widehat{\phi}_t, \phi)(s, a)} + \sqrt{\rho_3(\widehat{\phi}_t, \phi)} + \sqrt{\rho_4(\widehat{\phi}_t, \phi)} < 1\right) \tag{35}$$

$$= \bigcap_{s, a\neq\widehat{\pi}_t^\star(s)} \bigcap_{s', s"\in\mathcal{S}} \left(\sqrt{\rho_1(\widehat{\phi}_t, \phi)(s, a)} + \sqrt{\rho_2(\widehat{\phi}_t, \phi)(s, a)} + \sqrt{\rho_3(\widehat{\phi}_t, \phi)(s')} + \sqrt{\rho_4(\widehat{\phi}_t, \phi)(s")} < 1\right). \tag{36}$$

Define the stopping event:

$$\begin{aligned}\text{STOP}_t = &\left\{ \max_{s,a\neq\widehat{\pi}_t^\star(s)} \frac{\sqrt{\widehat{T_1}(s,a)x(\delta',n_t(s,a),2)}+\sqrt{\widehat{T_2}(s,a)x(\delta',n_t(s,a),S)}}{\sqrt{n_t(s,a)}} \right.\\ &\left. + \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_3}x(\delta',n_t(s,\widehat{\pi}_t^*(s)),2)}+\sqrt{\widehat{T_4}x(\delta',n_t(s,\widehat{\pi}_t^*(s)),S)}}{\sqrt{n_t(s,\widehat{\pi}_t^*(s))}} < 1 \right\}\\ = &\left\{ \max_{s,a\neq\widehat{\pi}_t^\star(s)} \frac{\sqrt{\widehat{T_1}(s,a)x(\delta',n_t(s,a),2)}+\sqrt{\widehat{T_2}(s,a)x(\delta',n_t(s,a),S)}}{\sqrt{n_t(s,a)}} \right.\\ &\left. + \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_3}x(\delta',n_t(s,\widehat{\pi}_t^*(s)),2)}}{\sqrt{n_t(s,\widehat{\pi}_t^*(s))}} + \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_4}x(\delta',n_t(s,\widehat{\pi}_t^*(s)),S)}}{\sqrt{n_t(s,\widehat{\pi}_t^*(s))}} < 1 \right\}\end{aligned} \tag{37}$$

where the last equality stems from the fact that both $n \to \frac{\sqrt{\widehat{T_3}x(\delta',n,2)}}{\sqrt{n}}$ and $n \to \frac{\sqrt{\widehat{T_4}x(\delta',n,S)}}{\sqrt{n}}$ are decreasing as soon as $n \geq 7(\mathcal{S}-1)$, therefore reaching their maximum at the same point. From the proof of Theorem 1 (refer to Equations (19)-(20)-(21)-(24)-(23)), we have the following "correctness' property:

$$\left(\phi \in \text{Alt}(\widehat{\phi}_t)\right) \subset \mathcal{E}_t^c, \tag{38}$$

where $\mathcal{E}_t^c$ stands for the complement of event $\mathcal{E}$. Therefore:

$$\begin{aligned}&(\tau_\delta < \infty) \cap (\widehat{\pi}^\star_{\tau_\delta} \neq \pi^\star) = \left(\exists t \geq 1,\ \text{STOP}_t \text{ and } \phi \in \text{Alt}(\widehat{\phi}_t)\right)\\ &\subset \left(\exists t \geq 1,\ \text{STOP}_t \cap \mathcal{E}_t^c\right)\\ &= \left(\exists t \geq 1,\ \bigcup_{s,a\neq\widehat{\pi}_t^\star(s)} \bigcup_{s',s"\in\mathcal{S}} \left(\left(\sqrt{\rho_1(\widehat{\phi}_t,\phi)(s,a)}+\sqrt{\rho_2(\widehat{\phi}_t,\phi)(s,a)}+\sqrt{\rho_3(\widehat{\phi}_t,\phi)(s')}+\sqrt{\rho_4(\widehat{\phi}_t,\phi)(s")} \geq 1\right)\right.\right.\\ &\qquad\qquad\left.\left. \cap\, \text{STOP}_t\right)\right)\\ &\subset \left(\exists t \geq 1,\ \bigcup_{s,a\neq\widehat{\pi}_t^\star(s)} \bigcup_{s',s"\in\mathcal{S}} \left(\mathcal{E}_{1,t}(s,a) \cup \mathcal{E}_{2,t}(s,a) \cup \mathcal{E}_{3,t}(s') \cup \mathcal{E}_{4,t}(s")\right)\right)\\ &\subset \bigcup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bigcup_{s',s"\in\mathcal{S}} \left(\left(\exists t \geq 1,\ \mathcal{E}_{1,t}(s,a) \cap \left(a = \widehat{\pi}_t^\star(s)\right)\right) \cup \left(\exists t \geq 1,\ \mathcal{E}_{2,t}(s,a) \cap \left(a = \widehat{\pi}_t^\star(s)\right)\right)\right.\\ &\qquad\left. \cup \left(\exists t \geq 1,\ \mathcal{E}_{3,t}(s')\right) \cup \left(\exists t \geq 1,\ \mathcal{E}_{4,t}(s")\right)\right),\end{aligned}$$

