Rapport de stage de M1
Identification de stratégie optimale dans les Processus de Décision de Markov Linéaires

Jérôme Taupin

Février-Juin 2022

Présentation générale du stage

J'ai effectué mon stage de M1 à l'université KTH à Stockholm dans une équipe spécialisée en machine learning. J'étais durant les quatre mois sur la supervision d'Alexandre Proutière, en collaboration avec un de ses doctorants, Yassir Jedra. En terme d'organisation j'étais sur site presque tous les jours mais avec beaucoup de liberté. Je travaillais avec Yassir très régulièrement et nous faisions des réunions avec Alexandre une ou deux fois par semaine.

Notre cadre d'étude était celui des Processus de Décisions de Markov (MDP) : On dispose d'un ensemble $\mathcal{S}$ d'états et d'un ensemble $\mathcal{A}$ d'actions. Dans un état donné le joueur doit choisir une action, et à la paire (état, action) sont associés un gain aléatoire et une transition aléatoire vers un autre état. On appelle *policy* une fonction $\pi : \mathcal{S} \mapsto \mathcal{A}$, c'est-à-dire la décision pour chaque état d'une action à choisir. L'efficacité d'une *policy* est évaluée via sa *value function*, définie comme l'espérance de la somme cumulée des gains suivant cette *policy*. Une *policy* est dite optimale si elle maximise cette quantité et ε-optimale si sa *value function* est proche de l'optimale à ε près. Il existe deux modèles dans la littérature pour les MDPs, *discounted* et *episodic*, qui sont définis précisément dans l'article que nous avons écrit qui suit cette introduction. Nous avons essentiellement étudié le premier modèle mais avons ensuite adapté nos résultats au second.

L'objectif de notre travail durant mon stage était de concevoir un algorithme (δ, ε)-PAC efficace, c'est-à-dire testant un certain nombre de fois le résultat de paires $(s, a) \in \mathcal{S} \times \mathcal{A}$ pour une instance $\mathcal{M}$ de MDP inconnue puis retournant une *policy* qui soit ε-optimale avec probabilité au moins $1 - \delta$. Par "efficace" on entend qu'on cherche à minimiser la *sample complexity*, i.e. le nombre total de *samples* effectués. Nous avons d'abord travaillé sans contraintes de navigation, c'est-à-dire que l'algorithme a une liberté totale dans le choix des paires à tester. Nous avons ensuite adapté notre algorithme pour suivre ces contraintes, ce qui signifie qu'après avoir testé une paire l'algorithme doit obligatoirement observer l'état obtenu comme transition. Ce cadre d'étude correspond à un apprentissage "en temps réel" et est celui qui serait utilisé en vue d'un objectif de minimisation de regret. Nous ne nous intéressons pas ici au regret mais ces contraintes restent plus applicables en pratique d'où l'intérêt de les étudier. L'ajout de ces contraintes n'est pas présent dans l'article car il nous manquait quelques détails techniques mais le raisonnement global est le même.

Plus précisément, un tel travail avait déjà été fait par Alexandre auparavant et nous l'avons adapté en ajoutant une hypothèse de structure, dite de linéarité : nous considérons que les gains moyens et probabilités de transition sont paramétrés par des *features* $\phi(s, a)$ encodant la dépendance en la paire (s, a) de la manière suivante :

— Pour tout $s' \in \mathcal{S}$ il existe un vecteur $\mu(s')$ tel que la probabilité de transition vers s' depuis une paire $(s, a) \in \mathcal{S} \times \mathcal{A}$ est $\phi(s, a)^\top \mu(s')$.

— Il existe un vecteur θ tel que le gain moyen depuis une paire $(s, a) \in \mathcal{S} \times \mathcal{A}$ est $\phi(s, a)^\top \theta$.

Les vecteurs $\phi(\cdot, \cdot)$, $\mu(\cdot)$ et θ appartiennent à un espace de dimension $d \ll |\mathcal{S}| \cdot |\mathcal{A}|$. De plus, θ et μ sont propres à un MDP mais les *features* en sont indépendantes et connues à l'avance. Cette hypothèse de linéarité

s'apparente à une hypothèse de faible rang, le cas général correspondant simplement à $d = |\mathcal{S}| \cdot |\mathcal{A}|$. Notre objectif est alors de retrouver les mêmes résultats que sans cette hypothèse, mais en remplaçant la dépendance en $|\mathcal{S}| \cdot |\mathcal{A}|$ par d.

La structure globale de notre raisonnement est la suivante : Dans un premier temps nous avons établi une borne inférieure de la forme $T(\mathcal{M}, \omega) \log(\frac{1}{\delta})$ sur la *sample complexity* de n'importe quel algorithme (δ, ε)-PAC, dépendant de l'instance $\mathcal{M}$ du MDP et de l'allocation ω de l'algorithme, c'est à dire des fréquences auxquelles chaque paire a été testée. Notons qu'en minimisant par rapport à ω et en maximisant par rapport à $\mathcal{M}$ on obtient une borne minimax. Nous avons ensuite majoré pour tout $\mathcal{M}'$ et ω', $T(\mathcal{M}', \omega')$ par une quantité $U(\mathcal{M}', \omega')$ plus simple qui peut être calculée en pratique pour être utilisée dans un algorithme si $\mathcal{M}'$ et ω' sont connus. Nous avons remarqué que, sans contraintes de navigation, l'allocation $\omega^\star$ minimisant $U(\mathcal{M}, \cdot)$ ne dépend pas de $\mathcal{M}$ mais seulement des *features*. Ceci suggère que $\omega^\star$ est l'allocation optimale pour apprendre n'importe quel MDP. De plus elle peut être calculée en résolvant un problème d'optimisation avant même de débuter l'apprentissage.

Notre algorithme calcule donc cette allocation optimale, puis *sample* des paires aléatoirement suivant cette dernière. À partir de ces *samples* il considère les estimateurs des moindres carrés (LSE) $\hat{\theta}_t$ de $\theta_{\mathcal{M}}$ et $\hat{\mu}_t(\cdot)$ de $\mu_{\mathcal{M}}(\cdot)$ après t *samples*, ainsi que $\widehat{\mathcal{M}}_t$ le MDP associée à ces estimateurs. Il s'arrête une fois que $t\, U(\widehat{\mathcal{M}}_t, \omega_t)^{-1} \gtrsim \log(\frac{1}{\delta}) + d \log(t)$ où ω_t est l'allocation des *samples* effectuées jusqu'à t et $\gtrsim$ cache des constantes numériques connues. Le fait que cette règle d'arrêt est (δ, ε)-PAC, c'est-à-dire quelle permet de retourner une *policy* ε-optimale avec probabilité $1-\delta$, se réduit à un résultat de concentration sur l'erreur des LSE. Cette implication est obtenue en utilisant une inégalité intermédiaire entre T et U faisant apparaître θ et μ. La forme de cette condition d'arrêt permet de montrer à l'aide d'un argument de continuité sur U^{-1} que le nombre de *samples* τ de notre algorithme vérifie $\mathbb{E}[\tau] = \mathcal{O}\big(U(\mathcal{M}, \omega^\star)\big(\log(\frac{1}{\delta}) + d\big)\big)$ où $\mathcal{O}$ cache les termes logarithmiques. Cette borne diffère de la borne minimax connue uniquement par ces termes logarithmiques, mais conserve une adaptation à l'instance particulière $\mathcal{M}$. De plus elle est valable pour tout δ et non seulement asymptotiquement pour $\delta \to 0$ comme c'est souvent le cas dans la littérature.

Il m'a fallu environ deux mois pour bien intégrer le sujet et parvenir à une première ébauche de l'algorithme décrit précédemment. Nous avons passé le temps restant à régler tous les détails et adapter ce raisonnement au modèle *episodic* ainsi qu'à ajouter les contraintes de navigation.

La suite de ce rapport est l'article que nous avons écrit et publié sur arXiv.

Best Policy Identification in Linear MDPs

Jérôme Taupin*[1,2], Yassir Jedra†[1], and Alexandre Proutière‡[1]

[1]KTH Royal Institute of Technology
[2]ENS

Abstract

We investigate the problem of best policy identification in discounted linear Markov Decision Processes in the fixed confidence setting under a generative model. We first 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. One of these exhibits a sample complexity upper bounded by $\mathcal{O}(\frac{d}{(\varepsilon+\Delta)^2}(\log(\frac{1}{\delta})+d))$ 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. We extend our algorithm to episodic linear MDPs.

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 or episodic 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. [14, 12, 8, 5, 21, 3, 18, 10, 6, 7] and more recently adopting an instance-specific analysis [20, 19]. According to the aforementioned 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 RL algorithms can address very small problems only), 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, and more specifically of deep RL algorithms, our theoretical understanding of these methods remain limited. In this paper, we investigate the *linear* MDPs, where linear functions are used to approximate the system dynamics and rewards. We propose computationally simple algorithms solving the best policy identification problem in the fixed confidence setting, and analyze their sample complexity. More precisely our contributions are as follows.

*jerome.taupin@ens.psl.eu
†jedra@kth.se
‡alepro@kth.se

1. For linear MDPs with discount factor γ, we first derive instance-specific sample complexity lower bounds satisfied by any (ε, δ)-PAC algorithm. Inspired by these lower bounds, we develop GSS (G-Sampling-and-Stop), an (ε, δ)-PAC algorithm that blends G-optimal design method and Least-Squares estimators. In the generative model (when in each round, the algorithm can sample a transition and reward from any (state, action) pair), we show that the expected sample complexity of GSS scales at most as $\frac{d(1-\gamma)^{-4}}{(\Delta(\mathcal{M})+\varepsilon)^2}(\log(\frac{1}{\delta}) + d)$ (up to logarithmic factors), where $\Delta(\mathcal{M})$ is an appropriately defined instance-specific sub-optimality gap that depends on the MPD $\mathcal{M}$.

2. For linear MDPs with finite time horizon H, we apply the same approach as that used for discounted MDPs. We derive instance-specific sample complexity lower bounds, and based on these bounds, extend the design of GSS to the episodic setting. The analysis of GSS reveals that its expected sample complexity scales at most as $\frac{dH^4}{(\Delta(\mathcal{M})+\varepsilon)^2}(\log(\frac{1}{\delta}) + d)$ (up to logarithmic factors).

The paper is organized as follows. We start with a literature review in the next section. Sections 3, 4, and 5 are devoted to discounted linear MDPs in the generative model. Finally in Section 6, we extend our results to episodic linear MDPs.

2 Related Work

Linear models in RL have attracted a lot of attention over the last few years. We distinguish episodic and discounted MDPs.

