Best-policy Identification in Discounted Linear MDPs

Jerome Taupin[*]
ENS-PSL
Paris, France
jerome.taupin@ens.psl.eu

Yassir Jedra[†]
KTH
Stockholm, Sweden
jedra@kth.se

Alexandre Proutiere[†]
KTH
Stockholm, Sweden
alepro@kth.se

Abstract

We consider the problem of best policy identification in discounted Linear Markov Decision Processes in the fixed confidence setting, under both generative and forward models. We derive an instance-specific lower bound on the expected number of samples required to identify an ε-optimal policy with probability $1-\delta$. The lower bound characterizes the optimal sampling rule as the solution of an intricate non-convex optimization program, but can be used as the starting point to devise simple and near-optimal sampling rules and algorithms. We devise such algorithms. In the generative model, our algorithm exhibits a sample complexity upper bounded by $\mathcal{O}((d(1-\gamma)^{-4}/(\varepsilon+\Delta)^2)(\log(1/\delta)+d))$ where Δ denotes the minimum reward gap of sub-optimal actions and d is the dimension of the feature space. This upper bound holds in the moderate-confidence regime (i.e., for all δ), and matches existing minimax and gap-dependent lower bounds. In the forward model, we determine a condition under which learning approximately optimal policies is possible; this condition is weak and does not require the MDP to be ergodic nor communicating. Under this condition, the sample complexity of our algorithm is asymptotically (as δ approaches 0) upper bounded by $\mathcal{O}((\sigma^\star(1-\gamma)^{-4}/(\varepsilon+\Delta)^2)(\log(\frac{1}{\delta})))$ where $\sigma^\star$ is an instance-specific constant, value of an optimal experiment-design problem. To derive this bound, we establish novel concentration results for random matrices built on Markovian data.

1 Introduction

In Reinforcement Learning (RL), an agent interacts with an unknown controlled stochastic dynamical system, with the objective of identifying as quickly as possible an approximately optimal control policy. In this paper, we consider dynamical systems modelled through discounted Markov Decision Processes (MDPs), and investigate the problem of best policy identification in the *fixed confidence* setting. More precisely, we aim at devising (ε,δ)-PAC RL algorithms, i.e., algorithms identifying ε-optimal policies with a level of certainty greater than $1-\delta$, using as few samples as possible. Such a learning objective has been considered extensively in tabular MDPs both in the discounted and episodic settings, most often using a minimax approach, see e.g. [15, 13, 7, 4, 22, 2, 19, 9, 5, 6] and more recently adopting an instance-specific analysis [21, 20]. According to the aforementioned

[*] This work was done when J. Taupin was a research intern at KTH.

[†] Y. Jedra and A. Proutiere are both supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

16th European Workshop on Reinforcement Learning (EWRL 2023).

work, in tabular MDPs, the minimal sample complexity for identifying an ε-optimal policy with probability at least $1-\delta$ scales as $\frac{SA}{\varepsilon^2}\log(1/\delta)$ (ignoring the dependence in the time-horizon or discount factor), where S and A represent the sizes of the state and action spaces, respectively. These results illustrate the curse of dimensionality (tabular MDPs with limited state and action spaces only are learnable), and highlight the need for the use of function approximation towards the design of scalable RL algorithms.

Despite the empirical successes of RL algorithms leveraging function approximation, our theoretical understanding of these methods remain limited. In this paper, we investigate the so-called *linear* MDPs, where linear functions are used to approximate the system dynamics and rewards. We aim at devising statistically and computationally efficient algorithms for the best policy identification with fixed confidence learning task. We address this task under both (i) the *generative model*, where in each round, a sample of the transition and reward from any given state-action pair can be observed; and (ii) the *forward model*, where the learner has access to a single controlled trajectory of the system. Our contributions are summarized below.

(a) Sample complexity lower bounds. We derive instance-specific lower bounds that any (ε,δ)-PAC algorithm must satisfy, for both the generative and forward models. These lower bounds are characterized by the solution of an intricate optimization problem. We propose a careful relaxation of these optimization problems. These relaxations suggest an *experiment design* approach based on G-optimal design to define the sampling strategy used to explore the MDP.

(b) Algorithms with a generative model. When the learner has access to a generative model, inspired by our sample complexity lower bounds, we devise G-Sample-and-Stop (GSS), a simple (ε,δ)-PAC algorithm that relies on G-optimal design [17, Chap. 21], least-squares estimators, and a proper stopping rule. We show that the expected sample complexity of GSS scales at most as $((d(1-\gamma)^{-4})/(\Delta_{\mathcal{M}}+\varepsilon)^2)(\log(1/\delta)+d)$ (up to logarithmic factors), where $\Delta_{\mathcal{M}}$ is an appropriately defined instance-specific sub-optimality gap that depends on the MDP $\mathcal{M}$. This upper bound holds in the moderate-confidence regime (i.e., for all $\delta\in(0,1)$), and matches existing minimax and gap-dependent lower bounds.

(c) Algorithms with the forward model. Again inspired by our sample complexity lower bounds, we propose G-Navigate-and-Stop (GNS). The analysis of GNS or other algorithms for the forward model presents many challenges: (i) In contrast with *episodic* setting, we do not have the ability to restart the trajectory at each episode. Hence, suitable conditions are required to ensure that learning is even possible from a single controlled trajectory. (ii) Because of the linear structure, the uniqueness of the optimal sampling policy that arise from our lower bounds is not guaranteed, and the set of such optimal policies does not have nice properties such as convexity. Therefore, a careful sampling scheme is required. (iii) The data generated when exploring the MDP is Markovian, which implies that new concentration results for random matrices with Markovian data must be derived. We overcome these challenges. First, we determine conditions under which learning approximately optimal policies is possible; these conditions are weak and do not require the MDP to be ergodic nor communicating. Then, under such conditions, we establish concentration bounds on random matrices with Markovian data. Finally, we show that the sample complexity of GNS, under the learnability conditions, is asymptotically (as δ approaches 0) upper bounded by $\mathcal{O}\left(((1-\gamma)^{-4}\sigma^{\star}_{\mathcal{M},\text{for}}/(\Delta_{\mathcal{M}}+\varepsilon)^2)\log(1/\delta)\right)$ where $\sigma^{\star}_{\mathcal{M},\text{for}}$ is an instance-specific constant, value of an optimal *experiment design* problem.

2 Related Work

Linear models in RL have attracted a lot of attention over the last few years. We summarize below the recent results, related to first episodic MDPs and then discounted MDPs.

Episodic linear MDPs. Most of the studies have aimed at devising algorithms minimizing regret. Jin et al. [12] propose an optimistic Least Squares Value Iteration (LSVI) algorithm that achieves a regret upper bound of order $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ and that can be implemented in polynomial time. [10] presents UCRL-VTR, a confidence based algorithm adapted to the linear MDP setting. The algorithm achieves a gap dependent regret of order $\tilde{\mathcal{O}}(((d^2H^5)/\Delta_{\min})\log(T/\delta)^3)$. When it comes to best policy identification problems, in [27], Wagenmaker et al. establish a sample complexity minimax lower bound for the task of identifying an ϵ-optimal policy. the lower bound scales as $\Omega(d^2H^2/\epsilon^2)$. The authors also propose an a reward-free algorithm with sample complexity of order $\tilde{\mathcal{O}}(d/\epsilon^2)(\log(1/\delta)+d)H^5$.

In a subsequent work, Wagenmaker et al. [28] introduce PEDEL, an elimination based algorithm with instance-specific sample complexity guarantees. In the worst case, the sample complexity upper bound scales as $\tilde{\mathcal{O}}((dH^5/\epsilon^2)(dH^2 + \log(1/\delta)))$. This bound hides a dependence on $\lambda^\star_{\min}$, the maximal minimum eigenvalue of the covariates matrix that can be induced by a policy. As in our work, the derived instance-specific sample complexity guarantees are related to G-optimal design and take the following form: $C_0\mathcal{G} + C_1$ where $\mathcal{G} = H^4 \sum_{h=1}^H \inf_{\Lambda_{exp}} \max_{\pi\in\Pi} \frac{\|\phi_{\pi,h}\|_{\Lambda_{exp}^{-1}} \log(\frac{|\Pi|}{\delta})}{\max\{V^\star(\Pi) - V^\pi, \Delta_{\min}(\Pi), \epsilon^2\}}$ with $C_0 = \log\left(\frac{1}{\epsilon}\right) \text{polylog}\left(H, \log(1/\epsilon)\right)$ and $C_1 = \text{poly}\left(d, H, 1/\lambda^\star_{\min}, \log(1/\delta), \log(1/\varepsilon), \log(|\Pi|)\right)$. Note that PEDEL requires as input a set of policies Π. The authors propose a way to approximate the set of all policies using restricted linear soft-max policies Π_ϵ which leads to an overall sample complexity of order $C_0 H^4 \sum_{h=1}^H \inf_{\Lambda_{exp}} \max_{\pi\in\Pi_\epsilon} \frac{\|\phi_{\pi,h}\|_{\Lambda_{exp}^{-1}} (dH^2 + \log(\frac{1}{\delta}))}{\max\{V^\star - V^\pi, \epsilon^2\}} + C_1$. In Zanette et al. [30], the authors also investigate the problem of identifying an ϵ-optimal policy with a generative model and propose a Linear Approximate Value Iteration algorithm (LAVI). They leverage the idea of anchor (state, action) pairs but require a set of such anchor pairs for each layer $h \in [H]$.

Discounted linear MDPs. In [29], Yang et al. focus on the ϵ-optimal policy identification problem in the generative setting and present Phased Parametric Q-Learning (PPQ-learning), an algorithm with sample complexity of order $\tilde{\mathcal{O}}(\frac{d}{(1-\gamma)^3\epsilon^2}\log(\frac{1}{\delta}))$ under the restrictive assumption that a so-called set of (state, action) anchor pairs exist (see Assumption 2) and that it is of size d. More precisely, this assumption states that there exists $\mathcal{K} \subset \mathcal{S} \times \mathcal{A}$, a set of anchor (state, action) pairs such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\phi(s, a)$ can be written as convex combination of features of anchor pairs. The authors further assume that $|\mathcal{K}| = d$ and that all features have non-negative entries and that the features correspond to probability vectors. The authors finally provide a matching minimax lower bound of order $\tilde{\Omega}(\frac{d}{\epsilon^2(1-\gamma)^3})$.

Lattimore et al. [18] also consider the ϵ-optimal policy identification problem in the generative setting. They devise a sampling rule based on G-optimal design and use an approximate policy iteration algorithm to recover the optimal policy. Their algorithm seeks to estimate the Q function directly at each iteration, by first evaluating the value of Q at anchor (state, action) pairs (determined by the G-optimal design) via rollout, and by then generalizing using least squares. The sample complexity of their algorithm is of the order $\tilde{\mathcal{O}}(\frac{d\sqrt{d}}{\epsilon^2(1-\gamma)^8}\log(\frac{1}{\delta}))$.

Finally it is worth mentioning [31], where Zhou et al. consider the regret minimization problem in the forward model. The notion of regret for discounted MDPs is not easy to define. Here, the authors consider the accumulated difference of rewards between an Oracle policy and the proposed policy but along the trajectory followed under the latter policy (this policy could well lead the system into regions of the state space). The proposed algorithm achieves a regret scaling at most as $\hat{\mathcal{O}}(d\sqrt{T}/(1-\gamma)^2)$.

3 Models and Objectives

3.1 Notation

We denote by $\|x\|$ the Euclidean norm of a vector $x \in \mathbb{R}^d$. For a given definite positive matrix $M \in \mathbb{R}^{d\times d}$, we denote by $\|x\|_M = \sqrt{x^\top M x}$ the weighted Euclidean norm of the vector $x \in \mathbb{R}^d$. We denote by $\|M\|$ the operator norm of a matrix $M \in \mathbb{R}^{d\times p}$. For a positive definite matrix $M \in \mathbb{R}^{d\times d}$, we denote its highest (resp. smallest) eigenvalue by $\lambda_{\max}(M)$ (resp. $\lambda_{\min}(M)$), respectively. For a given pair of two symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, we write $A \succ B$ (resp. $A \succeq B$) to mean that $A - B$ is positive definite (resp. positive semi-definite).

3.2 Discounted linear MDPs

We consider an infinite time-horizon discounted MDP, denoted $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p_\mathcal{M}, q_\mathcal{M}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, $p_\mathcal{M}$ and $q_\mathcal{M}$ are the dynamics and reward distributions, respectively, and $\gamma \in (0, 1)$ is the discount factor. More precisely, starting from state s and given that action a is selected, the probability to move to state s' is $p_\mathcal{M}(s, a, s')$ and the distribution of the collected reward is $q_\mathcal{M}(s, a)$. We assume that $q_\mathcal{M}(s, a)$ has support on $[0, 1]$, and we denote by $r_{\mathcal{M}}(s,a)$ the expected reward of $q_{\mathcal{M}}(s,a)$. $\mathcal{S}$ and $\mathcal{A}$ are finite sets of respective cardinalities S and A.

Each state-action pair (s,a) is associated to a feature vector $\phi(s,a) \in \mathbb{R}^d$. We assume that the feature map ϕ is known to the learner, that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\phi(s,a)\| \leq 1$, and that the features $(\phi(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ span $\mathbb{R}^d$. We are interested in the class of the so-called Linear MDPs, denoted by $\mathbb{M}$, and defined as follows [12]:

Definition 3.1 (Linear MDPs)**.** We say $\mathcal{M}$ is a Linear MDP if there exists $\mu_{\mathcal{M}}$, a family of d measures over $\mathcal{S}$, seen as $S \times d$-dimensional matrix, and $\theta_{\mathcal{M}} \in \mathbb{R}^d$, such that for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$p_{\mathcal{M}}(s,a,s') = \phi(s,a)^\top \mu_{\mathcal{M}}(s'), \quad \text{and} \quad r_{\mathcal{M}}(s,a) = \phi(s,a)^\top \theta_{\mathcal{M}}, \tag{1}$$

with $\max\{\|\theta_{\mathcal{M}}\|, \|\mu_{\mathcal{M}}(\mathcal{S})\|\} \leq \sqrt{d}$.

A deterministic stationary control policy π maps states to actions. We denote by s_t^π the state at time t under the policy π, and by $\pi(s)$ the action selected by π. The performance of a policy π is expressed through its state value function $V_{\mathcal{M}}^\pi$ and its state-action value function $Q_{\mathcal{M}}^\pi$ defined by: for all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$V_{\mathcal{M}}^\pi(s) = \mathbb{E}_{\mathcal{M}}\left[\sum_{t=0}^{+\infty} \gamma^t r_{\mathcal{M}}(s_t^\pi, \pi(s_t^\pi)) | s_0^\pi = s\right],$$

and

$$Q_{\mathcal{M}}^\pi(s,a) = r_{\mathcal{M}}(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_{\mathcal{M}}(s,a,s') V_{\mathcal{M}}^\pi(s').$$

An optimal policy $\pi_{\mathcal{M}}^\star$ for the MDP $\mathcal{M}$ maximizes the value function for any state, i.e., for any policy π, we have $V_{\mathcal{M}}^{\pi_{\mathcal{M}}^\star}(s) \geq V_{\mathcal{M}}^\pi(s)$ for all $s \in \mathcal{S}$. The state and state-action value functions of $\pi_{\mathcal{M}}^\star$ are referred to as the value function $V_{\mathcal{M}}^\star$ and the Q function $Q_{\mathcal{M}}^\star$, respectively. A policy π is said ε-optimal if $\max_{s\in\mathcal{S}} V_{\mathcal{M}}^\pi(s) - V_{\mathcal{M}}^\star(s) \leq \varepsilon$ point-wise, and we denote by $\Pi_\varepsilon^\star(\mathcal{M})$ the set of ε-optimal policies of $\mathcal{M}$.

3.3 Best policy identification

We aim at designing a learning algorithm interacting with the MDP $\mathcal{M}$ so as to identify an ε-optimal policy as quickly as possible. We formalize this objective in a PAC framework, where a learning algorithm consists of *(i) a sampling rule*, *(ii) a stopping rule* and *(iii) a decision rule*.

(i) *Sampling rule:* We distinguish between the generative and the forward model:

1. *Generative model:* In each round t, the sampling rule may select any (state, action) (s_t, a_t) to explore depending on past observations.
2. *Forward model:* Under this model, the learner is forced to follow the trajectory of the system, and only the action may be selected.

Under both models, from the selected pair, the learner observes the next state and receives a sample of the corresponding reward.

(ii) *Stopping rule:* This rule is defined through a stopping time τ deciding when the learner stops gathering information and wishes to output an estimated ε-optimal policy.

(iii) *Decision rule:* Based on the observations gathered before stopping, the learner outputs an estimated optimal policy $\hat{\pi}_\tau$.

We are interested in learning algorithms that are (ε,δ)-PAC in the following sense:

Definition 3.2 ((ε,δ)-PAC algorithms)**.** An algorithm is said (ε,δ)-PAC if at the time it stops τ, it ouputs a policy $\hat{\pi}_\tau$ satisifying:

$$\mathbb{P}_{\mathcal{M}}\left(\max_{s\in\mathcal{S}}\left(V_{\mathcal{M}}^\star(s) - V_{\mathcal{M}}^{\hat{\pi}_\tau}(s)\right) < \varepsilon\right) \geq 1-\delta$$

Our goal is to design (ε,δ)-PAC algorithms with minimal sample complexity $\mathbb{E}_{\mathcal{M}}[\tau]$. In contrast with most existing analyses, we will derive *instance-specific* lower and upper bounds on the sample complexity of such algorithms (ε,δ)-PAC algorithms. In particular, we wish these bounds to depend on the sub-optimality gap of the MDP $\mathcal{M}$ defined by $\Delta_{\mathcal{M}} = \min_{s\in\mathcal{S}, a\neq\pi_{\mathcal{M}}^\star(s)}(V_{\mathcal{M}}^\star(s) - Q_{\mathcal{M}}^\star(s,a))$.

4 Sample Complexity Lower Bounds

To state our instance-specific lower bounds, we first introduce the following notation. Given two MDPs $\mathcal{M}$ and $\mathcal{M}'$ in $\mathbb{M}$, we write $\mathcal{M} \ll \mathcal{M}'$ if for every pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have[3] $p_{\mathcal{M}}(s,a,\cdot) \ll p_{\mathcal{M}'}(s,a,\cdot)$ and $q_{\mathcal{M}}(s,a) \ll q_{\mathcal{M}'}(s,a)$. In this case, we define the Kullback-Leibler divergence between $\mathcal{M}$ and $\mathcal{M}'$ by:

$$\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) = \mathrm{KL}(q_{\mathcal{M}}(s,a)\|q_{\mathcal{M}'}(s,a)) + \mathrm{KL}(p_{\mathcal{M}}(s,a,\cdot)\|p_{\mathcal{M}'}(s,a,\cdot)).$$

We also denote by $\mathrm{kl}(a,b)$ the Kullback-Leibler divergence of two Bernoulli distributions of respective means a and b. Finally, we introduce the following set of MDPs. This set includes MDPs for which the set of ε-optimal policies does not contain an ε-optimal policy for $\mathcal{M}$.

$$\mathrm{Alt}_\varepsilon(\mathcal{M}) = \left\{\mathcal{M}' \in \mathbb{M} : \begin{cases}\mathcal{M} \ll \mathcal{M}' \\ \Pi^\star_\varepsilon(\mathcal{M}) \cap \Pi^\star_\varepsilon(\mathcal{M}') = \emptyset\end{cases}\right\}$$

We refer to $\mathrm{Alt}_\varepsilon(\mathcal{M})$ as the *set of alternative MDPs w.r.t.* $\mathcal{M}$. Let $\Sigma_{\mathcal{S}\times\mathcal{A}}$ be the probability simplex in $\mathbb{R}^{SA}$, and define for all $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$,

$$T_{\mathcal{M}}(\omega)^{-1} = \inf_{\mathcal{M}'\in\mathrm{Alt}_\varepsilon(\mathcal{M})} \sum_{s,a} \omega_{s,a}\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a). \tag{2}$$

With a generative model. For the *generative* model, we establish the following lower bound.