where

$$\begin{cases} \mathcal{E}_{1,t}(s,a) \triangleq \left\{ \sqrt{\rho_1(\widehat{\phi}_t,\phi)(s,a)} > \dfrac{\sqrt{\widehat{T_1}(s,a)x(\delta',n_t(s,a),2)}}{\sqrt{n_t(s,a)}} \right\}, \quad \forall (s,a) \notin \mathcal{O}(\widehat{\phi}_t), \\ \mathcal{E}_{2,t}(s,a) \triangleq \left\{ \sqrt{\rho_2(\widehat{\phi}_t,\phi)(s,a)} > \dfrac{\sqrt{\widehat{T_2}(s,a)x(\delta',n_t(s,a),S)}}{\sqrt{n_t(s,a)}} \right\}, \quad \forall (s,a) \notin \mathcal{O}(\widehat{\phi}_t), \\ \mathcal{E}_{3,t}(s) \triangleq \left\{ \sqrt{\rho_3(\widehat{\phi}_t,\phi)(s)} > \dfrac{\sqrt{\widehat{T}_3x(\delta',n_t(s,\widehat{\pi}_t^*(s)),2)}}{\sqrt{n_t(s,\widehat{\pi}_t^*(s))}} \right\}, \quad \forall s \in \mathcal{S}, \\ \mathcal{E}_{4,t}(s) \triangleq \left\{ \sqrt{\rho_4(\widehat{\phi}_t,\phi)(s)} > \dfrac{\sqrt{\widehat{T}_4x(\delta',n_t(s,\widehat{\pi}_t^*(s)),S)}}{\sqrt{n_t(s,\widehat{\pi}_t^*(s))}} \right\}, \quad \forall s \in \mathcal{S}. \end{cases}$$

Therefore:

$$\begin{aligned} \mathbb{P}_\phi(\tau_\delta < \infty, \widehat{\pi}^\star_{\tau_\delta} \neq \pi^\star_\phi) &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s',s"\in\mathcal{S}} \Big[\mathbb{P}\Big(\exists t \geq 1,\ \mathcal{E}_{1,t}(s,a) \cap \big(a = \widehat{\pi}^\star_t(s)\big) \Big) \\ &\quad + \mathbb{P}\Big(\exists t \geq 1,\ \mathcal{E}_{2,t}(s,a) \cap \big(a = \widehat{\pi}^\star_t(s)\big) \Big) + \mathbb{P}\Big(\exists t \geq 1,\ \mathcal{E}_{3,t}(s') \Big) + \mathbb{P}\Big(\exists t \geq 1,\ \mathcal{E}_{4,t}(s") \Big) \Big] \\ &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s',s"\in\mathcal{S}} 4\delta' \\ &= 4S^3A\delta' \triangleq \delta, \end{aligned}$$

where in the second inequality we have used the concentration inequalities (39), (40), (41) and (42). We detail the derivation of this second inequality below:

First term. Using Lemma 8, for $\delta' = \frac{\delta}{4S^3A}$, we have:

$$\begin{aligned} &\mathbb{P}\left(\exists t \geq 1,\ \sqrt{\rho_1(\widehat{\phi}_t,\phi)(s,a)} > \frac{\sqrt{\widehat{T_1}(s,a)x(\delta',n_t(s,a),2)}}{\sqrt{n_t(s,a)}} \right) \\ &= \mathbb{P}\Big(\exists t \geq 1,\ n_t(s,a)\mathrm{kl}(\widehat{r}_{n_t(s,a)}(s,a), r(s,a)) > x\left(\delta', n_t(s,a), 2\right) \Big) \\ &\leq \mathbb{P}\Big(\exists n \in \mathbb{N},\ n\mathrm{kl}(\widehat{r}_n(s,a), r(s,a)) > x\left(\delta', n, 2\right) \Big) \\ &\leq \delta'. \end{aligned} \tag{39}$$

Second term. Using Lemma 9, we get:

$$\begin{aligned} &\mathbb{P}\Big(\exists t \geq 1,\ \sqrt{\rho_2(\widehat{\phi}_t,\phi)(s,a)} > \frac{\sqrt{\widehat{T_2}(s,a)x(\delta',n_t(s,a),S)}}{\sqrt{n_t(s,a)}} \Big) \\ &= \mathbb{P}\Big(\exists t \geq 1,\ n_t(s,a)KL\left(\widehat{p}_{n_t(s,a)}(s,a) \| p(s,a)\right) > x\left(\delta', n_t(s,a), S\right) \Big) \\ &\leq \mathbb{P}\Big(\exists n \in \mathbb{N},\ KL\left(\widehat{p}_n(s,a) \| p(s,a)\right) > x\left(\delta', n, S\right) \Big) \\ &\leq \delta'. \end{aligned} \tag{40}$$

Third term. Following the same reasoning as in the first term we get:

$$\forall s \in \mathcal{S},\ \mathbb{P}\left(\exists t \geq 1,\ \sqrt{\rho_3(\widehat{\phi}_t, \phi)(s)} > \frac{\sqrt{\widehat{T}_{3,t} x(\delta', n_t(s, \widehat{\pi}_t(s)), 2)}}{\sqrt{n_t(s, \widehat{\pi}^*(s))}}\right) \leq \delta'. \tag{41}$$

Fourth term. Following the same reasoning as in the second term we get:

$$\forall s \in \mathcal{S},\ \mathbb{P}\left(\exists t \geq 1,\ \sqrt{\rho_4(\widehat{\phi}_t, \phi)(s)} > \frac{\sqrt{\widehat{T}_{4,t} x(\delta', n_t(s, \widehat{\pi}_t(s)), S)\Big)}}{\sqrt{n_t(s, \widehat{\pi}^*(s))}}\right) \leq \delta'. \tag{42}$$

□

G. Sample complexity of KLB-TS

In the following, we use the notation: $y(n,m) \triangleq (m-1) + (m-1)\log(1+n/(m-1))$. Hence the threshold function can be rewritten as: $x(\delta,n,m) = \log(1/\delta) + y(n,m)$.

We start this section by a technical lemma that is later used in the proof of Proposition 2 and Theorem 3.