Episodic linear MDPs. Most of the studies have aimed at devising algorithms minimizing regret. Jin et al. [11] 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. He et al. [9] present UCRL-VTR, a confidence based algorithm adapted to the linear MDP setting. The algorithm achieves a gap dependent regret of order $\tilde{\mathcal{O}}(\frac{d^2H^5}{\Delta_{\min}}\log\left(\frac{T}{\delta}\right)^3)$.
When it comes to best policy identification problems, researchers have used different approaches. In [26], Wagenmaker et al. aim at identifying an ϵ-optimal policy identification objective. They establish a sample complexity minimax lower bound of order $\Omega(\frac{d^2H^2}{\epsilon^2})$, and propose an a reward-free algorithms with sample complexity of order $\tilde{\mathcal{O}}(\frac{d(\log(1/\delta)+d)H^5}{\epsilon^2})$.
In a subsequent work, Wagenmaker et al. [25] introduce PEDEL, an elimination based algorithm with instance-specific sample complexity guarantees. In the worst case, the sample complexity upper bound scales as $\tilde{\mathcal{O}}(\frac{dH^5(dH^2+\log(1/\delta))}{\epsilon^2})$. 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:

$$C_0 H^4 \sum_{h=1}^{H} \inf_{\Lambda_{exp}} \max_{\pi\in\Pi} \frac{\|\phi_{\pi,h}\|_{\Lambda_{exp}^{-1}}}{\max\{V^\star(\Pi) - V^\pi, \Delta_{\min}(\Pi), \epsilon^2\}} \log\left(\frac{|\Pi|}{\delta}\right) + C_1,$$

where $C_0 = \log\left(\frac{1}{\epsilon}\right)\text{polylog}\left(H, \log(1/\epsilon)\right)$ and $C_1 = \text{poly}\left(d, H, \frac{1}{\lambda^\star_{\min}}, \log(1/\delta), \log(1/\varepsilon), \log(|\Pi|)\right)$. 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

$$C_0 H^4 \sum_{h=1}^{H} \inf_{\Lambda_{exp}} \max_{\pi\in\Pi_\epsilon} \frac{\|\phi_{\pi,h}\|_{\Lambda_{exp}^{-1}}}{\max\{V^\star - V^\pi, \epsilon^2\}}(dH^2 + \log(\frac{1}{\delta})) + C_1.$$

In Zanette et al. [28], 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 [27], 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. [16] 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 [29], 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 Discounted linear MDPs

We consider an infinite time-horizon MDP with a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$. Each (state, action) pair (s,a) is associated to a feature $\phi(s,a) \in \mathbb{R}^d$, and we assume that the feature map ϕ is known to the learner. Without loss of generality, we assume that $\|\phi(s,a)\| \leq 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and that the features $(\phi(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ cover $\mathbb{R}^d$. An MDP $\mathcal{M}$ is defined through its dynamics $p_{\mathcal{M}}$ and reward distributions $q_{\mathcal{M}}$. 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 $q_{\mathcal{M}}(s,a,\cdot)$. We assume that $q_{\mathcal{M}}(s,a,\cdot)$ is absolutely continuous w.r.t. a measure ν with support included in $[0,1]$, and we denote by $r_{\mathcal{M}}(s,a) = \mathbb{E}[q_{\mathcal{M}}(s,a)]$ the expected reward of the (state, action) pair (s,a). The MDP is linear, which means that its dynamics and expected rewards can be parametrized as follows:

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

where $\mu_{\mathcal{M}}$ is a family of d measures over $\mathcal{S}$, seen as a $\mathcal{S} \times d$ dimensional matrix, and $\theta_{\mathcal{M}} \in \mathbb{R}^d$. We will assume that $\|\theta_{\mathcal{M}}\| \leq \sqrt{d}$ and $\left\| \sum_{s\in\mathcal{S}} |\mu_{\mathcal{M}}(s)| \right\| \leq \sqrt{d}$. We denote by $\mathbb{M}$ the set of linear MDPs, i.e., satisfying (1).

A control policy π maps states to actions. We denote by s_t^π the state at time t under the policy π. For a given discount factor $\gamma \in (0,1)$, 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 $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$V_{\mathcal{M}}^\pi(s) = \mathbb{E}_{\mathcal{M}}\left[\sum_{t=0}^{+\infty} \gamma^t r_{\mathcal{M}}(s_t^\pi, \pi(s_t^\pi)) | s_0^\pi = s\right] \quad \text{and} \quad Q_{\mathcal{M}}^\pi(s,a) = r_{\mathcal{M}}(s,a) + \gamma \sum_{s'} p_{\mathcal{M}}(s,a,s') V_{\mathcal{M}}^\pi(s').$$

A policy π is said optimal for the MDP $\mathcal{M}$ if it maximizes the value function for any state, i.e., for any policy π', we have $V_{\mathcal{M}}^\pi \geq V_{\mathcal{M}}^{\pi'}$ point-wise. Throughout the paper, we assume that the optimal policy $\pi_{\mathcal{M}}^\star$ is unique. 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 $V_{\mathcal{M}}^\pi \geq V_{\mathcal{M}}^\star - \varepsilon$ point-wise, and we denote by $\Pi_\varepsilon^\star(\mathcal{M})$ the set of ε-optimal policies of $\mathcal{M}$. Note that all our results apply to optimal policies by choosing $\varepsilon = 0$.

3.2 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 design in a PAC framework. A learning algorithm consists of (i) a sampling rule, (ii) a stopping rule and (iii) a decision rule.

Sampling rule. We distinguish between the *generative* and the *forward* model. In the former, in each round t, the sampling rule may select any (state, action) (s_t, a_t) to explore depending on past observations. In the latter, the learner is forced to follow the trajectory of the system, and only the action may be selected. From the selected pair, the learner observes the next state and receives a sample of the corresponding reward.

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.

Decision rule. Based on the observations gathered before stopping, the learner outputs an estimated optimal policy $\hat{\pi}$.

An algorithm is (δ, ε)-PAC if it outputs a ε-optimal policy with probability at least $1-\delta$. Our goal is to design such algorithm with minimal expected 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 (δ, ε)-PAC algorithms. In particular, we wish these bounds to depend on the sub-optimality gap of the MDP $\mathcal{M}$ defined by $\Delta(\mathcal{M}) = \min_{s\in\mathcal{S}, a\neq\pi^\star_{\mathcal{M}}(s)}(V^\star_{\mathcal{M}}(s) - Q^\star_{\mathcal{M}}(s,a))$.

4 Sample Complexity Lower Bounds

To state our instance-specific lower bounds, we first introduce the following notations. 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 $p_{\mathcal{M}}(s,a,\cdot) \ll p_{\mathcal{M}}(s,a,\cdot)$ and $q_{\mathcal{M}}(s,a,\cdot) \ll q_{\mathcal{M}'}(s,a,\cdot)$. In this case, we define the Kullback-Leibler divergence between $\mathcal{M}$ and $\mathcal{M}'$ by:

$$\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) = \mathrm{KL}(q_{\mathcal{M}}(s,a,\cdot)\|q_{\mathcal{M}'}(s,a,\cdot)) + \mathrm{KL}(p_{\mathcal{M}}(s,a,\cdot)\|p_{\mathcal{M}'}(s,a,\cdot)). \tag{2}$$

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

$$\mathrm{Alt}_\varepsilon(\mathcal{M}) = \{\mathcal{M}' \in \mathbb{M} : \mathcal{M} \ll \mathcal{M}', \Pi^\star_\varepsilon(\mathcal{M}) \cap \Pi^\star_\varepsilon(\mathcal{M}') = \emptyset\}. \tag{3}$$

We refer to $\mathrm{Alt}_\varepsilon(\mathcal{M})$ as the *set of alternative MDPs w.r.t.* $\mathcal{M}$. The next proposition is established using classical change-of-measure arguments, and specifically, considering that the observations are generated under an MDP in the set of alternative MDPs w.r.t. $\mathcal{M}$.

Proposition 1. *Let $(\varepsilon,\delta) \in (0,1)^2$. The sample complexity τ of any (δ,ε)-PAC algorithm satisfies $\mathbb{E}_{\mathcal{M}}[\tau] \geq T^\star(\mathcal{M})\mathrm{kl}(\delta, 1-\delta)$, where $T^\star(\mathcal{M})^{-1} = \sup_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}} T(\mathcal{M},\omega)^{-1}$ and*

$$T(\mathcal{M},\omega)^{-1} = \inf_{\mathcal{M}'\in\mathrm{Alt}_\varepsilon(\mathcal{M})} \sum_{s,a} \omega_{s,a}\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a). \tag{4}$$

The vector $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$ solving the optimization problem leading to $T^\star(\mathcal{M})$ can be interpreted as the optimal proportions of times an optimal algorithm should sample the various (state, action) pairs. As it turns out, as in the case of tabular MDPs [20], analyzing and computing this allocation is difficult. Instead, our strategy will be to derive (tight) instance-specific and tractable upper bounds of the lower bound, and devise algorithms based on these upper bounds and the corresponding allocations. To state the upper bounds, we introduce the following quantities: let $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$,

$$\Lambda(\omega) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega_{s,a}\phi(s,a)\phi(s,a)^\top, \tag{5}$$

$$\sigma(\omega) = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{\Lambda(\omega)^{-1}}. \tag{6}$$

$\Lambda(\omega)$ is referred to as the *feature matrix*. Further introduce $\mathrm{Var}_{s,a}(\mathcal{M})$ the variance of the r.v. $R+\gamma V^\star_{\mathcal{M}}(s')$ where R is the random reward collected with (state, action) pair (s,a) and s' is a state with distribution $p_{\mathcal{M}}(s,a,\cdot)$. This variance is bounded by $(1-\gamma)^{-2}/4$ in the worst case. Finally, define $\mathrm{Var}(\mathcal{M}) = \max_{s,a}\mathrm{Var}_{s,a}(\mathcal{M})$. The next theorem presents two upper bounds on the sample complexity lower bound. The first bound depends on the sub-optimality gap only, whereas the refined second upper bound also exhibits a dependence in the variance.

Theorem 1. *We have for all $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$,*

$$T(\mathcal{M},\omega) \leq \frac{10\sigma(\omega)}{3(1-\gamma)^4(\Delta(\mathcal{M})+\varepsilon)^2} := U(\mathcal{M},\omega), \tag{7}$$

$$T^\star(\mathcal{M}) \leq \frac{10d}{3(1-\gamma)^4(\Delta(\mathcal{M})+\varepsilon)^2} := U^\star(\mathcal{M}). \tag{8}$$

Note that the optimal allocation $\omega^\star$ such that $U^\star(\mathcal{M}) = \inf_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}} U(\mathcal{M},\omega) = U(\mathcal{M},\omega^\star)$ is characterized by $\sigma(\omega^\star) = \inf_{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}}\sigma(\omega)$ and hence depends on the MDP $\mathcal{M}$ through its features only. The allocation $\omega^\star$ can be easily computed even before the learning process starts, and as shown in the next section, an algorithm just tracking it yields an expected sample complexity roughly no greater than $U^\star(\mathcal{M})\mathrm{kl}(\delta,1-\delta)$.

5 The GSS Algorithm

This section presents the GSS (G-Sampling-and-Stop) algorithm. It samples (state, action) pair according to the allocation $\omega^\star$ minimizing $\sigma(\omega)$, and hence can be seen as a classical G-sampling strategy. The design of the stopping rule is driven by the upper bound of the sample complexity lower bound $U^\star(\mathcal{M})\mathrm{kl}(\delta,1-\delta)$. However since $\mathcal{M}$ is unknown, we will replace, in this upper bound, $\mathcal{M}$ by the MDP obtained by plugging the least-squares estimators of $\theta_{\mathcal{M}}$ and $\mu_{\mathcal{M}}$. The components of GSS and their analysis is detailed below.

