Exploration-Exploitation with Thompson Sampling in Linear Systems

Marc Abeille

Université des Sciences et des Technologies de Lille
École Doctorale Sciences Pour l'Ingénieur

Thèse de Doctorat

préparée au sein du CRIStAL, UMR 9189 Lille 1/CNRS
et du centre de recherche Inria Lille - Nord Europe

Spécialité : **Mathématiques Appliquées**

présentée par
Marc ABEILLE

Exploration-Exploitation with Thompson Sampling in Linear Systems

---

sous la direction de **Rémi MUNOS**, **Alessandro LAZARIC** et **Emmanuel SERIE**

Soutenue publiquement à **Villeneuve d'Ascq**, le **13 décembre 2017** devant le jury composé de :

Mme. Shipra **AGRAWAL**	Université de Columbia	Rapporteur
M. Csaba **SZEPESVÁRI**	Université d'Alberta	Rapporteur
M. Olivier **GUÉANT**	Université Paris 1	Président
M. Rémi **MUNOS**	Inria Lille	Directeur
M. Alessandro **LAZARIC**	Inria Lille	Encadrant
M. Emmanuel **SÉRIÉ**	Capital Fund Management	Encadrant

Short english abstract: This dissertation is dedicated to the study of the *Thompson Sampling* (TS) algorithms designed to address the exploration-exploitation dilemma that is inherent in sequential decision-making under uncertainty. As opposed to algorithms based on the *optimism-in-the-face-of-uncertainty* (OFU) principle, where the exploration is performed by selecting the most favorable model within the set of plausible one, TS algorithms rely on *randomization* to enhance the exploration, and thus are much more computationally efficient. We focus on linearly parametrized problems that allow for continuous state-action spaces, namely the Linear Bandit (LB) problems and the Linear Quadratic (LQ) control problems. We derive two novel analyses for the regret of TS algorithms in those settings. While the obtained regret bound for LB is similar to previous results, the proof sheds new light on the functioning of TS, and allows us to extend the analysis to LQ problems. As a result, we prove the first regret bound for TS in LQ, and show that the frequentist regret is of order $O(\sqrt{T})$ which matches the existing guarantee for the regret of OFU algorithms in LQ. Finally, we propose an application of exploration-exploitation techniques to the practical problem of portfolio construction, and discuss the need for active exploration in this setting.

Titre en français : Thompson Sampling pour l'exploration-exploitation dans les systèmes linéaires.

Résumé court en français : Cette thèse est dédiée à l'étude du *Thompson Sampling* (TS), une heuristique qui vise à surmonter le dilemme entre exploration et exploitation qui est inhérent à tout processus décisionnel face à l'incertain. Contrairement aux algorithmes issus de l'heuristique *optimiste face à l'incertain* (OFU), où l'exploration provient du choix du modèle le plus favorable possible au vu de la connaissance accumulée, les algorithmes TS introduisent de l'aléa dans le processus décisionnel en sélectionnant aléatoirement un modèle plausible, ce qui les rend bien moins coûteux numériquement. Cette étude se concentre sur les problèmes paramétriques linéaires, qui autorisent les espaces état-action continus (infinis), en particulier les problèmes de Bandits Linéaires (LB) et les problèmes de contrôle Linéaire et Quadratique (LQ). Nous proposons dans cette thèse de nouvelles analyses du regret des algorithmes TS pour chacun de ces deux problèmes. Bien que notre démonstration pour les LB garantisse une borne supérieure identique aux résultats préexistants, la structure de la preuve offre une nouvelle vision du fonctionnement de l'algorithme TS, et nous permet d'étendre cette analyse aux problèmes LQ. Nous démontrons la première borne supérieure pour le regret de l'algorithme TS dans les problèmes LQ, qui garantie dans le cadre fréquentiste un regret au plus d'ordre $O(\sqrt{T})$. Enfin, nous proposons une application des méthodes d'exploration-exploitation pour les problèmes d'optimisation de portefeuille, et discutons dans ce cadre le besoin ou non d'explorer activement.

Key words: machine learning, decision-making, algorithm, thompson sampling, multi-armed bandit, linear bandit, linear quadratic control, reinforcement learning.

Mots clés: apprentissage automatique, processus décisionnel, algorithme, thompson sampling, bandit multi-bras, bandit linéaire, contrôle linéaire quadratique, apprentissage par renforcement.

Acknowledgement & Remerciements

Mes premiers remerciements vont naturellement à Emmanuel Sérié et Alessandro Lazaric qui ont encadré ma thèse. Merci Emmanuel de m'avoir poussé à suivre la voie du doctorat. Ton enthousiasme et ta determination à toujours aller plus loin dans la compréhension des phénomènes et méthodes en jeux ont été une aide précieuse. Alessandro, working with you has been a real pleasure. I cannot thank you enough for all I have learned at your side, your availability and your support. I hope I have taken from you a bit of your vision and your rigor. Enfin, merci à Rémi Munos qui a accepté la supervision de mon doctorat.

Thanks to Shipra Agrawal, Csaba Szepesvári and Olivier Guéant for accepting to be part of my committee, especially to Shipra and Csaba who agreed to review this dissertation. It is a great honor for me.

Je souhaite aussi exprimer ma gratitude aux chercheurs de Capital Fund Management, avec qui j'ai eu la chance d'interagir durant ces années passées à CFM. En particulier, merci à Charles-Albert Lehalle pour les discussions riches et variées, et à Jean-Philippe Bouchaud, sans qui cette thèse n'aurait pu avoir lieu.

Je remercie tout les stagiaires, doctorants et post-docs que j'ai eu la joie de côtoyer à CFM, plus particulièrement Joel et Jonathan avec qui j'ai partagé ces années de stage et de thèse. Nos discussions, scientifiques ou non, ont rendu ces années aussi profitables qu'agréables. Je remercie également les membres de l'équipe Sequel, avec qui j'ai eu la chance d'échanger lors de mes passages à l'Inria.

Enfin, merci à mes proches pour leur indefectible soutient, et surtout à Marion, qui m'a accompagné dans ce projet, et dans tout les autres.

Contents

Chapter 1
Introduction

This dissertation is dedicated to the study of Thompson Sampling algorithms, designed to address the exploration-exploitation dilemma that is inherent in sequential decision-making under uncertainty. Sequential decision-making is a complex and general process that consists in selecting a sequence of actions in order to achieve a task. Consider for instance the parking problem, where a driver (the agent) wants to park: each time the driver sees an available spot, it balances between parking immediately or waiting for another spot closer to its destination, running the risk of finding none and turning back. When the agent does not have the complete knowledge of its environment (e.g., the number of available parking spots), the agent has to deal with the uncertainty about the outcomes of its actions and adapts its decisions to its observations i.e., the empirical knowledge accumulated during the decision process. Therefore, the learning agent faces a dilemma between exploring (go further see whether there is any available spot) and exploiting (there was few available spots so far, it seems unlikely to find one further, park now). How to build intelligent agents that properly address this trade-off is one of the major goals of Artificial Intelligence, and is at the heart of Reinforcement Learning (RL). In RL, each time the agent selects an action, it receives a reward generated in an unknown fashion, and aims at cumulating as much reward as possible. While a large number of methods have been proposed so far to deal with the exploration-exploitation trade-off, two main principles emerged as the most effective and flexible solution to this problem: the *optimism-in-the-face-uncertainty* (OFU) and the *Thompson sampling* (TS). The former consists in selecting action that seems the more rewarding in the most favorable environment that is coherent with observations so far, while the latter introduces randomness in the way actions are chosen to enhance the exploration of the unknown environment with respect to the objective. The OFU principle has been intensively investigated in the last decades and led to the celebrated family of Upper Confidence Bound (UCB) algorithms, for which theoretical guarantees have been provided. TS is an old principle based upon Bayesian ideas, that has recently attracted lot of consideration because of its impressive empirical performances. However, theoretical guarantees for TS algorithms are still limited. Despite the success of RL in numerous domains, one of its main limitations is that the action and/or the state spaces are usually assumed to be finite (and of small cardinality) whereas real-world applications often deal with infinite (or large) state-action spaces. In order to overcome this limitation, a natural approach is to consider parametrized systems with continuous state-action spaces. However, this makes the control and the learning problems significantly harder. To this end, linearly parametrized systems are of major interest as they remain tractable but reflect the key difficulties in designing and studying exploration-exploitation algorithms for continuous state-action spaces

settings. Additionally, they consist in robust, yet flexible models and thus are widely used in practice.

In this thesis, we analyse TS algorithms for the Linear Bandit (LB) problem and the Linear Quadratic (LQ) control problem. LB is a natural extension of the celebrated Multi-Armed Bandit (MAB) framework, where an agent has to sequentially select actions from an infinite set, at each round, and receives a reward that is randomly generated independently from the previous rounds, according to a linear model. Formally, the reward function is a noisy linear mapping between the action and an unknown parameter. One of the main limitations of the LB setting is that it cannot model problems where the environment is affected by the actions chosen by the agent over time. For such systems, one has to consider the environment's dynamics that makes both the learning and the control problems harder. LQ is a specific instance of Markov Decision Processes (MDP) with infinite state-action spaces, that assumes the dynamics to be linear in the state (which characterizes the system) and the action (chosen by the agent), and where the reward function is quadratic. LQ is a standard in control theory and offer the advantage of having explicit solutions for the optimal policy (i.e., the mapping from observations to actions) when the dynamics and the reward model are known.

Contents

1.1 Our approach

The original motivation of this thesis lies in addressing the dynamical portfolio allocation problem, where the dynamics of the market exhibit both return predictability i.e, the fact that traders have "views" on the future prices' movements, and price impact i.e., the fact that the trading orders affect the prices in an adverse way. When the volume of the transactions is large compared to the available liquidity, the execution of such orders drastically changes the supply and demand, and thus the prices. Moreover, empirical studies suggest that this effect is dynamical, i.e., that the market has memory of past trades and digest them slowly. While the objective is not to construct a complete and functional trading robot, we consider this use-case as a tool that helps us highlighting the practical issues that arise when constructing autonomous agent, and address them theoretically.

In the concrete example of portfolio allocation, the challenges are numerous:

1. The dynamical nature of both the return predictability and the price impact effects requires one to use dynamic programming techniques in order to compute the optimal allocation. Moreover, the problem is intrinsically characterized by a large or a continuous state-action space: at short timescales (e.g., high-frequency trading), the prices and the traded volumes are discrete with fine-mesh that induces a very large state-action space; for longer horizons and larger volumes (e.g., asset management), the state-action space becomes so huge that both are usually modeled as continuous variables.

2. By nature, the dynamics of the financial market are unknown and have to be estimated. Moreover, to observe the price impact effects, the learner has to trade in a so-called *bandit feedback* scenario since it does not have access to the other participants' actions, and only observes the outcomes of its own decisions. This implies that one has to consider the online learning setting, where the task (i.e., the trading process) and the estimation are performed simultaneously.

3. Finally, for practical applications, it is of crucial importance that the models and the methods be robust (e.g., that two similar problems have similar solutions and performances) and tractable (i.e., computationally efficient).

We address the first point by considering parametrized systems, which can handle continuous state-action spaces. Nevertheless, this usually makes the control problem hard or impossible to solve. We overcome this issue showing that one can cast the portfolio optimization problem into a Linear Quadratic control problem for which closed-form solutions have been derived when the parameters of the dynamics are known. This reformulation can be performed for any linear Markovian market model and thus encompasses numbers of econometric models used in the industry. Then, we consider online learning procedures for generic LQ systems that would allow us to perform the control and the estimation jointly. To tackle the exploration-exploitation trade-off, algorithms have been derived from the two popular principles that are *optimism-in-the-face-of-uncertainty* and *Thompson sampling*. While the optimistic instance has been recently proved to suffer a low regret (i.e., have good performance), the computational cost of this strategy is prohibitive in this setting. Indeed, it requires at each policy re-evaluation to solve a non-convex, high-dimensional, optimization problem. On the other hand, TS stands as a good candidate to maintain tractability since it just requires a random sampling at each time step. Unfortunately, limited theoretical guarantee has been provided for this strategy which motivates our theoretical analysis for TS in LQ. However, the existing analysis for TS sampling in the simpler Linear Bandit setting (that does not take into account the dynamical effects) cannot be extended to LQ, mostly because of the structure of the proof. In particular, the analysis requires that the performance associated with any action under any parametrization concentrates as soon as the parameter concentrates around the true one. Unfortunately, while in LB the performance is well defined for every actions under any parametrization, it is no longer

the case in LQ where this performance can diverge to $-\infty$ and thus the concentration property does not hold. To this end, we derive an alternative analysis for TS in LB that does not rely on this property and thus that can be extended to LQ. Moreover, it provides some intuitions about the functioning of the TS algorithm, that we leverage together with the existing proof for the optimistic algorithm in LQ to guarantee the first regret bound for TS in LQ.

1.2 Outline and contributions

In this section, we give an outline of the thesis and summarize the contributions.

Chapter 2: Exploration-Exploitation Dilemma in Sequential Decision-Making

The objective of this chapter is to introduce the concept of *exploration-exploitation*, present the state-of-the-art and the material that we need for the analyses of Ch. 3 and 4. We first introduce the exploration-exploitation dilemma in the well-known Multi-Armed Bandit setting, explaining the challenges that the learner faces and presenting the two most popular principles to address this issue, based respectively on *optimism-in-face-of-uncertainty* and *Thompson sampling*. Then, we present two extensions, namely the Linear Bandit problem and the Linear Quadratic control problem, that address the main limitations of MAB, i.e., the finite action space and the independency between rounds. We present algorithms based on the two principles in those frameworks and recall the available regret guarantees.

Chapter 3: Thompson Sampling in Linear Bandit

We derive an alternative proof for the regret of Thompson sampling (TS) in the stochastic linear bandit setting, where the reward is linear in the chosen arm (the selected action) according to an unknown parameter. While we obtain a regret bound of order $\widetilde{O}(d^{3/2}\sqrt{T})$ as in previous results, the proof sheds new light on the functioning of the TS. We leverage the structure of the problem to show how the regret is related to the sensitivity (i.e., the gradient) of the objective function and how selecting optimal arms associated to *optimistic* parameters does control it. Thus, we show that TS can be seen as a generic randomized algorithm where the sampling distribution is designed to have a fixed probability of being optimistic, at the cost of an additional $\sqrt{d}$ regret factor compared to a UCB-like approach. Furthermore, we show that our proof can be readily applied to regularized linear optimization and generalized linear model problems for which we prove the first $\widetilde{O}(\sqrt{T})$ regret bound for the TS algorithm.

Chapter 4: Thompson Sampling in Linear Quadratic System

We consider the exploration-exploitation trade-off in Linear Quadratic control problems, where the state dynamics is linear and the cost function is quadratic in the state

and control. We analyze the regret of TS (a.k.a. posterior-sampling for reinforcement learning) in the frequentist setting, i.e., when the parameters characterizing the LQ dynamics are fixed. Despite the empirical and theoretical success in a wide range of problems from MAB to LB, extending those results to the LQ setting is highly challenging: 1) the standard line of proof that relies on classifying arms into saturated/unsaturated pool cannot be applied here as their associated optimal value could be infinite; 2) the TS functioning requires frequent policy updates, which is in contrast with the usual lazy update scheme used in most RL algorithm. As a consequence, it raises the issue of bounding the gap in the optimal value at the policy switches.
We prove that TS achieves a $\widetilde{O}(\sqrt{T})$ regret, thus matching the performance of the OFU approach and confirming the conjecture of Osband and Van Roy (2016). We address the first point leveraging the ideas introduced in Ch. 3, stressing the link between the actual actions chosen by TS and the gradient of the optimal value function. We exhibit the need to trade-off the frequency of sampling optimistic parameters and the frequency of switches in the control policy, and show that lazy update schemes induce at best an overall regret of $O(T^{2/3})$. Finally, we derive a novel bound on the regret due to policy switches, thus allowing to update parameters and the policy at each step and overcome the limitations due to lazy updates.

Chapter 5: Application to Portfolio Construction

We propose an application of exploration-exploitation strategies for the concrete example of optimal portfolio allocation under price impact. We introduce a novel LQ formulation for the portfolio allocation problem, under the assumption of linear price dynamics, from which we obtain the optimal control and discuss the exploration-exploitation trade-off arising from the presence of unknown parameters. We consider two problem instances with or without risk constraint and show that this affects the need for exploration. In the unconstrained case, a greedy strategy fails to achieve sub-linear regret, while Thompson Sampling or optimism-based algorithms effectively trade-off exploration and exploitation. On the other hand, the risk constraint modifies the structure of the policy, removing somehow the need for active exploration, and a greedy strategy is optimal. We discuss this counter-intuitive result and support it with numerical experiments.

Chapter 6: Summary and Future Work

We summarize the dissertation and discuss some directions for future research.

Chapter 2

Exploration-Exploitation Dilemma in Sequential Decision-Making

The objective of this chapter is to introduce the *exploration-exploitation* trade-off which is at the core of our work and is a standard problem in decision-making. We present the state-of-the-art methods and results and the material that we need for the analyses of Ch. 3 and 4. We first introduce the exploration-exploitation dilemma in the well-known Multi-Armed Bandit (MAB) setting, explaining the challenges that the learner faces and presenting the two most popular principles to address this issue, based respectively on *optimism-in-face-of-uncertainty* and *randomness*. Then, we present two extensions, namely the Linear Bandit (LB) problem and the Linear Quadratic (LQ) control problem, that address the main limitations of MAB, i.e., the finite action space and the independency between rounds. LB is a natural extension of MAB with continuous action space, where the reward depends linearly on an unknown parameter and the chosen arm. Despite the fact that the action set is allowed to change with time (in the contextual setting), it cannot be affected by the actions chosen by the learner. On the other hand, Markov Decision Processes allow the system to be dynamically affected by the actions, and thus overcome this limitation. However, it usually assumes the state-action space to be finite. To this end, LQ stands as a powerful framework to extend the MDP model to continuous state-action spaces, while imposing the dynamic of the system to be linearly parametrized. We present algorithms based on the two principles in those frameworks and recall the available regret guarantees.

Contents

2.1 The multi-armed bandit problem

The Multi-Armed Bandit (MAB) problem is a sequential decision-making problem where an agent chooses at each time step an action to play and it receives a reward drawn from an unknown distribution. The aim of the agent is to maximize the cumulative

reward, i.e., the global payoff received from the chosen sequence of actions. The name Multi-Armed Bandit comes from the problem instance where a gambler (agent) faces multiple slot machines (one-armed bandits) with different unknown reward distributions and has to sequentially decide which machine (arm) to play. MAB problems have been originally introduced to study the problem of clinical trials (which treatment to use on patients suffering from the same disease) and is now used in a large number of applications (see e.g., Bubeck and Cesa-Bianchi 2012) such as allocation in finance, web-advertisement, routing etc... From a theoretical perspective, its popularity is due to the fact that it offers a simple framework to study the *exploration-exploitation* dilemma that is inherent of sequential decision making in unknown environment, where the agent balances between selecting highly rewarding actions based on the knowledge acquired so far, and playing poorly estimated actions (with potential low reward) to enhance his knowledge and select better actions in the future. This allocation problem had been extensively studied in statistics and became a standard in Reinforcement Learning (RL). In this section, we present the MAB setting and the challenges induced by the *exploration-exploitation* trade-off, the main principles used to tackle this issue and the algorithms that have been derived from those principles together with their theoretical guarantees.

2.1.1 Setting and challenges

We consider the stochastic MAB setting (see e.g., Bubeck and Cesa-Bianchi 2012, Agrawal 1995, Auer et al. 2002a) where an agent selects at each time step $t = 1, 2, \dots$ one arm (action) I_t from a finite set of K arms i.e., $I_t \in \{1, \dots, K\}$. Each arm i is associated with a distribution ν_i of mean μ_i so that when the agent plays arm I_t, it receives a reward X_{I_t} randomly generated (independently from the past) according to ν_{I_t}.

Setting. The objective of the agent is to select a sequence of arms $(I_1, I_2, \dots)$ to maximize the associated cumulative reward. Denoting the optimal arm and optimal average reward as

$$i^\star \in \operatorname*{argmax}_{i=1,\dots,K} \mu_i \quad \text{and} \quad \mu^\star = \max_{i=1,\dots,K} \mu_i,$$

the equivalent objective is to minimize the regret of the strategy i.e., the difference between the optimal reward that would have been collected playing arm $i^\star$ at each time step and the reward actually collected. Formally, we consider the expected sudo-regret defined as

$$R_n = n\mu^\star - \mathbb{E}\left(\sum_{t=1}^{n} \mu_{I_t}\right),$$

where n is the total number of rounds. The expectation is taken w.r.t. any randomization in the choice of the sequence of arms. Finally, introducing $T_i(t) = \sum_{s=1}^{t} \mathbb{1}\{I_s = i\}$ the number of times the agent selected arm i up to time t, and $\Delta_i = \mu^\star - \mu_i$ the sub-optimality gap between arm i and $i^\star$, it is possible to re-write the pseudo-regret as

$$R_n = \mu^\star \sum_{i=1}^{K} \mathbb{E}\big(T_i(n)\big) - \mathbb{E}\left(\sum_{i=1}^{K} T_i(n)\mu_i\right) = \sum_{i=1}^{K} \Delta_i \mathbb{E}\big(T_i(n)\big). \tag{2.1}$$

The agent does not know about the reward distributions $\{\nu_i\}_{i=1,\dots,K}$, and hence does not know about the average rewards $\{\mu_i\}_{i=1,\dots,K}$, but collects knowledge about their value by observing the sequence of reward $(X_{I_1}, X_{I_2}, \dots)$ generated by its own sequence of actions $(I_1, I_2, \dots)$. As a result, the agent faces an *exploration-exploitation* trade-off, where it balances between playing the most rewarding arm given its current knowledge to minimize its instantaneous regret, and playing badly estimated arms to collect information that would help it to improve its future performance.

Lower bound. Before introducing the main strategies used to tackle the exploration-exploitation trade-off in the MAB problem, we highlight the inherent difficulty of this sequential decision making problem, by recalling the existing lower-bounds. In the parametric case where $\nu_i = \nu_i(\theta_i^\star)$ are function of a unknown parameter $\theta^\star = (\theta_1^\star, \dots, \theta_K^\star) \in \Theta$, under mild assumptions on the set of parameter Θ, one has:

Theorem 2.1.1 (Lai and Robbins (1985))**.** *For any adaptive strategy whose regret satisfies, for each $\theta \in \Theta$, the condition that as $n \to \infty$,*

$$R_n(\theta) = o(n^a) \quad \text{for every } a > 0, \tag{2.2}$$

one has that for any $\theta \in \Theta$,

$$\liminf_{n\to\infty} \frac{R_n(\theta)}{\log(n)} \geq \sum_{i \neq i^\star} \frac{\Delta_i}{KL(\nu_i||\nu_{i^\star})}, \tag{2.3}$$

where $KL(\nu_i||\nu_j) = \int_{-\infty}^{\infty} \nu_i(x) \log \frac{\nu_i(x)}{\nu_j(x)} dx$ is the Kullback-Leibler divergence.

Amongst the set of *consistent* strategies i.e., allocation rules satisfying Eq. 2.2, Thm. 2.1.1 guarantees that the regret scales at best as $\Omega(\log n)$: from Eq. 2.1, it implies that an optimal strategy must select sub-optimal arms at most $\log n$ times. Additionally, Thm. 2.1.1 provides us with the optimal problem dependent constant that is function of the unknown distributions $\{\nu_i\}_{i=1,\dots,K}$. To gain intuition about this result, consider the problem instance where rewards are generated according to Bernoulli distributions with parameters $(\mu_1, \dots, \mu_K) \in [0,1]^K$ pairwise disjoint ($\mu_i \neq \mu_j$ for any $i \neq j$) and suppose that $\mu^\star = 1/2$ for sake of simplicity. Then, for all $i \neq i^\star$, the Kullback-Leibler divergence can be lower and upper bounded as

$$2\Big(\Delta_i\Big)^2 \leq KL(\nu_i||\nu_{i^\star}) \leq 4\Big(\Delta_i\Big)^2,$$

and thus the constant in Eq. 2.3 scales as $\sum_{i\neq i^\star} \frac{1}{\Delta_i}$.
Finally, a non-asymptotic minimax lower bound (see e.g., Cesa-Bianchi and Lugosi 2006), i.e., a problem-independent regret bound is given by the following theorem:

Theorem 2.1.2. *Let $\sup_\nu$ denote the supremum over all distribution of rewards (i.e., over the MAB problem instances), and let $\inf_{algo}$ denote the infimum over any adaptive strategy. Then,*

$$\inf_{algo} \sup_{\nu} R_n = \Omega\Big(\sqrt{nK}\Big). \tag{2.4}$$

2.1.2 Heuristics

In this subsection, we introduce and illustrate the two main principles that address the *exploration-exploitation* in MAB in particular and in RL in general, namely the *optimism in face of uncertainty* that leads to the celebrated Upper Confidence Bound (UCB) algorithms, and the *Thompson Sampling* that relies on randomization to enhance the exploration.

To highlight the difficulty of the problem, we first describe the behavior of a naive strategy that is known to suffer linear regret, where the choice of I_t is made greedily given the current estimates of $\{\nu_i\}_{i=1,\dots,K}$. First, notice that at time step $t \leq n$, given the sequence of arms $(I_1, \dots, I_{t-1})$ selected so far, the learner observed the sequence of associated rewards $(X_{I_1,1}, \dots, X_{I_{t-1},t-1})$ where $X_{i,s}$ is randomly drawn from ν_i at time step s independently from the past. Thus, for each arm i, it can compute the empirical estimate $\widehat{\mu}_i^t$ of the mean μ_i as

$$\widehat{\mu}_i^t = \frac{1}{T_i(t-1)} \sum_{s=1}^{t-1} X_{i,s} \mathbb{1}\{i = I_s\}, \text{ for all } i = 1, \dots, K. \tag{2.5}$$

The greedy strategy then consists in selecting the arm as

$$I_t = \underset{i=1,\dots,K}{\operatorname{argmax}} \widehat{\mu}_i^t.$$

As claimed above, this strategy is known to fail, as the algorithm can be stuck in a configuration where it keeps on selecting a sub-optimal arm. Consider for instance the 2-armed case where $i^\star = 1$ and suppose that, at some time step t, due to the randomness of the rewards, $\widehat{\mu}_{1,t} < \widehat{\mu}_{2,t}$. Then, the algorithm will pick arm 2, improving the accuracy of $\widehat{\mu}_{2,t}$ but leaving $\widehat{\mu}_{1,t}$ unchanged. Thus, it may persistently pull arm 2, which unfortunately is the sub-optimal arm. We illustrate this configuration in Fig. 2.1. This sub-optimal behavior comes from the fact that the accuracy of the estimation is only improved over chosen arms, which speaks in favor of adjusting the score of arms w.r.t. the number of pull, or alternatively of adding some random perturbation to force the exploration over badly estimated arms.

Leveraging this intuition, *optimistic* algorithms no longer select the arm with the maximum empirical average but the one with the maximum adjusted empirical average as

$$I_t = \underset{i=1,\dots,K}{\operatorname{argmax}} \left[\widehat{\mu}_i^t + B_{i,t}\right],$$

where $B_{i,t}$ is the exploration bonus that quantifies how often arm i has been pulled up to time t. Since the agent still wants to take advantage of the accumulated knowledge, $B_{i,t}$ is designed to be an upper-bound for μ_i i.e., such that $\mu_i \in [\widehat{\mu}_i^t - B_{i,t}, \widehat{\mu}_i^t + B_{i,t}]$ with high probability, which explains the Upper Confidence Bound (UCB) name for this class of algorithms. As a result, such method overcomes the difficulty faced by the greedy strategy, as illustrated in Fig. 2.2, where we consider the same initial configuration as the one of Fig. 2.1.

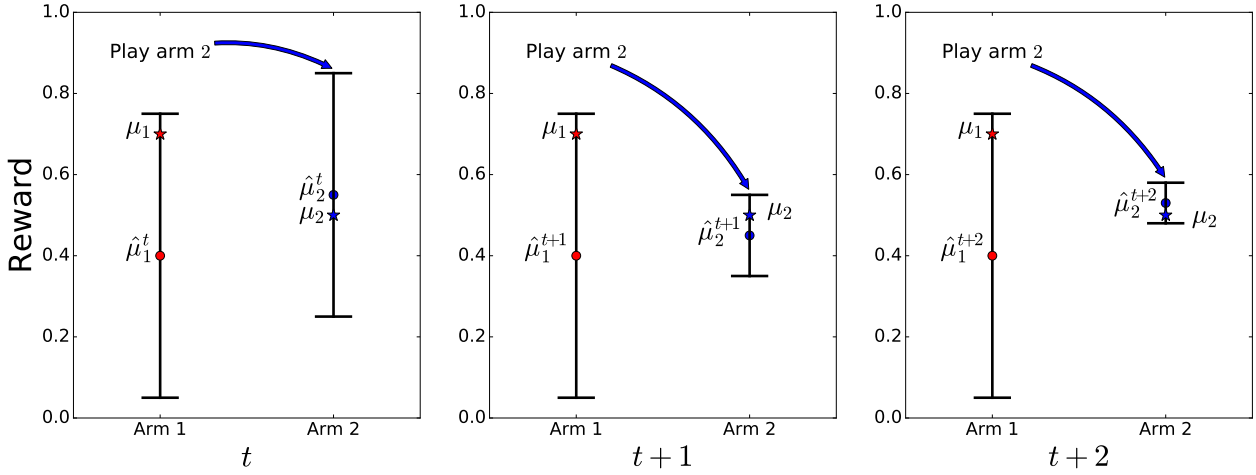

Figure 2.1 – Illustration of the functioning of the greedy strategy. *Left:* at time step t, the empirical means $\widehat{\mu}_1^t$ and $\widehat{\mu}_2^t$ are computed and deviate from the true means μ_1 and μ_2 due to the lack of accuracy of the estimation (the black segments represent the h.p. confidence intervals). Here, $\widehat{\mu}_2^t \geq \widehat{\mu}_1^t$ so the greedy strategy selects the sub-optimal arm 2. *Center:* The agent observes a reward for arm 2 randomly generated according to ν_2. Thus, $\widehat{\mu}_2$ is re-evaluated and its accuracy improves while $\widehat{\mu}_1$ are left unchanged. As a result, arm 2 is selected once again. *Right:* The greedy strategy keeps on selecting arm 2, improving the accuracy of its estimate, but is never able to discriminate that $\mu_1 \geq \mu_2$, and thus is persistently sub-optimal.

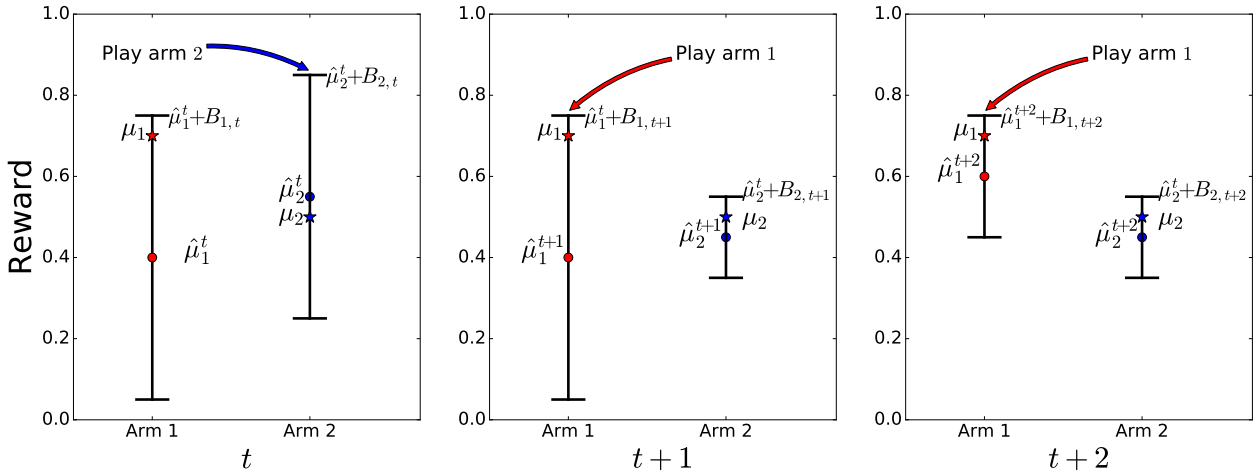

Figure 2.2 – Illustration of the functioning of the *optimistic* strategy. *Left:* The agent associates an optimistic score to each arm by adjusting the empirical means $\widehat{\mu}_{(1,2)}^t$ with a bonus $B_{(1,2),t}$ which is designed so that the score corresponds to the upper bound of the confidence interval (black segment). In this configuration, arm 2 is selected. *Center:* A reward is generated for arm 2, and its empirical mean and upper bound are re-evaluated while the one of arm 1 are left unchanged. As a result, the score of arm 1 is significantly boosted by the bonus $B_{1,t+1}$ compared to arm 2, and despite the fact that $\widehat{\mu}_2^{t+1} \geq \widehat{\mu}_1^{t+1}$, arm 1 is selected. *Right:* Similarly, a reward is generated for arm 1, and its empirical mean and upper bound are re-evaluated while the one of arm 2 are left unchanged. Improving the accuracy of the estimation implies that $\widehat{\mu}_1^{t+2}$ gets closer to μ_1 (and hence higher) so that arm 1 gets the best score and is selected by the optimistic strategy.

Finally, another principle, named Thompson Sampling (TS), is built on Bayesian ideas where one assumes a prior over the average reward μ_i's and maintains this distribution as new data are collected. Then, a prediction $\widetilde{\mu}_i$ is sample for each arm i according to the posterior and the arm is selected as

$$I_t = \operatorname*{argmax}_{i=1,\dots,K} \widetilde{\mu}_i^t.$$

The original idea dates back from Thompson (1933) and has attracted considerable attention recently because of the impressive empirical performance of TS (see e.g., Chapelle and Li 2011). Furthermore, despite the Bayesian construction of this method, TS has also been shown to perform well in the *frequentist* setting, where parameters of the reward distribution of every arms are fixed, though unknown, implying that TS can be seen as a *randomized* algorithm and that the prior assumption only acts as a convenient tool to derive the sampling distribution. The underlying idea is that the posterior plays the role of the high-probability confidence interval of UCB i.e., w.h.p. $\widetilde{\mu}_i \in [\widehat{\mu}_i^t - B_{i,t}, \widehat{\mu}_i^t + B_{i,t}]$, so TS can be seen as a randomized counterpart of UCB and hence avoids getting trapped in bad decisions. This point of view is somehow in contrast with the original Bayesian construction of TS and is at the core of our new analysis for TS in LB (see Ch. 3).

2.1.3 Algorithms and theoretical guarantees

We now present the main algorithms that had been derived from those heuristics and provide their associated theoretical guarantees. We first present the seminal UCB1 algorithm of Auer et al. (2002a) and its extensions UCB-V and KL-UCB introduced respectively by Audibert et al. (2007) and Cappé et al. (2013), Maillard et al. (2011). Then, we present two instances of TS algorithms with Bernoulli and arbitrary reward distributions respectively.

UCB algorithms. The UCB algorithm of Auer et al. (2002a) is based on the index

$$\widehat{\mu}_i^t + \sqrt{\frac{2\log t}{T_i(t-1)}} \text{ for each arm } i = 1, \dots, K,$$

where $\sqrt{\frac{2\log t}{T_i(t-1)}}$ plays the role of the exploration bonus $B_{i,t}$. This is motivated by the Chernoff-Hoeffding's inequality which ensures that

$$\mathbb{P}\left(\widehat{\mu}_{i,t} + \sqrt{\frac{2\log t}{T_i(t-1)}} \leq \mu_i\right) \leq \frac{1}{t^4}, \quad \text{for all } i = 1, \dots, K.$$

As a consequence, each index is an upper bound of the true means with high-probability. We report the formal algorithm in Fig. 2.3.

Initialization: Play each arm once, store the rewards in $X = (X_{1,1}, \dots, X_{K,1})$ and set $T_i = 1$ for all $i = 1, \dots, K$

for $t = \{1, \dots, n\}$ **do**
  Compute $\widehat{\mu}_i^t$ according to Eq. 2.5 for all $i = 1, \dots, K$
  Play arm $I_t = \text{argmax}_i \, \widehat{\mu}_i^t + \sqrt{\frac{2 \log t}{T_i}}$
  Receive reward $X_{I_t,t}$
  Update $T_{I_t} = T_{I_t} + 1$ and $X = (X, X_{I_t,t})$
end for

Figure 2.3 – UCB1 algorithm for the MAB problem with K arms.

Auer et al. (2002a) guarantee the following expected regret bound for the UCB1 algorithm:

Theorem 2.1.3. *For all $K > 1$, if policy UCB1 is run on the MAB problem with K arms, with arbitrary reward distributions $\nu_1, \dots, \nu_K$ with support in $[0, 1]$, its expected regret over n steps is at most*

$$R_n \leq 8 \log(n) \sum_{i \neq i^\star} \frac{1}{\Delta_i} + \left(1 + \pi^2/3\right) \sum_{i \neq i^\star} \Delta_i.$$

The derivation of UCB1 relies on the Chernoff-Hoeffding's inequality which does not take into account the higher moments of the distributions but the empirical average only. To refine this result, Audibert et al. (2007) introduced a variant named UCB-V which takes into account the empirical variance and is based on a more advanced concentration inequality, namely an empirical version of the Bernstein's inequality instead of Hoeffding's. Formally, the arm selection is made as

$$I_t = \underset{i=1,\dots,K}{\text{argmax}} \left(\widehat{\mu}_i^t + \sqrt{\frac{2 V_i^t \log(t)}{T_i(t-1)}} + 3 \frac{\log(t)}{T_i(t-1)} \right),$$

where V_i^t is the empirical variance of arm i at time t. They show that the regret of UCB-V is bounded as

$$R_n \leq 10 \log(n) \sum_{i \neq i^\star} \left(\frac{\sigma_i^2}{\Delta_i} + 2 \right),$$

where σ_i^2 is the variance of the distribution of arm i. This is a major improvement over Thm. 2.1.3 since it reflects the fact that an arm with low variance is easy to estimate and thus does not contribute much in the regret. Refining further this result, Cappé et al. (2013) and Maillard et al. (2011) introduced the KL-UCB algorithm using the Kullback-Leibler divergence $kl(p, q)$ of two Bernoulli distributions with parameter p and q. Formally, one has:

$$kl(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q},$$

and the arms are selected according to the index

$$u_i(t) = \max_q \left\{ q \in \left]\widehat{\mu}_i^t, 1\right] \ : \ kl(\widehat{\mu}_i^t, q) \leq \frac{\log t}{T_i(t-1)} \right\}$$

and $I_t = \underset{i=1,\dots,K}{\operatorname{argmax}}\, u_i(t)$. They show that the regret suffered by KL-UCB is bounded as

$$R_n \leq \log(n) \sum_{i \neq i^\star} \frac{\Delta_i}{kl(\mu_i, \mu^\star)} + O\Big(\sqrt{\log n}\Big),$$

which implies that KL-UCB is asymptotically optimal as it matches the lower bound in Thm. 2.1.1.

TS algorithms. In order to provide intuition about the functioning of the Thompson Sampling algorithm, we first introduce it for the Bernoulli bandit problem of Chapelle and Li (2011) i.e., we assume that the reward distributions ν_i's are Bernoulli with unknown parameters μ_i's. In this setting, the beta distribution stands as a convenient tool, since this distribution is a conjugate prior w.r.t. the Bernoulli distribution. Formally, we denote as $\text{Beta}(\alpha, \beta)$ the beta distribution with parameters $\alpha > 0$, $\beta > 0$, whose pdf is given by

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

When the prior over the reward distribution is $\text{Beta}(\alpha, \beta)$, after observing a Bernoulli trial, its posterior is simply $\text{Beta}(\alpha + 1, \beta)$ if the trial is a success, or $\text{Beta}(\alpha, \beta + 1)$ if the trial is a failure. Agrawal and Goyal (2012a) propose a specific instance of TS, where it is initially assumed that arm i has a prior $\text{Beta}(1, 1)$ over μ_i (which corresponds to the uniform distribution over $[0, 1]$), and maintain this distribution as $\text{Beta}(S_i(t) + 1, F_i(t) + 1)$ where $S_i(t)$ and $F_i(t)$ are respectively the number of success $(X_i = 1)$ and failure $(X_i = 0)$ of arm i observed so far. Then, the algorithm randomly samples from these posteriors and plays the arm with the largest sample mean. The algorithm in summarized in Fig. 2.4.

Initialization: For each arm $i = 1, \dots, K$, set $S_i = 0$ and $F_i = 0$
1: **for** $t = \{1, \dots, n\}$ **do**
2: For each arm $i = 1, \dots, K$, sample $\widetilde{\mu}_i(t) \sim \text{Beta}(S_i + 1, F_i + 1)$
3: Play arm $I_t = \text{argmax}_i\, \widetilde{\mu}_i(t)$
4: Receive reward $X_{I_t,t}$
5: If $X_{I_t,t} = 1$, update $S_{I_t} = S_{I_t} + 1$, else update $F_{I_t} = F_{I_t} + 1$
6: **end for**

Figure 2.4 – Thompson Sampling for Bernoulli Bandits.

When the reward distributions are not Bernoulli, Agrawal and Goyal (2012a) propose an extension of this algorithm based on Bernoulli trials: after observing a reward $X_{i,t} \in [0, 1]$ that is generated from an arbitrary distribution, the algorithm performs a Bernoulli trial with probability of success $X_{i,t}$, and updates the quantities $S_i(t)$ and

Initialization: For each arm $i = 1, \dots, K$, set $S_i = 0$ and $F_i = 0$
1: **for** $t = \{1, \dots, n\}$ **do**
2: For each arm $i = 1, \dots, K$, sample $\widetilde{\mu}_i(t)$ from Beta$(S_i + 1, F_i + 1)$
3: Play arm $I_t = \text{argmax}_i \, \widetilde{\mu}_i(t)$ and receive reward $X_{I_t,t}$
4: Perform a Bernoulli trial with success probability $X_{I_t,t}$ and observe output $\tilde{X}_t$
5: If $\tilde{X}_t = 1$, update $S_{I_t} = S_{I_t} + 1$, else update $F_{I_t} = F_{I_t} + 1$
6: **end for**

Figure 2.5 – Thompson Sampling for general MAB problem

$F_i(t)$ depending on whether this Bernoulli trial is a success or a failure. We summarize the algorithm in Fig. 2.5.

Agrawal and Goyal (2012a) proved the following guarantee:

Theorem 2.1.4. *The regret of the* TS *algorithm in Fig 2.5 is bounded w.h.p. as*

$$R_n \leq O\left(\left(\sum_{i \neq i^\star} \frac{1}{\Delta_i^2}\right)^2 \log n\right).$$

The bound in Thm. 2.1.4 is optimal w.r.t. the dependency on n, but sub-optimal w.r.t. the dependency on Δ_i in the constant. To overcome this, Kaufmann et al. (2012) derive a Bayes-UCB algorithm for Bernoulli rewards, which uses both the idea of UCB and TS, that achieves the lower bound of Lai and Robbins (1985). Further, Korda et al. (2013) study an instance of TS wih Jeffreys prior, and show that the induced algorithm has a regret bounded by $O\Big(\sum_{i \neq i^\star} \frac{\Delta_i}{KL(\nu_i || \nu_{i^\star})} \log n\Big)$, thus is asymptotically optimal, when the reward distributions ν_i's belong to a $1-$dimensional canonical exponential family. Finally, Agrawal and Goyal (2013) prove a $O(\sqrt{nK \log(n)})$ problem-independent regret bound for TS in MAB, that is near-optimal w.r.t. the problem-independent lower bound in Eq. 2.4, up to a factor $\log(n)$.

2.1.4 Extensions

The ability of the MAB problem to encode in a simple framework the hardness of the *exploration-exploitation* dilemma explains its popularity, and numerous variants and extensions have been studied over the last decade. For instance, while we focused in this section on the stochastic MAB problem, its *adversarial* counterpart have been studied by Bubeck and Cesa-Bianchi (2012) and Auer et al. (2002b), where the rewards are no longer assumed to be stochastic but chosen by an oblivious adversary. On the other hand, TS has also been studied in the *Bayesian* setting, where the parameters of the reward distributions are assumed to be generated from a true, yet unknown prior, and the regret is measured in average w.r.t. this prior (Bubeck and Liu, 2013). Further, enriching the structure of the rewards, May et al. (2012) considered the *contextual* bandit problem, where the reward is function of the action I_t and a context x_t. Russo et al. (2017) address the discounted regret to take into account time preference when

the optimal action is costly to learn compared to near-optimal actions and propose a variation of the TS algorithm in this setting. Finally, despite existing results for the MAB problem with more arms than the possible number of rounds i.e., $K \gg n$ (Wang et al., 2009), the main limitation of this framework is that the number of arms is finite, thus the action space cannot be continuous. This motivates one of the major extension, the Linear Bandit (LB) problem, where the action set is embedded in $\mathbb{R}^d$ and the reward is a noisy linear combination between the action and an unknown parameter. We present the LB problem in the next section.

2.2 The linear bandit problem

Despite the richness of the MAB framework and the flexibility of the model that allows very different reward distributions, one of its major limitation is its inability to model problems where the number of arms is infinite e.g., when the arm set is embedded in $\mathbb{R}^d$: without any further assumption on the problem, the lower bound of Eq. 2.3 clearly states that the regret would grow infinite (both because of infinite elements in the sum and because $\Delta \to 0$). To address this issue, Dani et al. (2008) formalized the Linear Bandit (LB) extension, that was introduced by Auer et al. (2002a), where the arms are vector of $\mathbb{R}^d$ and the payoff is a noisy and unknown linear function of the arm. Given this additional structure, they derived an *optimistic* algorithm that relies on least square estimation but shares the same structure as UCB. Further, Li et al. (2010) propose the LinUCB algorithm that uses tighter confidence bounds and has been proved later by Abbasi-Yadkori et al. (2011b). To do so, they introduce a new concentration inequality for the least square estimates which allows them to improve the regret bound of Dani et al. (2008) (see Abbasi-Yadkori et al. 2011a) since, as hinted in Sec. 2.1, the performance of optimistic algorithms are determined by the tightness of the confidence bounds. Similarly, a TS algorithm can be derived for the LB problem which has been shown to offer very good empirical performance. Agrawal and Goyal (2012b) provided the first regret analysis for the LB problem, and showed that up to a factor $\sqrt{d}$, the TS algorithm offers the same performance as the UCB-like algorithm. We present here the LB setting and the least square concentration inequality, and provide the two algorithms together with their theoretical regret bounds.

2.2.1 Setting

We consider the stochastic linear bandit extension of Dani et al. (2008). Let $\mathcal{X} \subset \mathbb{R}^d$ be an arbitrary (finite or infinite) bounded set of arms. When an arm $x \in \mathcal{X}$ is pulled, a reward is generated as

$$r(x) = x^\mathsf{T}\theta^\star + \xi,$$

where $\theta^\star \in \mathbb{R}^d$ is a fixed but unknown parameter and ξ is a zero-mean noise. An arm $x \in \mathcal{X}$ is evaluated according to its expected reward $x^\mathsf{T}\theta^\star$ and for any $\theta \in \mathbb{R}^d$ we denote

the optimal arm and its value by

$$x^\star(\theta) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta, \qquad J(\theta) = \sup_{x\in\mathcal{X}} x^\mathsf{T}\theta.$$

Then $x^\star = x^\star(\theta^\star)$ is the optimal arm for $\theta^\star$ and $J(\theta^\star)$ is its optimal value. At each step t, the learner selects an arm $x_t \in \mathcal{X}$ based on the past observations (and possibly additional randomization), it observes the reward $r_{t+1} = x_t^\mathsf{T}\theta^\star + \xi_{t+1}$, and it suffers a *regret* equal to the difference in expected reward between the optimal arm $x^\star$ and the arm x_t. All the information observed up to time t is encoded in the filtration $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, r_2, \dots, r_t, x_t))$, where $\mathcal{F}_1$ contains any prior knowledge. The objective of the learner is to minimize the *cumulative regret* up to step T, i.e.,

$$R(T) = \sum_{t=1}^{T} \left(x^{\star,\mathsf{T}}\theta^\star - x_t^\mathsf{T}\theta^\star\right).$$

2.2.2 RLS estimation

The stochastic LB problem is characterized by bandit feedback, in the sense that the learner only observes the rewards without any additional information about the components of $\theta^\star$. However, an estimate $\widehat{\theta}_t$ can be computed at each time step using the standard least square procedure. Formally, let $(x_1, \dots, x_t) \in \mathcal{X}^t$ be a sequence of arms chosen so far and $(r_2, \dots, r_{t+1})$ be the corresponding rewards, then $\theta^\star$ can be estimated by regularized least-squares (RLS). For any regularization parameter $\lambda \in \mathbb{R}^+$, the design matrix and the RLS estimate are defined as

$$V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\mathsf{T}, \qquad \widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} x_s r_{s+1}.$$

One of the many advantages of the RLS is that it offers strong theoretical guarantees. Leveraging the theory of self-normalized processes (see De La Pena et al. 2009), Abbasi-Yadkori et al. (2011b) derived a new concentration inequality for the RLS estimate. Notice that the analysis in the online setting is non-trivial because of the correlations between the data points (the covariates x_t's are chosen by the learner based on his knowledge and thus mutually dependent). Formally, they show the following result:

Proposition 2.2.1 (Thm. 8 in (Abbasi-Yadkori et al., 2011b))**.** *Assume that the noise sequence $\{\xi_{t+1}\}_{t\geq 1}$ is a $\mathcal{F}_t$-martingale difference sequence, conditionally subgaussian with constant R. Then, for any $\delta \in (0,1)$, for any $\mathcal{F}_t$-adapted sequence $(x_1, \dots, x_t)$, the RLS estimator $\widehat{\theta}_t$ is such that for any fixed $t \geq 1$,*

$$\|\widehat{\theta}_t - \theta^\star\|_{V_t} \leq \beta_t(\delta)$$

w.p. $1-\delta$ (w.r.t. the noise $\{\xi_t\}_t$ and any source of randomization in the choice of the arms), with

$$\beta_t(\delta) = R\sqrt{2\log\frac{(\lambda + tX^2)^{d/2}\lambda^{-d/2}}{\delta}} + \sqrt{\lambda}S.$$

where X and S are constants such that $\|x\| \leq X$ for all $x \in \mathcal{X}$ and $\|\theta^\star\| \leq S$.

2.2.3 Algorithms and theoretical guarantees

As for the MAB framework, two type of algorithms can be derived from the *optimistic* and *random* principles presented in Sec. 2.1. We first present the Optimism in Face of Uncertainty for Linear bandit (OFUL) algorithm of Abbasi-Yadkori et al. (2011a) which is a refined version of the ConfidenceBall algorithm of Dani et al. (2008), and then present the TS algorithm for LB of Agrawal and Goyal (2012b) together with their respective theoretical guarantees.

OFUL algorithm. According to the concentration inequality for RLS estimates, one can defined a confidence ellipsoid $\mathcal{E}_t^{\text{RLS}}$ at each time step as

$$\mathcal{E}_t^{\text{RLS}} = \Big\{\theta \in \mathbb{R}^d \mid \|\theta - \widehat{\theta}_t\|_{V_t} \leq \beta_t(\delta)\Big\},$$

which is centered around $\widehat{\theta}_t$ with orientation defined by V_t and radius $\beta_t(\delta)$. Thus, with high probability $\theta^\star \in \mathcal{E}_t^{\text{RLS}}$ so that $\mathcal{E}_t^{\text{RLS}}$ plays the role of the confidence bound of the UCB algorithm. Formally, the OFUL algorithm selects at each time step the *optimistic* parameter

$$\widetilde{\theta}_t = \underset{\theta \in \mathcal{E}_t^{\text{RLS}}}{\text{argmax}}\, J(\theta),$$

and then chooses the optimal arm w.r.t. $\widetilde{\theta}_t$ as $x_t = x^\star(\widetilde{\theta}_t)$. We summarize the algorithm in Fig. 2.6.

Initialization: Set $\widehat{\theta}_1 = 0$ and $V_1 = \lambda I$
1: **for** $t = \{1, \dots, n\}$ **do**
2: Define the confidence ellipsoid $\mathcal{E}_t^{\text{RLS}} = \{\theta \in \mathbb{R}^d \mid \|\theta - \widehat{\theta}_t\|_{V_t} \leq \beta_t(\delta)\}$
3: Select the optimal parameter $\widetilde{\theta}_t = \text{argmax}_{\theta \in \mathcal{E}_t^{\text{RLS}}}\, J(\theta)$
4: Play the arm $x_t = x^\star(\widetilde{\theta}_t) = \text{argmax}_{x \in \mathcal{X}}\, x^\mathsf{T}\widetilde{\theta}_t$
5: Observe reward $r_{t+1} = x_t^\mathsf{T}\theta^\star + \xi_{t+1}$
6: Update the RLS estimate $\widehat{\theta}_{t+1}$ and design matrix V_{t+1}
7: **end for**

Figure 2.6 – OFUL algorithm.

Under the assumptions of Prop. 2.2.1, Abbasi-Yadkori et al. (2011a) proved the following result:

Theorem 2.2.1. *For any $0 < \delta < 1$, with probability at least $1 - \delta$, the regret of the OFUL algorithm in Fig. 2.6 is bounded as*

$$R(T) = O\Big(d \log(T)\sqrt{T} + \sqrt{dT \log(T/\delta)}\Big).$$

Therefore, apart for logarithmic factors, the OFUL algorithm is optimal.

TS algorithm. To design the TS algorithm, Agrawal and Goyal (2012b) use a Gaussian prior for the unknown parameter $\theta^\star$. The motivation is that whenever the reward noise ξ_t is conditionally Gaussian, the linear model ensures that the posterior is also Gaussian. Notice that none of those assumptions are required to be true but that they only provide a useful tool to obtain the sampling distribution. Formally, at each time step, a parameter $\widetilde{\theta}_t$ is randomly sampled according to $\mathcal{N}(\widehat{\theta}_t, vV_t^{-1})$, where v is a parameter, and the optimal arm w.r.t. $\widetilde{\theta}_t$ is chosen i.e., $x_t = x^\star(\widetilde{\theta}_t)$. The TS algorithm is summarized on Fig. 2.7.

Initialization: Set $\widehat{\theta}_1 = 0$ and $V_1 = \lambda I$
1: **for** $t = \{1, \dots, n\}$ **do**
2: Sample the parameter as $\widetilde{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, vV_t^{-1})$
3: Play the arm $x_t = x^\star(\widetilde{\theta}_t) = \text{argmax}_{x\in\mathcal{X}}\, x^\mathsf{T}\widetilde{\theta}_t$
4: Observe reward $r_{t+1} = x_t^\mathsf{T}\theta^\star + \xi_{t+1}$
5: Update the RLS estimate $\widehat{\theta}_{t+1}$ and design matrix V_{t+1}
6: **end for**

Figure 2.7 – TS algorithm.