Lemma 11. *For all ϕ in Φ,*

$$\left(\max_{s,a\neq\pi^\star(s)} \frac{\sqrt{T_1(s,a;\phi)} + \sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} + \max_{s\in\mathcal{S}} \frac{\sqrt{T_3(\phi)} + \sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2 \leq 4U(\phi).$$

Proof. Denote by LHS the left-hand side term above. Using $(A+B)^2 \leq 2(A^2+B^2)$ twice, and $(\max_x f(x))^2 = \max_x f(x)^2$ for non-negative f, we write:

$$\begin{aligned}
\text{LHS} &\leq 2\left(\left(\max_{s,a\neq\pi^\star(s)} \frac{\sqrt{T_1(s,a;\phi)} + \sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}}\right)^2 + \left(\max_{s\in\mathcal{S}} \frac{\sqrt{T_3(\phi)} + \sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2\right) \\
&= 2\left(\max_{s,a\neq\pi^\star(s)} \left(\frac{\sqrt{T_1(s,a;\phi)} + \sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}}\right)^2 + \max_{s\in\mathcal{S}} \left(\frac{\sqrt{T_3(\phi)} + \sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2\right) \\
&\leq 4\left(\max_{s,a\neq\pi^\star(s)} \frac{T_1(s,a;\phi) + T_2(s,a;\phi)}{\overline{\omega}_{s,a}} + \max_{s\in\mathcal{S}} \frac{T_3(\phi) + T_4(\phi)}{\overline{\omega}_{s,\pi^\star(s)}}\right) \\
&\leq 4U(\phi),
\end{aligned}$$

where the last inequality comes from Corollary 1. □

G.1. Proof of Proposition 2

Proof. Recall the stopping condition:

$$\begin{aligned}
\tau_\delta = \inf\Bigg\{ t\in\mathbb{N} \,:\, &\max_{s,a\neq\widehat{\pi}_t^\star(s)} \frac{\sqrt{\widehat{T_1}(s,a)x(\delta',n_t(s,a),2)} + \sqrt{\widehat{T_2}(s,a)x(\delta',n_t(s,a),S)}}{\sqrt{n_t(s,a)}} \\
&+ \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_3}x(\delta',n_t(s,\widehat{\pi}_t^\star(s)),2)} + \sqrt{\widehat{T_4}x\Big(\delta',n_t(s,\widehat{\pi}_t^\star(s)),S\Big)}}{\sqrt{n_t(s,\widehat{\pi}_t^\star(s))}} \leq 1\Bigg\}.
\end{aligned}$$

First we derive a convenient upper-bound of the left-hand-side term of the inequality above (which we denote by LHS_t). Rewrite the definition of $x(\delta,n,m) = \log(1/\delta) + (m-1) + (m-1)\log(1+n/(m-1)) \triangleq \log(1/\delta) + y(n,m)$. Then, using the fact that $\sqrt{A+B} \leq \sqrt{A} + \sqrt{B}$, we have:

$$\begin{aligned}
\text{LHS}_t \leq &\sqrt{\log(\delta')}\left(\max_{s,a\neq\widehat{\pi}_t^\star(s)} \frac{\sqrt{\widehat{T_1}(s,a)} + \sqrt{\widehat{T_2}(s,a)}}{\sqrt{n_t(s,a)}} + \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_3}} + \sqrt{\widehat{T_4}}}{\sqrt{n_t(s,\widehat{\pi}_t^\star(s))}}\right) \\
&+ \max_{s,a\neq\widehat{\pi}_t^\star(s)} \frac{\sqrt{\widehat{T_1}(s,a)y(n_t(s,a),2)} + \sqrt{\widehat{T_2}(s,a)y(n_t(s,a),S)}}{\sqrt{n_t(s,a)}} \\
&+ \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_3}y(n_t(s,\widehat{\pi}_t^\star(s)),2)} + \sqrt{\widehat{T_4}y(n_t(s,\widehat{\pi}_t^\star(s)),S)}}{\sqrt{n_t(s,\widehat{\pi}_t^\star(s))}} \qquad (43) \\
&\triangleq \sqrt{\log(\delta')}\left(\max_{s,a\neq\widehat{\pi}_t^\star(s)} \frac{\sqrt{\widehat{T_1}(s,a)} + \sqrt{\widehat{T_2}(s,a)}}{\sqrt{n_t(s,a)}} + \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_3}} + \sqrt{\widehat{T_4}}}{\sqrt{n_t(s,\widehat{\pi}_t^\star(s))}}\right) + f(n_t,\widehat{\phi}_t),
\end{aligned}$$

where $n_t = (n_t(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ denotes the number of visits vector. Note that when the terms $(\widehat{T}_i)_{1\le i\le 4}$ are bounded and $\lim_{t\to\infty} n_t(s,a) = \infty$, which we will soon establish, then we have $\lim_{t\to\infty} f(n_t, \widehat{\phi}_t) = 0$.

Next define the convergence event:

$$\mathcal{C} = \Big\{\forall (s,a)\in\mathcal{S}\times\mathcal{A}, \lim_{t\to\infty}\frac{n_t(s,a)}{t} = \overline{\omega}_{s,a},\ \widehat{\phi}_t \to \phi\Big\}.$$

Then by assumptions of the theorem and since $\forall (s,a),\ \overline{\omega}_{s,a} > 0$, we have $\lim_{t\to\infty} n_t(s,a) = \infty$ which implies $\mathbb{P}_\phi(\mathcal{C}) = 1$. Under $\mathcal{C}$, by continuity of the involved functionals of the MDP, we have:

$$\forall \varepsilon > 0,\ \exists t_1(\varepsilon)\in\mathbb{N},\ \forall t \ge t_1 : \begin{cases} \widehat{\pi}_t^\star = \pi^*, \text{ as soon as } \left\|Q^\star_{\widehat{\phi}_t} - Q^\star_\phi\right\|_\infty < \Delta_{\min}/2, \\ \widehat{T}_{1,t}(s,a) < (1+\varepsilon)T_1(s,a), \quad \forall s,a \ne \pi^\star(s), \\ \widehat{T}_{2,t}(s,a) < (1+\varepsilon)T_2(s,a), \quad \forall s,a \ne \pi^\star(s), \\ \widehat{T}_{3,t} \le (1+\varepsilon)T_3, \\ \widehat{T}_{4,t} \le (1+\varepsilon)T_4, \\ n_t(s,a)/t \ge (1-\varepsilon)\overline{\omega}_{s,a},\ \forall s,a \ne \pi^\star(s), \\ n_t(s,\widehat{\pi}_t^\star(s))/t \ge (1-\varepsilon)\overline{\omega}_{s,\pi^*(s)},\ \forall s\in\mathcal{S}, \\ f(n_t,\widehat{\phi}_t) \le \varepsilon. \end{cases}$$