5.1 Sampling rule

Under the GSS algorithm, we start by computing the allocation $\omega^\star$ minimizing $\sigma(\omega)$ over ω. Then in each round t, the algorithm samples the pair (s_t,a_t) according to $\omega^\star$. We define $\Phi_t = \begin{pmatrix}\phi(s_1,a_1) & \cdots & \phi(s_t,a_t)\end{pmatrix}^\top \in \mathbb{R}^{t\times d}$ the matrix of the first t sampled features and $\omega_t = N_t/t$ the frequency of sampling of each pairs up to time t (N_t is the SA-dimensional vector recording the numbers of times each (state, action) pair has been selected up to round t). Notice that

$$\frac{1}{t}\Phi_t^\top\Phi_t = \frac{1}{t}\sum_{\ell=1}^{t}\phi(s_\ell,a_\ell)\phi(s_\ell,a_\ell)^\top = \sum_{s,a}\frac{N_t(s,a)}{t}\phi(s,a)\phi(s,a)^\top = \Lambda(\omega_t). \tag{9}$$

Following this sampling rule, the above matrix will converge towards $\Lambda(\omega^\star)$ and in particular $\sigma(\omega_t)$ will converge towards $\sigma(\omega^\star)$. Specifically, we have:

Proposition 2. *Let $\delta\in(0,1)$. For any $t \geq 10d\log\left(\frac{2d}{\delta}\right)$,*

$$\mathbb{P}\big[\sigma(\omega_t) \leq 2\sigma(\omega^\star)\big] \geq 1-\delta. \tag{10}$$

5.2 Least-squares estimators

GSS leverages 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). Define $R_t = \begin{pmatrix}r_1 & \cdots & r_t\end{pmatrix}^\top$

and $S_t(s) = (1_{s=s'_1}, \cdots, 1_{s=s'_t})^\top$. The regularized least-squares estimator with parameter λ of $\theta_{\mathcal{M}}$ after t experiences is given by

$$\hat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{\ell=1}^{t} \left(r_\ell - \phi(s_\ell, a_\ell)^\top \theta\right)^2 + \lambda \|\theta\|^2 = \left(\Phi_t^\top \Phi_t + \lambda I_d\right)^{-1} \Phi_t^\top R_t \tag{11}$$

and the regularized least-squares estimator of $\mu_{\mathcal{M}}(s)$ with parameter λ by

$$\hat{\mu}_t(s) = \underset{\mu \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{\ell=1}^{t} \left(1_{\{s=s'_\ell\}} - \phi(s_\ell, a_\ell)^\top \mu\right)^2 + \lambda \|\mu\|^2 = \left(\Phi_t^\top \Phi_t + \lambda I_d\right)^{-1} \Phi_t^\top S_t(s). \tag{12}$$

We will choose $\lambda = 1/d$ and denote $\widehat{\mathcal{M}}_t$ the associated MDP and $\widehat{V}_t$ and $\widehat{Q}_t$ its value functions. The least square estimators can be controlled in the following sense :

Proposition 3. *Let $\delta \in (0,1)$. Regardless of the sampling rule, we have with probability at least $1-\delta$ that for all $t \geq 1$,*

$$\left\| \hat{\theta}_t - \theta_{\mathcal{M}} + \gamma (\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}_t^\star \right\|^2_{t\Lambda(\omega_t)} \leq \frac{2}{(1-\gamma)^2} \left(2 \log \left(\frac{\sqrt{e}\zeta(2) t^2}{\delta} \right) + d \log \left(8 e^4 d t^2 \right) \right). \tag{13}$$

5.3 Stopping and decision rules

Denote $Z(t) = t\, U\left(\widehat{\mathcal{M}}_t, \omega_t\right)^{-1}$ the quantity we seek to control in order to achieve the desired sample complexity. The stopping rule of GSS is defined through the threshold

$$\beta(\delta, t) = \frac{12}{5} \left(2 \log \left(\frac{\sqrt{e}\zeta(2) t^2}{\delta} \right) + d \log \left(8 e^4 d t^2 \right) \right) \tag{14}$$

and the stopping time

$$\tau = \inf \left\{ t \geq 1 : Z(t) > \beta(\delta, t) \right\}. \tag{15}$$

This stopping rule is inspired by classical log-likelihood based stopping rules. Usually such stopping rules are proved to be correct by controlling the Kullback-Leibler divergence between the empirical estimator of the MDP and the MDP itself. Here we will take advantage of the linear assumptions to reduce the correctness of the stopping rule to proposition 3, which eventually leads to dependencies in the size of the feature space d instead of $|\mathcal{S}||\mathcal{A}|$. We use the least-squares estimators of the MDP parameters to compute $\widehat{\mathcal{M}}_t$ and implement the stopping time. When the algorithm stops, it computes $\hat{\pi}$ the optimal policy for the MDP $\widehat{\mathcal{M}}_\tau$. The description of GSS is now complete and summarized in Algorithm 1.

Algorithm 1: The GSS algorithm

Compute $\omega^\star = \underset{\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}}{\operatorname{argmin}} \sigma(\omega)$
while $Z(t) \leq \beta(\delta, t)$ **do**
  Sample (s_t, a_t) according to $\omega^\star$ and observe the corresponding experience
  Update $\hat{\theta}_t$ and $\hat{\mu}_t$ according to (11) and (12)
  $t = t + 1$
end
return $\hat{\pi} = \pi_t^\star$ the optimal policy of $\widehat{\mathcal{M}}_t$

5.4 Performance analysis

The next theorem states that GSS is (ε, δ)-PAC as long as it stops and is a direct consequence of proposition 3.

Theorem 2. *Under the GSS algorithm, we have:* $\mathbb{P}\left[\tau < +\infty, \hat{\pi} \notin \Pi^\star_\varepsilon(\mathcal{M})\right] \leq \delta$.

Proof of Theorem 2. The proof of theorem 1 presents an intermediate bound

$$U\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1} \leq \inf_{\mathcal{M}' : \pi_t^\star \notin \Pi^\star_\varepsilon(\mathcal{M}')} \frac{6(1-\gamma)^2}{5} \left\| \hat{\theta}_t - \theta_{\mathcal{M}'} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}'})^\top \widehat{V}_t^\star \right\|^2_{\Lambda(\omega_t)} \leq T\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1}.$$

Under the event that $\pi_t^\star \notin \Pi^\star_\varepsilon(\mathcal{M})$, we can then write

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

It follows that

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

The fact that this last probability is bounded by δ is exactly the statement of proposition 3. □

Finally, we analyze the expected sample complexity of GSS. In the following theorem, we are actually able to derive sample complexity upper bounds for any level of confidence δ (which contrasts with most existing analyses even in the case of best-arm identification problems in bandits). Asymptotically when δ goes to 0, the sample complexity upper bound scales as $U^\star(\mathcal{M}) \log(1/\delta)$.

Theorem 3. *There exists universal constants* c_1, c_2, c_3 *such that under GSS,*

$$\mathbb{E}[\tau] \leq c_1 \frac{d(1-\gamma)^{-4}}{(\Delta(\mathcal{M}) + \varepsilon)^2} \left(\log\left(\frac{c_2}{\delta}\right) + d \log\left(c_3 \frac{d^2(1-\gamma)^{-4}}{(\Delta(\mathcal{M}) + \varepsilon)^2} \right) \right). \tag{16}$$

In particular this result implies that the algorithm stops almost surely. Together with theorem 2 this proves that the GSS algorithm is (δ, ε)-PAC.

6 Episodic Linear MDPs

In this section, we extend our results to episodic MDPs with a fixed time horizon H. Such an MDP $\mathcal{M}$ is characterized through its dynamics and rewards distributions at each round $h \in [H]$. We denote by $p_{\mathcal{M},h}(s, a, s')$ the probability to move to state s' in step h given that the previous state is s and that action a is selected. The mean reward corresponding to this (state, action) pair at this step is denoted by $r_{\mathcal{M},h}(s, a)$. We assume that $\mathcal{M}$ is linear in the sense that: for all $h \in [H]$,

$$\forall (s, s') \in \mathcal{S}, a \in \mathcal{A}, \quad p_{\mathcal{M},h}(s, a, s') = \phi(s, a)^\top \mu_{\mathcal{M},h}(s') \quad \text{and} \quad r_{\mathcal{M},h}(s, a) = \phi(s, a)^\top \theta_{\mathcal{M},h}, \tag{17}$$

where $\mu_{\mathcal{M},h}$ is a family of d measures over $\mathcal{S}$ seen as a $\mathcal{S} \times d$ dimensional matrix, and $\theta_{\mathcal{M},h} \in \mathbb{R}^d$. We will assume that $\|\theta_{\mathcal{M},h}\| \leq \sqrt{d}$ and $\left\| \sum_{s \in \mathcal{S}} |\mu_{\mathcal{M},h}(s)| \right\| \leq \sqrt{d}$. A control policy π maps, in each step, states to actions. The performance of a policy is captured through its state value functions and its (state, action) value functions respectively defined by: $\forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$,

$$\begin{aligned}
V^\pi_{\mathcal{M},h}(s) &= \mathbb{E}_{\mathcal{M}}\left[\sum_{k=h}^{H} r_{\mathcal{M},k}(s_k^\pi, \pi(s_k^\pi)) | s_h^\pi = s \right], \\
Q^\pi_{\mathcal{M},h}(s, a) &= r_{\mathcal{M},h}(s, a) + \sum_{s'} p_{\mathcal{M},h}(s, a, s') V^\pi_{\mathcal{M},h+1}(s').
\end{aligned}$$

The definitions of optimal policies, ε-optimal policies and alternative MDP are the same as in the discounted setup, but only with regard to the value functions at the first step $V^\pi_{\mathcal{M},1}$ and $Q^\pi_{\mathcal{M},1}$ only (we are not interested in $V_{\mathcal{M},h}$ and $Q^\pi_{\mathcal{M},h}$ for $h > 1$). The gap however is defined considering all steps, that is

$$\Delta(\mathcal{M}) = \min_{h, s, a \neq \pi^\star_{\mathcal{M}}(s)} V^\star_{\mathcal{M},h}(s) - Q^\star_{\mathcal{M},h}(s, a). \tag{18}$$

6.1 Sample complexity lower bounds

Proposition 4. *Let $\delta \in (0,1)$ and $\varepsilon > 0$. The sample complexity τ of any (δ,ε)-PAC algorithm satisfies $\mathbb{E}_{\mathcal{M}}[\tau] \geq T^\star(\mathcal{M})\mathrm{kl}(\delta, 1-\delta)$ where*

$$T^\star(\mathcal{M})^{-1} = \sup_{\omega \in (\Sigma_{\mathcal{S}\times\mathcal{A}})^H} \inf_{\mathcal{M}' \in \mathrm{Alt}(\mathcal{M})} \sum_{h,s,a} \omega_{h,s,a} \mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(h,s,a). \tag{19}$$

Theorem 4. *We have for all $\omega \in (\Sigma_{\mathcal{S}\times\mathcal{A}})^H$,*

$$T(\mathcal{M},\omega) \leq \frac{10H^2 \sum_{h=1}^H \sigma(\omega_h)}{3(\Delta(\mathcal{M})+\varepsilon)^2} = U(\mathcal{M},\omega). \tag{20}$$

In particular maximizing over ω yields

$$T^\star(\mathcal{M}) \leq \frac{10H^3 d}{3(\Delta(\mathcal{M})+\varepsilon)^2} = U^\star(\mathcal{M}). \tag{21}$$

6.2 The GSS-E algorithm

Next we adapt the GSS algorithm to the episodic setting. The new algorithm is referred to as GSS-E ('E' stands for episodic), and its pseudo-code is presented in Algorithm 2. the sampling strategy in GSS-E is the same as in GSS at each step. Specifically, GSS-E selects in each step $h \in [H]$, a (state, action) pair according to the allocation $\omega^\star$. The same pair is selected in each step; and hence, the realized allocation ω_t is the same at each step. For this reason from now on in the episodic setting and for any allocation $\omega \in \Sigma_{\mathcal{S}\times\mathcal{A}}$, $T(\cdot,\omega)$ and $U(\cdot,\omega)$ will denote the corresponding functions with the same allocation ω for every step.