Proposition 4.1. *Let $\varepsilon > 0$, $\delta \in (0,1)$. The sample complexity τ of any (δ,ε)-PAC algorithm must satisfy: $\mathbb{E}_{\mathcal{M}}[\tau] \ge T^\star_{\mathcal{M},\mathrm{gen}}\,\mathrm{kl}(\delta, 1-\delta)$ where $T^\star_{\mathcal{M},\mathrm{gen}} = \inf_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}} T_{\mathcal{M}}(\omega)$.*

The derivation of the lower bound in Proposition 4.1 relies on standard change-of-measure arguments. We defer the proof to Appendix A. The vector $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$ solving the optimization problem and leading to $T^\star_{\mathcal{M},\mathrm{gen}}$ can be interpreted as the optimal proportions of times an optimal algorithm should sample the various (state, action) pairs. It turns out, as in the case of tabular MDPs (see [21]), that analyzing and computing this allocation is difficult. Instead, our strategy will be to derive instance-specific upper bounds of the $T^\star_{\mathcal{M},\mathrm{gen}}$ that can be computed in a computationally efficient manner. To state the upper bounds, we introduce the following quantities: let $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$, $\Lambda(\omega) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega_{s,a}\phi(s,a)\phi(s,a)^\top$, and $\sigma(\omega) = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{\Lambda(\omega)^{-1}}$. $\Lambda(\omega)$ is referred to as the *feature matrix*. Furthermore, observe that the function $\sigma(\cdot)$ corresponds to the so-called G-optimality criterion (see e.g. Chap. 21 in [17]). Our next result is to establish a link between $T_{\mathcal{M}}(\cdot)$ and $\sigma(\cdot)$.

Theorem 4.2. *For all $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$, it holds that*

$$T_{\mathcal{M}}(\omega) \le \frac{10\sigma(\omega)}{3(1-\gamma)^4(\Delta_{\mathcal{M}}+\varepsilon)^2}. \tag{3}$$

Consequently, we have $T^\star_{\mathcal{M},\mathrm{gen}} \le U^\star_{\mathcal{M},\mathrm{gen}} \triangleq \frac{10d}{3(1-\gamma)^4(\Delta_{\mathcal{M}}+\varepsilon)^2}$.

Theorem 4.2 relates the *experiment-design* approach based on G-optimality to our instance dependent lower bound. A similar link has been established in the case of best-arm identification in linear bandits [23]. However, establishing such a link in the case of Linear Discounted MDPs is more challenging and requires a careful relaxation of the optimization problem leading to the definition of $T_{\mathcal{M}}(\omega)$ in (2). The proof of Theorem 4.2 is deferred to Appendix A.

From an algorithmic perspective, Theorem 4.2 tells us that sampling according to a G-optimal design $\omega^\star$ (i.e., $\omega^\star \in \arg\min_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}} \sigma(\omega)$) is sufficient to identify an ϵ-optimal policy with a sample complexity upper bounded by the gap-dependent quantity $U^\star_{\mathcal{M},\mathrm{gen}}\log(1/\delta)$. $\omega^\star$ only depends on the feature map ϕ and not the uknowns $\mu_{\mathcal{M}}$ and $\theta_{\mathcal{M}}$, and therefore may be computed prior to the learning process.

[3] Here $\ll$ refers to the standard symbol for absolute continuity between probability measures.

With a forward model. Proposition 4.1 and Theorem 4.2 can be immediately extendend to the *forward* model. To simplify the exposition, we will restrict our attention to the asymptotic lower bounds when $\delta \to 0$. As in [20], we can establish that if ω_{sa} denotes the expected proportion of rounds where the state-action pair (s,a) is visited, then the allocation ω, asymptotically, must satisfy the balance equations of the Markov chain induced by the controlled system dynamics: for all $s \in \mathcal{S}$,

$$\sum_{a \in \mathcal{A}} \omega_{s,a} = \sum_{(s',a') \in \mathcal{S}\times\mathcal{A}} p_{\mathcal{M}}(s',a',s)\omega_{s',a'}. \tag{4}$$

Define $\Omega(\mathcal{M}) = \{\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$: the constraints (4) hold$\}$.

Proposition 4.3. *Let $\varepsilon > 0$, $\delta \in (0,1)$. In the forward model, the sample complexity τ of any (ε,δ)-PAC algorithm must satisfy: $\mathbb{E}_{\mathcal{M}}[\tau] \geq T^\star_{\mathcal{M},\text{for}}\, \text{kl}(\delta, 1-\delta)$ where $T^\star_{\mathcal{M},\text{for}} = \inf_{\omega\in\Omega(\mathcal{M})} T_{\mathcal{M}}(\omega)$.*

Theorem 4.4. *Let $\sigma^\star_{\mathcal{M},\text{for}} = \inf_{\omega\in\Omega(\mathcal{M})}\sigma(\omega)$. Then, we have*

$$T^\star_{\mathcal{M},\text{for}} \leq U^\star_{\mathcal{M},\text{for}} \triangleq \frac{10\ \sigma^\star_{\mathcal{M},\text{for}}}{3(1-\gamma)^4(\Delta_{\mathcal{M}}+\varepsilon)^2}. \tag{5}$$

The proof of Proposition 4.3 and Theorem 4.4 are presented in Appendix A. The upper bound we obtain on $T^\star_{\mathcal{M},\text{for}}$, suggests an *experiment design* approach where the objective is to sample according to an allocation $\omega^\star \in \arg\min_{\omega\in\Omega(\mathcal{M})}\sigma(\omega)$. This objective is similar in spirit to that considered in [28] for Episodic Linear MDPs.

5 The G-Sample-and-Stop Algorithm

We propose G-Sample-and-Stop (GSS), an algorithm whose sample complexity matches the complexity measure $U^\star_{\mathcal{M},\text{gen}}\log(1/\delta)$ presented in Theorem 4.2. The algorithm samples the state-action pairs according to a G-optimal design, and stops when it has gathered enough information. The adaptive nature of the stopping rule ensures a gap-dependent sample complexity upper bound.

5.1 Sampling rule

Prior to the learning process, under the GSS algorithm, we start by finding[4] an optimal allocation $\omega^\star \in \arg\min_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}}\sigma(\omega)$. Then, at each round t, the algorithm proceeds by sampling a state-action pair (s_t,a_t) according to $\omega^\star$. Define $P_t = \sum_{\ell=1}^t \phi(s_\ell,a_\ell)\phi(s_\ell,a_\ell)^\top$. Standard concentration arguments on random matrices ensure that the random matrix P_t converges to the matrix $t\Lambda(\omega^\star)$. In particular, $t\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|^2_{P_t^{-1}}$ will converge towards $\sigma(\omega^\star)$. We present this fact in the following proposition, and its proof is deferred to Appendix B.

Proposition 5.1. *Let $\delta \in (0,1)$. We have*

$$\mathbb{P}\left(t \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{P_t^{-1}} \leq 2\sigma(\omega^\star)\right) \geq 1-\delta,$$

provided $t \geq 10d\log\left(\frac{2d}{\delta}\right)$.

5.2 Least-squares estimation

The stopping and decision rules of GSS leverage the least-squares estimators of the parameters $\mu_{\mathcal{M}}$ and $\theta_{\mathcal{M}}$. We provide below explicit expressions for these estimators and derive concentration inequalities characterizing their performance. When the algorithm selects (state, action) pair (s_t,a_t) in round t, it observes the next state s'_t and receives the reward r_t. Overall, in round t, the algorithm gathers the experience (s_t,a_t,r_t,s'_t). The regularized least-squares estimators with parameter $\lambda > 0$ of $\mu_{\mathcal{M}}$ and $\theta_{\mathcal{M}}$ after t experiences are given by: for all $s \in \mathcal{S}$,

$$\hat{\mu}_t(s) = (P_t + \lambda I_d)^{-1}\sum_{\ell=1}^t \phi(s_\ell,a_\ell)\mathbb{1}_{\{s'_\ell = s\}}, \quad \text{and} \quad \hat{\theta}_t = (P_t+\lambda I_d)^{-1}\sum_{\ell=1}^t \phi(s_\ell,a_\ell)r_\ell. \tag{6}$$

[4] Finding a G-optimal design is well studied problem. We refer the reader to Chap. 21 in [17] and further computational considerations are discussed in Appendix B.

In what follows, we choose $\lambda = 1/d$ and denote by $\widehat{\mathcal{M}}_t$ the MDP associated to the corresponding least-squares estimators. Let $\widehat{V}_t$ and $\widehat{Q}_t$ be its value functions. The performance of the least-squares estimators can be controlled in the following sense:

Proposition 5.2. *Irrespective of the sampling rule, we have for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\forall t \geq 1, \quad \left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}_t^\star\right\|_{P_t}^2 \leq \beta(\delta, t)\right) \geq 1 - \delta \tag{7}$$

with the threshold $\beta(\delta, t) = \frac{C}{(1-\gamma)^2}\left(\log\left(e/\delta\right) + d\log\left(dt\right)\right)$ for some universal constant $C > 0$.

The proof of Proposition 5.2 is presented in Appendix C along with the precise constants. Importantly, the threshold β does not exhibit any dependence in S but only in d. This is thanks to the linear structure that characterizes the value function. Such a structure allows us to use a net argument on the space of all possible optimal value functions. This idea is borrowed from [12] and repurposed to our needs.

5.3 Stopping and decision rules

Let us start by describing the stopping rule. For all $t \geq 1$, we define the random variable $Z(t)$ and the threshold $\beta(\delta, t)$ as follows

$$Z(t) = \frac{3(1-\gamma)^4(\Delta_{\widehat{\mathcal{M}}_t} + \varepsilon)^2}{10\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|_{P_t^{-1}}^2}, \qquad \text{and} \qquad \beta(\delta, t) = C\left(\log\left(\frac{e}{\delta}\right) + d\log\left(dt\right)\right).$$

The random variable $Z(t)/t$ may be interpreted as an empirical estimator of the lower bound on $U^\star_{\mathcal{M},\mathrm{gen}}$ established in Theorem 4.2. The choice of the threshold $\beta(\delta, t)$ is motivated by the concentration result of Propostion 5.2 with C being the universal constant in the statement of the proposition. Finally, the stopping rule of GSS is defined by the stopping time

$$\tau = \inf\left\{t \geq 1 : Z(t) > \beta(\delta, t)\right\}. \tag{8}$$

This stopping rule is inspired by classical log-likelihood based stopping rules. When the algorithm stops, it computes $\hat{\pi}_\tau$, an optimal policy for the MDP $\widehat{\mathcal{M}}_\tau$. The description of GSS is now complete and summarized in Algorithm 1.

Algorithm 1: G-Sample-and-Stop (GSS)

```
Compute $\omega^\star = \arg\min_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}} \sigma(\omega)$
while $Z(t) \leq \beta(\delta, t)$ do
    sample $(s_t, a_t)$ according $\omega^\star$
    observe the experience $(s_t, a_t, r_t, s'_t)$
    update $(\mu_t, \theta_t)$ according to (6) and set $t = t + 1$
end
return $\hat{\pi} = \pi_t^\star$ the optimal policy of $\widehat{\mathcal{M}}_t$
```

The following Lemma establishes the (ε, δ)-PAC correctness of GSS. It is a consequence of Propositon 5.2 and its proof is deferred to Appendix C.

Lemma 5.3. *Under the GSS algorithm, we have: $\mathbb{P}\left(\tau < +\infty, \hat{\pi}_\tau \notin \Pi_\varepsilon^\star(\mathcal{M})\right) \leq \delta$.*

5.4 Sample complexity guarantees under GSS

Finally, in Theorem 5.4 we present the sample complexity guarantee enjoyed by GSS.

Theorem 5.4. *The sample complexity of GSS satisfies, for all $\varepsilon > 0$, $\delta \in (0,1)$,*

$$\mathbb{E}[\tau] \leq CU^\star_{\mathcal{M},\mathrm{gen}}\left(\log\left(\frac{e}{\delta}\right) + d\log\left(U^\star_{\mathcal{M},\mathrm{gen}}\right)\right) \tag{9}$$

where $C > 0$ is a universal constant. Furthermore, GSS is an (ε, δ)-PAC algorithm.

The proof of Theorem 5.4 is presented in Appendix D. First, observe that the sample complexity guarantee is valid for all $\delta \in (0,1)$ which contrasts with most existing asymptotic results in best policy identification. Additionally, our guarantee is matching, up to a constant multiplicative factor, the upper bound established in Theorem 4.2 as $\delta \to 0$.

6 The G-Navigate-and-Stop algorithm

In this section, we present G-Navigate-and-Stop (GNS), an algorithm whose sample complexity matches the complexity measure $U^{\star}_{\mathcal{M},\text{for}}\log(1/\delta)$ presented in Theorem 4.4. The design of GNS, as that of GSS, is guided by our lower bounds. In particular, the stopping and decision rules are the same as those of GSS and all guarantees related to these components also hold for GNS, namely Proposition 5.2 and Lemma 5.3. The major difference lies in the sampling rule where now we have to account for navigation constraints.

6.1 Sampling rule

In what follows, we denote, for ease of notations, for all $\ell \geq 1$, $\phi_\ell = \phi(s_\ell, a_\ell)$. Recall that $P_t = \sum_{\ell=1}^t \phi_\ell \phi_\ell^\top$. As already mentioned in the study of the generative model, this random matrix plays a crucial role. In the forward model, the role P_t is more pronounced and in fact all our learnability conditions concern this matrix.

6.1.1 Forced exploration

Learning from a single trajectory requires the existence of at least a policy that explores the MDP sufficiently. Additionally if there is any hope for finding an optimal exploration strategy then we need at least to guarantee that while searching for such a policy, we do not get trapped in states that irrevocably limit our exploration. This motivates the definition of (m,λ)-*covering policies.*

Definition 6.1 ((m,λ)-covering policy)**.** A policy π is said to be an (m,λ)-covering policy of $\mathcal{M}$ if there exists $m \geq 1$ and $\lambda > 0$ such that:

$$\min_{s\in\mathcal{S}} \lambda_{\min}\left(\frac{1}{m}\mathbb{E}^{\pi}_{\mathcal{M}}\left[\sum_{t=1}^{m}\phi_t\phi_t^\top \Big| s_1 = s\right]\right) > \lambda. \tag{10}$$

We make the following assumption, which is necessary to ensure that learnability is possible.

Assumption 6.2. There exists an (m,λ)-covering policy π_e. Furthermore, the learner is aware of the policy π_e and of m.

It is worth noting that Assumption 6.2 does not require a priori that the MDP $\mathcal{M}$ is ergodic nor communicating. For a more detailed discussion on this assumption, refer to Appendix B. We are now ready to present our forced exploration scheme.

Lemma 6.3 (Forced exploration)**.** *Let $(b_t)_{t\geq 1}$ be an arbitrary sequence of actions, possibly adversarially chosen. Under Assumption 6.2, let $(a_t)_{t\geq 1}$ be a sequence of actions sampled according to:*

$$a_t = (1-x_t)b_t + x_t\pi_e(\cdot|s_t) \tag{11}$$

where $x_t = 1$ with probability $t^{-1/2m}$ and $x_t = 0$ with probability $1 - t^{-1/2m}$. Then, we have $\mathbb{P}\left(\lambda_{\min}\left(\sum_{\ell=1}^t \phi_\ell\phi_\ell^\top\right) \geq \frac{\lambda}{2}\sqrt{t}\right) \geq 1-\delta$, provided that $t \geq \left(\frac{8m}{\lambda^2}\log\left(\frac{2d}{\delta}\right)\right)^2$.

The proof of Lemma 6.3 relies on a careful decomposition of P_t and using a matrix martingale Bernstein concentration bound. We refer the reader to Appendix B for the proof. As it turns out, the high probability guarantee on the growth of the smallest eigenvalue in Lemma 6.3 is sufficient to ensure consistency of the least-squares estimator of $\mu_{\mathcal{M}}$. This is required for the sample complexity analysis of GNS.

6.1.2 Tracking

Before we present our tracking procedure, we present what we refer to as the oracle policy of a given allocation ω.

Oracle policy. As in [3], given an allocation $\omega \in \Omega(\mathcal{M})$, we define the oracle policy $\pi^o(\omega)$ as follows: for all $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\pi^o(\omega)(a|s) = \begin{cases} \frac{\omega_{s,a}}{\sum_{a\in\mathcal{A}}\omega_{s,a}} & \text{if} \quad \sum_{a\in\mathcal{A}}\omega_{s,a} > 0, \\ \frac{1}{|\mathcal{A}|} & \text{otherwise.}\end{cases} \tag{12}$$

It is not difficult to verify that the policy $\pi^o(\omega)$ indeed induces the allocation ω.

Optimal allocations. Next, we make the following assumption to avoid unnecessary technical issues that may arise characterizing the set of optimal allocations.

Assumption 6.4. There exists $\eta > 0$ such that $\big(\arg\min_{\omega\in\Omega(\mathcal{M})}\sigma(\omega)\big) \cap \Omega_\eta(\mathcal{M}) = \emptyset$ where $\Omega_\eta(\mathcal{M}) \triangleq \{\omega \in \Omega(\mathcal{M}) : \Lambda(\omega) \succeq 2\eta I_d\}$. Furthermore, the learner has access to η.

Under Assumption 6.4, we can characterize the set of optimal allocations[5] as being non-empty, compact and convex. However, it is not guaranteed that the optimal allocation is unique[6]. This complicates the design of the tracking procedure. In particular, we cannot use the C-tracking rule used for instance in tabular MDPs as in [3]. To circumvent this issue, we use lazy updates or a doubling trick. Let $\mathcal{T} = \{2^k : k \in \mathbb{N}\}$. The allocation ω_t that GNS tracks is updated only when $t \in \mathcal{T}$.

Optimization oracle. We assume that the learner has access to an optimization oracle that given a model $\widehat{\mathcal{M}}$, outputs an allocation $\omega^\star \in \arg\min_{\omega\in\Omega_{\eta/2}(\widehat{\mathcal{M}})}\sigma(\omega)$. This optimization problem is convex and therefore computationally tractable.

We are now ready present the sampling rule of GNS. When $t \in \mathcal{T}$, the alogrithm computes $\omega^\star \in \arg\min_{\omega\in\Omega_{\eta/2}(\widehat{\mathcal{M}})}\sigma(\omega)$, and updates π_t as $\pi^o(\omega_t)$. Now in each round t, b_t is sampled according to $\pi_t(\cdot|s_t)$ and GNS selects the action a_t defined in (11). The pseudo-code of GNS is presented in Algorithm 2.

Algorithm 2: The G-Navigate-and-Stop

```
Initialize $\pi_1$ to be the uniform policy
while $Z(t) \le \beta(\delta, t)$ do
    If $t \in \mathcal{T}$ then compute $\omega_t \in \arg\min_{\omega\in\Omega_{\eta/2}(\widehat{\mathcal{M}}_t)}\sigma(\omega)$ and set $\pi_t \leftarrow \pi^o(\omega_t)$ following (12)
    sample $b_t \sim \pi_t(\cdot|s_t)$, and $a_t$ according to (11)
    update $(\mu_t, \theta_t)$ according to (6), and set $t = t + 1$
end
return $\hat{\pi}_t$ an optimal policy of $\widehat{\mathcal{M}}_t$
```

Next, we provide tools for the sample complexity analysis of GNS. One crucial step is to guarantee that under our sampling scheme, certain random matrices concentrate.

Assumption 6.5. There exists $\kappa > 0$, such that for all $\omega \in \Omega_{\eta/2}(\mathcal{M})$, $u \in \mathbb{S}^{SA-1}$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, the following holds

$$\mathbb{E}^{\pi^o(\omega)}\Big(\lim_{t\to\infty} M_t(u)|s_1 = s, a_1 = a\Big) \le \kappa,$$

where $\mathbb{E}^{\pi^o(\omega)}$ means that the expectation is taken with respect to trajectories generated by policy $\pi^o(\omega)$, and $M_t(u) = \sum_{\ell=1}^t \big(|u^\top\Lambda(\omega)^{-1/2}\phi_\ell|^2 - 1\big)$.

Essentially, Assumption 6.5 guarantees the convergence of $\frac{1}{t}\sum_{\ell=1}^t \phi_\ell\phi_\ell$ towards $\Lambda(\omega)$ when the sample trajectory is generated with the fixed policy $\pi^o(\omega)$. The uniform bound κ across state-action pairs in $\mathcal{S}\times\mathcal{A}$, allocations in $\Omega_\eta(\mathcal{M})$, and all unit vector in $\mathbb{R}^{SA}$ may appear strong. However, it can be shown for instance that if $\mathcal{M}$ is ergodic then $\kappa = \mathcal{O}(t_{\text{mix}}/\eta)$ where t_{mix} is a the mixing time of the MDP $\mathcal{M}$. We provide further discussions on this assumption in appendix B. Now, we present our concentration bounds on random matrices with Markovian data.