Notice that in line with MAB, the sampling of the TS algorithm for LB is made so that $\widetilde{\theta}_t$ spans the confidence ellipsoid $\mathcal{E}_t^{\text{RLS}}$. The variance V_t^{-1} is rescaled by the tuning parameter v, which is here to ensure that the sampling covers the whole ellipsoid with sufficient probability. In practice $v = R\sqrt{9d\log(T/\delta)}$ if T is known, or replaced by $v_t = R\sqrt{9d\log(t/\delta)}$ at time t if the horizon T is unknown. Leveraging the proof structure of TS for MAB (Agrawal and Goyal, 2013), Agrawal and Goyal (2012b) proved the following guarantee:

Theorem 2.2.2. *For any $0 < \delta < 1$, with probability at least $1-\delta$, the regret of the* TS *algorithm in Fig 2.7 is bounded as*

$$R(T) = O\Big(d^{3/2}\log(T)\sqrt{T} + d\sqrt{dT\log(T/\delta)}\Big).$$

The bound in Thm. 2.2.2 is $\sqrt{d}$ worst than the bound of OFUL. Despite the very good empirical performance of TS, whether this bound is tight or not is an open question. On the other hand, notice that the TS algorithm is computationally more efficient than OFUL, as the latter requires solving a bilinear optimization problem (i.e., $\arg\max_\theta \max_x x^\mathsf{T}\theta$) whereas the computational complexity of the former is dominated by the sampling and the computation of the optimal action (i.e., $\arg\max_x x^\mathsf{T}\theta$).

2.2.4 Extensions

We presented in this section the LB problem with fixed arm set $\mathcal{X}$. However, both Abbasi-Yadkori et al. (2011a) and Agrawal and Goyal (2012b) show that their results are still valid in the so-called *contextual* setting, where the arm set is allowed to vary with time,

i.e., replacing at each time step $\mathcal{X}$ by $\mathcal{X}_t$. While this setting is richer, it is limited to the case where the changes in the arm set are independent of the learner's actions and cannot handle the case where the learner's decisions affect the environment (and thus $\mathcal{X}_t$). To do so, one has to consider a more complicated setting, such as Markov Decision Processes (MDP), that we present in the next section. On the other hand, efforts have been made to relax the linear assumption on the reward model which led to the Generalized Linear Model extension of LB that is widely used in practice (see Filippi et al. 2010, Li et al. 2017, Jun et al. 2017). Finally, Lattimore and Szepesvari (2017) recently analyzed the asymptotic problem-dependent regret in LB and showed that no algorithm based on TS or optimism can achieve the lower bound in the finite arm setting.

2.3 Markov decision processes and linear quadratic control

In this section, we consider the more challenging setting where the agent's actions influence the environment's dynamics, which permits to overcome one of the limitations of the *contextual* bandit framework. To leverage the existing result of dynamic control, we restrict to the case where the environment's dynamics are Markovian, i.e., we focus on Markov Decision Processes (MDP) (Sutton and Barto, 1998). This is motivated by the fact that it allows one to use Bellman operators to compute the optimal policy (see (Bertsekas, 1995) for an introduction), i.e., the mapping from observations to actions which maximizes the total reward. We illustrate the agent/environment interactions in Fig. 2.8.

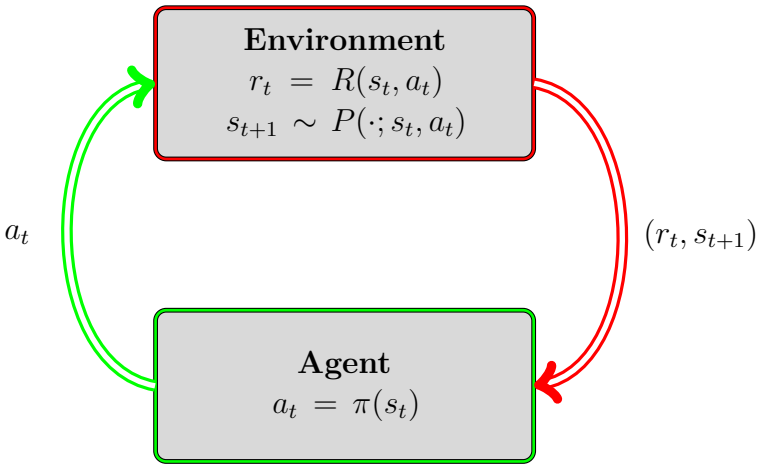

Figure 2.8 – Agent/environment interactions and dynamics in MDP.

At each time step t, the agent decides which action a_t to take, based on his current knowledge, encoded in the state s_t which characterizes the environment, according to a policy π. For each state-action pair, the environment returns a reward $r_t = R(s_t, a_t)$ and the system evolves to the next state s_{t+1} according to a transition model P which

is function of the current state and action (since we restrict to Markovian dynamics). Finally, the agent's objective is to find a policy π that maximizes the cumulative reward. When the transition and reward models are known, it is possible to compute the optimal policy using standard technics via the resolution of the well-known associated Bellman equation. For instance in the discrete case with finite state-action pairs, one can use Policy Iteration (PI) and Value Iteration (VI). On the other hand, when the transition and/or reward models are unknown, the agent faces an *exploration-exploitation* trade-off since it selects actions both to maximize the rewards and to get knowledge about the transition and/or reward model.
In this section, we first provide a brief overview of the standard finite state and action space MDP's for which algorithms have been derived using the *optimistic* and *random* principles. However, as for MAB, the finite state-action space property is a major limitation, that can be overcome by looking at *parametrized* MDPs with continuous state-action space. To this end, we present the Linear Quadratic (LQ) control problem, where the transition model is parametrized according to a linear model, which stands as a standard in control theory. One of the main advantages of LQ is that the Bellman equation can be solved efficiently, and that the unknown dynamic can be estimated via least square. We will present the setting and the optimistic algorithm introduced by Abbasi-Yadkori and Szepesvári (2011).

2.3.1 Markov decision process

Setting. A MDP is defined as a tuple $M = (\mathcal{S}, \mathcal{A}, P, R)$ where $\mathcal{S}$ and $\mathcal{A}$ are respectively the state and action space that are assumed to be finite with cardinality S and A, P is the transition model that defines the underlying Markov chain modeling the environment dynamic and R is the reward function. Formally, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ such that for any $(s, s') \in \mathcal{S}^2$, for any $a \in \mathcal{A}$,

$$P(s', s, a) = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a),$$

is the probability of observing the next state s' when action a is taken at state s and $R(s, a)$ is the associated reward. The objective is to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ mapping states to actions that maximizes the expected cumulative reward. Several instances of this problem have been studied depending on the horizon (e.g., finite horizon, infinite horizon with discount). Here, we focus on the *infinite horizon with average reward*, that consists in maximizing the expected average reward

$$\rho(M, \pi, s) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left(\sum_{t=1}^{T-1} r(s_t, a_t) \Big| s_0 = s, a_t = \pi(x_t) \; \forall \, t \geq 1 \right).$$

The difficulty of learning the optimal policy in MDP depends on size of the state-action space $S \times A$ and on the transition model P. To measure this transition structure, Jaksch et al. (2010) propose to consider the *diameter* D of the MDP which is defined as the time it takes to move from any state s to any other state s' under an appropriate policy.

Definition 2.3.1. *Let $T(s'|M,\pi,s)$ be the random variable for the first time step in which s' is reached from state s, for MDP M and stationary policy π. Then, the diameter of M is defined as:*

$$D(M) := \max_{s \neq s' \in \mathcal{S}} \min_{\pi} \mathbb{E}\Big[T(s'|M,\pi,s)\Big].$$

We further consider MDPs with finite diameter which are usually known as *communicating*. In this case, the optimal average reward $\rho^\star(M)$ does not depend on the initial state (see e.g., Puterman 2014), i.e.,

$$\rho^\star(M) := \rho^\star(M,s) = \max_{\pi} \rho(M,\pi,s).$$

Finally, we denote as $\pi^\star(M) = \arg\max_\pi \rho(M,\pi)$ the optimal policy w.r.t. M and define the regret of any adaptive strategies, i.e., a sequence of policy $(\pi_1,\dots,\pi_T)$ by:

$$R(T) = T\rho^\star(M) - \sum_{t=1}^{T} r\Big(s_t, \pi_t(s_t)\Big),$$

for which Jaksch et al. (2010) prove a lower bound of $\Omega\Big(\sqrt{DSAT}\Big)$.

UCRL2 algorithm. We now describe the UCRL2 algorithm, based on the *optimistic* principle, proposed by Jaksch et al. (2010) (a similar algorithm named REGAL has also been introduced by Bartlett and Tewari (2009)). In the general setting, the agent does not know about the transition model P and the reward function R of the true MDP M. However, it observes, at each time step t, the current state of the system s_t and the associated reward $r_t = R(s_t,a_t)$. As a result, it has access to empirical estimate $\widehat{P}_t(s',s,a)$ and $\widehat{R}_t(s,a)$ for any $s',s,a \in \mathcal{S}\times\mathcal{S}\times\mathcal{A}$ defined as

$$\widehat{R}_t = \frac{\sum_{\tau=1}^{t-1} r_t \mathbb{1}\{s_\tau = s, a_\tau = a\}}{\sum_{\tau=1}^{t-1} \mathbb{1}\{s_\tau = s, a_\tau = a\}} \;;\; \widehat{P}_t(s',s,a) = \frac{\sum_{\tau=1}^{t-1} \mathbb{1}\{s_{\tau+1} = s', s_\tau = s, a_\tau = a\}}{\sum_{\tau=1}^{t-1} \mathbb{1}\{s_\tau = s, a_\tau = a\}}. \quad (2.6)$$

Using concentration inequalities, Jaksch et al. (2010) introduce a confidence set $\mathcal{C}_t$ for the MDP at time t, that contains the MDPs with transition model $\widetilde{P}$ and reward function $\widetilde{R}$ such that, for any state-action pair (s,a),

$$\Big|\widetilde{R}(s,a) - \widehat{R}(s,a)\Big| \leq \sqrt{\frac{7\log(2SAt/\delta)}{2\sum_{\tau=1}^{t-1}\mathbb{1}\{s_\tau = s, a_\tau = a\}}},$$

$$\Big\|\widetilde{P}(\cdot,s,a) - \widehat{P}(\cdot,s,a)\Big\|_1 \leq \sqrt{\frac{14S\log(2At/\delta)}{2\sum_{\tau=1}^{t-1}\mathbb{1}\{s_\tau = s, a_\tau = a\}}},$$

where $\delta \in (0,1)$, so that, with probability at least $1-\delta$, $M \in \mathcal{C}_t$.
The structure of UCRL2 is very similar to UCB and OFUL, the only difference being that the chosen policy is kept constant for episodes instead of being re-evaluated at each time step: at the beginning of each episode, the agent selects the most optimistic MDP $\widetilde{M}$ within the confidence set, and compute the optimal policy $\pi^\star(\widetilde{M})$ associated with $\widetilde{M}$; then it follows this policy for a whole episode, it observes the states and

rewards, and refines the confidence set given the new observations at the end of the episode. An episode ends as soon as the number of visit doubles for a state-action pair. This way of updating the policy is used both for theoretical purpose, and for reducing the computational complexity. We summarize the UCRL2 algorithm in Fig. 2.9.

Initialization: Set $\nu(s,a) = 0$ and $N(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, s_1, π_0
for $t = \{1, \dots, T\}$ **do**
  if $\nu(s_t, \pi_{t-1}(s_t)) > N(s_t, \pi_{t-1}(s_t))$ **then**
    Update $\widehat{P}_t$ and $\widehat{R}_t$ by Eq. 2.6
    Find $\widetilde{M}_t = \arg\max_{M \in \mathcal{C}_t} \rho^\star(M)$
    Compute $\pi_t = \arg\max_\pi \rho(\widetilde{M}_t, \pi)$
    Let $\nu(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
  else
    $\pi_t = \pi_{t-1}$
    Choose action $a_t = \pi_t(s_t)$
    Obtains reward r_t and observe next state s_{t+1}
    Update $\nu(s_t, a_t) = \nu(s_t, a_t) + 1$ and $N(s_t, a_t) = N(s_t, a_t) + 1$
  end if
end for

Figure 2.9 – UCRL2 algorithm.

Notice that as opposed to UCB or OFUL, finding the optimistic MDP $\widetilde{M}_t$ and computing its optimal policy π_t is a complicated task. To solve this issue, Jaksch et al. (2010) introduced an extended value iteration procedure that allows them to compute directly the policy π_t, and proved the following regret bound which is a significant improvement over the bound of the UCRL algorithm previously introduced by Auer and Ortner (2007).

Theorem 2.3.1. *For any $0 < \delta < 1$, with probability at least $1 - \delta$, for any initial state s and any $T > 1$, the regret of the UCRL2 algorithm in Fig 2.9 is bounded as*

$$R(T) \leq 34DS\sqrt{AT \log\left(\frac{T}{\delta}\right)},$$

PSRL algorithm. In line with MAB and LB, a Thompson Sampling algorithm for MDP, known as Posterior Sampling for Reinforcement Learning (PSRL) have been derived (Strens, 2000). The idea is to assume a prior over the true MDP M i.e., over the transition model and reward function T and R and to maintain this distribution as new data are collected using prior-posterior update. Then, at each time step t, a MDP $\widetilde{M}_t$ is sampled from the posterior, and the actions are chosen according to the policy π_t that is optimal w.r.t. $\widetilde{M}_t$. From a practical perspective, PSRL is therefore much more efficient than UCRL2 since it does not require an extended value iteration procedure (to find the optimistic MDP) but a standard value iteration procedure to extract π_t from

$\widetilde{M}_t$. Additionally, PSRL has been shown to offer very good empirical performances. However, the analysis of PSRL in the same frequentist setting as UCRL2 is significantly more difficult, and most theoretical guarantees are provided for the *Bayesian* regret, i.e., the expected regret w.r.t. to the prior over the true MDP, and are restricted to finite and episodic MDP. Osband et al. (2013) proved the first regret bound for PSRL in finite MDPs of order $O(S\sqrt{AT})$. Osband and Roy (2014) studied finite factored MDPs (i.e., MDPs where dynamics and rewards are factored over the multidimensional representation of the state space), showing that their structure can be exploited to reduce the dependency on the number of states and actions in the final bound. The more general setting of learning in parameterized MDPs is studied in (Osband and Van Roy, 2014), where it is shown that the regret of PSRL depends on the dimensionality of the space of parameters rather its cardinality. Recently, Osband and Van Roy (2017) compared the behavior of randomized and optimism-based algorithm, showing that existing optimistic algorithms trade-off statistical efficiency with tractability while randomized approaches may enable simultaneous statistical and computational efficiency.
Unfortunately, when moving from the episodic to infinite horizon setting the results of PSRL are very limited. While most of the results for UCRL hold in both cases, Osband and Van Roy (2016) reviewed in detail the challenges of extending episodic results to infinite horizon showing how previous attempts in proving regret for infinite horizon problem were possibly flawed (Abbasi-Yadkori and Szepesvári, 2015). Notable exceptions are the work of Gopalan and Mannor (2015) who proved frequentist regret bounds in a slightly more general non-episodic setting under the assumption that the MDP is ergodic and that the initial state is positive recurrent under any policy, and the recent result of Agrawal and Jia (2017) who proved a $\widetilde{O}(D\sqrt{SAT})$ high-probability frequentist regret bound for any communicating MDP in the infinite horizon with average reward setting.

2.3.2 Linear quadratic control

As for MAB, the main limitation of finite MDPs is their inability to model systems with continuous state-action space, or even large but finite state-action space, since as stressed by the lower bound of Jaksch et al. (2010), the regret scales at best as $\sqrt{SA}$ w.r.t. the cardinality. To move from finite to continuous state-action space, the natural idea is to parametrize the MDP (see e.g., Abbasi-Yadkori and Szepesvári 2015), i.e., to impose a structure over the transition model and reward function so that one can learn the model without observing all state-action pairs. On the other hand, this makes the control problem, i.e., the resolution of the Bellman equation, much harder as Value Iteration (VI) and Policy Iteration (PI) cannot be applied. In order to overcome this issue, standard approaches consist in approximating the value function or discretizing the state-space to perform (VI) and (PI), thus narrowing down the interest of the continuous state-action space formulation.
A notable exception is the Linear Quadratic (LQ) control problem which is a specific instance of parametrized MDP with continuous state-action space where the dynamics of the environment is linear in the state and control, and the cost function is quadratic

(LQ is generally introduced as a cost minimization problem rather than a reward maximization problem). This setting stands as a standard in the control literature because the Bellman equation can be turned into a discrete Riccati equation, which can be solved efficiently. From a practical perspective, LQ problems have been intensively used and studied to address problems in many different fields such as robotics, economics, bioengineering, finance etc... We provide in this section an overview of the LQ theory. The interested reader may refer to (Bertsekas, 1995) for a first introduction and to (Lancaster and Rodman, 1995) for a complete survey of the underlying Riccati equation. Finally, we present the OFU-LQ algorithm introduced by Abbasi-Yadkori and Szepesvári (2011) that uses the *optimistic* principle to address the *exploration-exploitation* trade-off in LQ system.

Setting. We adopt the standard notations of the LQ theory and consider the dynamic system of Fig. 2.8 where the environment is characterized by a state $x_t \in \mathbb{R}^n$ and control $u_t \in \mathbb{R}^d$, and evolves according to the linear dynamic

$$x_{t+1} = Ax_t + Bu_t + \epsilon^x_{t+1},$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times d}$ and $\{\epsilon^x_{t+1}\}_{t\geq 0}$ is a martingale difference sequence w.r.t. the filtration $\mathcal{F}_t = \sigma(x_0, \dots, x_t)$ such that $\mathbb{V}(\epsilon^x_{t+1}|\mathcal{F}_t) = \Sigma^x$. The objective of the agent is to find a policy $\pi : \mathbb{R}^n \to \mathbb{R}^d$ mapping state to control that minimizes the average expected cost

$$J_\pi = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T-1} c(x_t, u_t) \Big| x_0 = s, u_t = \pi(x_t),\ \forall\, t \geq 1 \right).$$

The cost function is quadratic in the state and action as:

$$c(x, u) = x^\mathsf{T} Q x^\mathsf{T} + 2x^\mathsf{T} N u + u^\mathsf{T} R u,$$

where Q, R, N are matrices of appropriate dimensions. For sake of convenience, we collect them into a single matrix $\mathcal{Q} = \begin{pmatrix} Q & N \\ N^\mathsf{T} & R \end{pmatrix}$. Before stating the main result of the LQ theory, we introduce several properties characterizing the linear system.

Definition 2.3.2 (Ch.4 in (Lancaster and Rodman, 1995))**.** *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times d}$ be two real-valued matrices and let $\mathcal{C} = \left(B, AB, \dots, A^{n-1}B\right)$ be the associated controllability matrix. Then,*

- *The pair (A, B) is said to be **controllable** if and only if the controllability matrix $\mathcal{C}$ is of rank n (full row rank).*
- *The pair (A, B) is said to be **stabilizable** if and only if for any $x \in Ker(\mathcal{C})$, $\|Ax\| \leq \|x\|$.*

The *controllability* of the pair (A, B) ensures that every state x' is reachable from any state x in less than n time steps by using a suitable sequence of control. As a

consequence, this property can be seen as the continuous counterpart of the finite diameter in MDP. *Stabilizability* is a weaker notion that ensures that the uncontrollable part of the linear system is stable. Formally, a matrix M is said to be stable if $\|M\|_2 \leq 1$, where $\|\cdot\|_2$ is the spectral norm. Symmetrically, we introduce the notion of *observability* and *detectability* that ensures that it is possible to retrieve the initial state of the system driven by $x_{t+1} = Ax_t + \epsilon^x_{t+1}$ from any sequence of n observations given by $y_t = Cx_t$.

Definition 2.3.3 (Ch.4 in (Lancaster and Rodman, 1995))**.** *Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{d\times n}$ be two real-valued matrices. Then,*

- *The pair (C, A) is said to be **observable** if and only if the pair $(A^\mathsf{T}, C^\mathsf{T})$ is controllable.*
- *The pair (C, A) is said to be **detectable** if and only if the pair $(A^\mathsf{T}, C^\mathsf{T})$ is stabilizable.*

To ensure the existence and uniqueness of an optimal solution to the LQ problem, we consider the following assumption.

Assumption 2.3.1. *The pair (A, B) is stabilizable and the matrix $\mathcal{Q}$ is symmetric positive definite.*

In some case, the positive definiteness of $\mathcal{Q}$ is a too restrictive assumption, which can be relaxed by the weaker assumption:

Assumption 2.3.2. *The pair (A, B) is stabilizable, the matrix $\mathcal{Q}$ is symmetric non singular and the pair (Q, A) is observable.*

Under those assumptions, it is well known that the LQ problem admits a unique optimal policy which is linear in the state.

Theorem 2.3.2 (Th.16.6.4 in (Lancaster and Rodman, 1995))**.** *Under Asm. 2.3.1 or 2.3.2, the optimal solution of the LQ problem is unique and given by*

$$\begin{aligned}
&\pi^\star(x) = Kx,\\
&K = -(R + B^\mathsf{T}PB)^{-1}(B^\mathsf{T}PA + N^\mathsf{T}),\\
&P = Q + A^\mathsf{T}PA - (A^\mathsf{T}PB + N)(R + B^\mathsf{T}PB)^{-1}(B^\mathsf{T}PA + N^\mathsf{T}).
\end{aligned}$$

Further, the closed-loop matrix $A + BK$ is asymptotically stable and the optimal cost is given by $J_{\pi^\star} = Tr(P\Sigma^x)$.

Notice that the optimal control matrix K is function of a matrix P, that is the solution of a so-called *discrete Riccati equation*, which can be computed efficiently thus making the LQ solution tractable (see Laub 1991, Van Dooren 1981, Chun-hua 1998, Laub 1979).

RL in LQ systems. While Thm. 2.3.2 ensures that the optimal solution to the LQ problem is tractable for a given system with known matrices, it is often the case that the matrices are unknown and have to be estimated online, i.e., while controlling the system. To tackle this problem, Abbasi-Yadkori and Szepesvári (2011) introduced an optimistic algorithm called OFU-LQ, which addresses the induced *exploration-exploitation* trade-off in LQ system. We recall here their setting, we present the OFU-LQ algorithm, and its theoretical guarantees.
They consider the LQ system with cost function $c(x,u) = x^\mathsf{T} Q x + u^\mathsf{T} R u$ and assume that the cost matrices Q and R are known to the agent while the dynamic of the system is unknown. Additionally, they assume for sake of simplicity that $\Sigma^x = I$. Formally, they denote as (A_*, B_*) the unknown matrices of the true dynamics and collect them in a matrix $\theta_*^\mathsf{T} = (A_*, B_*)$. The objective of the learning strategy is to find a sequence of policy $\{\pi_t\}_t$ that minimizes the regret w.r.t. the unknown optimal average cost $J_* = J_{\pi^\star(\theta_*)} = \mathrm{Tr} P(\theta_*)$ defined as:

$$R(T) = \sum_{t=0}^{T} x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - T J_*,$$

where $u_t = \pi_t(x_t)$ for all $t = 0, \dots, T$.
Leveraging the linear structure of the state dynamic, at each time step, Abbasi-Yadkori and Szepesvári (2011) estimate θ_* given the past control sequence $(u_0, \dots, u_{t-1})$ and associated states $(x_0, \dots, x_t)$ using RLS, for any regularization parameter $\lambda > 0$, as:

$$V_t = \lambda I + \sum_{s=0}^{t-1} z_s z_s^\mathsf{T}; \qquad \widehat{\theta}_t = V_t^{-1} \sum_{s=0}^{t-1} z_s x_{s+1}^\mathsf{T}, \tag{2.7}$$

where $z_t = (x_t, u_t)^\mathsf{T}$. A concentration inequality for this matrix RLS estimate can be derived in a straightforward manner from Prop. 2.2.1.

Proposition 2.3.1. *For any $\delta \in (0,1)$ and any $\mathcal{F}_t$-adapted sequence $(z_0, \dots, z_t)$, the RLS estimator $\hat{\theta}_t$ is such that*

$$Tr\left((\hat{\theta}_t - \theta_*)^\mathsf{T} V_t (\hat{\theta}_t - \theta_*)\right) \leq \beta_t^2(\delta); \quad \beta_t(\delta) = n\sqrt{2\log\left(\frac{\det(V_t)^{1/2}}{\det(\lambda I)^{1/2}\delta}\right)} + \lambda^{1/2} S, \tag{2.8}$$

w.p. $1-\delta$ (w.r.t. the noise $\{\epsilon_t^x\}_t$ and any randomization in the choice of the control). S is a positive constant such that $\|\theta_*\| \leq S$.

Let $\mathcal{E}_t^{\mathrm{RLS}} := \{\theta \in \mathbb{R}^{n(n+d)} \text{ s.t. } \mathrm{Tr}\left((\hat{\theta}_t - \theta)^\mathsf{T} V_t (\hat{\theta}_t - \theta)\right) \leq \beta_t^2(\delta)\}$ be the confidence ellipsoid such that $\theta_* \in \mathcal{E}_t^{\mathrm{RLS}}$ with probability $1-\delta$. Abbasi-Yadkori and Szepesvári (2011) propose an algorithm that follows the same structure as UCRL2: at the beginning of each episode, an optimistic parameter $\widetilde{\theta}_t$ is chosen as

$$\widetilde{\theta}_t = \arg \min_{\theta \in \Theta \cap \mathcal{E}_t^{\mathrm{RLS}}} \mathrm{Tr} P(\theta),$$

where they add the constraint $\theta \in \Theta$ (we refer to (Abbasi-Yadkori and Szepesvári, 2011) for the formal definition of Θ) to guarantee the controllability of $\widetilde{\theta}_t^\mathsf{T} = (\widetilde{A}_t, \widetilde{B}_t)$. Then,

the system is optimally controlled w.r.t. $\widetilde{\theta}_t$ for the episode as $u_t = K(\widetilde{\theta}_t)x_t$. Finally, they use a *doubling schedule* to determine the length of the episodes. In this specific LQ setting, Abbasi-Yadkori and Szepesvári (2011) propose to update the policy when the determinant of the design matrix $\det(V_t)$ doubles. We summarize the OFU-LQ algorithm in Fig. 2.10.

Initialization: Set $\widehat{\theta}_0 = O$ and $V_0 = \lambda I$
for $t = \{1, \dots, T\}$ **do**
  if $\det(V_t) > 2\det(V_0)$ **then**
    Calculate $\widehat{\theta}_t$ by Eq. 2.7
    Find $\widetilde{\theta}_t = \arg\min_{\theta \in \Theta \cap \mathcal{E}_t^{\text{RLS}}} \text{Tr}P(\theta)$
    Let $V_0 = V_t$
  else
    $\widetilde{\theta}_t = \widehat{\theta}_{t-1}$
  end if
  Choose optimal control u_t based on the current parameters, $u_t = K(\widetilde{\theta}_t)x_t$
  Execute control, observe next state x_{t+1}
  Save (x_{t+1}, z_t) into the dataset, where $z_t^\mathsf{T} = (x_t^\mathsf{T}, u_t^\mathsf{T})$
  Update $V_{t+1} = V_t + x_t x_t^\mathsf{T}$
end for

Figure 2.10 – OFU-LQ algorithm.

Abbasi-Yadkori and Szepesvári (2011) prove the following regret bound for the OFU-LQ algorithm:

Theorem 2.3.3. *For any $0 < \delta < 1$, for any time T, with probability at least $1 - \delta$, the regret of* OFU-LQ *algorithm is bounded as*

$$R(T) = \widetilde{O}\Big(\sqrt{T \log(1/\delta)}\Big)$$

where $\widetilde{O}$ hides logarithmic factors and problem dependent constant.

We conclude this section by noting that finding the optimistic parameter $\widetilde{\theta}_t$ is a computationally expensive task as $\theta :\to \text{Tr}P(\theta)$ is a non-convex function, while computing the optimal control $K(\theta)$ is cheap thanks to the Riccati equation solver. As a result, a TS algorithm would be much more efficient, as the optimistic step would be replaced by a sampling from the posterior. Abbasi-Yadkori and Szepesvári (2015) proposed a lazy PSRL algorithm for smoothly parametrized MDP, for which LQ is a specific instance, that they study in the Bayesian regret setting. Unfortunately, as hinted by Osband and Van Roy (2016), the proof suffers a flaw coming from the difficulty to move from episodic to non-episodic settings. Applying PSRL to LQ and providing regret bounds for this algorithm is one of main questions that motivates our work. We address it in Ch. 4.

Chapter 3

Thompson Sampling in Linear Bandit

In this chapter[1] , we derive an alternative proof for the regret of Thompson sampling (TS) in the stochastic linear bandit setting. While we obtain a regret bound of order $\widetilde{O}(d^{3/2}\sqrt{T})$ as in previous results, the proof sheds new light on the functioning of the TS. We leverage the structure of the problem to show how the regret is related to the sensitivity (i.e., the gradient) of the objective function and how selecting optimal arms associated to *optimistic* parameters does control it. Thus, we show that TS can be seen as a generic randomized algorithm where the sampling distribution is designed to have a fixed probability of being optimistic, at the cost of an additional $\sqrt{d}$ regret factor compared to a UCB-like approach. Furthermore, we show that our proof can be readily applied to regularized linear optimization and generalized linear model problems.

Contents

[1] This chapter is an extended version of our AI&Stats paper (Abeille and Lazaric, 2017a), which has been accepted for publication in the Electronic Journal of Statistics.

3.1 Introduction

In this chapter, we focus on the Linear Bandit (LB) problem introduced in Sec. 2.2 and draw novel insights on the functioning of Thompson Sampling (TS) in this setting, where the value of an arm is obtained as the inner product between an arm feature vector x and an unknown global parameter $\theta^\star$. While TS has been originally introduced as a Bayesian heuristic (Thompson, 1933), it has been proved to offer good performance in the *frequentist* setting (Agrawal and Goyal, 2012b). We propose here an alternative frequentist analysis for the regret of TS in LB that stresses the *randomized* nature of the exploration performed by the algorithm. As opposed to the *optimistic* approaches, the main technical difficulty in analyzing TS lies in controlling the deviation in performance due to the randomness of the algorithm. Agrawal and Goyal (2012b) follows the MAB proof structure (as in (Agrawal and Goyal, 2012a)) classifying arms as saturated and unsaturated depending on whether their standard deviation is smaller or bigger than their gap to the optimal arm.[2] While for unsaturated arms the regret is related to their standard deviation that decreases over time, they prove that TS has a small (but constant) probability to select saturated arms and it achieves a regret $\widetilde{O}\big(d^{3/2}\sqrt{T}\big)$.

Contributions. The major contributions of this paper are: **1)** Following the intuition of Agrawal and Goyal (2012b), we show that the TS does not need to sample from an actual Bayesian posterior distribution and that any distribution satisfying suitable concentration and anti-concentration properties guarantees a small regret. In particular, we show that the distribution should *over-sample* w.r.t. the standard least-squares confidence ellipsoid by a factor $\sqrt{d}$ to guarantee a constant probability of being optimistic. **2)** We provide an alternative proof of TS achieving the same result as Agrawal and Goyal (2012b). One of our major finding is that, leveraging the properties of support functions from convex geometry, we are able to prove that the regret is related to the gradient of the objective function, that is ultimately controlled by the norm of the optimal arms associated to any optimistic parameter θ. This provides a novel insight on the fact that whenever an optimistic parameter θ_t is chosen, not only is its instantaneous regret small but the corresponding optimal arm $x_t = \arg\max_x x^\mathsf{T}\theta_t$ represents a *useful exploration* step that improves the accuracy of the estimation of $\theta^\star$ over dimensions which are relevant to reduce regret in any subsequent non-optimistic step. This approach allows us to avoid the introduction of saturated/unsaturated arms and it illustrates why any TS-like algorithm (not necessarily Bayesian) with a constant probability of being optimistic has a bounded regret. **3**) Finally, we show how our proof can be easily adapted to regularized linear optimization (with arbitrary penalty) and to the generalized linear model, for which we derive the first frequentist regret bound for TS, which was first suggested by Agrawal and Goyal (2012b) as a venue to explore.

[2]Here we refer to the definition introduced in the *arXiv* paper, which slightly differs from the original ICML paper.

3.2 Preliminaries

The setting. We briefly recall the LB setting introduced in Sec. 2.2, and detail the assumptions that we impose on the problem structure as well as the additional material needed for our analysis.
Let $\mathcal{X} \subset \mathbb{R}^d$ be an arbitrary (finite or infinite) set of arms. When an arm $x \in \mathcal{X}$ is pulled, a reward is generated as $r(x) = x^\mathsf{T}\theta^\star + \xi$, where $\theta^\star \in \mathbb{R}^d$ is a fixed but unknown parameter and ξ is a zero-mean noise. An arm $x \in \mathcal{X}$ is evaluated according to its expected reward $x^\mathsf{T}\theta^\star$ and, for any $\theta \in \mathbb{R}^d$, we denote the optimal arm and its value by

$$x^\star(\theta) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta, \qquad J(\theta) = \sup_{x\in\mathcal{X}} x^\mathsf{T}\theta. \tag{3.1}$$

At each step t, the learner selects an arm $x_t \in \mathcal{X}$ based on the past observations (and possibly additional randomization), it observes the reward $r_{t+1} = x_t^\mathsf{T}\theta^\star + \xi_{t+1}$, and it suffers a *regret* equal to the difference in expected reward between the optimal arm $x^\star(\theta^\star)$ and the arm x_t. The objective of the learner is to minimize the *cumulative regret* up to step T, i.e.,

$$R(T) = \sum_{t=1}^{T} \left(x^{\star\mathsf{T}}\theta^\star - x_t^\mathsf{T}\theta^\star\right).$$

Notation. We use $\|\cdot\|$ to denote the 2-norm and x^T to denote the transpose of $x \in \mathbb{R}^d$. For a positive definite matrix $M \in \mathbb{R}^{d\times d}$, we denote as $\|\cdot\|_M$ the weighted 2-norm defined by $\|x\|_M^2 = x^\mathsf{T}Mx$ for any $x \in \mathbb{R}^d$. We use $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ to denote the minimum and maximum eigenvalues of the positive semi-definite matrix M, respectively. We use $\mathbb{1}\{E\}$ to denote the indicator function of the event E. Finally, we encode all the information observed up to time t in the filtration $\mathcal{F}_t^x = (\mathcal{F}_1, \sigma(x_1, r_2, \ldots, r_t, x_t))$, where $\mathcal{F}_1$ contains any prior knowledge.

We impose the following assumptions on the problem structure and the noise ξ_{t+1}.

Assumption 3.2.1 (Arm set)**.** *The arm set $\mathcal{X}$ is a bounded closed (and hence compact) subset of $\mathbb{R}^d$ such that $\|x\| \le X$ for all $x \in \mathcal{X}$. We also assume $X = 1$.*

We focus here on the fixed arm set setting where $\mathcal{X}$ does not change with time while the original analysis of Agrawal and Goyal (2012b) has been derived for the *contextual* LB problem where the arm set $\mathcal{X}_t$ can be chosen by an oblivious adversary at each time step t. However, our analysis still holds in this case, replacing $\mathcal{X}$ by $\mathcal{X}_t$ and the optimal value function J by J_t in every steps of the proof.

Assumption 3.2.2 (Bandit parameter)**.** *There exists $S \in \mathbb{R}^+$ such that $\|\theta^\star\| \le S$ and S is known.*

Assumption 3.2.3 (Noise)**.** *The noise process $\{\xi_t\}_t$ is a martingale difference sequence given $\mathcal{F}_t^x$ and it is conditionally R-subgaussian for some constant $R \ge 0$,*

$$\forall t \ge 1,\ \mathbb{E}\left[\xi_{t+1}|\mathcal{F}_t^x\right] = 0,$$
$$\forall \alpha \in \mathbb{R},\ \mathbb{E}\left[e^{\alpha\xi_{t+1}} \mid \mathcal{F}_t^x\right] \le \exp\left(\alpha^2R^2/2\right).$$

Technical tools. Let $(x_1, \dots, x_t) \in \mathcal{X}^t$ be a sequence of arms and $(r_2, \dots, r_{t+1})$ be the corresponding rewards, then $\theta^\star$ can be estimated by regularized least-squares (RLS). For any regularization parameter $\lambda \in \mathbb{R}^+$, the design matrix and the RLS estimate are defined as

$$V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\mathsf{T}, \qquad \widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} x_s r_{s+1}. \tag{3.2}$$

For any $0 < \delta < 1$, we make use of Prop. 2.2.1 to define the high-probability ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$ at each time step t as

$$\mathcal{E}_t^{\mathrm{RLS}} = \left\{ \theta \in \mathbb{R}^d \;\middle|\; \|\theta - \widehat{\theta}_t\|_{V_t} \le \beta_t(\delta') \right\}, \text{ with } \beta_t(\delta') = R\sqrt{2\log\frac{(1+t/\lambda)^{d/2}}{\delta'}} + \sqrt{\lambda} S, \tag{3.3}$$

that is centered around $\widehat{\theta}_t$ with orientation defined by V_t and radius $\beta_t(\delta')$, where $\delta' = \delta/4T$. Under Asm. 3.2.1, 3.2.2, and 3.2.3, Prop. 2.2.1 guarantees that $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$ for all $t \le T$, with probability at least $1 - \delta/4$.
Finally, we report a standard result of RLS that, together with Prop. 2.2.1, shows that the prediction error on the x_t's used to construct the estimator $\widehat{\theta}_t$ is cumulatively small.

Proposition 3.2.1. *Let $\lambda \ge 1$, for any arbitrary sequence $(x_1, x_2, \dots, x_t) \in \mathcal{X}^t$ let V_{t+1} be the corresponding design matrix (Eq. 3.2), then*

$$\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \le 2\log\frac{\det(V_{t+1})}{\det(\lambda I)} \le 2d\log\left(1 + \frac{t}{\lambda}\right).$$

This result plays a central role in most of the proofs for linear bandit, since the regret is usually related to $\|x_s\|_{V_s^{-1}}$ and Prop. 3.2.1 is used to bound its cumulative sum. While Agrawal and Goyal (2012b) achieve this by dividing arms in saturated and unsaturated, we follow a different path that leverages the core features of the problem (structure of $J(\theta)$) and of TS (probability of being optimistic).

3.3 Linear Thompson sampling

Agrawal and Goyal (2012b) define TS for linear bandit as a Bayesian algorithm where a Gaussian prior over $\theta^\star$ is updated according to the observed rewards, a random sample is drawn from the posterior, and the corresponding optimal arm is selected at each step. As hinted by Agrawal and Goyal (2012b), we show that TS can be defined as a generic randomized algorithm constructed on the RLS-estimate rather than an algorithm sampling from a Bayesian posterior (see Fig. 3.1). At any step t, given RLS-estimate $\widehat{\theta}_t$ and the design matrix V_t, TS samples a *perturbed* parameter $\widetilde{\theta}_t$ as

$$\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta') V_t^{-1/2} \eta_t,$$

where η_t is a random sample drawn i.i.d. from a suitable multivariate distribution $\mathcal{D}^{\mathrm{TS}}$, which does not need to be associated with an actual posterior over $\theta^\star$. Then the optimal arm $x_t = x^\star(\widetilde{\theta}_t)$ is chosen, a reward r_{t+1} is observed and V_t and $\widehat{\theta}_t$ are updated

according to Eq. 3.2. Notice that the resulting distribution on $\widetilde{\theta}_t$ is obtained by rotation of η_t according to the design matrix V_t and by a rescaling $\beta_t(\delta')$. The computational complexity of TS is dominated by computation of $x^\star(\widetilde{\theta}_t)$, which requires solving a linear optimization problem and by the sampling process from $\mathcal{D}^{\mathrm{TS}}$. This is in contrast with OFUL (Abbasi-Yadkori et al., 2011a) presented in Fig. 2.6, which requires solving a bilinear optimization problem (i.e., $\arg\max_\theta \max_x x^\mathsf{T}\theta$).

Input: $\widehat{\theta}_1$, $V_1 = \lambda I$, δ, T
1: Set $\delta' = \delta/(4T)$
2: **for** $t = \{1, \dots, T\}$ **do**
3: Sample $\eta_t \sim \mathcal{D}^{\mathrm{TS}}$ and compute parameter $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t$
4: Compute optimal arm $x_t = x^\star(\widetilde{\theta}_t) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\widetilde{\theta}_t$
5: Pull arm x_t and observe reward r_{t+1}
6: Compute V_{t+1} and $\widehat{\theta}_{t+1}$ using Eq. 3.2
7: **end for**

Figure 3.1 – Thompson sampling algorithm for LB.

The key aspect to ensure small regret is that the perturbation η_t is distributed so that TS explores *enough* but *not too much*. This translates into the following conditions on $\mathcal{D}^{\mathrm{TS}}$.

Definition 3.3.1. *$\mathcal{D}^{\mathrm{TS}}$ is a multivariate distribution on $\mathbb{R}^d$ absolutely continuous with respect to the Lebesgue measure which satisfies the following properties:*

1. *(anti-concentration) there exists a strictly positive probability p such that for any $u \in \mathbb{R}^d$ with $\|u\| = 1$,*
$$\mathbb{P}_{\eta\sim\mathcal{D}^{\mathrm{TS}}}\left(u^\mathsf{T}\eta \geq 1\right) \geq p,$$
2. *(concentration) there exists c, c' positive constants such that $\forall\delta \in (0,1)$*
$$\mathbb{P}_{\eta\sim\mathcal{D}^{\mathrm{TS}}}\left(\|\eta\| \leq \sqrt{cd\log\frac{c'd}{\delta}}\right) \geq 1-\delta.$$

Once interpreted in the construction of $\widetilde{\theta}_t$, the definition of $\mathcal{D}^{\mathrm{TS}}$ basically requires TS to explore far enough from $\widehat{\theta}_t$ (anti-concentration) but not too much (concentration). This implies that TS performs "useful" exploration with enough frequency (notably it performs optimistic steps), but without selecting arms with too large regret. We introduce the high-probability ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ as

$$\mathcal{E}_t^{\mathrm{TS}} = \{\theta \in \mathbb{R}^d \mid \|\theta - \widehat{\theta}_t\|_{V_t} \leq \gamma_t(\delta')\}, \text{ with } \gamma_t(\delta') = \beta_t(\delta')\sqrt{cd\log(c'd/\delta')}.$$

The difference between $\mathcal{E}_t^{\mathrm{RLS}}$ and $\mathcal{E}_t^{\mathrm{TS}}$ lies in the additional factor $\sqrt{d}$ in the definition of $\gamma_t(\delta')$ and it is crucial for both concentration and anti-concentration to hold at the same time. In Sect. 3.5 we prove that any distribution satisfying the conditions in Def. 3.3.1 introduces the right amount of randomness to achieve the desired regret

without actually satisfying any Bayesian assumption. Def. 3.3.1 includes the Gaussian prior used by Agrawal and Goyal (2012b), but also other types of distributions such as the uniform on the unit ball $\mathcal{B}_d(0, \sqrt{d})$ or distributions concentrated on the boundary of $\mathcal{E}_t^{\mathrm{TS}}$ (refer to App. 3.A for exact values of c, c', and p for uniform and Gaussian distributions).

3.4 Sketch of the proof

In this section we report a sketch of the proof providing a geometric intuition on the behavior of TS and how its actions (i.e., the sampled $\widetilde{\theta}_t$ and the corresponding x_t) influence the regret. For the sake of illustration, we consider the unit ball $\mathcal{X} = \{\|x\| \leq 1\}$, such that the optimal arm is just the projection of θ on the ball ($x^\star(\theta) = \theta/\|\theta\|$), and the optimal value is $J(\theta) = \theta^\mathsf{T}\theta/\|\theta\| = \|\theta\|$. We start by decomposing the regret using the definition of $J(\theta)$ as

$$R(T) = \sum_{t=1}^{T} \left(x^{\star\mathsf{T}}\theta^\star - x_t^\mathsf{T}\widetilde{\theta}_t \right) + \left(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \right) = \underbrace{\sum_{t=1}^{T} \left(J(\theta^\star) - J(\widetilde{\theta}_t) \right)}_{R^{\mathrm{TS}}(T)} + \underbrace{\sum_{t=1}^{T} \left(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \right)}_{R^{\mathrm{RLS}}(T)},$$

where R^{TS} depends on the randomization of TS and R^{RLS} mostly depends on the properties of RLS.

3.4.1 Bounding $R^{\mathrm{RLS}}(T)$.

We first show that both RLS estimate $\widehat{\theta}_t$ and TS parameter $\widetilde{\theta}_t$ should concentrate appropriately, by decomposing the regret $R^{\mathrm{RLS}}(T)$ as

$$R^{\mathrm{RLS}}(T) = \sum_{t=1}^{T} \left(x_t^\mathsf{T}\widehat{\theta}_t - x_t^\mathsf{T}\theta^\star \right) + \sum_{t=1}^{T} \left(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\widehat{\theta}_t \right).$$

Since at each step t, $\widetilde{\theta}_t$ is sampled from $\mathcal{D}^{\mathrm{TS}}$, the second term is kept under control by construction, while the first sum deals with the prediction error of RLS. As opposed to R^{TS}, this error is not related to the exploration scheme and it is small for any sequence of arms. Intuitively, this is due to the fact that the RLS estimate is the minimizer of the regularized cumulative squared error $\widehat{\theta}_{T+1} = \arg\min_\theta \left(\sum_{t=1}^{T} |r_{t+1} - x_t^\mathsf{T}\theta|^2 + \lambda\|\theta\|^2 \right)$, so that $x_t^\mathsf{T}\widehat{\theta}_{T+1}$ is an accurate prediction *on the arms observed so far*. The RLS minimizes the error in "hindsight" (i.e., after observing all rewards up to T) and therefore it also

controls the *online* error $|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_{t+1}|^2$. By induction,

$$\begin{aligned}
\sum_{t=1}^{T} |r_{t+1} - x_t^\mathsf{T}\widehat{\theta}_{T+1}|^2 + \lambda\|\widehat{\theta}_{T+1}\|^2 &= |r_{T+1} - x_T^\mathsf{T}\widehat{\theta}_{T+1}|^2 + \sum_{t=1}^{T-1} |r_{t+1} - x_t^\mathsf{T}\widehat{\theta}_{T+1}|^2 + \lambda\|\widehat{\theta}_{T+1}\|^2 \\
&\geq |r_{T+1} - x_T^\mathsf{T}\widehat{\theta}_{T+1}|^2 + \min_{\theta} \left(\sum_{t=1}^{T-1} |r_{t+1} - x_t^\mathsf{T}\widehat{\theta}|^2 + \lambda\|\widehat{\theta}\|^2 \right) \\
&= |r_{T+1} - x_T^\mathsf{T}\widehat{\theta}_{T+1}|^2 + \sum_{t=1}^{T-1} |r_{t+1} - x_t^\mathsf{T}\widehat{\theta}_{T}|^2 + \lambda\|\widehat{\theta}_{T}\|^2 \\
&\geq \cdots \geq \sum_{t=1}^{T} |r_{t+1} - x_t^\mathsf{T}\widehat{\theta}_{t+1}|^2 + \lambda\|\widehat{\theta}_1\|^2.
\end{aligned}$$

Having a small *online* error also implies a small *prediction* error $|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_t|^2$. In fact, using a recursive version of Eq. 3.2, we have

$$\hat{\theta}_{t+1} = \hat{\theta}_t + V_t^{-1}x_t(1 + \|x_t\|^2_{V_t^{-1}})^{-1}(r_{t+1} - x_t^\mathsf{T}\hat{\theta}_t),$$

which, together with $\|x_t\|^2_{V_t^{-1}} \leq 1/\lambda$, leads to

$$|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_{t+1}| \geq \frac{\lambda}{1+\lambda}|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_t|.$$

Since the cumulative prediction error is small, then the associated regret $\sum_{t=1}^{T} |x_t^\mathsf{T}\widehat{\theta}_t - x_t^\mathsf{T}\theta^\star|$ is also small. This result can be seen as an intrinsic *on-policy* error guarantee of RLS. Nonetheless, notice that while RLS minimizes the prediction error for any sequence of arms, this does not imply the consistency of the estimator. For instance, when the same arm x is repeatedly played, the unknown parameter $\theta^\star$ is well-estimated in the direction of x (thus making $R^{\mathrm{RLS}}(T)$ small) but it is poorly estimated in any other directions. This shows the need for a careful exploration strategy to recover consistency and hence a sub-linear regret.

3.4.2 Bounding $R^{\mathbf{TS}}(T)$.

We denote by $R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t)$ each term in $R^{\mathrm{TS}}(T)$. For optimistic algorithms this term is bounded by 0 at any step since w.h.p. $J(\widetilde{\theta}_t) \geq J(\theta^*)$ by construction. In the Bayesian regret analysis of TS, this term is equal to 0 by assumption that θ^* is drawn from the same prior as $\widetilde{\theta}_t$. On the other hand, in the frequentist analysis, we have to control the deviations caused by the random sampling of $\widetilde{\theta}_t$. This is achieved by showing that the arms selected by TS provide "useful" information about $\theta^\star$ and contribute to keep the regret small. We follow three steps: **1)** we show that the regret is related to the sensitivity of J w.r.t. the errors in estimating $\theta^\star$ and we bound the regret with the gradient of $J(\theta)$ at any *optimistic* θ; **2)** we show how the gradient in a point θ is intrinsically related to its corresponding optimal arm $x^\star(\theta)$; **3)** since we prove that TS is frequently optimistic, then we can finally link $x^\star(\theta)$ to $x_t = x^\star(\widetilde{\theta}_t)$ and Prop. 3.2.1 allows us to finally bound the overall regret.

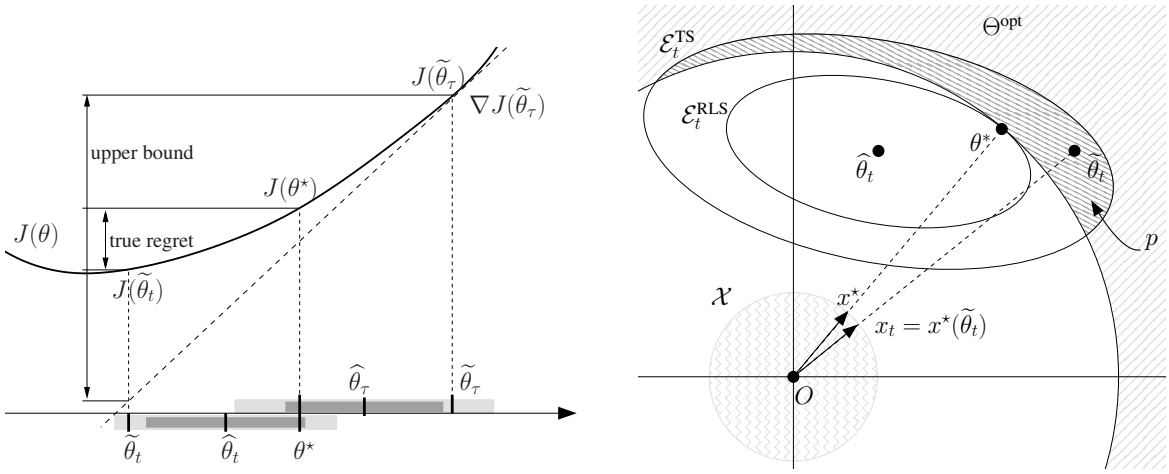

Figure 3.2 – Illustration of the steps **2)** and **3)** of the proof in $\mathbb{R}^1$ and $\mathbb{R}^2$. *Left:* The regret at step t could be bounded by the gradient of the function J at a previous optimistic $\widetilde{\theta}_\tau$ times the distance between $\widetilde{\theta}_\tau$ and the current $\widetilde{\theta}_t$. Notice that $\theta^\star$ is always included in $\mathcal{E}_t^{\mathrm{RLS}}$ (in dark gray) and thus $\widetilde{\theta}$s sampled from $\mathcal{E}_t^{\mathrm{TS}}$ (in light gray) are never too far. *Right:* TS has a constant probability of being optimistic thanks to the over-sampling of $\mathcal{D}^{\mathrm{TS}}$.

Step 1 (regret and sensitivity of J). We first show why the exploration of TS should be *well adapted* to $J(\theta)$. Using the definition of $J(\theta) = \|\theta\|$ we have

$$R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t) = \|\theta^\star\| - \|\widetilde{\theta}_t\| \leq \|\theta^\star - \widetilde{\theta}_t\| \leq \frac{\|\theta^\star - \widetilde{\theta}_t\|_{V_t}}{\sqrt{\lambda_{\min,t}}},$$

where $\lambda_{\min,t}$ is the smallest eigenvalue of V_t. This bound shows that it is sufficient to estimate $\theta^\star$ accurately over all its components (i.e., $\lambda_{\min,t}$ tends to infinity) to obtain a no-regret algorithm. Nonetheless, the desired regret bound of $O(\sqrt{T})$ is obtained only if $\lambda_{\min,t}$ increases as $O(t)$. While this could be achieved by a fully explorative algorithm (e.g., a round robin over the canonic vectors e_i reduces the ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ to a ball of radius $\lambda_{\min,t}$), it would severely increase the second term of $R^{\mathrm{RLS}}(T)$ and cause an overall linear regret[3]. Fortunately, inspecting the definition of R_t^{TS} reveals that not all components of $\theta^\star$ must be equally well estimated. In fact, we have w.h.p. that

$$R_t^{\mathrm{TS}} \leq \sup_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \Big(J(\theta) - J(\theta') \Big).$$

This shows that R_t^{TS} is determined by the *diameter* of ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ w.r.t. J, which suggests that the estimation of $\theta^\star$ should be more accurate on the dimensions on which J is more sensitive. In the case of $\mathcal{X}$ unit ball, the most sensitive direction of J is $\theta^\star/\|\theta^\star\|$ itself and Fig. 3.3 illustrates two opposite cases where the accuracy in the estimation of $\theta^\star$ is the same (i.e., V_t has the same eigenvalues) but the regret may be very different.

[3]This happens because x_t would be optimal w.r.t. a $\widetilde{\theta}_t$, which is *not* in the ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$.

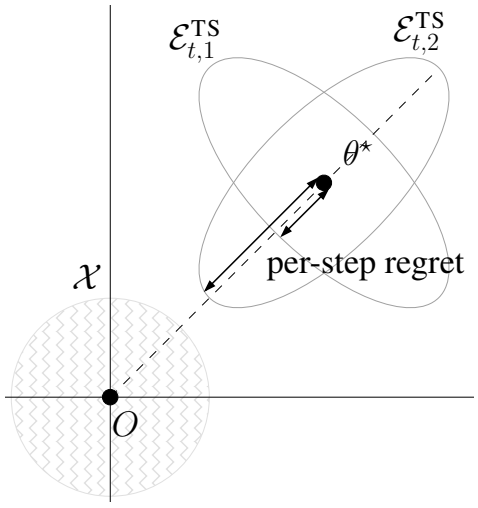

Figure 3.3 – While $\mathcal{E}_{t,1}^{\mathrm{TS}}$ and $\mathcal{E}_{t,2}^{\mathrm{TS}}$ have an equivalent accurate estimation of $\theta^\star$, $\mathcal{E}_{t,1}^{\mathrm{TS}}$ has smaller regret than $\mathcal{E}_{t,2}^{\mathrm{TS}}$.

The numerical experiment reported in Fig. 3.4 supports the fact that TS handles correctly the consistency of the estimates w.r.t. the *sensitivity* of J. In the 2-dimensional case, the eigenvalues $\lambda_{\max,t}$ and $\lambda_{\min,t}$ exhibit different divergence rate. While $1/\lambda_{\max,t}$ decreases as $1/t$, $1/\lambda_{\min,t}$ decreases as $1/\sqrt{t}$, which according to a direct consistency argument, is not enough to guarantee a $\sqrt{T}$ regret. However, the picture on the r.h.s shows that the diameter of the ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ w.r.t J decreases as $1/\sqrt{t}$. This is due to the fact that the sampling ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ tends to align with the first configuration of Fig. 3.3. It implies that, on direction where J is *very sensitive*, the diameter of the ellipsoid shrinks appropriately (scaling with $1/\lambda_{\max,t} \approx 1/t$) whereas on direction where J is *less sensitive*, a slower decay (scaling with $1/\lambda_{\min,t} \approx 1/\sqrt{t}$) of the diameter of the ellipsoid still guarantees the deviation in J to be small. Therefore, the overall deviation in any direction decreases as $1/\sqrt{t}$, inducing a $\sqrt{T}$ regret.