Algorithm 2: The GSS-E algorithm

Compute $\omega^\star = \underset{\omega\in\Sigma_{\mathcal{S}\times\mathcal{A}}}{\mathrm{argmin}}\, \sigma(\omega)$
while $Z(t) \leq H\beta(\delta/H, t)$ **do**
  Choose randomly (s_t, a_t) according to $\omega^\star$ and sample this pair for each step $h \in [H]$
  Update $\hat{\theta}_{t,h}$ and $\hat{\mu}_{t,h}$ According to (11) and (12)
  $t = t+1$
end
return $\hat{\pi} = \pi_t^\star$ the optimal policy of $\widehat{\mathcal{M}}_t$

As in the discounted setting, the parameters of the MDP are inferred using the Least-Squares Estimators. We can analyzed the error made by these estimators as previously and get a result analogous to proposition 3:

Proposition 5. *Let $\delta \in (0,1)$ and $h \in [H]$. Regardless of the sampling rule, we have with probability at least $1-\delta$ that for all $t \geq 1$,*

$$\left\| \hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}^\star_{t,h+1} \right\|^2_{t\Lambda(\omega_t)} \leq 2H^2 \left(2\log\left(\frac{\sqrt{e}\zeta(2)t^2}{\delta} \right) + d\log\left(8e^4 dt^2\right) \right). \tag{22}$$

GSS-E applies the stopping rule defined by:

$$\tau = \inf\left\{ t \geq 1 : Z(t) > H\beta(\delta/H, t) \right\}, \tag{23}$$

The additional H terms in the threshold come from the fact that we have H LSE running simultaneously. The performance of GSS-E is summarized in the following theorem.

Theorem 5. *Let $\delta \in (0,1)$. Under the GSS-E algorithm, we have: $\mathbb{P}\left[\tau < +\infty, \hat{\pi} \notin \Pi^\star_\varepsilon(\mathcal{M})\right] \leq \delta$. Furthermore,*

$$\mathbb{E}[\tau] = \tilde{O}\left(\frac{dH^4}{(\Delta(\mathcal{M})+\varepsilon)^2}\left(\log\left(\frac{1}{\delta}\right) + d\log\left(\frac{d^2H^4}{(\Delta(\mathcal{M})+\varepsilon)^2}\right)\right)\right). \tag{24}$$

Proof of the first statement of Theorem 5. The proof of theorem 4 presents an intermediate bound

$$U\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1} \leq \inf_{\mathcal{M}' : \pi^\star_t \notin \Pi^\star_\varepsilon(\mathcal{M}')} \frac{6}{5H^2} \sum_{h=1}^{H} \left\| \hat{\theta}_{t,h} - \theta_{\mathcal{M}',h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M}',h})^\top \widehat{V}^\star_{t,h+1} \right\|^2_{\Lambda(\omega_t)} \leq T\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1}.$$

Under the event that $\pi^\star_t \notin \Pi^\star_\varepsilon(\mathcal{M})$, we can then write

$$Z(t) = t\, U\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1} \leq \frac{6}{5H^2} \sum_{h=1}^{H} \left\| \hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}^\star_{t,h+1} \right\|^2_{t\Lambda(\omega_t)}.$$

It follows that

$$\begin{aligned}
\mathbb{P}\left[\tau < +\infty, \hat{\pi} \notin \Pi^\star_\varepsilon(\mathcal{M})\right] &= \mathbb{P}\left[\exists t \geq 1 : Z(t) > H\beta(\delta/H, t), \pi^\star_t \notin \Pi^\star_\varepsilon(\mathcal{M})\right] \\
&\leq \mathbb{P}\left[\exists t \geq 1 : \sum_{h=1}^{H} \left\| \hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}^\star_{t,h+1} \right\|^2_{t\Lambda(\omega_t)} > \frac{5H^3}{6}\beta(\delta/H, t)\right] \\
&\leq \sum_{h=1}^{H} \mathbb{P}\left[\exists t \geq 1 : \left\| \hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}^\star_{t,h+1} \right\|^2_{t\Lambda(\omega_t)} > \frac{5H^2}{6}\beta(\delta/H, t)\right].
\end{aligned}$$

The fact that each terms inside the sum is bounded by $\frac{\delta}{H}$ is exactly the statement of proposition 5. ☐

The proof of the upper bound on the sample complexity is deferred to appendix D.2.

References

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

[2] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d5

[3] Alekh Agarwal, Sham Kakade, and Lin F. Yang. "Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal". In: ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 67–83. URL: http://proceedings.mlr.press/v125/agarwa

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

[5] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model". In: *Machine learning* 91.3 (2013), pp. 325–349.

[6] Christoph Dann and Emma Brunskill. *Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning*. 2015. DOI: 10.48550/ARXIV.1510.08906. URL: https://arxiv.org/abs/1510.08906.

[7] Omar Darwiche Domingues et al. "Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited". In: *CoRR* abs/2010.03531 (2020). arXiv: 2010.03531. URL: https://arxiv.org/abs/2010.03531.

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

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

[10] Jiafan He, Dongruo Zhou, and Quanquan Gu. "Minimax Optimal Reinforcement Learning for Discounted MDPs". In: *CoRR* abs/2010.00587 (2020). arXiv: 2010.00587. URL: https://arxiv.org/abs/2010.00587.

[11] Chi Jin et al. "Provably efficient reinforcement learning with linear function approximation". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2137–2143.

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

[13] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. "On the complexity of best-arm identification in multi-armed bandit models". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1–42.

[14] Michael Kearns and Satinder Singh. "Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms". In: *Advances in Neural Information Processing* 11 (Apr. 1999).

[15] Jack Kiefer and Jacob Wolfowitz. "The equivalence of two extremum problems". In: *Canadian Journal of Mathematics* 12 (1960), pp. 363–366.

[16] 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*. PMLR. 2020, pp. 5662–5670.

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

[18] Gen Li et al. "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model". In: *arXiv preprint* arXiv:2005.12900 (2020).

[19] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. *Navigating to the Best Policy in Markov Decision Processes*. 2021. URL: https://arxiv.org/abs/2106.02847.

[20] Aymen Al Marjani and Alexandre Proutiere. *Adaptive Sampling for Best Policy Identification in Markov Decision Processes*. 2020. DOI: 10.48550/ARXIV.2009.13405. URL: https://arxiv.org/abs/2009.13405.

[21] Aaron Sidford et al. "Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 5186–5196. URL: http://papers.nips.cc/paper/7765-near-optimal-

[22] Marta Soare, Alessandro Lazaric, and Rémi Munos. "Best-arm identification in linear bandits". In: *Advances in Neural Information Processing Systems* 27 (2014).

[23] Joel A Tropp et al. "An introduction to matrix concentration inequalities". In: *Foundations and Trends® in Machine Learning* 8.1-2 (2015), pp. 1–230.

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

[25] Andrew Wagenmaker and Kevin Jamieson. "Instance-Dependent Near-Optimal Policy Identification in Linear MDPs via Online Experiment Design". In: *arXiv preprint arXiv:2207.02575* (2022).

[26] Andrew Wagenmaker et al. "Reward-Free RL is No Harder Than Reward-Aware RL in Linear Markov Decision Processes". In: *arXiv preprint arXiv:2201.11206* (2022).

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

[28] Andrea Zanette et al. "Limiting extrapolation in linear approximate value iteration". In: *Advances in Neural Information Processing Systems* 32 (2019).

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

A Sample Complexity Lower bounds

A.1 Proof of Proposition 1

The proof follows a standard change of measure argument to obtain instance specific sample complexity lower bounds (see [13, 4] and references therein).

A.2 Gap bounds and value difference lemmas

Here we present key difference lemmas which are useful to relax the optimization problem that appears in the lower bound.

Lemma 6. *Let $\varepsilon > 0$ and $\mathcal{M}'$ a MDP such that $\pi^\star_{\mathcal{M}} \notin \Pi^\star_\varepsilon(\mathcal{M}')$. Then, we have:*

(i) In the discounted setting, it holds that

$$\Delta(\mathcal{M}) + \varepsilon \leq \|V^\star_{\mathcal{M}} - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}'}\|_\infty + \|Q^\star_{\mathcal{M}} - Q^\star_{\mathcal{M}'}\|_\infty. \tag{25}$$

(ii) In the episodic setting, there exists $h \in [H]$ such that the following holds

$$\Delta(\mathcal{M}) + \varepsilon \leq \|V^\star_{\mathcal{M},h} - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}',h}\|_\infty + \|Q^\star_{\mathcal{M},h} - Q^\star_{\mathcal{M}',h}\|_\infty. \tag{26}$$

Proof of Lemma 6. We present the proofs of *(i)* and *(ii)* separately.

Discounted setting - proof of (i). $\pi^\star_{\mathcal{M}} \notin \Pi^\star_\varepsilon(\mathcal{M}')$ 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

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

Episodic setting - proof of (ii). For each step $h \in [H]$ we denote $s_h = \arg\max_s V^\star_{\mathcal{M}',h}(s) - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}',h}(s)$. Since $\pi^\star_{\mathcal{M}} \notin \Pi^\star_\varepsilon(\mathcal{M}')$, we have $V^\star_{\mathcal{M}',1}(s_1) - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}',1}(s_1) \geq \varepsilon$. Note that if $\pi^\star_{\mathcal{M}',1}(s_1) = \pi^\star_{\mathcal{M},1}(s_1)$ then

$$V^\star_{\mathcal{M}',1}(s_1) - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}',1}(s_1) = p_{\mathcal{M}',1}(s_1, \pi^\star_{\mathcal{M}',1}(s_1))^\top (V^\star_{\mathcal{M}',2} - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}',2}) \leq V^\star_{\mathcal{M}',2}(s_2) - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}',2}(s_2).$$

Iterating this reasoning we can show that there exists a step h such that $\pi^\star_{\mathcal{M}',h}(s_h) \neq \pi^\star_{\mathcal{M},h}(s_h)$ (else we would end up with $\varepsilon \leq 0$) and that $\varepsilon \leq V^\star_{\mathcal{M}',h}(s_h) - V^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}',h}(s_h)$. We can then write

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

□

Lemma 7. *Let π be any deterministic policy. We have:*

(i) In the discounted setting, it holds that

$$\|V^\pi_{\mathcal{M}} - V^\pi_{\mathcal{M}'}\|_\infty \leq \|Q^\pi_{\mathcal{M}} - Q^\pi_{\mathcal{M}'}\|_\infty \leq \frac{1}{1-\gamma} \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\pi_{\mathcal{M}}\right)\right|. \quad (27)$$

(ii) In the episodic setting, it holds for all $h_0 \in [H]$ that

$$\|V^\pi_{\mathcal{M},h_0} - V^\pi_{\mathcal{M}',h_0}\|_\infty \leq \|Q^\pi_{\mathcal{M},h_0} - Q^\pi_{\mathcal{M}',h_0}\|_\infty \leq \sum_{h=h_0}^{H} \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M},h} - \theta_{\mathcal{M}',h} + (\mu_{\mathcal{M},h} - \mu_{\mathcal{M}',h})^\top V^\pi_{\mathcal{M},h}\right)\right|. \quad (28)$$

Proof of Lemma 7. We present the proofs of *(i)* and *(ii)* separately.

Discounted setting - Proof of (i). For any $s \in \mathcal{S}$ we have $V^\pi_{\mathcal{M}}(s) - V^\pi_{\mathcal{M}'}(s) = Q^\pi_{\mathcal{M}}(s,\pi(s)) - Q^\pi_{\mathcal{M}'}(s,\pi(s))$ thus the first inequality. 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

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

which implies the second inequality.