Thus when $t \ge t_1(\varepsilon)$, inequality (43) implies:

$$\text{LHS}_t \le \sqrt{\frac{(1+\varepsilon)\log(\delta')}{(1-\varepsilon)t}}\left(\max_{s,a\ne\pi^\star(s)}\frac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} + \max_{s\in\mathcal{S}}\frac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right) + \varepsilon. \tag{44}$$

Next we define :

$$\begin{aligned} t_2(\delta,\varepsilon) = \inf\Bigg\{ t>0 \;\Bigg|\; &\sqrt{\frac{(1+\varepsilon)\log(\delta')}{(1-\varepsilon)t}}\Bigg(\max_{s,a\ne\pi^\star(s)}\frac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} \\ &+ \max_{s\in\mathcal{S}}\frac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\Bigg) \le 1-\varepsilon\Bigg\} \\ = \frac{(1+\varepsilon)\log(\delta')}{(1-\varepsilon)^3}&\left(\max_{s,a\ne\pi^\star(s)}\frac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} + \max_{s\in\mathcal{S}}\frac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2. \end{aligned} \tag{45}$$

Combining (44) and (45), we have for $t \ge \max(t_1(\varepsilon), t_2(\delta,\varepsilon))$, $LHS_t \le 1$. Therefore:

$$\begin{aligned} \tau_\delta &\le \max\left(t_1(\varepsilon), t_2(\varepsilon,\delta)\right) \\ &= \max\left(t_1(\varepsilon), \frac{(1+\varepsilon)\log(\delta')}{(1-\varepsilon)^3}\left(\max_{s,a\ne\pi^\star(s)}\frac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} + \max_{s\in\mathcal{S}}\frac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2\right). \end{aligned}$$

Thus $\forall\delta\in(0,1),\ \tau_\delta$ is finite on $\mathcal{C}$ and we have:

$$\forall\varepsilon>0,\ \limsup_{\delta\to 0}\frac{\tau_\delta}{\log(1/\delta)} \le \frac{1+\varepsilon}{(1-\varepsilon)^3}\left(\max_{s,a\ne\pi^\star(s)}\frac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} + \max_{s\in\mathcal{S}}\frac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2.$$

Taking the limit when $\varepsilon\to 0$, we get:

$$\limsup_{\delta\to 0}\frac{\tau_\delta}{\log(1/\delta)} \le \left(\max_{s,a\ne\pi^\star(s)}\frac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} + \max_{s\in\mathcal{S}}\frac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2.$$

We conclude by applying Lemma 11. □

G.2. Proof of Theorem 3

For a kernel u in $\mathbb{R}^{S\times SA}$, we define the norm $\|u\|_{1,\infty} \triangleq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} |u(s'|s,a)|$. Next, we define the following distance on MDPs:

$$\|\psi - \phi\| = \max_{s,a} \left(\|q_\psi(.|s,a) - q_\phi(.|s,a)\|_1 \ \vee \ \|p_\psi(.|s,a) - p_\phi(.|s,a)\|_1 \right).$$

Based on this distance, we can define balls on the set of MDPs:

$$\mathcal{B}_{\|.\|}(\phi,\xi) \triangleq \{\psi \ : \ \|\psi - \phi\| \leq \xi\}.$$

Let $\varepsilon > 0$. By recursively bounding Bellman operator, one can prove that $Q^\star$ is Liptschitz w.r.t. rewards and transitions:

$$\begin{aligned}\left\|Q^\star_\phi - Q^\star_\psi\right\|_\infty &\leq \left(1 + \frac{1}{1-\gamma}\right)\left(\|r_\phi - r_\psi\|_\infty + \frac{\gamma}{(1-\gamma)}\|p_\phi - p_\psi\|_{1,\infty}\right) \\ &\leq \left(1 + \frac{1}{1-\gamma}\right)\left(\max_{s,a}\|q_\psi(.|s,a) - q_\phi(.|s,a)\|_1 + \frac{\gamma}{(1-\gamma)}\|p_\phi - p_\psi\|_{1,\infty}\right).\end{aligned}$$

Thus, there exists $\xi = \xi(\varepsilon) > 0$ such that:

$$\forall \psi \in \mathcal{B}_{\|.\|}(\phi,\xi), \quad \left\|Q^\star_\phi - Q^\star_\psi\right\|_\infty < \Delta_{\min}/2 \ \text{ and } \ \max_{s,a}|\overline{\omega}_{s,a}(\psi) - \overline{\omega}_{s,a}(\phi)| \leq \varepsilon.$$

Crucially, the first inequality implies that $\pi^\star_\psi = \pi^\star_\phi$. For $T \in \mathbb{N}$, consider the concentration event:

$$\mathcal{E}_T = \bigcap_{t=T^{1/4}}^{T} \left(\widehat{\phi}_t \in \mathcal{B}_{\|.\|}(\phi,\xi)\right).$$

We will be using the following technical lemmas. The first corresponds to Lemma 20 in (Garivier & Kaufmann, 2016), which we reformulate in our case by replacing the number of arms of the bandit by the number of (state, action) pairs of the MDP.

Lemma 12. *There exists a constant T_ε such that for $T \geq T_\varepsilon$, it holds on $\mathcal{E}_T$, for C-Tracking:*

$$\forall t \geq T_\varepsilon, \ \max_{s,a}\left|\frac{n_t(s,a)}{t} - \overline{\omega}_{s,a}\right| \leq 3(SA-1)\varepsilon.$$

The second lemma is a concentration inequality similar to that of Lemma 19 in (Garivier & Kaufmann, 2016) (we defer its proof to the end of this appendix).