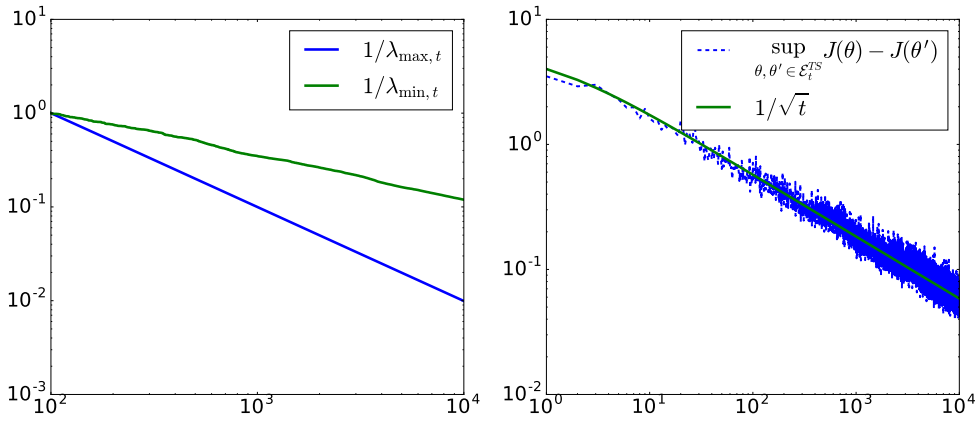

Figure 3.4 – Numerical illustration on how TS adapts the eigenvalues rate of divergence with the sensitivity of J. While $\lambda_{\min,t}$ and $\lambda_{\max,t}$ have different divergence rate, the diameter of the ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ shrinks fast, which stresses that $\lambda_{\min,t}$ has been associated with the less sensitive direction while $\lambda_{\max,t}$ has been associated with the most sensitive direction. *Left:* loglog plot of the inverse of the eigenvalues of the design matrix V_t w.r.t. t. Rates of convergence are $1/t$ and $1/\sqrt{t}$. *Right:* loglog plot of the deviation in J over $\mathcal{E}_t^{\mathrm{TS}}$ w.r.t. t (blue dashed line). The rate of convergence is $1/\sqrt{t}$ (green line).

Let $\Theta^{\mathrm{opt}} = \{\theta : J(\theta) \geq J(\theta^\star)\}$ be the set of optimistic parameters. In our example $J(\theta) = \|\theta\|$ is convex thus we can make explicit the dependency of the regret on the sensitivity of J through its gradient evaluated at any $\theta \in \Theta^{\mathrm{opt}}$ as (see Prop. 3.5.1 for the general case)

$$R_t^{\mathrm{TS}} \leq \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} J(\theta) - J(\theta') \leq \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \nabla J(\theta)^\mathsf{T}(\theta - \theta'),$$

which shows that the regret of non-optimistic $\widetilde{\theta}_t$ is bounded by the gradient of $J(\theta)$ at any optimistic θ and its distance to any other point in the TS ellipsoid.

Step 2 (sensitivity of J and optimal arm). According to Prop. 2.2.1, the difference $\theta - \theta'$ in the previous inequality is well controlled whenever θ belongs to the ellipsoid, while the first term cannot be immediately controlled by the algorithm. Nonetheless, we notice that since $J(\theta) = \|\theta\|$, then $\nabla J(\theta) = \theta/\|\theta\| = x^\star(\theta)$ (see Lem. 3.5.2 for the general case). This shows how selecting the optimal arm associated to an optimistic θ is equivalent to controlling the gradient of J, which results in

$$R_t^{\mathrm{TS}} \leq \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} x^\star(\theta)^\mathsf{T}(\theta - \theta').$$

From Prop. 3.2.1, we could conclude that the regret would be cumulatively small if $x^\star(\theta)$ corresponded to the arms chosen by the TS ($x_t = x^\star(\widetilde{\theta}_t)$). As a result, we need a θ **1)** that is optimistic (i.e., $\theta \in \Theta^{\mathrm{opt}}$), **2)** it belongs or is close to the ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ and **3)** it is used to select an arm x_t. The first two requirements are at the core of the choice of the TS distribution in Def. 3.3.1 where the anti-concentration property guarantees enough probability to be optimistic, while the concentration property implies that $\widetilde{\theta}$s are within a small ellipsoid. Let $\tau < t$ be any step when TS selects $\widetilde{\theta}_\tau \in \Theta^{\mathrm{opt}}$ with corresponding arm $x_\tau = x^\star(\widetilde{\theta}_\tau)$, then we have (see an illustration of this bound in Fig. 3.2 in the 1-d case)

$$R_t^{\mathrm{TS}} \leq \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} x_\tau^\mathsf{T}(\widetilde{\theta}_\tau - \theta') \leq \|x_\tau\|_{V_\tau^{-1}} \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta}_\tau - \theta'\|_{V_\tau}.$$

Introducing $\theta^\star$ and using the fact that the design matrices forms a non-decreasing sequence (i.e., $V_\tau \preccurlyeq V_t$), we decompose

$$\begin{aligned}\sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta}_\tau - \theta'\|_{V_\tau} &\leq \|\widetilde{\theta}_\tau - \theta^\star\|_{V_\tau} + \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\theta^\star - \theta'\|_{V_\tau}, \\ &\leq \|\widetilde{\theta}_\tau - \theta^\star\|_{V_\tau} + \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\theta^\star - \theta'\|_{V_t}.\end{aligned}$$

Since by Prop. 2.2.1, $\theta^\star$ is contained in all ellipsoids $\mathcal{E}_t^{\mathrm{RLS}}$ with high probability, then

$$\begin{aligned}R_t^{\mathrm{TS}} &\leq \Big(\beta_\tau(\delta') + \gamma_\tau(\delta') + \beta_t(\delta') + \gamma_t(\delta')\Big)\|x_\tau\|_{V_\tau^{-1}}, \\ &\leq \Big(2\beta_T(\delta') + 2\gamma_T(\delta')\Big)\|x_\tau\|_{V_\tau^{-1}}.\end{aligned}$$

Let K be the number of times $\widetilde{\theta}_t \in \Theta^{\text{opt}}$, t_k the corresponding steps, and $\nu_k = t_k - t_{k-1}$, then the final regret can be written as

$$R^{\text{TS}}(T) \leq 2\Big(\beta_T(\delta') + \gamma_T(\delta')\Big) \sum_{k=1}^{K} \nu_k \|x_{t_k}\|_{V_{t_k}^{-1}}.$$

Step 3 (optimism). This bound shows the importance that TS is optimistic with high frequency. In fact, whenever $\widetilde{\theta}_t$ is in Θ^{opt}, not only the corresponding instantaneous regret R_t^{TS} is upper-bounded by 0, but the exploration performed by playing arm $x^\star(\widetilde{\theta}_t)$ has also a positive impact in controlling the regret for any subsequent non-optimistic step. Consider the extreme case when TS is never optimistic. Then, $K = 1$, $\nu_1 = T$ and $R^{\text{TS}}(T) = O(T)$. On the other hand, if TS is optimistic with a constant frequency, then we can easily show that $R^{\text{TS}}(T)$ is bounded by $\widetilde{O}(\sqrt{T})$. Consider the case where an optimistic θ is chosen with probability p. Since $\mathbb{E}[\nu_k] = 1/p$, we can prove that w.h.p. $R^{\text{TS}}(T) \leq \widetilde{O}(1/p\sqrt{T})$ by Cauchy-Schwarz and Prop. 3.2.1 applied to $\sum_{k=1}^{K} \|x_{t_k}\|^2_{V_{t_k}^{-1}}$, where $K \approx T$. Unfortunately, sampling $\widetilde{\theta}_t$ from the RLS ellipsoid $\mathcal{E}_t^{\text{RLS}}$ may have a very small probability of being optimistic (see e.g., Fig. 3.2, where sampling uniformly in $\mathcal{E}_t^{\text{RLS}}$ has zero probability to return a $\widetilde{\theta}_t \in \Theta^{\text{opt}}$). For this reason, TS is required to draw $\widetilde{\theta}_t$ from a distribution *over-sampling* by a factor $\sqrt{d}$ w.r.t. $\mathcal{E}_t^{\text{RLS}}$ as in the definition of $\mathcal{D}^{\text{TS}}$. This guarantees a fixed probability p of being optimistic (see Lem. 3.5.3) and the final desired regret.

3.5 Formal proof

In this section we report the main steps of the regret analysis, while we postpone technical lemmas to the supplementary material. We prove the following result.

Theorem 3.5.1. *Under assumptions 3.2.1,3.2.2,3.2.3, the regret of* TS *is bounded w.p.* $1 - \delta$ *as*

$$R(T) \leq \Big(\beta_T(\delta') + \gamma_T(\delta')(1 + 4/p)\Big)\sqrt{2Td\log\Big(1 + \frac{T}{\lambda}\Big)} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}},$$

where $\delta' = \frac{\delta}{4T}$.

As anticipated in introduction, this bound is of order $\widetilde{O}(d^{3/2}\sqrt{T})$ and it entirely matches the result of Agrawal and Goyal (2012b). The analysis of the regret requires extra care in the definition of the filtrations. While in analyzing R^{RLS} we consider all the knowledge up to step t (i.e., including the sampled parameter $\widetilde{\theta}_t$), in R^{TS} we need to study the randomness of $\widetilde{\theta}_t$ conditional on all the information before sampling η_t. We introduce an additional filtration besides $\mathcal{F}_t^x$.

Definition 3.5.1. *We define the filtration* $\mathcal{F}_t$ *as the accumulated information up to time* t *before the sampling procedure, i.e.,* $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, r_2, \dots, x_{t-1}, r_t))$.

Notice that $\widehat{\theta}_t$ and V_t^{-1} are both $\mathcal{F}_t$ and $\mathcal{F}_t^x$ adapted, while $\widetilde{\theta}_t$ is a random variable w.r.t. $\mathcal{F}_t$ and it is fixed when considering $\mathcal{F}_t^x$. Hence we have $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_2^x \subset \mathcal{F}_3 \subset \mathcal{F}_3^x, \dots$. We are now ready to introduce the high-probability events we use in the rest of the proof.

Definition 3.5.2. *Let $\delta \in (0,1)$ and $\delta' = \delta/(4T)$ and $t \in [1,T]$. We define $\widehat{E}_t$ as the event where the RLS estimate concentrates around $\theta^\star$ for all steps $s \le t$, i.e.,*

$$\widehat{E}_t = \Big\{\forall s \le t,\ \ \|\widehat{\theta}_s - \theta^\star\|_{V_s} \le \beta_s(\delta')\Big\}.$$

We also define $\widetilde{E}_t$ as the event where the sampled parameter $\widetilde{\theta}_s$ concentrates around $\widehat{\theta}_s$ for all steps $s \le t$, i.e.,

$$\widetilde{E}_t = \Big\{\forall s \le t,\ \ \|\widetilde{\theta}_s - \widehat{\theta}_s\|_{V_s} \le \gamma_s(\delta')\Big\}.$$

Then we have that $\widehat{E} := \widehat{E}_T \subset \cdots \subset \widehat{E}_1$, $\widetilde{E} := \widetilde{E}_T \subset \cdots \subset \widetilde{E}_1$ and we use $E_t = \widehat{E}_t \cap \widetilde{E}_t$ and $E = \widehat{E} \cap \widetilde{E}$.

Lemma 3.5.1. *[see proof in App. 3.D] Under Asm. 3.2.2, 3.2.3 we have*

$$\mathbb{P}(\widehat{E} \cap \widetilde{E}) \ge 1 - \frac{\delta}{2}.$$

Conditioned on $\mathcal{F}_t$ and event $\widehat{E}_t$, we have $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$, while on event $\widetilde{E}_t$ we have $\widetilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}$, then we directly bound the regret on event E as

$$\begin{aligned}
R(T) = R(T)\mathbb{1}\{E\} &= \sum_{t=1}^{T}\Big(J(\theta^\star) - J(\widetilde{\theta}_t)\Big)\mathbb{1}\{E\} + \sum_{t=1}^{T}\Big(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star\Big)\mathbb{1}\{E\} \\
&\le \sum_{t=1}^{T}\Big(J(\theta^\star) - J(\widetilde{\theta}_t)\Big)\mathbb{1}\{E_t\} + \sum_{t=1}^{T}\Big(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star\Big)\mathbb{1}\{E_t\} \\
&= \sum_{t=1}^{T} R_t^{\mathrm{TS}} + \sum_{t=1}^{T} R_t^{\mathrm{RLS}},
\end{aligned}$$

w.p. $1 - \delta/2$. Notice that the formal definitions of R_t^{TS} and R_t^{RLS} involve the indicator of the high-probability events $\mathbb{1}\{E_t\}$, which is the quantity of interest since we aim at bounding the regret on E. As discussed in Sec. 3.4, the main difficulty lies in bounding the regret term specific to TS. We first report the formal proof to bound $R^{\mathrm{TS}}(T)$, while the bound on $R^{\mathrm{RLS}}(T)$ and the overall regret is postponed to Sec. 3.5.4.

Similar to the sketch in Sect. 3.4, the proof follows three steps: **1)** we use the convexity of J to upper-bound the regret by its expectation conditioned on being optimistic and to relate it to the gradient of J, **2)** we relate the gradient of J to the arms chosen by TS over time, **3)** we show that despite the randomization, TS has a constant probability of being optimistic.

3.5.1 Step 1 (regret and gradient of $J(\theta)$).

On event E_t, $\widetilde{\theta}_t$ belongs to $\mathcal{E}_t^{\mathrm{TS}}$ and thus,

$$R_t^{\mathrm{TS}} \leq \Big(J(\theta^\star) - \inf_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} J(\theta)\Big)\mathbb{1}\{\widehat{E}_t\}.$$

Recalling that Θ^{opt} is the set of all optimistic θs, we can bound the previous expression by the expectation over any random choice of $\widetilde{\theta}$ in $\Theta_t^{\mathrm{opt}} := \Theta^{\mathrm{opt}} \cap \mathcal{E}_t^{\mathrm{TS}}$ where we restrict the optimistic set to the high-probability sampling ellipsoid, that is

$$R_t^{\mathrm{TS}} \leq \mathbb{E}\Big[\Big(J(\widetilde{\theta}) - \inf_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} J(\theta)\Big)\mathbb{1}\{\widehat{E}_t\}\Big|\mathcal{F}_t, \widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\Big],$$

where $\widetilde{\theta} = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta$ with $\eta \sim \mathcal{D}^{\mathrm{TS}}$ is the TS sampling distribution. We now rely on the following characterization of $J(\theta)$ (see App. 3.C).

Proposition 3.5.1. *For any set of arm $\mathcal{X}$ satisfying Asm. 3.2.1, $J(\theta) = \sup_x x^\mathsf{T}\theta$ has the following properties: **1)** J is real-valued as the supremum is attained in $\mathcal{X}$, **2)** J is convex on $\mathbb{R}^d$, **3)** J is continuous with continuous first derivative except for a zero-measure set w.r.t. the Lebesgue's measure.*

These properties follow from the fact that J is the *support function* of $\mathcal{X}$ and it shows that J is convex for any arm set $\mathcal{X}$. As a result, we can directly relate R_t^{TS} to the gradient of J as

$$\begin{aligned} R_t^{\mathrm{TS}} &\leq \mathbb{E}\Big[\sup_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} \nabla J(\widetilde{\theta})^\mathsf{T}(\widetilde{\theta} - \theta)\mathbb{1}\{\widehat{E}_t\}\Big|\mathcal{F}_t, \widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\Big] \\ &\leq \mathbb{E}\Big[\|\nabla J(\widetilde{\theta})\|_{V_t^{-1}} \sup_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta} - \theta\|_{V_t}\mathbb{1}\{\widehat{E}_t\}\Big|\mathcal{F}_t, \widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\Big] \\ &\leq 2\gamma_t(\delta')\mathbb{E}\Big[\|\nabla J(\widetilde{\theta})\|_{V_t^{-1}}\Big|\mathcal{F}_t, \widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\Big]\mathbb{1}\{\widehat{E}_t\}, \end{aligned} \tag{3.4}$$

where we used Cauchy-Schwarz from line 1 to line 2, the fact that $\widetilde{\theta} \in \mathcal{E}_t^{\mathrm{TS}}$ and that $\widehat{E}_t$ is $\mathcal{F}_t$ measurable from line 2 to line 3.

3.5.2 Step 2 (from gradient of $J(\theta)$ to optimal arm $x^\star(\theta)$).

In the sketch of the proof there was a direct relationship between $\nabla J(\theta)$ and the optimal arm corresponding to θ by direct construction. In the next lemma, we show that this connection is true for any arm set $\mathcal{X}$.

Lemma 3.5.2. *Under Asm. 3.2.1, for any $\theta \in \mathbb{R}^d$, we have $\nabla J(\theta) = x^\star(\theta)$ except for a zero-measure set w.r.t. the Lebesgue's measure.*

Proof. The proof relies on the fact that, for any arm set $\mathcal{X}$, J is the support function of $\mathcal{X}$. Recalling Eq. 3.1,

$$x^\star(\theta) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta, \qquad J(\theta) = \sup_{x\in\mathcal{X}} x^\mathsf{T}\theta.$$

By Asm. 3.2.1, $x^\star(\theta)$ is well defined (since $\mathcal{X}$ is bounded) and $x^\star(\theta) \in \mathcal{X}$ (since $\mathcal{X}$ is closed). By Prop. 3.C.1, we know that J is twice differentiable almost everywhere. Thus, denoting as $\partial J(\theta)$ the sub-gradient of J in $\theta \in \mathbb{R}^d$, it guarantees that it reduces to a singleton almost everywhere. This translate in $\partial J(\theta) = \{\nabla J(\theta)\}$ except for a zero-measure set w.r.t. the Lebesgue's measure.
Thus, one just has to ensure that $x^\star(\theta) \in \partial J(\theta)$ for all $\theta \in \mathbb{R}^d$ to conclude the proof. By definition, $x^\star(\theta)^\mathsf{T}\theta = J(\theta)$ and for any $\bar{\theta} \in \mathbb{R}^d$, $J(\bar{\theta}) \geq x^\star(\theta)^\mathsf{T}\bar{\theta}$. Therefore,

$$\begin{aligned} J(\bar{\theta}) - x^\star(\theta)^\mathsf{T}\bar{\theta} \geq 0 &:= J(\theta) - x^\star(\theta)^\mathsf{T}\theta \\ J(\bar{\theta}) \geq J(\theta) + x^\star(\theta)^\mathsf{T}\left(\bar{\theta} - \theta\right), &\quad \forall \bar{\theta} \in \mathbb{R}^d \end{aligned}$$

which is the definition of the sub-gradient. □

Using Lem. 3.5.2 in Eq. 3.4, one obtains:

$$R_t^{\mathrm{TS}} \leq 2\gamma_t(\delta')\mathbb{E}\left[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}} \middle| \mathcal{F}_t, \widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\right] \mathbb{1}\{\widehat{E}_t\}. \tag{3.5}$$

This property strongly connects the exploration of TS to the actual regret. In fact, together with Prop. 3.2.1, it implies that selecting the optimal arm associated with any optimistic θ is equivalent to reducing the weighted norm of the gradient of J and ultimately the regret R_t^{TS}. This motivates the next step where we show that since TS is often optimistic, then the arm $x_t = x^\star(\widetilde{\theta}_t)$ contributes to the reduction of the regret.

3.5.3 Step 3 (optimism).

The optimism of TS is a direct consequence of the convexity of J and the fact that the distribution of η is oversampling by a factor $\sqrt{d}$ w.r.t. the ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$. Since this is at the core of the TS sampling analysis, we detail the proof here but postpone convexity results in App. 3.B.

Lemma 3.5.3. *Let* $\Theta_t^{opt} := \{\theta \in \mathbb{R}^d | J(\theta) \geq J(\theta^\star)\} \cap \mathcal{E}_t^{\mathrm{TS}}$ *be the set of optimistic parameters,* $\widetilde{\theta} = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta$ *with* $\eta \sim \mathcal{D}^{\mathrm{TS}}$*, then, on* $\widehat{E}_t$*,*

$$\forall t \geq 1,\ \mathbb{P}\left(\widetilde{\theta} \in \Theta_t^{opt} | \mathcal{F}_t\right) \geq p/2.$$

Proof. We need to study the probability that a $\widetilde{\theta}$ drawn at time t from the TS sampling distribution is optimistic, i.e., $J(\widetilde{\theta}) \geq J(\theta^\star)$, under event $\widehat{E}_t$. More formally let

$$p_t = \mathbb{P}\left(J(\widetilde{\theta}) \geq J(\theta^\star) | \mathcal{F}_t\right).$$

Using the definition of $\widehat{E}_t$ we have that $\theta^\star \in \mathcal{E}_t^{\text{RLS}}$ (i.e., the true parameter vector belongs to the RLS ellipsoid) and then, under event $\widehat{E}_t$, we can replace $J(\theta^\star)$ by the supremum over the ellipsoid as

$$p_t \geq \mathbb{P}\Big(J(\widetilde{\theta}) \geq \sup_{\theta \in \mathcal{E}_t^{\text{RLS}}} J(\theta) \Big| \mathcal{F}_t\Big).$$

By recalling the definition of the TS sampling process, we can write $\widetilde{\theta} = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta$, where $\eta \sim \mathcal{D}^{\text{TS}}$ and for notational convenience, we define the function $f_t(\eta) = J(\widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta)$. Let $\overline{\theta}_t = \arg\max_{\theta \in \mathcal{E}_t^{\text{RLS}}} J(\theta)$ and $\overline{\eta}_t$ be the corresponding η (i.e., $\overline{\eta}_t$ is such that $\overline{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\overline{\eta}_t$). Since the supremum is taken within $\mathcal{E}_t^{\text{RLS}}$, $\overline{\eta}_t$ belongs to the unit ball (i.e., $\overline{\eta}_t \in \mathcal{B}_d(0,1)$). As a result, we can rewrite the previous expression as

$$p_t \geq \mathbb{P}\Big(f_t(\eta) \geq f_t(\overline{\eta}_t) \Big| \mathcal{F}_t\Big).$$

Since the function f_t inherits all the properties of J, notably its convexity in η, we know that the supremum on a convex closed set is reached at least at one point $\bar{\eta}_t$ and that it belongs to the boundary (see Prop. 3.B.1), which in our case corresponds to $\|\overline{\eta}_t\| = 1$. Moreover, let $\mathcal{H}_t(\overline{\eta}_t)$ be the hyperplane tangent to $\overline{\eta}_t$. $\mathcal{H}_t(\bar{\eta}_t)$ splits $\mathbb{R}^d$ in two complementary subsets $\mathcal{G}_t$ and $\mathcal{G}_t^\perp$ where $\mathcal{G}_t$ does not contain the unit ball by convention. Formally, one has:

$$\mathcal{H}_t(\overline{\eta}_t) := \{\eta \in \mathbb{R}^d \text{ s.t. } \eta^\mathsf{T}\bar{\eta}_t = 1\}, \quad \mathcal{G}_t := \{\eta \in \mathbb{R}^d \text{ s.t. } \eta^\mathsf{T}\bar{\eta}_t \geq 1\}.$$

Again, the convexity of f_t ensures that $f_t(\eta) \geq f_t(\bar{\eta}_t)$ for all $\eta \in \mathcal{G}_t$ as proved in Prop. 3.B.2. As illustrated in Fig. 3.5 the probability of being optimistic is now reduced to the probability that η drawn from $\mathcal{D}^{\text{TS}}$ falls into $\mathcal{G}_t$, which corresponds to

$$p_t \geq \mathbb{P}\Big(\eta \in \mathcal{G}_t \Big| \mathcal{F}_t\Big) = \mathbb{P}\Big(\eta^\mathsf{T}\bar{\eta}_t \geq 1 \Big| \mathcal{F}_t\Big).$$

Notice that $\bar{\eta}_t$ is entirely defined by the filtration $\mathcal{F}_t$ and the event $\widehat{E}_t$ and it is thus independent from η. As a result, we obtain from property 1 of Def. 3.3.1 of the TS sampling distribution, that

$$\mathbb{P}\Big(\eta^\mathsf{T}\bar{\eta}_t \geq 1 \Big| \mathcal{F}_t\Big) \geq p.$$

Finally, we show that this property is not affected, up to a second order term, by the high-probability concentration event. It relies on the fact that the chosen confidence level $\delta' = \delta/4T$ is small compared to the anti-concentration probability p of Def. 3.3.1. For sake of simplicity, we assume that $T \geq 1/2p$ which implies that $\delta' \leq p/2$.
For any events A and B, one has

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$$

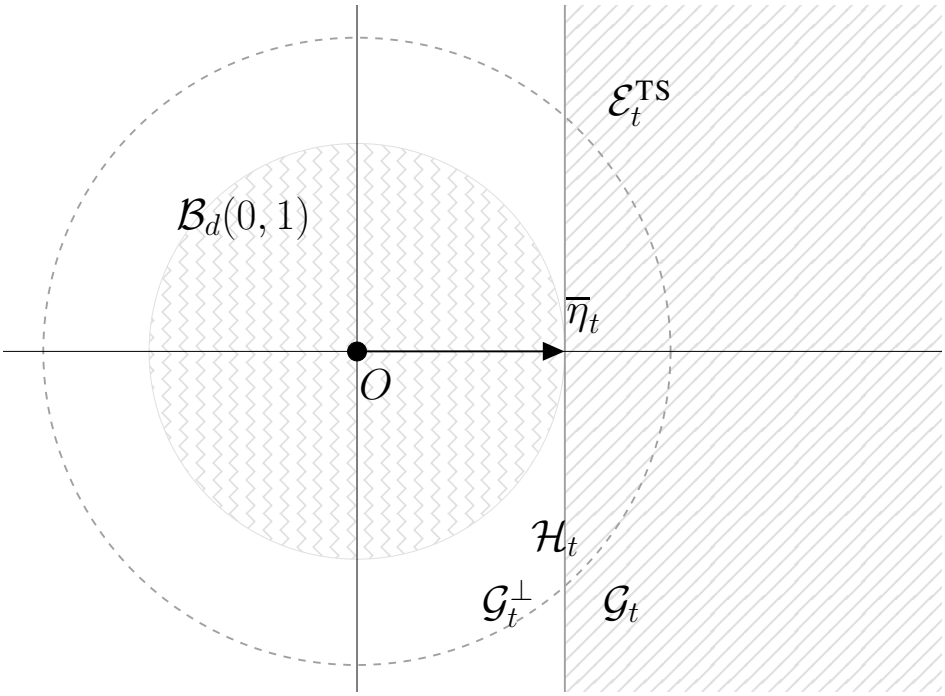

Figure 3.5 – Illustration of the probability of selecting an optimistic $\widetilde{\theta}_t$.

Since, by Def. 3.3.1, $\mathbb{P}(\widetilde{\theta} \in \mathcal{E}_t^{\mathrm{TS}}) \geq 1 - \delta'$, applying the previous inequality to $A := \{J(\widetilde{\theta}) \geq J(\theta^\star)\}$ and $B := \{\widetilde{\theta} \in \mathcal{E}_t^{\mathrm{TS}}\}$ leads to

$$\mathbb{P}(\widetilde{\theta}_t \in \Theta^{\mathrm{opt}} \cap \mathcal{E}_t^{\mathrm{TS}} | \mathcal{F}_t) \geq p - \delta' \geq p/2.$$

□

To control the regret, we now make use of the fact that the probability of being optimistic is constant under the event $\widehat{E}_t$. Let $g(\widetilde{\theta})$ be an arbitrary non-negative function of $\widetilde{\theta}$, then we can write the full expectation as

$$\begin{aligned}
\mathbb{E}[g(\widetilde{\theta})|\mathcal{F}_t]\mathbb{1}\{\widehat{E}_t\} &\geq \mathbb{E}[g(\widetilde{\theta}_t)\mathbb{1}\{\widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\}|\mathcal{F}_t]\mathbb{1}\{\widehat{E}_t\} \\
&\geq \mathbb{E}\Big[g(\widetilde{\theta})\Big|\widetilde{\theta} \in \Theta_t^{\mathrm{opt}}, \mathcal{F}_t\Big]\mathbb{P}\Big(\widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\Big|\mathcal{F}_t\Big)\mathbb{1}\{\widehat{E}_t\} \\
&\geq p/2\mathbb{E}\Big[g(\widetilde{\theta})\Big|\widetilde{\theta} \in \Theta_t^{\mathrm{opt}}, \mathcal{F}_t\Big]\mathbb{1}\{\widehat{E}_t\}.
\end{aligned}$$

Setting $g(\widetilde{\theta}) = 2\gamma_t(\delta')\|x^\star(\widetilde{\theta})\|_{V_t^{-1}}$, we obtains an upper bound for Eq. 3.5 as

$$R_t^{\mathrm{TS}} \leq 4\gamma_t(\delta')/p\,\mathbb{E}\Big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}}|\mathcal{F}_t\Big]\mathbb{1}\{\widehat{E}_t\} \leq 4\gamma_t(\delta')/p\,\mathbb{E}\Big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}}|\mathcal{F}_t\Big],$$

where $2/p$ can be interpreted as the expected time between any two optimistic samples. Finally, by construction $\widetilde{\theta} \overset{d}{=} \widetilde{\theta}_t|\mathcal{F}_t$, thus $x^\star(\widetilde{\theta}) \overset{d}{=} x^\star(\widetilde{\theta}_t)|\mathcal{F}_t$. As a result,

$$R_t^{\mathrm{TS}} \leq 4\gamma_t(\delta')/p\,\mathbb{E}\Big[\|x^\star(\widetilde{\theta}_t)\|_{V_t^{-1}}|\mathcal{F}_t\Big],$$

and we can use Azuma's inequality to obtain the final bound with probability at least $1-\delta/2$

$$R^{\mathrm{TS}}(T) \leq \frac{4\gamma_T(\delta')}{p}\Big(\sum_{t=1}^{T}\|x_t\|_{V_t^{-1}} + \sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}\Big), \tag{3.6}$$

where x_t is the optimal arm $x^\star(\widetilde{\theta}_t)$ selected by TS. The proof is concluded using Cauchy-Schwarz and Prop. 3.2.1 to bound $R^{\mathrm{TS}}(T)$ and Prop. 2.2.1 to bound $R^{\mathrm{RLS}}(T)$.

3.5.4 Final bound

Bounding $R^{\mathbf{RLS}}(T)$. Similar to the sketch in Sect. 3.4, the proof relies on the fact that the RLS guarantees a small *on-policy error*. Let

$$R^{\mathrm{RLS}}(T) = \sum_{t=1}^{T} R_t^{\mathrm{RLS}} = \sum_{t=1}^{T}\big(x_t^T\widetilde{\theta}_t - x_t^T\theta^\star\big)\mathbb{1}\{E_t\},$$

the bound on R^{RLS} is derived as in previous analysis (Abbasi-Yadkori et al., 2011b, Agrawal and Goyal, 2012b). We decompose the term in a *sampling prediction error* and a RLS *prediction error* as follow

$$R^{\mathrm{RLS}}(T) \leq \sum_{t=1}^{T}|x_t^\mathsf{T}(\widetilde{\theta}_t - \widehat{\theta}_t)|\mathbb{1}\{E_t\} + \sum_{t=1}^{T}|x_t^\mathsf{T}(\widehat{\theta}_t - \theta^\star)|\mathbb{1}\{E_t\}$$

By definition of the concentration event E_t,

$$|x_t^\mathsf{T}(\widetilde{\theta}_t - \widehat{\theta}_t)|\mathbb{1}\{E_t\} \leq \|x_t\|_{V_t^{-1}}\gamma_t(\delta'), \qquad |x_t^\mathsf{T}(\widehat{\theta}_t - \theta^\star)|\mathbb{1}\{E_t\} \leq \|x_t\|_{V_t^{-1}}\beta_t(\delta'),$$

thus, one obtains:

$$\sum_{t=1}^{T} R_t^{\mathrm{RLS}} \leq \big(\gamma_T(\delta') + \beta_T(\delta')\big)\sum_{t=1}^{T}\|x_t\|_{V_t^{-1}} \tag{3.7}$$

Plugging everything together. Collecting bounds in Eq. 3.6 and Eq. 3.7, applying Cauchy-Schwarz and using Prop. 3.2.1 , one has, with probability at least $1-\delta/2$,

$$\begin{aligned}
R^{\mathrm{RLS}}(T) + R^{\mathrm{TS}}(T) &\leq \big((4/p+1)\gamma_T(\delta') + \beta_T(\delta')\big)\sum_{t=1}^{T}\|x_t\|_{V_t^{-1}} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}} \\
&\leq \big((4/p+1)\gamma_T(\delta') + \beta_T(\delta')\big)\sqrt{T}\Big(\sum_{t=1}^{T}\|x_t\|^2_{V_t^{-1}}\Big)^{1/2} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}} \\
&\leq \big((4/p+1)\gamma_T(\delta') + \beta_T(\delta')\big)\sqrt{2d\log(1+T/\lambda)}\sqrt{T} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}
\end{aligned}$$

Finally, by Lem. 3.5.1, the event $E = \bigcap_{t\leq T} E_t$ holds with probability at least $1-\delta/2$. Hence, a union bound argument ensures that with probability at least $1-\delta$,

$$\begin{aligned}
R(T) &\leq R^{\mathrm{TS}}(T) + R^{\mathrm{RLS}}(T) \\
&\leq \big(\beta_T(\delta') + \gamma_T(\delta')(1+4/p)\big)\sqrt{2Td\log\Big(1+\frac{T}{\lambda}\Big)} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}
\end{aligned}$$

where $\delta' = \frac{\delta}{4T}$ which proves Thm. 3.5.1.

3.6 Extensions

We provide an alternative proof for TS in LB, leveraging the properties of the optimal value function, the properties of the RLS estimate and the *concentration/anti-concentration* property of the sampling. As the functioning of the proof does not rely on the *specific* shape of the J function, we can readily apply it to similar, yet, more general linear problems. We present here two extensions: the regularized linear optimization problem and the generalized linear bandit.

3.6.1 Regularized linear optimization

Our proof holds for any arm set $\mathcal{X}$ and the corresponding constrained optimization problem $\max_{x\in\mathcal{X}} x^\mathsf{T}\theta^\star$. Similarly, we can apply it to any regularized linear optimization problem $\max_{x\in\mathbb{R}^d} f_{\mu,c}(x;\theta)$, with $f_{\mu,c}(x;\theta) = x^\mathsf{T}\theta + \mu c(x)$, where μ is a constant and $c(x)$ is an arbitrary penalty function of x (e.g., norm-regularization). While there always exists a set of constraints (corresponding to a set of arms $\mathcal{X}_{c,\mu,\theta}$) such that the solution to the constrained and regularized problems coincides, such mapping is often unknown (e.g., $c(x) = \|x\|_1$) and thus TS cannot be run on $\mathcal{X}_{c,\mu,\theta}$ but we need to directly deal with the regularized problem (i.e., sampling $\widetilde{\theta}_t$ and pulling arm $x_t = \arg\max_x f_{\mu,c}(x;\widetilde{\theta}_t)$). In this case, it can be seen that the three main steps of our proof still hold. In fact **1)** $J(\theta)$ is convex, **2)** the gradient of $J(\theta)$ corresponds to the optimal arm $x^*(\theta)$, **3)** Lemma 3.5.3 holds unchanged since it relies on the convexity of $J(\theta)$ and the TS distribution $\mathcal{D}^{\mathrm{TS}}$ is the same. As a result, the regret bound follows.

Setting. We consider here the Regularized Linear Optimization (RLO) problem as an extension of the Linear Bandit problem. Given a set of arms $\mathcal{X} \subset \mathbb{R}^d$ and an unknown parameter $\theta^\star \in \mathbb{R}^d$, a learner aims at each time step $t = 1, \dots, T$ to select action $x_t \in \mathcal{X}$ which maximizes its associated reward $x_t^\mathsf{T}\theta^\star + \mu c(x_t)$ where μ is a known constant and c an arbitrary (yet known) real-valued function. Whenever arm x is pulled, the learner receives a noisy observation $y = x^\mathsf{T}\theta^\star + \xi$. As for LB, we introduce the function $f(x;\theta) = x^\mathsf{T}\theta + \mu c(x)$, and denote as $x^\star(\theta) = \arg\max_{x\in\mathcal{X}} f(x;\theta)$ and $J(\theta) = \max_{x\in\mathcal{X}} f(x;\theta)$ the optimal action and optimal reward associated with θ. The regret is therefore defined as

$$R^{RLO}(T) = \sum_{t=1}^{T} f(x^\star(\theta^\star);\theta^\star) - f(x_t;\theta^\star).$$

Since this problem is just the regularized extension of the Linear Bandit, the TS algorithm is similar to Fig. 3.1 where r_t is replaced y_t and $x_t = \arg\max_{x\in\mathcal{X}} f(x,\widetilde{\theta}_t)$. Under the same assumptions, the regret shares the same bound and our line of proof holds.

Sketch of the proof. First, we decompose the regret

$$R^{RLO}(T) = \sum_{t=1}^{T} \left[(f(x^\star(\theta^\star);\theta^\star) - f(x_t;\widetilde{\theta}_t)) + (f(x_t;\widetilde{\theta}_t) - f(x_t;\theta^\star)) \right]$$
$$= \underbrace{\sum_{t=1}^{T} \left[J(\theta^\star) - J(\widetilde{\theta}_t) \right]}_{=R^{\mathrm{TS}}(T)} + \underbrace{\sum_{t=1}^{T} \left[x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star \right]}_{=R^{\mathrm{RLS}}(T)}.$$

Since Prop. 2.2.1 holds thanks to the linear observations y_t, $R^{\mathrm{RLS}}(T)$ is bounded as in the LB analysis. Finally, to bound $R^{\mathrm{TS}}(T)$, one just need to ensure that Prop. 3.5.1, Lem. 3.5.2 and Lem. 3.5.3 hold.

The convexity of the function f with respect to θ implies the convexity of J: $\forall x \in \mathcal{X}$, $\forall \theta, \theta' \in \mathbb{R}^d$, $\forall \alpha \in (0,1)$,

$$J(\alpha\theta + (1-\alpha)\theta') = \max_{x\in\mathcal{X}} f(x;\alpha\theta + (1-\alpha)\theta')$$
$$\leq \max_{x\in\mathcal{X}} \Big(\alpha f(x;\theta) + (1-\alpha) f(x;\theta')\Big) \leq \alpha J(\theta) + (1-\alpha) J(\theta').$$

Then, J is real-valued and convex which implies its continuous differentiability thanks to Alexandrov's theorem. As a consequence, the first step of the proof holds.
The equality between the gradient $\nabla J(\theta)$ and the optimal arm $x^\star(\theta)$ can be derived as in Lem. 3.5.2: for any $\theta, \bar{\theta} \in \mathbb{R}^d$, by definition, $J(\theta) = f(x^\star(\theta);\theta)$ and $J(\bar{\theta}) \geq f(x^\star(\theta);\bar{\theta})$. Then,

$$J(\bar{\theta}) - f(x^\star(\theta),\bar{\theta}) \geq 0 := J(\theta) - f(x^\star(\theta),\theta),$$
$$J(\bar{\theta}) \geq J(\theta) + f(x^\star(\theta),\bar{\theta}) - f(x^\star(\theta),\theta) = J(\theta) + x^\star(\theta)^\mathsf{T}\left(\bar{\theta} - \theta\right), \quad \forall \bar{\theta} \in \mathbb{R}^d,$$

which is the definition of the sub-gradient. Finally, the almost everywhere differentiability of J ensures the sub-gradient to be a singleton and hence equals the gradient. Therefore, Lem. 3.5.2 holds and so is step 2. Finally, since the optimism just relies on the convexity of J and on the over-sampling, it is satisfied in the RLO and step 3 holds. As a result, we obtain the same regret bound as in the LB.

On the other hand, the original proof by Agrawal and Goyal (2012b) could be less readily applied to this case. First notice that the mapping from μ and $c(x)$ to the constrained set $\mathcal{X}_{c,\mu,\theta^\star}$ requires the unknown parameter $\theta^\star$. This means that if we pass from the regularized problem to the constrained problem at each time step t, we would be working on a set $\mathcal{X}_{c,\mu,\widetilde{\theta}_t}$ which keeps changing over time. While Agrawal and Goyal (2012b) study the contextual bandit problem where $\mathcal{X}_t$ changes arbitrarily over time, in this case $\mathcal{X}_t$ would change in response to $\widetilde{\theta}_t$ itself (i.e., it would not be available in advance) and the analysis would bound the per-step regret $r_t = \max_{x\in\mathcal{X}_{c,\mu,\widetilde{\theta}_t}} x^\mathsf{T}\theta^\star - x_t^\mathsf{T}\theta$, which does not correspond to the desired regret on $f_{\mu,c}$ (the true optimal arm $x^\star(\theta^\star)$ may not even be in $\mathcal{X}_{c,\mu,\widetilde{\theta}_t}$). Alternatively, we need to formulate a suitable definition of saturated and unsaturated arms for $f_{\mu,c}(x;\theta)$, which does not seem trivial and it may require developing a more *ad-hoc* analysis.

3.6.2 Generalized linear bandit

Another interesting extension is the generalized linear bandit (GLM) problem of Filippi et al. (2010). In this setting, the reward associated to arm $x \in \mathcal{X}$ is no longer drawn from the linear regression model but is generated as $r(x) = \mu(x^\mathsf{T}\theta^\star) + \xi$, where μ is the so-called *link function*, $\theta^\star \in \mathbb{R}^d$ is a fixed but unknown parameter vector and ξ is a random zero-mean noise. One of the major advantage of this setting is that it encompasses *logistic regression*. It can model the case when the reward is in $[0, 1]$ and thus became very popular in recommender system where the reward represents the probability of click.
Similarly to the regularized optimization problem, a regret bound can be derived for the GLM problem using the same line of proof that we use for LB. It first relies on the fact that, under suitable assumptions about the link function μ, consistent estimates are available for $\theta^\star$ together with high probability confidence ellipsoids. Then, we can show that the GLM optimal value function $J^{\mathrm{GLM}}(\theta) := \sup_{x\in\mathcal{X}} \mu(x^\mathsf{T}\theta)$ is related to the LB optimal value function $J(\theta)$ as $J^{\mathrm{GLM}}(\theta) = \mu(J(\theta))$ and thus, link the regret of GLM to the regret of LB.
We present here how to apply our derivation to the generalized linear bandit (GLM) problem of Filippi et al. (2010). The regret bound is obtained by basically showing that the GLM problem can be reduced to studying the linear case.

The setting. Let $\mathcal{X} \subset \mathbb{R}^d$ be an arbitrary (finite or infinite) set of arms. Every time an arm $x \in \mathcal{X}$ is pulled, a reward is generated as $r(x) = \mu(x^\mathsf{T}\theta^\star) + \xi$, where μ is the so-called *link function*, $\theta^\star \in \mathbb{R}^d$ is a fixed but unknown parameter vector and ξ is a random zero-mean noise. The value of an arm $x \in \mathcal{X}$ is evaluated according to its expected reward $\mu(x^\mathsf{T}\theta^\star)$ and for any parameter $\theta \in \mathbb{R}^d$ we denote the optimal arm and its optimal value as

$$x^{\star,\mathrm{GLM}}(\theta) = \arg\max_{x\in\mathcal{X}} \mu(x^\mathsf{T}\theta), \qquad J^{\mathrm{GLM}}(\theta) = \sup_{x\in\mathcal{X}} \mu(x^\mathsf{T}\theta).$$

Then $x^\star = x^{\star,\mathrm{GLM}}(\theta^\star)$ is the optimal arm associated with the true parameter $\theta^\star$ and $J^{\mathrm{GLM}}(\theta^\star)$ its optimal value. At each step t, a learner chooses an arm $x_t \in \mathcal{X}$ using all the information observed so far (i.e., sequence of arms and rewards) but without knowing $\theta^\star$ and $x^\star$. At step t, the learner suffers an *instantaneous regret* corresponding to the difference between the expected rewards of the optimal arm $x^\star$ and the arm x_t played at time t. The objective of the learner is to minimize the *cumulative regret* up to a finite step T,

$$R^{\mathrm{GLM}}(T) = \sum_{t=1}^{T} \Big(\mu(x^{\star,\mathsf{T}}\theta^\star) - \mu(x_t^\mathsf{T}\theta^\star)\Big).$$

Assumptions. The assumptions associated with this more general problem are the same as in the linear bandit problem plus one regarding the link function. Formally, we require assumption 3.2.1, 3.2.2 and 3.2.3 and add:

Assumption 3.6.1 (link function)**.** *The link function $\mu : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, Lipschitz with constant k_μ and such that $c_\mu = \inf_{\theta\in\mathbb{R}^d, x\in\mathcal{X}} \mu(x^\mathsf{T}\theta) > 0$.*

Technical tools. Let $(x_1, \dots, x_t) \in \mathcal{X}^t$ be a sequence of arms and $(r_2, \dots, r_{t+1})$ be the corresponding observed (random) rewards, then the unknown parameter $\theta^\star$ can be estimated by GLM estimator. Following Filippi et al. (2010) one gets, for any regularization parameter $\lambda \in \mathbb{R}^+$,

$$\widehat{\theta}_t^{\mathrm{GLM}} = \arg\min_{\theta \in \mathbb{R}^d} \| \sum_{s=1}^{t-1} \Big(r_{s+1} - \mu(x_s^\mathsf{T} \theta) \Big) x_s \|_{V_t^{-1}}^2, \tag{3.8}$$

where V_t is the same design matrix as in the linear case. Similarly to Prop. 2.2.1, we have a concentration inequality for the GLM estimate.

Proposition 3.6.1 (Prop. 1 in appendix.A in (Filippi et al., 2010))**.** *For any $\delta \in (0,1)$, under assumptions 3.2.1, 3.2.2, 3.2.3 and 3.6.1, for any $\mathcal{F}_t^x$-adapted sequence $(x_1, \dots, x_t, \dots)$, the prediction returned by the GLM estimator $\widehat{\theta}_t^{GLM}$ (Eq. 3.8) is such that for any fixed $t \geq 1$,*

$$\|\widehat{\theta}_t^{GLM} - \theta^\star\|_{V_t} \leq \frac{\beta_t(\delta)}{c_\mu},$$

and

$$\forall x \in \mathbb{R}^d, \quad \|\mu(x^\mathsf{T}\widehat{\theta}_t^{GLM}) - \mu(x^\mathsf{T}\theta^\star)\| \leq \frac{k_\mu \beta_t(\delta)}{c_\mu} \|x\|_{V_t^{-1}},$$
$$\|x^\mathsf{T}\widehat{\theta}_t^{GLM} - x^\mathsf{T}\theta^\star\| \leq \frac{\beta_t(\delta)}{c_\mu} \|x\|_{V_t^{-1}},$$

with probability $1-\delta$ (w.r.t. the noise sequence $\{\xi_t\}_t$ and any other source of randomization in the definition of the sequence of arms), where $\beta_t(\delta)$ is defined as in Eq. 3.3.

The Asm. 3.6.1 on the link function together with the properties of the GLM estimator implies the following: **1)** since the first derivative is strictly positive, μ is strictly increasing and $x^{\star,\mathrm{GLM}}(\theta) = \arg\max_{x \in \mathcal{X}} x^\mathsf{T}\theta = x^\star(\theta)$ so we retrieve the optimal arm of the linear case. Similarly, $J^{\mathrm{GLM}}(\theta) = \mu(J(\theta))$; **2)** the concentration inequality of the GLM estimate involves the same ellipsoid as for the RLS (multiplied by a factor $\frac{1}{c_\mu}$). These two facts suggest to use then exactly the same TS algorithm as for the linear case (with a $\beta_t(\delta')$ multiplied by a factor $\frac{1}{c_\mu}$, where $\delta' = \frac{\delta}{4T}$).

Sketch of the proof. From the previous comments, making use of the property of μ, one just need to reduce the GLM case to the standard linear case.

$$\begin{aligned} R^{\mathrm{GLM}}(T) &= \sum_{t=1}^{T} \Big(\mu(x^\star\theta^\star) - \mu(x_t^\mathsf{T}\theta^\star) \Big), \\ &= \sum_{t=1}^{T} \Big(\mu(x^\star\theta^\star) - \mu(x_t^\mathsf{T}\tilde{\theta}_t) \Big) + \sum_{t=1}^{T} \Big(\mu(x_t^\mathsf{T}\tilde{\theta}_t) - \mu(x_t^\mathsf{T}\theta^\star) \Big) \\ &\leq \sum_{t=1}^{T} \Big(\mu(x^\star\theta^\star) - \mu(x_t^\mathsf{T}\tilde{\theta}_t) \Big) + \sum_{t=1}^{T} k_\mu \|x\|_{V_t^{-1}} \|\tilde{\theta}_t - \theta^\star\|_{V_t}. \end{aligned}$$

The second term is bounded exactly as $R^{\mathrm{RLS}}(T)$. To bound the first one, we make use of the fact that

$$\mu(x^\star\theta^\star) - \mu(x_t^\mathsf{T}\tilde{\theta}_t) \le k_\mu\Big(J(\theta^\star) - J(\tilde{\theta}_t)\Big), \quad \text{if} J(\theta^\star) - J(\tilde{\theta}_t) \ge 0,$$
$$\mu(x^\star\theta^\star) - \mu(x_t^\mathsf{T}\tilde{\theta}_t) \le c_\mu\Big(J(\theta^\star) - J(\tilde{\theta}_t)\Big), \quad \text{otherwise.}$$

Following the proof of the linear case, with high probability, for all $t \ge 1$,

$$J(\theta^\star) - J(\tilde{\theta}_t) \le \frac{2\gamma_t(\delta')}{c_\mu p}\mathbb{E}\Big(\|x_t\|_{V_t^{-1}}|\mathcal{F}_t\Big).$$

Since the r.h.s is strictly positive one can bound the first part of the regret, independently of the sign by,

$$\sum_{t=1}^{T}\Big(\mu(x^\star\theta^\star) - \mu(x_t^\mathsf{T}\tilde{\theta}_t)\Big) \le \frac{2k_\mu\gamma_T(\delta')}{c_\mu p}\sum_{t=1}^{T}\mathbb{E}\Big(\|x_t\|_{V_t^{-1}}|\mathcal{F}_t\Big).$$

Finally, the same proof as in the linear case leads to the following bound for the Generalized Linear Bandit regret.

Lemma 3.6.1. *Under assumptions 3.2.1,3.2.2,3.2.3 and 3.6.1, the cumulative regret of* TS *over T steps is bounded as*

$$R^{GLM}(T) \le \frac{k_\mu}{c_\mu}\Big(\beta_T(\delta') + \gamma_T(\delta')(1+2/p)\Big)\sqrt{2Td\log\Big(1+\frac{T}{\lambda}\Big)} + \frac{2k_\mu\gamma_T(\delta')}{pc_\mu}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}$$

with probability $1-\delta$ where $\delta' = \frac{\delta}{4T}$.

3.6.3 Other extensions

To go further, we can generalize our proof to the other convex optimization problems $\max_{x\in\mathcal{X}} f(x,\theta)$, with linear observations (i.e., $y = x^\mathsf{T}\theta + \xi$). If $f(x,\theta)$ is convex in θ, then $J(\theta)$ is convex as well, thus enabling the possibility to apply our line of proof. More precisely, the gradient of J to the arms played by TS should be related (step 2, Lem. 3.5.2) and the on-policy prediction error R^{RLS} measured w.r.t. f should be bounded (Prop. 2.2.1). Whenever these properties are satisfied, the regret result follows. Notice that while the original proof by Agrawal and Goyal (2012b) may be extended to cover some of these problems, its requirements are slightly stronger. In fact, the definition of saturated and unsaturated arms relies on the fact that $f(x,\widehat{\theta}_n)$ concentrates to $f(x,\theta)$ for any x, while in our case, we only need to bound R^{RLS}, which corresponds to an *on-policy* error, where prediction errors are measured *on* the specific arms selected by the algorithm. While this advantage may appear abstract, let consider the reinforcement learning case, where $f(x,\theta)$ is the value function of a policy x in an environment θ. In this case, $f(x,\theta^\star)$ may actually be unbounded for some x (i.e., the policy x does not control the system) and the definition of saturated/unsaturated arms could not be easily adjusted. This suggests that our proof could enable covering special RL cases as well. Finally, we remark that defining TS as a randomized algorithm and using convex geometry arguments in its analysis bears a strong resemblance with follow-the-perturbed-leader algorithm and its regret analysis in adversarial linear bandit (Abernethy et al., 2015), suggesting that the two approaches may be strongly related.

3.7 Discussion

We developed an alternative proof for TS in LB with novel insights on the core elements of the algorithm (*optimism*) and the structure of the problem (*support function* $J(\theta)$). There are a number of possible applications of our results and future directions of investigation. The main open question is whether or not oversampling is needed to guarantee a $\sqrt{T}$ regret bound for TS. Since this worsens the bound by $\sqrt{d}$, answering this question could improve the current frequentist bound from $\tilde{O}(d^{3/2}\sqrt{T})$ to $\tilde{O}(d\sqrt{T})$, thus matching the bound achieved by OFUL. We first present numerical experiments that compare the Bayesian and frequentist versions of TS and exhibit the dependency of the constant of the frequentist regret w.r.t d. Then, we stress why oversampling is needed in the current analysis and discuss how to relax it.

Numerical experiments. To understand the impact of the oversampling, we compare two instances of the TS algorithm:

1) we denote as $FreqTS$ the instance of the algorithm where, at each time step, the parameter is sampled as $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta$ with $\eta \sim \mathcal{N}(0, I)$, for which we prove a $\tilde{O}\big(d^{3/2}\sqrt{T}\big)$ regret bound,

2) we denote as $BayesTS$ its Bayesian counterpart where, at each time step, the parameter is sampled as $\widetilde{\theta}_t = \widehat{\theta}_t + V_t^{-1/2}\eta$ with $\eta \sim \mathcal{N}(0, I)$, for which no frequentist regret guarantee exists.

We compute the regret over trajectories of length $T = 200000$ for values of d spanning $[0, 30]$ and present the results in Fig. 3.6. The motivation for such long trajectories is that the regret curves exhibit slightly different regimes (w.r.t. t). Since we focus on the dependency on d, we discard this effect ensuring that each trajectory reaches the asymptotic regime. The parameter $\theta^\star$ is fixed at the beginning of each trajectory as $\theta^\star = (1, 0, \dots, 0)$. The reason for imposing $\|\theta^\star\| = 1$ is to remove the dependency on d in the norm of $\theta^\star$, which affects the regret through the constant S of Asm. 3.2.2. The RLS estimation is initialized as $V_0 = \lambda I$ with $\lambda = 1$ and $\widehat{\theta}_0$ is randomly chosen on the unite sphere. Finally, the reward noise sequences $\{\xi_t\}_t$ are generated i.i.d according to $\xi_t \sim \mathcal{N}(0, 1)$.

The intuition provided by this experiment is twofold: first, it stresses that the $\tilde{O}\big(d^{3/2}\sqrt{T}\big)$ regret bound of the $FreqTS$ algorithm is tight, so a factor $\sqrt{d}$ cannot be removed by a different analysis; secondly, it suggests that no oversampling is needed to guarantee a $\sqrt{T}$ regret and that a $\tilde{O}\big(d\sqrt{T}\big)$ regret bound could be derived for the $BayesTS$ algorithm.