Episodic setting - Proof of (ii). It is immediate that for any $h \in [H]$ $\|V^\pi_{\mathcal{M},h} - V^\pi_{\mathcal{M}',h}\|_\infty \leq \|Q^\pi_{\mathcal{M},h} - Q^\pi_{\mathcal{M}',h}\|_\infty$. Now, as in *(i)* we can write for any h

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

and conclude by iterating this inequality for $h = h_0$ to H. ∎

Lemma 8. *We have*

(i) In the discounted setting, it holds that

$$\|V^\star_{\mathcal{M}} - V^\star_{\mathcal{M}'}\|_\infty \leq \|Q^\star_{\mathcal{M}} - Q^\star_{\mathcal{M}'}\|_\infty \leq \frac{1}{1-\gamma} \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right)\right|. \quad (29)$$

(ii) In the episodic setting, it holds for all $h_0 \in [H]$ that

$$\|V^\star_{\mathcal{M},h_0} - V^\star_{\mathcal{M}',h_0}\|_\infty \leq \|Q^\star_{\mathcal{M},h_0} - Q^\star_{\mathcal{M}',h_0}\|_\infty \leq \sum_{h=h_0}^{H} \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M},h} - \theta_{\mathcal{M}',h} + (\mu_{\mathcal{M},h} - \mu_{\mathcal{M}',h})^\top V^\star_{\mathcal{M},h}\right)\right|. \quad (30)$$

Proof of Lemma 8. We present the proof of *(i)* and *(ii)* separately.

Discounted setting - Proof of (i). Let $s \in \mathcal{S}$. We have by optimality of $\pi^\star_{\mathcal{M}'}$ that

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

$V^\star_{\mathcal{M}'}(s) - V^\star_{\mathcal{M}}(s)$ 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

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

which implies the result.

Episodic setting - proof of (ii). For any $h \in [H]$ we can write with the same reasoning as in the proof of *(i)* that $\|V^\star_{\mathcal{M},h} - V^\star_{\mathcal{M}',h}\|_\infty \leq \|Q^\star_{\mathcal{M},h} - Q^\star_{\mathcal{M}',h}\|_\infty$ and

$$\|Q^\star_{\mathcal{M},h} - Q^\star_{\mathcal{M}',h}\|_\infty \leq \max_{s,a} \left|\phi(s,a)^\top \left(\theta_{\mathcal{M},h} - \theta_{\mathcal{M}',h} + (\mu_{\mathcal{M},h} - \mu_{\mathcal{M}',h})^\top V^\star_{\mathcal{M},h+1}\right)\right| + \|Q^\star_{\mathcal{M},h+1} - Q^\star_{\mathcal{M}',h+1}\|_\infty$$

and conclude by iterating this inequality for $h = h_0$ to H. □

Remark 1. *In lemmas 7 and 8 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 $\widehat{\mathcal{M}}_t$ which transitions probabilities, defined as $\phi(s,a)^\top \hat{\mu}_t^\top$, 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 Theorem 1

The goal of this section is to show the bound $T(\mathcal{M},\omega) \leq U(\mathcal{M},\omega)$ for a given MDP $\mathcal{M}$ and a given allocation ω. In other words, the goal is to show that

$$T(\mathcal{M},\omega)^{-1} = \inf_{\mathcal{M}' \in \mathrm{Alt}_\varepsilon(\mathcal{M})} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega_{s,a} \mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \geq \frac{3(1-\gamma)^4(\Delta(\mathcal{M})+\varepsilon)^2}{10\sigma(\omega)}.$$

We are actually going to show this bound but with 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 Kullback-Leibler divergence can be lower bounded using lemma 10. For a given pair (s,a), we choose $f = r + \gamma V^\star_{\mathcal{M}}(s')$ where r and s' are respectively the random reward and the random next step after playing the pair (s,a). f is almost surely bounded by $(1-\gamma)^{-1}$ and the lemma gives

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

Summing over all state action pairs,

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega_{s,a} \mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \geq \frac{6(1-\gamma)^2}{5} \left\|\theta_{\mathcal{M}} - \theta_{\mathcal{M}'} + \gamma(\mu_{\mathcal{M}} - \mu_{\mathcal{M}'})^\top V^\star_{\mathcal{M}}\right\|^2_{\Lambda(\omega)} \tag{31}$$

Putting together Lemma 6, Lemma 7 and Lemma 8 (and choosing $\pi = \pi^\star_{\mathcal{M}}$ in Lemma 7), 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 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 this together with equation (33) and then taking the infimum over $\mathcal{M}'$, we have

$$T(\mathcal{M}, \omega)^{-1} \geq \frac{3(1-\gamma)^4(\Delta(\mathcal{M}) + \varepsilon)^2}{10\sigma(\omega)}. \tag{32}$$

Now, optimizing over $\omega \in \Sigma_{S\times A}$, we obtain that

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

Now, applying Kiefer-Wolfowitz theorem (see Theorem 6) entails 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 1.

Theorem 6 (Kiefer-Wolfowitz [15])**.** *Let $\Phi \subseteq \mathbb{R}^d$ be a finite set and $\mathrm{span}(\Phi) = d$. Let Σ be the set of probability distributions supported on Φ, then 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 2. *The statement of the Kiefer-Wolfowitz theorem in [15] 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.*

A.4 Proof of Theorem 4

Our goal is to show that

$$T(\mathcal{M}, \omega)^{-1} = \inf_{\mathcal{M}'\in\mathrm{Alt}_\varepsilon(\mathcal{M})}\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\omega_{h,s,a}\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(h,s,a) \geq \frac{3(\Delta(\mathcal{M}) + \varepsilon)^2}{10H^2\sum_{h=1}^H\sigma(\omega_h)}.$$

We are actually going to show this bound but with 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 term $\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(h,s,a)$ can be lower bounded using lemma 10 like we did in the discounted model by choosing the function $f = r + \gamma V^\star_{\mathcal{M},h+1}(s')$ where r and s' are respectively the random reward and the random next step after playing the pair (s,a) at step h. f is almost surely bounded by H and the lemma gives

$$\sum_{h=1}^H\sum_{s,a}\omega_{s,a}\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \geq \frac{6}{5H^2}\sum_{h=1}^H\sum_{s,a}\omega_{s,a}\left(\phi(s,a)^\top\left(\theta_{\mathcal{M},h} - \theta_{\mathcal{M}',h} + (\mu_{\mathcal{M},h} - \mu_{\mathcal{M}',h})^\top V^\star_{\mathcal{M},h+1}\right)\right)^2 \tag{33}$$

$$= \frac{6}{5H^2}\sum_{h=1}^H\left\|\theta_{\mathcal{M},h} - \theta_{\mathcal{M}',h} + (\mu_{\mathcal{M},h} - \mu_{\mathcal{M}',h})^\top V^\star_{\mathcal{M},h+1}\right\|^2_{\Lambda(\omega_h)}. \tag{34}$$

Putting together Lemma 6, Lemma 7 and Lemma 8 (and choosing $\pi = \pi^\star_{\mathcal{M}}$ in Lemma 7), obtain a bound on the quantity $\Delta(\mathcal{M}) + \varepsilon$ as follows

$$\Delta(\mathcal{M}) + \varepsilon \leq 2 \sum_{h=1}^{H} \max_{s,a} \left| \phi(s,a)^\top \left(\theta_{\mathcal{M},h} - \theta_{\mathcal{M}',h} + (\mu_{\mathcal{M},h} - \mu_{\mathcal{M}',h})^\top V^\star_{\mathcal{M},h+1} \right) \right|.$$

Now, we can apply lemma 9 with $n = H$, $\Delta = \frac{1}{2}(\Delta(\mathcal{M}) + \varepsilon)$, $\Lambda_h = \Lambda(\omega_h)$ and ϕ_h the feature maximizing the h-th term in the sum above, and deduce that

$$\begin{aligned} \sum_{h=1}^{H} \left\| \theta_{\mathcal{M},h} - \theta_{\mathcal{M}',h} + (\mu_{\mathcal{M},h} - \mu_{\mathcal{M}',h})^\top V^\star_{\mathcal{M},h+1} \right\|^2_{\Lambda(\omega_h)} &\geq \frac{(\Delta(\mathcal{M}) + \varepsilon)^2}{4 \sum_{h=1}^{H} \|\phi_h\|^2_{\Lambda(\omega_h)^{-1}}} \\ &\geq \frac{(\Delta(\mathcal{M}) + \varepsilon)^2}{4 \sum_{h=1}^{H} \sigma(\omega_h)}. \end{aligned}$$

Putting this together with equation (33) and then taking the infimum over $\mathcal{M}'$, we have

$$T(\mathcal{M}, \omega)^{-1} \geq \frac{3(\Delta(\mathcal{M}) + \varepsilon)^2}{10 H^2 \sum_{h=1}^{H} \sigma(\omega_h)}. \tag{35}$$

A.5 Technical lemmas

Lemma 9. *Let $(\phi_i)_i \in (\mathbb{R}^d)^n$, $\Delta > 0$ and $(\Lambda_i)_i \in (\mathbb{R}^{d\times d})^n$ some definite symmetric matrices. We have the following optimisation problem exact solution*

$$\inf_{\substack{x \in \mathbb{R}^{n\times d} \\ \sum_{i=1}^n |\phi_i^\top x_i| \geq \Delta}} \sum_{i=1}^{n} \|x_i\|^2_{\Lambda_i} = \frac{\Delta^2}{\sum_{i=1}^n \|\phi_i\|^2_{\Lambda_i^{-1}}}. \tag{36}$$

Proof of Lemma 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

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

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

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

Finally we have

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

and the solution of the problem is

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

□

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

$$\mathrm{KL}(\alpha\|\beta) \geq \frac{6}{5\|f\|^2_\infty}(\mathbb{E}_\alpha[f] - \mathbb{E}_\beta[f])^2. \tag{37}$$

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

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

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 regard 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,

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

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

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

□

B Sampling rules

B.1 Proof of Proposition 2

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

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{\Lambda(\tilde{\omega}^\star)^{-1}} \leq (1+\epsilon) \inf_{\omega\in\Sigma} \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|^2_{\Lambda(\omega)^{-1}} = (1+\epsilon)d. \tag{38}$$

Obtaining 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 [16], or using efficient rounding procedures coupled with a doubling trick [22]. In our case, we don't use a doubling trick, and our budget of samples is random since we are using a stopping time. Our approach is to directly sample from an approximate G-optimal design, and we shall see that in fact this is sufficient.

A matrix concentration result. We will prove a slightly stronger result that is valid for all ϵ-approximate G-optimal designs.

Lemma 11. *Let $\tilde{\omega}^\star \in \Sigma_{S\times A}$, be an ϵ-approximate G-optimal design for some $\epsilon > 0$ (i.e., satisfying (38)). Assume that the sequence of state action pairs $(s_t, a_t)_{t\geq 1}$ are sampled according to $\tilde{\omega}^\star$, then, for all $\delta \in (0,1)$, $\rho > 0$, we have*

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

Remark 3. *Note that the statement of Lemma 11, along with the fact that $\tilde{\omega}^\star$ is an ϵ-approximate G-optimal design, ensures that the event*

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

holds with probability at least $1-\delta$, provided $t \geq 2(1+\epsilon)\left(\frac{1}{\rho^2}+\frac{1}{3\rho}\right)d\log\left(\frac{2d}{\delta}\right)$. 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.

Remark 4. *In particular, specializing 11 to the G-optimal design $\omega^\star$ and choosing $\rho = 1/2$ gives*

$$\forall t \geq \frac{28d}{3}\log\left(\frac{2d}{\delta}\right), \qquad \mathbb{P}\left[\sigma(\omega_t) \leq 2\sigma(\omega^\star)\right] \geq 1-\delta. \tag{39}$$

This is exactly the statement of Proposition 2.