Lemma 13. *Denote by $\mathcal{E}_T^c$ the complementary of the event $\mathcal{E}_T$. There exists two constants B, C (that depend on ϕ and ε) such that:*

$$\forall T \geq 1, \mathbb{P}\left(\mathcal{E}_T^c\right) \leq BT\exp(-CT^{1/8}).$$

Recall inequality (43), which gives an upper bound of the left-hand-side of the stopping condition:

$$\text{LHS}_t \leq \sqrt{\log(\delta')}\left(\max_{s,a\neq\widehat{\pi}^\star_t(s)} \frac{\sqrt{\widehat{T_1}(s,a)} + \sqrt{\widehat{T_2}(s,a)}}{\sqrt{n_t(s,a)}} + \max_{s\in\mathcal{S}} \frac{\sqrt{\widehat{T_3}} + \sqrt{\widehat{T_4}}}{\sqrt{n_t(s,\widehat{\pi}^\star_t(s))}}\right) + f(n_t,\widehat{\phi}_t)$$

where $f(.,.)$ is a continuous function in both arguments. Define:

$$\begin{cases} D(\phi,\varepsilon) = \sup_{\substack{\psi \in \mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon)) \\ \|\omega' - \omega(\phi)\| \leq 3(SA-1)\varepsilon}} \max_{s,a \neq \pi^\star(s)} \frac{\sqrt{T_1(s,a;\psi)} + \sqrt{T_2(s,a;\psi)}}{\sqrt{\omega'_{sa}}}, \\ E(\phi,\varepsilon) = \sup_{\substack{\psi \in \mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon)) \\ \|\omega' - \omega(\phi)\| \leq 3(SA-1)\varepsilon}} \max_{s \in \mathcal{S}} \frac{\sqrt{T_3(\psi)} + \sqrt{T_4(\psi)}}{\sqrt{\omega'_{s,\pi^\star(s)}}}, \\ F(\phi,\varepsilon,t) = \sup_{\substack{\psi \in \mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon)) \\ \|\omega' - \omega(\phi)\| \leq 3(SA-1)\varepsilon}} f(t \times \omega', \psi). \end{cases}$$

For $T \geq T_\varepsilon$, on the event $\mathcal{E}_T$, we have: $\forall t \geq T^{1/4}, \quad \widehat{\pi}^\star_t = \pi^\star$, and using Lemma 12, $\left\| \frac{n_t(s,a)}{t} - \overline{\omega}_{s,a} \right\|_\infty \leq 3(SA-1)\varepsilon$. Therefore, for the stopping condition $\text{LHS}_t \leq 1$ to be satisfied, it is sufficient to have:

$$\frac{\sqrt{\log(\delta')}}{\sqrt{t}} \Big(D(\phi,\varepsilon) + E(\phi,\varepsilon) \Big) + F(\phi,\varepsilon,t) \leq 1. \tag{46}$$

By Lemma 14, $\lim_{t\to\infty} F(\phi,\varepsilon,t) = 0$. Hence, we can define the following times :

$$\begin{cases} t_1(\phi,\varepsilon,\eta,\delta) = \inf \left\{ t > 0 \mid \forall x > t, \quad \frac{\sqrt{\log(\delta')}}{\sqrt{x}} \Big(D(\phi,\varepsilon) + E(\phi,\varepsilon) \Big) \leq 1 - \eta \right\} \\ \qquad\qquad = \frac{\log(\delta') \Big(D(\phi,\varepsilon) + E(\phi,\varepsilon) \Big)^2}{(1-\eta)^2} \\ t_2(\phi,\varepsilon,\eta) = \inf \left\{ t > 0 \mid \forall x > t, \quad F(\phi,\varepsilon,t) \leq \eta \right\}. \end{cases},$$

It is easy to see that for $T \geq \max(T_\varepsilon, t_1, t_2)$, condition (46) is verified and consequently: $\tau_\delta \leq T$. In other words, we just proved that:

$$\forall T \geq \max(T_\varepsilon, t_1, t_2), \quad \mathcal{E}_T \subset (\tau_\delta \leq T).$$

Therefore:

$$\begin{aligned} \mathbb{E}_\phi[\tau_\delta] &= \sum_{T=1}^{\infty} \mathbb{P}(\tau_\delta > T) \\ &\leq \sum_{T=1}^{\max(T_\varepsilon,t_1,t_2)} 1 + \sum_{T=\max(T_\varepsilon,t_1,t_2)}^{\infty} \mathbb{P}(\mathcal{E}_T^c) \\ &\leq T_\varepsilon + t_1(\phi,\varepsilon,\eta,\delta) + t_2(\phi,\varepsilon,\eta) + \sum_{T=1}^{\infty} BT \exp(-CT^{1/8}), \end{aligned}$$

where the last inequality comes from Lemma 13. Thus, $\mathbb{E}[\tau_\delta]$ is finite and we have:

$$\limsup_{\delta\to 0}\frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \le \limsup_{\delta\to 0}\frac{t_1(\phi,\varepsilon,\eta,\delta)}{\log(1/\delta)} = \frac{\Big(D(\phi,\varepsilon)+E(\phi,\varepsilon)\Big)^2}{(1-\eta)^2}.$$

Letting η and ε go to zero, and noting that:

$$\begin{cases} \lim_{\varepsilon\to 0} D(\phi,\varepsilon) = \max_{s,a\neq\pi^\star(s)} \dfrac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}}, \\ \lim_{\varepsilon\to 0} E(\phi,\varepsilon) = \max_{s\in\mathcal{S}} \dfrac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}, \\ \left(\max_{s,a\neq\pi^\star(s)} \dfrac{\sqrt{T_1(s,a;\phi)}+\sqrt{T_2(s,a;\phi)}}{\sqrt{\overline{\omega}_{s,a}}} + \max_{s\in\mathcal{S}} \dfrac{\sqrt{T_3(\phi)}+\sqrt{T_4(\phi)}}{\sqrt{\overline{\omega}_{s,\pi^\star(s)}}}\right)^2 \le 4U(\phi), \ \text{(Lemma 11)}, \end{cases}$$

we get the desired result.