Proposition 6.6. *Let ω_k be the optimal allocation used by GNS between $t_k < t \le t_{k+1}$ for some $k \ge 1$. Furthermore, assume that $\omega_k \in \Omega_\eta(\mathcal{M})$. Then, under GNS with the forced exploration* (11)*,*

[5] This claim follows from Berge's maximum theorem. Refer to appendix D for a formal statement.

[6] The non-unicity of the optimal allocation also occurs in best arm identification for linear bandits (see e.g.,[11]). The non-unicity is a consequence of Caratheodory's theorem.

under Assumption 6.5, we have, for all $\varepsilon > 0$, $\delta \in (0,1)$

$$\mathbb{P}\left(\frac{1}{t_{k+1}-t_k}\sum_{t=t_k+1}^{t_{k+1}} \phi_t\phi_t^\top \succeq (1-\varepsilon)\Lambda(\omega_k)\right) \geq 1-\delta,$$

provided that $t_{k+1} - t_k \geq C \max\left(\left(\frac{16\kappa}{\varepsilon}\right)^2, \frac{16\kappa}{3\varepsilon}\right)\left(\log\left(\frac{e}{\delta}\right) + d\right)$ for some universal constant $C > 0$.

The proof of Proposition 6.6 relies on decomposing $\sum_{t=t_k+1}^{t_{k+1}} \phi_t\phi_t^\top$ using Poisson's equation [8] so as to obtain a martingale that can be easily controlled under Assumption 6.5. We defer the proof to Appendix B along with the precise constants.

6.2 Sample complexity guarantees under GNS

Finally, in Theorem 6.7, we present a sample complexity upper bound for GNS.

Theorem 6.7. *The sample complexity of GNS, satisfies for all $\varepsilon > 0$,*

$$\mathbb{E}[\tau] \leq C U^\star_{\mathcal{M},\text{for}}\left(\log\left(\frac{e}{\delta}\right)\right) + o\left(\log\left(\frac{e}{\delta}\right)\right) \tag{13}$$

for some universal constant $C > 0$. Furthermore, the GNS algorithm is (ε, δ)-PAC.

The proof of Theorem 6.7 is slightly more complex than that of Theorem 5.4 due the navigation constraints. We present the proof in Appendix D. Observe that the GNS algorithm attains a sample complexity that matches, up to some multiplication constant and assymptotically (as $\delta \to 0$), the complexity measure $U^\star_{\mathcal{M},\text{for}}\log(1/\delta)$ presented in Theorem 4.4.

7 Conclusion

In this paper, we have first derived instance-dependent lower bounds on the sample complexity of best policy identification in discounted linear MDPs. As of now, these instance-dependent bounds remain challenging to exploit algorithmically. Instead, we proposed a relaxation that links these lower bounds to experiment-design criteria based on the G-optimal design. These criteria lead to the sample complexity measures $U^\star_{\mathcal{M},\text{gen}}\log(1/\delta)$ and $U^\star_{\mathcal{M},\text{for}}\log(1/\delta)$ for the generative model and forward model, respectively. Importantly, these complexity measures are instance-dependent as they exhibit a dependence on the minimum gap $\Delta_{\mathcal{M}}$.

Furthermore, we have established that these experiment design criteria can be exploited algorithmically by proposing the algorithms GSS amd GNS with sample complexity upper bounds matching asymptotically $U^\star_{\mathcal{M},\text{gen}}\log(1/\delta)$ and $U^\star_{\mathcal{M},\text{for}}\log(1/\delta)$, respectively, as $\delta \to 0$. In fact, GSS enjoys a stronger guarantee that holds for all $\delta \in (0,1)$ and matches existing minimax lower bounds (in the episodic case, these bounds are of the order $\Omega(d^2/\varepsilon^2)$). In the forward model, we are the first, to the best of our knowledge, to investigate the problem of ε-best policy indentification for discounted linear MDPs. Notably, we establish, for this model, conditions under which learnability is possible. These conditions are a priori weaker than ergodicity and communication.

As a future direction, we believe that it would be interesting to improve the relaxations of the lower bounds, as well as devising, for the forward model, algorithms with sample complexity guarantees in the moderate confidence regimes.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. volume 125 of *Proceedings of Machine Learning Research*, pages 67–83. PMLR, 09–12 Jul 2020.

[3] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34:25852–25864, 2021.

[4] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[5] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning, 2015.

[6] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. *CoRR*, abs/2010.03531, 2020.

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

[8] Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012.

[9] Jiafan He, Dongruo Zhou, and Quanquan Gu. Minimax optimal reinforcement learning for discounted mdps. *CoRR*, abs/2010.00587, 2020.

[10] Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

[11] Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

[12] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[13] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London, England, 2003.

[14] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

[15] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in Neural Information Processing*, 11, 04 1999.

[16] Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[17] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[18] Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

[19] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint*, arXiv:2005.12900, 2020.

[20] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes, 2021.

[21] Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes, 2020.

[22] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5186–5196. Curran Associates, Inc., 2018.

[23] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

[24] Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[25] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[26] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[27] Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. *arXiv preprint arXiv:2201.11206*, 2022.

[28] Andrew Wagenmaker and Kevin Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *arXiv preprint arXiv:2207.02575*, 2022.

[29] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[30] Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Limiting extrapolation in linear approximate value iteration. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

A Sample Complexity Lower Bounds

A.1 Proof of Propositions 4.1 and 4.3

The proofs of the two propositions follows a standard change-of-measure argument (see [14, 3] and references therein). In our case, this argument is summarized in the following lemma.

Lemma A.1 (Change-of-measure lemma)**.** *For any (ε, δ)-PAC algorithm, for all $\mathcal{M}' \in \mathrm{Alt}_\varepsilon(\mathcal{M})$, we have*

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}[N_{s,a}(\tau)]\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \geq \mathrm{kl}(\delta, 1-\delta)$$

where $N_{s,a}(\tau) = \sum_{t=1}^{\tau} \mathbb{1}\{(s_t, a_t) = (s,a)\}$.

Lemma A.1 is borrowed from [3] (see Lemma 9 in [3]), therefore we omit its proof. Now introducing, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\omega_{s,a} = \mathbb{E}_\mathcal{M}[N_{s,a}(\tau)]/\mathbb{E}_\mathcal{M}[\tau]$, we immediately see that

$$\mathbb{E}_\mathcal{M}[\tau] T_\mathcal{M}(\omega)^{-1} = \mathbb{E}_\mathcal{M}[\tau] \inf_{\mathcal{M}'\in\mathrm{Alt}_\varepsilon(\mathcal{M})} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_\mathcal{M}[N_{s,a}(\tau)]}{\mathbb{E}_\mathcal{M}[\tau]} \mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \geq \mathrm{kl}(\delta, 1-\delta) \tag{14}$$

where we recall that $T_\mathcal{M}(\cdot)$ is defined in (2). From here, to obtain the statements of Proposition 4.1, it suffices to optimize the quantity $T_\mathcal{M}(\omega)$, more precisely we take $T^\star_{\mathcal{M},\text{gen}} = \inf_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}} T_\mathcal{M}(\omega)$. The proof of Propoisition 4.3 follows similarly, but this time we optimize $T_\mathcal{M}(\omega)$ over $\omega \in \Omega(\mathcal{M})$, namely by setting $T^\star_{\mathcal{M},\text{for}} = \inf_{\omega\in\Omega(\mathcal{M})} T_\mathcal{M}(\omega)$. However, note that for thiscase, we require $\delta \to 0$, so that $\mathbb{E}[\tau] \to \infty$, to ensure that $\omega \in \Omega(\mathcal{M})$. Therefore, the lower bound in Propoisition 4.3 is asymptotic in δ.

A.2 Gap bounds and value difference lemmas

Next, we present key *difference* lemmas, that will be useful to relax the optimization problem that appears in the lower bound.

Lemma A.2. *Let $\varepsilon > 0$ and let $\mathcal{M}'$ be an MDP such that $\pi^\star_\mathcal{M} \notin \Pi^\star_\varepsilon(\mathcal{M}')$. Then, we have:*

$$\Delta_\mathcal{M} + \varepsilon \leq \|V^\star_\mathcal{M} - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}\|_\infty + \|Q^\star_\mathcal{M} - Q^\star_{\mathcal{M}'}\|_\infty. \tag{15}$$

Proof of Lemma A.2. $\pi^\star_\mathcal{M} \notin \Pi^\star_\varepsilon(\mathcal{M}')$ implies that $\varepsilon \leq \max_{s\in\mathcal{S}} V^\star_{\mathcal{M}'}(s) - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}(s)$. Denote s the state maximizing this quantity. We have $\pi^\star_{\mathcal{M}'}(s) \neq \pi^\star_\mathcal{M}(s)$. Indeed if it was not the case then

$$\begin{aligned}
V^\star_{\mathcal{M}'}(s) - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}(s) &= Q^\star_{\mathcal{M}'}(s, \pi^\star_{\mathcal{M}'}(s)) - Q^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}(s, \pi^\star_{\mathcal{M}'}(s)) \\
&= \gamma p_{\mathcal{M}'}(s, \pi^\star_{\mathcal{M}'}(s))^\top (V^\star_{\mathcal{M}'} - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}) \\
&\leq \gamma \max_{s'\in\mathcal{S}} (V^\star_{\mathcal{M}'}(s') - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}(s')) \\
&= \gamma (V^\star_{\mathcal{M}'}(s) - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}(s))
\end{aligned}$$

which is a contradiction since $\gamma < 1$. Now, since $\pi^\star_{\mathcal{M}'}(s) \neq \pi^\star_\mathcal{M}(s)$, we have $\Delta_\mathcal{M} \leq V^\star_\mathcal{M}(s) - Q^\star_\mathcal{M}(s, \pi^\star_{\mathcal{M}'}(s))$. We can then write

$$\begin{aligned}
\Delta_\mathcal{M} + \varepsilon &\leq V^\star_\mathcal{M}(s) - Q^\star_\mathcal{M}(s, \pi^\star_{\mathcal{M}'}(s)) + V^\star_{\mathcal{M}'}(s) - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}(s) \\
&= V^\star_\mathcal{M}(s) - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}(s) + Q^\star_{\mathcal{M}'}(s, \pi^\star_{\mathcal{M}'}(s)) - Q^\star_\mathcal{M}(s, \pi^\star_{\mathcal{M}'}(s)) \\
&\leq \|V^\star_\mathcal{M} - V^{\pi^\star_\mathcal{M}}_{\mathcal{M}'}\|_\infty + \|Q^\star_\mathcal{M} - Q^\star_{\mathcal{M}'}\|_\infty.
\end{aligned}$$

□

Lemma A.3. *Let π be any deterministic policy. We have:*

$$\|V^\pi_\mathcal{M} - V^\pi_{\mathcal{M}'}\|_\infty \leq \|Q^\pi_\mathcal{M} - Q^\pi_{\mathcal{M}'}\|_\infty \leq \frac{1}{1-\gamma} \max_{s,a} \left| \phi(s,a)^\top \left(\theta_\mathcal{M} - \theta_{\mathcal{M}'} + \gamma(\mu_\mathcal{M} - \mu_{\mathcal{M}'})^\top V^\pi_\mathcal{M}\right)\right|. \tag{16}$$

Proof of Lemma A.3. For any $s \in \mathcal{S}$, we have $V^\pi_{\mathcal{M}}(s) - V^\pi_{\mathcal{M}'}(s) = Q^\pi_{\mathcal{M}}(s, \pi(s)) - Q^\pi_{\mathcal{M}'}(s, \pi(s))$. Hence the first inequality holds. Now, we can write for any pair $(s, a) \in \mathcal{S} \times \mathcal{A}$

$$Q^\pi_{\mathcal{M}}(s,a) - Q^\pi_{\mathcal{M}'}(s,a) = \phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\pi_{\mathcal{M}}\right) + \gamma p_{\mathcal{M}'}(s,a)^\top (V^\pi_{\mathcal{M}} - V^\pi_{\mathcal{M}'}),$$

so that

$$\begin{aligned} \|Q^\pi_{\mathcal{M}} - Q^\pi_{\mathcal{M}'}\|_\infty &\le \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\pi_{\mathcal{M}}\right)\right| + \gamma \|V^\pi_{\mathcal{M}} - V^\pi_{\mathcal{M}'}\|_\infty \\ &\le \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\pi_{\mathcal{M}}\right)\right| + \gamma \|Q^\pi_{\mathcal{M}} - Q^\pi_{\mathcal{M}'}\|_\infty, \end{aligned}$$

which implies the second inequality. □

Lemma A.4. *We have:*

$$\|V^\star_{\mathcal{M}} - V^\star_{\mathcal{M}'}\|_\infty \le \|Q^\star_{\mathcal{M}} - Q^\star_{\mathcal{M}'}\|_\infty \le \frac{1}{1-\gamma} \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right)\right|. \tag{17}$$

Proof of Lemma A.4. Let $s \in \mathcal{S}$. We have by optimality of $\pi^\star_{\mathcal{M}'}$ that

$$\begin{aligned} V^\star_{\mathcal{M}}(s) - V^\star_{\mathcal{M}'}(s) &= Q^\star_{\mathcal{M}}(s, \pi^\star_{\mathcal{M}}(s)) - Q^\star_{\mathcal{M}'}(s, \pi^\star_{\mathcal{M}'}(s)) \\ &\le Q^\star_{\mathcal{M}}(s, \pi^\star_{\mathcal{M}}(s)) - Q^\star_{\mathcal{M}'}(s, \pi^\star_{\mathcal{M}}(s)) \\ &\le \|Q^\star_{\mathcal{M}} - Q^\star_{\mathcal{M}'}\|_\infty. \end{aligned}$$

$V^\star_{\mathcal{M}'}(s) - V^\star_{\mathcal{M}}(s)$ can be bounded the same way using the optimality of $\pi^\star_{\mathcal{M}}$, so that this inequality is true in absolute value which gives the first inequality. Now, we can write for any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$Q^\star_{\mathcal{M}}(s,a) - Q^\star_{\mathcal{M}'}(s,a) = \phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right) + \gamma p_{\mathcal{M}'}(s,a)^\top (V^\star_{\mathcal{M}} - V^\star_{\mathcal{M}'}),$$

so that

$$\begin{aligned} \|Q^\star_{\mathcal{M}} - Q^\star_{\mathcal{M}'}\|_\infty &\le \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right)\right| + \gamma \|V^\star_{\mathcal{M}} - V^\star_{\mathcal{M}'}\|_\infty \\ &\le \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right)\right| + \gamma \|Q^\star_{\mathcal{M}} - Q^\star_{\mathcal{M}'}\|_\infty \end{aligned}$$

which implies the result. □

Remark A.5. In Lemma A.3 and Lemma A.4 we have used the fact that for any (s,a), $\|p_{\mathcal{M}'}(s,a)\|_1 = 1$, but only with $\mathcal{M}'$ and not with $\mathcal{M}$. When working with the LSE estimators $\hat{\theta}_t$ and $\hat{\mu}_t$, we will construct a MDP $\widehat{\mathcal{M}}_t$ whose transitions probabilities, defined as $\phi(s,a)^\top \mu_t$, may not be actual probability vectors. This is not an issue since these lemmas will only be used with $\widehat{\mathcal{M}}_t$ taking the place of the first MDP which does not require such property.

A.3 Proof of Theorems 4.2 and 4.4

In this section, we prove Theorems 4.2 and 4.4. First, we establish the following Lemma.

Lemma A.6. *Let $\mathcal{M} \in \mathbb{M}$ be a discounted linear MDP. Then, for all $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$, we have:*

$$T_{\mathcal{M}}(\omega) \le U_{\mathcal{M}}(\omega) := \frac{10\sigma(\omega)}{3(1-\gamma)^4(\Delta_{\mathcal{M}} + \varepsilon)^2}. \tag{18}$$

Proof. We are actually going to show (18), but by considering in the definition of $T_{\mathcal{M}}(\omega)$ an infimum over the set of MDPs $\mathcal{M}'$ such that $\pi^\star_{\mathcal{M}} \notin \Pi^\star_\varepsilon(\mathcal{M}')$ – which is larger than $\mathrm{Alt}_\varepsilon(\mathcal{M})$ and thus gives a smaller infimum than $T_{\mathcal{M}}(\omega)^{-1}$. From now on, we consider one such MDP $\mathcal{M}'$. The proof proceeds in two steps:

(Step 1) Lower bounds on the terms $\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a)$. The Kullback-Leibler divergence can be lower bounded using Lemma A.10 (see below). For a given pair (s,a), let $f = r + \gamma V^\star_{\mathcal{M}}(s')$ where r and

s' denote the random reward and the random next state after playing the pair (s,a), respectively. f is almost surely bounded by $(1-\gamma)^{-1}$ and applying Lemma A.10 with this choice for f yields:

$$\begin{aligned}\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) &\geq \frac{6(1-\gamma)^2}{5}\left(\mathbb{E}_{\mathcal{M}(s,a)}[r+\gamma V^\star_{\mathcal{M}}(s')] - \mathbb{E}_{\mathcal{M}'(s,a)}[r+\gamma V^\star_{\mathcal{M}}(s')]\right)^2 \\ &= \frac{6(1-\gamma)^2}{5}\left(\phi(s,a)^\top\left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right)\right)^2.\end{aligned}$$

Summing over all state-action pairs,

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega_{s,a}\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \geq \frac{6(1-\gamma)^2}{5}\left\|\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right\|^2_{\Lambda(\omega)}. \tag{19}$$

(Step 2) Introducing the gaps $\Delta_{\mathcal{M}}$. Putting together Lemma A.2, Lemma A.3 and Lemma A.4 (and choosing $\pi = \pi^\star_{\mathcal{M}}$ in Lemma A.3), we obtain a bound on the quantity $\Delta_{\mathcal{M}} + \varepsilon$ as follows

$$\Delta_{\mathcal{M}} + \varepsilon \leq \frac{2}{1-\gamma}\max_{s,a}\left|\phi(s,a)^\top\left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right)\right|.$$

Now, we can apply Lemma A.9 with $n=1$, $\Delta = \frac{1-\gamma}{2}(\Delta_{\mathcal{M}}+\varepsilon)$, $\Lambda_1 = \Lambda(\omega)$ and ϕ_1 the feature maximizing the term above, and deduce that

$$\begin{aligned}\left\|\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right\|^2_{\Lambda(\omega)} &\geq \frac{(1-\gamma)^2(\Delta_{\mathcal{M}}+\varepsilon)^2}{4\|\phi\|^2_{\Lambda(\omega)^{-1}}} \\ &\geq \frac{(1-\gamma)^2(\Delta_{\mathcal{M}}+\varepsilon)^2}{4\sigma(\omega)}.\end{aligned}$$

Putting the above inequality together with (19) and then taking the infimum over $\mathcal{M}'$, we have:

$$T_{\mathcal{M}}(\omega)^{-1} \geq \frac{3(1-\gamma)^4(\Delta_{\mathcal{M}}+\varepsilon)^2}{10\sigma(\omega)}. \tag{20}$$

□

Proof of Theorem 4.2. Now, to obtain the statement of Theorem 4.2, we optimize the inequality (18) obtained from Lemma A.6 over $\omega \in \Sigma_{S\times A}$, and get

$$T^\star_{\mathcal{M},\text{gen}} = \inf_{\omega\in\Sigma_{S\times A}} T_{\mathcal{M}}(\omega) \leq \frac{10\inf_{\omega\in\Sigma_{S\times A}}\sigma(\omega)}{3(1-\gamma)^4(\Delta_{\mathcal{M}}+\varepsilon^2)} = U^\star_{\mathcal{M},\text{gen}}$$

Now, applying Kiefer-Wolfowitz theorem (see Theorem A.7 below) implies that $\inf_{\omega\in\Sigma_{S\times A}}\sigma(\omega) = d$, and that $\omega^\star(\mathcal{M})$ which achieves the minimum is the so-called G-optimal design (see [17] and references therein). This concludes the proof of Theorem 4.2. □

Proof of Theorem 4.4. To obtain the statement of Theorem 4.4, we use the inequality (18) obtained from Lemma A.6, and optimize it over $\omega \in \Omega(\mathcal{M})$ to get

$$T^\star_{\mathcal{M},\text{for}} = \inf_{\omega\in\Omega(\mathcal{M})} T_{\mathcal{M}}(\omega) \leq \frac{10\sigma^\star_{\mathcal{M},\text{for}}}{3(1-\gamma)^4(\Delta_{\mathcal{M}}+\varepsilon^2)} = U^\star_{\mathcal{M},\text{gen}}$$

□

A.4 Technical lemmas

Theorem A.7 (Kiefer-Wolfowitz [16]). *Let* $\Phi \subseteq \mathbb{R}^d$ *be a finite set such that* span$(\Phi) = d$. *Let* Σ *be the set of probability distributions on* Φ. *The following statements are equivalent:*

(i) $\omega^\star = \arg\min_{\omega\in\Sigma}\max_{\phi\in\Phi}\phi^\top(\sum_{\phi\in\Phi}\omega(\phi)\phi\phi^\top)^{-1}\phi$,

(ii) $\omega^\star = \arg\max_{\omega\in\Sigma}\log\det(\sum_{\phi\in\Phi}\omega(\phi)\phi\phi^\top)$,

(iii) $\max_{\phi\in\Phi} \phi^\top (\sum_{\phi\in\Phi} \omega^\star(\phi)\phi\phi^\top)^{-1}\phi = d$.