About optimism and oversampling. As illustrated in Sect. 3.4, in the current proof optimistic steps allows to bound the regret of non-optimistic steps. Nonetheless, it can be shown that some non-optimistic steps (even very *pessimistic*!) may indeed be as "informative" as optimistic steps and allow reducing the regret as well. Let consider

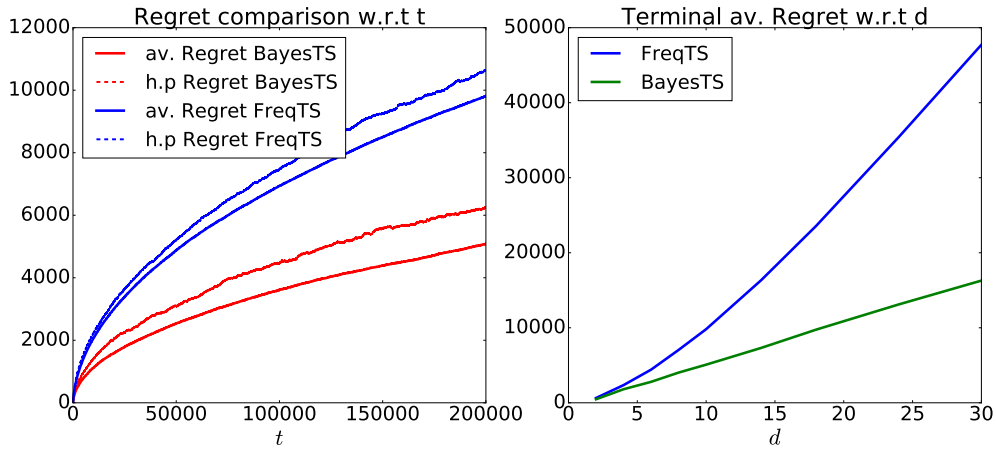

Figure 3.6 – Regret comparison between *BayesTS* and *FreqTS* algorithm. *Left:* empirical average and high probability regret w.r.t t for 100 trajectories with $d = 10$. *Right:* empirical average terminal regret $R(T)$ for $T = 200000$ w.r.t d for 100 trajectories.

a minor change in the line of proof, anticipating the use of the convexity of J, i.e.,

$$\begin{aligned} R_t^{\mathrm{TS}} &\leq \sup_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} \nabla J(\theta^\star)^\mathsf{T}(\theta^\star - \theta)\mathbb{1}\{E_t\} \\ &\leq \|\nabla J(\theta^\star)\|_{V_t^{-1}} 2\gamma_t(\delta')\mathbb{1}\{E_t\}. \end{aligned}$$

If we sample a $\widetilde{\theta}$ such that the gradient at it $\nabla J(\widetilde{\theta})$ (i.e., which coincides with the corresponding optimal action $x^\star(\widetilde{\theta})$) has the same V_t^{-1}-norm as $\nabla J(\theta^\star)$, then we could apply the same reasoning as in the original sketch of the proof and bound the regret of any subsequent step. More formally, we can define the set $\Theta_t^{\mathrm{grad}} = \{\theta : \|\nabla J(\theta)\|_{V_t^{-1}} \geq \|\nabla J(\theta^\star)\|_{V_t^{-1}}\}$ of parameters that have larger gradient than $\theta^\star$'s. Similar to Θ^{opt}, if the probability of sampling $\widetilde{\theta}$ in Θ_t^{grad} is lower-bounded by a constant p', then the proof can be reproduced with exactly the same arguments and result. Even further, we could relax the requirement and define $\Theta_t^{\mathrm{grad}}(\alpha) = \{\theta : \|\nabla J(\theta)\|_{V_t^{-1}} \geq \alpha\|\nabla J(\theta^\star)\|_{V_t^{-1}}\}$, with $\alpha < 1$, which would allow even a bigger probability at the cost of an extra constant factor α in the final regret.

As illustrated in Fig. 3.7, in the case $\mathcal{X} = B_d(0, 1)$, $\Theta_t^{\mathrm{grad}}(\alpha)$ corresponds to a cone whose overlap with $\mathcal{E}^{\mathrm{TS}}$ may actually be even larger than for Θ^{opt}. This illustration shows that the set of *useful* explorative actions does not necessarily coincide with the set of optimistic parameters and that many more parameters in $\mathcal{E}^{\mathrm{TS}}$ may contribute to reduce the regret. This may explain the empirical success of TS and it may suggest that the oversampling by a factor $\sqrt{d}$ to ensure optimism may be a too strong requirement. Finally, we remark that a similar optimistic argument is employed by Agrawal and Goyal (2013) in MAB. Nonetheless, in Lemma 2 they prove that the probability of being optimistic increases over time. This may suggest that $\mathcal{E}^{\mathrm{TS}}$ needs to be only a *constant* fraction bigger than $\mathcal{E}^{\mathrm{RLS}}$, since the initial small probability of being optimistic would tend to a constant (or even to 1) later on during the learning process. Whether this argument holds and how to prove it remains an open question.

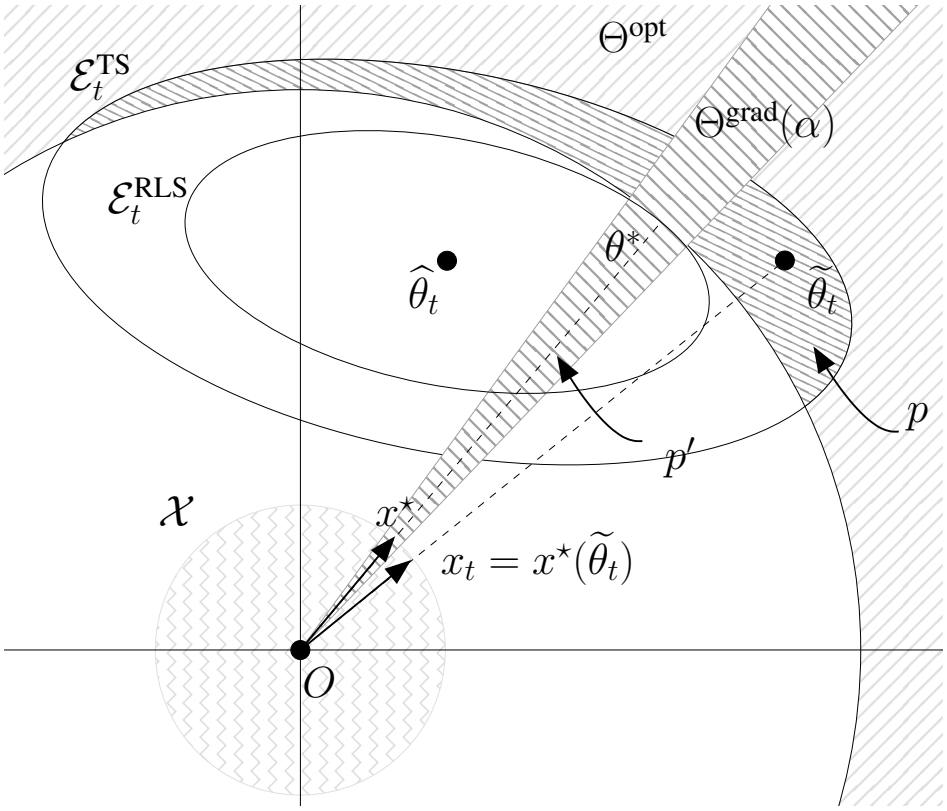

Figure 3.7 – Illustration of the non-optimistic region that could contribute to reduce the regret.

Appendix

3.A Examples of TS distributions

Example 1: Uniform distribution $\eta \sim \mathcal{U}_{B_d(0,\sqrt{d})}$. The uniform distribution satisfies the concentration property with constants $c = 1$ and $c' = \frac{e}{d}$ by definition. Since the set $\{\eta | u^\mathsf{T}\eta \geq 1\} \cap B_d(0, \sqrt{d})$ is an hyper-spherical cap for any direction u of $\mathbb{R}^d$, the the anti-concentration property is satisfied provided that the ratio between the volume of an hyper-spherical cap of height $\sqrt{d} - 1$ and the volume of the ball of radius $\sqrt{d}$ is constant (i.e., independent from d). Using standard geometric results (see Prop. 3.E.1), one has that for any vector $\|u\| = 1$

$$\mathbb{P}(u^\mathsf{T}\eta \geq 1) = \frac{1}{2} I_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right),$$

where $I_x(a, b)$ is the incomplete regularized beta function. In Prop. 3.E.2 we prove that

$$I_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \geq \frac{1}{8\sqrt{3\pi}},$$

and hence we obtain $p = \frac{1}{16\sqrt{3\pi}}$. □

Example 2: Gaussian case $\eta \sim \mathcal{N}(0, I_d)$. The concentration property comes directly from the Chernoff bound for standard Gaussian random variable together with union bound argument. For any $\alpha > 0$, we have

$$\mathbb{P}(\|\eta\| \leq \alpha\sqrt{d}) \geq \mathbb{P}(\forall 1 \leq i \leq d,\ |\eta_i| \leq \alpha) \geq 1 - d\mathbb{P}(|\eta_i| \geq \alpha).$$

Standard concentration inequality for Gaussian random variable gives, $\forall \alpha > 0$,

$$\mathbb{P}(|\eta_i| \geq \alpha) \leq 2e^{-\alpha^2/2}.$$

Plugging everything together with $\alpha = \sqrt{2 \log \frac{2d}{\delta}}$ gives the desired result with $c = c' = 2$. Let η_i be the i-th component of η for any $1 \leq i \leq d$. Then $\eta_i \sim \mathcal{N}(0, 1)$. Since η is rotationally invariant, for any direction u of $\mathbb{R}^d$ and an appropriate choice of basis, we have $\mathbb{P}(u^\mathsf{T}\eta \geq 1) \geq \mathbb{P}(\eta_1 \geq 1)$. From standard Gaussian properties (see Thm 2 of Chang et al. (2011)) we have

$$\mathbb{P}(\eta_1 \geq 1) = \frac{1}{2}\mathrm{erfc}\left(\frac{1}{\sqrt{2}}\right) \geq \frac{1}{4\sqrt{e\pi}}$$

which ensures the anti-concentration property with $p = \frac{1}{4\sqrt{e\pi}}$. □

3.B Properties of convex function

Proposition 3.B.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function and C be a closed convex set of $\mathbb{R}^d$. Then, on C, f reaches its maximum on the boundary of C.*

Proof. Let's denote as $int(C)$ and $bound(C)$ the interior and the boundary of the closed convex set C respectively. Assume that $\exists x^\star \in int(C)$ such that $f(x^\star) > f(x)$ for any $x \in bound(C)$ and $f(x^\star) \geq f(y)$ for any $y \in int(C)$.

Then define $y = x^\star + \epsilon(x^\star - x)$ for some $x \in bound(C)$. By definition of the open set $int(C)$, $\exists \epsilon > 0$ such that $y \in int(C)$. Moreover, $x^\star \in [y, x]$ i.e.,

$$x^\star = (1-t)x + ty, \quad t = \frac{1}{1+\epsilon} \in]0,1[$$

Using the convexity of f, one has

$$f(x^\star) \leq (1-t)f(x) + tf(y) < (1-t)f(x^\star) + tf(y) \quad \Rightarrow \quad f(x^\star) < f(y)$$

which is impossible by assumption. □

Proposition 3.B.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Let $B_d(0,1)$ be the unit $d-$dimensional ball and $S_d(0,1)$ the associated unit sphere.*
Let $x^\star \in S_d(0,1)$ such that $f(x^\star) \geq f(x)$ for all $x \in B_d(0,1)$, and let $\mathcal{H}(x^\star)$ be the hyperplan tangent to $B_d(0,1)$ at the point $x^\star$, which splits $\mathbb{R}^d$ into two complementary subsets $\mathcal{G}(x^\star)$ and $\mathcal{G}^\perp(x^\star)$ defined respectively by

$$\begin{aligned}\mathcal{H}(x^\star) &= \{x \in \mathbb{R}^d \ s.t. \ x^\mathsf{T} x^\star = 1\},\\ \mathcal{G}(x^\star) &= \{x \in \mathbb{R}^d \ s.t. \ x^\mathsf{T} x^\star \geq 1\},\\ \mathcal{G}(x^\star)^\perp &= \{x \in \mathbb{R}^d \ s.t. \ x^\mathsf{T} x^\star < 1\}.\end{aligned}$$

Then, $\forall y \in \mathcal{G}(x^\star), \quad f(y) \geq f(x^\star)$.

Proof. We first notice that from Proposition 3.B.1 $x^\star$ is well defined since the maximum is reached on the boundary. The associated subspace $\mathcal{G}(x^\star)$ is then

$$\mathcal{G}(x^\star) := \{y = x^\star + u, u \in \mathbb{R}^d \ | \ u^\mathsf{T} x^\star \geq 0\}.$$

We want to show that $f(y) \geq f(x^\star)$ for any $y \in \mathcal{G}(x^\star)$. We introduce the increasing sequence of subspace

$$\mathcal{G}_n = \left\{ y = x^\star + u, u \in \mathbb{R}^d \ | \ u^\mathsf{T} x^\star \geq \frac{\|u\|}{2(n-1)} \right\}, \quad n \geq 2.$$

For any $y = x^\star + u$ in $\mathcal{G}_n$, we associate $x = x^\star - \frac{1}{2(n-1)} \frac{u}{\|u\|}$. By definition of y (and hence u), we have

$$\|x\|^2 = 1 + \frac{1}{2(n-1)}^2 - \frac{1}{2(n-1)\|u\|} u^\mathsf{T} x^\star = 1 + \frac{1}{2(n-1)} \left[\frac{1}{2(n-1)} - \frac{u^\mathsf{T}}{\|u\|} x^\star \right] \leq 1,$$

which means that $x \in \mathcal{B}_d(0,1)$. Moreover let $t = [2(n-1)\|u\| + 1]^{-1}$, $t \in]0,1[$ one has $x^\star = (1-t)x + ty$. Since $x \in \mathcal{B}_d(0,1)$ then

$$f(x^\star) \leq (1-t)f(x) + tf(y) \leq (1-t)f(x^\star) + tf(y) \quad \Rightarrow \quad f(x^\star) \leq f(y).$$

Since the statement of the proposition holds for any $\mathcal{G}_n$, then we obtain the desired result for $\mathcal{G}(x^\star)$ by continuity of f. Let $y \in \mathcal{G}(x^\star)$, $y = x^\star + u$. If $u^\mathsf{T} x^\star > 0$, then $\exists n \geq 2$ such that $y \in \mathcal{G}_n$ and the proposition is satisfied. Otherwise, if $u^\mathsf{T} x^\star = 0$, we introduce the sequences $\{u_n\}$ and $\{y_n\}$ defined as:

$$u_n = u + \frac{\|u\|}{\sqrt{1 - \frac{1}{2(n-1)}^2}} \frac{x^\star}{2(n-1)} = u + \frac{\|u_n\|}{2(n-1)} x^\star,$$
$$y_n = x^\star + u_n.$$

By construction, $y_n \in \mathcal{G}_n$ and $y_n \to y$ as $n \to \infty$. Since the $f(y_n) \geq f(x^\star)$ for any $n \geq 2$ we obtain the desired result taking the limit since f is continuous as a convex function on $\mathbb{R}^d$. □

Theorem 3.B.1 (A.D. Alexandrov)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, then it is twice differentiable almost everywhere with respect to the Lebesgue's measure.*

Proof. This result is an extension of the Rademacher's theorem for convex functions. A proof can be found in (Niculescu and Persson, 2006), Thm. 3.11.2. □

3.C Properties of support function (proof of Proposition 3.5.1)

We study the *support function* of a set C, which is a function $f_C : \mathbb{R}^d \to \mathbb{R}$ such that

$$f_C(\theta) = \sup_{x \in C} x^\mathsf{T} \theta$$

Those functions are at the core of convex geometry analysis.

Proposition 3.C.1. *Let $C \subset \mathbb{R}^d$ be a non-empty compact set and f_C the associated support function. Then,*

1. *f_C is real-valued and $\sup_{x \in C} x^\mathsf{T} \theta$ is attained in C,*
2. *f_C is convex,*
3. *f_C is continuous on $\mathbb{R}^d$ and twice differentiable almost everywhere with respect to the Lebesgue's measure.*

Proof.

1. This comes directly from the compactness of C: since C is bounded, the support function is real-valued and since C is closed, the supremum is attained in C,
2. Let θ_1, θ_2 two vectors of $\mathbb{R}^d$, and $t \in (0, 1)$. By definition of the supremum, since f_C is real-valued:
$$f_C(t\theta_1 + (1-t)\theta_2) = \sup_{x \in C} \left(t x^\mathsf{T} \theta_1 + (1-t) x^\mathsf{T} \theta_2 \right) \leq t \sup_{x \in C} x^\mathsf{T} \theta_1 + (1-t) \sup_{x \in C} x^\mathsf{T} \theta_2$$
3. The continuity is a consequence of the convexity of f_C on the open convex set $\mathbb{R}^d$ and the second order differentiability comes from Alexandrov's theorem 3.B.1.

□

3.D Proof of Lemma 3.5.1

We first bound the two events separately.

Bounding $\widehat{E}$. This bound is a straightforward application of Proposition 2.2.1 together with a union bound argument. Let $\delta' = \delta/(4T)$, then

$$\begin{aligned}
\forall 1 \leq t \leq T, \quad & \mathbb{P}\left(\|\widehat{\theta}_t - \theta^\star\|_{V_t} \leq \beta_t(\delta')\right) \geq 1 - \delta' \\
\text{from union bound,} \quad & \mathbb{P}\left(\bigcap_{t=1}^{T}\left\{\|\widehat{\theta}_t - \theta^\star\|_{V_t} \leq \beta_t(\delta')\right\}\right) \geq 1 - \sum_{t=1}^{T}\mathbb{P}\left(\|\widehat{\theta}_t - \theta^\star\|_{V_t} \geq \beta_t(\delta')\right) \\
\Rightarrow \quad & \mathbb{P}\left(\bigcap_{t=1}^{T}\left\{\|\widehat{\theta}_t - \theta^\star\|_{V_t} \leq \beta_t(\delta')\right\}\right) \geq 1 - \sum_{t=1}^{T}\delta' \\
\Rightarrow \quad & \mathbb{P}\big(\widehat{E}\big) \geq 1 - T\delta' = 1 - \frac{\delta}{4}.
\end{aligned}$$

Bounding $\widetilde{E}$. This bound comes directly from the concentration property of the TS sampling distribution. From the expression of $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t$ where η_t is drawn i.i.d. from $\mathcal{D}^{\mathrm{TS}}$, we have

$$\forall 1 \leq t \leq T, \quad \mathbb{P}\left(\|\widetilde{\theta}_t - \widehat{\theta}_t\|_{V_t} \leq \beta_t(\delta')\sqrt{cd\log\frac{c'd}{\delta'}}\right) = \mathbb{P}\left(\|\eta_t\| \leq \sqrt{cd\log\frac{c'd}{\delta'}}\right).$$

Then from Definition 3.3.1, we have

$$\mathbb{P}\left(\|\eta_t\| \leq \sqrt{cd\log\frac{c'd}{\delta'}}\right) \geq 1 - \delta'.$$

As before, a union bound over the two bounds ensures that $\mathbb{P}(\widetilde{E}) \geq 1 - T\delta' = 1 - \frac{\delta}{4}$. Finally, a union bound argument between the two terms leads to $\mathbb{P}(\widehat{E} \cap \widetilde{E}) \geq 1 - \frac{\delta}{2}$.

3.E Hyperspherical cap and beta function

Proposition 3.E.1. *Let $V_d(R)$ be the volume of the $d-$dimensional ball of radius R and let $V_d^{cap}(h)$ the volume of the hyperspherical cap of height $h = R - r > 0$. Then,*

$$V_d^{cap}(h) = \frac{1}{2}V_d(R)I_{1-(\frac{r}{R})^2}\left(\frac{d+1}{2}, \frac{1}{2}\right)$$

where $I_x(a, b)$ is the incomplete regularized beta function.

Proof. The proof can be found in (Li, 2011). □

Proposition 3.E.2. *Let $I_x(a, b)$ is the incomplete regularized beta function,*

$$\forall d \geq 2, \quad I_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \geq \frac{1}{8\sqrt{3\pi}}$$

Proof. The incomplete regularized beta function can be expressed in terms of the beta function $B(a,b)$ and the incomplete beta function $B_x(a,b)$ where

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt, \qquad B(a,b) = B_1(a,b), \qquad I_x(a,b) = \frac{B_x(a,b)}{B(a,b)}$$

Hence we seek for a lower bound on $B_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right)$ and an upper bound for $B\left(\frac{d+1}{2}, \frac{1}{2}\right)$.

1. Let first find an lower bound for the incomplete beta function. Since $t \to t^{\frac{d-1}{2}}(1-t)^{-1/2}$ is positive and increasing on $[0,1]$, for any $d \geq 2$,

$$\begin{aligned} B_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) &\geq \int_{1-\frac{3}{2d}}^{1-\frac{d}{2}} t^{\frac{d-1}{2}}(1-t)^{-1/2}dt \geq \frac{1}{2d}\left(\frac{3}{2d}\right)^{-1/2}\left(1-\frac{3}{2d}\right)^{\frac{d-1}{2}} \\ &\geq \frac{1}{\sqrt{6d}}\left(1-\frac{3}{2d}\right)^{\frac{d-1}{2}} \geq \frac{1}{\sqrt{6d}}\left(1-\frac{3}{2d}\right)^{\frac{d}{2}} \end{aligned}$$

   From the increasing property of $x \to (1-\frac{\alpha}{x})^x$ for any $\alpha < 1$ the sequence $\left\{\left(1-\frac{3}{2d}\right)^{\frac{d}{2}}\right\}_{d\geq 2}$ is increasing and

$$B_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \geq \frac{1}{\sqrt{6d}}\left(1-\frac{3}{2\times 2}\right)^{\frac{2}{2}} = \frac{1}{4\sqrt{6d}}$$

2. Now we seek for an upper bound for $B\left(\frac{d+1}{2}, \frac{1}{2}\right)$. Since $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ one has:

$$B\left(\frac{d+1}{2}, \frac{1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)} = \sqrt{\pi}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)}$$

   From Chen and Qi (2005), we have the following inequalities for the gamma function $\forall n \geq 1$:

$$\begin{aligned} \frac{\Gamma(n+1/2)}{\Gamma(n+1)} &\leq (n+1/4)^{-1/2} \\ \frac{\Gamma(n+1/2)}{\Gamma(n+1)} &\geq (n+4/\pi-1)^{-1/2} \end{aligned}$$

   Together with $\Gamma(x+1) = x\Gamma(x)$ and treating separately cases where d is even or not, one gets $\forall d \geq 2$

$$\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)} \leq \sqrt{\frac{2}{d}}$$

3. Using the obtained upper and lower bound we get:

$$I_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \geq \frac{\sqrt{d}}{\sqrt{2\pi}\times 4\sqrt{6d}} \geq \frac{1}{8\sqrt{3\pi}}$$

□

Chapter 4

Thompson Sampling in Linear Quadratic System

We now consider the exploration-exploitation tradeoff in linear quadratic (LQ) control problems, where the state dynamics is linear and the cost function is quadratic in the state and control. We analyze the regret of Thompson sampling (TS) (a.k.a. posterior-sampling for reinforcement learning) in the frequentist setting, i.e., when the parameters characterizing the LQ dynamics are fixed. Despite the empirical and theoretical success in a wide range of problems from multi-armed bandit to linear bandit, extending those results to the LQ setting is highly challenging: **1)** standard line of proof that relies on classifying arms into saturated/unsaturated pool cannot be applied here as their associated optimal value could be infinite; **2)** the TS functioning requires frequent policy updates, which is in contrast with the usual lazy update scheme used in most RL algorithm. As a consequence, it raises the issue of bounding the gap in the optimal value at the policy switches.
In the chapter[1], we prove that indeed TS achieves a $O(\sqrt{T})$ regret in LQ problems, thus matching the performance of the OFU approach and confirming the conjecture of Osband and Van Roy (2016). We address the first point leveraging the ideas introduced in Ch. 3, stressing the link between the actual actions chosen by TS and the gradient of the optimal value function. We exhibit the need to trade-off the frequency of sampling optimistic parameters and the frequency of switches in the control policy, and show that lazy update schemes induces at best an overall regret of $\Omega(T^{2/3})$. Finally, we derive novel bound on the regret due to policy switches, thus allowing to update parameters and the policy at each step and overcome the limitations due to lazy updates.

Contents

[1]This chapter is an extended version of our AI&Stats paper (Abeille and Lazaric, 2017b).

4.1 Introduction

Designing algorithms to properly trade off *exploration* of an unknown environment and *exploitation* of the estimated optimal control policy is one of the most important challenges towards scaling reinforcement learning (RL) (Sutton and Barto, 1998) to problems with large and/or continuous state and action spaces. To this end, we focus in this chapter, on a specific family of continuous state-action MDPs, the linear quadratic (LQ) control problems introduce in Sec. 2.3, where the state transition is linear and the cost function is quadratic in the state and the control. Despite their specific structure, LQ models are very flexible and widely used in practice (e.g., to track a reference trajectory). If the parameter θ defining dynamics and cost is known, the optimal control can be computed explicitly as a linear function of the state with an appropriate gain. On the other hand, when θ is unknown, an exploration-exploitation trade-off needs to be solved. Bittanti et al. (2006) and Campi and Kumar (1998), first proposed an optimistic approach to this problem, showing that the performance of an adaptive control strategy asymptotically converges to the optimal control. Building on this approach and the OFU principle, Abbasi-Yadkori and Szepesvári (2011) proposed a learning algorithm (OFU-LQ) with $O(\sqrt{T})$ cumulative regret. Abbasi-Yadkori and Szepesvári (2015) further studied how the TS strategy, could be adapted to work in the LQ control problem, but due to the difficulty to move from episodic to infinite horizon, no regret guarantee is available, either in a Bayesian or a frequentist sense.

Contributions. In this chapter, we analyze the regret of TS in LQ problems in the frequentist case[2], where θ is a fixed parameter, with no prior assumption of its value, and prove a $O(\sqrt{T})$ regret bound for the 1-dimensional case (i.e., states are one dimensional). The analysis of OFU-LQ relies on three main ingredients: **1)** optimistic parameters, **2)** lazy updates (the control policy is updated only a logarithmic number of times) and **3)** concentration inequalities for regularized least-squares used to estimate the unknown parameter θ. While we build on previous results for the least-squares estimates of the parameters, points **1)** and **2)** should be adapted for TS. Unfortunately, the existing frequentist regret analysis for TS in linear bandit due to Agrawal and Goyal (2012b) cannot be generalized to the LQ case. Leveraging the novel analysis of TS for LB presented in Ch. 3, we first prove that TS has a constant probability to sample an optimistic parameter (i.e., an LQ system whose optimal expected average cost is smaller than the true one) and then we exploit the LQ structure to show how being optimistic allows to directly link the regret to the controls operated by TS over time and eventually bound them. Nonetheless, this analysis reveals a critical trade-off between the frequency with which new parameters are sampled (and thus the chance of being optimistic) and the regret cumulated every time the control policy changes. In OFU-LQ this trade-off is easily solved by construction: the lazy update guarantees that the control policy changes very rarely and whenever a new policy is computed, it is guaranteed to be optimistic. On the other hand, TS relies on the *random* sampling

[2]This setting is more challenging than its Bayesian counterpart in general as it encompasses it.

process to obtain optimistic models and if this is not done *frequently enough*, the regret can grow unbounded which forces TS to favor short episodes. We first show that, sticking to lazy updates, the regret guarantee scales in $O(T^{2/3})$ at best, and then prove a $O(\sqrt{T})$ bounds in the frequentist regret when the policy updates are performed at each time step, thus confirming the conjecture in (Osband and Van Roy, 2016). This result is enabled by a novel lemma that bounds the regret suffered at the switch between two episodes. We show that the regret incurred at policy switches is somehow related to the overall regret and that it can be bounded following similar steps. As a result, we are able to reduce the length of episodes even further (i.e., constant length or even one single step) at no additional cost and fully exploit the optimism of TS.

4.2 Preliminaries

We briefly recall the setting for the LQ problem introduced in Subsec. 2.3.2 and detail the assumptions that we impose on the problem structure as well as the additional material needed for our analysis. Most of the notations in this section are adapted from (Abbasi-Yadkori and Szepesvári, 2011).

The control problem. We consider the discrete-time infinite-horizon linear quadratic (LQ) control problem with state $x \in \mathbb{R}^n$ and control $u \in \mathbb{R}^d$. Given state x_t and control u_t at time t, the next state and cost are computed as:

$$x_{t+1} = A_* x_t + B_* u_t + \epsilon_{t+1}; \qquad c(x_t, u_t) = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t, \tag{4.1}$$

where A_*, B_*, Q, R are matrices of appropriate dimension and $\{\epsilon_{t+1}\}_t$ is a zero-mean process. Following the setting of Abbasi-Yadkori and Szepesvári (2011), we assume Q and R are known, while the unknown parameters are summarized in $\theta_*^\mathsf{T} = (A_*, B_*)$. The objective of the learner is to find a stationary deterministic control policy $\pi : \mathbb{R}^n \to \mathbb{R}^d$ mapping states to controls that minimizes the asymptotic (i.e., infinite horizon) average expected cost

$$J_\pi(\theta_*) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T} c(x_t, u_t)\right], \tag{4.2}$$

with $x_0 = 0$ and $u_t = \pi(x_t)$. We denote as $\pi_*(\theta_*)$ the optimal policy of the LQ problem parametrized by θ_*. We impose the following assumptions over the noise process and the linear system of Eq. 4.1.

Assumption 4.2.1 (Noise)**.** *The noise $\{\epsilon_t\}_t$ is a $\mathcal{F}_t-$martingale difference sequence, where $\mathcal{F}_t$ is the filtration which represents the information knowledge up to time t. Furthermore, the noise is conditionally Gaussian, i.e., $\epsilon_t|\mathcal{F}_t \sim \mathcal{N}(0, I)$ for all $t \leq T$.*

Assumption 4.2.2 (LQ)**.** *The cost matrices Q and R are symmetric p.d. and (A_*, B_*) is stabilizable.*[3]

[3] (A, B) is stabilizable if there exists a gain matrix K s.t. $A + BK$ is stable, i.e., all eigenvalues are in $(-1, 1)$. A formal characterization is given in Def. 2.3.2.

Under Asm. 4.2.1 and 4.2.2, Thm. 2.3.2 guarantees the existence and uniqueness of an optimal policy, such that $\pi_*(\theta_*) = K(\theta_*)x$, where,

$$K(\theta_*) = -\left(R + B^\mathsf{T} P(\theta_*) B\right)^{-1} B^\mathsf{T} P(\theta_*) A,$$
$$P(\theta_*) = Q + A^\mathsf{T} P(\theta_*) A + A^\mathsf{T} P(\theta_*) B K(\theta_*),$$

The optimal average cost is $J_* = J_{\pi_*}(\theta_*) = \mathrm{Tr}(P(\theta_*))$. Finally, we also have that the closed-loop matrix $A_* + B_* K(\theta_*)$ is asymptotically stable. For sake of compactness, we introduce the matrix $H(\theta_*) = \begin{pmatrix} I & K(\theta_*)^\mathsf{T} \end{pmatrix}^\mathsf{T}$ and rewrite the closed-loop matrix as $A_* + B_* K(\theta_*) = \theta_*^\mathsf{T} H(\theta_*)$.

We now construct a constraint set $\mathcal{S}$ such that for any $\theta \in \mathcal{S}$, there exists an optimal control $K(\theta)$. This is done by rejecting the pair (A, B) that are non-stabilizable, i.e., the one for which there exists no linear controller K such that the closed-loop matrix $A+BK$ is asymptotically stable (i.e., has eigenvalues in the open unit disk). Fortunately, this set is of zero Lebesgue measure as provided by the following proposition.

Proposition 4.2.1 (Cor. 12 in (Klamka, 2016))**.** *For given dimensions n and d, the set of dynamical systems which are controllable is open and dense in the space $\mathbb{R}^{n(n+d)}$ of all dynamical system of the form (4.1).*

Since controllability implies stabilizability (see Def. 2.3.2), the set of uncontrollable system is of zero Lebesgue measure and so is the set of unstabilizable system. Moreover, when the pair (A, B) is not stabilizable, there exists no control K such that $A + BK$ is stable, thus, under any linear controller, the state process x_t diverges exponentially. As a result, the associated "optimal" average cost is $J(\theta) = +\infty$.
This property has two major implications: First, it is possible to define the optimal value function over the whole space $\mathbb{R}^{n(n+d)}$ in a continuous manner by setting its value to $+\infty$ wherever θ is a non-stabilizable pair; Second, for any sampling distribution absolutely continuous w.r.t the Lebesgue measure, the associated optimal value function is finite with probability one. However, it might sample parameters that are almost non-stabilizable (and hence of large optimal cost), which is still harmful from both theoretical and practical perspective. This motivates the introduction of the constraint set $\mathcal{S}$ defined as:

Definition 4.2.1. $\mathcal{S} = \{\theta \in \mathbb{R}^{(n+d)\times n} \text{ s.t. } J(\theta) = Tr(P(\theta)) \leq D \text{ and } Tr(\theta\theta^\mathsf{T}) \leq S^2\}$

This definition is implicit since it involves the optimal average cost but offers the advantages of unifying assumption A-2 and A-4 in (Abbasi-Yadkori and Szepesvári, 2011) in a tight way and implies the following guarantees.

Proposition 4.2.2. *$\mathcal{S}$ is a compact set. For any $\theta \in \mathcal{S}$, θ is a stabilizable pair (since $J(\theta) = +\infty$ otherwise) and there exist $\rho < 1$ and $C < \infty$ positive constants such that $\rho = \sup_{\theta\in\mathcal{S}} \|A + BK(A, B)\|_2$ and $C = \sup_{\theta\in\mathcal{S}} \|K(\theta)\|_2$.*

Further, we assume that the true parameter belongs to the constraint set. Formally,

Assumption 4.2.3. *Let $\mathcal{S}$ be defined as in Def. 4.2.1, then $\theta_* \in \mathcal{S}$.*

Finally, we recall a result about the regularity of the Riccati solution.

Proposition 4.2.3 (proof in App. 4.A)**.** *Under Asm. 4.2.1 and for any LQ with parameters $\theta^\mathsf{T} = (A, B)$ and cost matrices Q and R satisfying Asm. 4.2.2, let $J(\theta) = Tr(P(\theta))$ be the optimal solution of Eq. 4.2. Then, the mapping $\theta \in \mathcal{S} \to Tr(P(\theta))$ is continuously differentiable. Furthermore, let $A_c(\theta) = \theta^\mathsf{T} H(\theta)$ be the closed-loop matrix, then the directional derivative of $P(\theta)$ in a direction $\delta\theta$, denoted as $dP(\theta)(\delta\theta)$, where $dP(\theta)(\delta\theta) \in \mathbb{R}^{n\times n}$, is the solution of the Lyapunov equation*

$$dP(\theta)(\delta\theta) = A_c(\theta)^\mathsf{T} dP(\theta)(\delta\theta) A_c(\theta) + C(\theta, \delta\theta) + C(\theta, \delta\theta)^\mathsf{T},$$

where $C(\theta, \delta\theta) = A_c(\theta)^\mathsf{T} P(\theta) \delta\theta^\mathsf{T} H(\theta)$.

The learning problem. We consider the standard online learning setting where at each step t the learner receives the current state x_t as input, it executes a control u_t and it observes the associated cost $c(x_t, u_t)$; the system then transitions to the next state x_{t+1} according to Eq. 4.1. The learning performance is measured by the cumulative regret over T steps, where the costs cumulated over time are compared to the minimal cost obtained on average by the optimal policy. Formally we define

$$R_T(\theta_*) = \sum_{t=0}^{T} \Big(c_t - J_*(\theta_*)\Big).$$

Independently from the control problem, we need basic tools for the estimation of the parameter θ_*. Let $(u_0, \dots, u_t)$ be a sequence of controls and $(x_0, x_1, \dots, x_{t+1})$ be the corresponding states generated according to Eq. 4.1. For any regularization parameter $\lambda \in \mathbb{R}_+^*$ the regularized least-squares (RLS) and the associated design matrix are defined as

$$V_t = \lambda I + \sum_{s=0}^{t-1} z_s z_s^\mathsf{T}; \qquad \widehat{\theta}_t = V_t^{-1} \sum_{s=0}^{t-1} z_s x_{s+1}^\mathsf{T}, \tag{4.3}$$

where $z_t = (x_t, u_t)^\mathsf{T}$. While the RLS estimate of Eq. 4.3 slightly differs from the one of Ch. 3 since it is derived in matrix form, the concentration inequality still hold (see Prop. 2.3.1). Thus, for any $0 < \delta < 1$, we define the high-probability ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$ at each time step t as

$$\mathcal{E}_t^{\mathrm{RLS}} = \Big\{\theta \in \mathbb{R}^{n(n+d)} \Big| \|\theta - \widehat{\theta}_t\|_{V_t} \le \beta_t(\delta')\Big\},\ \beta_t(\delta') = n\sqrt{2\log\left(\frac{\det(V_t)^{1/2}}{\det(\lambda I)^{1/2}\delta'}\right)} + \lambda^{1/2} S, \tag{4.4}$$

where $\delta' = \frac{\delta}{8T}$ and $\|\cdot\|_M$ denote the weighted Frobenius norm associated with any positive definite matrix M, such that, for any $\theta \in \mathbb{R}^{n(n+d)}$, $\|\theta\|_M^2 = \mathrm{Tr}\Big(\theta^\mathsf{T} M \theta\Big)$. Under Asm. 4.2.1 and 4.2.3, Prop. 2.3.1 guarantees that $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$ for all $t \le T$, with probability at least $1 - \delta/8$.

Finally, we also have the standard result of RLS that, together with Prop. 2.3.1, shows that the prediction error on the points z_t used to construct the estimator $\widehat{\theta}_t$ is cumulatively small.

Proposition 4.2.4 (Lem. 10 in (Abbasi-Yadkori and Szepesvári, 2011)). *Let $\lambda \geq 1$, for any arbitrary $\mathcal{F}_t$-adapted sequence $(z_0, z_1, \ldots, z_t)$, let V_{t+1} be the corresponding design matrix, then*

$$\sum_{s=0}^{t} \min\left(\|z_s\|^2_{V_s^{-1}}, 1\right) \leq 2 \log \frac{\det(V_{t+1})}{\det(\lambda I)}.$$

Moreover when $\|z_t\| \leq Z$ for all $t \geq 0$, then

$$\text{and} \quad \sum_{s=0}^{t} \|z_s\|^2_{V_s^{-1}} \leq 2\frac{Z^2}{\lambda}(n+d)\log\left(1 + \frac{(t+1)Z^2}{\lambda(n+d)}\right)$$
$$\beta_t \leq nZ\left(\frac{(n+d)}{\lambda}\log\left(1 + \frac{(t+1)Z^2}{\lambda(n+d)}\right) + 2\log(1/\delta)\right)^{1/2} + \lambda^{1/2}S.$$

4.3 Thompson sampling for LQ

We introduce a specific instance of TS for learning in LQ problems obtained as a modification of the algorithm proposed by Abbasi-Yadkori and Szepesvári (2015), where we replace the Bayesian structure and the Gaussian prior assumption with a randomized Gaussian process and we modify the update rule. The algorithm is summarized in Fig. 4.1. At any step t, given the RLS-estimate $\widehat{\theta}_t$ and the design matrix V_t, TS samples a *perturbed* parameter $\widetilde{\theta}_t$. In order to ensure that the sampling parameter is indeed admissible, we re-sample it until a valid $\widetilde{\theta}_t \in \mathcal{S}$ is obtained.We define $\widetilde{\theta}_t$ as

$$\widetilde{\theta}_t = \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta_t), \tag{4.5}$$

where $\mathcal{R}_{\mathcal{S}}$ is the rejection sampling operator associated with the admissible set $\mathcal{S}$, $\widehat{\theta}_t$ is the RLS-estimate, V_t is the design matrix and each entry of the perturbation matrix $\eta_t \in \mathbb{R}^{(n+d)\times n}$ is a random sample drawn i.i.d. from $\mathcal{N}(0,1)$. Then the control $u_t = K(\widetilde{\theta}_t)x_t$ is executed and the next state x_{t+1} and c_t are observed. The new samples are then used to update $\widetilde{\theta}_t$ and V_t.

Input: $\hat{\theta}_0$, $V_0 = \lambda I$, δ, T, $t_0 = 0$
1: Set $\beta_t = \beta_t(\delta')$ where $\delta' = \frac{\delta}{8T}$ according to Eq. 4.4
2: **for** $t = \{0, \ldots, T\}$ **do**
3: Sample $\widetilde{\theta}_t = \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta_t)$
4: Execute control $u_t = K(\widetilde{\theta}_t)x_t$
5: Observe state x_{t+1} and cost $c_t = c(x_t, u_t)$
6: Compute V_{t+1} and $\widehat{\theta}_{t+1}$ using Eq. 4.3
7: **end for**

Figure 4.1 – Thompson sampling algorithm for LQ.

The major difference of this instance of TS w.r.t. OFU-LQ and the algorithms of Abbasi-Yadkori and Szepesvári (2015) is that we are no longer using a lazy-update

scheme and the policy is updated at each step. As shown in the Subsec. 4.5.4, this is possible because of the novel analysis of the regret suffered at each policy switch.
The algorithm in Fig. 4.1 can be turned into a "proper" Bayesian algorithm if we follow the same approach as in (Abbasi-Yadkori and Szepesvári, 2015) and define a prior $\mathcal{P}_0$ over θ_* as a multivariate Gaussian conditioned on $\theta_* \in \mathcal{S}$. More formally, let $[\theta_*]_i$ be the i-th column of θ_* then the density of the prior is proportional to $\prod_{i=1}^{n} \exp\left(\lambda [\theta_*]_i^\mathsf{T} [\theta_*]_i\right) \mathbb{1}\{\theta_* \in \mathcal{S}\}$ where the indicator function guarantees that the resulting system is in the admissible set. As a result, Eq. 4.5 with $\beta_t = 1$ is indeed the posterior over θ_* at time t. On the other hand, TS can be seen as a randomized algorithm as suggested in Ch. 3. In this case, no prior is needed and the factor β_t in Eq. 4.4 is allowed to change as in Eq. 2.8. We consider the latter instance here and use $\beta_t = \beta_t(\delta')$ with $\delta' = \frac{\delta}{8T}$.

4.4 Challenges and sketch of the proof

In this section, we highlights the challenges of proving a regret bound for TS in LQ problem, stressing the differences with the Bayes and OFU-LQ analysis. Then, we report a sketch of the proof, leveraging ideas presented in Ch. 3, to provide intuition on the behavior of TS. For the sake of illustration, we postpone the rigorous proof to Sec. 4.5. The following analysis only holds on some high probability events and we denote as $\square$, a numerical constant that varies from line to line.
We start by decomposing the regret similarly to Abbasi-Yadkori and Szepesvári (2011, Sec. 4.2).

$$
\begin{aligned}
R(T) \leq & \sum_{t=0}^{T} \{J(\widetilde{\theta}_t) - J(\theta_*)\} && := R^{\mathrm{TS}} \\
& + \sum_{t=0}^{T} \{z_t^\mathsf{T} \widetilde{\theta}_t P(\widetilde{\theta}_t) \widetilde{\theta}_t^\mathsf{T} z_t - z_t^\mathsf{T} \theta_* P(\widetilde{\theta}_t) \theta_*^\mathsf{T} z_t\} && := R^{\mathrm{RLS}} \\
& + \sum_{t=0}^{T} \{x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t - \mathbb{E}\left[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_{t+1}) x_{t+1} | \mathcal{F}_t\right]\} && := R^{mart} \\
& + \sum_{t=0}^{T} \mathbb{E}\left[x_{t+1}^\mathsf{T} \left(P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\right) x_{t+1} | \mathcal{F}_t\right] && := R^{Gap}
\end{aligned}
$$

Before entering into the details of how to bound each of these components, we discuss what are the main challenges in bounding the regret.

4.4.1 Related work and challenges

Since the RLS estimator is the same in both TS and OFU, the regret terms R^{RLS} and R^{mart} can be bounded as in (Abbasi-Yadkori and Szepesvári, 2011). In fact, R^{mart} is a martingale by construction and it can be bounded by Azuma's inequality. The term R^{RLS} is related to the difference between the *true* next expected state $\theta_\star^\mathsf{T} z_t$ and the *predicted* next expected state $\widetilde{\theta}_t^\mathsf{T} z_t$. A direct application of RLS properties makes

this difference small by construction, thus bounding R^{RLS}. Finally, the R^{gap} term is directly affected by the changes in model from any two time instants (i.e., $\widetilde{\theta}_t$ and $\widetilde{\theta}_{t+1}$), while R^{TS} measures the difference in optimal average expected cost between the true model θ_* and the sampled model $\widetilde{\theta}_t$. In the following, we denote by R_t^{gap} and R_t^{TS} the elements at time t of these two regret terms and we refer to them as *consistency regret* and *optimality regret* respectively.

Optimistic approach. OFU-LQ explicitly bounds both regret terms directly by construction. In fact, the lazy update of the control policy allows to set to zero the consistency regret R_t^{gap} in all steps but when the policy changes between two episodes. Since in OFU-LQ an episode terminates only when the determinant of the design matrix is doubled, the number of episodes is bounded by $O(\log(T))$, which bounds R^{gap} as well (with a constant depending on the state bound X and other parameters specific of the LQ system).[4] At the same time, at the beginning of each episode an optimistic parameter $\widetilde{\theta}_t$ is chosen, i.e., $J(\widetilde{\theta}_t) \leq J(\theta_*)$, which directly ensures that R_t^{TS} is upper bounded by 0 at each time step.

Bayesian regret. The lazy PSRL algorithm of Abbasi-Yadkori and Szepesvári (2015) has the same lazy update as OFUL and thus R^{gap} should be controlled the same way. Unfortunately, as hinted in (Osband and Van Roy, 2016), challenges of extending episodic TS results to infinite horizon introduced a flaw in this approach, so even in the Bayesian analysis, bounding R^{gap} remains an open question. On the other hand, the random choice of $\widetilde{\theta}_t$ does not guarantee optimism at each step anymore. Nonetheless, the regret is analyzed in the Bayesian setting, where θ_* is drawn from a known prior and the regret is evaluated *in expectation* w.r.t. the prior. Since $\widetilde{\theta}_t$ is drawn from a posterior constructed from the same prior as θ_*, in expectation its associated $J(\widetilde{\theta}_t)$ is the same as $J(\theta_*)$, thus ensuring that $\mathbb{E}[R_t^{\text{TS}}] = 0$.

Frequentist regret. When moving from Bayesian to frequentist regret, this argument does not hold anymore and the (positive) deviations of $J(\widetilde{\theta}_t)$ w.r.t. $J(\theta_*)$ has to be bounded in high probability. Abbasi-Yadkori and Szepesvári (2011) exploits the linear structure of LQ problems to reuse arguments originally developed in the linear bandit setting. Similarly, we could leverage the analysis of TS for linear bandit by Agrawal and Goyal (2012b) to derive a frequentist regret bound. Agrawal and Goyal (2012b) partition the (potentially infinite) arms into *saturated* and *unsaturated* arms depending on their estimated value and their associated uncertainty (i.e., an arm is saturated when the uncertainty of its estimate is smaller than its performance gap w.r.t. the optimal arm). In particular, the uncertainty is measured using confidence intervals derived from a concentration inequality similar to Prop. 2.3.1. This suggests to use a similar argument and classify policies as saturated and unsaturated depending on their value. Unfortunately, this proof direction cannot be applied in the case of LQR. In fact, in an LQ system θ the performance of a policy π is evaluated by the function $J_\pi(\theta)$ and the policy uncertainty should be measured by a confidence interval constructed as

[4]Notice that the consistency regret is not specific to LQ systems but it is common to all regret analyses in RL (see e.g., UCRL2 (Jaksch et al., 2010)) except for episodic MDPs and it is always bounded by keeping under control the number of switches of the policy (i.e., number of episodes).

$|J_\pi(\theta_*) - J_\pi(\widehat{\theta}_t)|$. Despite the concentration inequality in Prop. 2.3.1, we notice that neither $J_\pi(\theta_*)$ nor $J_\pi(\widehat{\theta}_t)$ may be finite, since π may not stabilize the system θ_* (or $\widehat{\theta}_t$) and thus incur an infinite cost. As a result, it is not possible to introduce the notion of saturated and unsaturated policies in this setting and another line of proof is required. Another key element in the proof of Agrawal and Goyal (2012b) for TS in linear bandit is to show that TS has a constant probability p to select optimistic actions and that this contributes to reduce the regret of any non-optimistic step. In our case, this translates to requiring that TS selects a system $\widetilde{\theta}_t$ whose corresponding optimal policy is such that $J(\widetilde{\theta}_t) \leq J(\theta_*)$. Lem. 4.5.5 shows that this happens with a constant probability p. Furthermore, we can show that optimistic steps reduce the regret of non-optimistic steps, thus effectively bounding the optimality regret R^{TS}. Nonetheless, this is not compatible with lazy updates. In fact, while R^{TS} is small when optimistic parameters $\widetilde{\theta}_t$ are sampled *often enough*, a crude bound of the consistency regret R^{gap} requires to reduce the switches between policies as much as possible (i.e., number of episodes). If we keep the same number of episodes as with the lazy update of OFUL (i.e., about $\log(T)$ episodes), then the number of sampled points is as small as $T/(T - \log(T))$. While OFU-LQ guarantees that any policy update is optimistic by construction, with TS, only a fraction $T/(p(T - \log(T))$ of steps would be optimistic *on average*. Unfortunately, such small number of optimistic steps is no longer enough to derive a bound on the optimality regret R^{TS}. Summarizing, in order to derive a frequentist regret bound for TS in LQ systems, we need the following ingredients. **1)** constant probability of optimism, **2)** connection between optimism and R^{TS} without using the saturated and unsaturated argument, **3)** a novel approach to bound the deviation in the optimal value at the policy switch to guarantee small consistency regret.

4.4.2 Sketch of the proof

The outline of the proof is the following. We prove first the main result that states that TS keeps the *average absolute deviation* of the optimal value function small. Formally, let

$$\Delta_t = \mathbb{E}\Big(|J(\widetilde{\theta}_t) - \mathbb{E}\big(J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\big)|\Big|\mathcal{F}_t^x, E_t\Big),$$

where E_t is some high probability event, and $\mathcal{F}_t^x$ is a filtration encoding the knowledge up to time t (we postpone the formal definitions to Sec. 4.5), we show that

$$\sum_{t=0}^{T} \Delta_t = \sum_{t=0}^{T} \mathbb{E}\Big(|J(\widetilde{\theta}_t) - \mathbb{E}\big(J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\big)|\Big|\mathcal{F}_t^x, E_t\Big) = \widetilde{O}\big(\sqrt{T}\big).$$

Then, we show how this implies bounds for the terms R^{gap} and R^{TS}. Thanks to the constant probability of being optimistic and the fact that sampling distributions does not change much between two subsequent steps, we relate R^{TS} and R^{Gap} to the *average absolute deviation* and thus bound the overall regret. Finally, we recall the bounds from (Abbasi-Yadkori and Szepesvári, 2011) derived for the OFUL strategy, but which still apply to R^{RLS} and R^{mart}. This is due to the

fact that the TS strategy shares the principle of choosing a point within a high probability confidence ellipsoid (deterministically for OFUL, randomly for TS), and then of controlling the system with an optimal control w.r.t the chosen parameter. We sketch here the steps that are used to bound the regret terms specific to TS.

Average absolute deviation. The proof consists in the following chain of inequalities:

$$\begin{aligned}
\sum_{t=0}^{T}\mathbb{E}\Big(|J(\widetilde{\theta}_t)-\mathbb{E}\big(J(\widetilde{\theta}_t)|\mathcal{F}_t^x,E_t\big)|\Big|\mathcal{F}_t^x,E_t\Big) &\leq \square\sum_{t=0}^{T}\mathbb{E}\Big(\|\nabla J(\widetilde{\theta}_t)\|_{V_t^{-1}}|\mathcal{F}_t^x,E_t\Big) && (1)\\
&\leq \square\sum_{t=0}^{T}\mathbb{E}\Big(\|H(\widetilde{\theta}_t)\|_{V_t^{-1}}|\mathcal{F}_t^x,E_t\Big) && (2)\\
&\leq \square\sum_{t=0}^{T}\mathbb{E}\Big(\|z_t\|_{V_t^{-1}}|\mathcal{F}_{t-1},E_{t-1}\Big) && (3)\\
&\leq \square\sum_{t=0}^{T}\|z_t\|_{V_t^{-1}}+\sqrt{T} && (4)\\
&\leq \square\sqrt{T}.
\end{aligned}$$

The objective of the first inequality is to link the *average absolute deviation* of the optimal value function to its gradient, inspired by the discussion of Ch. 3 about the *sensitivity* of the performance w.r.t the randomness of the sampling. However, as opposed to LB, in the LQ setting no convexity argument can be used. We tackle this limitation introducing a modified Poincaré inequality (see Lem. 4.5.3). The Poincaré inequality is a major result in Sobolev spaces that allows to bound a function u by its derivative, as $\|u\|_{L^p}\leq C\|\nabla u\|_{L^p}$. We adapt it to our problem, by modifying the sharp L^1-inequality derived by Acosta and Durán (2004) and apply it to the function $f_t(\eta)=J(\widehat{\theta}_t+\beta V_t^{-1/2}\eta)$ so that $\nabla_\eta f_t=\beta V_t^{-1/2}\nabla_\theta J$.
The second inequality relates the gradient of the optimal value function to the *control* matrix. Thanks to Prop. 4.A.1, we show that for any $\theta\in\mathcal{S}$, for any p.s.d. matrix V, $\|\nabla J(\theta)\|_V\leq\square\,\|H(\theta)\|_V$. This stresses how the *control* chosen by TS over time is related to the *sensitivity* of the optimal value function. However, bounding the absolute deviation Δ_t by the *control matrix* chosen by TS is not enough to bound the cumulative sum, as Prop. 4.2.4 concerns the *actual control*. Formally, we would like to move from $H(\widetilde{\theta}_t)$ to z_t. Noticing that $z_t=H(\widetilde{\theta}_t)x_t$, one can associate the *control* $H(\widetilde{\theta}_t)$ to the chosen *direction* while x_t plays the role of the *amplitude*. We rely on the weak dependence between those two quantities to show their equivalence. For sake of illustration, suppose for now that, conditionally to $\mathcal{F}_{t-1}$, $\widetilde{\theta}_t$ and x_t are independent. Making use of the fact that the state x_t is excited by the noise ϵ_t (see Eq. 4.1), one has $\mathbb{E}(x_tx_t^\mathsf{T}|\mathcal{F}_{t-1},\widetilde{\theta}_t)=\mathbb{E}(x_tx_t^\mathsf{T}|\mathcal{F}_{t-1})\succcurlyeq I$ so

$$\begin{aligned}
\|H(\widetilde{\theta}_t)\|_{V_t^{-1}} &\leq \|H(\widetilde{\theta}_t)\mathbb{E}\Big(x_tx_t^\mathsf{T}|\mathcal{F}_{t-1},\widetilde{\theta}_t\Big)\|_{V_t^{-1}}\\
&\leq \mathbb{E}\Big(\|H(\widetilde{\theta}_t)x_tx_t^\mathsf{T}\|_{V_t^{-1}}|\mathcal{F}_{t-1},\widetilde{\theta}_t\Big)\\
&\leq X\mathbb{E}\Big(\|z_t\|_{V_t^{-1}}|\mathcal{F}_{t-1},\widetilde{\theta}_t\Big),
\end{aligned}$$

where X is an upper bound on $\|x_t\|$. As a consequence, using Azuma's inequality and the law of expectation, one has

$$\mathbb{E}\Big(\|H(\widetilde{\theta}_t)\|_{V_t^{-1}}|\mathcal{F}_t^x\Big) \lesssim \mathbb{E}\Big(\|H(\widetilde{\theta}_t)\|_{V_t^{-1}}|\mathcal{F}_{t-1}\Big) \lesssim X\mathbb{E}\Big(\|z_t\|_{V_t^{-1}}|\mathcal{F}_{t-1}\Big).$$

In practice, $\widetilde{\theta}_t$ and x_t are weakly linearly dependent, but this dependency scales in $\|z_t\|_{V_t^{-1}}$. When both are distributed as Gaussian random variable (conditionally), it is possible to handle this weak dependence and derive a similar result. The last difficulty comes from the conditioning $\widetilde{\theta}_t \in \mathcal{S}$ and $\|x_t\| \leq X$ that breaks this Gaussian property. We overcome this issue by showing that this result still holds for truncated Gaussian random variable, given that the truncations are far enough from the means of the distributions.