G.3. Second technical lemma

Lemma 14. *Let $\pi^\star=\pi^\star_\phi$ and let $y(n,m)=(m-1)+(m-1)\log(1+n/(m-1))$. Define:*

$$f(n,\psi) = \max_{s,a\neq\pi^\star(s)} \frac{\sqrt{T_1(s,a;\psi)y(n(s,a),2)}+\sqrt{T_2(s,a;\psi)y(n(s,a),S)}}{\sqrt{n(s,a)}} + \max_{s\in\mathcal{S}} \frac{\sqrt{T_3(\psi)y(n(s,\pi^\star(s)),2)}+\sqrt{T_4(\psi)y(n(s,\pi^\star(s)),S)}}{\sqrt{n(s,\pi^\star(s))}}$$

and

$$F(\phi,\varepsilon,t) = \sup_{\substack{\psi\in\mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon)) \\ \|\omega'-\omega(\phi)\|\le 3(SA-1)\varepsilon}} f(t\times\omega',\psi).$$

Then, there exists ε_0 such that: $\forall\varepsilon\le\varepsilon_0$, $\lim_{t\to\infty} F(\phi,\varepsilon,t)=0$.

Proof. Define:

$$\begin{cases} T_1(s,a,\phi,\varepsilon) \triangleq \sup_{\psi\in\mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon))} T_1(s,a;\psi), \\ T_2(s,a,\phi,\varepsilon) \triangleq \sup_{\psi\in\mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon))} T_2(s,a;\psi), \\ T_3(\phi,\varepsilon) \triangleq \sup_{\psi\in\mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon))} T_3(\psi), \\ T_4(\phi,\varepsilon) \triangleq \sup_{\psi\in\mathcal{B}_{\|.\|}(\phi,\xi(\varepsilon))} T_4(\psi). \end{cases}$$

By continuity of the functionals $(T_i)_{1\le i\le 4}$ in ϕ, there exists $\varepsilon_0>0$, such that for all $\varepsilon\le\varepsilon_0$, the supremums defined above are upper bounded by $M = 2\times\max_{s,a\neq\pi^\star(s)}(T_1(s,a;\phi),T_2(s,a;\phi),T_3(\phi),T_4(\phi))$. Furthermore, if $\|\omega'-\omega(\phi)\|\le 3(SA-1)\varepsilon$, then for all (s,a): $\omega_{sa}(\phi)-3(SA-1)\varepsilon\le\omega'_{sa}\le\omega_{sa}(\phi)+3(SA-1)\varepsilon$. Summing up these inequalities we get, for ε small enough:

$$F(\phi,\varepsilon,t) \le \sqrt{M}\max_{s,a\neq\pi^\star(s)} \frac{\sqrt{y(t[\omega_{sa}(\phi)+3(SA-1)\varepsilon],2)}+\sqrt{y(t[\omega_{sa}(\phi)+3(SA-1)\varepsilon],S)}}{\sqrt{t[\omega_{sa}(\phi)-3(SA-1)\varepsilon]}} + \max_{s\in\mathcal{S}} \frac{\sqrt{y(t[\omega_{s,\pi^\star(s)}(\phi)+3(SA-1)\varepsilon],2)}+\sqrt{y(t[\omega_{s,\pi^\star(s)}(\phi)+3(SA-1)\varepsilon],S)}}{\sqrt{t[\omega_{s,\pi^\star(s)}(\phi)-3(SA-1)\varepsilon]}}. \tag{47}$$

Since $\forall a > 0\ \forall m \geq 2,\ \lim_{x\to\infty} \frac{\sqrt{y(ax,m)}}{\sqrt{x}} = \lim_{x\to\infty} \frac{\sqrt{(m-1)+(m-1)\log(1+ax/(m-1))}}{\sqrt{x}} = 0$, and the maximums in (47) are taken over finite sets, then $\lim_{t\to\infty} F(\phi, \varepsilon, t) = 0$. □

G.4. Proof of Lemma 13

Proof. We have:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_T^c\right) &\leq \sum_{t=T^{1/4}}^{T} \mathbb{P}\left(\widehat{\phi}_t \notin \mathcal{B}_{\|.\|}(\phi, \xi)\right) \\
&\leq \sum_{t=T^{1/4}}^{T} \sum_{s,a} \Bigg[\mathbb{P}\Big(\widehat{r}_t(s,a) - r(s,a) > \xi\Big) + \mathbb{P}\Big(\widehat{r}_t(s,a) - r(s,a) < -\xi\Big) \\
&\qquad + \sum_{s'} \mathbb{P}\Big(\widehat{p}_t(s'|s,a) - p(s'|s,a) > \xi/S\Big) + \mathbb{P}\Big(\widehat{p}_t(s'|s,a) - p(s'|s,a) < -\xi/S\Big)\Bigg].
\end{aligned}$$