Remark A.8. The statement of the Kiefer-Wolfowitz theorem in [16] holds under a much weaker assumption than that of a finite set Φ. For example, if $\Phi = \{\phi(x) : x \in \mathcal{X}\}$ where $\phi : \mathcal{X} \to \mathbb{R}^d$ is a continuous map on some compact set $\mathcal{X}$, then the equivalence between the three statements *(i), (ii)* and *(iii)* still holds.

Lemma A.9. *Let $\Delta > 0$, $\phi_i \in \mathbb{R}^d$ and $\Lambda_i \in \mathbb{R}^{d\times d}$ some positive definite symmetric matrices for $i = 1, \dots, n$. We have:*

$$\inf_{\substack{x\in\mathbb{R}^{n\times d} \\ \sum_{i=1}^n |\phi_i^\top x_i| \geq \Delta}} \sum_{i=1}^n \|x_i\|_{\Lambda_i}^2 = \frac{\Delta^2}{\sum_{i=1}^n \|\phi_i\|_{\Lambda_i^{-1}}^2}. \tag{21}$$

Proof of Lemma A.9. The absolute values can be removed from the constraint $\sum_i |\phi_i^\top x_i| \geq \Delta$, as we can then apply it adding arbitrary signs before each ϕ_i and get the same result since $\|-\phi_i\|_{\Lambda_i^{-1}} = \|\phi_i\|_{\Lambda_i^{-1}}$. The Lagrangian of the problem without the absolute value is

$$\mathcal{L}(x,\nu) = \sum_{i=1}^n \|x_i\|_{\Lambda_i}^2 - \nu\left(\sum_{i=1}^n \phi_i^\top x_i - \Delta\right)$$

and the KKT conditions for optimality are

$$\begin{aligned}
&\forall i,\ 2\Lambda_i x_i - \nu\phi_i = 0,\\
&\nu\left(\Delta - \sum_{i=1}^n \phi_i^\top x_i\right) = 0,\\
&\Delta \leq \sum_{i=1}^n \phi_i^\top x_i,\\
&\nu \geq 0.
\end{aligned}$$

The first one gives $2x_i = \nu\Lambda_i^{-1}\phi_i$. This formula together with the third condition imply that $\nu > 0$, so that the third condition is an equality and

$$\nu = \frac{2\Delta}{\sum_{i=1}^n \phi_i^\top \Lambda_i^{-1}\phi_i} = \frac{2\Delta}{\sum_{i=1}^n \|\phi_i\|_{\Lambda_i^{-1}}^2}.$$

Finally we have

$$x_i = \Delta \cdot \frac{\Lambda_i^{-1}\phi_i}{\sum_{i=1}^n \|\phi_i\|_{\Lambda_i^{-1}}^2},$$

and the solution of the optimization problem is

$$\frac{\sum_{i=1}^n \phi_i^\top \Lambda_i^{-1}\Lambda_i\Lambda_i^{-1}\phi_i}{\left(\sum_{i=1}^n \|\phi_i\|_{\Lambda_i^{-1}}^2\right)^2}\Delta^2 = \frac{\Delta^2}{\sum_{i=1}^n \|\phi_i\|_{\Lambda_i^{-1}}^2}.$$

□

Lemma A.10. *Let α and β be two probability measures and f be a bounded random variable such that $f \geq 0$. Then we have the following inequality:*

$$\mathrm{KL}(\alpha\|\beta) \geq \frac{6}{5\|f\|_\infty^2}(\mathbb{E}_\alpha[f] - \mathbb{E}_\beta[f])^2. \tag{22}$$

Proof of Lemma A.10. We prove that if $\mathbb{E}_\beta[f] = 0$ then

$$\mathrm{KL}(\alpha\|\beta) \geq \frac{6}{5\|f\|_\infty^2}\mathbb{E}_\alpha[f]^2.$$

It then suffices to apply this result to $f - \mathbb{E}_\beta[f]$ and to notice that if $f \geq 0$ then $\|f - \mathbb{E}_\beta[f]\|_\infty \leq \|f\|_\infty$.

Let f be centered with respect to β. Using Donsker-Varadhan's inequality, we know that for any $\lambda > 0$,

$$\mathrm{KL}(\alpha\|\beta) \geq E_\alpha[\lambda f] - \log(E_\beta[\exp(\lambda f)]).$$

Now,

$$\begin{aligned}\mathbb{E}_\beta[\exp(\lambda f)] &\leq E_\beta\left[1 + \lambda f + f^2 \sum_{k=2}^{+\infty} \frac{\lambda^k \|f\|_\infty^{k-2}}{k!}\right] \\ &\leq 1 + \frac{\mathbb{V}_\beta[f]}{\|f\|_\infty^2}\left(e^{\lambda\|f\|_\infty} - \lambda\|f\|_\infty - 1\right) \\ &\leq 1 + \frac{1}{4}\left(e^{\lambda\|f\|_\infty} - \lambda\|f\|_\infty - 1\right).\end{aligned}$$

Using $\log(1+u) \leq u$,

$$\mathrm{KL}(\alpha\|\beta) \geq \mathbb{E}_\alpha[\lambda f] - \frac{1}{4}\left(e^{\lambda\|f\|_\infty} - \lambda\|f\|_\infty - 1\right).$$

Optimizing over λ, by choosing $\lambda = \frac{1}{\|f\|_\infty}\log\left(1 + 4\frac{\mathbb{E}_\alpha[f]}{\|f\|_\infty}\right)$, we get:

$$\mathrm{KL}(\alpha\|\beta) \geq \frac{1}{4}\left(\left(1 + 4\frac{\mathbb{E}_\alpha[f]}{\|f\|_\infty}\right)\log\left(1 + 4\frac{\mathbb{E}_\alpha[f]}{\|f\|_\infty}\right) - 4\frac{\mathbb{E}_\alpha[f]}{\|f\|_\infty}\right).$$

Using Bernstein's inequality $(1+u)\log(1+u) - u \geq \frac{u^2}{2(1+u/3)}$, we finally have

$$\mathrm{KL}(\alpha\|\beta) \geq \frac{\left(4\frac{\mathbb{E}_\alpha[f]}{\|f\|_\infty}\right)^2}{8\left(1 + \frac{4}{3}\frac{\mathbb{E}_\alpha[f]}{\|f\|_\infty}\right)} = \frac{2\mathbb{E}_\alpha[f]^2}{\|f\|_\infty^2 + \frac{2}{3}\|f\|_\infty\mathbb{E}_\alpha[f]} \geq \frac{6}{5\|f\|_\infty^2}\mathbb{E}_\alpha[f]^2.$$

□

B Concentration of Random Matrices and Sampling Rules

In this section, we present all the results related to our sampling rules for both the generative and forward models. An important quantity that arises in in the analysis is the random matrix $P_t = \sum_{\ell=1}^{t} \phi(s_t, a_t)\phi(s_t, a_t)^\top$. To guarantee any form of learnability, the minimum eigenvalue of the matrix P_t has to grow with t sufficiently fast.

B.1 The generative model – Proof of Proposition 5.1

Sampling under the G-optimal design. It may be ambitious to target a sampling allocation that corresponds exactly to the G-optimal design. Instead, we may focus on a solution that is only approximately optimal. We will say that an allocation (or design) $\tilde{\omega}^\star \in \Sigma_{S\times A}$ is an ϵ-approximate G-optimal design if it satisfies

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{\Lambda(\tilde{\omega}^\star)^{-1}} \le (1+\epsilon) \inf_{\omega\in\Sigma} \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{\Lambda(\omega)^{-1}} = (1+\epsilon)d. \tag{23}$$

Such a solution may be obtained efficiently using a Frank-Wolfe algorithm (see [17] and references therein). Classically, existing procedures, that use G-optimal design as a basis for their sampling schemes, do that in a deterministic fashion by requiring a budget of samples ahead [18], or using efficient rounding procedures coupled with a doubling trick [23]. For our purposes, we will simply sample according to the obtained G-optimal design and that will be enough thanks to the concentration results presented below.

We prove the following matrix concentration result, valid for all ϵ-approximate G-optimal designs.

Lemma B.1. *Let $\tilde{\omega}^\star \in \Sigma_{S\times A}$, be an ϵ-approximate G-optimal design for some $\epsilon > 0$ (i.e., satisfying* (23)*). Assume that the sequence of state action pairs $(s_t, a_t)_{t\ge 1}$ are sampled according to $\tilde{\omega}^\star$, then, for all $\delta \in (0,1)$, $\rho > 0$, we have:*

$$\forall t \ge 2(1+\epsilon)\left(\frac{1}{\rho^2} + \frac{1}{3\rho}\right) d \log\left(\frac{2d}{\delta}\right), \qquad \mathbb{P}\left((1-\rho)\Lambda(\tilde{\omega}^\star) \preceq \Lambda(\omega_t) \preceq (1+\rho)\Lambda(\tilde{\omega}^\star)\right) \ge 1-\delta.$$

Remark B.2. Note that the statement of Lemma B.1, along with the fact that $\tilde{\omega}^\star$ is an ϵ-approximate G-optimal design, ensures that the event

$$\frac{d}{1+\rho} \le \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{\Lambda(\omega_t)^{-1}} \le \frac{(1+\epsilon)d}{1-\rho}$$

holds with probability at least $1-\delta$, provided $t \ge 2(1+\epsilon)\big(\frac{1}{\rho^2} + \frac{1}{3\rho}\big) d \log\big(\frac{2d}{\delta}\big)$. Note that the maximum over $\mathcal{S}\times\mathcal{A}$ came for free thanks to the matrix concentration, and this concentration did not require a priori any condition on the finiteness of the set $\mathcal{S}\times\mathcal{A}$. Actually, the above generalizes immediately for any continuous and compact state-action spaces $\mathcal{S}\times\mathcal{A}$, provided we can compute an ϵ-approximate G-optimal design.

Proof of Proposition 5.1. Specializing the result of Lemma B.1 to the G-optimal design $\omega^\star$ and choosing $\rho = 1/2$ gives

$$\forall t \ge \frac{28d}{3} \log\left(\frac{2d}{\delta}\right), \qquad \mathbb{P}\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{P_t^{-1}} \le 2\sigma(\omega^\star)\right) \ge 1-\delta. \tag{24}$$

This is exactly the statement of Proposition 5.1. □

Proof of Lemma B.1. The proof is an application of Matrix Bernstein's inequality [25]. Let $\delta \in (0,1)$ and $t \ge 1$. First, we have:

$$(\tilde{\Lambda}^\star)^{-1/2}\Lambda(\omega_t)(\tilde{\Lambda}^\star)^{-1/2} - I_d = \sum_{\ell=1}^{t} \frac{1}{t}\Big(\big((\tilde{\Lambda}^\star)^{-1/2}\phi(s_\ell, a_\ell)\big)\big((\tilde{\Lambda}^\star)^{-1/2}\phi(s_\ell, a_\ell)\big)^\top - I_d\Big).$$

where we denote $\tilde{\Lambda}^\star = \Lambda(\tilde{\omega}^\star)$. Denote $(X_\ell)_{1\le \ell \le t}$ the summands appearing in the sum above. Note that X_ℓ is a symmetric random matrix that satisfies for all $\ell \ge 1$, $\|X_\ell\| \le \frac{(1+\epsilon)d}{t}$ a.s. and

$\|\mathbb{E}[X_\ell^2]\| \le \frac{(1+\epsilon)d}{t^2}$ for the operator norm. Indeed, we have for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}\left\| \left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)^\top \right\| &= \max_{\|x\|=1} x^\top \left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)^\top x \\ &= \max_{\|x\|=1} \left(\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)^\top x\right)^2 \\ &\le \|\phi(s,a)\|^2_{(\tilde{\Lambda}^\star)^{-1}} \\ &\le (1+\epsilon)d\end{aligned}$$

so that a.s.

$$\|X_\ell\| \le \frac{1}{t}\max\left(\left\|\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s_\ell,a_\ell)\right)\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s_\ell,a_\ell)\right)^\top\right\|, \|I_d\|\right) \le \frac{(1+\epsilon)d}{t}$$

and, since $\mathbb{E}_{(s,a)\sim\tilde{\omega}^\star}\left[\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)^\top\right] = (\tilde{\Lambda}^\star)^{-1/2}\tilde{\Lambda}^\star(\tilde{\Lambda}^\star)^{-1/2} = I_d$,

$$\begin{aligned}\mathbb{E}[X_\ell^2] &\preceq \mathbb{E}_{(s,a)\sim\tilde{\omega}^\star}\left[\left(\frac{1}{t}\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)^\top\right)^2\right] \\ &\preceq \frac{1}{t^2}\mathbb{E}_{(s,a)\sim\tilde{\omega}^\star}\left[\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)^\top\right] \\ &\qquad \times \max_{s,a}\left\|\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)\left((\tilde{\Lambda}^\star)^{-1/2}\phi(s,a)\right)^\top\right\| \\ &\preceq \frac{(1+\epsilon)d}{t^2} I_d.\end{aligned}$$

Now, using Matrix Bernstein's inequality (more precisely, we use Theorem 5.4.1. in [26], see also [25]), we obtain that for all $\rho > 0$,

$$\mathbb{P}\left(\left\|\sum_{\ell=1}^t X_\ell\right\| > \rho\right) \le 2d\exp\left(-\frac{t\rho^2}{2(1+\epsilon)(1+\rho/3)d}\right).$$

This implies that

$$\forall t \ge 2(1+\epsilon)\left(\frac{1}{\rho^2}+\frac{1}{3\rho}\right)d\log\left(\frac{2d}{\delta}\right), \qquad \mathbb{P}\left(\left\|\sum_{\ell=1}^t X_\ell\right\| > \rho\right) \le \delta.$$

Finally, in order to conclude, observe that

$$\|(\tilde{\Lambda}^\star)^{-1/2}\Lambda(\omega_t)(\tilde{\Lambda}^\star)^{-1/2} - I_d\| \le \rho \implies (1-\rho)\tilde{\Lambda}^\star \preceq \Lambda(\omega_t) \preceq (1+\rho)\tilde{\Lambda}^\star.$$

Thus, provided $t \ge 2(1+\epsilon)\big(\frac{1}{\rho^2}+\frac{1}{3\rho}\big)d\log\big(\frac{2d}{\delta}\big)$, it follows that

$$\mathbb{P}((1-\rho)\tilde{\Lambda}^\star \preceq \Lambda(\omega_t) \preceq (1+\rho)\tilde{\Lambda}^\star) \ge \mathbb{P}\left(\|(\tilde{\Lambda}^\star)^{-1/2}\Lambda(\omega_t)(\tilde{\Lambda}^\star)^{-1/2} - I_d\| \le \rho\right) \ge 1-\delta.$$

□

B.2 The forward model

This part is devoted to the proof of Lemma 6.3 and Proposition 6.6. We start by some remarks on Assumption 6.2.

B.2.1 Discussion on Assumption 6.2

We establish that assuming the existence of an (m,λ)-covering policy is weaker then assuming that the underlying MDP $\mathcal{M}$ is ergodic or communicating. Lemma B.3 shows indeed that the former assumption implies the latter.

Lemma B.3. *If an $\mathcal{M}$ is ergodic or communicating, then there exists an (m, λ)-covering policy π for some $m \geq 1$ and $\lambda > 0$.*

Proof of Lemma B.3. If an MDP $\mathcal{M}$ is communicating, then for each state-pair $i = (s, s') \in \mathcal{S} \times \mathcal{S}$, there exists a policy π_i, and an integer $m_i \geq 1$ such that $\mathbb{E}^{\pi_i}[\mathbb{1}\{s_{m_i} = s'\}|s_1 = s] > 0$. Now defining $\pi = \frac{1}{S^2}\sum_{i \in \mathcal{S}\times\mathcal{S}} \pi_i$, we can clearly see that there exists an $m \geq 1$, such that for all $s, s', a' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $\mathbb{E}^{\pi_i}[\mathbb{1}\{s^\pi_m = s', a^\pi_m = a'\}|s_1 = s] > 0$. Now, under the assumption that the feature maps $(\phi(s, a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ span $\mathbb{R}^d$, it follows that

$$\mathbb{E}^\pi \left[\sum_{\ell=1}^{m} \phi(s_\ell, a_\ell)\phi(s_\ell, a_\ell)^\top |s_1 = s\right] > 0$$

Finally, to conclude we note that any ergodic MDP is also communicating, and therefore our results are proven. □

Remark B.4. We can also construct a linear MDP $\mathcal{M}$ that admits an (m, λ)-covering policy, but that is not ergodic nor communicating. A simple example is when the underlying $\mathcal{M}$ has disjoints communicating classes $(\mathcal{C}_1, \dots, \mathcal{C}_n)$, in the sense that whatever the policy used, if the state is initially in C_i then it remains in this class forever. So $\mathcal{M}$ is not communicating. To get the covering property, we need to ensure that for any i, there is a policy such that $\mathbb{E}[\sum_{\ell=1}^{m_i} \phi(s_\ell, a_\ell)\phi(s_\ell, a_\ell)^\top |s_1 = s] > 0$ for all $s \in C_i$, and for some $m_i \geq 1$.

Remark B.5. Assume that there exists an (m, λ)-covering policy, then we make the following observations.

- In Assumption 6.2, the knowledge π_e is without loss of generality. Indeed, if existence of an (m, λ)-covering policy is guaranteed, then it is not difficult to see that the uniform policy is also an (m', λ')-covering policy for some $m' \geq 1$ and $\lambda' > 0$.
- In Assumption 6.2, the knowledge of m may be relaxed. In fact, when using the forced exploration scheme (11), we only need an upper bound on m.

In view of the discussion above, we have established a minimal assumption under which $\lambda_{\min}(P_t)$ grows sufficiently fast (i.e. $\lambda_{\min}(P_t) = \Omega(t^\gamma)$ for some $\gamma > 0$ with high probability).