Probability of being optimistic. The objective of this step is to relate $J(\widetilde{\theta}_t) - J(\theta_*)$ to Δ_t i.e., the *point-wise deviation* from θ_* w.r.t. J to the *average absolute deviation*. Up to a martingale term, we equivalently aim to bound $\mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta_*)$. To this end, we rely on the probability of sampling optimistic parameters i.e., whose optimal average cost is smaller than the one of θ_*. Formally, let

$$\Theta^{\text{opt}} = \{\theta \in \mathbb{R}^{n(n+d)} \text{ s.t. } J(\theta) \leq J(\theta_*)\},$$

for any $\theta \in \Theta^{\text{opt}}$, one has:

$$\mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta_*) \leq \mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta) \leq |\mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta)|.$$

This implies that

$$\begin{aligned}
\mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta_*) &\leq \mathbb{E}\Big(|J(\widetilde{\theta}_t) - \mathbb{E}(J(\widetilde{\theta}_t)|\Big|\mathcal{F}_t^x\Big)\Big|\mathcal{F}_t^x, \widetilde{\theta}_t \in \Theta^{\text{opt}}\Big) \\
&\leq \mathbb{E}\Big(|J(\widetilde{\theta}_t) - \mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x)|\Big|\mathcal{F}_t^x\Big)\Big/\mathbb{P}(\widetilde{\theta}_t \in \Theta^{\text{opt}}|\mathcal{F}_t^x), \\
&= \Delta_t\Big/\mathbb{P}(\widetilde{\theta}_t \in \Theta^{\text{opt}}|\mathcal{F}_t^x).
\end{aligned}$$

Therefore, R^{TS} is bounded as soon as the probability of being optimistic is constant which is provided by Lem. 4.5.5. Intuitively, this means that optimistic samples induce controls that contribute to bound the deviation in the performance, and the constant probability of being optimistic ensures that we sample those parameters at a fixed frequency i.e., often enough.

Subsequent sampling distributions. The objective of this step is to bound the regret due to the gap at the policy switch R^{gap}. While standard approaches rely on lazy updates to control this term, this is in contrast with the functioning of TS that requires frequent updates in order to keep R^{TS} small. This motivate our new line of proof, which relies on the fact that subsequent sampling distributions are close to each other. Unfortunately, due to the $\mathcal{S}$ constraint, this result is only provided for 1-dimensional LQ system (we discuss the general case in Sec. 4.6). The objective is again to show that

$\mathbb{E}\Big(|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|\Big|\mathcal{F}_t\Big)$ is bounded by Δ_t. Supposed that $\widetilde{\theta}_{t+1}$ and $\widetilde{\theta}_t$ have the same distribution, then, by definition, $\mathbb{E}\Big(|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|\Big|\mathcal{F}_t\Big) \leq 2\Delta_t$. The idea of the proof is to show that this still holds, provided that the distributions are close enough. For sake of illustration, we only sketch the proof when $\beta_t = \beta$ here. Let $\phi_t(\theta)$ be the pdf of the sampling step without the rejection sampling procedure (which can be done by appropriate extension of the J function), one has:

$$\phi_t(\theta) = \frac{\det(V_t)^{1/2}}{\beta(2\pi)^{n(n+d)/2}} \exp\left(-\frac{1}{2\beta^2}\|\theta - \widehat{\theta}_t\|^2_{V_t}\right).$$

Noticing that $V_{t+1} \geq V_t$ and that $\det(V_{t+1}) \leq \square\ \det(V_t)$, one obtains

$$\phi_{t+1}(\theta) \leq \square\ \phi_t(\theta - \widehat{\theta}_{t+1} + \widehat{\theta}_t).$$

We use this inequality to re-write the expectation (together with some manipulation of the conditioning) and get

$$\mathbb{E}\Big(|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|\Big|\mathcal{F}_t\Big) \leq \mathbb{E}\Big(|J(\widetilde{\theta}_t - \widehat{\theta}_{t+1} + \widehat{\theta}_t) - \mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}^x_t)|\Big|\mathcal{F}^x_t\Big) + \Delta_t.$$

We conclude by using the Lipschitz property of J and Prop. 4.5.5, which guarantees that least square increments are cumulatively bounded.

4.5 Theoretical analysis

We prove the first frequentist $\sqrt{T}$ regret bound for TS in LQ systems of dimension $n = 1$ and arbitrary d. In order to isolate the steps which explicitly rely on this restriction, whenever possible we derive the proof in the general $(n + d)$-dimensional case. We discuss limitation and possible extension to the $(n + d)$-dimensional case in Sec. 4.6.

Theorem 4.5.1. *Consider the LQ system in Eq. 4.1 of dimension $n = 1$ and arbitrary d. Under Asm. 4.2.1 and 4.2.2 for any $0 < \delta < 1$, the cumulative regret of* TS *(Algorithm 4.1) over T steps is bounded w.p. at least $1 - \delta$ as* [5]

$$R(T) = \widetilde{O}\left(\sqrt{\log(1/\delta)T}\right).$$

4.5.1 Setting the stage

Filtration and sampling. To take into account the randomness of the sampling, we define the "extended" filtration $\mathcal{F}^x_t = (\mathcal{F}_{t-1}, x_t)$. Note that both $\widehat{\theta}_t$ and V_t are $\mathcal{F}^x_t$-measurable while $\widetilde{\theta}_t|\mathcal{F}^x_t$ is a random variable (since it is not conditioned on $\mathcal{F}_t$). Additionally, to handle the constraint $\mathcal{S}$, we introduce $\bar{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})$ so that $\widetilde{\theta}_t|\mathcal{F}^x_t \overset{d}{=} \mathcal{R}_{\mathcal{S}}(\bar{\theta}_t)$, where $\mathcal{R}_{\mathcal{S}}$ is a rejection sampling operator.

High-probability events. We introduce the following high probability events.

[5] Further details can be recovered from the proof.

Definition 4.5.1. *Let $\delta \in (0,1)$ and $\delta' = \delta/(8T)$ and $t \in [0,T]$. We define the confidence ellipsoids (RLS estimate concentration)*

$$\mathcal{E}_t^{\mathrm{RLS}} := \left\{\theta \in \mathbb{R}^d,\ \|\theta - \widehat{\theta}_t\|_{V_t} \leq \beta_t(\delta')\right\},$$

and (parameter $\widetilde{\theta}_t$ concentration around $\widehat{\theta}_t$)

$$\mathcal{E}_t^{\mathrm{TS}} := \left\{\theta \in \mathbb{R}^d,\ \|\theta - \widehat{\theta}_t\|_{V_t} \leq \gamma_t(\delta')\right\},$$

where $\gamma_t(\delta') = \beta_t n\sqrt{2(n+d)\log\left(2n(n+d)/\delta'\right)}$. We also introduce their associated high probability events $\widehat{E}_t = \left\{\forall s \leq t,\ \theta_* \in \mathcal{E}_s^{\mathrm{RLS}}\right\}$ and $\widetilde{E}_t = \left\{\forall s \leq t,\ \widetilde{\theta}_s \in \mathcal{E}_s^{\mathrm{TS}}\right\}$ respectively.

We also introduce a high probability event on which the states x_t are bounded almost surely.

Definition 4.5.2. *Let $\delta \in (0,1)$, X_1, X_2 be two problem-dependent positive constants and $t \in [0,T]$ and let $X = X_1 \log \frac{X_2}{\delta}$. We define the event (bounded states) $\bar{E}_t = \left\{\forall s \leq t,\ \|x_s\| \leq X\right\}$.*

Then we have that $\widehat{E} := \widehat{E}_T \subset \cdots \subset \widehat{E}_1$, $\widetilde{E} := \widetilde{E}_T \subset \cdots \subset \widetilde{E}_1$ and $\bar{E} := \bar{E}_T \subset \cdots \subset \bar{E}_1$. We show that these events do hold with high probability.

Lemma 4.5.1. $\mathbb{P}(\widehat{E} \cap \widetilde{E}) \geq 1 - \delta/4$.

Corollary 4.5.1. *On $\widehat{E} \cap \widetilde{E}$, $\mathbb{P}(\bar{E}) \geq 1 - \delta/4$. Thus, $\mathbb{P}(\widehat{E} \cap \widetilde{E} \cap \bar{E}) \geq 1 - \delta/2$.*

Lem. 4.5.1 leverages Prop. 2.3.1 and the sampling distribution to ensure that $\widehat{E} \cap \widetilde{E}$ holds w.h.p. Furthermore, Corollary 4.5.1 ensures that the states remain bounded w.h.p. on the events $\widehat{E} \cap \widetilde{E}$.[6] As a result, the proof can be derived considering that both parameters concentrate and that states are bounded, which we summarize in the sequence of events $E_t = \widehat{E}_t \cap \widetilde{E}_t \cap \bar{E}_t$, which holds with probability at least $1 - \delta/2$ for all $t \in [0,T]$.

Regret decomposition. Conditioned on the filtration $\mathcal{F}_t$ and event E_t, we have $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$, $\widetilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}$ and $\|x_t\| \leq X$. We decompose the regret and bound it on this event in line with (Abbasi-Yadkori and Szepesvári, 2011) (see details in App. 4.D).

[6] This non-trivial result is directly collected from the bounding-the-state section in (Abbasi-Yadkori and Szepesvári, 2011). While our algorithm differs, it shares the same mechanism of picking a parameter $\widetilde{\theta}_t$ within a confidence ellipsoid w.h.p., which is the core idea of the proof. Finally, TS uses the ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ instead of $\mathcal{E}_t^{\mathrm{RLS}}$ which is harmless to the proof because the scaling remains the same in terms of x_t's.

Proposition 4.5.1. *Let $\widehat{E}_t$, $\widetilde{E}_t$ and $\bar{E}_t$ be the high probability events introduced in Def. 4.5.1-4.5.2, let $E_t = \widehat{E}_t \cap \widetilde{E}_t \cap \bar{E}_t$, then*

$$\begin{aligned} R(T)\mathbb{1}\{E_T\} \leq & \sum_{t=0}^{T} \Big\{ J(\widetilde{\theta}_t) - J(\theta_*) \Big\} \mathbb{1}\{E_t\} && := R^{\mathrm{TS}} \\ & + \sum_{t=0}^{T} \Big\{ z_t^\mathsf{T} \widetilde{\theta}_t P(\widetilde{\theta}_t) \widetilde{\theta}_t^\mathsf{T} z_t - z_t^\mathsf{T} \theta_* P(\widetilde{\theta}_t) \theta_*^\mathsf{T} z_t \Big\} \mathbb{1}\{E_t\} && := R^{\mathrm{RLS}} \\ & + \sum_{t=0}^{T} \Big\{ x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t \mathbb{1}\{E_t\} - \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_{t+1}) x_{t+1} \mathbb{1}\{E_{t+1}\} | \mathcal{F}_t \Big] \Big\} && := R^{mart} \\ & + \sum_{t=0}^{T} \mathbb{E}\Big[x_{t+1}^\mathsf{T} \Big(P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t) \Big) x_{t+1} \mathbb{1}\{E_{t+1}\} | \mathcal{F}_t \Big] && := R^{Gap} \end{aligned} \tag{4.6}$$

4.5.2 Bounding the absolute deviation of the optimal value function

We prove in this subsection the main result of the proof that states that the conditional *average absolute deviation* of the performance J w.r.t. the TS distribution is cumulatively bounded. This is critical in the TS analysis: since the controls selected at each time step are based on random choices of $\widetilde{\theta}_t$, one cannot expect to control the deviation in the performance everywhere but only in expectation. Those results holds for LQ systems of arbitrary dimension n and d.

Lemma 4.5.2. *Consider the LQ system in Eq. 4.1. Under Asm. 4.2.1 and 4.2.2 for any $0 < \delta < 1$, the absolute deviation of the performance J w.r.t. the sampling of* TS *algorithm 4.1 is cumulatively bounded w.p. at least $1 - \delta/12$ as*

$$\sum_{t=0}^{T} \mathbb{E}\left(|J(\widetilde{\theta}_t) - \mathbb{E}\Big(J(\widetilde{\theta}_t) | \mathcal{F}_t^x, E_t \Big)| \Big| \mathcal{F}_t^x, E_t \right) \leq \gamma_{abs} \sqrt{T},$$

where

$$\gamma_{abs} = 16(1 + 1/\beta_0^2)(1 + C)\alpha^2 \Big[\sqrt{2(n+d)/\lambda \log\left(1 + ((1+C)\alpha)^2 T/\lambda(n+d)\right)} + \sqrt{2\log(24/\delta)} \Big],$$

$$\gamma = 4\sqrt{n(n+d)}\beta_T D n \rho / (1 - \rho^2),$$

$$\alpha = (1 + 1/\beta_0^2)\Big(\sqrt{2n\log(3n)} + (1+C)X(2S + \sqrt{n(n+d)}/\beta_0) \Big).$$

The absolute deviation is taken w.r.t the actual sample parameter $\widetilde{\theta}_t$ i.e., the one with rejection sampling. However, in bounding the gap in the policy switch R^{Gap}, we will use a slightly modified version of Lem. 4.5.2 for the parameter $\bar{\theta}_t$ without rejection sampling. This is given by the following corollary:

Corollary 4.5.2. *Consider the LQ system in Eq. 4.1. Let $\bar{J}$ be the continuous extension of J over $\mathbb{R}^{n(n+d)}$ defined by $\bar{J}(\theta) = J(\theta)\mathbb{1}_{\mathcal{S}}(\theta) + D\mathbb{1}_{\mathcal{S}^c}(\theta)$. Let $\bar{\theta}_t$ be a random variable*

defined by $\bar{\theta}_t = \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta_t$ *where each component of* η_t *is conditionally distributed as* $\mathcal{N}(0,1)$. *Then, for any* $0 < \delta < 1$, *the absolute deviation of the extended performance* $\bar{J}$ *of* TS *algorithm 4.1 w.r.t. to the random variable* $\bar{\theta}_t$ *is cumulatively bounded w.p. at least* $1 - \delta/12$ *as*

$$\sum_{t=0}^{T} \mathbb{E}\Big(|\bar{J}(\bar{\theta}_t) - \mathbb{E}\big(\bar{J}(\bar{\theta}_t)|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\text{TS}}, E_t\big)| \Big| \mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\text{TS}}, E_t \Big) \leq \gamma_{abs}\sqrt{T},$$

where

$$\gamma_{abs} = 16(1 + 1/\beta_0^2)(1 + C)\alpha^2\Big[\sqrt{2(n+d)/\lambda \log\left(1 + ((1+C)\alpha)^2 T/\lambda(n+d)\right)} + \sqrt{2\log(24/\delta)}\Big],$$

$$\gamma = 4\sqrt{n(n+d)}\beta_T D n\rho/(1-\rho^2),$$

$$\alpha = (1 + 1/\beta_0^2)\Big(\sqrt{2n\log(3n)} + (1+C)X\left(2S + \sqrt{n(n+d)}/\beta_0\right)\Big).$$

We first prove Lem. 4.5.2 and then show how it can be extended to Cor. 4.5.2 thanks to the extension of the performance function. Let

$$\Delta_t = \mathbb{E}\Big(|J(\widetilde{\theta}_t) - \mathbb{E}\big(J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\big)| \Big| \mathcal{F}_t^x, E_t \Big).$$

The proof follows four steps: **1)** we introduce a modified Poincaré inequality to show that Δ_t is bounded at each time step by a quantity which depends on ∇J; **2)** we link the gradient of the performance $\nabla J(\theta)$ to the LQ control matrix $H(\theta)$; **3)** we introduce the state x_t, making use of the fact that $H(\widetilde{\theta}_t)x_t = z_t$, to bound Δ_t by $\|z_t\|_{V_t^{-1}}$; **4)** we conclude by using Prop. 4.2.4 to bound the cumulative sum.

Step 1) Absolute deviation and gradient. Let $d' = \sqrt{n(n+d)}$, we introduce the mapping f_t from the ball $\mathcal{B}(0, d')$ to $\mathbb{R}_+$ defined in Eq. 4.7,

$$f_t(\eta) = J(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t], \tag{4.7}$$

where the restriction on the ball is here to meet the $\mathcal{E}_t^{\text{TS}}$ confidence ellipsoid of the sampling. By definition of the sampling distribution, we can rewrite Δ_t as

$$\Delta_t = \mathbb{E}_{\eta_t}\Big[|f_t(\eta_t)| \Big| \eta_t \in \mathcal{B}(0,d'), \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta_t \in \mathcal{S} \Big] \mathbb{1}\{\hat{E}_t, \bar{E}_t\}.$$

We now need to show that this formulation of the regret is related to the policy executed by TS. We prove the following result (proof in App. 4.C), that is a generic result of Sobolev's space. In particular, we use the notation of Acosta and Durán (2004) which are unrelated to the Chapter's notation.

Lemma 4.5.3. *Let* $\Omega \subset \mathbb{R}^d$ *be a convex domain with finite diameter* diam *and denote as* $W^{1,1}(\Omega)$ *the Sobolev space of order 1 in* $L^1(\Omega)$. *Let* p *be a non-negative log-concave function on* Ω *with continuous derivative up to the second order. Then, for all* $u \in W^{1,1}(\Omega)$ *such that* $\int_\Omega u(z)p(z)dz = 0$ *one has*

$$\int_\Omega |u(z)|p(z)dz \leq 2\text{diam}\int_\Omega ||\nabla u(z)||p(z)dz$$

Using Lem. 4.5.3 allows us to link Δ_t to ∇f_t and thus to ∇J since, for any η and any $\theta = \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta$, $\nabla f_t(\eta) = \beta_t V_t^{-1/2}\nabla J(\theta)$.

Step 2) From gradient to control. To obtain a bound on the norm of ∇f_t, we apply Prop. 4.A.1 (proof in App. 4.A) to get a bound on $\|\nabla J(\theta)\|_{V_t^{-1}}$ for any $\theta \in \mathcal{S}$. First, notice that

$$\|\nabla J(\theta)\|_{V_t^{-1}} \leq \sup_{\|\delta\theta\|=1} \mathrm{Tr}\Big(\delta\theta^\mathsf{T} V_t^{-1/2}\nabla J(\theta)\Big) = \sup_{\|\delta\theta\|=1} \mathrm{Tr}\Big(dP(\theta)(V_t^{-1/2}\delta\theta)\Big),$$

where the first inequality comes from the fact that the equality holds for $\delta\theta = V_t^{-1/2}\nabla J(\theta)/\|V_t^{-1/2}\nabla J(\theta)\|$ and the second equality comes directly from the definition of the differential. Then, making use of $\mathrm{Tr}(A) \leq n\|A\|_2$ for any matrix $A \in \mathbb{R}^{n\times n}$, we obtain:

$$\|\nabla J(\theta)\|_{V_t^{-1}} \leq n \sup_{\|\delta\theta\|=1} \|dP(\theta)(V_t^{-1/2}\delta\theta)\|_2.$$

We conclude using Prop. 4.A.1 which ensures that, for any $\theta \in \mathcal{S}$,

$$\forall \|\delta\theta\| = 1, \qquad \|dP(\theta)(V_t^{-1/2}\delta\theta)\|_2 \leq 2D\rho/(1-\rho^2)\|H(\theta)\|_{V_t^{-1}}.$$

As a result, we have that $\theta \in \mathcal{S}$,

$$\|\nabla J(\theta)\|_{V_t^{-1}} \leq 2D\rho/(1-\rho^2)\|H(\theta)\|_{V_t^{-1}}. \tag{4.8}$$

We are now ready to use the weighted Poincaré inequality of Lem. 4.5.3 to link the expectation of $|f_t|$ to the expectation of the norm of its gradient. From Lem. 4.2.3, we have $f_t \in W^{1,1}(\Omega)$, where $\Omega = \mathcal{B}(0, d')$ and its expectation is zero by construction. On the other hand, the rejection sampling introduces the conditioning $\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta \in \mathcal{S}$ which is unfortunately not convex ($\mathcal{S}$ is not convex). However, we can still apply Lem. 4.5.3 considering the function $\tilde{f}_t(\eta) = f_t(\eta)\mathbb{1}(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta \in \mathcal{S})$ and diameter $diam = d'$. As a result, we finally obtain,

$$\Delta_t \leq \gamma\mathbb{E}\Big[\Big\|H(\bar{\theta}_t)\Big\|_{V_t^{-1}} | \mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{S}, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}\Big], \tag{4.9}$$

where $\bar{\theta}_t = \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta_t$ and $\gamma = 4\sqrt{n(n+d)}\beta_T Dn\rho/(1-\rho^2)$.

Step 3) From gradient to actions. Recalling the definition of $H(\theta) = \begin{pmatrix} I & K(\theta)^\mathsf{T}\end{pmatrix}^\mathsf{T}$ we notice that the previous expression bound the regret Δ_t with a term involving the gain $K(\theta)$ of the optimal policy for the sampled parameter θ. This shows that the *absolute deviation* is directly related to the policies chosen by TS. To make such relationship more apparent, we now elaborate the previous expression to reveal the sequence of state-control pairs z_t induced by the policy with gain $K(\widetilde{\theta}_t)$.
By noticing that $z_t = H(\widetilde{\theta}_t)x_t$ one just needs to include the state x_t in Eq. 4.9. The intuition why this does not have a big impact on the bound is the following. The main contribution is coming from the *direction* $H(\widetilde{\theta}_t)$ that intercepts the design matrix V_t^{-1}.

On the other hand, the state x_t can be seen as an *amplitude* i.e., once the direction is chosen, the exploration is made proportionally to the state. This is true thanks to the relative independence between x_t and $\widetilde{\theta}_t$. Finally, since x_t is driven by the dynamic of Eq. 4.1, which is excited by ϵ_t, its *amplitude* is lower bounded. Moreover, this property still holds when we constrain the state to be bounded, given that the bound is large enough. This is formalized in the following property on the conditional second order moment.

Proposition 4.5.2. *Let $\bar{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})$, let x_t be the state generated by any $\mathcal{F}_t^x$-measurable sequence of control $\{u_t\}_t$, let $\alpha_t = \sqrt{2n\log(3n)} + \|\bar{x}_t\|$ where $\bar{x}_t = \mathbb{E}(x_t|\mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1})$. Then,*

$$\mathbb{E}\Big(x_t x_t^\mathsf{T} \mathbb{1}_{\{\|x_t\| \leq \alpha_t\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}\Big) \succcurlyeq \frac{1}{8(1+1/\beta_t^2)} I.$$

We prove Prop. 4.5.2 in two steps. First, we deal with the conditioning and then with the boundedness of the state.

Proposition 4.5.3. *Let $\bar{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})$, let x_t be the state generated by any $\mathcal{F}_t^x$-measurable sequence of control $\{u_t\}_t$, then,*

$$\mathbb{V}\Big(x_t|\mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}\Big) \succcurlyeq \frac{1}{1+1/\beta_t^2} I.$$

Proof. This proposition is based on the following property for Gaussian random variables: let $X \sim \mathcal{N}(\mu_x, \Sigma_x)$, $Y \sim \mathcal{N}(\mu_y, \Sigma_y)$ and $\mathbb{C}ov(X, Y) = \Sigma_{xy}$, then,

$$\mathbb{V}(X|Y) = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^\mathsf{T} \text{ and } \mathbb{E}(X|Y) = \mu_x + \Sigma_{xy}\Sigma_y^{-1}(Y - \mu_y).$$

This property still holds for matrix Gaussian distribution (by vectorization).
To exhibit the dependency, we write $\bar{\theta}_t = a_{t-1} + V_t^{-1} z_{t-1} x_t^\mathsf{T} + \beta_t V_t^{-1/2} \eta_t$ where a_{t-1}, z_{t-1} and V_t are $\mathcal{F}_{t-1}$-measurable quantities and $\eta_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0, I)$. Then, applying the Gaussian property to $X = x_t|\mathcal{F}_{t-1}$ and $Y = \bar{\theta}_t|\mathcal{F}_{t-1}$ one obtains by vectorization, re-ordering and a little bit of algebra:

$$\begin{aligned}
\mathbb{V}\Big(x_t|\mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}\Big) &= \Big(1 + \frac{1}{\beta_t^2}\|z_{t-1}\|^2_{V_t^{-1}}\Big)^{-1} I, \\
\mathbb{E}\Big(x_t|\mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}\Big) &= \mathbb{E}(x_t|\mathcal{F}_{t-1}, E_{t-1}) \\
&\quad + \Big(\bar{\theta}_t - \mathbb{E}(\bar{\theta}_t|\mathcal{F}_{t-1}, E_{t-1})\Big)^\mathsf{T} \Big(V_t^{-1} z_{t-1} z_{t-1}^\mathsf{T} V_t^{-1} + \beta_t^2 V_t^{-1}\Big)^{-1} V_t^{-1} z_{t-1}.
\end{aligned} \tag{4.10}$$

Finally, since $\|z_{t-1}\|^2_{V_t^{-1}} = z_{t-1}^\mathsf{T}(V_{t-1} + z_{t-1}z_{t-1}^\mathsf{T})^{-1} z_{t-1} = \frac{\|z_{t-1}\|^2_{V_{t-1}^{-1}}}{1+\|z_{t-1}\|^2_{V_{t-1}^{-1}}} \leq 1$, one obtains the desired result. □

Proposition 4.5.4. *Let $\bar{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})$, let x_t be the state generated by any $\mathcal{F}_t^x$-measurable sequence of control $\{u_t\}_t$, let $\alpha_t = \sqrt{2n\log(3n)} + \|\bar{x}_t\|$ where $\bar{x}_t = \mathbb{E}(x_t|\mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1})$, then,*

$$\mathbb{E}\Big(x_t x_t^\mathsf{T} \mathbb{1}_{\{\|x_t\| \leq \alpha_t\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}\Big) \succcurlyeq 1/8 \mathbb{V}(x_t|\mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}).$$

The proof of Prop. 4.5.4 relies on properties of truncated Gaussian random variable. The main ingredient is that, if the truncation takes place far enough from the mean, the Gaussian properties are preserved and the second order moment is greater than the variance. We postpone the proof to App. 4.D.

Thanks to Prop. 4.5.2, one has

$$\begin{aligned}\|H(\bar{\theta}_t)\|_{V_t^{-1}} &\leq 8(1+1/\beta_t^2)\|H(\bar{\theta}_t)\mathbb{E}\Big(x_t x_t^\mathsf{T}\mathbb{1}_{\{\|x_t\|\leq\alpha_t\}}|\mathcal{F}_{t-1},\bar{\theta}_t,E_{t-1}\Big)\|_{V_t^{-1}}\\ &\leq 8(1+1/\beta_t^2)\mathbb{E}\bigg(\|H(\bar{\theta}_t)x_t x_t^\mathsf{T}\|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha_t\}}|\mathcal{F}_{t-1},\bar{\theta}_t,E_{t-1}\bigg),\end{aligned}$$

which, plugged in Eq. 4.9 leads to

$$\begin{aligned}\Delta_t &\leq \gamma 8(1+1/\beta_t^2)\mathbb{E}\bigg[\mathbb{E}\bigg(\|H(\bar{\theta}_t)x_t x_t^\mathsf{T}\|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha_t\}}|\mathcal{F}_{t-1},\bar{\theta}_t,E_{t-1}\bigg)\bigg|\mathcal{F}_t^x,\bar{\theta}_t\in\mathcal{S},\bar{\theta}_t\in\mathcal{E}_t^{\mathrm{TS}}\bigg]\\ &\leq \gamma 8(1+1/\beta_0^2)\mathbb{E}\bigg[\mathbb{E}\bigg(\|H(\bar{\theta}_t)x_t x_t^\mathsf{T}\|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha_t\}}|\mathcal{F}_{t-1},\bar{\theta}_t,E_{t-1}\bigg)\bigg|\mathcal{F}_t^x,\bar{\theta}_t\in\mathcal{S},\bar{\theta}_t\in\mathcal{E}_t^{\mathrm{TS}}\bigg].\end{aligned}$$

Plugging $\bar{\theta}_t = a_{t-1} + V_t^{-1}z_{t-1}x_t^\mathsf{T} + \beta_t V_t^{-1/2}\eta_t$ into Eq. 4.10, one can obtain the following bound for α_t, conditioned on $\eta_t\in\mathcal{B}(0,d')$:

$$\begin{aligned}\alpha_t &\leq \sqrt{2n\log(3n)} + \|\mathbb{E}(x_t|\mathcal{F}_{t-1},\bar{\theta}_t,E_{t-1})\|,\\ &\leq \sqrt{2n\log(3n)} + 2(1+C)SX + 1/\beta_t d'(1+C)X + (1-\frac{1}{1+\|z_{t-1}\|^2_{V_t^{-1}}/\beta_t^2})\|x_t\|.\end{aligned}$$

Thus, making use of $\|z_{t-1}\|^2_{V_t^{-1}}\leq 1$, one has

$$\begin{aligned}\|x_t\|\leq\alpha_t \quad\Longrightarrow\quad \|x_t\| &\leq (1+1/\beta_t^2)(\sqrt{2n\log(3n)} + (1+C)X(2S+d'/\beta_t))\\ &\leq (1+1/\beta_0^2)(\sqrt{2n\log(3n)} + (1+C)X(2S+d'/\beta_0)).\end{aligned}$$

Let $\alpha = (1+1/\beta_0^2)(\sqrt{2n\log(3n)} + (1+C)X(2S+d'/\beta_0))$, one obtains:

$$\begin{aligned}\Delta_t &\leq \gamma 8(1+1/\beta_0^2)\mathbb{E}\bigg[\mathbb{E}\bigg(\|H(\bar{\theta}_t)x_t x_t^\mathsf{T}\ \|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha_t\}}|\mathcal{F}_{t-1},\bar{\theta}_t,E_{t-1}\bigg)\bigg|\mathcal{F}_t^x,\bar{\theta}_t\in\mathcal{S},\bar{\theta}_t\in\mathcal{E}_t^{\mathrm{TS}}\bigg],\\ &\leq \gamma 8(1+1/\beta_0^2)\alpha\mathbb{E}\bigg[\mathbb{E}\bigg(\|H(\bar{\theta}_t)x_t\|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha\}}|\mathcal{F}_{t-1},\bar{\theta}_t,E_{t-1}\bigg)\bigg|\mathcal{F}_t^x,\bar{\theta}_t\in\mathcal{S},\bar{\theta}_t\in\mathcal{E}_t^{\mathrm{TS}}\bigg],\\ &\leq \gamma 8(1+1/\beta_0^2)\alpha\mathbb{E}\bigg[\mathbb{E}\bigg(\|H(\widetilde{\theta}_t)x_t\|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha\}}|\mathcal{F}_{t-1},\widetilde{\theta}_t,E_{t-1}\bigg)\bigg|\mathcal{F}_t^x,\widetilde{\theta}_t\in\mathcal{E}_t^{\mathrm{TS}}\bigg],\\ &\leq \gamma 8(1+1/\beta_0^2)\alpha\mathbb{E}\bigg[\mathbb{E}\bigg(\|z_t\|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha\}}|\mathcal{F}_{t-1},\widetilde{\theta}_t,E_{t-1}\bigg)\bigg|\mathcal{F}_t^x,\widetilde{\theta}_t\in\mathcal{E}_t^{\mathrm{TS}}\bigg],\\ &\leq \gamma 16(1+1/\beta_0^2)\alpha\mathbb{E}\bigg[\mathbb{E}\bigg(\|z_t\|_{V_t^{-1}}\mathbb{1}_{\{\|x_t\|\leq\alpha\}}|\mathcal{F}_{t-1},\widetilde{\theta}_t,E_{t-1}\bigg)\bigg|\mathcal{F}_t^x\bigg].\end{aligned}$$

where we used that $\widetilde{\theta}_t \overset{d}{=} \bar{\theta}_t|\mathcal{S}$ from line 2 to line 3, that $H(\widetilde{\theta}_t)x_t = z_t$ from line 3 to line 4, and that $\mathbb{P}(\widetilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}|\mathcal{F}_t^x) \geq \mathbb{P}(\widetilde{E}_t) \geq 1/2$ to discard the conditioning from line 4 to line 5.

Step 4) Bounding the cumulative sum. Let

$$Y_t = \mathbb{E}\Big[\mathbb{E}\Big(\|z_t\|_{V_t^{-1}} \mathbb{1}_{\{\|x_t\| \leq \alpha\}} | \mathcal{F}_{t-1}, \widetilde{\theta}_t, E_{t-1}\Big)\Big|\mathcal{F}_t^x\Big],$$

summing the previous bound over T leads to:

$$\sum_{t=0}^{T} \Delta_t \leq 16(1+1/\beta_0^2)\alpha \sum_{t=0}^{T} Y_t, \quad \text{where} \quad \alpha = (1+1/\beta_0^2)(\sqrt{2n\log(3n)}+(1+C)X(2S+d'/\beta_0)).$$

Noticing that $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = \mathbb{E}(\|z_t\|_{V_t^{-1}} \mathbb{1}_{\{\|x_t\| \leq \alpha\}}|\mathcal{F}_{t-1})$, by the law of iterated expectation, $\Big\{Y_t - \|z_t\|_{V_t^{-1}} \mathbb{1}_{\{\|x_t\| \leq \alpha\}}\Big\}_{t \geq 1}$ is a $\mathcal{F}_t$-martingale difference sequence, bounded almost surely at each time step by $(1+C)\alpha$. Hence, using Azuma's inequality, with probability at least $1 - \delta/12$,

$$\begin{aligned}
\sum_{t=0}^{T} Y_t &\leq \sum_{t=0}^{T} \|z_t\|_{V_t^{-1}} \mathbb{1}_{\{\|x_t\| \leq \alpha\}} + (1+C)\alpha\sqrt{2T\log(24/\delta)}, \\
&\leq \sum_{t=0}^{T} \|z_t\|_{V_t^{-1}} + (1+C)\alpha\sqrt{2T\log(24/\delta)}, \\
&\leq (1+C)\alpha\sqrt{2(n+d)T/\lambda \log\Big(1 + ((1+C)\alpha)^2 \frac{T}{\lambda(n+d)}\Big)} + (1+C)\alpha\sqrt{2T\log(24/\delta)}
\end{aligned}$$

where we used Cauchy-Schwarz inequality and Prop. 4.2.4 in the last inequality. This ends the proof of Lem. 4.5.2.

Proof of Cor. 4.5.2. We conclude this section by proving Cor. 4.5.2, showing that

$$\bar{\Delta}_t = \mathbb{E}\Big(|\bar{J}(\bar{\theta}_t) - \mathbb{E}\Big(\bar{J}(\bar{\theta}_t)|\mathcal{F}_t^x, E_t\Big)|\Big|\mathcal{F}_t^x, E_t\Big)$$

is upper bounded as Δ_t. The difference lies in the fact that no rejection sampling is considered here (i.e., $\bar{\theta}_t$ versus $\widetilde{\theta}_t$) but the optimal value function J is extended to $\bar{J}$ in a constant fashion (i.e., $\bar{J}(\theta) = J(\theta)\mathbb{1}_{\mathcal{S}}(\theta) + D\mathbb{1}_{\mathcal{S}^c}(\theta)$).
The proof is similar to the one of Lem. 4.5.2, the only difference lying in Step 1. Since $\bar{J}$ is a continuous extension of J, differentiable almost everywhere, we can apply the modified Poincaré inequality, which leads to

$$\bar{\Delta}_t \leq 2d'\beta_t \mathbb{E}\Big[\|\nabla \bar{J}(\bar{\theta}_t)\|_{V_t^{-1}}|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}\Big].$$

However, by definition $\nabla\bar{J}(\theta) = \nabla J(\theta)\mathbb{1}_{\mathcal{S}}(\theta)$. Therefore,

$$\begin{aligned}
\bar{\Delta}_t &\leq 2d'\beta_t \mathbb{E}\Big[\|\nabla J(\bar{\theta}_t)\|_{V_t^{-1}} \mathbb{1}_{\mathcal{S}}(\bar{\theta}_t)|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}\Big], \\
&\leq 2d'\beta_t \mathbb{E}\Big[\|\nabla J(\bar{\theta}_t)\|_{V_t^{-1}}|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}, \bar{\theta}_t \in \mathcal{S}\Big].
\end{aligned}$$

Finally, using Eq. 4.8, we obtain $\bar{\Delta}_t \leq \gamma\mathbb{E}\Big[\Big\|H(\bar{\theta}_t)\Big\|_{V_t^{-1}}|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{S}, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}\Big]$, which is identical to Eq. 4.9 so the rest of the proof follows.

4.5.3 Bounding the optimality regret $R^{\mathbf{TS}}$

We prove in this section a $\sqrt{T}$ bound for the regret term R^{TS}. The idea of the proof is to show that, thanks to the constant probability of being optimistic (see Lem. 4.5.5), one can link R^{TS} to the *average absolute deviation* of the optimal value function and thus bound it with high probability according to Lem. 4.5.2. Unfortunately, since Lem. 4.5.5 only holds when $n = 1$, this restricts the result to the 1d case. However, we believe that this comes mainly from technical difficulties, and that both the ideas and the structure of the proof should hold in any dimension. We discuss its extension in Sec. 4.6.

Lemma 4.5.4. *Consider the LQ system in Eq. 4.1 of dimension $n = 1$ and arbitrary d. Under Asm. 4.2.1 and 4.2.2, for any $0 < \delta < 1$, the regret R^{TS} incurred by running the* TS *algorithm 4.1 is cumulatively bounded w.p. at least $1 - \delta/6$ as*

$$R^{\mathrm{TS}} = \sum_{t=0}^{T} \{J(\widetilde{\theta}_t) - J(\theta_*)\} \mathbb{1}\{E_t\} \leq \gamma_{\mathrm{TS}} \sqrt{T}$$

where $\gamma_{\mathrm{TS}} = 2D\sqrt{2\log(24/\delta)} + \frac{\gamma_{abs}}{p}$.

The structure of the proof is the following: **1)** we first decompose the regret, introducing the average performance $\mathbb{E}\Big[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\Big]$, and bound the martingale part of it, **2)** we show that the probability of being optimistic is constant and we link the remaining term to the *average absolute deviation* of the performance function, **3)** we conclude by using Lem. 4.5.2.

$R^{\mathbf{TS}}$ decomposition. Let $R_t^{\mathrm{TS}} := \{J(\widetilde{\theta}_t) - J(\theta_*)\}\mathbb{1}\{E_t\}$. Introducing $\mathbb{E}\Big[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\Big]$, one can split R^{TS} as

$$R^{\mathrm{TS}} = \sum_{t=0}^{T} R_t^{\mathrm{TS},1} + \sum_{t=0}^{T} R_t^{\mathrm{TS},2} \quad \text{where} \quad R_t^{\mathrm{TS},1} := \Big\{J(\widetilde{\theta}_t) - \mathbb{E}\Big[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\Big]\Big\}\mathbb{1}\{E_t\},$$
$$R_t^{\mathrm{TS},2} := \Big\{\mathbb{E}\Big[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\Big] - J(\theta_*)\Big\}\mathbb{1}\{E_t\}.$$

By definition, $\{R_t^{\mathrm{TS},1}\}_{t\geq 1}$ is a $\mathcal{F}_t^x$-martingale difference sequence, bounded almost surely at each time step by $2D$ (thanks to Def. 4.2.1). Therefore, applying Azuma's inequality guarantees that, w.p. at least $1 - \delta/12$,

$$\sum_{t=0}^{T} R_t^{\mathrm{TS},1} = \sum_{t=0}^{T} \Big\{J(\widetilde{\theta}_t) - \mathbb{E}\Big[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t\Big]\Big\}\mathbb{1}_{E_t} \leq 2D\sqrt{2T\log(24/\delta)}.$$

Optimism and expectation. We now focus on the second term $R_t^{\mathrm{TS},2}$. Let

$$\Theta^{\mathrm{opt}} = \{\theta : J(\theta) \leq J(\theta_*)\}$$

be the set of optimistic parameters (i.e., LQ systems whose optimal average expected cost is lower than the true one). Then, for any $\theta \in \Theta^{\mathrm{opt}}$, the per-step regret $R_t^{\mathrm{TS},2}$ is bounded by:

$$R_t^{\mathrm{TS},2} \leq \Big(\mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t] - J(\theta)\Big)\mathbb{1}\{E_t\} \leq \Big|J(\theta) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\Big|\mathbb{1}\{E_t\},$$

which implies that, for any random variable $\widetilde{\theta}$,

$$R_t^{\mathrm{TS},2} \leq \mathbb{E}\Big[\Big|J(\widetilde{\theta}) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\Big|\mathcal{F}_t^x, E_t, \widetilde{\theta} \in \Theta^{\mathrm{opt}}\Big],$$

where we use first the definition of the optimistic parameter set and bound the resulting quantity by its absolute value. Since this inequality holds for any optimistic parameter, it still holds in expectation, conditioned on $\widetilde{\theta} \in \Theta^{\mathrm{opt}}$. While the last inequality is true for any sampling distribution, it is convenient to select it equivalent to the sampling distribution of TS. Thus, we set $\widetilde{\theta} = \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta)$ with η is component wise Gaussian $\mathcal{N}(0,1)$ and obtain

$$\begin{aligned} R_t^{\mathrm{TS},2} &\leq \mathbb{E}\Big[\Big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\Big|\mathcal{F}_t^x, E_t, \widetilde{\theta}_t \in \Theta^{\mathrm{opt}}\Big], \\ &= \mathbb{E}\Big[\Big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\mathbb{1}\{\Theta^{\mathrm{opt}}\}\Big|\mathcal{F}_t^x, E_t\Big] \Big/ \mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{\mathrm{opt}} \mid \mathcal{F}_t^x, E_t\Big) \qquad (4.11) \\ &\leq \mathbb{E}\Big[\Big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\Big|\mathcal{F}_t^x, E_t\Big] \Big/ \mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{\mathrm{opt}} \mid \mathcal{F}_t^x, E_t\Big). \end{aligned}$$

Probability of being optimistic. At this point we need to show that the probability of sampling an optimistic parameter $\widetilde{\theta}_t$ is constant at any step t. This is provided by the following lemma (proof in App. 4.B).

Lemma 4.5.5. *Let $\Theta^{opt} := \{\theta \in \mathbb{R}^d \mid J(\theta) \leq J(\theta^\star)\}$ be the set of optimistic parameters and $\widetilde{\theta}_t = \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta)$ with η be component-wise normal $\mathcal{N}(0,1)$, then in the one-dimensional case ($n{=}1$ and $d{=}1$)*

$$\forall t \geq 0,\ \mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{opt} \mid \mathcal{F}_t^x, E_t\Big) \geq p,$$

where p is a strictly positive constant.

Summing up. Integrating the result of Lem. 4.5.5 into Eq. 4.11 gives

$$R_t^{\mathrm{TS},2} \leq \frac{1}{p}\mathbb{E}\Big[\Big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\Big| \mid \mathcal{F}_t^x, E_t\Big]. \qquad (4.12)$$

The most interesting aspect of this result is that the constant probability of being optimistic allows us to bound the worst-case non-stochastic quantity $\mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x] - J(\theta_*)$ depending on $J(\theta_*)$ by an expectation $\mathbb{E}\Big[\Big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x]\Big| \mid \mathcal{F}_t^x\Big]$ up to a multiplicative constant (we drop the events E for notational convenience). The last term is the conditional *average absolute deviation* of the performance J w.r.t. the TS distribution. This connection provides a major insight about the functioning of TS, since it shows that TS does not need to have an accurate estimate of θ_* but it should rather reduce the estimation errors of θ_* only on the directions that may translate in larger errors in estimating the objective function J.

From Eq. 4.12, and applying Lem. 4.5.2, we obtain, by Azuma, that with probability at least $1-\delta/12$,

$$\sum_{t=0}^{T} R_t^{\mathrm{TS},2} \leq 1/p \sum_{t=0}^{T} \Delta_t \leq \frac{\gamma_{abs}}{p}\sqrt{T}.$$

Moreover, since with probability at least $1-\delta/12$, one has $\sum_{t=1}^{T} R_t^{\mathrm{TS},1} \leq 2D\sqrt{2T\log(24/\delta)}$, a union bound argument guarantees that, with probability $1-\delta/6$,

$$R^{\mathrm{TS}} \leq 2D\sqrt{2T\log(24/\delta)} + \frac{\gamma_{abs}}{p}\sqrt{T}.$$

4.5.4 Bounding the gap at policy switch R^{gap}

We derive here a $\sqrt{T}$ bound for the regret term R^{gap} which takes into account the deviation in the performance between two subsequent sampling $P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)$. The standard approach to control this term is to modify the algorithm by keeping the policy constant over episodes and performing policy update from time to time. Such approach offers two advantages: from a computational point of view, the fewest policy are updated, the cheapest (see Abbasi-Yadkori and Szepesvári 2011 for a lazy update scheme in the LQ setting for OFUL algorithm); from a technical point of view, it avoids quantifying and bounding the gap at policy switch.
However, this technique does not suit well with the TS algorithm. We first show that TS has to solve a trade-off between frequently updating the policy to guarantee enough optimistic samples (and hence bound R^{TS}) and reducing the number of policy switches to limit the regret incurred at each change (and hence bound R^{gap}). This gives rise to a final bound of $O\left(T^{2/3}\right)$. Then, we derive a new line proof that overcomes the trade-off of frequent versus lazy policy updates and thus leads to a $\sqrt{T}$ bound.

Trading-off frequent and lazy updates. We discuss in this paragraph the issue that arises when using lazy updates. Consider the modification of the TS algorithm 4.1 where the policy is kept constant over an episode. Denote as $\{t_k\}_{k=1,\cdots,K}$ the time steps at which the policy switch occurs, K the number of switches and $\tau_k = t_{k+1} - t_k + 1$ the length of each episode. As a result, in line with (Abbasi-Yadkori and Szepesvári, 2011), the regret term R^{gap} is directly bounded by $2DX^2K$.
However, the bound of the optimality regret is modified accordingly: since the policy is kept constant over episode, R^{TS} becomes

$$\begin{aligned} R^{\mathrm{TS}} &= \sum_{k=1}^{K} \tau_k R_{t_k}^{\mathrm{TS}} = \sum_{k=1}^{K} \tau_k R_{t_k}^{\mathrm{TS},1} + \sum_{k=1}^{K} \tau_k R_{t_k}^{\mathrm{TS},2}, \\ &\leq \sup_k \tau_k \left(\sum_{k=1}^{K} R_{t_k}^{\mathrm{TS},1} + \sum_{k=1}^{K} R_{t_k}^{\mathrm{TS},2} \right). \end{aligned}$$

Following the same proof as in Sec. 4.5.3, one obtains $\sum_{k=1}^{K} R_{t_k}^{\mathrm{TS},1} + \sum_{k=1}^{K} R_{t_k}^{\mathrm{TS},2} \leq \square\sqrt{K}$, where $\square$ is the appropriate numerical constant, and $\sqrt{K}$ comes directly from the use of Azuma's inequality.

Finally, consider for sake of simplicity the case where the length of episode is constant (e.g. $\tau_k = \tau = T/K$ for all $k \leq K$), on obtains:

$$R^{\mathrm{TS}} + R^{gap} \leq \square\Big(\tau\sqrt{K} + K\Big) \leq \square\Big(T/\sqrt{K} + K\Big),$$

which is minimized by $K = T^{2/3}$ and leads to a $\widetilde{O}(T^{2/3})$ overall regret. This synthesizes the trade-off faced by TS: by nature, thanks to randomization, the algorithm selects *on average* useful policies which contribute to control the regret. Thus the more it samples, the closer it gets to this *average behavior*. This speaks in favor of frequent updates. On the other hand, the available bound for R^{gap} scales linearly with the number of policy updates, and hence is not compatible with frequent updates.To overcome this issue, we derive a new proof for R^{gap} that allows us to change the policy at each time step without scaling linearly with the number of update. We prove the following result:

Lemma 4.5.6. *Consider the LQ system in Eq. 4.1 of dimension $n = 1$ and arbitrary d. Under Asm. 4.2.1 and 4.2.2, for any $0 < \delta < 1$, the regret R^{gap} incurred by running the* TS *algorithm 4.1 is cumulatively bounded w.p. at least $1 - \delta/6$ as*

$$R^{gap} = \sum_{t=0}^{T} \mathbb{E}\Big[x_{t+1}^{\mathsf{T}}\Big(P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\Big)x_{t+1}\mathbb{1}\{E_{t+1}\}|\mathcal{F}_t\Big] \leq \gamma_{gap}\sqrt{T} + \gamma'_{gap}$$

$$\textit{where} \quad \begin{cases} \gamma_{gap} := 4D\sqrt{2\log(24/\delta)} + X^2(1 + 2\gamma_{gap,2})\gamma_{abs}/p + 2X^2\gamma_{gap,1}\gamma_{lip}\gamma_{gap,3}/p, \\ \gamma'_{gap} := X^2\gamma_{gap,2}D\delta/(2p) + \dfrac{n^2}{p\lambda}\gamma_{gap,3}^2, \\ \gamma_{lip} := 2D\rho/(1-\rho^2)(1+C), \\ \gamma_{gap,1} := \beta_T(1+C)X + \sqrt{n}, \\ \gamma_{gap,2} := \Big(1 + \dfrac{(1+C)X}{\lambda}\Big)^{1/2}, \\ \gamma_{gap,3} := (1+C)X\sqrt{2(n+d)/\lambda \log\Big(1 + ((1+C)X)^2T/\lambda(n+d)\Big)}. \end{cases}$$

As for Lem. 4.5.2, this result only holds in the 1d case. This is due to the rejection sampling procedure that introduces boundary issue (w.r.t. the change in subsequent sampling distribution) and requires the use of the extended optimal value function $\bar{J}$. However, we believe that this comes mainly from technical difficulties, and that both the ideas and the structure of the proof should hold in any dimension. We discuss its extension in Sec. 4.6.

The structure of the proof is the following: **1)** we decompose the regret and show that the probability of sampling an admissible parameter (i.e., $\theta \in \mathcal{S}$) is constant. This allows us to replace the actual TS sampling distribution by the unconstrained one (i.e., we replace $\widetilde{\theta}_t$ by $\bar{\theta}_t$). **2)** We show that subsequent sampling distributions change slowly and make use of the Lipschitz property of J to link R^{gap} to the *average absolute deviation* of the extended optimal value function. **3)** We conclude by using Cor. 4.5.2.

Regret decomposition. We consider here the 1-d case with $n = 1$. As a result, $P(\theta) = J(\theta)$ and R^{gap} can be re-written as

$$\begin{aligned} R^{gap} &= \sum_{t=0}^{T} \mathbb{E}\Big[x_{t+1}^{\mathsf{T}}\Big(P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\Big)x_{t+1}\mathbb{1}\{E_{t+1}\}\Big|\mathcal{F}_t\Big], \\ &\leq X^2 \sum_{t=0}^{T} \mathbb{E}\Big[\|P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\|_2 \mathbb{1}\{E_{t+1}\}\Big|\mathcal{F}_t\Big], \\ &\leq X^2 \sum_{t=0}^{T} \mathbb{E}\Big[|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|\mathbb{1}\{E_{t+1}\}\Big|\mathcal{F}_t\Big]. \end{aligned}$$

Applying Azuma's inequality to the $\mathcal{F}_t$-martingale difference sequence

$$\Big\{\mathbb{E}\Big[|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|\mathbb{1}\{E_{t+1}\}\Big|\mathcal{F}_t^x\Big] - \mathbb{E}\Big[|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|\mathbb{1}\{E_{t+1}\}\Big|\mathcal{F}_t\Big]\Big\}_{t\geq 1}$$

implies that, w.p. at least $1 - \delta/12$,

$$R^{gap} \leq X^2 \sum_{t=0}^{T} \mathbb{E}\Big[|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|\mathbb{1}\{E_{t+1}\}\Big|\mathcal{F}_t^x\Big] + 4D\sqrt{2T\log(24/\delta)}.$$

Let $\bar{J} = J(\theta)\mathbb{1}_{\mathcal{S}}(\theta) + D\mathbb{1}_{\mathcal{S}^c}(\theta)$ be the continuous extension of the optimal value function. Let $\bar{J}_t = \mathbb{E}\Big(\bar{J}(\bar{\theta}_t)|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}, E_t\Big)$ where $\bar{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})$ then, w.p. at least $1-\delta/12$,

$$\begin{aligned} R^{gap} &\leq 4D\sqrt{2T\log(24/\delta)} + X^2 \sum_{t=0}^{T} R_t^{gap,1} + X^2 \sum_{t=0}^{T} R_t^{gap,2}, \\ \text{where} \quad & R_t^{gap,1} := \mathbb{E}\Big[|J(\widetilde{\theta}_t) - \bar{J}_t||\mathcal{F}_t^x, E_t\Big], \\ & R_t^{gap,2} := \mathbb{E}\Big[|J(\widetilde{\theta}_{t+1}) - \bar{J}_t||\mathcal{F}_t^x, E_{t+1}\Big], \end{aligned} \tag{4.13}$$

where we used that $\mathbb{1}\{E_{t+1}\} \leq \mathbb{1}\{E_t\}$, and pushed the events in the conditioning to obtain the expressions of $R_t^{gap,1}$ and $R_t^{gap,2}$.

Removing the $\mathcal{S}$ constraint. Let $\bar{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})$. Given that on E_t, $\widetilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}$ and that $\widetilde{\theta}_t \overset{d}{=} \bar{\theta}_t|\mathcal{S}$ for all $t \geq 1$, the conditioning $\widetilde{\theta}_t \in E_t$ implies that $\bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}} \cap \mathcal{S}$. Since, on $\mathcal{S}$, $J(\theta) = \bar{J}(\theta)$, one has

$$\begin{aligned} R_t^{gap,1} &= \mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}} \cap \mathcal{S}, E_t\Big], \\ R_t^{gap,2} &= \mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1}) - \bar{J}_t|\Big|\mathcal{F}_t^x, \bar{\theta}_{t+1} \in \mathcal{E}_{t+1}^{\mathrm{TS}} \cap \mathcal{S}, E_{t+1}\Big]. \end{aligned}$$

We now rely on the following corollary to handle the rejection sampling.