Proof of Lemma 11. Let $\delta \in (0,1)$ and $t \geq 1$. First, we have

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

where we denote $\tilde{\Lambda}^\star = \Lambda(\tilde{\omega}^\star)$. Denote $(X_\ell)_{1\leq\ell\leq t}$ the summands appearing in the sum above. Note that X_ℓ is a symmetric random matrix that satisfies for all $\ell \geq 1$, $\|X_\ell\| \leq \frac{(1+\epsilon)d}{t}$ a.s. and $\|\mathbb{E}[X_\ell^2]\| \leq \frac{(1+\epsilon)d}{t^2}$ for the

operator norm. Indeed, we have for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

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

so that a.s.

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

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

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

Now, using Matrix Bernstein's inequality (more precisely, we use Theorem 5.4.1. in [24], see also [23]), we obtain that for all $\rho > 0$,

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

which implies that

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

Finally, in order to conclude, observe that

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

thus, provided $t \geq 2(1+\epsilon)\left(\frac{1}{\rho^2}+\frac{1}{3\rho}\right)d\log\left(\frac{2d}{\delta}\right)$, it follows that

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

. □

C Least Square Estimation and Stopping Rules

In this section we show the correctness of our stopping rules through the concentration inequalities of the least square estimators, i.e. propositions 3 and 5. The least square error term depends on the optimal value function of $\widehat{\mathcal{M}}_t$. To get rid of this dependency we decide to control uniformly the error for any optimal value function, which is possible thanks to a net argument. We introduce a first general lemma that establish a self-normalized martingale uniformly on a set of parameters via a net argument.

Lemma 12. *Consider a sequence of feature vectors $(\phi_t)_{t\geq 1}$ in $\mathbb{R}^d$, a set of parameters $\mathcal{V} \subset \mathbb{R}^{\mathcal{S}}$ and for each $V \in \mathcal{V}$ consider a martingale $(x_t(V))_{t\geq 1}$. Finally consider a sequence of positive scalars $(\lambda_t)_{t\geq 1}$. Denote $\Phi_t = \begin{pmatrix}\phi_1 & \dots & \phi_t\end{pmatrix}^\top \in \mathbb{R}^{t\times d}$ and $X_t(V) = \begin{pmatrix}x_1(V) & \dots & x_t(V)\end{pmatrix}^\top \in \mathbb{R}^t$. Under the following assumptions :*

(i) there exists a constant $L > 0$ such that for any $t \geq 1$ and $V \in \mathcal{V}$, $\|x_t(V)\|_\infty \leq L$,

(ii) for any $V, V' \in \mathcal{V}$ and $t \geq 1$, $\|x_t(V) - x_t(V')\| \leq \|V - V'\|_\infty$,

(iii) for any $\epsilon > 0$, $\mathcal{V}$ admits a ϵ-net $\mathcal{V}_\epsilon$ of finite cardinality $\mathcal{N}_\epsilon$, i.e. for any $V \in \mathcal{V}$ there exists $V' \in \mathcal{V}_\epsilon$ such that $\|V - V'\|_\infty \leq \epsilon$,

we have for any $\delta > 0$, $\epsilon \in (0, L)$ and $t \geq 1$ that

$$\mathbb{P}\left[\max_{V\in\mathcal{V}} \|\Phi_t^\top X_t(V)\|^2_{(\Phi_t^\top\Phi_t+\lambda_t I_d)^{-1}} \leq 2L^2 \log\left(\frac{\mathcal{N}_\epsilon}{\delta}\right) + L^2 \log\det\left(\left(\Phi_t^\top\Phi_t + \lambda_t I_d\right)(\lambda_t I_d)^{-1}\right) + td\epsilon^2\right] \geq 1-\delta. \tag{40}$$

Remark 5. *With the same assumptions as above, if we add that for any $\epsilon \in (0, L)$, $\mathcal{N}_\epsilon \leq \left(1 + \frac{2L\sqrt{d}}{\epsilon}\right)^d$, then by choosing $\epsilon = \frac{2L}{\sqrt{t}}$ the threshold becomes*

$$L^2\left(2\log\left(\frac{1}{\delta}\right) + 2d\log(e^2(1+\sqrt{dt})) + \log\det\left(\left(\Phi_t^\top\Phi_t + \lambda_t I_d\right)(\lambda_t I_d)^{-1}\right)\right). \tag{41}$$

Remark 6. *Notice that by linearity of the trace, $\operatorname{tr}\left(\Phi_t^\top\Phi_t + \lambda_t I_d\right) = \sum_{\ell=1}^t \operatorname{tr}\left(\phi_\ell\phi_\ell^\top\right) + d\lambda_t = \sum_{\ell=1}^t \|\phi_\ell\|^2 + d\lambda_t \leq t + d\lambda_t$. Now, the trace of a matrix is the sum of its eigenvalues and the determinant their product. The maximum possible product positive scalars, given their sum, is attained when they are all equal. Hence*

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

Therefore, choosing the regularization $\lambda_t = \frac{1}{d}$ and upper bounding $1 + \sqrt{dt} \leq 2\sqrt{dt}$ and $1 + t \leq 2t$, the threshold in lemma 12 can be replaced with an upper bound

$$L^2\left(2\log\left(\frac{1}{\delta}\right) + d\log\left(8e^4dt^2\right)\right). \tag{42}$$

Proof of Lemma 12. 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 and 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

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

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

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

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

(i) $\forall V \in \mathcal{V}_\epsilon, \mathbb{P}[\mathcal{C}_3(V)] \geq 1 - \frac{\delta}{\mathcal{N}_\epsilon}$. This result is a concentration inequality on self-normalized processes. It can be found as lemma 9 in [2] 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) $\mathbb{P}[\mathcal{C}_2] \geq 1-\delta$. We can immediately see that $\mathcal{C}_2 = \bigcap_{V\in\mathcal{V}_\epsilon} \mathcal{C}_3(V)$. Then an union bound gives

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

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

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

and we can finally write

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

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

In order to use this result we will need to have access to such nets. First we introduce a parametrization of the value function in the linear model.

Lemma 13. *Under the linear assumptions, any value function can be parametrized by a d-dimensional vector the same way the expected rewards and transition probabilities can. Specifically :*

- *In the discounted setup, for any policy π, there exists a vector $\xi^\pi_{\mathcal{M}} \in \mathbb{R}^d$ such that for any pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, $Q^\pi_{\mathcal{M}}(s,a) = \phi(s,a)^\top \xi^\pi_{\mathcal{M}}$. Moreover, we have $\xi^\pi_{\mathcal{M}} = \theta_{\mathcal{M}} + \gamma \mu_{\mathcal{M}}^\top V^\pi_{\mathcal{M}}$ and $\|\xi^\pi_{\mathcal{M}}\| \leq \frac{\sqrt{d}}{1-\gamma}$.*
- *In the episodic setup, for any policy π and any $h \in [H]$, there exists a vector $\xi^\pi_{\mathcal{M},h} \in \mathbb{R}^d$ such that for any pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, $Q^\pi_{\mathcal{M},h}(s,a) = \phi(s,a)^\top \xi^\pi_{\mathcal{M},h}$. Moreover, we have $\xi^\pi_{\mathcal{M},h} = \theta_{\mathcal{M},h} + \mu_{\mathcal{M},h}^\top V^\pi_{\mathcal{M},h+1}$ and $\|\xi^\pi_{\mathcal{M}}\| \leq H\sqrt{d}$.*

Proof of lemma 13. Discounted Setup Using the Bellman equation together with the linear assumptions we directly have $Q^\pi_\mathcal{M}(s,a) = \phi(s,a)^\top \left(\theta_\mathcal{M} + \gamma \mu_\mathcal{M}^\top V^\pi_\mathcal{M}\right)$. Then

$$\left\|\theta_\mathcal{M} + \gamma \mu_\mathcal{M}^\top V^\pi_\mathcal{M}\right\| \leq \|\theta_\mathcal{M}\| + \gamma \left\| \sum_{s\in\mathcal{S}} |\mu_\mathcal{M}(s)| \right\| \|V^\pi_\mathcal{M}\|_\infty \leq \sqrt{d} + \gamma \frac{\sqrt{d}}{1-\gamma} = \frac{\sqrt{d}}{1-\gamma}.$$

Episodic Setup Using the Bellman equation together with the linear assumptions we directly have $Q^\pi_{\mathcal{M},h}(s,a) = \phi(s,a)^\top \left(\theta_{\mathcal{M},h} + \mu_{\mathcal{M},h}^\top V^\pi_{\mathcal{M},h+1}\right)$. Then

$$\left\|\theta_{\mathcal{M},h} + \mu_{\mathcal{M},h}^\top V^\pi_{\mathcal{M},h+1}\right\| \leq \|\theta_{\mathcal{M},h}\| + \left\| \sum_{s\in\mathcal{S}} |\mu_{\mathcal{M},h}(s)| \right\| \|V^\pi_{\mathcal{M},h+1}\|_\infty \leq \sqrt{d} + \sqrt{d}(H-h) \leq H\sqrt{d}.$$

□

The existence of the nets is then ensured by the following lemma.

Lemma 14. *Let $\epsilon > 0$. The set of optimal value functions admits a finite ϵ-net. Specifically :*

(i) In the discounted setup, the set $\mathcal{V}$ of all optimal value functions admits an ϵ-net $\mathcal{V}_\epsilon$ of cardinality $\mathcal{N}_\epsilon \leq \left(1 + \frac{2\sqrt{d}}{(1-\gamma)\epsilon}\right)^d$.

(ii) In the episodic setup, the set $\mathcal{V}(h)$ of all optimal value functions at step $h \in [H]$ admits an ϵ-net $\mathcal{V}_\epsilon$ of cardinality $\mathcal{N}_\epsilon \leq \left(1 + \frac{2H\sqrt{d}}{\epsilon}\right)^d$.

Proof of lemma 14. Discounted setup. Using the parametrization given by lemma 13 and adding the fact that optimal value functions (V, Q) verify $V(\cdot) = \max_{a\in\mathcal{A}} Q(\cdot, a)$, we can write

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

If $V, V' \in \mathcal{V}$ then, considering the corresponding ξ, ξ' as above,

$$\|V - V'\|_\infty \leq \max_{s,a} \|\phi(s,a)^\top (\xi - \xi')\| \leq \|\xi - \xi'\|.$$

Therefore, using this parametrization by ξ, a ϵ-net of $\mathcal{V}$ can be obtained through a ϵ-net of the euclidean ball of radius $\frac{\sqrt{d}}{1-\gamma}$ in $\mathbb{R}^d$. Such net exists with cardinality $\left(1 + \frac{2\sqrt{d}}{(1-\gamma)\epsilon}\right)^d$.