B.2.2 Proof of Lemma 6.3

Proof. First, let us assume w.l.o.g. that $K := t/m \in \mathbb{N} \setminus \{0\}$. For all $k \in [K]$, $i \in [m]$, let $\phi_{k,i} = \phi(s_{(k-1)m+i}, a_{(k-1)m+i})$, and $x_{k,i} = x_{(k-1)m+i}$. Let us denote for all Denote $k \in [K]$,

$$\Lambda_k = \frac{1}{m}\left(\sum_{i=1}^{m} \phi_{k,i}\phi_{k,i}^\top\right)\left(\prod_{i=1}^{m} x_{k,i}\right)$$

and note that

$$\sum_{\ell=1}^{t} \phi(s_\ell, a_\ell)\phi(s_\ell, a_\ell)^\top = m\sum_{k=1}^{K}\frac{1}{m}\left(\sum_{i=1}^{m}\phi_{k,i}\phi_{k,i}^\top\right) \succeq m\sum_{k=1}^{K}\Lambda_k$$

Furthermore, let $(\mathcal{F}_k)_{k\geq 1}$ denote σ-algebra generated by the sequence $(s_1, a_1, \varepsilon_1, \dots, s_{km-1}, a_{km-1}, x_{km-1})$. By successive use of the tower rule we can easily verify that:

$$\begin{aligned}
\mathbb{E}[\Lambda_k|\mathcal{F}_k] &= \frac{1}{m}\mathbb{E}\left[\left(\sum_{i=1}^{m}\phi_{k,i}\phi_{k,i}^\top\right)\left(\prod_{i=1}^{m}x_{k,i}\right)\Bigg|\mathcal{F}_k\right] \\
&= \mathbb{E}^\pi\left[\Lambda_k|\mathcal{F}_k\right]\left(\prod_{i=1}^{m}\mathbb{P}(x_{k,i} = 1)\right) \\
&\succeq \lambda\prod_{i=1}^{m}\left(\frac{1}{t}\right)^{\frac{1}{2m}} I_d \\
&= \frac{\lambda}{\sqrt{t}} I_d
\end{aligned}$$

which implies that

$$\sum_{k=1}^{K} \mathbb{E}[\Lambda_k | \mathcal{F}_k] \succeq \frac{K\lambda}{\sqrt{t}} \lambda = \frac{\sqrt{t}\lambda}{m} I_d$$

We also have

$$\|\Lambda_k - \mathbb{E}[\Lambda_k | \mathcal{F}_k]\| \le 2 \qquad a.s.$$

and

$$\begin{aligned}
\|\mathbb{E}[(\Lambda_k - \mathbb{E}[\Lambda_k | \, \mathcal{F}_k])^2 \, | \mathcal{F}_k]\| &\le \mathbb{E}[\|\Lambda_k\|^2 | \mathcal{F}_k] \\
&\le \mathbb{E}\left[\left\| \frac{1}{m} \sum_{i=1}^{m} \phi_{k,i} \phi_{k,i}^\top \right\|^2 \left(\prod_{i=1}^{m} x_{k,i} \right)^2 | \mathcal{F}_k \right] \\
&\le \mathbb{E}\left[\left(\prod_{i=1}^{m} x_{k,i} \right) | \mathcal{F}_k \right] \\
&\le \prod_{i=1}^{m} \mathbb{P}(x_{k,i} = 1) \\
&\le \frac{1}{\sqrt{(k-1)m + 1}}.
\end{aligned}$$

Observe that $\sum_{k=1}^{K} \Lambda_k - \mathbb{E}[\Lambda_k | \mathcal{F}_k]$ is a matrix martingale difference. Therefore, applying Matrix Bernstein's inequality for martingales (see e.g. Theorem 1.2 in [24]) gives

$$\mathbb{P}\left(\left\| \sum_{k=1}^{K} \Lambda_k - \mathbb{E}[\Lambda_k | \mathcal{F}_k] \right\| \ge \epsilon \right) \le 2d \exp\left(- \frac{\epsilon^2}{2\sigma_K + 4\epsilon/3} \right),$$

with

$$\sigma_K = \sum_{k=1}^{K} \frac{1}{\sqrt{(k-1)m + 1}} \le \frac{2\sqrt{m(K-1)+1}}{m} \le \frac{2\sqrt{t}}{m}.$$

This leads to:

$$\mathbb{P}\left(\left\| \sum_{k=1}^{K} \Lambda_k - \mathbb{E}[\Lambda_k | \mathcal{F}_k] \right\| \ge \sigma_K \epsilon \right) \le 2d \exp\left(-\sigma_K \min\left(\epsilon^2, \frac{3\epsilon}{2} \right) \right).$$

Now observe that

$$\left\| \sum_{k=1}^{K} \Lambda_k - \mathbb{E}[\Lambda_k | \mathcal{F}_k] \right\| \le \sigma_K \epsilon \implies \sum_{k=1}^{K} \Lambda_k \succeq \sum_{k=1}^{K} \mathbb{E}[\Lambda_k | \mathcal{F}_k] - \sigma_K \epsilon I_d \succeq \frac{\sqrt{t}}{m} (\lambda - 2\epsilon) I_d,$$

Setting $\epsilon = \lambda/4$ gives

$$\mathbb{P}\left(m\lambda_{\min}\left(\sum_{k=1}^{K} \Lambda_k \right) < \frac{\sqrt{t}\lambda}{2} \right) \le 2d \exp\left(- \frac{\sqrt{t}\lambda^2}{8m} \right).$$

Finally, we get:

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{\ell=1}^{t} \phi(s_\ell, a_\ell) \phi(s_\ell, a_\ell)^\top \right) < \frac{\sqrt{t}\lambda}{2} \right) \le 2d \exp\left(- \frac{\sqrt{t}\lambda^2}{8m} \right).$$

□

B.2.3 Proof of Proposition 6.6

We present and establish Proposition B.6, which provides a stronger result than that of Proposition 6.6.

Proposition B.6. *Under Assumption 6.5, under the sampling rule* (11)*, denote by π_k the oracle policy used by GNS between $t_k < t \leq t_{k+1}$, and denote by ω_k its induced allocation. Furthermore, assume that $\omega_k \in \Omega_\eta(\mathcal{M})$. Then, we have for all $\delta > 0$, $\varepsilon \in (0,1)$*

$$\mathbb{P}\left((1+\varepsilon)\Lambda(\omega_k) \succeq \frac{1}{t_{k+1}-t_k}\sum_{t=t_k+1}^{t_{k+1}} \phi(s_t,a_t)\phi(s_t,a_t)^\top \succeq (1-\varepsilon)\Lambda(\omega_k)\right) \geq 1-\delta$$

provided that

$$t_{k+1}-t_k \geq C\max\left(\left(\frac{16\kappa}{\varepsilon}\right)^2, \frac{16\kappa}{3\varepsilon}\right)\left(\log\left(\frac{e}{\delta}\right)+d\right)$$

for some positive constant $C > 0$.

Proof of Proposition B.6. We seek to establish a concentration bound on the $\sum_{t=t_k}^{t_{k+1}}\left(\phi(s_t,a_t)\phi(s_t,a_t)^\top - \Lambda(\omega_k)\right)$. First, we renormalize and instead find a concentration result on the random matrix

$$W \triangleq \sum_{t=t_k}^{t_{k+1}}\left(\left(\Lambda(\omega_k)^{-1/2}\phi(s_t,a_t)\right)\left(\Lambda(\omega_k)^{-1/2}\phi(s_t,a_t)\right)^\top - I_d\right).$$

We know that $\|W\| = \sup_{u\in\mathbb{S}^{SA-1}}|u^\top W u|$ since W is a symmetric matrix. We will use a net argument to establish a concentration on $\|W\|$. We introduce for all $u \in \mathbb{S}^{SA-1}$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, $f_u(s,a) = |u^\top\Lambda(\omega_k)^{-1/2}\phi(s,a)|^2 - 1$, so that we may simply write

$$\sup_{u\in\mathbb{S}^{SA-1}}|u^\top W u| = \sup_{u\in\mathbb{S}^{SA-1}}\left|\sum_{t=t_k}^{t_{k+1}} f_u(s_t,a_t)\right|$$

(Step 1) Using Poisson's equation: We use Poisson's equation to rewrite $\sum_{t=t_k}^{t_{k+1}} f_u(s_t,a_t)$ in a convenient form. First, let us denote the transition kernel under policy $\pi := \pi_k$ (to simplify the notations), $p^\pi(s',a'|s,a) = \pi(a'|s')p(s'|s,a)$, for all $s,s' \in \mathcal{S}$,$a,a' \in \mathcal{A}$. We have to find $g_u : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ such that Poisson's equation $(I-p^\pi)g_u = f_u$ holds. A natural candidate for g_u (see [8]) is choosing

$$g_u = \sum_{j=0}^{\infty}(p^\pi)^j f_u.$$

We note that under Assumption 6.5, we have

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}|g_u(s,a)| \leq \kappa \tag{25}$$

which clearly implies that Poisson's equation is satisfied. To shorten the notation, let $X_t = (s_t,a_t)$. Now by leveraging Poisson's equation, we may write:

$$\begin{aligned}\sum_{t_k<t\leq t_{k+1}} f_u(X_t) &= \sum_{t_k<t\leq t_{k+1}} g_u(X_t) - \mathbb{E}_{X\sim p^\pi(\cdot|X_t)}[g_u(X)]\\ &= \sum_{t_k<t\leq t_{k+1}} g_u(X_t) - \mathbb{E}_{X\sim p^\pi(\cdot|X_{t-1})}[g_u(X)] + \mathbb{E}[g(X_{t_k-1})] - \mathbb{E}[g(X_{t_{k+1}})]\\ &= S_1+S_2+S_3+S_4\end{aligned}$$

where we define

$$
\begin{aligned}
S_1 &= \sum_{t_k < t \leq t_{k+1}} \left(g_u(X_t) - \mathbb{E}_{X \sim p^\pi(\cdot|X_{t-1})}[g_u(X)]\right)(1 - x_t), \\
S_2 &= \sum_{t_k < t \leq t_{k+1}} \left(g_u(X_t) - \mathbb{E}_{X \sim p^\pi(\cdot|X_{t-1})}[g_u(X)]\right)(x_t - p_t) \\
S_3 &= \sum_{t_k < t \leq t_{k+1}} \left(g_u(X_t) - \mathbb{E}_{X \sim p^\pi(\cdot|X_{t-1})}[g_u(X)]\right) p_t \\
S_4 &= \left|\mathbb{E}[g(X_{t_k-1})] - \mathbb{E}[g(X_{t_{k+1}})]\right|
\end{aligned}
$$

where $p_t = \mathbb{E}[x_t] = t^{-1/(2m)}$. Recall that x_t is the Bernoulli r.v. involved in the definition of our policy.

(Step 2) Bounding** S_1, S_2, S_3, S_4**: The terms S_1 and S_2 can be bounded immediately via standard concentration inequalities and bounds on S_3 and S_4 follow immediately from the inequality (25). Indeed:

- Observe that S_1 is a martingale. Therefore, using a standard peeling argument combined with Hoeffding's Lemma and (25), gives a bound on the moment generating function of S_1:

  $$\forall \lambda > 0, \qquad \mathbb{E}[\exp(\lambda S_1)] \leq \exp\left(\frac{(t_{k+1} - t_k)\lambda^2 \kappa^2}{2}\right).$$

  This leads via Markov inequality to the following bound:

  $$\mathbb{P}\left(S_1 > (t_{k+1} - t_k)\varepsilon\right) \leq \exp\left(-\frac{(t_{k+1} - t_k)\varepsilon^2}{2\kappa^2}\right). \tag{26}$$

- Bounding S_2 is immediate via Bernstein's inequality together with (25) by noting that x_t is independent of $x_1, \dots, x_{t-1}$ and $X_1, \dots X_t$ for all $t \geq 1$. Indeed, we have:

  $$
  \begin{aligned}
  \mathbb{P}\left(S_2 > \varepsilon\right) &\leq \exp\left(-\frac{\varepsilon^2}{2\kappa^2 \sum_{t_k < t \leq t_{k+1}} p_t + \frac{4}{3}\kappa\varepsilon}\right) \\
  &\leq \exp\left(-\frac{\varepsilon^2}{2\kappa^2 (t_{k+1} - t_k) p_{t_k} + \frac{4}{3}\kappa\varepsilon}\right),
  \end{aligned}
  $$

  where we used the fact that $(p_t)_{t \geq 1}$ is a non-increasing sequence. After reparametrization, we obtain the equivalent inequality:

  $$\mathbb{P}\left(S_2 > (t_{k+1} - t_k)\varepsilon\right) \leq \exp\left(-\frac{(t_{k+1} - t_k)\varepsilon^2}{2\kappa^2 p_{t_k} + \frac{4}{3}\kappa\varepsilon}\right). \tag{27}$$

- We can easily bound S_3 using (25) as follows

  $$S_3 \leq 2\kappa(t_{k+1} - t_k) p_{t_k}. \tag{28}$$

- We can easily bound S_4 using (25) as follows

  $$S_4 \leq 2\kappa. \tag{29}$$

(Step 3) Putting everything toegether: We use a union bound on and to obtain:

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{t_k < t \leq t_{k+1}} f_u(X_t) > 2(t_{k+1} - t_k)\varepsilon + 2\kappa(t_{k+1} - t_k)p_{t_k} + 2\kappa\right) \\
&\qquad\qquad \leq \exp\left(-\frac{(t_{k+1} - t_k)\varepsilon^2}{2\kappa^2}\right) + \exp\left(-\frac{(t_{k+1} - t_k)\varepsilon^2}{2\kappa^2 p_{t_k} + \frac{4}{3}\kappa\varepsilon}\right) \\
&\qquad\qquad \leq 2\exp\left(-\frac{(t_{k+1} - t_k)}{2} \min\left(\frac{\varepsilon^2}{\kappa^2}, \frac{3\varepsilon}{\kappa}\right)\right).
\end{aligned}
$$

Similarly we have:

$$\mathbb{P}\left(-\sum_{t_k<t\leq t_{k+1}} f_u(X_t) > 2(t_{k+1}-t_k)\varepsilon + 2\kappa(t_{k+1}-t_k)p_{t_k} + 2\kappa\right) \\ \leq 2\exp\left(-\frac{(t_{k+1}-t_k)}{2}\min\left(\frac{\varepsilon^2}{\kappa^2}, \frac{3\varepsilon}{\kappa}\right)\right).$$

Thus, by union bound, we have:

$$\mathbb{P}\left(\left|\sum_{t_k<t\leq t_{k+1}} f_u(X_t)\right| > 4\kappa(t_{k+1}-t_k)(\varepsilon + p_{t_k}) + 2\kappa\right) \leq 4\exp\left(-\frac{(t_{k+1}-t_k)\min\left(\varepsilon^2, 3\varepsilon\right)}{2}\right).$$

(Step 4) Concluding with a net argument: We use an ϵ-net argument with $\epsilon = 1/2$ (see Chap. 4 in [26]) to obtain

$$\mathbb{P}\left(\sup_{u\in\mathbb{S}^{SA-1}}\left|\sum_{t_k<t\leq t_{k+1}} f_u(X_t)\right| > 8\kappa(t_{k+1}-t_k)(\varepsilon + p_{t_k}) + 4\kappa\right) \\ \leq 4(5^d)\exp\left(-\frac{(t_{k+1}-t_k)\min\left(\varepsilon^2, 3\varepsilon\right)}{2}\right).$$

We recall that $t_k = 2^k = t_{k+1} - t_k$, therefore $p_{t_k} = (t_{k+1}-t_k)^{1/2m}$, thus as long as

$$t_k = t_{k+1} - t_k \geq \frac{1}{\varepsilon^{2m}},$$

we obtain the desired result. □

C Least Square Estimation and Stopping Rules

In this appendix, we show the correctness of our stopping rule presented in Lemma 5.3. This result relies fundamentally on our proposed approach of relaxing of the optimization problem characterizing the optimal sample complexity lower bounds. We also present the present the proof of Proposition 5.2 which corresponds to a certain concentration result for the least-squares estimators we use and relies on the properties of linear MDPs. Finally, we highlight that Lemma 5.3 and Proposition 5.2 hold regardless of our the sampling strategy we use, consequently they hold under both the generative and forward model.

C.1 Correctness of the stopping rule - Proof of Lemma 5.3

In the proof of Lemma 5.3, the fact that $U_{\mathcal{M}}\big(\omega\big)^{-1}$ can be upper bounded as follows

$$U_{\mathcal{M}}\big(\omega\big)^{-1} \le \inf_{\mathcal{M}':\pi_t^\star \notin \Pi_\varepsilon^\star(\mathcal{M})} \frac{6(1-\gamma)^2}{5} \left\|\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu'_{\mathcal{M}})^\top V^\star_{\mathcal{M}}\right\|^2_{\Lambda(\omega)} \le T_{\mathcal{M}}(\omega)^{-1} \tag{30}$$

is critical in the analysis and allows us to construct a stopping rule even when $\varepsilon = 0$. This upper bound is the fruit of our relaxation of the lower bound (see proof of Lemma A.6) and justifies the design of our stopping rule as a relaxed generalized likelihood test $\tau = \inf\{t \ge 1 : Z(t) = tU_{\widehat{\mathcal{M}}_t}(\omega_t)^{-1} \ge \beta(\delta, t)\},$. It is worth mentioning that even though the form of $U^\star_{\mathcal{M}}$ can be caracterized by the G-optimal design, it does not a priori tell us how to design a stopping rule without the inequality (30), especially if $\varepsilon = 0$.

Proof of Lemma 5.3. First, let us recall that $\omega_t(s,a) = N_{s,a}(t)/t$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ where $N_{s,a}(t)$ is the number of times the state-action pair (s,a) has been visited up to time t. Now assuming that $\widehat{\mathcal{M}}_t$ is a valid model, then following a similar reasoning as in the proof of Lemma A.6, we can establish that

$$\begin{aligned} U_{\widehat{\mathcal{M}}_t}\big(\omega_t\big)^{-1} &= \frac{3(1-\gamma)^4(\Delta_{\widehat{\mathcal{M}}_t} + \varepsilon)^2}{10\sigma(\omega_t)} \\ &\le \inf_{\mathcal{M}':\pi_t^\star \notin \Pi_\varepsilon^\star(\widehat{\mathcal{M}}_t)} \frac{6(1-\gamma)^2}{5} \left\|\hat{\theta}_t - \theta_{\mathcal{M}'} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}'})^\top \widehat{V}_t^\star\right\|^2_{\Lambda(\omega_t)} \end{aligned} \tag{31}$$

From the inequality (31), we immediately observe that under the event $\pi_t^\star \notin \Pi_\varepsilon^\star(\mathcal{M})$, we have:

$$Z(t) = t\,U_{\widehat{\mathcal{M}}_t}(\omega_t)^{-1} \le \frac{6(1-\gamma)^2}{5} \left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}_t^\star\right\|^2_{t\Lambda(\omega_t)}.$$

Hence, we have:

$$\begin{aligned} \mathbb{P}\left(\tau < +\infty, \hat{\pi} \notin \Pi_\varepsilon^\star(\mathcal{M})\right) &= \mathbb{P}\left(\exists t \ge 1 : Z(t) > \beta(\delta,t), \pi_t^\star \notin \Pi_\varepsilon^\star(\mathcal{M})\right) \\ &\le \mathbb{P}\left(\exists t \ge 1 : \left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}_t^\star\right\|^2_{t\Lambda(\omega_t)} > \frac{5}{6(1-\gamma)^2}\beta(\delta,t)\right) \\ &\le \delta. \end{aligned}$$

The fact that this last probability is bounded by δ is exactly the statement of Proposition 5.2 (proven later in this appendix). □

Remark C.1. Note that in the proof we assumed that $\widehat{\mathcal{M}}_t$ is a valid linear model. This is not required for the derivation inequality 31 (see Remark A.5). However, this is required for Proposition 5.2 to be applied. We provide further remarks on how to adress this technical detail later on when presenting the proof of Proposition 5.2.

C.2 Properties of linear MDPs

Under the linear MDP assumption, the value and action-value functions of any policy admit a linear representation. This is presented in Lemma C.2 below, borrowed from [12]. We provide a proof for completeness.

Lemma C.2. *Let $\mathcal{M}$ be a discounted linear MDP. For any policy π, there exists a vector $\xi_{\mathcal{M}}^{\pi} \in \mathbb{R}^d$ such that for any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, $Q_{\mathcal{M}}^{\pi}(s,a) = \phi(s,a)^\top \xi_{\mathcal{M}}^{\pi}$. Moreover, we have $\xi_{\mathcal{M}}^{\pi} = \theta_{\mathcal{M}} + \gamma \mu_{\mathcal{M}}^\top V_{\mathcal{M}}^{\pi}$ and $\|\xi_{\mathcal{M}}^{\pi}\| \le \sqrt{d}/(1-\gamma)$.*

Proof of lemma C.2. Using the Bellman equation together with the linear assumptions, we directly have $Q_{\mathcal{M}}^{\pi}(s,a) = \phi(s,a)^\top \left(\theta_{\mathcal{M}} + \gamma \mu_{\mathcal{M}}^\top V_{\mathcal{M}}^{\pi}\right)$ (see also [12]). Then

$$\|\theta_{\mathcal{M}} + \gamma \mu_{\mathcal{M}}^\top V_{\mathcal{M}}^{\pi}\| \le \|\theta_{\mathcal{M}}\| + \gamma \| \sum_{s \in \mathcal{S}} |\mu_{\mathcal{M}}(s)| \| \|V_{\mathcal{M}}^{\pi}\|_\infty \le \sqrt{d} + \gamma \frac{\sqrt{d}}{1-\gamma} = \frac{\sqrt{d}}{1-\gamma}.$$

□

In view of Lemma C.2, we know that the set of optimal value functions under the linear MDP assumption (see Definition 3.1) all belong to the following set:

$$\mathcal{V}^\star = \left\{ V \in \mathbb{R}^S : \exists \xi \in \mathbb{R}^d, V(\cdot) = \max_{a \in \mathcal{A}} \phi(\cdot, a)^\top \xi, \ \|\xi\| \le \frac{\sqrt{d}}{1-\gamma} \right\}. \tag{32}$$

A key observation is that we may construct an ϵ-net of $\mathcal{V}^\star$ with respect to the infinity norm $\|\cdot\|_\infty$ with minimal cardinality that only depends exponentially on the dimension d and not the size of the state space S. This observation is made precise in the following lemma, which is borrowed from [12]. We provide a proof for completeness.