Corollary 4.5.3. *Let $\bar{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})$, then $\mathbb{P}(\bar{\theta}_t \in \mathcal{S}|\mathcal{F}_t^x, \bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}, E_t) \geq p$ where p is the same constant as in Lem. 4.5.5.*

Proof. The proof relies on Lem. 4.5.5, as in the worst-case configuration, the set of admissible parameters may coincide with the set of optimistic parameters (it can happen if $D = J(\theta_*)$). Formally, one just need to notice that, by definition of the

admissible parameter set, $\forall\theta\in\mathcal{S}$, $J(\theta)\leq D$. Moreover $\theta_*\in\mathcal{S}$. Therefore, $\forall\theta\in\Theta^{\text{opt}}$, $J(\theta)\leq J(\theta_*)\leq D$ and $\Theta^{\text{opt}}\subset\mathcal{S}$. As a consequence,

$$\begin{aligned}\mathbb{P}(\bar{\theta}_t\in\mathcal{S}|\mathcal{F}_t^x,\bar{\theta}_t\in\mathcal{E}_t^{\text{TS}},E_t)&\geq\mathbb{P}(\bar{\theta}_t\in\Theta^{\text{opt}}|\mathcal{F}_t^x,\bar{\theta}_t\in\mathcal{E}_t^{\text{TS}},E_t),\\&\geq\mathbb{P}(\widetilde{\theta}_t\in\Theta^{\text{opt}}|\mathcal{F}_t^x,E_t)\geq p.\end{aligned}$$

□

Making use of Cor. 4.5.3, we get rid of the admissible constraint set

$$\begin{aligned}R_t^{gap,1}&\leq\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t)-\bar{J}_t|\mathbb{1}\{\mathcal{S}\}\Big|\mathcal{F}_t^x,\mathcal{E}_t^{\text{TS}},E_t\Big]\Big/\mathbb{P}(\bar{\theta}_t\in\mathcal{S}\Big|\mathcal{F}_t^x,\mathcal{E}_t^{\text{TS}},E_t)\\&\leq 1/p\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t)-\bar{J}_t|\mathbb{1}\{\mathcal{S}\}\Big|\mathcal{F}_t^x,\mathcal{E}_t^{\text{TS}},E_t\Big]\\R_t^{gap,2}&=\mathbb{E}\bigg(\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1})-\bar{J}_t|\Big|\mathcal{F}_{t+1}^x,\mathcal{E}_{t+1}^{\text{TS}},E_{t+1},\mathcal{S}\Big]\Big|\mathcal{F}_t^x,\mathcal{E}_{t+1}^{\text{TS}},E_{t+1}\bigg)\\&\leq\mathbb{E}\bigg(\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1})-\bar{J}_t|\mathbb{1}\{\mathcal{S}\}\Big|\mathcal{F}_{t+1}^x,\mathcal{E}_{t+1}^{\text{TS}},E_{t+1}\Big]\Big/\mathbb{P}(\bar{\theta}_{t+1}\in\mathcal{S}\Big|\mathcal{F}_{t+1}^x,\mathcal{E}_{t+1}^{\text{TS}},E_{t+1})\bigg|\mathcal{F}_t^x,\mathcal{E}_{t+1}^{\text{TS}},E_{t+1}\bigg)\\&\leq 1/p\mathbb{E}\bigg(\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1})-\bar{J}_t|\mathbb{1}\{\mathcal{S}\}\Big|\mathcal{F}_{t+1}^x,\mathcal{E}_{t+1}^{\text{TS}},E_{t+1}\Big]\bigg|\mathcal{F}_t^x,\mathcal{E}_{t+1}^{\text{TS}},E_{t+1}\bigg)\end{aligned}$$

$$\implies\quad\begin{cases}R_t^{gap,1}\leq 1/p\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t)-\bar{J}_t|\Big|\mathcal{F}_t^x,\bar{\theta}_t\in\mathcal{E}_t^{\text{TS}},E_t\Big]\\R_t^{gap,2}\leq 2/p\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1})-\bar{J}_t|\Big|\mathcal{F}_t^x,\bar{\theta}_{t+1}\in\mathcal{E}_{t+1}^{\text{TS}},E_t\Big]\end{cases}\tag{4.14}$$

where we replaced E_{t+1} by E_t in the last expression, making use of the fact that $\mathbb{P}(E_{t+1}\cap E_t^c|\mathcal{F}_t^x)\geq 1-\delta/8T\geq 1/2$.
Thanks to Cor. 4.5.2, one directly obtains a bound for $R_t^{gap,1}$. The objective of next step is to show that $R_t^{gap,2}$ is closed to $R_t^{gap,1}$ because subsequent sampling distribution changes slowly.

Bounding $R_t^{gap,2}$. This step relies on the following propositions.

Proposition 4.5.5. *Let $\{\widehat{\theta}_t\}_{t\geq 1}$ be the sequence of least square estimate, $\forall t\geq 1$,*

$$\mathbb{E}\Big[\|\widehat{\theta}_{t+1}-\widehat{\theta}_t\|\Big|\mathcal{F}_t^x,E_t\Big]\leq\gamma_{gap,1}\|z_t\|_{V_t^{-1}}\quad\textit{where}\quad\gamma_{gap,1}:=\Big[(1+C)X+\sqrt{n}\Big]$$

Proof. Least Square updates can be written explicitly as $\widehat{\theta}_{t+1}=\widehat{\theta}_t+\frac{V_t^{-1/2}z_t}{1+\|z_t\|^2_{V_t^{-1}}}(x_{t+1}-\widehat{\theta}_t^{\mathsf{T}}z_t)$.
Thus,

$$\|\widehat{\theta}_{t+1}-\widehat{\theta}_t\|\leq\frac{\|z_t\|_{V_t^{-1}}}{1+\|z_t\|^2_{V_t^{-1}}}\|\theta_*^{\mathsf{T}}z_t-\widehat{\theta}_t^{\mathsf{T}}z_t+\epsilon_{t+1}\|\leq\|\widehat{\theta}_{t+1}-\widehat{\theta}_t\|\leq\frac{\|z_t\|_{V_t^{-1}}}{1+\|z_t\|^2_{V_t^{-1}}}(\beta_t\|z_t\|_{V_t^{-1}}+\|\epsilon_{t+1}\|)$$

$$\mathbb{E}\Big[\|\widehat{\theta}_{t+1}-\widehat{\theta}_t\|\Big|\mathcal{F}_t^x,E_t\Big]\leq\frac{\|z_t\|_{V_t^{-1}}}{1+\|z_t\|^2_{V_t^{-1}}}(\beta_t\|z_t\|_{V_t^{-1}}+\sqrt{n})\leq\|z_t\|_{V_t^{-1}}\Big[\beta_t(1+C)X+\sqrt{n}\Big]$$

□

Proposition 4.5.6. *Let $\{\beta_t\}_{t\geq 1} = \{\beta_t(\delta')\}_{t\geq 1}$ be the sequence defined in Eq. 4.4, then, $\forall t \geq 1$,*

$$|\beta_{t+1} - \beta_t| \leq \frac{n\|z_t\|^2_{V_t^{-1}}}{2}$$

Proof. Using the expression of $\beta_t = n\sqrt{2\log\left(\frac{|V_t|^{1/2}}{|V_0|^{1/2}\delta'}\right)} + \sqrt{\lambda}S$ and that $|V_{t+1}| = |V_t|(1 + \|z_t\|^2_{V_t^{-1}})$, one has

$$\beta_{t+1} \leq \beta_t + n\sqrt{\log(1 + \|z_t\|^2_{V_t^{-1}})} \leq \beta_t + n\sqrt{1 + \|z_t\|^2_{V_t^{-1}}} \leq \beta_t + \frac{n\|z_t\|^2_{V_t^{-1}}}{2}.$$

□

First, we get rid of the conditioning in Eq. 4.14 which is a high probability event:

$$\forall t \geq 1, \quad \mathbb{P}(\bar{\theta}_{t+1} \in \mathcal{E}^{\mathrm{TS}}_{t+1} | \mathcal{F}^x_t, E_t) \geq 1 - \delta/8T \geq 1/2.$$

Hence,

$$R^{gap,2}_t \leq 2/p\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1}) - \bar{J}_t|\Big|\mathcal{F}^x_t, E_t\Big] \leq 2/p\mathbb{E}\left(\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1}) - \bar{J}_t|\Big|\mathcal{F}^x_{t+1}, E_t\Big]\Big|\mathcal{F}^x_t, E_t\right).$$

Let $\bar{\theta}_t \overset{d}{=} \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta$, where $\eta \sim \mathcal{N}(0, I)$. We denote as $\phi_t(\theta)$ and $\phi(\eta)$ the gaussian pdf of $\bar{\theta}_t|\mathcal{F}^x_t$ and η respectively defined as

$$\phi_{t+1}(\theta) = \frac{\det(V_{t+1})^{1/2}}{\beta^2(2\pi)^{(n+d)n/2}} e^{-1/2\beta^2\|\theta - \widehat{\theta}_{t+1}\|^2_{V_{t+1}}}$$
$$\phi(\eta)(\eta) = \frac{1}{1(2\pi)^{(n+d)n/2}} e^{-1/2\|\eta\|^2}.$$

Rewriting the expectation in the integral form, one has:

$$\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1}) - \bar{J}_t|\Big|\mathcal{F}^x_{t+1}, E_t\Big] = \int_{R^{(n+d)n}} |\bar{J}(\bar{\theta}) - \bar{J}_t|\phi_{t+1}(\bar{\theta})d\bar{\theta} = \int_{R^{(n+d)n}} |\bar{J}(\widehat{\theta}_{t+1} + \beta_{t+1}V_{t+1}^{-1/2}\eta) - \bar{J}_t|\phi(\eta)d\eta.$$

Using Prop. 4.2.2 and Prop. 4.2.3, it is clear that, for any $\theta \in \mathcal{S}$, $\|\nabla J(\theta)\| \leq 2D\rho/(1 - \rho^2)(1 + C) := \gamma_{lip}$. Therefore, J is Lipschitz on $\mathcal{S}$ and by construction, so is $\bar{J}$ (with the same Lipschitz constant). Thanks to the Lipschitz property of $\bar{J}$ and using. Prop. 4.5.5 and 4.5.6, we have:

$$\begin{aligned}
\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1}) - \bar{J}_t|\Big|\mathcal{F}^x_{t+1}, E_t\Big] &= \int_{R^{(n+d)n}} |\bar{J}(\widehat{\theta}_{t+1} + \beta_{t+1}V_{t+1}^{-1/2}\eta) - \bar{J}_t|\phi(\eta)d\eta \\
&\leq |\bar{J}(\widehat{\theta}_{t+1}) - \bar{J}(\widehat{\theta}_t)| \\
&\quad + \int_{R^{(n+d)n}} |\bar{J}(\widehat{\theta}_t + \beta_{t+1}V_{t+1}^{-1/2}\eta) - \bar{J}(\widehat{\theta}_t + \beta_t V_{t+1}^{-1/2}\eta)|\phi(\eta)d\eta \\
&\quad + \int_{R^{(n+d)n}} |\bar{J}(\widehat{\theta}_t + \beta_t V_{t+1}^{-1/2}\eta) - \bar{J}_t|\phi(\eta)d\eta \\
&\leq \gamma_{lip}\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\| + |\beta_{t+1} - \beta_t| \int_{R^{(n+d)n}} \|V_{t+1}^{-1/2}\eta\|\phi(\eta)d\eta \\
&\quad + \int_{R^{(n+d)n}} |\bar{J}(\widehat{\theta}_t + \beta_t V_{t+1}^{-1/2}\eta) - \bar{J}_t|\phi(\eta)d\eta.
\end{aligned}$$

Using that $\int_{R^{(n+d)n}} \|V_{t+1}^{-1/2}\eta\|\phi(\eta)d\eta \leq \frac{n}{\sqrt{\lambda}}$ for the second term and a change of variable for the third one, we get:

$$\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1}) - \bar{J}_t|\Big|\mathcal{F}_{t+1}^x, E_t\Big] \leq \gamma_{lip}\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\| + |\beta_{t+1} - \beta_t|\frac{n}{\sqrt{\lambda}} + \frac{|V_{t+1}|^{1/2}}{|V_t|^{1/2}}\int_{R^{(n+d)n}} |\bar{J}(\bar{\theta}) - \bar{J}_t|\phi_t(\bar{\theta})d\bar{\theta}$$

Let $\gamma_{gap,2} := \left(1 + \frac{(1+C)X}{\lambda}\right)^{1/2}$, using Prop. 4.5.5 and 4.5.6, we finally have that

$$\begin{aligned}\mathbb{E}\Big[|\bar{J}(\bar{\theta}_{t+1}) - \bar{J}_t|\Big|\mathcal{F}_{t+1}^x, E_t\Big] &\leq \gamma_{lip}\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\| + \frac{n^2}{2\lambda}\|z_t\|_{V_t^{-1}} + \gamma_{gap,2}\int_{R^{(n+d)n}} |\bar{J}(\bar{\theta}) - \bar{J}_t|\phi_t(\bar{\theta})d\bar{\theta} \\ &= \gamma_{lip}\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\| + \frac{n^2}{2\lambda}\|z_t\|_{V_t^{-1}} + \gamma_{gap,2}\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x\Big].\end{aligned}$$

Thus,

$$R_t^{gap,2} \leq 2/p\left(\gamma_{lip}\mathbb{E}\Big[\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|\Big|\mathcal{F}_t^x, E_t\Big] + \frac{n^2}{2\lambda}\|z_t\|_{V_t^{-1}} + \gamma_{gap,2}\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x, E_t\Big]\right).$$

Re-introducing the constraint $\bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}$ gives

$$\begin{aligned}\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x, E_t\Big] &= \mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\mathbb{1}\{\mathcal{E}_t^{\mathrm{TS}}\}\Big|\mathcal{F}_t^x, E_t\Big] + \mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\mathbb{1}\{\mathcal{E}_t^{\mathrm{TS},c}\}\Big|\mathcal{F}_t^x, E_t\Big] \\ &\leq \mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x, \mathcal{E}_t^{\mathrm{TS}}, E_t\Big] + 2D\Big(1 - \mathbb{P}(\bar{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}|\mathcal{F}_t^x, E_t)\Big) \\ &\leq \mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x, \mathcal{E}_t^{\mathrm{TS}}, E_t\Big] + D\delta/(4T),\end{aligned}$$

and using Prop. 4.5.5, one has

$$\mathbb{E}\Big[\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|\Big|\mathcal{F}_t^x, E_t\Big]\Big) \leq \gamma_{gap,1}\|z_t\|_{V_t^{-1}}.$$

Finally, one obtains

$$R_t^{gap,2} = \frac{2}{p}\left(\gamma_{gap,2}\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x, \mathcal{E}_t^{\mathrm{TS}}, E_t\Big] + \gamma_{gap,1}\gamma_{lip}\|z_t\|_{V_t^{-1}} + \gamma_{gap,2}D\delta/(2pT) + \frac{n^2}{2\lambda}\|z_t\|_{V_t^{-1}}\right).$$

Summing up. Applying Cor. 4.5.2 provides us with a bound on the cumulative sum of $\mathbb{E}\Big[|\bar{J}(\bar{\theta}_t) - \bar{J}_t|\Big|\mathcal{F}_t^x, \mathcal{E}_t^{\mathrm{TS}}, E_t\Big]$ and applying Prop. 4.2.4 provides us with a bound on $\sum \|z_t\|_{V_t^{-1}}$. Plugging it into Eq. 4.13 ensures that, with probability at least $1 - \delta/6$, $R^{gap} \leq \gamma_{gap}\sqrt{T} + \gamma'_{gap}$.

4.5.5 Final bound

Bounding $R^{\mathbf{RLS}}$ and R^{mart} The regret term R^{RLS} is related to the prediction error of the least square. The following lemma provides an upper bound similar to the one derived in (Abbasi-Yadkori and Szepesvári, 2011) with a minor modification due to the different policy update rule between our TS sampling algorithms. We postpone the proofs to App. 4.D.

Lemma 4.5.7. *Let If the* TS *is run over T time step according to algorithm 4.1, then,*

$$R^{\mathrm{RLS}} \leq \gamma_{\mathrm{RLS}}\sqrt{T}$$

where

$$\gamma_{\mathrm{RLS}} := 2/\sqrt{\lambda}(1+C^2)X^2SD(\gamma_T(\delta')+\beta_T(\delta'))\sqrt{2(n+d)\log\left(1+\frac{T(1+C^2)X^2}{\lambda(n+d)}\right)}.$$

On the other hand, as discussed in Sec. 4.4, the term R^{mart} is, up to minor modification, a martingale sequence and thus is bounded w.h.p.

Lemma 4.5.8. *With probability at least $1-\delta/6$, $R^{mart} \leq \gamma_{mart}\sqrt{T}$ where $\gamma_{mart} := 2DX^2\sqrt{2\log(12/\delta)}$.*

Plugging everything together. We are now ready to bring all the regret terms together. Collecting all the results, one has

$$\begin{aligned}
R^{\mathrm{RLS}} &\leq \gamma_{\mathrm{RLS}}\sqrt{T} && \text{a.s.}\\
R^{mart} &\leq \gamma_{mart}\sqrt{T} && \text{w.p. at least } 1-\delta/6\\
R^{Gap} &\leq \gamma_{gap}\sqrt{T}+\gamma'_{gap} && \text{w.p. at least } 1-\delta/6\\
R^{\mathrm{TS}} &\leq \gamma_{\mathrm{TS}}\sqrt{T} && \text{w.p. at least } 1-\delta/6
\end{aligned}$$

Therefore, w.p. at least $1-\delta/2$,

$$R^{\mathrm{RLS}}+R^{mart}+R^{Gap}+R^{\mathrm{TS}} \leq (\gamma_{\mathrm{RLS}}+\gamma_{mart}+\gamma_{gap}+\gamma_{\mathrm{TS}})\sqrt{T}+\gamma'_{gap}.$$

Finally, notice that the regret decomposition Eq. 4.6, is conditioned on the high probability sequence of events $\{E_t\}_{t\leq T}$ which holds w.p. at least $1-\delta/2$ according to Cor. 4.5.1. Thus, a union bound argument ensures that, w.p. at least $1-\delta$, $R(T) = \widetilde{O}\left(\sqrt{T}\right)$ where $\widetilde{O}$ hides some logarithmic factor and can be recovered from the proof.

4.6 Discussion

We derived the first $\widetilde{O}(\sqrt{T})$ frequentist regret for TS in LQ control systems. Despite the existing results in LQ for optimistic approaches (OFU-LQ), the Bayesian analysis of TS in LQ, and its frequentist analysis in linear bandit, we showed that controlling the frequentist regret induced by the randomness of the sampling process in LQ systems is considerably more difficult and it requires developing a novel approach, inspired from Ch. 3 that directly relates the regret of TS and the controls executed over time. Furthermore, we stress the need for TS to sample new parameters (and hence choose new policies) at a high frequency, which is in contrast with the lazy update approach of OFU-LQ. The major implication is that the available bound for the gap at policy switch only guarantees a $\widetilde{O}(T^{2/3})$ overall regret bound. To overcome this issue, we introduced a new line of proof that allows us to control the gap at policy switch, thus making possible to update parameters at each time step while ensuring a $\widetilde{O}(\sqrt{T})$ overall regret.
Despite the fact that most of the proof is derived in the general n dimensional case, two key steps are unfortunately restricted to the 1d case (i.e., when the state is 1 dimensional), which narrows down the final results. We believe that this comes mainly from technical difficulties and that it can be extended to n dimensions. We discuss here these limitations and how to relax it.

4.6.1 Probability of being optimistic

The first limitation comes from the need for optimism. The current proof of Lem. 4.5.5 uses the 1 dimension restriction to exhibit an optimistic set of parameters. Thanks to over-sampling, we show that the probability of sampling within this set is constant. While, inspired by the analysis of Ch. 3, we believe the way over-sampling is performed should guarantee the extension to n dimensions, the main limitation comes from the shape of the optimistic set. We highlight this issue and then try to tackle the problem at a higher level, removing the need for optimism.

Extension to n dimension. For sake of illustration, consider the n-dimensional problem which consists in n independent problems, of dimension 1. Formally, the structure of the problem is diagonal

$$A_* = \begin{pmatrix} a_*^1 & & 0 \\ & \ddots & \\ 0 & & a_*^n \end{pmatrix}, \quad B_* = \begin{pmatrix} b_*^1 & & 0 \\ & \ddots & \\ 0 & & b_*^n \end{pmatrix}, \quad Q = \begin{pmatrix} q^1 & & 0 \\ & \ddots & \\ 0 & & q^n \end{pmatrix}, \quad R = \begin{pmatrix} r^1 & & 0 \\ & \ddots & \\ 0 & & r^n \end{pmatrix}.$$

and assume that the learner is aware of this structure. In this case, each system can be estimated and sampled independently, so one can write $\widehat{\theta}_t = [\widehat{\theta}_t^1, \dots, \widehat{\theta}_t^n]$, and $\widetilde{\theta}_t = [\widetilde{\theta}_t^1, \dots, \widetilde{\theta}_t^n]$, with $\widetilde{\theta}_t^i$ conditionally independent. Accordingly, each systems can be controlled independently, so $\mathrm{Tr}(P(\widetilde{\theta}_t)) = \sum_{i=1}^n J(\widetilde{\theta}_t^i)$, $\mathrm{Tr}(P(\theta_*)) = \sum_{i=1}^n J(\theta_*^i)$. As a

consequence,

$$\mathbb{P}\Big(\mathrm{Tr}(P(\widetilde{\theta}_t)) \leq \mathrm{Tr}(P(\theta_*))\Big) \geq \mathbb{P}\Big(J(\widetilde{\theta}_t^i) \leq J(\theta_*^i),\ \forall i \in [1,\dots,n]\Big) \geq \prod_{i=1}^{n} \mathbb{P}\Big(J(\widetilde{\theta}_t^i) \leq J(\theta_*^i)\Big).$$

By independence, the probability of being optimistic can be narrowed down to the probability of being jointly optimistic *in each* direction. Finally, notice that Prop. 2.3.1 becomes

$$\mathrm{Tr}\left((\hat{\theta}_t^i - \theta_*^i)^\mathsf{T} V_t^i (\hat{\theta}_t^i - \theta_*^i)\right) \leq \beta_t^i(\delta)^2, \quad \forall i \in [1,\dots,n]$$

but that the sampling is made as $\widetilde{\theta}_t^i = \mathcal{R}_{\mathcal{S}}\Big(\widehat{\theta}_t^i + \beta_t V_t^{i,-1/2} \eta_t^i\Big)$ where $\beta_t \geq \sqrt{n}\beta_t^i$ for all $i \in [1,\dots,n]$. In line with the analysis of Ch. 3, this over-sampling by a factor $\sqrt{n}$ ensures the joint probability to be constant.

Despite the lack of generality due to the diagonal structure, this example stresses the intuition that over-sampling (coming from β_t) prevents a bad scaling of the probability of being optimistic with the dimension, and that the crucial difficulties lies in the characterization of the optimistic set Θ^{opt}. While in the diagonal case, one can look at the joint probability of each system being optimistic, it is no longer possible in the general case, as optimizing the cost in a direction could incur a larger cost in another direction.

In fact, the proof of Lem. 4.5.5 relies on the following steps. First, we exhibit a set Θ^{opt} that contains optimistic parameters θ, leveraging the shape of the optimal average cost function $J(\theta) = \mathrm{Tr}(P(\theta))$. Then, we show that the mass of this set, once transformed by the mapping $L : \theta \to \frac{1}{\beta_t} V_t^{1/2}(\theta - \widehat{\theta}_t)$, is constant. This is motivated by the fact that the randomness of the sampling is due to η_t, thus the probability of being optimistic is related to the volume of Θ^{opt} under the parametrization corresponding to $\eta_t = L(\widetilde{\theta}_t)$. In the 1-dimensional case, we are able to exhibit a set Θ^{opt} which corresponds to the area intercepted by two parallel hyperplanes. This specific shape is invariant under the mapping L, thus the constant mass of Θ^{opt} w.r.t. the Lebesgue measure guarantee a constant mass of $L(\Theta^{\mathrm{opt}})$ i.e., a fixed probability of being optimistic. On the other hand, in the n-dimensional case, we are only able to exhibit a set Θ^{opt} that is a convex n-dimensional cone, of constant mass. Despite the fact that, once transformed by L, the obtained set is still a convex cone, its volume can shrink to zero (think of the case where one eigenvalue of V_t diverge to ∞) and therefore does not guarantee a constant probability of being optimistic. Whether it is possible to exhibit a bigger set Θ^{opt} of invariant mass w.r.t. the mapping L is, up to our knowledge, still an open question.

On the other hand, we would like to point out the fact that this approach implicitly replaces the initial objective $\mathrm{Tr}(P(\theta)) \leq \mathrm{Tr}(P(\theta_*))$ by the more constrained version $P(\theta) \preccurlyeq P(\theta_*)$, where $\preccurlyeq$ is the inequality associated with p.s.d. matrices. It means that we require *every* eigenvalues to be smaller, or, equivalently to be more optimistic in *every* direction. This is, of course, more restrictive than comparing the trace where being very optimistic in a direction could compensate for a little pessimism in others. Comparing the eigenvalues rather than the trace is motivated by the fact that there exists no perturbation theory for trace of Riccati/Lyapunov solutions while some

material is available in the matrix case.

Removing the need for optimism. While the above discussion is specific to LQ problem, as it focuses on the shape of the Riccati solution, we try here to tackle the issue at a higher level by discussing the need for optimism. As explained in Sec. 4.4, the core of the analysis is to show that, thanks to the Poincaré inequality and the relationship between the gradient of the optimal value function and the actual actions selected by TS, the *average absolute deviation* is cumulatively small. Formally, we show that

$$\sum_{t=0}^{T} \Delta_t = \sum_{t=0}^{T} \mathbb{E}\Big[\big| J(\widetilde{\theta}_t) - \mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) \big| \Big| \mathcal{F}_t^x \Big] = \widetilde{O}(\sqrt{T}).$$

Notice that this bound holds for any sampling distribution, which stresses the intuition that it is a structural property of TS (as long as the gradient corresponds to actions). On the other hand, the initial objective is to provide a bound on $J(\widetilde{\theta}_t) - J(\theta_*)$ or, up to a martingale term, on $\mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta_*)$. Since the only knowledge one has about the unknown parameter is that w.h.p. $\theta_* \in \mathcal{E}_t^{\mathrm{RLS}}$ for all $t \leq T$, one aims to bound

$$\sup_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \Big(\mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta) \Big). \tag{4.15}$$

This motivates the structure of the sampling (see Eq. 4.5), which ensures that $\widetilde{\theta}_t$ spans $\mathcal{E}_t^{\mathrm{RLS}}$. Then the actual analysis makes use of the over-sampling (i.e., $\mathcal{E}_t^{\mathrm{RLS}} \subset \mathcal{E}_t^{\mathrm{TS}}$) to guarantee a fixed probability of being optimistic, which provides a link between Eq. 4.15 and Δ_t as

$$\sup_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \Big(\mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta) \Big) \leq \Delta_t / p.$$

However, this derivation is worst-case in the sense that it implicitly assumes that the function $f : \theta \to \mathbb{E}(J(\widetilde{\theta}_t)|\mathcal{F}_t^x) - J(\theta)$ is flat everywhere but over the optimistic set, where it is of high value, and in particular, completely discards *almost optimistic* points. Thanks to the regularity of the optimal value function J (and hence of f), it seems that such inequality could be derived in a more global and tighter way. Formally, we would like to have access to an inequality of the form

$$\sup_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} |f(\theta)| \leq c_0 \, \mathbb{E}_{\theta \sim \mathcal{E}_t^{\mathrm{TS}}} \Big(|f(\theta)| \Big) \tag{4.16}$$

Alternatively, we could also use an inequality of the form

$$\sup_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} |f(\theta)| \leq c_1 \, \mathbb{E}_{\theta \sim \mathcal{E}_t^{\mathrm{TS}}} \Big(|f(\theta)| \Big) + c_2 \, \mathbb{E}_{\theta \sim \mathcal{E}_t^{\mathrm{TS}}} \Big(\|\nabla f(\theta)\| \Big). \tag{4.17}$$

with c_0, c_1, c_2 constants that depend on the regularity of f, on the sampling distribution and on the dimension. Despite the fact that the converse of Eq. 4.16 holds with $c_0 = 1$, it is quite hard to construct counter-example for reasonably smooth function, since it requires f to be flat everywhere but on a small subset where it is very steep (and thus

not that smooth). Furthermore, notice that Eq. 4.16 holds for linear functions while Eq. 4.17 holds in 1d (Taylor expansion). From a theoretical perspective, Eq. 4.16 seems related to Converse Holder Inequality, in the extreme case L^∞ versus L^1, while Eq. 4.17 seems related to Sobolev inequalities. Proving those inequalities with small constant c_0, c_1, c_2 is a difficult and open question with implications way beyond the scope of this work. In particular, it will be a major breakthrough for TS analysis, stressing why its structural property (maintaining Δ_t small) ensures a small regret.

4.6.2 Bounding the gap at policy switch

The second limitation comes from the way we bound R^{gap}, and is deeply related to the constraint $\mathcal{S}$ and the rejection sampling. The current proof of Lem. 4.5.6 relies on the fact that, when $n = 1$, $P(\theta) = J(\theta)$ so that bounding $\|P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\|_2$ is equivalent to bounding $|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|$. This is motivated by the fact that J can be extended to $\bar{J}$ in a constant fashion, while P cannot and because the constant extension is necessary to get rid of the rejection sampling due to the $\mathcal{S}$ constraint. However, we believe that this is a technical detail of the proof and that the core idea, which consists in linking $|J(\widetilde{\theta}_{t+1}) - J(\widetilde{\theta}_t)|$ to the *average absolute deviation* is still valid for $\|P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\|_2$. To highlight this intuition, we provide a sketch of the proof under the assumption that no rejection sampling is needed, and then discuss why this conjecture may hold.

The Gaussian sampling case. Assume for now, that the Riccati solution $P(\theta)$ is bounded everywhere (i.e., $\|P(\theta)\|_2 \leq D$, $\forall \theta \in \mathbb{R}^{n(n+d)}$). Therefore, Prop. 4.2.3 guarantee that P is Lipschitz everywhere with constant γ_{lip} and no rejection sampling is needed, so $\widetilde{\theta}_t \overset{d}{=} \bar{\theta}_t$ is actually sampled according to a Gaussian distribution. Thus, following the proof of Sec. 4.5.4, introducing $\bar{P}_t = \mathbb{E}\Big(P(\bar{\theta}_t)|\mathcal{F}_t^x, \mathcal{E}_t^{\text{TS}}, E_t\Big)$, one has:

$$
\begin{aligned}
R^{gap} &\leq 4D\sqrt{2T\log(24/\delta)} + X^2\sum_{t=0}^{T} R_t^{gap,1} + X^2\sum_{t=0}^{T} R_t^{gap,2} \quad \text{where} \\
R_t^{gap,1} &:= \mathbb{E}\Big[\|P(\bar{\theta}_t) - \bar{P}_t\|_2|\mathcal{F}_t^x, \mathcal{E}_t^{\text{TS}}, E_t\Big] \\
R_t^{gap,2} &:= \mathbb{E}\Big[\|P(\bar{\theta}_{t+1}) - \bar{P}_t\|_2|\mathcal{F}_t^x, \mathcal{E}_{t+1}^{\text{TS}}, E_{t+1}\Big]
\end{aligned}
$$

As in Sec. 4.5.4, $R_t^{gap,1}$ is an *average absolute deviation* of the performance and can be bounded the same way through the use of a modified Poincaré inequality. Even though Lem. 4.5.3 is proved for real-valued function, at the cost of worsening a bit the constant, using algebraic manipulation and matrix norm equivalence, one can apply it component-wise and retrieve the same result as in Sec. 4.5.2, thus ensuring that w.h.p. $\sum_{t=1}^{T} R_t^{gap,1} = \widetilde{O}(\sqrt{T})$. Therefore, what remains is to show that $R_t^{gap,2} = O\Big(R_t^{gap,1} + \sqrt{T}\Big)$. Using the same line of proof, one obtains:

$$
R_t^{gap,2} \leq \square\mathbb{E}\Big[\|P(\bar{\theta}_t + \widehat{\theta}_{t+1} - \widehat{\theta}_t) - \bar{P}_t\|_2\Big|\mathcal{F}_t^x, E_t\Big]
$$

where $\square$ is the appropriate constant. Finally, the Lipschitz property of P together with $\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|$ cumulatively small (see Prop. 4.5.5) ensures that $R_t^{gap,2}$ is cumulatively

closed to $R_t^{gap,1}$ and hence bounded.

Removing the constraint $\mathcal{S}$. The above derivation requires P to be bounded everywhere and thus Lipschitz. Unfortunately, this property is not true, as stated in Prop. 4.2.1 Yet intuitively, we believe that the idea of the proof may still hold. A sufficient condition to deal with the boundedness of P would be to guarantee that (for t big enough) $\mathcal{E}_t^{\mathrm{TS}} \subset \mathcal{S}$ or that, w.h.p., $\bar{\theta}_t \in \mathcal{S}$. This is tricky to verify theoretically, because it requires quantifying at which rate does $\mathcal{E}_t^{\mathrm{TS}}$ shrink i.e., to look directly at the consistency of the estimates. On the other hand, it is likely to be true in practice since, the bigger the probability of sampling a parameter close to $\mathcal{S}^c$ is, the faster $\mathcal{E}_t^{\mathrm{TS}}$ shrinks. The reason is that parameters that are close to $\mathcal{S}^c$ have high optimal value and thus induce very aggressive control: each time such a point is selected, a lot of knowledge is collected from the system which significantly reduces the uncertainty about the parameter estimates.

Appendix

4.A Control theory

4.A.1 Proof of Prop. 4.2.2

Proposition 4.2.2. *$\mathcal{S}$ is a compact set. For any $\theta \in \mathcal{S}$, θ is a stabilizable pair (since $J(\theta) = +\infty$ otherwise) and there exist $\rho < 1$ and $C < \infty$ positive constants such that $\rho = \sup_{\theta\in\mathcal{S}} \|A + BK(A,B)\|_2$ and $C = \sup_{\theta\in\mathcal{S}} \|K(\theta)\|_2$.*

1. When $\theta^\top = (A, B)$ is not stabilizable, there exists no linear control K such that the controlled process $x_{t+1} = Ax_t + BKx_t + \epsilon_{t+1}$ is stationary. Thus, the positiveness of Q and R implies $J(\theta) = \mathrm{Tr}(P(\theta)) = +\infty$. As a consequence, $\theta^\top \notin \mathcal{S}$.

2. The mapping $\theta \to \mathrm{Tr}(P(\theta))$ is continuous (see Lem. 4.2.3). Thus, $\mathcal{S}$ is compact as the intersection between a closed and a compact set.

3. The continuity of the mapping $\theta \to K(\theta)$ together with the compactness of $\mathcal{S}$ justifies the finite positive constants ρ and C. Moreover, since every $\theta \in \mathcal{S}$ are stabilizable pairs, $\rho < 1$.

4.A.2 Proof of Prop. 4.2.3

Let $\theta^\mathsf{T} = (A, B)$ where A and B are matrices of size $n \times n$ and $n \times d$ respectively. Let $\mathcal{R} : \mathbb{R}^{n+d,n} \times \mathbb{R}^{n,n} \to \mathbb{R}^{n,n}$ be the Riccati operator defined by:

$$\mathcal{R}(\theta, P) := Q - P + A^\mathsf{T}PA - A^\mathsf{T}PB(R + B^\mathsf{T}PB)^{-1}B^\mathsf{T}PA, \tag{4.18}$$

where Q, R are positive definite matrices. Then, the solution $P(\theta)$ of the Riccati equation is the solution of $\mathcal{R}(\theta, P) = 0$. While Prop. 4.2.2 guarantees that there exists a unique admissible solution as soon as $\theta \in \mathcal{S}$, addressing the regularity of the function $\theta \to P(\theta)$ requires the use of the implicit function theorem.

Theorem 4.A.1 (Implicit function theorem (Krantz and Parks, 2012))**.** *Let E and F be two Banach spaces, let $\Omega \subset E \times F$ be an open subset. Let $f : \Omega \to F$ be a C^1-map and let (x_0, y_0) be a point of Ω such that $f(x_0, y_0) = 0$. We denote as $d_y f(x_0, y_0) : F \to F$ the differential of the function f with respect to the second argument at point (x_0, y_0). Assume that this linear transformation is bounded and invertible. Then, there exists*

1. *two open subsets U and V such that $(x_0, y_0) \in U \times V \subset \Omega$,*

2. *a function $g : U \to V$ such that $g(x) = y$ for all $(x, y) \in U \times V$.*

Moreover, g is C^1 and $dg(x) = -d_y f(x, g(x))^{-1} d_x f(x, g(x))$ for all $(x, y) \in U \times V$.

Since R is positive definite, the Riccati operator is clearly a C^1-map from $\mathcal{S} \times \mathbb{S}_n^{++}$ to $\mathbb{S}_n^{++}$. Moreover, thanks to Prop. 4.2.2, to any $\theta \in \mathcal{S}$, there exists $P \in \mathbb{S}_n^{++}$ such that $\mathcal{R}(\theta, P) = 0$. Thanks to Thm. 4.A.1, a sufficient condition for $\theta \to P(\theta)$ to be C^1 on $\mathcal{S}$ is that the linear map $d_P\mathcal{R}(\theta, P(\theta)) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is a bounded invertible transformation.

- **Bounded.** There exists M such that, for any $P \in \mathbb{R}^{n\times n}$, $\|d_P\mathcal{R}(\theta, P(\theta))(P)\| \leq M\|P\|$.
- **Invertible.** There exists a bounded linear operator $S : \mathbb{R}^{n\times n} \to R^{n\times n}$ such that $SP = I_{n,n}$ and $PS = I_{n,n}$.

Property 4.A.1. *Let $\theta^\mathsf{T} = (A, B)$ and $\mathcal{R}$ be the Riccati operator defined in equation (4.18). Then, the differential of $\mathcal{R}$ w.r.t P taken in (θ, P) denoted as $d_P\mathcal{R}(\theta, P)$ is given by:*

$$d_P\mathcal{R}(\theta, P)(\delta P) := A_c^T \delta P A_c - \delta P, \quad \text{for any } \delta P \in \mathbb{R}^{n\times n},$$

where $A_c = A - B(R + B^\mathsf{T}PB)^{-1}B^\mathsf{T}PA$.

Proof. The proof is straightforward using the standard composition/multiplication/inverse operations for the differential operator together with an appropriate rearranging. □

Clearly, $d_P\mathcal{R}(\theta, P)$ is a bounded linear map. Moreover, thanks to the Lyapunov theory, for any stable matrix $\|A_c\|_2 < 1$ and for any positive definite matrix Q, the Lyapunov equation $A_c^T X A_c - X = Q$ admits a unique solution. From Prop. 4.2.2, the optimal matrix $P(\theta)$ is such that the corresponding A_c is stable. This implies that $d_P\mathcal{R}(\theta, P)$ is an invertible operator, and $\theta \to P(\theta)$ is C^1 on $\mathcal{S}$.

Therefore, the differential of $\theta \to P(\theta)$ can be deduced from the implicit function theorem. After tedious yet standard operations, one gets that for any $\theta \in \mathcal{S}$ and direction $\delta\theta \in \mathbb{R}^{(n+d)\times n}$:

$$dJ(\theta)(\delta\theta) = \mathrm{Tr}(dP(\theta)(\delta\theta)) = \mathrm{Tr}(\nabla J(\theta)^\mathsf{T}\delta\theta),$$

where $\nabla J(\theta) \in \mathbb{R}^{(n+d)\times n}$ is the jacobian matrix of J in θ and, for any $\delta\theta \in \mathbb{R}^{(n+d)\times n}$, one has:

$$\begin{aligned} dP(\theta)(\delta\theta) = A_c(\theta)^\mathsf{T} dP(\theta)(\delta\theta) A_c(\theta) + C(\theta, \delta\theta) + C(\theta, \delta\theta)^\mathsf{T}, \quad \text{where} \\ C(\theta, \delta\theta) = A_c(\theta)^\mathsf{T} P(\theta) \delta\theta^\mathsf{T} H(\theta). \end{aligned} \tag{4.19}$$

Proposition 4.A.1. *For any $\theta \in \mathcal{S}$ and any positive definite matrix V, one has the following inequality for the differential of P:*

$$\sup_{\|\delta\theta\|=1} \|dP(\theta)(V^{1/2}\delta\theta)\|_2 \leq 2D\rho/(1-\rho^2)\|H(\theta)\|_V.$$

Proof. From Eq. 4.19, we have, for any $\theta \in \mathcal{S}$, for any $\|\delta\theta\|_F = 1$,

$$\begin{aligned}\|dP(\theta)(V_t^{-1/2}\delta\theta)\|_2 &\leq \|A_c(\theta)\|_2^2\|dP(\theta)(V_t^{-1/2}\delta\theta)\|_2 + 2\|A_c(\theta)\|_2\|P(\theta)\|_2\|\delta\theta^\mathsf{T}V_t^{-1/2}H(\theta)\|_2 \\ &\leq \|A_c(\theta)\|_2^2\|dP(\theta)(V_t^{-1/2}\delta\theta)\|_2 + 2\|A_c(\theta)\|_2\|P(\theta)\|_2\|\delta\theta^\mathsf{T}V_t^{-1/2}H(\theta)\| \\ &\leq \|A_c(\theta)\|_2^2\|dP(\theta)(V_t^{-1/2}\delta\theta)\|_2 + 2\|A_c(\theta)\|_2\|P(\theta)\|_2\|\delta\theta\|\|H(\theta)\|_{V_t^{-1}},\end{aligned}$$

where we used the matrix norm equivalence from line 1 to line 2 and Cauchy-Schwartz from line 2 to line 3. Finally, on $\mathcal{S}$, $\|A_c(\theta)\|_2 \leq \rho$ and $\|P(\theta)\|_2 \leq \mathrm{Tr}P(\theta) \leq D$. Thus, $\|dP(\theta)(V_t^{-1/2}\delta\theta)\|_2 \leq 2D\rho/(1-\rho^2)\|H(\theta)\|_{V_t^{-1}}$ which concludes the proof. □

4.B Proof of Lem. 4.5.5

We prove here that, on E, the sampling $\widetilde{\theta} \sim \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta_t)$ guarantees a fixed probability of sampling an optimistic parameter i.e., which belongs to $\Theta_t^{\mathrm{opt}} := \{\theta \in \mathbb{R}^d \mid J(\theta) \leq J(\theta^\star)\}$. However, our result only holds for the $1-$dimensional case as we deeply leverage the geometry of the problem. Figure 4.B.1 synthesizes the properties of the optimal value function and the geometry of the problem w.r.t the probability of being optimistic.

1) First, we introduce a simpler subset of optimistic parameters which involves hyperplanes rather than complicated J level sets. Without loss of generality we assume that $A_* + B_*K_* = \rho_* \geq 0$ and introduce $H_* = \begin{pmatrix} 1 \\ K_* \end{pmatrix} \in \mathbb{R}^{1+d}$ so that $A_* + B_*K_* = \theta^\mathsf{T}H_*$. Let $\Theta^{lin,\mathrm{opt}} = \{\theta \in \mathbb{R}^d \mid |\theta^\mathsf{T}H_*| \leq \rho_*\}$. Intuitively, $\Theta^{lin,\mathrm{opt}}$ consists in the set of systems θ which are more stable under control K_*. The following proposition ensures those systems to be optimistic.

Proposition 4.B.1. $\Theta^{lin,opt} \subset \Theta_t^{opt}$.

Proof. Leveraging the expression of J, one has when $n = 1$,

$$J(\theta) = \mathrm{Tr}(P(\theta)) = P(\theta) = \lim_{T\to\infty} \mathbb{E}\Big[\sum_{t=0}^{T} x_t^2(Q + K(\theta)^2R)\Big] = (Q + K(\theta)^2R)\mathbb{V}(x_t),$$

where $\mathbb{V}(x_t) = (1 - |\theta^\mathsf{T}H(\theta)|^2)^{-1}$ is the steady-state variance of the stationary first order autoregressive process $x_{t+1} = \theta^\mathsf{T}H(\theta)x_t + \epsilon_{t+1}$ where ϵ_t is zero mean noise of variance 1 and $H(\theta) = \begin{pmatrix} 1 \\ K(\theta) \end{pmatrix}$. Thus,

$$J(\theta) = \Big(Q + K(\theta)^2R\Big)\Big(1 - |\theta^\mathsf{T}H(\theta)|^2\Big)^{-1}.$$

Hence, for any $\theta \in \Theta^{lin,\mathrm{opt}}$, $(1 - |\theta^\mathsf{T}H_*|^2)^{-1} \leq (1 - |\theta_*^\mathsf{T}H_*|^2)^{-1}$ which implies that

$$(Q + K_*^2R)(1 - |\theta^\mathsf{T}H_*|^2)^{-1} \leq (Q + K_*^2R)(1 - |\theta_*^\mathsf{T}H_*|^2)^{-1} = J(\theta_*).$$

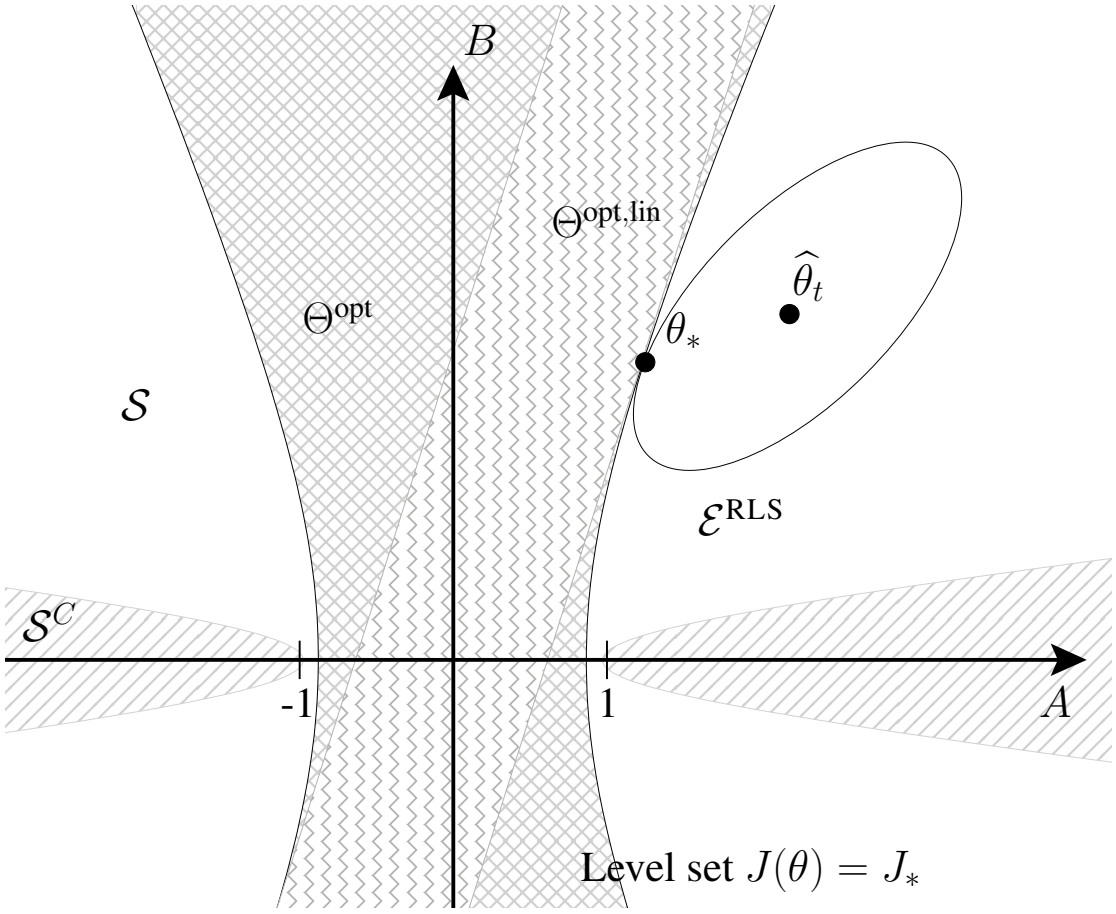

Figure 4.B.1 – **Optimism and worst case configuration. 1)** In 1-D, the Riccati solution is well-defined expect for $\{(A, B) \in]-\infty, -1] \cup [1, \infty[\times \{0\}\}$. The rejection sampling procedure into $\mathcal{S}$ ensures $P(\widetilde{\theta}_t)$ to be well-defined. Moreover, $\mathcal{S}^c$ does not overlap with Θ^{opt}. **2)** The introduction of the subset $\Theta^{lin,\text{opt}}$ prevents using the actual - yet complicated - optimistic set Θ^{opt} to lower bound the probability of being optimistic. **3)** Even if the event $\mathcal{E}^{\text{RLS}}$ holds, there exists an ellipsoid configuration which does not contain any optimistic point. This justifies the over-sampling to guarantee a fixed probability of being optimistic.

However, since $K(\theta)$ is the optimal control associated with θ,

$$\begin{aligned} J(\theta) &= (Q + K(\theta)^2 R)(1 - |\theta^\mathsf{T} H(\theta)|^2)^{-1} = \min_K (Q + K^2 R)(1 - |\begin{pmatrix} 1 & K \end{pmatrix} \theta|^2)^{-1} \\ &\leq (Q + K_*^2 R)(1 - |\theta^\mathsf{T} H_*|^2)^{-1} \leq J(\theta_*) \end{aligned}$$

□

As a result, $\mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{\text{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big) \geq \mathbb{P}\Big(\widetilde{\theta}_t \in \ \Theta^{lin,\text{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big)$ and we can focus on $\Theta^{lin,\text{opt}}$.

2) To ensure the sampling parameter to be admissible, we perform a rejection sampling until $\widetilde{\theta}_t \in \mathcal{S}$. Noticing that $\Theta^{lin,\text{opt}} \subset \Theta^{\text{opt}} \subset \mathcal{S}$ by construction, the rejection sampling

is always favorable in terms of probability of being optimistic. Since we seek for a lower bound, we can get rid of it and consider $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta$ where $\eta \sim \mathcal{N}(0, I_{1+d})$.[7]

3) On $\widehat{E}_t$, $\theta_\star \in \mathcal{E}_t^{\mathrm{RLS}}$, where $\mathcal{E}_t^{\mathrm{RLS}}$ is the confidence RLS ellipsoid centered in $\widehat{\theta}_t$. Since θ_* is fixed (by definition), we lower bound the probability by considering the worst possible $\widehat{\theta}_t$ such that $\widehat{E}_t$ holds. Intuitively, we consider the worst possible center for the RLS ellipsoid such that $\theta_\star$ still belong in $\mathcal{E}_t^{\mathrm{RLS}}$ and that the probability of being optimistic is minimal. Formally,

$$\begin{aligned}\mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big) &= \mathbb{P}_{\widetilde{\theta}_t \sim \mathcal{N}(\widehat{\theta}_t, \beta_t^2 V_t^{-1})}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big) \\ &\geq \min_{\widehat{\theta}: \|\widehat{\theta}-\theta_*\|_{V_t} \leq \beta_t(\delta')} \mathbb{P}_{\widetilde{\theta}_t \sim \mathcal{N}(\widehat{\theta}, \beta_t^2 V_t^{-1})}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x\Big)\end{aligned}$$

Moreover, by Cauchy-Schwarz inequality, for any $\widehat{\theta}$,

$$\Big|(\widehat{\theta} - \theta_*)^\mathsf{T} H_*\Big| \leq \|\widehat{\theta} - \theta_*\|_{V_t} \|H_*\|_{V_t^{-1}} \leq \beta_t(\delta')\|H_*\|_{V_t^{-1}},$$

thus,

$$\begin{aligned}\mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big) &\geq \min_{\widehat{\theta}: \|\widehat{\theta}-\theta_*\|_{V_t} \leq \beta_t(\delta')} \mathbb{P}_{\widetilde{\theta}_t \sim \mathcal{N}(\widehat{\theta}, \beta_t^2 V_t^{-1})}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x\Big) \\ &\geq \min_{\widehat{\theta}: |(\widehat{\theta}-\theta_*)^\mathsf{T} H_*| \leq \beta_t(\delta')\|H_*\|_{V_t^{-1}}} \mathbb{P}_{\widetilde{\theta}_t \sim \mathcal{N}(\widehat{\theta}, \beta_t^2 V_t^{-1})}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x\Big) \\ &= \min_{\widehat{\theta}: |\widehat{\theta}^\mathsf{T} H_* - \rho_*| \leq \beta_t(\delta')\|H_*\|_{V_t^{-1}}} \mathbb{P}_{\widetilde{\theta}_t \sim \mathcal{N}(\widehat{\theta}, \beta_t^2 V_t^{-1})}\Big(|\widetilde{\theta}_t^\mathsf{T} H_*| \leq \rho_* \mid \mathcal{F}_t^x\Big)\end{aligned}$$

Cor. 4.B.1 provides us with an explicit expression of the worst case ellipsoid. Introducing $x = \widetilde{\theta}_t^\mathsf{T} H_*$, one has $x \sim \mathcal{N}(\bar{x}, \sigma_x^2)$ with $\bar{x} = \widehat{\theta} H_*$ and $\sigma_x = \beta_t \|H_*\|_{V_t^{-1}}$. Applying Cor. 4.B.1 with $\alpha = \rho_*$, $\rho = \rho_*$ and $\beta = \beta_t(\delta')\|H_*\|_{V_t^{-1}}$, the last inequality becomes

$$\begin{aligned}\mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big) &\geq \min_{\widehat{\theta}: |\widehat{\theta}^\mathsf{T} H_* - \rho_*| \leq \beta_t(\delta')\|H_*\|_{V_t^{-1}}} \mathbb{P}_{\eta \sim \mathcal{N}(0, I_{1+d})}\Big(|\widehat{\theta}^\mathsf{T} H_* + \beta_t \eta^\mathsf{T} V_t^{-1/2} H_*| \leq \rho_* \mid \mathcal{F}_t^x\Big) \\ &\geq \mathbb{P}_{\eta \sim \mathcal{N}(0, I_{1+d})}\Big(|\rho_* + \beta_t(\delta')\|H_*\|_{V_t^{-1}} + \beta_t \eta^\mathsf{T} V_t^{-1/2} H_*| \leq \rho_* \mid \mathcal{F}_t^x\Big)\end{aligned}$$

Introducing the vector $u_t = \beta_t(\delta') V_t^{-1/2} H_*$, one can simplify

$$\begin{aligned}&|\rho_* + \beta_t(\delta')\|H_*\|_{V_t^{-1}} + \beta_t(\delta')\eta^\mathsf{T} V_t^{-1/2} H_*| \leq \rho_*, \\ \Leftrightarrow &-\rho_* \leq \rho_* + \|u_t\| + \eta^\mathsf{T} u_t \frac{\beta_t}{\beta_t(\delta')} \leq \rho_*, \\ \Leftrightarrow &-\frac{\rho_*}{\|u_t\|} - 1 \leq \eta^\mathsf{T} \frac{u_t}{\|u_t\|} \frac{\beta_t}{\beta_t(\delta')} \leq -1.\end{aligned}$$

Since $\eta \sim \mathcal{N}(0, I_{1+d})$ is rotationally invariant and since $\frac{\beta_t}{\beta_t(\delta')} := 1$,

$$\mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{lin,\mathrm{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big) \geq \mathbb{P}_{\epsilon \sim \mathcal{N}(0,1)}\Big(\epsilon \in \Big[1, 1 + \frac{2\rho_*}{\|u_t\|}\Big] \mid \mathcal{F}_t^x, \widehat{E}_t\Big).$$

[7] In the 1-dimensional case, η is just a $1 + d$ standard Gaussian r.v.