Let T be such that $T^{1/4} \geq (SA)^2$. Then for $t \geq T^{1/4}$, we have $\forall (s,a), \quad n_t(s,a) \geq (\sqrt{t} - SA/2)_+ - 1 \geq \sqrt{t} - SA$. Therefore, using a union bound and Chernoff inequality, one can write:

$$\begin{aligned}
\mathbb{P}\Big(\widehat{p}_t(s'|s,a) - p(s'|s,a) > \xi/S\Big) &= \mathbb{P}\Big(\widehat{p}_t(s'|s,a) - p(s'|s,a) > \xi/S,\ n_t(s,a) \geq \sqrt{t} - S\Big) \\
&\leq \sum_{t'=\sqrt{t}-SA}^{t} \mathbb{P}\Big(\widehat{p}_t(s'|s,a) - p(s'|s,a) > \xi/S,\ n_t(s,a) = t'\Big) \\
&\leq \sum_{t'=\sqrt{t}-SA}^{t} \exp\Big(-t' \cdot \mathrm{kl}\big(p(s'|s,a) + \xi/S,\ p(s'|s,a)\big)\Big) \\
&\leq \frac{\exp\Big(-(\sqrt{t} - SA)\mathrm{kl}\big(p(s'|s,a) + \xi/S,\ p(s'|s,a)\big)\Big)}{1 - \exp\Big(-\mathrm{kl}\big(p(s'|s,a) + \xi/S,\ p(s'|s,a)\big)\Big)}.
\end{aligned}$$

Using the same reasoning, we can prove that:

$$\begin{cases}
\mathbb{P}\Big(\widehat{p}_t(s'|s,a) - p(s'|s,a) < -\xi/S\Big) \leq \dfrac{\exp\Big(-(\sqrt{t} - SA)\mathrm{kl}\big(p(s'|s,a) - \xi/S,\ p(s'|s,a)\big)\Big)}{1 - \exp\Big(-\mathrm{kl}\big(p(s'|s,a) - \xi/S,\ p(s'|s,a)\big)\Big)}, \\
\mathbb{P}\Big(\widehat{r}_t(s,a) - r(s,a) > \xi\Big) \leq \dfrac{\exp\Big(-(\sqrt{t} - SA)\mathrm{kl}\big(r(s,a) + \xi,\ r(s,a)\big)\Big)}{1 - \exp\Big(-\mathrm{kl}\big(r(s,a) + \xi,\ r(s,a)\big)\Big)}, \\
\mathbb{P}\Big(\widehat{r}_t(s,a) - r(s,a) < -\xi\Big) \leq \dfrac{\exp\Big(-(\sqrt{t} - SA)\mathrm{kl}\big(r(s,a) - \xi,\ r(s,a)\big)\Big)}{1 - \exp\Big(-\mathrm{kl}\big(r(s,a) - \xi,\ r(s,a)\big)\Big)}.
\end{cases}$$

Thus, for the following choice of constants

$$\begin{aligned}
C = \min_{s,a}\Bigg(&\mathrm{kl}\big(r(s,a) - \xi,\ r(s,a)\big) \ \wedge\ \mathrm{kl}\big(r(s,a) + \xi,\ r(s,a)\big) \\
&\wedge\ \min_{s'}\Big(\mathrm{kl}\big(p(s'|s,a) - \xi/S,\ p(s'|s,a)\big) \ \wedge\ \mathrm{kl}\big(p(s'|s,a) + \xi/S,\ p(s'|s,a)\big)\Big)\Bigg)
\end{aligned}$$

and

$$B = \sum_{s,a} \left(\frac{\exp\Big(SA \cdot \mathrm{kl}\big(r(s,a)+\xi,\ r(s,a)\big)\Big)}{1-\exp\Big(-\mathrm{kl}\big(r(s,a)+\xi,\ r(s,a)\big)\Big)} + \frac{\exp\Big(SA \cdot \mathrm{kl}\big(r(s,a)-\xi,\ r(s,a)\big)\Big)}{1-\exp\Big(-\mathrm{kl}\big(r(s,a)-\xi,\ r(s,a)\big)\Big)} \right.$$
$$\left. + \sum_{s'} \left[\frac{\exp\Big(SA \cdot \mathrm{kl}\big(p(s'|s,a)+\xi/S,\ p(s'|s,a)\big)\Big)}{1-\exp\Big(-\mathrm{kl}\big(p(s'|s,a)+\xi/S,\ p(s'|s,a)\big)\Big)} + \frac{\exp\Big(SA \cdot \mathrm{kl}\big(p(s'|s,a)-\xi/S,\ p(s'|s,a)\big)\Big)}{1-\exp\Big(-\mathrm{kl}\big(p(s'|s,a)-\xi/S,\ p(s'|s,a)\big)\Big)} \right] \right),$$

we have

$$\mathbb{P}\left(\mathcal{E}_T^c\right) \leq \sum_{t=T^{1/4}}^{T} B\exp(-C\sqrt{t}) \leq BT\exp(-CT^{1/8}).$$

□

H. Comparison of KLB-TS and BESPOKE:

H.1. Design principles

As KLB-TS, BESPOKE is an algorithm that adapts its sampling strategy to the learnt MDP. The two algorithms have however different objectives: BESPOKE aims at returning an ε-optimal policy. BESPOKE starts with an intialization phase where each (state, action) pair is sampled $n_{\min} = \frac{2\times 625^2\times\gamma^2\times S\times\log(1/\delta)}{(1-\gamma)^2}$ times. After this first phase, the algorithm enters an inner loop. Each iteration of the loop aims at halving the sub-optimality gap $\left\|V_\phi^\star - V_\phi^{\widehat{\pi}^*}\right\|_\infty$ of the empirical best policy. The algorithm iterates until the gap becomes smaller than ε. At the beginning of each iteration, the algorithm solves a convex program whose solution provides the numbers of times each (state, action) pair should be sampled in this iteration. The program minimizes a weighted sum of "confidence intervals" of rewards and transitions estimates at each (state, action) pair, subject to a maximum budget constraint. This objective is known, thanks to the Simulation Lemma[4], to be an upper bound of the sub-optimality gap of the empirical optimal policy. BESPOKE uses a doubling trick to compute the maximum budget for each iteration (this budget is defined so that the gap is halved). We note the following important differences between KLB-TS and BESPOKE.

1. KLB-TS does not need to solve any convex program to update its sampling strategy, because given an estimate of the MDP, this strategy is explicit.