Lemma C.3. *Let $\mathcal{N}$ be an ϵ-net of $\mathcal{V}^\star$ with respect to the inifinity norm $\|\cdot\|_\infty$, with minimal cardinality. Then, we have*

$$|\mathcal{N}| \le \left(1 + \frac{2\sqrt{d}}{(1-\gamma)\epsilon}\right)^d$$

Proof of Lemma C.3. Let $V_1, V_2 \in \mathcal{V}^\star$, and let $\xi_1, \xi_2 \in \mathbb{R}^d$ be there corresponding representation as ensured by Lemma C.2. We have

$$\|V_1 - V_2\|_\infty \le \max_{s,a} \|\phi(s,a)^\top(\xi_1 - \xi_2)\| \le \|\xi_1 - \xi_2\|.$$

Therefore, using this parametrization by ξ, an ϵ-net of $\mathcal{V}$ can be constructed from an ϵ-net of an euclidean ball in $\mathbb{R}^d$ of radius $\sqrt{d}/(1-\gamma)$. Such net exists and has a cardinality that is at most $\left(1 + \frac{2\sqrt{d}}{(1-\gamma)\epsilon}\right)^d$ (see e.g., [26]). □

C.3 Self-normalized concentration tool

Proposition C.4 is the key concentration results we use to establish Proposition 5.2 and relies on the self-normalized concentration bound established in [1].

Proposition C.4. *Let $(\mathcal{F}_t)_{t\ge1}$ be a filtration. Let $(\eta_{v,t})_{v\in\mathcal{V},t\ge1}$ be a stochastic process indexed by a time and the subset $\mathcal{V} \subseteq \mathbb{R}^S$ and taking values in $\mathbb{R}^S$. For each $v \in \mathcal{V}$, the process $(\eta_{v,t})_{t\ge1}$ is martingale adapted to $(\mathcal{F}_t)_{t\ge1}$ and satisifies $\sup_{v\in\mathbb{V},t\ge1} \|\eta_{v,t}\|_\infty \le L$. Let $(\phi_t)_{t\ge1}$ be a predictable stochastic process with respect to $(\mathcal{F}_t)_{t\ge1}$, taking values in $\mathbb{R}^d$. Introducing the matrices $\Phi_t = [\phi_1 \quad \dots \quad \phi_t]^\top$ and $E_{v,t} = [\eta_{v,1} \quad \dots \quad \eta_{v,t}]^\top$ and assuming that the following holds:*

(i) for any $v, v' \in \mathcal{V}$, for all $t \ge 1$, $\|\eta_{v,t} - \eta_{v',t}\| \le \|v - v'\|_\infty$,

(ii) the set $\mathcal{V}$ admits for all $\epsilon \in (0, L)$ an ϵ-net $\mathcal{N}_\epsilon$ with respect to the infinity norm $\|\cdot\|_\infty$ of finite cardinality,

then, for all $\delta \in (0,1)$, $\epsilon \in (0,L)$, and $t \ge 1$, we have: the following event

$$\sup_{V \in \mathcal{V}} \left\|\Phi_t^\top E_{v,t}\right\|^2_{\left(\Phi_t \Phi_t^\top + \lambda_t I_d\right)^{-1}} \le 2L^2 \log \left(\frac{|\mathcal{N}_\epsilon| \det\left(\left(\Phi_t^\top \Phi_t + \lambda_t I_d\right) \left(\lambda_t I_d\right)^{-1} \right)^{1/2}}{\delta} \right) + td\epsilon^2 \tag{33}$$

holds with probability at least $1 - \delta$, and where $(\lambda_t)_{t\ge1}$ is a sequence of positive scalars.

Remark C.5. The threshold of the concentration result in Proposition C.4 can be simplied provided we have a good upper bound on $|\mathcal{N}_\epsilon|$, with a propoer choice of ϵ and λ_t. Indeed, considering that

$$\det\left(\left(\Phi_t^\top \Phi_t + \lambda_t I_d\right)\left(\lambda_t I_d\right)^{-1}\right) \leq \frac{1}{\lambda_t^d}\left(\frac{\operatorname{tr}\left(\Phi_t^\top \Phi_t + \lambda_t I_d\right)}{d}\right)^d \leq \frac{1}{\lambda_t^d}\left(\frac{t + d\lambda_t}{d}\right)^d = \left(1 + \frac{t}{d\lambda_t}\right)^d,$$

and assuming we have $|\mathcal{N}_\epsilon| \leq \left(1 + \frac{2L\sqrt{d}}{\epsilon}\right)^d$. Then, choosing $\epsilon = \frac{2L}{\sqrt{t}}$ and $\lambda_t = 1/d$, gives after further basic manipulations the following threshold

$$L^2\left(2\log\left(\frac{1}{\delta}\right) + d\log\left(8e^4 dt^2\right)\right). \tag{34}$$

Proof of Proposition C.4. The process can be easily controlled when focusing on a single $v \in \mathcal{V}$ due to a self-normalized martingale concentration result. In order to control uniformly over the whole set of parameters, we approximate it by a finite net, which raises an error term in the threshold. We then control each parameters individually and conclude with a union bound. In the following, $\delta > 0$, $\epsilon \in (0, L)$ and $t \geq 1$ are fixed. Define the events

$$\mathcal{C}_1 = \left\{\sup_{v\in\mathcal{V}} \|\Phi_t^\top E_{v,t}\|^2_{(\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1}} \leq 2L^2 \log\left(\frac{|\mathcal{N}_\epsilon| \det\left(\left(\Phi_t^\top \Phi_t + \lambda_t I_d\right)\left(\lambda_t I_d\right)^{-1}\right)^{1/2}}{\delta}\right) + td\epsilon^2\right\},$$

$$\mathcal{C}_2 = \left\{\max_{v\in\mathcal{N}_\epsilon} \|\Phi_t^\top E_{v,t}\|^2_{(\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1}} \leq 2L^2 \log\left(\frac{|\mathcal{N}_\epsilon| \det\left(\left(\Phi_t^\top \Phi_t + \lambda_t I_d\right)\left(\lambda_t I_d\right)^{-1}\right)^{1/2}}{\delta}\right)\right\},$$

$$\mathcal{C}_3(v) = \left\{\|\Phi_t^\top E_{v,t}\|^2_{(\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1}} \leq 2L^2 \log\left(\frac{|\mathcal{N}_\epsilon| \det\left(\left(\Phi_t^\top \Phi_t + \lambda_t I_d\right)\left(\lambda_t I_d\right)^{-1}\right)^{1/2}}{\delta}\right)\right\},$$

where the last event is defined for any $v \in \mathcal{N}_\epsilon$. Recall that our goal is to show that $\mathcal{C}_1$ holds with probability at least $1 - \delta$.

(i) Establishing $\forall v \in \mathcal{V}_\epsilon, \mathbb{P}(\mathcal{C}_3(v)) \geq 1 - \delta/|\mathcal{N}_\epsilon|$. This result is a concentration inequality on self-normalized processes. It can be found as Lemma 9 in [1] for example. To apply it, we use the fact that under all the assumptions, for any $V \in \mathcal{V}_\epsilon$, we have $\|x_t(V)\| \leq L + \epsilon \leq 2L$.

(ii) Establishing $\mathbb{P}(\mathcal{C}_2) \geq 1 - \delta$. We can immediately see that $\mathcal{C}_2 = \bigcap_{v\in\mathcal{N}_\epsilon} \mathcal{C}_3(v)$. Then an union bound gives

$$\mathbb{P}[\mathcal{C}_2] \geq 1 - \sum_{v\in\mathcal{N}_\epsilon}\left(1 - \mathbb{P}(\mathcal{C}_3(V))\right) \geq 1 - \sum_{v\in\mathcal{N}_\epsilon} \frac{\delta}{|\mathcal{N}_\epsilon|} = 1 - \delta.$$

(iii) Establishing $\mathbb{P}(\mathcal{C}_1) \geq 1 - \delta$**.** We want to show that $\mathcal{C}_2 \subset \mathcal{C}_1$. Notice that if $v \in \mathcal{V}$ and $v' \in \mathcal{N}_\epsilon$ such that $\|v - v'\|_\infty \leq \epsilon$, then by using assumption (ii) we have

$$\begin{aligned}
\left\|\Phi_t^\top (E_{v,t} - E_{v',t})\right\|^2_{(\Phi_t^\top \Phi_t \lambda_t I_d)^{-1}} &= \left\|(\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1/2} \Phi_t^\top (E_{v,t} - E_{v',t})\right\|^2 \\
&= \sum_{i=1}^d \left| \left((\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1/2}\right)_i^\top \Phi_t^\top (E_{v,t} - E_{v',t}) \right|^2 \\
&\leq \sum_{i=1}^d \left(\sum_{\ell=1}^t \left| \left((\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1/2}\right)_i^\top \phi_\ell \right| \right)^2 \|(E_{v,t} - E_{v',t})\|_\infty^2 \\
&\leq t \sum_{i=1}^d \sum_{\ell=1}^t \left(\left((\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1/2}\right)_i^\top \phi_\ell \right)^2 \max_{1 \leq \ell \leq t} \|(\eta_{v,t} - \eta_{v',t})\|^2 \\
&\leq t \sum_{\ell=1}^t \left\|(\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1/2} \phi_\ell \right\|^2 \|v - v'\|_\infty^2 \\
&\leq t\epsilon^2 \text{tr}\left(\Phi_t (\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1} \Phi_t^\top\right) \\
&= t\epsilon^2 \text{tr}\left((\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1} (\Phi_t^\top \Phi_t + \lambda_t I_d - \lambda_t I_d)\right) \\
&= t\epsilon^2 \left(d - \lambda_t \text{tr}\left((\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1}\right)\right) \\
&\leq t d \epsilon^2,
\end{aligned}$$

and we can finally write

$$\max_{v \in \mathcal{V}} \|\Phi_t^\top E_{v,t}\|^2_{(\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1}} \leq \max_{V \in \mathcal{V}_\epsilon} \|\Phi_t^\top E_{v,t}\|^2_{(\Phi_t^\top \Phi_t + \lambda_t I_d)^{-1}} + t d \epsilon^2,$$

which implies that $\mathcal{C}_2 \subset \mathcal{C}_1$ and concludes the proof. □

C.4 Proof of Proposition 5.2

Proof of proposition 5.2. We show that under any sampling rule the $(1/d)$-regularized least square estimators verify the following concentration inequality: For any $\delta \in (0, 1)$ the events

$$\mathcal{C}(t) = \left\{ \left\| \hat{\theta}_t - \theta_\mathcal{M} + \gamma(\hat{\mu}_t - \mu_\mathcal{M})^\top \widehat{V}_t^\star \right\|^2_{t\Lambda(\omega_t)} \leq \frac{2}{(1-\gamma)^2} \left(2 \log\left(\frac{\sqrt{e}\zeta(2)t^2}{\delta} \right) + d \log\left(8e^4 d t^2\right) \right) \right\}$$

for all $t \geq 1$ hold simultaneously with probability at least $1 - \delta$. More precisely we are going to show that for any $t \geq 1$, $\mathbb{P}(\mathcal{C}(t)) \geq 1 - \frac{\delta}{\zeta(2)t^2}$. The desired result is then shown via a simple union bound over t. It is hard to control this quantity with a dynamic value function, therefore we will control it for all optimal value functions by controlling $\sup_{v \in \mathcal{V}^\star} \left\| \hat{\theta}_t - \theta_\mathcal{M} + \gamma(\hat{\mu}_t - \mu_\mathcal{M})^\top v \right\|^2_{t\Lambda(\omega_t)}$ instead, assuming that $\widehat{\mathcal{M}}_t$ is a valid model, and use a net argument.

Denote $\delta_t = \frac{\delta}{\zeta(2)t^2}$ for clarity. Recall the definitions of the $\frac{1}{d}$-regularized least square estimators $\hat{\theta}_t$ and $\hat{\mu}_t$:

$$\hat{\theta}_t = \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \Phi_t^\top R_t, \qquad \hat{\mu}_t(s) = \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \Phi_t^\top S_t(s),$$

where $\Phi_t = (\phi(s_1, a_1) \;\; \cdots \;\; \phi(s_t, a_t))^\top$, $R_t = (r_1 \;\; \cdots \;\; r_t)^\top$ and $S_t(s) = \left(\delta_{s,s'_1} \;\; \cdots \;\; \delta_{s,s'_t}\right)^\top$. Recall that $t\Lambda(\omega_t) = \Phi_t^\top \Phi_t$. For any $v \in \mathcal{V}$

$$\begin{aligned}
&\hat{\theta}_t - \theta_\mathcal{M} + \gamma(\hat{\mu}_t - \mu_\mathcal{M})^\top v \\
&= \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \left(\Phi_t^\top R_t - \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right) \theta_\mathcal{M} + \gamma \left(\Phi_t^\top S_t^\top - \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right) \mu_\mathcal{M}^\top \right) v \right) \\
&= \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \Phi_t^\top \left(R_t - \Phi_t \theta_\mathcal{M} + \gamma(S_t^\top - \Phi_t \mu_\mathcal{M}^\top) v\right) - \frac{1}{d} \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} (\theta_\mathcal{M} + \gamma \mu_\mathcal{M}^\top v) \\
&= \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \Phi_t^\top E_{v,t} - \frac{1}{d} \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \xi(v)
\end{aligned}$$

where we denote $\xi(v) = \left(\theta_{\mathcal{M}} + \gamma \mu_{\mathcal{M}}^{\top} v\right)$ and define $\eta_{v,t} = r_t - \phi_t^{\top}\theta_{\mathcal{M}} + \gamma(v(s_t') - \phi_t^{\top}\mu_{\mathcal{M}}^{\top} v) = r_t - \mathbb{E}[r_t|\mathcal{F}_{t-1}] + \gamma(v(s_t') - \mathbb{E}[v(s_t')|\mathcal{F}_{t-1}])$ and $E_{v,t} = R_t - \Phi_t\theta_{\mathcal{M}} + \gamma(S_t^{\top} - \Phi_t\mu_{\mathcal{M}}^{\top})v = \begin{pmatrix}\eta_{v,a} & \cdots & \eta_{v,t}\end{pmatrix}^{\top}$. It follows that

$$\begin{aligned}
&\left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^{\top} v\right\|^2_{\Phi_t^{\top}\Phi_t} \\
&\qquad \leq \left\| \left(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d\right)^{-1} \Phi_t^{\top} E_{v,t} - \frac{1}{d}\left(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d\right)^{-1} \xi(v) \right\|^2_{\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d} \\
&\qquad = \left\| \Phi_t^{\top} E_{v,t} - \frac{1}{d}\xi(v) \right\|^2_{(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d)^{-1}} \\
&\qquad \leq 2\left\|\Phi_t^{\top} E_{v,t}\right\|^2_{(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d)^{-1}} + \frac{2}{d^2}\left\|\xi(v)\right\|^2_{(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d)^{-1}}
\end{aligned}$$

Lemma C.2 states that $\|\xi(v)\| \leq \frac{\sqrt{d}}{1-\gamma}$. Since the greatest eigenvalue of $(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d)^{-1}$ can be upper bounded by d, we can finally write

$$\sup_{v\in\mathcal{V}} \left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^{\top} v\right\|^2_{\Phi_t^{\top}\Phi_t} \leq 2\sup_{v\in\mathcal{V}} \left\|\Phi_t^{\top} E_{v,t}\right\|^2_{(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d)^{-1}} + \frac{2}{(1-\gamma)^2}.$$

It is immediate to see that the first two conditions in Proposition C.4 are satisfied by taking $L = (1-\gamma)^{-1}$ and the third one (i.e., (ii)) is given by Lemma C.3. Therefore we can apply the proposition with $\lambda_t = \frac{1}{d}$ and obtain for all $t \geq 1$

$$\mathbb{P}\left(\sup_{V\in\mathcal{V}} \left\|\Phi_t^{\top} E_{v,t}\right\|^2_{(\Phi_t^{\top}\Phi_t + \frac{1}{d}I_d)^{-1}} \leq \frac{1}{(1-\gamma)^2}\left(2\log\left(\frac{1}{\delta_t}\right) + d\log\left(8e^4 dt^2\right)\right)\right) \geq 1 - \delta_t.$$

The event in the bound above directly implies $\mathcal{C}(t)$ and we can finally conclude that $\mathbb{P}(\mathcal{C}(t)) \geq 1 - \delta_t$ for all $t \geq 1$.

□

Remark C.6. As we have mentioned earlier in the proof, we require that $\widehat{\mathcal{M}}_t$ is a valid linear MDP model, which might not be the case under the plain LSE. Luckily, this is not a big issue and can be addressed by estimating $\hat{V}_t^{\star}$ and $\hat{Q}_t^{\star}$ as follows. At time t, we obtain with the LSE the estimates $\hat{\mu}_t$ and $\hat{\theta}_t$. With these estimates we construct the rewards and transitions as follows: $\hat{r}_t(s,a) = \phi(s,a)^{\top}\hat{\theta}_t$, $\hat{p}_t(s'|s,a) = \phi(s,a)^{\top}\hat{\mu}_t(s')$, for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. However, observe that the estimates $\hat{p}_t(\cdot|s,a)$ are not necessarily transition probabilities. Therefore, there is no guarantee that a solution to the Bellman equation exist under $\widehat{\mathcal{M}}_t$. This issue can be solved by using the following truncated Bellman equation

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad \hat{Q}_{t,h}^{\star}(s,a) = \min\left\{\phi(s,a)^{\top}\left(\hat{\theta}_t + \gamma\mu_t^{\top}\hat{V}_{t,h+1}^{\star}\right), \frac{\sqrt{d}}{1-\gamma}\right\}$$

with $\hat{V}_{t,h}^{\star}(s) = \max_{a\in\mathcal{A}} \hat{Q}_{t,h}^{\star}(s,a)$ and $\hat{V}_{t,H+1}^{\star} = 0$ for some H large enough. Then we use $\hat{V}_t^{\star} \triangleq \hat{V}_{t,1}^{\star}$. With this construction, we can guarantee that all $\hat{V}_{t,h}^{\star}$ belong to a set of value functions that has a similar structure to $\mathcal{V}^{\star}$ but with a slightly larger radius. This construction is identical to that considered by Jin et al. [12] for the episodic setting with the LSE and can be viewed as a natural extension to the discounted setting. For our purposes and to keep the exposition simpler, we will simply assume that under the LSE there exists $\hat{V}_t^{\star}$ and $\hat{Q}_t^{\star}$ satisfying:

$$\begin{aligned}
\forall s,a \in \mathcal{S}\times\mathcal{A}, \qquad \hat{Q}_t^{\star}(s,a) &= \phi(s,a)^{\top}\left(\hat{\theta}_t + \gamma\mu_t^{\top}\hat{V}_t^{\star}\right) \\
\hat{V}_t^{\star}(s) &= \max_{a\in\mathcal{A}} \hat{Q}_t^{\star}(s,a).
\end{aligned}$$

and that $\hat{V}_t^{\star} \in \mathcal{V}^{\star}$. This only simplifies the analysis and it is without loss of generality.

D Sample Complexity Analysis

In this appendix, we present the sample complexity analysis of both GSS and GNS.