Finally, for all $t \leq T$, u_t is almost surely bounded: $\|u_t\| \leq \beta_T(\delta')\sqrt{(1+C^2)/\lambda}$. Therefore,

$$\mathbb{P}\Big(\widetilde{\theta}_t \in \Theta^{lin,\text{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\Big) \geq \mathbb{P}_{\epsilon\sim\mathcal{N}(0,1)}\Big(\epsilon \in \Big[1, 1+2\rho_*/\beta_T(\delta')\sqrt{(1+C^2)/\lambda}\Big]\Big) := p$$

Corollary 4.B.1. *For any $\rho, \sigma_x > 0$, for any $\alpha, \beta \geq 0$, $\arg\min_{\bar{x}:|\bar{x}-\alpha|\leq\beta} \mathbb{P}_{x\sim\mathcal{N}(\bar{x},\sigma_x^2)}\Big(|x| \leq \rho\Big) = \alpha + \beta$.*

This corollary is a direct consequence of the properties of standard Gaussian r.v.

Lemma 4.B.1. *Let x be a real random variable. For any $\rho, \sigma_x > 0$ Let $f : \mathbb{R} \to [0,1]$ be the continuous mapping defined by $f(\bar{x}) = \mathbb{P}_{x\sim\mathcal{N}(\bar{x},\sigma_x^2)}\Big(|x| \leq \rho\Big)$. Then, f is increasing on $\mathbb{R}_-$ and decreasing on $\mathbb{R}_+$.*

Proof. Without loss of generality, one can assume that $\sigma_x = 1/\sqrt{2}$ (otherwise, modify ρ), and that $\bar{x} \geq 0$ (by symmetry). Denoting as Φ and erf the standard Gaussian cdf and the error function, one has:

$$\begin{aligned}
f(\bar{x}) &= \mathbb{P}_{x\sim\mathcal{N}(\bar{x},\sigma_x^2)}\Big(-\rho \leq x \leq \rho\Big), = \mathbb{P}_{x\sim\mathcal{N}(\bar{x},\sigma_x^2)}\Big(x \leq \rho\Big) - \mathbb{P}_{x\sim\mathcal{N}(\bar{x},\sigma_x^2)}\Big(x \leq -\rho\Big),\\
&= \mathbb{P}_{x\sim\mathcal{N}(\bar{x},\sigma_x^2)}\Big((x-\bar{x})/\sigma_x \leq (\rho-\bar{x})/\sigma_x\Big) - \mathbb{P}_{x\sim\mathcal{N}(\bar{x},\sigma_x^2)}\Big((x-\bar{x})/\sigma_x \leq (-\rho-\bar{x})/\sigma_x\Big),\\
&= \Phi((\rho-\bar{x})/\sigma_x) - \Phi(-(\rho+\bar{x})/\sigma_x),\\
&= \frac{1}{2} + \frac{1}{2}\text{erf}((\rho-\bar{x})/\sqrt{2}\sigma_x) - \frac{1}{2} - \frac{1}{2}\text{erf}(-(\rho+\bar{x})/\sqrt{2}\sigma_x),\\
&= \frac{1}{2}\Big(\text{erf}(\rho-\bar{x}) - \text{erf}(-(\rho+\bar{x}))\Big).
\end{aligned}$$

Since erf is odd, one obtains $f(\bar{x}) = \frac{1}{2}\Big(\text{erf}(\rho-\bar{x}) + \text{erf}(\rho+\bar{x})\Big)$. The error function is differentiable with $\text{erf}'(z) = \frac{2}{\pi}e^{-z^2}$, thus

$$\begin{aligned}
f'(\bar{x}) &= \frac{1}{\pi}\Big(\exp\Big(-(\rho+\bar{x})^2\Big) - \exp\Big(-(\rho-\bar{x})^2\Big)\Big)\text{ă}\\
&= -\frac{2}{\pi}\sinh\Big((\rho-\bar{x})^2\Big) \leq 0
\end{aligned}$$

Hence, f is decreasing on $\mathbb{R}_+$ and by symmetry, is increasing on $\mathbb{R}_-$. □

4.C Weighted L1 Poincaré inequality (proof of Lem. 4.5.3)

This result is build upon the following theorem which links the function to its gradient in L^1 norm:

Theorem 4.C.1 (see Acosta and Durán 2004). *Let $W^{1,1}(\Omega)$ be the Sobolev space on $\Omega \subset \mathbb{R}^d$. Let Ω be a convex domain bounded with diameter D and $f \in W^{1,1}(\Omega)$ of zero average on Ω then*

$$\int_\Omega |f(x)|dx \leq \frac{D}{2}\int_\Omega ||\nabla f(x)||dx$$

Lem. 4.5.3 is an extension of Thm. 4.C.1. In practice, we show that their proof still holds for log-concave weight.

Theorem 4.C.2. *Let $L > 0$ and ρ any non negative and log-concave function on $[0, L]$. Then for any $f \in W^{1,1}(0, L)$ such that*

$$\int_0^L f(x)\rho(x)dx = 0$$

one has:

$$\int_0^L |f(x)|\rho(x)dx \leq 2L \int_0^L |f'(x)|\rho(x)dx \tag{4.20}$$

The proof is based on the following inequality for log-concave function.

Lemma 4.C.1. *Let ρ be any non negative log-concave function on $[0, 1]$ such that $\int_0^1 \rho(x) = 1$ then*

$$\forall x \in (0, 1), \quad H(\rho, x) := \frac{1}{\rho(x)} \int_0^x \rho(t)dt \int_x^1 \rho(t)dt \leq 1 \tag{4.21}$$

Proof. Since any non-negative log-concave function on $[0, 1]$ can be rewritten as $\rho(x) = e^{\nu(x)}$ where ν is a concave function on $[0, 1]$ and since $x \to e^x$ is increasing, the monotonicity of ν is preserved and as for concave function, ρ can be either increasing, decreasing or increasing then decreasing on $[0, 1]$.
Hence, $\forall x \in (0, 1)$, either

1. $\rho(t) \leq \rho(x)$ for all $t \in [0, x]$,
2. $\rho(t) \leq \rho(x)$ for all $t \in [x, 1]$.

Assume that $\rho(t) \leq \rho(x)$ for all $t \in [0, x]$ without loss of generality. Then,

$$\begin{aligned}
\forall x \in (0, 1), \quad H(\rho, x) &:= \frac{1}{\rho(x)} \int_0^x \rho(t)dt \int_x^1 \rho(t)dt \\
&= \int_0^x \frac{\rho(t)}{\rho(x)} \int_x^1 \rho(t)dt \\
&\leq \int_0^x dt \int_x^1 \rho(t)dt \\
&\leq x \int_0^1 \rho(t)dt \leq x \leq 1
\end{aligned}$$

□

Proof of theorem 4.C.2. This proof is exactly the same as Acosta and Durán (2004) where we use lemma 4.C.1 instead of a concave inequality. We provide it for sake of completeness.

A scaling argument ensures that it is enough to prove it for $L = 1$. Moreover, dividing both side of (4.20) by $\int_0^1 \rho(x)dx$, we can assume without loss of generality that

$\int_0^1 \rho(x)dx = 1$.
Since $\int_0^1 f(x)\rho(x)dx = 0$ by integration part by part one has:

$$f(y) = \int_0^y f'(x) \int_0^x \rho(t)dt - \int_y^1 f'(x) \int_x^1 \rho(t)dt$$
$$|f(y)| \leq \int_0^y |f'(x)| \int_0^x \rho(t)dt + \int_y^1 |f'(x)| \int_x^1 \rho(t)dt$$

Multiplying by $\rho(y)$, integrating on y and applying Fubini's theorem leads to

$$\int_0^1 |f(y)|\rho(y)dy \leq 2 \int_0^1 |f'(x)| \int_0^x \rho(t)dt \int_x^1 \rho(t)dt$$

and applying (4.21) of lemma 4.C.1 ends the proof. □

While theorem 4.C.2 provides a 1 dimensional weighted Poincaré inequality, we actually seek for one in $\mathbb{R}^d$. The idea of Acosta and Durán (2004) is to use arguments of Payne and Weinberger (1960) to reduce the $d-$dimensional problem to a $d-1$ dimensional problem by splitting any convex set Ω into subspaces Ω_i thin in all but one direction and such that an average property is preserved. We just provide their result.

Lemma 4.C.2. *Let $\Omega \subset \mathbb{R}^d$ be a convex domain with finite diameter D and $u \in L^1(\Omega)$ such that $\int_\Omega u = 0$. Then, for any $\delta > 0$, there exists a decomposition of Ω into a finite number of convex domains Ω_i satisfying*

$$\Omega_i \cap \Omega_j = \emptyset \text{ for } i \neq j, \quad \bar{\Omega} = \bigcup \bar{\Omega}_i, \quad \int_{\Omega_i} u = 0$$

and each Ω_i is thin in all but one direction i.e., in an appropriate rectangular coordinate system $(x, y) = (x, y_1, \ldots, y_{d-1})$ the set Ω_i is contained in

$$\{(x, y) : \ 0 \leq x \leq D, \quad 0 \leq y_i \leq \delta \text{ for } i = 1, \ldots, d-1\}$$

This decomposition together with Treheorem 4.C.2 allow us to prove the $d-$dimensional weighted Poincaré inequality.

Proof of Lem. 4.5.3. By density, we can assume that $u \in C^\infty(\bar{\Omega})$. Hence, $up \in C^2(\bar{\Omega})$. Let M be a bound for up and all its derivative up to the second order.
Given $\delta > 0$ decompose the set Ω into Ω_i as in lemma 4.C.2 and express $z \in \Omega_i$ into the appropriate rectangular basis $z = (x, y)$, where $x \in [0, d_i]$, $y \in [0, \delta]$. Define as $\rho(x_0)$ the $d-1$ volume of the intersection between Ω_i and the hyperplan $\{x = x_0\}$. Since Ω_i is convex, ρ is concave and from the smoothness of up one has:

$$\left| \int_{\Omega_i} |u(x,y)|p(x,y)dxdy - \int_0^{d_i} |u(x,0)|p(x,0)\rho(x)dx \right| \leq (d-1)M|\Omega_i|\delta \tag{4.22}$$
$$\left| \int_{\Omega_i} |\frac{\partial u}{\partial x}(x,y)|p(x,y)dxdy - \int_0^{d_i} |\frac{\partial u}{\partial x}(x,0)|p(x,0)\rho(x)dx \right| \leq (d-1)M|\Omega_i|\delta \tag{4.23}$$
$$\left| \int_{\Omega_i} u(x,y)p(x,y)dxdy - \int_0^{d_i} u(x,0)p(x,0)\rho(x)dx \right| \leq (d-1)M|\Omega_i|\delta \tag{4.24}$$

Those equation allows us to switch from $d-$dimensional integral to $1-$dimensional integral for which we can apply theorem 4.C.2 at the condition that $\int_0^{d_i} u(x,0)p(x,0)\rho(x)dx = 0$ (which is not satisfied here). On the other hand, we can apply theorem 4.C.2 to

$$g(x) = u(x,0) - \int_0^{d_i} u(x,0)p(x,0)\rho(x)dx / \int_0^{d_i} p(x,0)\rho(x)dx$$

with weighted function $x \to p(x,0)\rho(x)$. Indeed, $x \to p(x,0)$ is log-concave - as restriction along one direction of log-concave function, $x \to \rho(x)$ is log-concave - as a concave function, and so is $x \to p(x,0)\rho(x)$ - as product of log-concave function. Moreover, $g \in W^{1,1}(0,d_i)$ and $\int_0^{d_i} g(x)p(x,0)\rho(x)dx = 0$ by construction. Therefore, applying theorem 4.C.2 one gets:

$$\begin{aligned}
&\int_0^{d_i} |g(x)|p(x,0)\rho(x)dx \leq 2d_i \int_0^{d_i} |g'(x)|p(x,0)\rho(x)dx \\
&\int_0^{d_i} |u(x,0)|p(x,0)\rho(x)dx \leq 2d_i \int_0^{d_i} |\frac{\partial u}{\partial x}(x,0)|p(x,0)\rho(x)dx - \left|\int_0^{d_i} u(x,0)p(x,0)\rho(x)dx\right| \\
&\int_0^{d_i} |u(x,0)|p(x,0)\rho(x)dx \leq 2d_i \int_0^{d_i} |\frac{\partial u}{\partial x}(x,0)|p(x,0)\rho(x)dx + (d-1)M|\Omega_i|\delta
\end{aligned} \tag{4.25}$$

where we use equation (4.24) together with $\int_{\Omega_i} u(z)p(z)dz = 0$ to obtain the last inequality.

Finally, from (4.22)

$$\int_{\Omega_i} |u(x,y)|p(x,y)dxdy \leq \int_0^{d_i} |u(x,0)|p(x,0)\rho(x)dx + (d-1)M|\Omega_i|\delta$$

from (4.25)

$$\int_{\Omega_i} |u(x,y)|p(x,y)dxdy \leq 2d_i \int_0^{d_i} |\frac{\partial u}{\partial x}(x,0)|p(x,0)\rho(x)dx + (d-1)M|\Omega_i|\delta(1+2d_i)$$

from (4.23)

$$\begin{aligned}
&\int_{\Omega_i} |u(x,y)|p(x,y)dxdy \leq 2d_i \int_{\Omega_i} |\frac{\partial u}{\partial x}(x,y)|p(x,y)dxdy + (d-1)M|\Omega_i|\delta(1+4d_i) \\
&\int_{\Omega_i} |u(x,y)|p(x,y)dxdy \leq 2d_i \int_{\Omega_i} ||\nabla u(x,y)||p(x,y)dxdy + (d-1)M|\Omega_i|\delta(1+4d_i)
\end{aligned}$$

Summing up on Ω_i leads to

$$\int_{\Omega} |u(z)|p(z)dz \leq 2D \int_{\Omega} ||\nabla u(z)||p(z)dz + (d-1)M|\Omega|\delta(1+4D)$$

and since δ is arbitrary one gets the desired result. □

4.D Regret proofs

We collect here the regret proofs that are directly inspired or collected from (Abbasi-Yadkori and Szepesvári, 2011). Since our framework slightly differs, minors differences coming from a different conditioning or the fact that we do not consider lazy updates, we provide it for the sake of completeness.

Proof of Lem. 4.5.1. Let $\delta' = \delta/8T$.
1) From Prop. 2.3.1, $\mathbb{P}\Big(\|\widehat{\theta}_t - \theta_*\|_{V_t} \leq \beta_t(\delta')\Big) \geq 1 - \delta'$. Hence,

$$\begin{aligned}\mathbb{P}\Big(\widehat{E}\Big) &= \mathbb{P}\Big(\bigcap_{t=0}^{T}\Big(\|\widehat{\theta}_t - \theta_*\|_{V_t} \leq \beta_t(\delta')\Big)\Big) = 1 - \mathbb{P}\Big(\bigcup_{t=0}^{T}\Big(\|\widehat{\theta}_t - \theta_*\|_{V_t} \geq \beta_t(\delta')\Big)\Big) \\ &\geq 1 - \sum_{t=0}^{T}\mathbb{P}\Big(\|\widehat{\theta}_t - \theta_*\|_{V_t} \geq \beta_t(\delta')\Big) \geq 1 - T\delta' \geq 1 - \delta/8\end{aligned}$$

2) From Lem. A.2, let η be component-wise $\mathcal{N}(0,1)$ then, for any $\epsilon > 0$, making use of the fact that $\|\eta\| \leq n\sqrt{n+d}\max_{i\leq n+d, j\leq n}|\eta_{i,j}|$,

$$\begin{aligned}\mathbb{P}\Big(\|\eta\| \leq \epsilon\Big) &\geq \mathbb{P}\Big(n\sqrt{n+d}\max_{i,j}|\eta_{i,j}| \leq \epsilon\Big) \geq 1 - \prod_{i,j}\mathbb{P}\Big(|\eta_{i,j}| \geq \frac{\epsilon}{n\sqrt{n+d}}\Big) \\ &\geq 1 - n(n+d)\mathbb{P}_{X\sim\mathcal{N}(0,1)}\Big(|X| \geq \frac{\epsilon}{n\sqrt{n+d}}\Big).\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}\Big(\widetilde{E}\Big) &= \mathbb{P}\Big(\bigcap_{t=0}^{T}\Big(\|\widetilde{\theta}_t - \widehat{\theta}_t\|_{V_t} \leq \gamma(\delta')\Big)\Big) = 1 - \mathbb{P}\Big(\bigcup_{t=0}^{T}\Big(\|\widetilde{\theta}_t - \widehat{\theta}_t\|_{V_t} \geq \gamma(\delta')\Big)\Big) \\ &\geq 1 - \sum_{t=0}^{T}\mathbb{P}\Big(\|\widetilde{\theta}_t - \widehat{\theta}_t\|_{V_t} \geq \gamma(\delta')\Big) \geq 1 - \sum_{t=0}^{T}\mathbb{P}\Big(\|\eta\| \geq \gamma(\delta')/\beta_t(\delta')\Big) \\ &\geq 1 - \sum_{t=0}^{T}\mathbb{P}\Big(\|\eta\| \geq n\sqrt{2(n+d)\log\Big(2n(n+d)/\delta'\Big)}\Big) \geq 1 - T\delta' \geq 1 - \delta/8.\end{aligned}$$

Finally, a union bound argument ensures that $\mathbb{P}(\widehat{E} \cap \widetilde{E}) \geq 1 - \delta/4$.

Proof of Cor. 4.5.1. This result comes directly from Sec. 4.1. and App. D of Abbasi-Yadkori and Szepesvári (2011). The proof relies on the fact that, on $\widehat{E}$, because $\widetilde{\theta}_t$ is chosen within the confidence ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$, the number of time steps the true closed loop matrix $A_* + B_*K(\widetilde{\theta}_t)$ is unstable is small. Intuitively, the reason is that as soon as the true closed loop matrix is unstable, the state process explodes and the confidence ellipsoid is drastically changed. As the ellipsoid can only shrink over time, the state is well controlled expect for a small number of time steps.
Since the only difference is that, on $\widehat{E} \cap \widetilde{E}$, $\widetilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}$, the same argument applies and the same bound holds replacing β_t with γ_t. Therefore, there exists appropriate problem dependent constants X, X' such that $\mathbb{P}(\bar{E}|\widehat{E} \cap \widetilde{E}) \geq 1 - \delta/4$. Finally, a union bound

argument ensures that $\mathbb{P}(\widehat{E} \cap \widetilde{E} \cap \bar{E}) \geq 1 - \delta/2$.

Regret decomposition. Let $R(T) = \sum_{t=0}^{T} x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J(\theta_*)$. Since $\{E_t\}_{t \leq T}$ is an decreasing sequence of events, one has

$$R(T)\mathbb{1}\{E_T\} \leq \sum_{t=0}^{T} \Big(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J(\theta_*) \Big) \mathbb{1}\{E_t\}.$$

From the Bellman optimality equations for LQ problems, we get that

$$\begin{aligned} J(\widetilde{\theta}_t) + x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t &= \min_u \left\{ x_t^\mathsf{T} Q x_t + u^\mathsf{T} R u + \mathbb{E}\Big[\tilde{x}_{t+1}^{u,\mathsf{T}} P(\widetilde{\theta}_t) \tilde{x}_{t+1}^{u} \Big| \mathcal{F}_t \Big] \right\} \\ &= x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t + \mathbb{E}\Big[\tilde{x}_{t+1}^{u_t,\mathsf{T}} P(\widetilde{\theta}_t) \tilde{x}_{t+1}^{u_t} \Big| \mathcal{F}_t \Big], \end{aligned}$$

where $\tilde{x}_{t+1}^{u} = \widetilde{\theta}_t^\mathsf{T} z_t + \epsilon_{t+1}$. Hence,

$$\begin{aligned} \Big(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J(\theta_*) \Big) \mathbb{1}\{E_t\} &= \Big(J(\widetilde{\theta}_t) - J(\theta_*) \Big) \mathbb{1}\{E_t\} \\ &\quad + \Big(z_t^\mathsf{T} \widetilde{\theta}_t P(\widetilde{\theta}_t) \widetilde{\theta}_t^\mathsf{T} z_t - z_t^\mathsf{T} \theta_* P(\widetilde{\theta}_t) \theta_*^\mathsf{T} z_t \Big) \mathbb{1}\{E_t\} \\ &\quad + x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t \mathbb{1}\{E_t\} - \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_t) x_{t+1} \mathbb{1}\{E_t\} \Big| \mathcal{F}_t \Big] \end{aligned}$$

where we used in the last line that E_t is $\mathcal{F}_t-$measurable. Noticing that $E_{t+1} \subset E_t$, one has $\mathbb{1}\{E_{t+1}\}\mathbb{1}\{E_t\} = \mathbb{1}\{E_{t+1}\}$ and since $P(\widetilde{\theta}_t)$ is positive definite, $x_{t+1}^\mathsf{T} P(\widetilde{\theta}_t) x_{t+1} \geq 0$. Therefore,

$$\begin{aligned} \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_t) x_{t+1} \mathbb{1}\{E_t\} \Big| \mathcal{F}_t \Big] &= \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_t) x_{t+1} \mathbb{1}\{E_t\} (\mathbb{1}\{E_{t+1}\} + \mathbb{1}\{E_{t+1}^c\}) \Big| \mathcal{F}_t \Big] \\ &= \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_t) x_{t+1} \mathbb{1}\{E_{+1}\} \Big| \mathcal{F}_t \Big] + \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_t) x_{t+1} \mathbb{1}\{E_t\} \mathbb{1}\{E_{t+1}^c\} \Big| \mathcal{F}_t \Big] \\ &\geq \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_t) x_{t+1} \mathbb{1}\{E_{+1}\} \Big| \mathcal{F}_t \Big] \\ &= \mathbb{E}\Big[x_{t+1}^\mathsf{T} \Big(P(\widetilde{\theta}_t) - P(\widetilde{\theta}_{t+1}) \Big) x_{t+1} \mathbb{1}\{E_{+1}\} \Big| \mathcal{F}_t \Big] \\ &\quad + \mathbb{E}\Big[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_{t+1}) x_{t+1} \mathbb{1}\{E_{+1}\} \Big| \mathcal{F}_t \Big] \end{aligned}$$

Bounding R^{mart}. Re-ordering the term, using the fact that $x_0 = 0$ and that $P(\widetilde{\theta}_t)$ is definite positive by definition, one obtains:

$$\begin{aligned} R^{mart} \leq & \sum_{t=1}^{T} \left\{ x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t \mathbb{1}\{E_t\} - \mathbb{E}\Big[x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t \mathbb{1}\{E_t\} | \mathcal{F}_{t-1} \Big] \right\} + \\ & x_0^\mathsf{T} P(\widetilde{\theta}_0) x_0 - \mathbb{E}\Big[x_{T+1}^\mathsf{T} P(\widetilde{\theta}_{T+1}) x_{T+1} \mathbb{1}\{E_{T+1}\} | \mathcal{F}_T \Big] \\ \leq & \sum_{t=1}^{T} \left\{ x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t \mathbb{1}\{E_t\} - \mathbb{E}\Big[x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t \mathbb{1}\{E_t\} | \mathcal{F}_{t-1} \Big] \right\} \end{aligned}$$

which is turns to be a martingale. On E, $\|x_t\| \leq X$ for all $t \in [0, T]$. Moreover, since $\widetilde{\theta}_t \in \mathcal{S}$ for all $t \in [0, T]$ due to the rejection sampling, $\mathrm{Tr}(P(\widetilde{\theta}_t)) \leq D$. From the definition of the matrix 2-norm,

$$\sup_{\|x\| \leq X} x^\mathsf{T} P(\widetilde{\theta}_t) x \leq X^2 \|P(\widetilde{\theta}_t)^{1/2}\|_2^2.$$

Matrix norm equivalence ensures that for any $A \in \mathbb{R}^{m,n}$, $\|A\|_2 \leq \|A\|$. Therefore, $\|P(\widetilde{\theta}_t)^{1/2}\|_2^2 \leq \|P(\widetilde{\theta}_t)^{1/2}\|^2 = \mathrm{Tr}P(\widetilde{\theta}_t)$ and, for any $t \in [0,T]$, $\sup_{\|x\|\leq X} x^\mathsf{T} P(\widetilde{\theta}_t)x \leq X^2 D$ so the martingale increments are bounded almost surely on E by $2DX^2$.
Applying Azuma's inequality to R^{mart} one obtains that, w.p. at least $1-\delta/6$,

$$R_1^{\mathrm{RLS}} = \sum_{t=0}^{T} \left\{ x_t^\mathsf{T} P(\widetilde{\theta}_t) x_t \mathbb{1}\{E_t\} - \mathbb{E}\left[x_{t+1}^\mathsf{T} P(\widetilde{\theta}_{t+1}) x_{t+1} \mathbb{1}\{E_{t+1}\} | \mathcal{F}_t\right] \right\} \leq 2DX^2\sqrt{2T\log(12/\delta)}.$$

Bounding $R^{\mathbf{RLS}}$. The whole derivation is performed on the event E.

$$\begin{aligned} R^{\mathrm{RLS}} &= \sum_{t=0}^{T} \left\{ z_t^\mathsf{T} \widetilde{\theta}_t P(\widetilde{\theta}_t) \widetilde{\theta}_t^\mathsf{T} z_t - z_t^\top \theta_* P(\widetilde{\theta}_t) \theta_*^\mathsf{T} z_t \right\} = \sum_{t=0}^{T} \left\{ \|\widetilde{\theta}_t^\mathsf{T} z_t\|^2_{P(\widetilde{\theta}_t)} - \|\theta_*^\mathsf{T} z_t\|^2_{P(\widetilde{\theta}_t)} \right\}, \\ &= \sum_{t=0}^{T} \left(\|\widetilde{\theta}_t^\mathsf{T} z_t\|_{P(\widetilde{\theta}_t)} - \|\theta_*^\mathsf{T} z_t\|_{P(\widetilde{\theta}_t)} \right) \left(\|\widetilde{\theta}_t^\mathsf{T} z_t\|_{P(\widetilde{\theta}_t)} + \|\theta_*^\mathsf{T} z_t\|_{P(\widetilde{\theta}_t)} \right) \end{aligned}$$

By the triangular inequality,

$$\|\widetilde{\theta}_t^\mathsf{T} z_t\|_{P(\widetilde{\theta}_t)} - \|\theta_*^\mathsf{T} z_t\|_{P(\widetilde{\theta}_t)} \leq \|P(\widetilde{\theta}_t)^{1/2}(\widetilde{\theta}_t^\mathsf{T} z_t - \theta_*^\mathsf{T} z_t)\| \leq \|P(\widetilde{\theta}_t)\| \|(\widetilde{\theta}_t^\mathsf{T} - \theta_*^\mathsf{T}) z_t\|.$$

Making use of the fact that $\widetilde{\theta}_t \in \mathcal{S}$ by construction of the rejection sampling, $\theta_\star \in \mathcal{S}$ by Asm. 4.2.2 and that $\sup_{t\in[0,T]} \|z_t\| \leq \sqrt{(1+C^2)X^2}$ thanks to the conditioning on E and Prop. 4.2.2, one gets:

$$\begin{aligned} R^{\mathrm{RLS}} &\leq \sum_{t=0}^{T} \left(\sqrt{D} \|(\widetilde{\theta}_t^\mathsf{T} - \theta_*^\mathsf{T}) z_t\| \right) \left(2S\sqrt{D}\sqrt{(1+C^2)X^2} \right) \\ &\leq 2SD\sqrt{(1+C^2)X^2} \sum_{t=0}^{T} \|(\widetilde{\theta}_t^\mathsf{T} - \theta_*^\mathsf{T}) z_t\| \end{aligned}$$

and one just has to bound $\sum_{t=0}^{T} \|(\widetilde{\theta}_t^\mathsf{T} - \theta_*^\mathsf{T}) z_t\|$. Using Cauchy-Schwarz inequality, one has:

$$\sum_{t=0}^{T} \|(\widetilde{\theta}_t^\mathsf{T} - \theta_*^\mathsf{T}) z_t\| = \sum_{t=0}^{T} \|(V_t^{1/2})(\widetilde{\theta}_t - \theta_*))^\mathsf{T} V_t^{-1/2} z_t\| \leq \sum_{t=0}^{T} \|\widetilde{\theta}_t - \theta_*\|_{V_t} \|z_t\|_{V_t^{-1}}$$

However, on E, $\|\widetilde{\theta}_t - \theta_*\|_{V_t} \leq \|\widetilde{\theta}_t - \widehat{\theta}_t\|_{V_t} + \|\theta_* - \widehat{\theta}_t\|_{V_t} \leq \beta_t(\delta') + \gamma(\delta') \leq \beta_T(\delta') + \gamma(\delta')$. Therefore,

$$R^{\mathrm{RLS}} \leq 4SD\sqrt{(1+C^2)X^2} \left(\beta_T(\delta') + \gamma(\delta') \right) \sum_{t=0}^{T} \|z_t\|_{V_t^{-1}}.$$

The proof is conclude by using Cauchy-Schwarz inequality and Prop. 4.2.4.

Proof of Prop. 4.5.4 We rely on the following properties of Gaussian/truncated Gaussian random variable.

Property 4.D.1. *Let $\epsilon \sim \mathcal{N}(0,1)$, for any $a \geq \sqrt{2\log(6)}$, $\mathbb{V}(\epsilon \mid |\epsilon| \leq a) \geq 1/2$.*

Proof. Explicit formula for the truncated normal distribution moment leads to

$$\mathbb{V}(\epsilon \mid |\epsilon| \leq a) = 1 - \frac{2a}{\sqrt{2\pi}} \frac{e^{-a^2/2}}{\mathbb{P}(|\epsilon| \leq a)} \geq 1 - \frac{ae^{-a^2/2}}{\mathbb{P}(|\epsilon| \leq a)}.$$

Standard inequality for the Gaussian cdf guarantees that $\mathbb{P}(|\epsilon| \leq a) \leq e^{-a^2/2}$. Hence, for all $a \geq \sqrt{2\log(6)}$,

$$\mathbb{V}(\epsilon \mid |\epsilon| \leq a) \geq 1 - \frac{ae^{-a^2/2}}{1 - e^{-a^2/2}} \geq \frac{\sqrt{2\log(6)}}{5} \geq 1/2$$

□

Property 4.D.2. *Let $\epsilon \sim \mathcal{N}(0,1)$, for any $n \geq 2$, for any $a \geq \sqrt{2\log(n)}$, $\mathbb{P}(|\epsilon| \leq a)^n \geq 1/4$.*

Proof. Again, since $\mathbb{P}(|\epsilon| \leq a) \leq e^{-a^2/2}$, one has, for any $a \geq \sqrt{2\log(n)}$, $\mathbb{P}(|\epsilon| \leq a) \leq (1 - 1/n)^n \geq 1/4$ □

Proof of Prop. 4.5.4. Denote as $\bar{x}_t = \mathbb{E}(x_t|\bar{\theta}_t, \mathcal{F}_{t-1}, E_{t-1})$ and $\Sigma_t = \mathbb{V}(x_t|\bar{\theta}_t, \mathcal{F}_{t-1}, E_{t-1})$. By Prop. 4.5.3 $\|\Sigma_t\|_2 \leq 1$, hence, one has,

$$\begin{aligned}
\mathbb{E}(x_t x_t^\mathsf{T} \mathbb{1}_{\{\|x_t\| \leq \alpha_t\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}) &\geq \mathbb{E}(x_t x_t^\mathsf{T} \mathbb{1}_{\{\|x_t - \bar{x}_t\| \leq \alpha_t - \rho\|\bar{x}_t\|\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}), \\
&\geq \Sigma_t^{1/2} \mathbb{E}(y_t y_t^\mathsf{T} \mathbb{1}_{\{\|y_t - \bar{y}_t\| \leq \alpha_t - \rho\|\bar{x}_t\|\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}) \Sigma_t^{1/2}, \\
&\geq \Sigma_t^{1/2} \mathbb{E}(y_t y_t^\mathsf{T} \mathbb{1}_{\{|y_t^i - \bar{y}_t^i| \leq \frac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}}, \forall i \leq n\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}) \Sigma_t^{1/2},
\end{aligned}$$

where $y_t = \Sigma_t^{-1/2} x_t$, $\bar{y}_t = \Sigma^{-1/2}\bar{x}_t$ and y^i denotes the i^{th} coordinate of the n-dimensional vector y. By definition $y_t|_{\bar{\theta}_t, \mathcal{F}_{t-1}, E_{t-1}} \sim \mathcal{N}(\bar{y}_t, I)$ thus,

$$\begin{aligned}
&\mathbb{E}(y_t y_t^\mathsf{T} \mathbb{1}_{\{|y_t^i - \bar{y}_t^i| \leq \frac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}}, \forall i \leq n\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}) \\
&\qquad = \mathbb{P}(|y_t^i - \bar{y}_t^i| \leq \tfrac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}}, \forall i \leq n) \mathbb{E}(y_t y_t^\mathsf{T} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}, \{|y_t^i - \bar{y}_t^i| \leq \tfrac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}}, \forall i \leq n\}) \\
&\qquad \geq \mathbb{P}(|y_t^i - \bar{y}_t^i| \leq \tfrac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}}, \forall i \leq n) \mathbb{V}(y_t | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}, \{|y_t^i - \bar{y}_t^i| \leq \tfrac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}}, \forall i \leq n\}) \\
&\qquad = \mathbb{P}(|\epsilon| \leq \tfrac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}})^n \mathbb{V}(\epsilon \mid |\epsilon| \leq \tfrac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}}) I,
\end{aligned}$$

where $\epsilon \sim \mathcal{N}(0,1)$. Noticing that $\frac{\alpha_t - \rho\|\bar{x}_t\|}{\sqrt{n}} = \sqrt{2\log(3n)}$, Properties. 4.D.1- 4.D.2 holds and $\mathbb{E}(y_t y_t^\mathsf{T} \mathbb{1}_{\{|y_t^i - \bar{y}_t^i| \leq \frac{\alpha - \rho\|\bar{x}_t\|}{\sqrt{n}}, \forall i \leq n\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}) \geq 1/8 I$. As a results,

$$\mathbb{E}(x_t x_t^\mathsf{T} \mathbb{1}_{\{\|x_t\| \leq \alpha_t\}} | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1}) \geq 1/8 \Sigma_t = 1/8 \mathbb{V}(x_t | \mathcal{F}_{t-1}, \bar{\theta}_t, E_{t-1})$$

□

Chapter 5

Application to Portfolio Construction

In this chapter[1], we apply exploration-exploitation techniques to the problem of portfolio construction. While this is a standard in finance since the seminal work of Markowitz, several approaches have been recently proposed to integrate the price impact effects i.e., the fact that buying and selling shares modifies the supply and demand and hence the price at which transactions are made. This makes the control problem significantly more difficult since one has to anticipate the future costs implied by price impact. Moreover, by nature, the dynamics of financial markets are unknown and have to be estimated. However, in order to observe the price impact effects, asset managers have to trade directly on the market and this may induce the exploration-exploitation trade-off problem to balance between trading to make profit and trading to gain knowledge about the market dynamics. We introduce a novel LQ formulation for the portfolio allocation problem, under the assumption of linear price dynamics, from which we obtain the optimal control and discuss the exploration-exploitation trade-off arising from the presence of unknown parameters. We consider two problem instances with or without risk constraint and show that this affects the need for exploration. In the unconstrained case, a greedy strategy fails to achieve sub-linear regret, while Thompson Sampling or optimism-based algorithms effectively trade-off exploration and exploitation. On the other hand, the risk constraint modifies the structure of the policy, removing somehow the need for active exploration, and a greedy strategy is optimal. We discuss this counter-intuitive result and support it with numerical experiments.

Contents

[1]This chapter is based on our paper (Abeille et al., 2016) for generic portfolio construction using LQG systems that is submitted to International Journal of Theoretical and Applied Finance (under review).

5.1 Introduction

Modern finance theory is often thought to have started with the mean-variance approach of Markowitz (1952). This approach provides portfolio managers with a systematic treatment of the risk-return tradeoff by maximizing their own utility. This started intensive research to further develop the basic mean-variance theory. In particular, it raised questions about the relationship between risk and return, leading to the celebrated CAPM model (Sharpe, 1964, Jensen et al., 1972) as well as finer modeling for the risk structure and the return predictability (Fama and French, 1993). However, one of the limitations of these approaches is their inability to take transaction costs into account: in a multi-step setting, performing such a strategy may be highly suboptimal as the rebalancing cost can be worse than the expected gain. This observation, of crucial importance for practitioners, led to dynamic allocation rules where portfolio managers anticipate this additional cost and track the Markowitz position by constraining the turnover (see e.g., Constantinides 1979, Taksar et al. 1988, Morton and Pliska 1995, Grinold 2010).
When the volume of the transaction is large compared to the available liquidity, another effect known as price impact induces transaction costs: the execution of a large order drastically changes the supply and demand and thus affects the price in an adverse manner. The understanding of the market impact and the way to minimize it is an important topic for large investors and a large amount of literature addresses this question from different perspectives. Motivated by stylized facts and empirical studies which stress that markets digest very slowly modifications induced by large trade (Bouchaud et al., 2008, Brokmann et al., 2014), Mastromatteo et al. (2014), Donier et al. (2014) derived a microstructure based model for the price impact from the dynamic of the latent order book. Following the work of Kyle (1985), another stream of literature considers agent-based model to understand how information is incorporated into the prices and how it affects the liquidity. Huberman and Stanzl (2004), Gatheral (2010) study the effect of the price impact on the absence of price manipulation and derive various inequalities about the shape of the price impact function. On the other hand, a large part of the literature is dedicated to the minimization of the price impact. Two types of problem are usually considered: optimal execution (see e.g., Bertsimas and Lo 1998, Almgren and Chriss 2001, Guéant 2012, Obizhaeva and Wang 2013) where investors seek to liquidate a given position within a certain period, and optimal allocation (see e.g., Gârleanu 2009, Lataillade et al. 2012, Gârleanu and Pedersen 2013, Kallsen and Muhle-Karbe 2013, Moreau et al. 2014) where investors try to dynamically control a portfolio to maximize their risk-profit utility under price impact. Finally, Park and Van Roy (2015) consider the optimal allocation problem of Gârleanu and Pedersen (2013) and introduce an adaptive algorithm, called CTRACE, that performs both the estimation and control when the impact model is unknown.

In this chapter, we consider the optimal portfolio allocation problem when the market exhibits dynamical return predictability and price impact. Our setting is close to the one of Park and Van Roy (2015), the main difference being that they assume the predictable part of the returns model to be known and focus only on the estimation

of the impact, while we assume both to be unknown and estimate them jointly. We consider an investor who wants to dynamically allocate a portfolio of N assets in order to optimize a multi-horizon Markowitz cost function. Under the assumption that the market dynamics are linear, we propose a novel approach that allows us to cast this problem as a LQ problem, from which the optimal controller can be computed in an efficient way by solving the associated Riccati equation. Our method holds for any linear Markovian return dynamics (Abeille et al., 2016) and thus can handle most of the standard financial model. In order to illustrate and discuss the exploration-exploitation issue in portfolio allocation, we focus here a synthetic example that exhibits the key features of the prices dynamics while remaining simple enough to provide intuition. We first describe the setting that we consider, how to encode the problem into a LQ framework and the obtained model, as well as its optimal solution. Additionally, we stress that the existence and uniqueness of the solution is directly related to the non-arbitrage property of the market model, highlighting the one-to-one relationship between the LQ theory and the financial intuition.
Despite the generality and the flexibility of this LQ formulation, the main issue is that the parameters of the state-space are unknown and have to be estimated online i.e., while trading, since the impact effects are only generated by trades. Leveraging ideas presented in Ch. 4, we estimate the unknown parameters by Least Square (RLS) and investigate the exploration-exploitation trade-off in this setting. We compare three different algorithms: Certainty Equivalence (CE) which is the greedy strategy that consists in trading optimally given the current RLS estimate, Thompson Sampling (TS) where the exploration is based on randomization, and Optimism in Face of Uncertainty (OFU-LQ), where the exploration is based on optimism. We focus on two instances of the problem: at first, we consider the portfolio allocation problem without risk constraint, and show that the CE algorithm incurs a linear regret, while both TS and OFU-LQ incur a $O(\sqrt{T})$ regret. Then, we address the portfolio allocation problem with risk constraint and show that the regret of the greedy strategy is $O(\log T)$, thus optimal, which means that no additional exploration is needed in this setting. Finally, we discuss why this surprising result is implied by the structure of the controller.

5.2 Setting the stage

5.2.1 The portfolio allocation control problem

Setting. We consider an investor whose objective is to dynamically construct a portfolio of N assets: at each time step, it can decide to rebalance his portfolio, using his current knowledge, in order to optimize his gain - encoded in the Profit and Loss (PnL) measure - with respect to a cost function that represents the risk-return trade-off. The seminal work of Markowitz suggests to balance between minimizing the PnL variance (i.e., the risk) and maximizing the PnL expectation (i.e., the return). Formally, one aims to

minimize

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\Big(\gamma\mathbb{V}(PnL_{t,t+1} \mid \mathcal{F}_t) - \mathbb{E}(PnL_{t,t+1} \mid \mathcal{F}_t)\Big) \;\Big|\; \mathcal{F}_0\right] \tag{5.1}$$

where $PnL_{t,t+1}$ represents the increment in the investor wealth, γ is a risk-tuning parameter and $\mathcal{F}_t$ represents the information accumulated so far. To take into account the transaction costs, we explicitly do the distinction between the *decision prices* p_t that are observed at every time step t and used to take the trading decision and the *execution prices* $\bar{p}_{t,t+1}$, at which transactions occur, that are observed after the trade. Let Q_t be the inventory position at time step t i.e., the number of shares of each asset hold at time t, and q_t be the trade executed at time t one has:

$$PnL_{t,t+1} = Q_{t+1}^\mathsf{T} p_{t+1} - Q_t^\mathsf{T} p_t - q_t^\mathsf{T}\bar{p}_{t,t+1}, \qquad Q_{t+1} = Q_t + q_t. \tag{5.2}$$

The introduction of the *execution prices* is of crucial importance when considering impact effects due to large trade: indeed, the *decision prices* correspond to the prices at which the supply and demand meet, which by definition is of zero liquidity. Whenever an investor seeks to buy (resp. sell) a large amount of shares, it has to offer a higher (resp. lower) price in order to find enough liquidity. A reasonable model is to assume that the execution prices are on average of the current and next decision prices as $\bar{p}_{t,t+1} = \eta p_{t+1} + (1-\eta)p_t$. Finally, for sake of simplicity we assume that $\eta = 1$ which corresponds to a setting where the investor executes at the next decision price. As a consequence, one can rewrite Eq. 5.2 as

$$PnL_{t,t+1} = Q_t^\mathsf{T} r_{t+1}, \qquad r_{t+1} = p_{t+1} - p_t, \qquad Q_{t+1} = Q_t + q_t.$$

As claimed in the introduction, we assume that returns r_{t+1} follow a linear model, and exhibit both impact effect and return predictability. We encode this into a state space of the form:

$$\begin{aligned} \alpha_{t+1} &= \Phi_\alpha^\star \alpha_t + \epsilon_{t+1}^\alpha \\ I_{t+1} &= \Phi_I^\star I_t + q_t \\ r_{t+1} &= \beta_p^\star \alpha_t + \beta_I^\star I_t + \beta_q^\star q_t + \epsilon_{t+1}^r \end{aligned} \tag{5.3}$$

where ϵ_{t+1}^α and ϵ_{t+1}^r and zero-mean noises, conditionally independent and of identity variance, for sake of simplicity. $\Phi_\alpha^\star$, $\Phi_I^\star$, $\beta_p^\star$, $\beta_I^\star$ and $\beta_q^\star$ are matrices of parameters of appropriate dimensions. $\{\alpha_t\}_{t\geq 0}$ is a stochastic process that represents the predictable part of the returns, while $\{I_t\}_{t\geq 0}$ is a deterministic process (w.r.t. trades) which takes into account the trades executed so far and quantifies the price impact effects. To visualize the dynamic of such systems, Impulse Response (IR) stands as a convenient tool: it consists in separately perturbing the inputs of the system (here ϵ_{t+1}^α and q_t) and let the system evolves. We plot them in Fig. 5.1 for 1 dimensional system ($r_{t+1} \in \mathbb{R}$), and describe the system dynamic. Similarly, to visualize the PnL dynamic, we plot in Fig. 5.2 the round trip response i.e., the PnL trajectory induced by a round-trip that consists in buying 1 share for the first 10 time steps, let the system evolve for the next 10 time steps, and selling back the 10 shares for the last 10 time steps to have a zero position at the end.

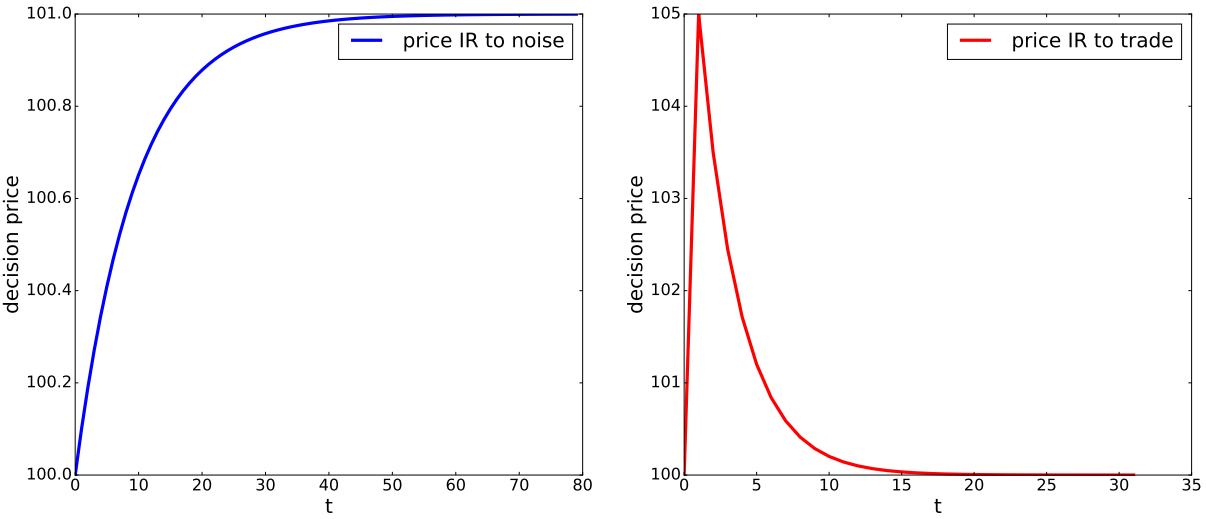

Figure 5.1 – Impulse Response of the decision price. *Left:* An impulse of the predictor noise pushes the price up and the growth is exponential thanks to the auto-regressive dynamic of α_t. *Right:* The impulse of the trade summarizes the impact model. The price is pushed up by the instantaneous impact γq_t, then decreases exponentially, thank to the mean-reverting dynamic of I_t.

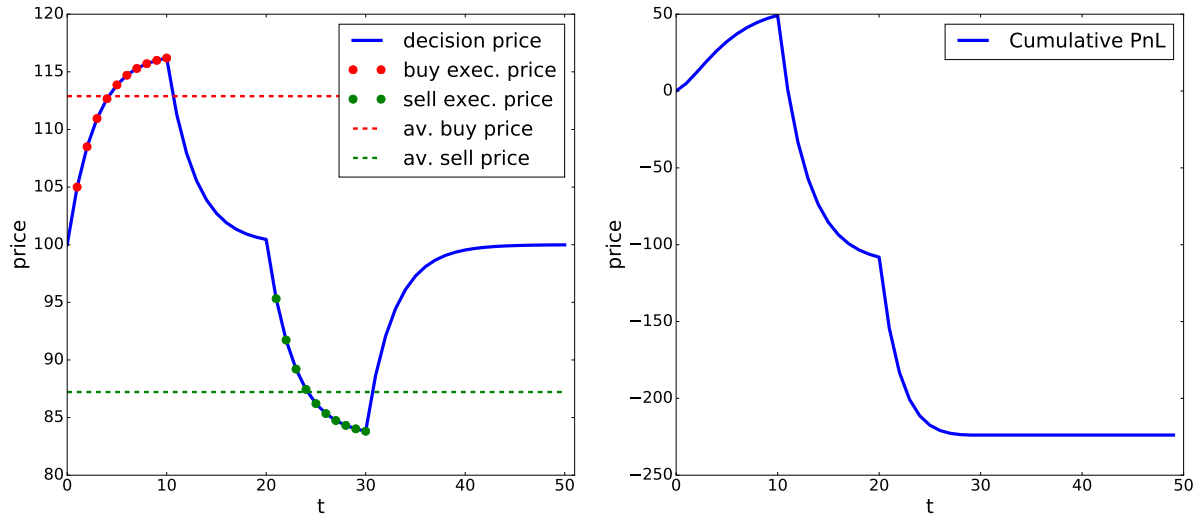

Figure 5.2 – Round trip response of the open loop system. The trade sequence consists in buying 1 share for each of the first 10 time steps, do nothing for 10 time steps and sell back the position the same way. *Left:* Decision and execution prices trajectories. The decision price is pushed up by the purchase and exhibits a concavity induced by the impact relaxation of past trades. The execution is made at the next decision price. Symmetrically, the decision price is pushed down when selling. On average over the trajectory, the execution price is high for buy transactions, low for sell transactions, which stresses the adverse effect of price impact. *Right:* PnL trajectory. Since the impact effects act in an adverse manner, the overall PnL is negative. If it increases at the beginning of the trajectory due to the purchase, it is purely artificial as the price mean-revert to zero when the trading stops, thus inducing a loss. This effect is accelerated when selling, because the price impact is symmetric.

LQ formulation. We introduce the state variable $x_t = \begin{pmatrix} Q_t \\ \alpha_t \\ I_t \end{pmatrix}$, and write Eq. 5.3 as

$$\begin{aligned} x_{t+1} &= A^\star x_t + B^\star q_t + \epsilon^x_{t+1}, \\ r_{t+1} &= \begin{pmatrix} 0 & \beta^\star_\alpha & \beta^\star_I \end{pmatrix} x_t + \beta^\star_q q_t + \epsilon^r_{t+1}, \end{aligned} \tag{5.4}$$

where

$$A^\star := \begin{pmatrix} I & 0 & 0 \\ 0 & \Phi^\star_\alpha & 0 \\ 0 & 0 & \Phi^\star_I \end{pmatrix}, \qquad B^\star := \begin{pmatrix} I \\ 0 \\ I \end{pmatrix}, \qquad \epsilon^x_{t+1} = \begin{pmatrix} 0 \\ \epsilon^\alpha_{t+1} \\ 0 \end{pmatrix}.$$

Further, we express the cost function $c_t = \gamma \mathbb{V}(PnL_{t,t+1}|\mathcal{F}_t) - \mathbb{E}(PnL_{t,t+1}|\mathcal{F}_t)$ as a function of x_t and q_t:

$$\gamma \mathbb{V}(PnL_{t,t+1}|\mathcal{F}_t) - \mathbb{E}(PnL_{t,t+1}|\mathcal{F}_t) = x_t^\intercal Q^\star x_t + 2 x_t^\intercal N^\star q_t + q_t^\intercal R^\star q_t,$$

where

$$Q^\star := \frac{1}{2} \begin{pmatrix} 2\gamma I & -\beta^\star_p & -\beta^\star_I \\ -\beta^{\star,\intercal}_p & 0 & 0 \\ -\beta^{\star,\intercal}_I & 0 & 0 \end{pmatrix}, \quad R^\star := \begin{pmatrix} 0 \end{pmatrix}, \quad N^\star := \frac{1}{2} \begin{pmatrix} -\beta^\star_q \\ 0 \\ 0 \end{pmatrix}, \quad \mathcal{Q}^\star := \begin{pmatrix} Q^\star & N^\star \\ N^{\star,\intercal} & R^\star \end{pmatrix}.$$

Therefore, solving the dynamical Markowitz problem of Eq. 5.1 reduces to find the stationary deterministic control policy π mapping states to trades that minimizes the performance measured by the asymptotic (i.e., infinite horizon) average expected cost

$$J_\pi = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} x_t^\intercal Q^\star x_t + 2 x_t^\intercal N^\star q_t + q_t^\intercal R^\star q_t \right], \tag{5.5}$$

with $x_0 = 0$ and $q_t = \pi(x_t)$. Since x_t follows the linear dynamics of Eq. 5.4, we retrieve a similar LQ formulation to the one introduced in Subsec. 2.3.2.[2]

Optimal control. As opposed to the LQ problem of Ch. 4, the cost matrices $Q^\star, R^\star, N^\star$ are no longer known but depend on the parameters of the returns model. However, they still can be estimated through a linear regression thanks to Eq. 5.3. Additionally, even if $(A^\star, B^\star)$ is a stabilizable pair, the mapping of the portfolio allocation problem into a LQ problem implies that the matrix $\mathcal{Q}^\star$ is no longer positive definite which breaks the standard LQ assumptions (Asm. 2.3.1). Fortunately, one can still compute an optimal control under a more flexible criterion that we link to non-arbitrage.

Assumption 5.2.1 (Noise)**.** *The noises $\{\epsilon^\alpha_t\}_t$ and $\{\epsilon^r_t\}_t$ are $\mathcal{F}_t-$martingale difference sequences, conditionally independent, where $\mathcal{F}_t$ is the filtration which represents the accumulated information up to time t. Furthermore, we assume that $\mathbb{V}(\epsilon^\alpha_{t+1}|\mathcal{F}_t) = I$ and $\mathbb{V}(\epsilon^r_{t+1}|\mathcal{F}_t) = I$ for all $t \geq 0$.*

[2]This formulation slightly differs as it involves a cross-term $x_t^\intercal N q_t$, although it remains quadratic. Thanks to a change of variable, it is possible to retrieve the original formulation of Subsec. 2.3.2 (see (Lancaster and Rodman, 1995)).

Assumption 5.2.2 (LQ). *The matrices $\Phi_\alpha^\star$ and $\Phi_I^\star$ are stable and the matrices $A^\star, B^\star, Q^\star, R^\star, N^\star$ are such that there exists a control matrix K which stabilizes the system (i.e., $A^\star + B^\star K$ is stable) such that the hermitian matrix $\Psi_K(z) > 0$ for all $|z| = 1$, $z \in \mathbb{C}$, where*

$$\Psi_K(z) = Y_K^\intercal(z^{-1}) \begin{pmatrix} (Iz^{-1} - A^\star)^{-1} B^\star \\ I \end{pmatrix}^\intercal H^\star \begin{pmatrix} (Iz - A^\star)^{-1} B^\star \\ I \end{pmatrix} Y_K(z)$$
$$Y_K(z) = I + K(Iz - A^\star - B^\star K)^{-1} B^\star.$$

Asm. 5.2.2 relies on a criterion over the Popov function Ψ_K to guarantee the existence and uniqueness of an admissible controller (see Molinari 1975). While this stands as a technical tool for LQ in general, it is possible to link it to a financial argument in our specific portfolio allocation setting. In Lem. 5.2.1, we show that this is equivalent to the fact that the system allows no dynamical arbitrage i.e., that, in the absence of predictability, there exists no strategy which guarantees a positive profit. This "no free lunch" assumption is standard in finance, and justifies the validity of Asm. 5.2.2. We postpone the proof to App. 5.A.