Episodic setup. Let $h \in [H]$. With the same reasoning we have

$$\mathcal{V}(h) \subset \left\{ V \in \mathbb{R}^\mathcal{S} : \exists \xi \in \mathbb{R}^d, V(\cdot) = \max_{a\in\mathcal{A}} \phi(\cdot,a)^\top \xi, \|\xi\| \leq H\sqrt{d} \right\}$$

and a net exists with cardinality $\left(1 + \frac{2H\sqrt{d}}{\epsilon}\right)^d$. □

C.1 Proof of Proposition 3

In this section we show that under any sampling rule the $\frac{1}{d}$-regularized least square estimators verify the following concentration inequality : For any $\delta \in (0,1)$ the events

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

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 $\max_{V\in\mathcal{V}} \left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top V\right\|^2_{t\Lambda(\omega_t)}$ instead 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 $\Phi_t = \begin{pmatrix}\phi(s_1,a_1) & \cdots & \phi(s_t,a_t)\end{pmatrix}^\top$, $R_t = \begin{pmatrix}r_1 & \cdots & r_t\end{pmatrix}^\top$ and $S_t(s) = \begin{pmatrix}\delta_{s,s_1'} & \cdots & \delta_{s,s_t'}\end{pmatrix}^\top$. Recall that $t\Lambda(\omega_t) = \Phi_t^\top\Phi_t$. For any $V\in\mathcal{V}$

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

where we denote $\xi(V) = \left(\theta_{\mathcal{M}} + \gamma\mu_{\mathcal{M}}^\top V\right)$ and define $x_t(V) = r_t - \phi_t^\top\theta_{\mathcal{M}} + \gamma(V(s_t') - \phi_t^\top\mu_{\mathcal{M}}^\top V) = r_t - \mathbb{E}[r_t|\mathcal{F}_{t-1}] + \gamma(V(s_t') - \mathbb{E}[V(s_t')|\mathcal{F}_{t-1}])$ and $X_t(V) = R_t - \Phi_t\theta_{\mathcal{M}} + \gamma(S_t^\top - \Phi_t\mu_{\mathcal{M}}^\top)V = \begin{pmatrix}x_1(V) & \cdots & x_t(V)\end{pmatrix}^\top$. It follows that

$$\begin{aligned}
\left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top V\right\|^2_{\Phi_t^\top\Phi_t} &\le \left\|\left(\Phi_t^\top\Phi_t + \frac{1}{d}I_d\right)^{-1}\Phi_t^\top X_t(V) - \frac{1}{d}\left(\Phi_t^\top\Phi_t + \frac{1}{d}I_d\right)^{-1}\xi(V)\right\|^2_{\Phi_t^\top\Phi_t + \frac{1}{d}I_d} \\
&= \left\|\Phi_t^\top X_t(V) - \frac{1}{d}\xi(V)\right\|^2_{(\Phi_t^\top\Phi_t + \frac{1}{d}I_d)^{-1}} \\
&\le 2\left\|\Phi_t^\top X_t(V)\right\|^2_{(\Phi_t^\top\Phi_t + \frac{1}{d}I_d)^{-1}} + \frac{2}{d^2}\left\|\xi(V)\right\|^2_{(\Phi_t^\top\Phi_t + \frac{1}{d}I_d)^{-1}}
\end{aligned}$$

Lemma 13 states that $\|\xi(V)\| \le \frac{\sqrt{d}}{1-\gamma}$. 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

$$\max_{V\in\mathcal{V}}\left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top V\right\|^2_{\Phi_t^\top\Phi_t} \le 2\max_{V\in\mathcal{V}}\left\|\Phi_t^\top X_t(V)\right\|^2_{(\Phi_t^\top\Phi_t + \frac{1}{d}I_d)^{-1}} + \frac{2}{(1-\gamma)^2}.$$

It is immediate to see that the first two conditions in lemma 12 are satisfied by taking $L = (1-\gamma)^{-1}$ and the third one is given by lemma 14. Therefore we can apply the lemma with $\lambda_t = \frac{1}{d}$ and obtain for all $t \ge 1$

$$\mathbb{P}\left[\max_{V\in\mathcal{V}}\left\|\Phi_t^\top X_t(V)\right\|^2_{(\Phi_t^\top\Phi_t + \frac{1}{d}I_d)^{-1}} \le \frac{1}{(1-\gamma)^2}\left(2\log\left(\frac{1}{\delta_t}\right) + d\log\left(8e^4dt^2\right)\right)\right] \ge 1 - \delta_t.$$

Using the bound above, this event directly implies $\mathcal{C}(t)$ and we can finally conclude that $\mathbb{P}[\mathcal{C}(t)] \ge 1 - \delta_t$ for all $t \ge 1$.

C.2 Proof of Proposition 5

In this section we show that under any sampling rule the $\frac{1}{d}$-regularized least square estimators verify the following concentration inequality : For any $\delta \in (0,1)$ and any $h \in [H]$ the events

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

for all $t \geq 1$ hold simultaneously with probability at least $1-\delta$. As in the discounted setup it is enough to show that for any $t \geq 1$, $\mathbb{P}[\mathcal{C}(t)] \geq 1 - \frac{\delta}{\zeta(2)t^2}$ and we will follow the same reasoning.

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 $\hat{\mu}_t$:

$$\hat{\theta}_{t,h} = \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \Phi_t^\top R_{t,h}, \qquad \hat{\mu}_{t,h}(s) = \left(\Phi_t^\top \Phi_t + \frac{1}{d} I_d\right)^{-1} \Phi_t^\top S_{t,h}(s),$$

where $\Phi_t = \begin{pmatrix} \phi(s_1,a_1) & \cdots & \phi(s_t,a_t) \end{pmatrix}^\top$, $R_{t,h} = \begin{pmatrix} r_{1,h} & \cdots & r_{t,h} \end{pmatrix}^\top$ and $S_{t,h}(s) = \begin{pmatrix} \delta_{s,s'_{1,h}} & \cdots & \delta_{s,s'_{t,h}} \end{pmatrix}^\top$. Recall that $t\Lambda(\omega_t) = \Phi_t^\top \Phi_t$. For any $V \in \mathcal{V}(h+1)$

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

where we denote $\xi(V) = \left(\theta_{\mathcal{M}} + \mu_{\mathcal{M}}^\top V\right)$ and define $x_t(V) = r_t - \phi_t^\top \theta_{\mathcal{M}} + (V(s'_t) - \phi_t^\top \mu_{\mathcal{M}}^\top V) = r_t - \mathbb{E}[r_t|\mathcal{F}_{t-1}] + (V(s'_t) - \mathbb{E}[V(s'_t)|\mathcal{F}_{t-1}])$ and $X_t(V) = R_t - \Phi_t \theta_{\mathcal{M}} + (S_t^\top - \Phi_t \mu_{\mathcal{M}}^\top) V = \begin{pmatrix} x_1(V) & \cdots & x_t(V) \end{pmatrix}^\top$. It follows that

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

Lemma 13 states that $\|\xi(V)\| \leq H\sqrt{d}$. 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

$$\max_{V \in \mathcal{V}} \left\|\hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top V\right\|^2_{\Phi_t^\top \Phi_t} \leq 2 \max_{V \in \mathcal{V}} \left\|\Phi_t^\top X_t(V)\right\|^2_{(\Phi_t^\top \Phi_t + \frac{1}{d} I_d)^{-1}} + 2H^2.$$

It is immediate to see that the first two conditions in lemma 12 are satisfied by taking $L = H$ and the third one is given by lemma 14. Therefore we can apply the lemma with $\lambda_t = \frac{1}{d}$ and obtain for all $t \geq 1$

$$\mathbb{P}\left[\max_{V \in \mathcal{V}} \left\|\Phi_t^\top X_t(V)\right\|^2_{(\Phi_t^\top \Phi_t + \frac{1}{d} I_d)^{-1}} \leq H^2 \left(2 \log\left(\frac{1}{\delta_t}\right) + d \log\left(8e^4 d t^2\right)\right)\right] \geq 1 - \delta_t.$$

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

C.3 Concentration bounds for self-normalized processes

Proposition 15 is borrowed from [1].

Proposition 15. *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 $\mathbb{R}$. 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 [1]. Then, for all $\delta \in (0,1)$,*

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

Proposition is a direct consequence of Proposition 15 and may be obtained via a net argument.

Proposition 16. *Under the same assumptions as in Proposition 15, with the only exception that $(\eta_t)_{t\geq 1}$ is now taking values in $\mathbb{R}^p$, and for each $t \geq 1$, the random vector η_{t+1}, conditionally on $\mathcal{F}_t$, is a zero-mean, and σ^2-subgaussian [2]. Then,*

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

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

[2] We say that a random vector X taking values in $\mathbb{R}^d$, is σ^2-subgaussian if for all $\theta \in \mathbb{R}^d$, $\mathbb{E}[\exp(\theta^\top X)] \leq \exp(\|\theta\|^2\sigma/2)$

D Sample Complexity Analysis

In this section we establish the sample complexity bounds of theorems 3, **??** and 5. First, we establish a continuity-like bound on the quantity $U(\widehat{\mathcal{M}}_t, \omega_t)^{-1}$ which follow the same reasoning as appendix A. The first step is the continuity of the gap, given by the following lemma.

Lemma 17. *(i) In the discounted setting, it holds that*

$$|\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})| \leq \frac{2}{1-\gamma} \max_{s,a} \left| \phi(s,a) \left(\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}_t^\star \right) \right|. \tag{43}$$

(ii) In the episodic setting, there exists $h \in [H]$ such that the following holds

$$|\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})| \leq 2 \sum_{h=1}^{H} \max_{s,a} \left| \phi(s,a) \left(\hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}_{t,h+1}^\star \right) \right|. \tag{44}$$

Proof of Lemma 17. We present the proofs of *(i)* and *(ii)* separately.

Discounted setting - proof of (i). 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 $\Delta = \min_{s \in \mathcal{S}, a \neq \pi^\star(s)} \Delta_{s,a}$.
Let (s,a) be the pair such that $\Delta(\mathcal{M}) = \Delta_{s,a}(\mathcal{M})$. If $a \neq \pi_t^\star(s)$ then $\Delta(\widehat{\mathcal{M}}_t) \leq \Delta_{s,a}(\widehat{\mathcal{M}}_t)$ and $\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M}) \leq \Delta_{s,a}(\widehat{\mathcal{M}}_t) - \Delta_{s,a}(\mathcal{M})$. 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 $\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M}) \leq \Delta_{s',a'}(\widehat{\mathcal{M}}_t) = \Delta_{s',a'}(\widehat{\mathcal{M}}_t) - \Delta_{s',a'}(\mathcal{M})$. 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

$$\begin{aligned}
|\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})| &\leq |\Delta_{s,a}(\widehat{\mathcal{M}}_t) - \Delta_{s,a}(\mathcal{M})| \\
&= |\widehat{V}_t^\star(s) - \widehat{Q}_t^\star(s,a) - V_{\mathcal{M}}^\star(s) + Q_{\mathcal{M}}^\star(s,a)| \\
&= |\widehat{V}_t^\star(s) - V_{\mathcal{M}}^\star(s) + Q_{\mathcal{M}}^\star(s,a) - \widehat{Q}_t^\star(s,a)| \\
&\leq \|\widehat{V}_t^\star - V_{\mathcal{M}}^\star\|_\infty + \|\widehat{Q}_t^\star - Q_{\mathcal{M}}^\star\|_\infty.
\end{aligned}$$

We conclude with lemma 8.