D.1 A useful perturbation bound

Before, proceeding with the proof of Theorem 5.4 and Theorem 6.7, we present a useful result that allows us to carefully analyze the quantities $U^\star_{\mathcal{M},\text{gen}}$ and $U^\star_{\mathcal{M},\text{for}}$, as we vary the model $\mathcal{M}$. We present Lemma D.1 below, which is valid for both the generative and forward models. Therefore, in what follows, we will abuse notations and use $U^\star_{\mathcal{M}}$ to mean both $U^\star_{\mathcal{M},\text{gen}}$ and $U^\star_{\mathcal{M},\text{for}}$ for any linear MDP $\mathcal{M}$.

Lemma D.1. *For any $t \ge 1$, we have:*

$$\left|(U^\star_{\mathcal{M}})^{-1} - (U^\star_{\widehat{\mathcal{M}}_t})^{-1}\right| \le 6(1-\gamma)^2 \left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t\right\|^2_{\Lambda(\omega_t)} + \left(\frac{5}{4} - \frac{\sigma(\omega^\star)}{\sigma(\omega_t)}\right)(U^\star_{\mathcal{M}})^{-1}. \tag{35}$$

where $U^\star_{\mathcal{M}}$ can be either $U^\star_{\mathcal{M},\text{gen}}$ or $U^\star_{\mathcal{M},\text{for}}$ for any linear MDP model $\mathcal{M}$.

Before proving Lemma D.1, we present Lemma D.2, used as an intermediate step.

Lemma D.2. *we have*

$$|\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}| \le \frac{2}{1-\gamma} \max_{s,a} \left|\phi(s,a)\left(\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t\right)\right|. \tag{36}$$

Proof of Lemma D.2. For clarity, we denote for both MDPs, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\Delta_{s,a} = V^\star(s) - Q^\star(s,a)$, so that $\Delta_{\mathcal{M}} = \min_{s\in\mathcal{S}, a\neq\pi^\star(s)} \Delta_{s,a}$. Let (s,a) be the pair such that $\Delta_{\mathcal{M}} = \Delta_{\mathcal{M},s,a}$. If $a \neq \pi^\star_t(s)$ then $\Delta_{\widehat{\mathcal{M}}_t} \le \Delta_{\widehat{\mathcal{M}}_t,s,a}$ and $(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}) \le (\Delta_{\widehat{\mathcal{M}}_t,s,a} - \Delta_{\mathcal{M},s,a})$. Else, since both MDPs have exactly $|\mathcal{S}|$ optimal state/action pairs (one for each state), the fact that the pair (s,a) is optimal for $\widehat{\mathcal{M}}_t$ but not for $\mathcal{M}$ means that there exists a pair (s',a') optimal for $\mathcal{M}$ but not for $\widehat{\mathcal{M}}_t$, and we have $\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}} \le \Delta_{s',a'}(\widehat{\mathcal{M}}_t) = \Delta_{s',a'}(\widehat{\mathcal{M}}_t) - \Delta_{s',a'}(\mathcal{M})$. Either way, and doing the same reasoning to bound $\Delta_{\mathcal{M}} - \Delta_{\widehat{\mathcal{M}}_t}$, we can find a pair (s,a) such that

$$\begin{aligned}
|\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}| &\le |\Delta_{\widehat{\mathcal{M}}_t,s,a} - \Delta_{\mathcal{M},s,a}| \\
&= |\widehat{V}^\star_t(s) - \widehat{Q}^\star_t(s,a) - V^\star_{\mathcal{M}}(s) + Q^\star_{\mathcal{M}}(s,a)| \\
&= |\widehat{V}^\star_t(s) - V^\star_{\mathcal{M}}(s) + Q^\star_{\mathcal{M}}(s,a) - \widehat{Q}^\star_t(s,a)| \\
&\le \|\widehat{V}^\star_t - V^\star_{\mathcal{M}}\|_\infty + \|\widehat{Q}^\star_t - Q^\star_{\mathcal{M}}\|_\infty.
\end{aligned}$$

The result is then obtained combining the above inequality and Lemma A.4. □

Proof of Lemma D.1. Observe that by triangular inequality, we have:

$$\left|(U^\star_{\mathcal{M}})^{-1} - (U_{\widehat{\mathcal{M}}_t}(\omega_t))^{-1}\right| \le \left|(U^\star_{\mathcal{M}})^{-1} - (U_{\mathcal{M}}(\omega_t))^{-1}\right| + \left|(U_{\mathcal{M}}(\omega_t))^{-1} - (U_{\widehat{\mathcal{M}}_t}(\omega_t))^{-1}\right|$$

and the first term can be rewritten

$$\left|(U^\star_{\mathcal{M}})^{-1} - (U_{\mathcal{M}}(\omega_t))^{-1}\right| = \left(1 - U^\star_{\mathcal{M}}(U_{\mathcal{M}}(\omega_t))^{-1}\right)(U^\star_{\mathcal{M}})^{-1} = \left(1 - \frac{\sigma(\omega^\star)}{\sigma(\omega_t)}\right)(U^\star_{\mathcal{M}})^{-1}.$$

For the second term, setting $u(\omega_t)^{-1} = 3(1-\gamma)^4/(10\sigma(\omega_t))$, we obtain

$$\begin{aligned}
\left|(U_{\mathcal{M}}(\omega_t))^{-1} - (U_{\widehat{\mathcal{M}}_t}(\omega_t))^{-1}\right| &= u(\omega_t)^{-1}\left|(\Delta_{\mathcal{M}}+\varepsilon)^2 - (\Delta_{\widehat{\mathcal{M}}_t}+\varepsilon)^2\right| \\
&= u(\omega_t)^{-1}\left|\left(\Delta_{\widehat{\mathcal{M}}_t} + \Delta_{\mathcal{M}} + 2\varepsilon\right)\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)\right| \\
&= u(\omega_t)^{-1}\left|\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)^2 + 2(\Delta_{\mathcal{M}}+\varepsilon)\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)\right| \\
&\le u(\omega_t)^{-1}\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)^2 + 2\sqrt{\frac{1}{4}u(\omega_t)^{-1}(\Delta_{\mathcal{M}}+\varepsilon)^2}\sqrt{4u(\omega_t)^{-1}\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)^2} \\
&\le u(\omega_t)^{-1}\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)^2 + \frac{1}{4}(U_{\mathcal{M}}(\omega_t))^{-1} + 4u(\omega_t)^{-1}\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)^2 \\
&\le 5u(\omega_t)^{-1}\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)^2 + \frac{1}{4}(U^\star_{\mathcal{M}})^{-1}
\end{aligned}$$

using $(U_{\mathcal{M}}(\omega_t))^{-1} \le (U^\star_{\mathcal{M}})^{-1}$ for the last step. To conclude, it remains to show that

$$\frac{3(1-\gamma)^4}{2\sigma(\omega_t)}\left(\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right)^2 \le 6\left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t\right\|^2_{\Lambda(\omega_t)}.$$

From Lemma D.2, we get:

$$\left|\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right| \le \frac{2}{1-\gamma}\max_{s,a}\left|\phi(s,a)^\top\left(\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t\right)\right|.$$

The final result is obtained by applying Lemma A.9 with $n = 1$, ϕ_1 the feature maximizing the term above and $\Delta = \frac{1-\gamma}{2}\left|\Delta_{\widehat{\mathcal{M}}_t} - \Delta_{\mathcal{M}}\right|$. □

D.2 With the generative model - Proof of Theorem 5.4

Proof of Theorem 5.4. Recall the threshold

$$\beta(\delta,t) = \frac{12}{5}\left(2\log\left(\frac{\sqrt{e}\zeta(2)t^2}{\delta}\right) + d\log\left(8e^4dt^2\right)\right)$$

and the stopping time

$$\tau = \inf\left\{t \ge 1 : Z(t) > \beta(\delta,t)\right\},$$

where $Z(t) = t\,(U_{\widehat{\mathcal{M}}_t}(\omega_t))^{-1}$. In what follows we will use the notation $U^\star_{\mathcal{M}} = U^\star_{\mathcal{M},\text{gen}}$. In order to establish the sample complexity upper bound, we are going to find a time T such that for any $t \ge T$, $\mathbb{P}(\tau > t) = O\big(\frac{1}{t^2}\big)$, so that we can bound $\mathbb{E}[\tau]$ by T plus a constant. Thanks to Lemma D.1, we have $\{\tau > t\} \subset \left\{t\,(U_{\widehat{\mathcal{M}}_t}(\omega_t))^{-1} \le \beta(\delta,t)\right\} \subset \left\{t\,(U^\star_{\mathcal{M}})^{-1} \le \beta(\delta,t) + tB(t)\right\}$, where we set

$$B(t) = 6(1-\gamma)^2\left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t\right\|^2_{\Lambda(\omega_t)} + \left(\frac{5}{4} - \frac{\sigma(\omega^\star)}{\sigma(\omega_t)}\right)(U^\star_{\mathcal{M}})^{-1}.$$

Now, recall that when proving Proposition 5.2, we have shown that, for any $\delta' > 0$ and for any $t \ge 1$, we have:

$$\mathbb{P}\left(\left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t\right\|^2_{t\Lambda(\omega_t)} \le \frac{5}{6(1-\gamma)^2}\beta(\delta',t)\right) \ge 1 - \frac{\delta'}{\zeta(2)t^2}.$$

Moreover, Lemma B.1 states that if $t \ge \frac{28d}{3}\log\left(\frac{2\zeta(2)dt^2}{\delta'}\right)$, then with probability at least $1 - \frac{\delta'}{\zeta(2)t^2}$, we have $\sigma(\omega_t) \le 2\sigma(\omega^\star)$. Choosing $\delta' = 1$ and plugging both bounds in the definition of $B(t)$, we have with an union bound that, for all $t \ge T_1$,

$$\mathbb{P}\left(tB(t) \le 5\beta(1,t) + \frac{3t}{4}(U^\star_{\mathcal{M}})^{-1}\right) \ge 1 - \frac{2}{\zeta(2)t^2},$$

where we define

$$T_1 = \frac{56d}{3}\log(2\zeta(2)d) + \frac{112d}{3}\log\left(\frac{112d}{3}\right) = \frac{56d}{3}\log\left(\frac{6272\zeta(2)d^3}{3}\right),$$

so that according to Lemma D.9, $t \geq T_1$ implies $t \geq \frac{28d}{3}\log(2\zeta(2)dt^2)$. Now to conclude, we only need to show that this event implies $\left\{t\,(U^\star_{\mathcal{M}})^{-1} > \beta(\delta,t) + tB(t)\right\}$ when t is large enough. Assume that $tB(t) \leq 5\beta(1,t) + \frac{3t}{4}(U^\star_{\mathcal{M}})^{-1}$. Since $\delta < 1$ we have $\beta(1,t) < \beta(\delta,t)$ and $\beta(\delta,t) + tB(t) \leq 6\beta(\delta,t) + \frac{3t}{4}(U^\star_{\mathcal{M}})^{-1}$. To show that this is bounded by $t\,(U^\star_{\mathcal{M}})^{-1}$ is equivalent to show that $24\beta(\delta,t) \leq t\,(U^\star_{\mathcal{M}})^{-1}$. Again, we can show that this last bound is true when $t \geq T_2$ thanks to Lemma D.9, where we define

$$T_2 = U^\star_{\mathcal{M}}\frac{576}{5}\left(2\log\left(\frac{\sqrt{e}\zeta(2)}{\delta}\right) + d\log(8e^4 d)\right) + U^\star_{\mathcal{M}}\frac{576(d+2)}{5}\log\left(\frac{576(d+2)}{5}\right).$$

We have shown that, when $t \geq \max(T_1, T_2)$,

$$\mathbb{P}(\tau > t) \leq \mathbb{P}\left(t\,(U^\star_{\mathcal{M}})^{-1} \leq \beta(\delta,t) + tB(t)\right) \leq \mathbb{P}\left(tB(t) > 5\beta(1,t) + \frac{3t}{4}(U_{\widehat{\mathcal{M}}_t}(\omega_t))^{-1}\right) \leq \frac{2}{\zeta(2)t^2}.$$

Therefore, with $T = \max(T_1, T_2)$,

$$\mathbb{E}[\tau] = \sum_{t\geq 0}\mathbb{P}(\tau > t) = \sum_{t=0}^{T-1}\mathbb{P}(\tau > t) + \sum_{t=T}^{+\infty}\mathbb{P}(\tau > t) \leq T + \sum_{t=T}^{+\infty}\frac{2}{\zeta(2)t^2} \leq T + 2.$$

□

D.3 With the forward model - Proof of Theorem 6.7

The proof of Theorem 6.7 is more complex than that of Theorem 5.4 because of the navigation constraints. Indeed, the analysis of the sample complexity of GNS requires a careful definition of certain *good* events under which the stopping rule can be controlled.

D.3.1 Continuity properties of the optimal solution

Lemma D.3. *Let us denote* $\sigma^\star : \mathcal{M} \mapsto \inf_{\omega\in\Omega_\eta(\mathcal{M})}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|_{\Lambda(\omega)^{-1}}$ *and the set of optimal allocations* $C^\star_\eta : \mathcal{M} \mapsto \arg\min_{\omega\in\Omega(\mathcal{M})}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi\|_{\Lambda(\omega)^{-1}}$.

(i) σ *is continuous and convex on* $\Omega_\eta(\mathcal{M})$*, and attains its maximum on this set.*

(ii) $\sigma^\star$ *is continuous in* $\mu_{\mathcal{M}}$*.*

(iii) $C^\star_\eta(\mathcal{M}) \subseteq \Omega_\eta(\mathcal{M})$ *and is non-empty, compact and convex.*

Proof of Lemma D.3. Proof of (i). This result follows from the fact that $\omega \mapsto \operatorname{tr}\left(\Lambda^{-1}(\omega)\phi(s,a)\phi(s,a)^\top\right)$ is a continuous and convex map on $\Omega_\eta(\mathcal{M})$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Thus, taking the maximum of convex and continuous functions results in a continuous and convex function.

Proof of (ii) and (iii). This an immediate consequence of Berge's maximum theorem. Indeed, observe that $\Omega(\mathcal{M})$ is a set of linear constraints that are parametrized continuously by $\mu_{\mathcal{M}}$. Then, note that by assumption, it must hold that the optimums $C^\star(\mathcal{M}) \subseteq \Omega_\eta(\mathcal{M})$, and we know from (i) that σ is continuous on $\Omega_\eta(\mathcal{M})$. Therefore, Berge's maximum theorem applies and we obtain the desired result. □

The following remark is an immediate consequence of Lemma D.3 and the continuity of the mapping that associates to each allocation its oracle policy.

Remark D.4. There exists $\xi > 0$, where for all $\widehat{\mathcal{M}}_t$ such that $\max_{s\in\mathcal{S}}\|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\| \leq \xi$, then for any allocation $\hat{\omega}_t \in \arg\min_{\omega\in\Omega_\eta(\widehat{\mathcal{M}})}\sigma(\omega)$, it must hold that under $\mathcal{M}$, the oracle policy $\pi^o(\hat{\omega}_t)$ induces $\omega_t \in \Omega_{\eta/2}(\mathcal{M})$, and we have

$$\min_{\omega\in\Omega_\eta(\mathcal{M})}\sigma(\omega) \leq \sigma(\omega_t) \leq 2\min_{\omega\in\Omega_\eta(\mathcal{M})}\sigma(\omega).$$

D.3.2 The good events under the sampling rule of GNS

LSE consistency. In view of Remark D.4, whenever $\max_{s\in\mathcal{S}} \|\hat{\mu}_t(s)-\mu_{\mathcal{M}}(s)\| \leq \xi$, it is guaranteed that the allocation $\omega_t \in \arg\min_{\omega\in\Omega_\eta(\widehat{\mathcal{M}})} \sigma(\omega)$ satisifies $\sigma(\omega_t) \leq 2\min_{\omega\in\Omega(\mathcal{M})} \sigma(\omega)$. We will define the following set

$$\mathcal{E}_{1,T} = \bigcap_{t=\lceil\sqrt{T}\rceil}^{T} \left\{ \max_{s\in\mathcal{S}} \|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\| \leq \xi \right\} \tag{37}$$

Now we show that the event $\mathcal{E}_{1,T}$ holds with high probability. To this objective, we first establish Lemma D.5.

Lemma D.5. *Under assumption 6.2, under GNS with forced exploration* (11)*, we have: for all $\varepsilon > 0$, $\delta \in (0,1)$,*

$$\forall t \geq t_1(\delta), \qquad \mathbb{P}\left(\max_{s\in\mathcal{S}} \|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\| \leq \varepsilon \right) \geq 1-\delta,$$

with

$$t_1(\delta) = \frac{C_1}{\lambda^2} \max\left(m^2, \frac{1}{\varepsilon^4}\right) \left(d\log(d) + \log(S) + \log\left(\frac{e}{\delta}\right)\right)^2$$

where C_1 is some positive universal constant.

Lemma D.5 follows from our forced exploration scheme under the minimal learnability Assumption 6.2. We present the proof at the end of this subsection. Now, combining Lemma D.5 together with a union bound over $\lceil\sqrt{T}\rceil \leq t \leq T$, we immediately get:

Lemma D.6. *Under Assumption 6.2, under GNS with forced exploration* (11)*, we have: for all $\varepsilon > 0$, for all $T \geq 1$,*

$$\mathbb{P}\left(\mathcal{E}_{1,T}^c\right) \leq \delta_1(T) = T\exp\left(-\frac{\lambda T^{1/4}}{C_1} \min\left(\frac{1}{m}, \xi^2\right) + d\log(d) + \log(S) + 1 \right) \tag{38}$$

where we recall that (m,λ) are the paramaters of the covering policy π_e used by the sampling rule of GNS (11)*, and C_1 is a positive universal constant.*

Sampling optimally. Under the event $\mathcal{E}_{1,T}$, eventhough it is guaranteed that $\sigma(\omega_t) \leq 2\min_{\omega\in\Omega(\mathcal{M})}\sigma(\omega)$, the sampling rule under GNS still uses a forced exploration and whenever it does not, it samples according to $\pi^o(\omega_t)$. Therefore, we still have to guarantee, that eventually we will be sampling optimally. Define the event

$$\mathcal{E}_{2,T} = \left\{ T \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|_{\left(\sum_{t=1}^{T}\phi(s_t,a_t)\phi(s_t,a_t)^\top\right)^{-1}} \leq \frac{8\sigma^\star(\mathcal{M})}{(1-\varepsilon)} \right\}.$$

We establish Lemma D.7 below, which guarantees that indeed we will eventually sample according to an approximately optimal allocation. Lemma D.7 relies on the concentration result for random matrices with Markovian data established in appendix B (see Proposition 6.6).