2. It is also worth noting that the initialization phase of BESPOKE is extremely long: $\frac{2\times 625^2\times\gamma^2\times S^2A\times\log(1/\delta)}{(1-\gamma)^2}$ samples must be gathered. During this phase, the algorithm is not adaptive at all. As we have shown in our numerical experiments, even with small state and action spaces, the initialization phase constitutes a very large proportion of the sample complexity – which makes the algorithm less adaptive than it seems, and really leads to poor performance. KLB-TS has a much smaller initialization phase and is really adaptive. On Figure 3, we see that BESPOKE's large sample complexity is mainly due to the constant term corresponding to the minimum number of samples it allocates to each (state, action) pair in the initialization phase. Note that this minimum number of samples cannot be avoided as it is necessary to ensure that BESPOKE halves the accuracy of the empirical policy after each iteration[5]

3. BESPOKE's stopping rule is suited to identify $\varepsilon-$optimal policies. Unless it has access an oracle revealing $\Delta_{\min}$, it cannot perform best policy identification.

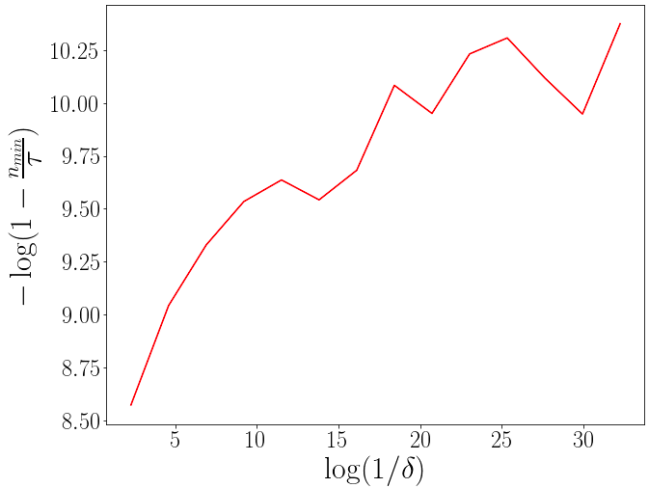

Figure 3. Comparing BESPOKE initialization phase duration $n_{\min}$ to its total sample complexity τ: $-\log(1-\frac{n_{\min}}{\tau})$ as a function of $\log(1/\delta)$.

[4] see Lemma 2 in (Zanette et al., 2019)

[5] see Lemma 16 and the proof of Theorem 1 in (Zanette et al., 2019).

H.2. Theoretical guarantees of BESPOKE and KLB-TS

Theorem 2 in (Zanette et al., 2019) states that with a probability at least $1-\delta$, the sample complexity of best-policy identification using BESPOKE with $\varepsilon = \Delta_{\min}$ is upper bounded by[6]:

$$\tau_\delta = \widetilde{\mathcal{O}}\Bigg(\sum_{s,a\neq\pi^\star(s)} \left(\frac{\text{Var}[R(s,a)] + \gamma^2 \text{Var}_{p(s,a)}[V_\phi^\star]}{\Delta_{sa}^2} + \frac{1}{(1-\gamma)\Delta_{sa}} \right) + \sum_{s\in\mathcal{S}} \min\left\{ \frac{1}{(1-\gamma)^3\Delta_{\min}^2}, \quad \frac{\text{Var}[R(s,\pi^*(s))] + \gamma^2 \text{Var}_{p(s,\pi^*(s))}[V_\phi^\star]}{\Delta_{\min}^2} + \frac{1}{(1-\gamma)^2\Delta_{\min}} \right\} + \frac{S^2 A}{(1-\gamma)^2} \Bigg).$$

In contrast, the sample complexity of KLB-TS scales as:

$$\tau_\delta = \mathcal{O}\Bigg(\sum_{s,a\neq\pi^\star(s)} \left(\max\left\{ \frac{\text{Var}_{p(s,a)}[V_\phi^\star]}{\Delta_{sa}^2}, \frac{\text{sp}[V_\phi^\star]^{4/3}}{\Delta_{sa}^{4/3}} \right\} + \frac{1}{\Delta_{sa}^2} \right) + S \times \min\left\{ \frac{1}{(1-\gamma)^3\Delta_{\min}^2}, \max\left\{ \frac{\text{Var}^*_{max}[V_\phi^\star]}{(1-\gamma)^2\Delta_{\min}^2}, \frac{\text{sp}[V_\phi^\star]^{4/3}}{(1-\gamma)^{4/3}\Delta_{\min}^{4/3}} \right\} \right\} + \frac{S}{(1-\gamma)^2\Delta_{\min}^2} \Bigg) \log(1/\delta) + o(\log(1/\delta)).$$

From the above upper bounds, we can make the following comments:

1. Both bounds depend on functionals of the particular MDP to be learnt, such as the minimum gap, the variance or maximum deviations of value functions. This means that BESPOKE and KLB-TS can adapt to the hardness of the problem, and in particular perform significantly better than minimax approaches when the MDP is easy (e.g. when the minimum gap is high or when the variances of the value function is low).
2. In the worst case, both sample complexities scale at most as $\widetilde{\mathcal{O}}\left(\frac{SA}{\Delta_{\min}^2(1-\gamma)^3} \right)$, which corresponds to the minimax bound.
3. When the rewards have strictly positive variances, then the two upper bounds are very similar, except for the large constant term $\frac{S^2 A \log(1/\delta)}{(1-\gamma)^2}$ for BESPOKE which comes from its very long initialization phase. We believe that this constant term makes BESPOKE impractical.
4. While BESPOKE's bound has the advantage of being non-asymptotic, it only holds with probability $1-\delta$. In contrast, KLB-TS comes with an asymptotic bound on the expected sample complexity, which we also proved to be finite for all confidence levels δ.

[6] $\widetilde{\mathcal{O}}(.)$ is used to indicate a quantity that depends on (.) up to a *polylog* expression at most polynomial in S, A, $\frac{1}{1-\gamma}$, $\frac{1}{\delta}$.