Episodic setting - proof of (ii). For clarity we denote for both MDPs, for any $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\Delta_{h,s,a} = V_h^\star(s) - Q_h^\star(s,a)$, so that $\Delta = \min_{h \in [H], s \in \mathcal{S}, a \neq \pi^\star(s)} \Delta_{h,s,a}$.
Let (h,s,a) be the parameters such that $\Delta(\mathcal{M}) = \Delta_{h,s,a}(\mathcal{M})$. If $a \neq \pi_{t,h}^\star(s)$ then $\Delta(\widehat{\mathcal{M}}_t) \leq \Delta_{h,s,a}(\widehat{\mathcal{M}}_t)$ and $\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M}) \leq \Delta_{h,s,a}(\widehat{\mathcal{M}}_t) - \Delta_{h,s,a}(\mathcal{M})$. Else, since both MDPs have exactly $H|\mathcal{S}|$ optimal state/action pairs (one for each (step, state) pair), the fact that (h,s,a) is optimal for $\widehat{\mathcal{M}}_t$ but not for $\mathcal{M}$ means that there exist (h',s',a') optimal for $\mathcal{M}$ but not for $\widehat{\mathcal{M}}_t$, and we have $\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M}) \leq \Delta_{h',s',a'}(\widehat{\mathcal{M}}_t) = \Delta_{h',s',a'}(\widehat{\mathcal{M}}_t) - \Delta_{s',a'}(\mathcal{M})$. Either way, and doing the same reasoning to bound $\Delta(\mathcal{M}) - \Delta(\widehat{\mathcal{M}}_t)$, we can find (h,s,a) such that

$$\begin{aligned}
|\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})| &\leq |\Delta_{h,s,a}(\widehat{\mathcal{M}}_t) - \Delta_{h,s,a}(\mathcal{M})| \\
&= |\widehat{V}_{t,h}^\star(s) - \widehat{Q}_{t,h}^\star(s,a) - V_{\mathcal{M},h}^\star(s) + Q_{\mathcal{M},h}^\star(s,a)| \\
&= |\widehat{V}_{t,h}^\star(s) - V_{\mathcal{M},h}^\star(s) + Q_{\mathcal{M},h}^\star(s,a) - \widehat{Q}_{t,h}^\star(s,a)| \\
&\leq \|\widehat{V}_t^\star - V_{\mathcal{M},h}^\star\|_\infty + \|\widehat{Q}_{t,h}^\star - Q_{\mathcal{M},h}^\star\|_\infty.
\end{aligned}$$

We conclude with lemma 8. ∎

Lemma 18. *For any $t \geq 1$ it holds that*

$$\left|U^\star(\mathcal{M})^{-1} - U\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1}\right| \leq B(t), \tag{45}$$

where $B(t)$ is defined:

(i) in the discounted setting as

$$B(t) = 6(1-\gamma)^2 \left\|\hat{\theta}_t - \theta_{\mathcal{M}} + \gamma(\hat{\mu}_t - \mu_{\mathcal{M}})^\top \widehat{V}_t^\star\right\|^2_{\Lambda(\omega_t)} + \left(\frac{5}{4} - \frac{\sigma(\omega^\star)}{\sigma(\omega_t)}\right) U^\star(\mathcal{M})^{-1}. \tag{46}$$

(ii) in the episodic setting as

$$B(t) = \frac{6}{H^2}\sum_{h=1}^{H} \left\|\hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}_{t,h+1}^\star\right\|^2_{\Lambda(\omega_t)} + \left(\frac{5}{4} - \frac{\sigma(\omega^\star)}{\sigma(\omega_t)}\right) U^\star(\mathcal{M})^{-1}. \tag{47}$$

Proof. For both settings we can write

$$\left|U^\star(\mathcal{M})^{-1} - U\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1}\right| \leq \left|U^\star(\mathcal{M})^{-1} - U(\mathcal{M}, \omega_t)^{-1}\right| + \left|U(\mathcal{M}, \omega_t)^{-1} - U\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1}\right|$$

and the first term can be rewritten

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

For the second term, regardless of the setting there exists a quantity $u(\omega_t)^{-1}$ such that $U(\cdot, \omega_t)^{-1} = u(\omega_t)^{-1}(\Delta(\cdot) + \varepsilon)^2$. In the discounted setting we have $u(\omega_t)^{-1} = \frac{3(1-\gamma)^4}{10\sigma(\omega_t)}$ and in the episodic setting we have $u(\omega_t)^{-1} = \frac{3}{10H^3\sigma(\omega_t)}$. Then

$$\begin{aligned}
&\left|U(\mathcal{M}, \omega_t)^{-1} - U\big(\widehat{\mathcal{M}}_t, \omega_t\big)^{-1}\right| \\
&= u(\omega_t)^{-1}\left|(\Delta(\mathcal{M}) + \varepsilon)^2 - (\Delta(\widehat{\mathcal{M}}_t) + \varepsilon)^2\right| \\
&= u(\omega_t)^{-1}\left|\Big(\Delta(\widehat{\mathcal{M}}_t) + \Delta(\mathcal{M}) + 2\varepsilon\Big)\Big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\Big)\right| \\
&= u(\omega_t)^{-1}\left|\big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\big)^2 + 2(\Delta(\mathcal{M}) + \varepsilon)\big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\big)\right| \\
&\leq u(\omega_t)^{-1}\big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\big)^2 + 2\sqrt{\frac{1}{4}u(\omega_t)^{-1}(\Delta(\mathcal{M}) + \varepsilon)^2}\sqrt{4u(\omega_t)^{-1}\big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\big)^2} \\
&\leq u(\omega_t)^{-1}\big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\big)^2 + \frac{1}{4}U(\mathcal{M}, \omega_t)^{-1} + 4u(\omega_t)^{-1}\big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\big)^2 \\
&\leq 5u(\omega_t)^{-1}\big(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\big)^2 + \frac{1}{4}U^\star(\mathcal{M})^{-1}
\end{aligned}$$

using $U(\mathcal{M}, \omega_t)^{-1} \leq U^\star(\mathcal{M})^{-1}$ for the last step. Now we separate the end of the proof for both settings.

Discounted setting. To conclude we need to show that

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

Recall that lemma 17 gives

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

Then the result is obtained by applying lemma 9 with $n = 1$, ϕ_1 the feature maximizing the term above and $\Delta = \frac{1-\gamma}{2}\left|\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\right|$.

Episodic setting. To conclude we need to show that

$$\frac{3}{2H^3\sigma(\omega_t)}\left(\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\right)^2 \leq 6\sum_{h=1}^{H}\left\|\hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}^\star_{t,h+1}\right\|^2_{\Lambda(\omega_t)}.$$

Recall that lemma 17 gives

$$\left|\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\right| \leq 2\sum_{h=1}^{H}\max_{s,a}\left|\phi(s,a)^\top\left(\hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}^\star_{t,h+1}\right)\right|.$$

Then the result is obtained by applying lemma 9 with $n = H$, ϕ_h the feature maximizing the h-term in the sum above and $\Delta = \frac{1}{2}\left|\Delta(\widehat{\mathcal{M}}_t) - \Delta(\mathcal{M})\right|$. □

D.1 Proof of Theorem 3

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

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

where $Z(t) = t\,U(\widehat{\mathcal{M}}_t, \omega_t)^{-1}$. In order to establish the sample complexity upper bound, we are going to find a time T such that for any $t \geq T$, $\mathbb{P}[\tau > t] = O\left(\frac{1}{t^2}\right)$, so that we can bound $\mathbb{E}[\tau]$ by T plus a constant. Thanks to lemma 18, we have $\{\tau > t\} \subset \left\{t\,U(\widehat{\mathcal{M}}_t, \omega_t)^{-1} \leq \beta(\delta, t)\right\} \subset \left\{t\,U^\star(\mathcal{M})^{-1} \leq \beta(\delta, t) + tB(t)\right\}$. Recall that when proving proposition 3 we have shown that for any $\delta' > 0$ and for any $t \geq 1$ we have with probability at least $1 - \frac{\delta'}{\zeta(2)t^2}$ that

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

Moreover, lemma 11 states that if $t \geq \frac{28d}{3}\log\left(\frac{2\zeta(2)dt^2}{\delta'}\right)$ then with probability at least $1 - \frac{\delta'}{\zeta(2)t^2}$ we have $c(\omega_t) \leq 2c(\omega^\star)$. Choosing $\delta' = 1$ and plugging both bounds in the definition of $B(t)$ we have with an union bound that for all $t \geq T_1$

$$\mathbb{P}\left[tB(t) \leq 5\beta(1, t) + \frac{3t}{4}U^\star(\mathcal{M})^{-1}\right] \geq 1 - \frac{2}{\zeta(2)t^2},$$

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 19 $t \geq T_1$ 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 19, where we define

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

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, denoting $T = \max(T_1, T_2)$,

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

D.2 Proof of Theorem 5

Recall the threshold

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

and the stopping time

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

where $Z(t) = t\,U(\widehat{\mathcal{M}}_t, \omega_t)^{-1}$. Thanks to lemma 18, we have $\{\tau > t\} \subset \left\{t\,U(\widehat{\mathcal{M}}_t, \omega_t)^{-1} \leq H\beta(\delta/H, t)\right\} \subset \left\{t\,U^\star(\mathcal{M})^{-1} \leq H\beta(\delta/H, t) + tB(t)\right\}$. Recall that from the proof of proposition 5 we can deduce that for any $\delta' > 0$ and for any $t \geq 1$ we have with probability at least $1 - \frac{\delta'}{\zeta(2)t^2}$ that

$$\sum_{h=1}^{H}\left\|\hat{\theta}_{t,h} - \theta_{\mathcal{M},h} + (\hat{\mu}_{t,h} - \mu_{\mathcal{M},h})^\top \widehat{V}^\star_{t,h+1}\right\|^2_{t\Lambda(\omega_t)} \leq \frac{5H^3}{6}\beta(\delta'/H, t).$$

Moreover, lemma 11 states that if $t \geq \frac{28d}{3}\log\left(\frac{2\zeta(2)dt^2}{\delta'}\right)$ then with probability at least $1 - \frac{\delta'}{\zeta(2)t^2}$ we have $c(\omega_t) \leq 2c(\omega^\star)$. Choosing $\delta' = 1$ and plugging both bounds in the definition of $B(t)$ we have with an union bound that for all $t \geq T_1$

$$\mathbb{P}\left[tB(t) \leq 5H\beta(1/H, t) + \frac{3t}{4}U^\star(\mathcal{M})^{-1}\right] \geq 1 - \frac{2}{\zeta(2)t^2},$$

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 19 $t \geq T_1$ implies $t \geq \frac{28d}{3}\log(2\zeta(2)dt^2)$. Now to conclude we only need to show that this event implies $\left\{t\,U^\star(\mathcal{M})^{-1} > H\beta(\delta/H, t) + tB(t)\right\}$ when t is large enough. Assume that $tB(t) \leq 5H\beta(1/H, t) + \frac{3t}{4}U^\star(\mathcal{M})^{-1}$. Since $\delta < 1$ we have $\beta(1/H, t) < \beta(\delta/H, t)$ and $H\beta(\delta/H, t) + tB(t) \leq 6H\beta(\delta/H, t) + \frac{3t}{4}U^\star(\mathcal{M})^{-1}$. To show that this is bounded by $t\,U^\star(\mathcal{M})^{-1}$ is equivalent to show that $24H\beta(\delta/H, t) \leq t\,U^\star(\mathcal{M})^{-1}$. Again, we can show that this last bound is true when $t \geq T_2$ thanks to lemma 19, where we define

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

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

$$\mathbb{P}[\tau > t] \leq \mathbb{P}\left[t\,U^\star(\mathcal{M})^{-1} \leq H\beta(\delta/H, t) + tB(t)\right] \leq \mathbb{P}\left[tB(t) > 5H\beta(1/H, t) + \frac{3t}{4}U^\star(\mathcal{M})^{-1}\right] \leq \frac{2}{\zeta(2)t^2}.$$

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

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

D.3 Technical lemmas

Lemma 19. *Let $a, b > 0$. A sufficient condition for $t > a\log(t) + b$ to hold is that $t \geq 2a\log(2a) + 2b$.*

Proof. Let $t \geq 2a\log(2a) + 2b$. Then

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

□