Lemma D.7. *Under Assumption 6.4 and Assumption 6.5, under GNS, we have for all $\varepsilon > 0$, for all $T \geq 1$,*

$$\mathbb{P}(\mathcal{E}_{2,T}^c | \mathcal{E}_{1,T}) \leq \delta_2(T) = \frac{\log(T)}{\log(2)} \exp\left(-\frac{\sqrt{T}}{C'} \min\left(\left(\frac{\varepsilon}{16\kappa}\right)^2, \frac{3\varepsilon}{16\kappa} \right) + d + 1 \right).$$

Proof of Lemma D.7. First, we analyze the deviations of the random matrix $\sum_{t=1}^{T}\phi(s_t,a_t)\phi(s_t,a_t)^\top$ in high probability. Let $L, K \geq 0$ be such that

$$t_K < T \leq t_{K+1} \qquad \text{and} \qquad t_{L-1} < \left\lceil\sqrt{T}\right\rceil \leq t_L. \tag{39}$$

We may write

$$\begin{aligned}\sum_{t=1}^{T}\phi(s_t,a_t)\phi(s_t,a_t)^\top &\succeq \sum_{t=1}^{t_L}\phi(s_t,a_t)\phi(s_t,a_t)^\top + \sum_{k=L}^{K-1}\sum_{t=t_k+1}^{t_{k+1}}\phi(s_t,a_t)\phi(s_t,a_t)^\top\\ &\succeq \sum_{k=L}^{K-1}\sum_{t=t_k+1}^{t_{k+1}}\phi(s_t,a_t)\phi(s_t,a_t)^\top,\end{aligned}$$

where we used the fact that $\|\phi(s,a)\| \le 1$ and $t_L \le 2\sqrt{T}$ (by definition of the set $\mathcal{T}$). Now, by Proposition 6.6, we have for all $L \le k \le K+1$

$$\mathbb{P}\left(\sum_{t=t_k+1}^{t_{k+1}}\phi(s_t,a_t)\phi(s_t,a_t)^\top \succeq (t_{k+1}-t_k)(1-\varepsilon)\Lambda(\omega_k)\right) \ge 1-\delta$$

provided that

$$t_k = t_{k+1}-t_k \ge C\max\left(\left(\frac{16\kappa}{\varepsilon}\right)^2, \frac{16\kappa}{3\varepsilon}\right)\left(\log\left(\frac{e}{\delta}\right)+d\right).$$

Thus, by a union bound, we have:

$$\mathbb{P}\left(\sum_{k=L}^{K-1}\sum_{t=t_k+1}^{t_{k+1}}\phi(s_t,a_t)\phi(s_t,a_t)^\top \succeq \sum_{k=L}^{K-1}(t_{k+1}-t_k)(1-\varepsilon)\Lambda(\omega_k)\right) \ge 1-(K-L)\delta$$

provided that

$$t_L \ge \sqrt{T} \ge C\max\left(\left(\frac{16\kappa}{\varepsilon}\right)^2, \frac{16\kappa}{3\varepsilon}\right)\left(\log\left(\frac{e}{\delta}\right)+d\right).$$

Now, note that under the event $\mathcal{E}_{1,T}$, we have:

$$\begin{aligned}\left(\sum_{k=L}^{K}\sum_{t=t_k+1}^{t_{k+1}}\phi(s_t,a_t)\phi(s_t,a_t)^\top\right)^{-1} &\preceq \frac{1}{t_K-t_L}\left(\sum_{k=L}^{K}\frac{(t_{k+1}-t_k)}{t_K-t_L}(1-\varepsilon)\Lambda(\omega_k)\right)^{-1}\\ &\preceq \frac{1}{(1-\varepsilon)(t_K-t_L)}\sum_{k=L}^{K}\frac{t_{k+1}-t_k}{t_K-t_L}\Lambda(\omega_k)^{-1}.\end{aligned}$$

This leads, under the event $\mathcal{E}_{1,T}$, to

$$\begin{aligned}&\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|^2_{\left(\sum_{t=1}^{T}\phi(s_t,a_t)\phi(s_t,a_t)^\top\right)^{-1}}\\ &\qquad\le \frac{1}{(1-\varepsilon)(t_K-t_L)}\sum_{k=L}^{K-1}\frac{t_{k+1}-t_k}{t_K-t_L}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|_{\Lambda(\omega_k)^{-1}}\\ &\qquad\le \frac{\max_{L\le k<K}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|_{\Lambda(\omega_k)^{-1}}}{(1-\varepsilon)(t_K-t_L)}\\ &\qquad\le \frac{2\sigma^\star(\mathcal{M})}{(1-\varepsilon)(t_K-t_L)}\\ &\qquad\le \frac{8\sigma^\star(\mathcal{M})}{(1-\varepsilon)T}\end{aligned}$$

where the last inequality comes from $t_K - t_L \ge T/2 - 2\sqrt{T} \ge T/4$ and holds whenever $T \ge 64$. Thus,

$$\mathbb{P}\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|^2_{\left(\sum_{t=1}^{T}\phi(s_t,a_t)\phi(s_t,a_t)^\top\right)^{-1}} \le \frac{8\sigma^\star(\mathcal{M})}{(1-\varepsilon)T}\middle|\mathcal{E}_{1,T}\right) \ge 1-\frac{\log(T)\delta}{\log(2)}$$

as long as

$$\sqrt{T} \geq C' \max\left(\left(\frac{16\kappa}{\varepsilon}\right)^2, \frac{16\kappa}{3\varepsilon}\right)\left(\log\left(\frac{e}{\delta}\right) + d\right).$$

Choosing $\delta = \delta_2(T)$ to be the parameter satisfying equality in the above inequality concludes the proof. □

Proof of Lemma D.5. Let $s \in \mathcal{S}$, we have:

$$\|\hat{\mu}(s) - \mu_{\mathcal{M}}(s)\|^2 \leq \frac{\|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\|^2_{t\Lambda(\omega_t)+\gamma I_d}}{\lambda_{\min}(t\Lambda(\omega_t) + \gamma I_d)}$$

Now, we know by Lemma 6.3, that

$$\forall t \geq \left(\frac{8m}{\lambda}\log\left(\frac{2d}{\delta}\right)\right)^2, \qquad \mathbb{P}\left(\lambda_{\min}(t\Lambda(\omega_t) + \gamma I_d) \geq \frac{t^{1/2}\lambda}{2} + \gamma\right) \geq 1 - \delta.$$

Next, as an immediate consequence of the self-normalized inequality in Proposition D.8, we have

$$\mathbb{P}\left(\|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\|^2_{t\Lambda(\omega_t)+\gamma I_d} \leq 2\left(2\log\left(\frac{(\gamma^{-1}t+1)^d}{\delta}\right) + \sqrt{\gamma}\|\mu_{\mathcal{M}}(s)\|^2\right)\right) \geq 1 - \delta$$

using the fact that $\|\mu(s)\| \leq \sqrt{d}$ and $\gamma = 1/d$. This gives, for all $t \geq 1$

$$\mathbb{P}\left(\|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\|^2_{t\Lambda(\omega_t)+\gamma I_d} \leq 4\log\left(\frac{e(2dt)^d}{\delta}\right)\right) \geq 1 - \delta.$$

Using a union bound, we obtain that:

$$\mathbb{P}\left(\max_{s\in\mathcal{S}}\|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\|^2_{t\Lambda(\omega_t)+\gamma I_d} \leq 4\log\left(\frac{e(2dt)^d S}{\delta}\right)\right) \geq 1 - \delta.$$

Thus, by a union bound again:

$$\forall t \geq \left(\frac{8m}{\lambda}\log\left(\frac{2d}{\delta}\right)\right)^2, \qquad \mathbb{P}\left(\max_{s\in\mathcal{S}}\|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\|^2 \leq \frac{8}{\sqrt{t}\lambda}\log\left(\frac{2e(2dt)^d S}{\delta}\right)\right) \geq 1 - \delta \tag{40}$$

which leads to

$$\forall t \geq \max\left(\left(\frac{8m}{\lambda}\log\left(\frac{2d}{\delta}\right)\right)^2, \left(\frac{8}{\varepsilon^2\lambda}\log\left(\frac{2e(2dt)^d S}{\delta}\right)\right)^2\right),$$
$$\mathbb{P}\left(\max_{s\in\mathcal{S}}\|\hat{\mu}_t(s) - \mu_{\mathcal{M}}(s)\| \leq \varepsilon\right) \geq 1 - \delta.$$

After further simplifications, using Lemma D.9, we obtain the desired result. □

D.3.3 Proof of Theorem 6.7

The proofs follows from the same arguments as those used in the proof of Theorem 5.4, with the only exception that now we have to analyze $\{\tau > t\}$ under the events $\mathcal{E}_{1,t} \cap \mathcal{E}_{2,t}$

Proof of Theorem 6.7. The proof follows the same strategy as in the proof of Theorem 5.4, except that we will use different events to control $\{\tau > t\}$. More precisely, we will use the events $\mathcal{E}_{1,t}$ and $\mathcal{E}_{2,t}$. In what follow, we will use $U^\star_{\mathcal{M}}$ to denote $U^\star_{\mathcal{M},\text{for}}$ and recall the notation $\omega_{t,s,a} = N_{s,a}(t)/t$ where $N_{s,a}(t)$ is the number of times the state-action pair (s,a) has been visited up to time t. For the sake of clarity, let us also introduce the event

$$\mathcal{E}_{3,t} = \left\{\left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t\right\|^2_{t\Lambda(\omega_t)} \leq \frac{5}{6(1-\gamma)^2}\beta(1,t)\right\}$$

and observe that by Proposition 5.2, we have $\mathbb{P}(\mathcal{E}_{3,t}) \geq 1 - delta_3(t)$ with $\delta_3(t) = 1/(\xi(2)t^2)$. We will analyze the event $\{\tau > t\}$ under the event $\mathcal{E}_{1,t} \cap \mathcal{E}_{2,t} \cap \mathcal{E}_{3,t}$. First, observe that by definition of the stopping rule τ, and by Lemma D.1, we have

$$\{\tau > t\} \subseteq \left\{ t(U^\star_{\mathcal{M}})^{-1} \leq \beta(\delta, t) + tB(t) \right\},$$

where we defined

$$B(t) = 6(1-\gamma)^2 \left\| \hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}^\star_t \right\|^2_{\Lambda(\omega_t)} + \left(\frac{5}{4} - \frac{\sigma(\omega^\star)}{\sigma(\omega_t)} \right) (U^\star_{\mathcal{M}})^{-1}.$$

Next under the event $\mathcal{E}_{1,t} \cap \mathcal{E}_{2,t} \cap \mathcal{E}_{3,t}$, we have

$$tB(t) \leq 5\beta(1, \delta) + \frac{3t}{4}(U^\star_{\mathcal{M}})^{-1}$$

which leads to

$$\{\tau > t\} \cap \mathcal{E}_{1,t} \cap \mathcal{E}_{2,t} \cap \mathcal{E}_{3,t} \subseteq \left\{ \frac{t}{4}(U^\star_{\mathcal{M}})^{-1} \leq \beta(\delta, t) + 5\beta(1, t) \right\}.$$

Following similar computations as in the proof of Theorem 5.4, the event $\left\{ \frac{t}{4}(U^\star_{\mathcal{M}})^{-1} \leq \beta(\delta, t) + 5\beta(1, t) \right\} = \emptyset$ whenever $t \geq t(\delta)$, where

$$t(\delta) = U^\star_{\mathcal{M}} \frac{576}{5} \left(2 \log\left(\frac{\sqrt{e}\zeta(2)}{\delta} \right) + d \log(8e^4 d) \right) + U^\star_{\mathcal{M}} \frac{576(d+2)}{5} \log\left(\frac{576(d+2)}{5} \right).$$

We have just shown that for all $t \geq t(\delta)$, we have

$$\{\tau > t\} \subseteq \mathcal{E}^c_{1,t} \cup \mathcal{E}^c_{2,t} \cup \mathcal{E}^c_{3,t} = (\mathcal{E}^c_{1,t} \cap \mathcal{E}_{2,t}) \cup \mathcal{E}^c_{2,t} \cup \mathcal{E}^c_{3,t}.$$

Hence, we obtain

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{t=1}^{\infty} \mathbb{P}(\tau > t) \\
&\leq \sum_{t=1}^{t(\delta)} \mathbb{P}(\tau > t) + \sum_{t=t(\delta)}^{\infty} \mathbb{P}((\mathcal{E}^c_{1,t} \cap \mathcal{E}_{2,t}) \cup \mathcal{E}^c_{2,t} \cup \mathcal{E}^c_{3,t}) \\
&\leq t(\delta) + \sum_{t=t(\delta)}^{\infty} \mathbb{P}(\mathcal{E}^c_{1,t} \cap \mathcal{E}_{2,t}) + \mathbb{P}(\mathcal{E}^c_{2,t}) + \mathbb{P}(\mathcal{E}^c_{3,t}) \\
&\leq t(\delta) + \sum_{t=t(\delta)}^{\infty} \mathbb{P}(\mathcal{E}^c_{1,t} | \mathcal{E}_{2,t}) + \mathbb{P}(\mathcal{E}^c_{2,t}) + \mathbb{P}(\mathcal{E}^c_{3,t}) \\
&\leq t(\delta) + \sum_{t=t(\delta)}^{\infty} \delta_1(t) + \delta_2(t) + \delta_3(t).
\end{aligned}$$

Now, note that $\sum_{t=t(\delta)}^{\infty} \delta_1(t) + \delta_2(t) + \delta_3(t) < \infty$. Therefore, we conclude by writing

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \leq \limsup_{\delta \to 0} \frac{t(\delta)}{\log(1/\delta)} + \limsup_{\delta \to 0} \frac{\sum_{t=t(\delta)}^{\infty} \delta_1(t) + \delta_2(t) + \delta_3(t)}{\log(1/\delta)} \lesssim U^\star_{\mathcal{M},\mathrm{for}}.$$

□

D.4 Miscelleneous results and concentration tools

Proposition D.8 is an immediate consequence of the self-normalized concentration result established in [1] (see Lemma 9 in [1]) together with a net argument. Therefore, we simply present the result below without a proof.

Proposition D.8. *Let $(\mathcal{F}_t)_{t\geq 1}$ be a filtration. Let $(\eta_t)_{t\geq 1}$ be a stochastic process adapted to $(\mathcal{F}_t)_{t\geq 1}$ and taking values in $\mathbb{R}^p$. Let $(\phi_t)_{t\geq 1}$ be a predictable stochastic process with respect to $(\mathcal{F}_t)_{t\geq 1}$, taking values in $\mathbb{R}^d$. Furthermore, assume that η_{t+1}, conditionally on $\mathcal{F}_t$, is a zero-mean, σ^2-sub-gaussian* [7]*. Then, for all $\delta \in (0,1)$, the following event*

$$\left\| \left(\sum_{\ell=1}^{t} \phi_\ell \phi_\ell^\top + \lambda I_d \right)^{-1/2} \left(\sum_{\ell=1}^{t} \phi_\ell \eta_\ell^\top \right) \right\|^2 \leq 4\sigma^2 \log \left(\frac{5^p \det((\lambda^{-1}(\sum_{\ell=1}^{t} \phi_\ell \phi_\ell^\top) + I_d))}{\delta} \right)$$

holds with probability at least $1-\delta$.

Lemma D.9. *Let $a, b > 0$. A sufficient condition for $t > a\log(t) + b$ to hold is that $t \geq 2a\log(2a) + 2b$.*

Proof of Lemma D.9. Let $t \geq 2a\log(2a) + 2b$. Then

$$t \geq a\frac{t}{2a} + \frac{t}{2} > a\log\left(\frac{t}{2a}\right) + a\log(2a) + b \geq a\log(t) + b.$$

□

[7] We say that a random variable is σ^2-sub-gaussian, if for all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda^2\sigma^2/2\right)$.

E Numerical Experiments

We assess the performance of GSS on toy examples.

Linear MDPs. We consider the following linear MDPs with dimension $d = S^2$, where S denotes the number of states. the state space is $\mathcal{S} = \{1, \dots, S\}$, and the actions are $\mathcal{A} = \{(a_s, i), s \in \mathcal{S}, i = 1, \dots, A_0\}$. Denote by $e_{s,s'}$ the unit vector in $\mathbb{R}^d$ in the direction (s, s'). The feature vectors are defined as follows:

$$\phi(s, (a_{s'}, i)) = \frac{A_0 + 1 - i}{A_0 + 1} e_{s,s'} + \frac{i}{A_0 + 1} e_{s,s'+1}.$$

The expected rewards are defined by the vector $\theta \in \mathbb{R}^d$ with for all $(s, s') \in \mathcal{S}$, $\theta_{(s,s')} = r_s$, i.e., the reward does not depend on the selected action. Rewards are Bernoulli with means θ. The transition probabilities are defined through μ by: for all s, s', s” $\in \mathcal{S}$,

$$\mu(s'')_{(s,s')} = (1 - \rho)\mathbb{1}_{s'=s''} + \frac{\rho}{S}.$$

Choosing for example $S = 4$ and $A_0 = 3$, when in state s the player will observe an expected reward of r_s regardless of the action, and the available transition distributions will be

$$(1-\rho)\begin{pmatrix} 0.75 & 0.25 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 \\ 0 & 0.75 & 0.25 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.25 & 0.75 & 0 \\ 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.25 & 0.75 \end{pmatrix} + \rho \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

where each row represents an action. Assume that the states are ordered by increasing expected reward. Then the optimal policy is the one choosing the last action from all states. Moreover, the gap can be shown to be equal to

$$\Delta = \gamma(1 - \rho)\frac{r_S - r_{S-1}}{(A_0 + 1)}.$$

Parameters. We are going to plot the sample complexity of the GSS algorithm as a function of d and Δ, which are respectively controlled by S and ρ. The values used for the non varying parameters are the following : $S = 4$, $A_0 = 3$, $\rho = 0.2$, $\gamma = 0.9$, $\delta = 0.05$ and $\varepsilon = 0$. Choosing $\varepsilon = 0$ allows us to highlight the capability of our algorithm to produce an instance-specific sample complexity for optimal policy identification. Finally, we choose the expected rewards to be $(r_s)_{s\in\mathcal{S}} = \left(\frac{1}{2S}, \frac{3}{2S}, \dots, \frac{2S-1}{2S}\right)$.

Stopping rules. We will compare the performance of GSS with that of an algorithm sharing the same sampling rule as GSS but with an optimal stopping rule. The latter stops whenever the algorithm has identified the best policy. For GSS, we use the a stopping threshold scaling as our theoretical threshold but with different constants. Specifically, the threshold is $\beta_{\text{mod}}(\delta, t) = \frac{12}{5}\left(2c_1 \log\left(\frac{1}{\delta}\right) + c_2 d \log\left(8e^4 dt^2\right)\right)$ with $c_1 = 10^{-4}$ and $c_2 = 1.25 \times 10^{-6}$. We run the GSS algorithm with this adjusted threshold and also replace $(1-\gamma)^{-4}$ with $(1-\gamma)^{-1}$ when computing $U_{\widehat{\mathcal{M}}_t}(\omega_t)$ as the dependency in γ is known to be sub-optimal.

In order to fasten the computation, $\widehat{\mathcal{M}}_t$ is computed and the stopping rule tested only every power of 1.2. Moreover, the transition matrix of the estimate $\widehat{\mathcal{M}}_t$ is projected onto the simplex before computing $\Delta(\widehat{\mathcal{M}}_t)$ by value iteration.

Results. Every point plotted corresponds to values averaged over $N = 500$ runs. The sample complexity and the performance of the optimal policy estimated at various steps are presented in Figure 1.

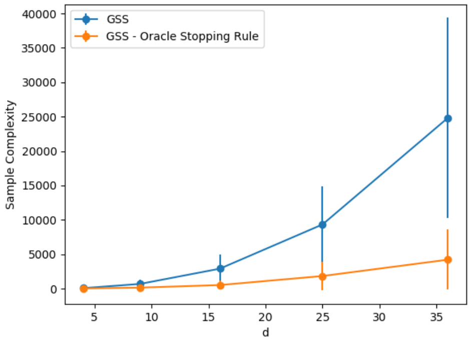

Figure 1: Sample complexity vs. d.

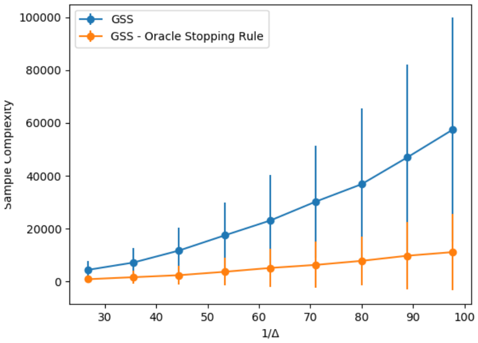

Figure 2: Sample complexity vs. Δ.

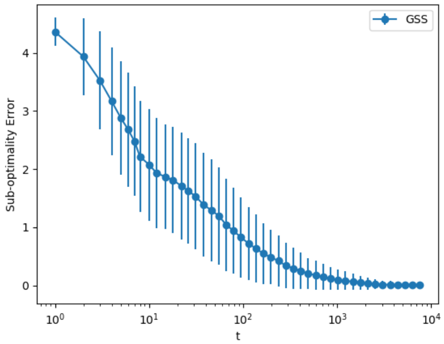

Figure 3: Sub-optimality of the optimal estimated policy after t steps.

The stopping rule of GSS leads to a sample complexity of the same order of magnitude as that of GSS with the oracle stopping rule. But GSS stops later, as expected. the sample complexity increases with both with the dimension d and the inverse of the gap Δ. For reference, the first curve is expected to show a growth of d^3 because the gap is proportional to $1/S$ so d^2/Δ^2 grows as $S^6 = d^3$. The second curve is expected to grow as $1/\Delta^2$. Note that we selected $\delta = 0.05$. In all experiments, the proportion of runs where the algorithm did not identify the best policy was under this threshold. Finally, for Figure 1(c), we plot the *sub-optimality* error of the optimal policy estimated at various steps. This error for the estimated policy $\hat{\pi}$ is defined as $\|V^{\star}\breve{}\,V^{\hat{\pi}}\|_{\infty}$.