Lemma 5.2.1. *We denote as $l^p := \{(x_n)_{n\geq 0} \in \mathbb{R}^{\mathbb{N}} \text{ s.t. } \sum_{n\geq 0} |x_n|^p < \infty\}$ Consider the deterministic system associated with Eq. 5.3 where the sequences of noise $\{\epsilon_t^\alpha\}_t$ and $\{\epsilon_t^r\}_t$ are set to zero. Let $\mathcal{RT}$ be the set of admissible round-trip defined as*

$$\mathcal{RT} = \{q = (q_0, q_1, \dots) \in l^1 \cap l^2\}, \qquad \text{if } \gamma = 0,$$
$$\mathcal{RT} = \{q = (q_0, q_1, \dots) \in l^1 \cap l^2 \quad s.t. \quad (Q_0, Q_1, \dots) \in l^1 \cap l^2\} \qquad \text{if } \gamma \in (0, \infty),$$

where $l^1 = \{(q_n)_{n\geq 0} \text{ s.t. } \sum_{n\geq 0} |q_n| < \infty\}$ and $l^2 = \{(q_n)_{n\geq 0} \text{ s.t. } \sum_{n\geq 0} |q_n|^2 < \infty\}$. Let $\Psi_K(z)$ be the Popov function of Asm. 5.2.2. Then, $\Psi_K(z) > 0$ if and only if, for any $q \in \mathcal{RT}$, the PnL trajectory induced by q is such that $\sum_{t=0}^\infty PnL_{t,t+1} \leq 0$.

Under these assumptions, a slightly modified version of Thm. 2.3.2 (see Ch. 2, Subsec. 2.3.2) provides that the optimal policy is linear with the state, i.e., $\pi_* = Kx$, where the optimal gain K is computed as

$$K = -\left(R + B^{\star,\intercal} P B^\star\right)^{-1} [B^{\star,\intercal} P A^\star + N^{\star,\intercal}],$$
$$P = Q^\star + A^{\star,\intercal} P A^\star + [A^{\star,\intercal} P B^\star + N^\star] K.$$

P is the unique solution to the Riccati equation associated with the control problem. The optimal average cost is $J = J_{\pi_*} = \mathrm{Tr}(P\Sigma^x)$ with $\Sigma^x = \mathbb{V}(\epsilon_{t+1}^x | \mathcal{F}_t) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Finally, we also have the closed-loop matrix $A^\star + B^\star K$ is asymptotically stable.

5.2.2 The learning problem

Structure of the parameters. As claimed in introduction, one of the major issue in this dynamic portfolio allocation problem is that the dynamic of the market is unknown by nature which translates in unknown parameters contained in $\Phi_\alpha^\star, \Phi_I^\star, \beta_\alpha^\star, \beta_I^\star, \beta_q^\star$. However, one can reasonably assume that part of those parameters are known to the controller, or at least can be estimated separately. We make the following assumptions and discuss their validity.

Assumption 5.2.3. *The predictable part of the returns model is observed at each time step i.e., α_t is $\mathcal{F}_t$-measurable and parameters contained in $\Phi_\alpha^\star$ are known. Furthermore, we assume that this representation is minimal i.e., the eigenvalues of $\Phi_\alpha^\star$ are pairwise disjoint and that $\beta_\alpha^\star$ is of full row-rank.*

Assumption 5.2.4. *The parameters contained in $\Phi_I^\star$ are known. As a consequence, the impact variable I_t can be reconstructed from past trades (which are $\mathcal{F}_t$-measurable) and thus, I_t is $\mathcal{F}_t$-measurable.*

Since the variable α_t encodes the prediction made by the investor about the future prices move, it is reasonable to assume that it is observed at each time step. Furthermore, its dynamics is independent from the sequence of trades as it takes into account the uncontrolled prices move. As a result, $\Phi_\alpha^\star$ can be estimated separately and its estimation converges almost surely to the true parameters (as provided by autoregressive process estimation) which implies that this does not participate to the exploration-exploitation dilemma. For sake of convenience, we assume that $\Phi_\alpha^\star$ is known beforehand. Finally, the minimal representation assumption is here to ensure that the prediction model is not over-parametrized (all predictors are different) and significant (all predictors influence the returns dynamic).
On the other hand, there is no reason for $\Phi_I^\star$ to be known a priori. Moreover, the impact variable I_t has no physical meaning but takes into account the fact that the market digests slowly the modification induced by trades. Said differently, it is not a hidden variable but a latent variable that is here to encode the structure of the impact decay. However, one can overcome this difficulty using an over-parametrized state-space model of higher dimension for the learning than the true one. This idea is already used by Park and Van Roy (2015): consider the 1-dimensional case where $r_t \in \mathbb{R}$ and $I_t \in \mathbb{R}$; focusing on the impact effect only, the return modeling has a dynamic characterized by

$$I_{t+1} = \Phi_I^\star I_t + q_t; \qquad r_{t+1} = \beta_I^\star I_t + \beta_q^\star q_t.$$

Under Asm. 5.2.2, $\Phi_I^\star$ is stable which means that $|\Phi_I^\star| < 1$ and by invariance, one can set $\Phi_I^\star \geq 0$. Hence, it is possible to approximate the above model using a vector $\Phi_I = (\Phi_I^0, \dots, \Phi_I^s)$ which spans $[0, 1[$ together with an augmented state $\tilde{I}_t \in \mathbb{R}^s$ whose dynamic follows

$$\tilde{I}_{t+1} = \begin{pmatrix} \Phi_I^0 & & 0 \\ & \ddots & \\ 0 & & \Phi_I^s \end{pmatrix} \tilde{I}_t + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} q_t; \qquad r_{t+1} = \beta_I^\intercal \tilde{I}_t + \beta_q^\star q_t.$$

In this case, each component of $\tilde{I}_t$ represents a different impact model whose effect on the return are weighted by the component of β_I. As a result, the underlying difficulty in learning parameters $\Phi_I^\star$, $\beta_I^\star$ and $\beta_q^\star$ can be summarized into the estimation of β_I and $\beta_q^\star$ as long as Φ_I is sufficiently dense in $[0, 1[$. In particular if Φ_I contains the true value $\Phi_I^\star$, the augmented model includes the true one (where β_I components are all zero but one equal to $\beta_I^\star$) and the two estimation problems are equivalent. This stresses that Asm. 5.2.4 does not reduce the learning complexity of the problem, but acts as a choice of representation for the impact model. In particular, this does not influence the exploration-exploitation problem. Neglecting the approximation error, it is therefore possible to assume that $\Phi_I^\star$ is known by the controller and since, given $\Phi_I^\star$, I_t can be reconstructed from past trades, that the impact variable is observed at each time step. Finally, under Asm. 5.2.3 and 5.2.4, the matrices $A^\star, B^\star$ of Eq. 5.4 are known while the unknown parameters are contained in $\beta_q^\star, \beta_\alpha^\star, \beta_I^\star$. We collect them in $\theta^{\star,\mathsf{T}} = [\beta_q^\star, \beta_\alpha^\star, \beta_I^\star]$. Since $\theta^\star$ is used to construct the costs matrices $Q^\star, R^\star, N^\star$, it determines the LQ problem. Thus, for any θ satisfying Asm. 5.2.2, we denote as $P(\theta)$, $K(\theta)$ and $J(\theta)$ the optimal LQ quantities associated with this parametrization.

Regret definition. We consider the standard online learning setting where at each step t the learner receives the current state x_t as input, the current return r_t as observation, it executes a control q_t and it suffers the associated cost

$$c_t = c_{\theta^\star}(x_t, q_t) = x_t^\mathsf{T} Q^\star x_t + 2x_t^\mathsf{T} N^\star q_t + q_t^\mathsf{T} R^\star q_t$$

The system then transitions to the next state x_{t+1} and the next observation r_{t+1} is generated according to Eq. 5.4. The learning performance is measured by the cumulative regret over T steps, where the costs cumulated over time are compared to the minimal cost obtained on average by the optimal policy. Formally we define

$$R_T(\theta^\star) = \sum_{t=0}^{T} \Big(c_t - J(\theta^\star) \Big)$$

RLS estimates. From a learning perspective, this LQ problem slightly differs from the one of Ch. 4 although from a control point of view they share the same structure. In particular, the unknown parameter $\theta^\star$ is used to defined the cost function rather than the state dynamics.[3] On the other hand, one still needs an estimate about its value to compute the optimal control, thus requiring basic tools for the estimation of the parameter $\theta^\star$. Let $(q_0, \dots, q_t)$ be the sequence of trades, $(\alpha_0, \dots, \alpha_t)$ the sequence of predictors, $(I_0, \dots, I_t)$ the sequence of impact variables, and let $(r_1, \dots, r_{t+1})$ be the corresponding returns (i.e., observations) generated according to Eq. 5.4. For any regularization parameter $\lambda \in \mathbb{R}_+^*$ the regularized least-squares estimate (RLS) and the associated design matrix are defined as

$$V_t = \lambda I + \sum_{s=0}^{t-1} z_s z_s^\mathsf{T}; \qquad \widehat{\theta}_t = V_t^{-1} \sum_{s=0}^{t-1} z_s r_{s+1}^\mathsf{T}, \tag{5.6}$$

[3] Using state-space manipulations, it is possible to re-write the problem in the same form as in Ch. 4. However, this requires to augment the dimension of the system, thus for sake of clarity, we prefer to stick with this minimal model.

where $z_t^\mathsf{T} = (q_t^\mathsf{T}, \alpha_t^\mathsf{T}, I_t^\mathsf{T})$.

5.3 Algorithms

In Sec. 5.2, we showed how the portfolio allocation problem of Eq. 5.1 can be cast as an LQ problem, which allows us to compute the optimal control given the parameters of the price dynamics. Additionally, we discussed how, under reasonable assumptions about the parameters knowledge, one can use RLS to estimate the unknown component, thus providing us with all the material needed to perform adaptive strategies. We present in this section the algorithms CE, TS and the OFU-LQ (derived from the algorithm presented in Fig. 2.10, for the specific portfolio allocation problem instance) that we consider to address the exploration-exploitation trade-off in portfolio optimization. All of them are based on policy updates that consists in choosing a new parameter and following the policy which is optimal w.r.t this choice. To ensure that those parameters are coherent with Asm. 5.2.2 and 5.2.3, we constrain them to belong to an admissible set $\mathcal{S}$ that will be specified in the next section. We report in Fig. 5.1 the common structure of the three algorithms and then precise separately the sub-routine TRIGGER and SELECT that are specific to CE, TS and OFU-LQ and determine respectively the frequency of updates and the choice of the parameters.

Input: $\hat{\theta}_0$, $V_0 = \lambda I$, δ, T, $\mathcal{S}$, τ
Set $\delta' = \delta/(8T)$
for $t = \{0, \dots, T\}$ **do**
  if TRIGGER$(t, t_0, \tau, V_0, V_t) = 1$ **then**
    Select a new parameter $\widetilde{\theta}_t = \text{SELECT}(\widehat{\theta}_t, V_t, \delta', \mathcal{S})$
  else
    $\widetilde{\theta}_{t+1} = \widetilde{\theta}_t$
  end if
  Execute control $q_t = K(\widetilde{\theta}_t)x_t$
  Observe state x_{t+1}, observation r_{t+1} and suffer cost $c_t = c_{\theta^\star}(x_t, q_t)$
  Compute V_{t+1} and $\widehat{\theta}_{t+1}$ using Eq. 5.6
end for

Figure 5.1 – Adaptive algorithm for Portfolio allocation.

Greedy strategy. First, we describe the method known as the Certainty Equivalence (CE) in the control literature (Kumar and Varaiya, 2015), that is a greedy strategy where the optimal control is computed with respect to the current RLS estimate. While this algorithm is known to fail to achieve sub-linear regret in standard bandit problems, we use it as a base case to highlight the need or not for explicit exploration, and surprisingly notice that it offers very good performance as soon as we consider the risk constrained problem (see Subsec. 5.4.2).
To reduce the computational complexity of the algorithm, we keep the same policy for

episode of constant length according to the TRIGGER sub-routine (see Fig. 5.2). Then, we select the greedy parameter $\widetilde{\theta}_t \in \mathcal{S}$ by projecting $\widehat{\theta}_t$ onto $\mathcal{S}$.

Input: t, t_0, τ
if $t \geq t_0 + \tau$ **then**
    TRIGGER$(t, t_0, \tau) = 1$ and $t_0 = t$
else
    TRIGGER$(t, t_0, \tau) = 0$
end if

Figure 5.2 – CE TRIGGER sub-routine.

Input: $\widehat{\theta}_t$, $\mathcal{S}$
if $\widehat{\theta}_t \in \mathcal{S}$ **then**
    $\widetilde{\theta}_t = \widehat{\theta}_t$
else
    $\widetilde{\theta}_t = \arg\min_{\theta \in \mathcal{S}} \|\theta - \widehat{\theta}_t\|^2$
end if

Figure 5.3 – CE SELECT sub-routine.

Randomized strategy. To overcome the potential failure of the CE algorithm, we consider the TS algorithm of Ch. 4. As discussed in the previous chapters, the idea is to sample new parameters $\widetilde{\theta}_t$ taking into account the knowledge acquire so far, around the current RLS estimate $\widehat{\theta}_t$ but on the basis of the uncertainty i.e., with a variance that depends on the design matrix V_t. Each time the policy is re-evaluted, a parameter $\widetilde{\theta}_t$ is sampled, and the optimal policy w.r.t to $\widetilde{\theta}_t$ is executed. Formally, $\widetilde{\theta}_t = \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t V_t^{-1/2} \eta_t)$, where $\mathcal{R}_{\mathcal{S}}$ is the rejection sampling operator associated with the admissible set $\mathcal{S}$, $\widehat{\theta}_t$ is the RLS-estimate, V_t is the design matrix and each entry of the perturbation matrix $\eta_t \in \mathbb{R}^{(n+d)\times n}$ is a random sample drawn i.i.d. from $\mathcal{N}(0,1)$. The SELECT sub-routine in presented in Fig. 5.4, while the TRIGGER sub-routine is identical to the CE (Fig. 5.2).

Input: $\widehat{\theta}_t$, V_t, δ',$\mathcal{S}$
Compute $\beta_t = \beta_t(\delta')$ from Eq. 2.8
while $\widetilde{\theta}_t \notin \mathcal{S}$ **do**
    Sample $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t V_t^{-1/2} \eta_t$ where η_t is component-wise $\mathcal{N}(0,1)$
end while

Figure 5.4 – TS SELECT sub-routine.

Optimistic strategy. Finally, we also compare the performance of CE and TS to the OFU-LQ algorithm introduced in (Abbasi-Yadkori and Szepesvári, 2011) presented in Fig. 2.10. We recall the TRIGGER and SELECT sub-routine in Fig. 5.5 and 5.6, based respectively on the *doubling-schedule* and on *optimism*.

Input: V_0, V_t
if $\det(V_t) \geq 2\det(V_0)$ **then**
  TRIGGER$(V_0, V_t) = 1$
  $V_0 = V_t$
else
  TRIGGER$(V_0, V_t) = 0$
end if

Figure 5.5 – OFU-LQ TRIGGER sub-routine.

Input: $\widehat{\theta}_t$, V_t, δ',$\mathcal{S}$
Compute $\beta_t = \beta_t(\delta')$ from Eq. 2.8
Define $\mathcal{E}_t^{\text{RLS}} = \{\theta \text{ s.t. } \text{Tr}[(\hat{\theta}_t - \theta)^\intercal V_t(\hat{\theta}_t - \theta)] \leq \beta_t^2\}$
Find $\widetilde{\theta}_t = \text{argmin}_{\theta \in \mathcal{E}_t^{\text{RLS}} \cap \mathcal{S}} J(\theta)$

Figure 5.6 – OFU-LQ SELECT sub-routine.

From a theoretical perspective, applying the results of Ch. 4 and the one of Abbasi-Yadkori and Szepesvári (2011), one obtains the following guarantees for the regret of TS of OFU-LQ.

Corollary 5.3.1 (From Thm. 4.5.1)**.** *Consider the portfolio allocation problem in Eq. 5.1 of dimension $n = 1$. Under Asm. 5.2.1, 5.2.2, 5.2.3 and 5.2.4 for any $0 < \delta < 1$, the cumulative regret of* TS *Algorithm 5.1 over T steps is bounded w.p. at least $1 - \delta$ as*

$$R(T) = \widetilde{O}\left(\sqrt{\log(1/\delta)T}\right).$$

Corollary 5.3.2 (From (Abbasi-Yadkori and Szepesvári, 2011))**.** *Consider the portfolio allocation problem in Eq. 5.1 of arbitrary dimension. Under Asm. 5.2.1, 5.2.2, 5.2.3 and 5.2.4 for any $0 < \delta < 1$, the cumulative regret of* OFU-LQ *Algorithm 5.1 over T steps is bounded w.p. at least $1 - \delta$ as*

$$R(T) = \widetilde{O}\left(\sqrt{\log(1/\delta)T}\right).$$

Those two results are stated as Corollary as they can be derived easily, following the same proof structure, from Ch. 4 and (Abbasi-Yadkori and Szepesvári, 2011). The minor differences come from the regret decomposition (the term R^{RLS} no longer appears), the fact that x_t is bounded w.h.p. since the matrices $A^\star + B^\star K(\widetilde{\theta}_t)$ are stable by Asm. 5.2.2, and from a similar relationship between gradient and control (Prop. 4.A.1). Finally, despite the fact that the CE algorithm is widely used in practice (see e.g., Polderman

1986), its performance relies on the consistency of the parameter estimation which may or may not occur. We exhibit this behavior through numerical experiments and discuss it in Sec. 5.5.

5.4 Experiments

In this section, we provide numerical experiments for the CE, TS and OFU-LQ algorithms. For sake of clarity, we consider the 1-dimensional case where $r_t \in \mathbb{R}$ and use the following values for the true parameters of the system: $\Phi_\alpha^\star = 0.9$, $\Phi_I^\star = 0.7$, $\beta_\alpha^\star = 0.1$, $\beta_I^\star = -1.5$ and $\beta_q^\star = 5$. Notice that the impact parameters $\beta_I^\star$ and $\beta_q^\star$ are significantly larger than the prediction parameter $\beta_\alpha^\star$ which takes into account the fact that for large orders, the impact effects are huge compare to the predictions. Finally, the TRIGGER sub-routine of CE and TS is run with $\tau = 50$. In the next subsections, we consider separately the case $\gamma = 0$ (no risk constraint) and $\gamma = 0.1$ (with risk constraint). We first detail the shape of the admissible set $\mathcal{S}$, and then present the results of the strategies both in term of regret and in term of consistency of the estimation.

5.4.1 Optimal allocation without risk constraint

Whenever $\gamma = 0$, the optimization is made without taking into account the risk of the allocation strategy i.e., no control is made about the amplitude of the position Q_t, and the objective is to maximize the profit only. This translates in the fact that the optimal policy only depends on the current prediction and impact variables α_t and I_t. Formally, the control matrix has the shape of $K = \begin{pmatrix} 0 & K_\alpha & K_I \end{pmatrix}$. However, the problem is still well defined due to impact effects: indeed, since trading is costly, it constrains the amplitude of the trades. The set of admissible parameters $\mathcal{S}$ which is the set of parameters such that there exists no dynamical arbitrage imposes the impact effect to be adversarial. Formally, one has

$$\mathcal{S} := \{\theta = (\beta_q, \beta_\alpha, \beta_I) \text{ s.t. } \beta_q > 0,\ \beta_\alpha \neq 0 \text{ and } \beta_q \leq -(1 - \Phi_I^\star)\beta_I\}$$

We first consider the CE algorithm and independently run multiple experiments. The estimate trajectories are plot in Fig. 5.1. It turns out that, on each trajectory, the sequence of estimates do converge, but that the convergence happens only in distribution. This is clear since running the same experiment, one obtains different limit points. Moreover, it is possible to characterize the support of this distribution, which depends on the parameter $\theta^\star$ (see Subsec. 5.5.2). Finally, despite the fact that $\theta^\star$ belongs to this set, the limit distribution does not concentrate around it, which means that no consistency is achieved and that the control can converge, with fixed probability, to a sub-optimal one. As a result, the regret of this strategy is linear (see Fig. 5.2). This is a well-known issue in RL and it motivates the use of more explorative algorithms, such as TS and OFU-LQ, that, thanks to randomness or optimism, are able to discriminate the true parameter and overcome the limitation of the CE. We plot in Fig. 5.2 the regret bounds for the 3 strategies, on average and in high probability.

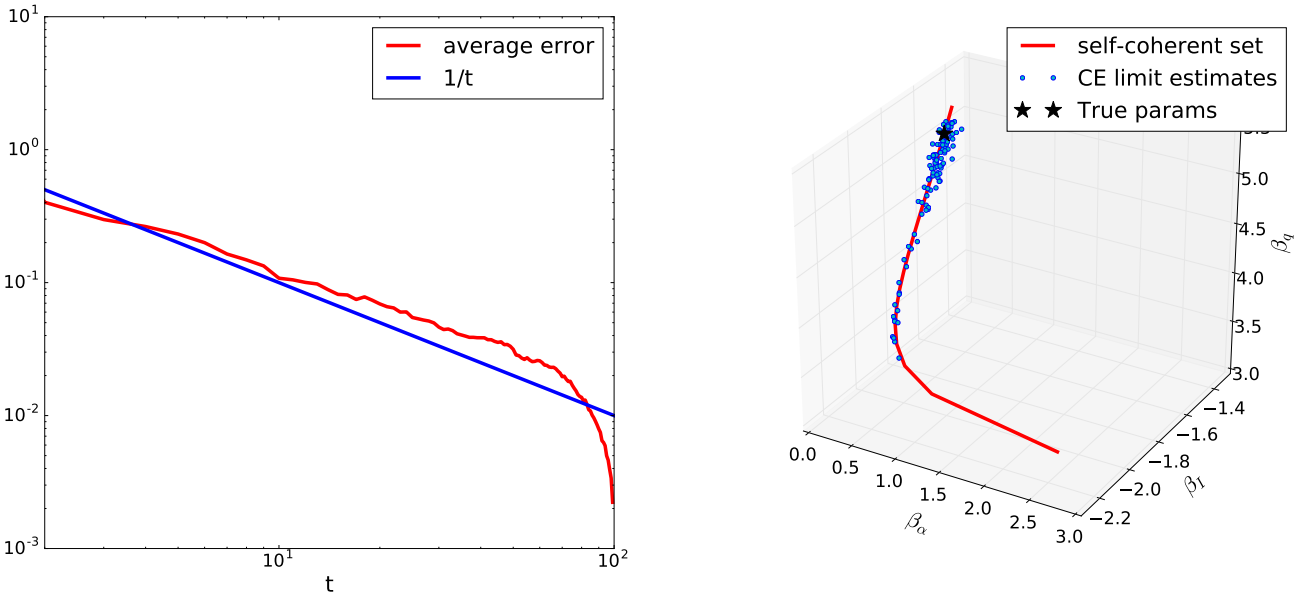

Figure 5.1 – RLS estimates of the CE algorithm for 100 trajectories. *Left:* Empirical mean squared error of the RLS estimate $\|\widehat{\theta}_t - \widehat{\theta}_\infty\|^2$ on average over the trajectories and rate of convergence in loglog plot. *Right:* Empirical distribution of estimates once the convergence is obtained. The line corresponds to the support of the distribution, which contains the true parameter $\theta^\star$.

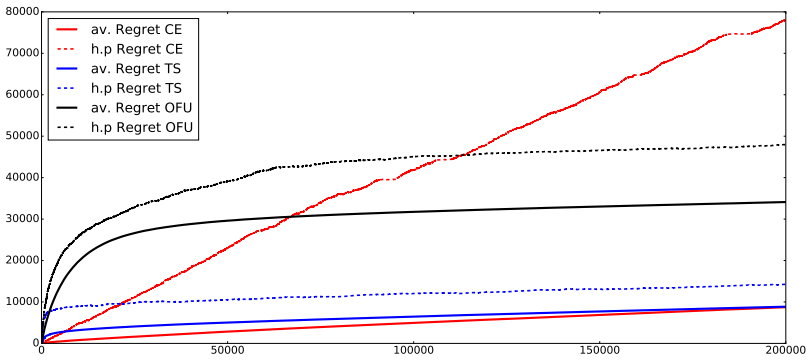

Figure 5.2 – Average and high probability regret bounds of the CE, TS and OFU-LQ algorithms.

First, as expected from Fig. 5.1, on sees that the regret of the CE strategy is linear and that the high probability bound is significantly worse than the average bound. This is due to the fact that for lots of trajectories, the estimates do converge to a point close to $\theta^\star$. However, there is a small, yet fixed, probability that $\widehat{\theta}_t$ converges to a value which is far from the true one, incurring point-wise a large linear regret. This is not the case of TS and OFU-LQ which, accordingly with Cor. 5.3.1 and 5.3.2, achieve a $O(\sqrt{T})$ regret, although, it may take some time to be smaller than the CE's because of the

"expensive" exploration phase. Finally, we note that TS performs better than OFU-LQ (by a constant), which is often observed empirically although not provided by the theory. This may be due to the fact that optimism induces a more aggressive exploration than randomness, and that such aggressive behavior is not needed in practice and mostly comes from the looseness of the theoretical bounds used to derive the algorithms.

5.4.2 Optimal allocation with risk constraint

We now consider the case $\gamma = 0.1$, which corresponds to the setting where the cost function includes a risk term. The optimal policy becomes function of the current position Q_t as well as the current prediction and impact variables α_t and I_t. Formally, the control matrix has the shape of $K = \begin{pmatrix} K_Q & K_\alpha & K_I \end{pmatrix}$. The set of admissible parameters $\mathcal{S}$ simplifies into

$$\mathcal{S} := \{\theta = (\beta_q, \beta_\alpha, \beta_I) \text{ s.t. } \beta_q > 0,\ \beta_\alpha \neq 0 \text{ and } \beta_I \leq 0\}$$

This modification in the structure of the control K has a major consequence since it implies that the CE estimation is now consistent. To illustrate this effect, we plot in Fig. 5.3 the average trajectories of the estimates starting from different initial points. Since they all converge to $\theta^\star$, the convergence no longer occurs in distribution but almost surely while the rate of convergence is still $1/t$. As a result, the regret of the CE is in $\log(T)$ which is way better than the regret of TS and OFU-LQ. We show the average and high probability regret bounds of the CE algorithm in Fig. 5.4.

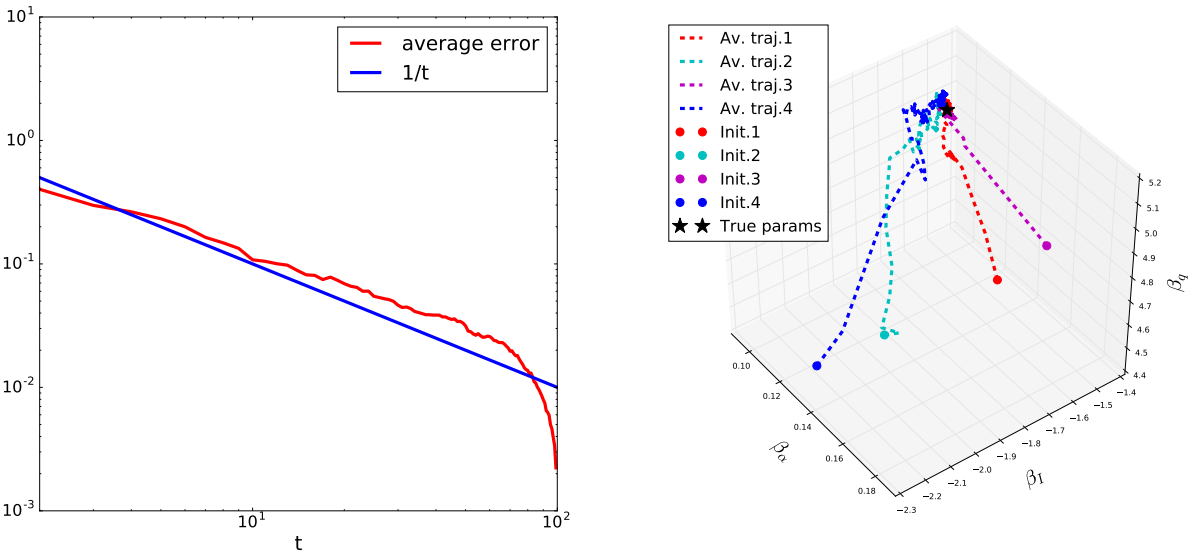

Figure 5.3 – RLS estimates of the CE algorithm for 100 trajectories. *Left:* average error of the RLS estimate and rate of convergence in loglog plot. *Right:* average trajectories of the RLS estimates starting from different initial points.

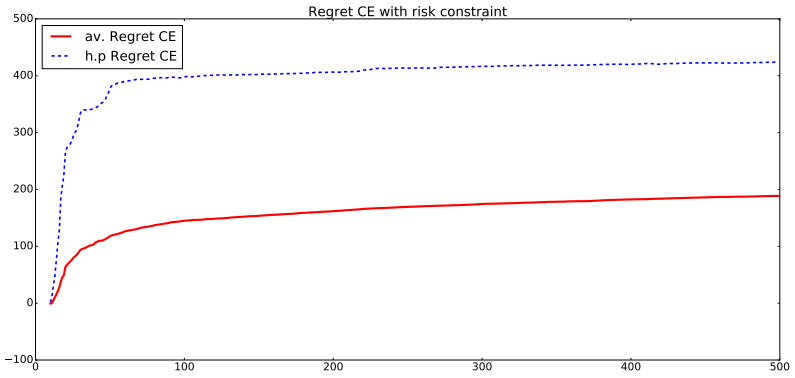

Figure 5.4 – Average and high probability bounds for the regret of the CE algorithm.

While this situation is unusual in RL, where greedy strategies often fail to achieve a sub-linear regret, it provides support to the fact that this strategy had been widely used by the control community. In the next section, we discuss this result and show why adding a risk constraint to the portfolio allocation problem makes the greedy strategy consistent, and thus optimal.

5.5 From closed-loop consistency to consistency

In this section, we discuss why the RLS estimates of the CE strategy converge or not to the true parameters $\theta^\star$. We first show that the limit estimate of the CE belongs to a specific set, and then that it reduces to $\theta^\star$, thus ensuring consistency, when risk constraint is considered.

5.5.1 Closed-loop consistency

We introduce the set of *self-coherent parameters* which ensures the *closed-loop consistency*. The *closed-loop consistency* corresponds to the fact that the closed-loop dynamics of the system i.e., the dynamics of the system once controlled, is well estimated. The set of *self-coherent parameters* corresponds to the set of θ such that the predicted closed-loop dynamics is asymptotically equal to the observed one. We denote this set as $\mathcal{SC}$ which is formally defined as

$$\mathcal{SC} := \{\theta \in \mathcal{S} \text{ s.t } \|(\theta^\star - \theta)^\mathsf{T} z_T\|^2 \underset{T\to\infty}{\longrightarrow} 0, \text{ where } q_t = K(\theta)x_t \text{ for all } t \geq 1\}, \qquad (5.7)$$

where the process $\{z_t\}_t \geq 0$ is generated by following the policy $q = K(\theta)x$. We first characterize the shape of $\mathcal{SC}$ and then show that the CE estimates converge to this set.

Lemma 5.5.1. *Under Asm. 5.2.1, 5.2.2, 5.2.3 and 5.2.4, the set of self-coherent parameters satisfies*

$$\mathcal{SC} = \left\{ \theta \in \mathcal{S} \; s.t \; \theta^\mathsf{T} \begin{pmatrix} & K(\theta) & \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} = \theta^{\star,\mathsf{T}} \begin{pmatrix} & K(\theta) & \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \right\}.$$

The result of Lem. 5.5.1 provides a characterization of the set $\mathcal{SC}$ which is independent of the covariate z_t. We present the main ideas of the proof and postpone the formal derivation to App. 5.B. We use the structure of z_t to introduce the control matrix $K(\theta)$ and the state process x_t and show that, under the assumptions of Lem. 5.5.1, the state process admits a stationary distribution with positive definite variance.
By construction, since $q_t = K(\theta)x_t$, $x_t^\mathsf{T} = (Q_t^\mathsf{T}, \alpha_t^\mathsf{T}, I_t^\mathsf{T})$, $z_t^\mathsf{T} = (q_t^\mathsf{T}, \alpha_t^\mathsf{T}, I_t^\mathsf{T})$, one has, for all $t \geq 0$,

$$z_t = \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} x_t + \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix} K(\theta) x_t = \begin{pmatrix} & K(\theta) & \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} x_t,$$

where $x_{t+1} = \Big(A^\star + B^\star K(\theta)\Big) x_t + \epsilon_{t+1}^x$. Under Asm. 5.2.2, the matrix $A_c(\theta) = A^\star + B^\star K(\theta)$ is stable and so is the process $\{x_t\}_t$ which admits a stationary distribution. Denoting as x_∞ the random variable following the stationary distribution, Asm. 5.2.1 guarantees that $\mathbb{E}(x_\infty) = 0$ and that $\mathbb{V}(x_\infty) = \Sigma_\infty$. Therefore, computing the expectatio of Eq. 5.7, one obtains that

$$\mathcal{SC} \subset \left\{ \theta \in \mathcal{S} \text{ s.t } \mathrm{Tr}\Big((\theta^\star - \theta)^\mathsf{T} \begin{pmatrix} & K(\theta) & \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \Sigma_\infty \begin{pmatrix} K^\mathsf{T}(\theta) & I & 0 \\ & 0 & I \end{pmatrix} (\theta^\star - \theta) \Big) = 0 \right\}.$$

A sufficient condition to ensure Lem. 5.5.1 is thus that Σ_∞ is positive definite. Intuitively, it requires all components of x_t to be persistently excited by ϵ_{t+1}^x and that there exists no linear relationship between them. Notice that ϵ_{t+1}^x has the specific structure $\epsilon_{t+1}^{x,\mathsf{T}} = (0, \epsilon_{t+1}^{\alpha,\mathsf{T}}, 0)$, which encodes the fact that the prediction α_t are naturally excited by the random noise ϵ_{t+1}^α, while the inventory position Q_t and the impact variable I_t are function of the prediction process $\{\alpha_t\}_t$, and thus of $\{\epsilon_t^\alpha\}_t$ through the trading policy $q = K(\theta)x$. As a result, all component of x_t are persistently excited by $\{\epsilon_t^\alpha\}_t$. Further, the structure of the dynamic and the control matrix $K(\theta)$ for any $\theta \in \mathcal{S}$ impose that there exists no linear relationship between them, provided that $\Phi_I^\star \neq I$ i.e., provided that the dynamic of Q_t and I_t are disjoint. Since this is guaranteed by Asm. 5.2.2, one obtains the desired result.

We now show that the sequence of the CE estimates converge to $\mathcal{SC}$.

Lemma 5.5.2. *Let $\widehat{\theta}_t$ be the sequence of* RLS *estimates of the CE algorithm. Then, $\widehat{\theta}_t$ converges in distribution to a random variable $\widehat{\theta}^\infty$ and $\widehat{\theta}^\infty \in \mathcal{SC}$ a.s.*

Applying the RLS properties of Prop. 4.2.4 to the portfolio estimation, one obtains:

Corollary 5.5.1. *Let $\lambda \geq 1$, for any arbitrary $\mathcal{F}_t$-adapted sequence of control $(q_0, \dots, q_t)$, let $\widehat{\theta}_t$ be the sequence of* RLS *estimates defined in Eq. 5.6, one has*

$$\sum_{s=0}^{t} \|(\theta^\star - \widehat{\theta}_t)^\mathsf{T} z_t\| = \widetilde{O}\Big(\sqrt{T}\Big).$$

Cor. 5.5.1 guarantees that the *on-policy* error of the estimation is cumulatively bounded and that this is a structural property of the RLS, since it holds for any sequence of trades. As a consequence it implies that asymptotically,

$$\|(\theta^\star - \widehat{\theta}_T)^\mathsf{T} z_T\| \underset{T\to\infty}{\longrightarrow} 0.$$

We now adopt an asymptotic reasoning. First, notice that without further assumption on the design matrix V_t, no consistency can be guarantee for the RLS estimates, although the convergence still holds in distribution. We introduce the random variable $\widehat{\theta}^\infty$ which corresponds to the limit estimate, so that $\widehat{\theta}_T \underset{T\to\infty}{\xrightarrow{d}} \widehat{\theta}^\infty$. Further, the regularity of the function $K(\theta)$ ensures that the control performed by the CE satisfy $K(\widehat{\theta}_T) \underset{T\to\infty}{\xrightarrow{d}} K(\widehat{\theta}^\infty)$. Using the fact that the trades executed by the CE algorithm follows, neglecting the projection onto $\mathcal{S}$, $q_t = K(\widehat{\theta}_t)x_t$, one obtains that $\widehat{\theta}^\infty$ satisfies Eq. 5.7 i.e., that $\widehat{\theta}^\infty \in \mathcal{SC}$.

5.5.2 Self-coherence set

Thanks to Lem. 5.5.2, we can exhibit, for any parametrization $\theta^\star$, a sufficient condition for the consistency of the CE estimates, or equivalently, a necessary condition for the failure of this greedy strategy. To support the experiments of Sec. 5.4, we illustrate both behavior depending on the presence of risk constraint.

Allocation without risk constraint. As discussed in Sec. 5.4.1, whenever $\gamma = 0$, the optimal policy does not depend on the current inventory position: for any $\theta \in \mathcal{S}$, $K(\theta) = \begin{pmatrix} 0 & K_\alpha(\theta) & K_I(\theta) \end{pmatrix}$. As a result, the set of *self-coherent parameters* $\mathcal{SC}$ becomes:

$$\mathcal{SC} = \left\{ \theta = (\beta_q, \beta_\alpha, \beta_I) \in \mathcal{S} \text{ s.t. } \begin{pmatrix} \beta_q K_\alpha(\theta) + \beta_\alpha \\ \beta_q K_I(\theta) + \beta_I \end{pmatrix} = \begin{pmatrix} \beta_q^\star K_\alpha(\theta) + \beta_\alpha^\star \\ \beta_q^\star K_I(\theta) + \beta_I^\star \end{pmatrix} \right\}.$$

Since $K(\theta)$ is known, the set $\mathcal{SC}$ is characterized block-wise by two equations of three variables, and thus not reduced to a singleton. As a consequence, the CE asymptotic estimates can potentially be distributed onto this set, which stands as a necessary condition for the failure of the greedy strategy. This condition is common to RL problems that require additional exploration: to illustrate this, we plot in Fig. 5.1 the shape of such sets for the portfolio allocation problem (similar to the r.h.s of Fig. 5.1), the LB problem of Ch. 3 and the LQ problem of Ch. 4.

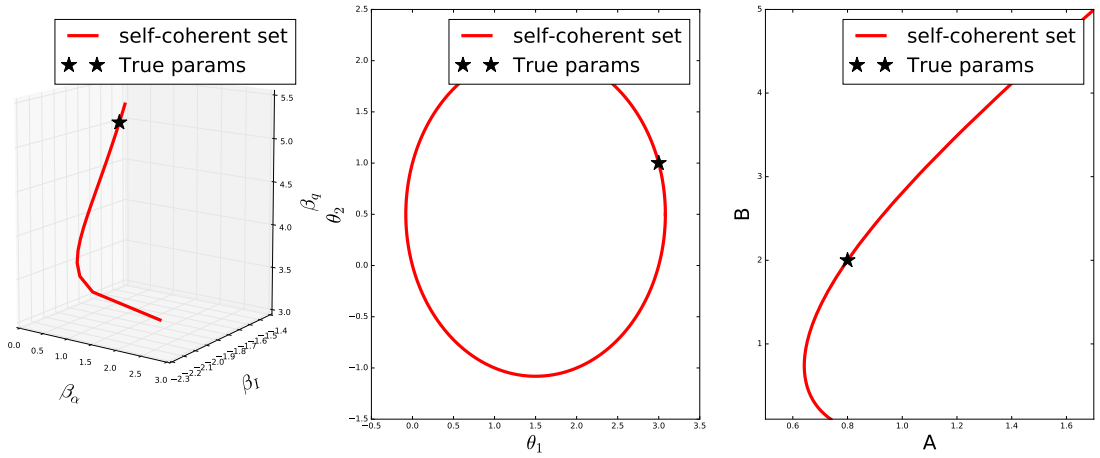

Figure 5.1 – Shape of the set of self-coherent parameters $\mathcal{SC}$, i.e., parameters for which the average predicted observation is equal to the observed one. *Left:* Portfolio allocation without risk constraint. *Center:* Linear Bandit problem with arm set $\mathcal{X} = \{x \in \mathbb{R}^2 \text{ s.t. } \|x\| \leq 1\}$. *Right:* LQ problem.

Notice that if this provides us with a necessary condition, it is not a sufficient one. Despite the fact that the support of the distribution of the asymptotic estimates is not a singleton, it is does not imply that those points are attracting w.r.t the RLS estimation. Moreover, for specific problems, it is possible that this set coincides with the set of parameters that induces optimal control defined as $\mathcal{K} := \{\theta \in \mathcal{S} \text{ s.t. } K(\theta) = K(\theta^\star)\}$. Therefore, this argument does not replace a regret analysis but provides some insight about the potential behavior of the estimation process. Finally, notice that the more we constrain the admissible parameter set $\mathcal{S}$, the more we may reduce $\mathcal{SC}$ which is coherent with the intuition that as we inject knowledge about the parameters in the learning process (through the constraint), we make the exploration-exploitation trade-off easier to solve.

Allocation with risk constraint. When the optimal allocation is made under risk constraint ($\gamma > 0$), the optimal policy depends on the current inventory position thus, for any $\theta \in \mathcal{S}$, $K(\theta) = \begin{pmatrix} K_Q(\theta) & K_\alpha(\theta) & K_I(\theta) \end{pmatrix}$ with $K_Q(\theta) \neq 0$. Therefore, the set of *self-coherent parameters* $\mathcal{SC}$ becomes:

$$\mathcal{SC} = \left\{ \theta = (\beta_q, \beta_\alpha, \beta_I) \in \mathcal{S} \text{ s.t. } \begin{pmatrix} \beta_q K_Q(\theta) \\ \beta_q K_\alpha(\theta) + \beta_\alpha \\ \beta_q K_I(\theta) + \beta_I \end{pmatrix} = \begin{pmatrix} \beta_q^\star K_Q(\theta) \\ \beta_q^\star K_\alpha(\theta) + \beta_\alpha^\star \\ \beta_q^\star K_I(\theta) + \beta_I^\star \end{pmatrix} \right\},$$

and is now characterized by three equations of three unknown variables. For sake of simplicity, consider the 1-dimensional case where $\theta \in \mathbb{R}^3$. Since $K(\theta)$ is known to the controller, solving the three equations in $\mathcal{SC}$ leads to $\beta_q = \beta_q^\star$, $\beta_\alpha = \beta_\alpha^\star$ and $\beta_I = \beta_I^\star$. As a consequence, $\mathcal{SC} = \{\theta^\star\}$, so Lem. 5.5.2 ensures that the CE estimates are *consistent*. Additionally, it allows us to characterize the convergence rate. Notice that by Least

Squares, in average,

$$\left\| \begin{pmatrix} \widehat{\beta}_{q,t} K_Q(\widehat{\theta}_t) \\ \widehat{\beta}_{q,t} K_\alpha(\widehat{\theta}_t) + \widehat{\beta}_{\alpha,t} \\ \widehat{\beta}_{q,t} K_I(\widehat{\theta}_t) + \widehat{\beta}_{I,t} \end{pmatrix} - \begin{pmatrix} \beta_q^\star K_Q(\widehat{\theta}_t) \\ \beta_q^\star K_\alpha(\widehat{\theta}_t) + \beta_\alpha^\star \\ \beta_q^\star K_I(\widehat{\theta}_t) + \beta_I^\star \end{pmatrix} \right\| \underset{t\to\infty}{\sim} \frac{1}{t}.$$

which implies that $\left\|\widehat{\theta}_t - \theta^\star\right\| \underset{t\to\infty}{\sim} \frac{1}{t}$. As a consequence, in the presence of risk constraint, the average regret of the CE strategy is bounded by $\widetilde{O}\Big(\log(T)\Big)$.

Discussion. We end this section stressing how the difference in the structure of the controller, and hence the trades, modifies the behavior of the CE estimation. First, looking at Eq. 5.4, one observes that the inventory position Q_t does not affect the *open-loop* i.e., uncontrolled dynamic of the returns, which only depends on α_t, I_t and q_t. When the trading policy is of the form $q_t = K_\alpha \alpha_t + K_I I_t$, which corresponds to the risk-free case, this creates an ambiguity in the prediction of the returns, that the CE strategy is unable to eliminate. A standard way to overcome this issue is to perturb or regularize the control to enhance the exploration, which is somehow what TS or OFU-LQ do. Following this intuition, one can understand trading policies of the form $q_t = (K_\alpha \alpha_t + K_I I_t) + K_Q Q_t$ as a perturbed or regularized control. Moreover, in the presence of risk constraint, it coincides with the shape of the optimal control, which means that perturbing/regularizing is now optimal, thus the additional exploration is given for free (in term of regret). While this point of view is clear when the perturbation is exogenous i.e., independent of the state dynamic, it is less obvious here as Q_t also depends on $\{\alpha_t\}_t$ and $\{I_t\}_t$. This is at the core of the proof of Lem. 5.5.1, which uses the fact that the dynamic of Q_t is by construction different from the one of α_t and I_t, and thus, relatively independent. As a result, it still acts as an "exogenous" perturbation, maintaining the validity of the intuition.

5.6 Conclusion

In this Chapter, we present an application of adaptive strategies in LQ problems for portfolio allocation. We first highlight the interest of this framework, showing that under the assumption of linear Markovian prices dynamics, one can efficiently encode and solve the allocation problem with price impact and return predictability. In addition, we show that the LQ technical assumptions can be rephrased in term of non-arbitrage, thus allowing to use the richness of the Riccati theory while maintaining financial intuition. Further, we discuss the structure of the parameters and present an approximation procedure that enables us to use RLS to estimate the model. We recall and compare three strategies, both in terms of regret and consistency of the estimation, respectively based on greedy updates (CE), randomness (TS) and optimism (OFU-LQ). The obtained results differ significantly depending on the presence of risk constraint: in the risk-free setting, we observe that the greedy strategy fails to tackle the exploration-exploitation trade-off, while specifically designed algorithm such as TS

and OFU-LQ does, and we retrieve the linear and square-root regret bounds expected from the theory. On the other hand, under risk constraint, we show that the greedy strategy becomes consistent and exhibit a $\log(T)$ average regret. We explain how this modification induces a different structure in the control, and why it guarantees the consistency, introducing the set of *self-coherent parameters* and comparing its shape in both cases. This has several implications: first, it stresses that the CE strategy may perform well in adaptive problems that exhibit a specific structure and explains its popularity in the control community. From a practical perspective, one shouldn't discard it a priori, even though it is usually expected to suffer a linear regret in standard RL problems. Finally, we believe that the study of the *self-coherent parameters* set provides a generic characterization of this issue in parametrized adaptive control problem, of interest beyond the scope of portfolio allocation.

Appendix

5.A Proof of Lem. 5.2.1.

We derive here the proof of theorem 5.2.1 which maps the existence and uniqueness guarantee of the LQR solution to a non-arbitrage criterion. The proof is structured as follow: first, we present the Popov criterion (see e.g., Molinari 1975) which guarantee the existence and uniqueness of a solution to the Riccati equation. Secondly, we show that the deterministic and stochastic LQR share the same Riccati equation and hence, share the same existence and uniqueness condition. Then, we translate the Popov frequency domain criterion in terms of the cost function of the deterministic LQR and thus, in terms of non-arbitrage for admissible trade sequence. Finally, we show that the set of admissible trade sequence is the set of round-trip sequence.

5.A.1 The Popov criterion

Since the cost matrix $\begin{pmatrix} Q & N \\ N^\intercal & R \end{pmatrix}$ associated with the LQR problem for portfolio construction presented in Section 5.2 is not positive definite by construction, the usual guarantee for the Riccati equation solution is violated. However, the existence of a unique admissible solution of 5.5 can still be provided using the Popov criterion. Introducing the hermitian matrix:

$$\Psi(z) = \begin{pmatrix} (Iz^{-1} - A)^{-1}B \\ I \end{pmatrix}' \begin{pmatrix} Q & N' \\ N & R \end{pmatrix} \begin{pmatrix} (Iz - A)^{-1}B \\ I \end{pmatrix},$$

and

$$\begin{aligned} \Psi_K(z) &= Y_K'(z^{-1})\Psi(z)Y_K(z), \\ Y_K(z) &= I + K(Iz - A - BK)^{-1}B, \end{aligned}$$

from (Molinari, 1975), we have the following theorem:

Theorem 5.A.1. *Assume that the pair (A, B) is stabilizable then there exists a (necessarily) unique symmetric stabilizing solution P satisfying the associated Riccati equation if and only if for some (and hence all) K such that $A + BK$ is asymptotically stable, $\Phi_K(z) > 0$ for all $|z| = 1$, $z \in \mathbb{C}$.*

This frequency-domain criterion guarantees the global convexity of the problem based on a fairly complete existence theory (see e.g., Molinari 1975, Ionescu et al. 1997, Van Dooren 1981, Wimmer 1984). The main drawback however is the use of the frequency-domain method involved. To get a better intuition about the existence of optimal solution, we link here the Popov criterion to a non-dynamical arbitrage criterion in line with (Gatheral, 2010).

5.A.2 Deterministic LQR

First, let's notice that the LQR solution of the stochastic problem 5.5 involves the same Riccati equation (see e.g., Bertsekas 1995) - and hence shares the same conditions - as the deterministic LQR problem (5.8):

$$\begin{aligned} \underset{\{q_t\}_{t=1,\dots,\infty}\in\mathcal{Q}}{\text{minimize}} \quad & \tilde{J}(q_0, q_1, \dots) := \sum_{t=0}^{\infty} x_t' Q x_t + 2x_t' N q_t + q_t' R q_t, \\ \text{subject to} \quad & x_{t+1} = Ax_t + Bq_t, \end{aligned} \tag{5.8}$$

where $\mathcal{Q} := \{q = (q_0, q_1, \dots) \in l^1 \cap l^2 \text{ such that } x = (x_0, x_1, \dots) \in l^1 \cap l^2\}$ is the admissible control space which are the stabilizing sequences.
Indeed, the stochastic LQR problem cost function is defined with an expectation regarding the noise process ϵ_t^x which is of zero conditional mean. Thanks to the linear structure of the dynamics, the noises vanish within the Bellman equation which is then the same as the one of the deterministic LQR. As a result, we can apply the Popov criterion on the deterministic problem to ensure the existence and uniqueness of a solution to the stochastic one.

5.A.3 From frequency to time domain

We now state the first corollary which derives directly from Thm. 5.A.1:

Corollary 5.A.1. *Assume that the pair (A, B) is stabilizable then there exists a (necessarily) unique symmetric stabilizing solution P if and only if $\tilde{J}(q_0, q_1, \dots) > 0$ for any $q \in \mathcal{Q}$.*

The proof is straightforward using the z-transform theory. Let $q \in \mathcal{Q}$ be any admissible control sequence and denote as $q(z)$ and $x(z)$ the z-transform of the control sequence and associated state sequence respectively. Then, applying the z-transform theory to (5.8) and using Parceval's theorem leads to:

$$\tilde{J}(q) = \oint_{|z|=1} q^{\mathsf{T}}(z^{-1}) \Phi(z) q(z) dz. \tag{5.9}$$

The following lemma provides another description for the admissible sequence q:

Lemma 5.A.1. *$q \in \mathcal{Q}$ if and only if there exists a stable control K i.e., such that $A + BK$ is stable, and a sequence $v \in l^1 \cap l^2$ such that*

$$q_t = Kx_t + v_t, \quad \forall t \geq 0.$$

Proof. Let K be a stable control and define $v_t = q_t - Kx_t$ for all $t \geq 1$. By definition of $\mathcal{Q}$, q and x belong to $l^1 \cap l^2$ and so does v.
On the other hand, let $v \in l^1 \cap l^2$, K be a stable control and define $q_t = Kx_t + v_t$ for all $t \geq 1$. Then, $x_{t+1} = (A + BK)x_t + v_t$ and since $A + BK$ is stable, $v \in l^1 \cap l^2$ implies that $x \in l^1 \cap l^2$ and so does q. □

We make use of Lemma 5.A.1 to rephrase (5.9) in terms of v sequence: for any $q \in \mathcal{Q}$, let K be a stable control and $v \in l^1 \cap l^2$ sequence such that $q_t = Kx_t + v_t$. Denoting as $v(z)$ the z-transform of v, the z-transform $q(z)$ is:

$$q(z) = Y_K(z)v(z).$$

Finally, equation (5.9) becomes:

$$\tilde{J}(q) = \oint_{|z|=1} q^\mathsf{T}(z^{-1})\Phi(z)q(z)dz = \oint_{|z|=1} v^\mathsf{T}(z^{-1})\Phi_K(z)v(z).$$

Therefore, rephrasing Thm. 5.A.1, there exists a unique symmetric stabilizing solution to the Riccati equation if and only if for some (and hence all) K such that $A + BK$ is stable $\Phi_K(z) > 0$ for all $|z| = 1$ if and only if $\oint_{|z|=1} v^\mathsf{T}(z^{-1})\Phi_K(z)v(z) > 0$ for all $v \in l^1 \cap l^2$ if and only if $\tilde{J}(q) > 0$ for any $q \in \mathcal{Q}$.

5.A.4 From admissible trade sequence to round-trip

Finally, to prove Lem. 5.2.1, one just has to show that the set of admissible sequence coincides with the one of round-trip trajectories. Recalling the definition, one has

$$\begin{aligned}
&\mathcal{RT} = \{q = (q_0, q_1, \dots) \in l^1 \cap l^2 \text{ s.t. } Q = (Q_0, Q_1, \dots) \in l^1 \cap l^2\} && \text{if } \gamma \in (0, \infty),\\
&\mathcal{RT} = \{q = (q_0, q_1, \dots) \in l^1 \cap l^2\}, && \text{if } \gamma = 0.
\end{aligned}$$

We first deal with the generic case where $\gamma \in (0, \infty)$ and then discuss the specific instance where $\gamma = 0$. By definition, $\mathcal{Q} \subset \mathcal{RT}$ so we just need to prove that for any $q \in \mathcal{RT}$, the associated state sequence is such that $x \in l^1 \cap l^2$. To do so, we denote as before $q(z)$, $Q(z)$ and $x(z)$ the z-transform of q_t, Q_t and x_t respectively. Then one has:

$$\begin{aligned}
x(z) &= (Iz - A)^{-1}Bq(z),\\
x(z) &= (Iz - A)^{-1}B(z-1)Q(z).
\end{aligned}$$

Multiplying by $(z-1)$ and taking the limit when $z \to 1$ one has, since $(z-1)(Iz-A)^{-1}$ converges to a constant matrix (finite) H:

$$(z-1)(Iz-A)^{-1} \underset{z\to 1}{\longrightarrow} H < \infty \quad \Longrightarrow \quad (z-1)x(z) \underset{z\to 1}{\sim} H(z-1)Q(z).$$

Thanks to the final value theorem, one gets $x_t \underset{t\to\infty}{\sim} Q_t$ and since $Q \in l^1 \cap l^2$ by definition of $\mathcal{RT}$ so does x. As a consequence, $\mathcal{Q} = \mathcal{RT}$. Finally, in the specific case where $\gamma = 0$, one does not need to stabilize the position Q_t and thus does not need to feedback the current trade q_t on the current position Q_t. As a result, the same derivation can be applied to the sub-system with internal state $x_t = \begin{pmatrix}\alpha_t^\mathsf{T} & I_t^\mathsf{T}\end{pmatrix}^\mathsf{T}$ which follows a stable dynamic by Asm. 5.2.2. Therefore, one directly obtains that $q \in \mathcal{RT}$ implies that $x \in l_1 \cap l_2$.

5.A.5 Plugging everything together

Thanks to the previous steps, we have that there exists a unique solution to the Riccati equation (since (A, B) is stabilizable) if and only if, for any $q \in \mathcal{RT}$, $\tilde{J}(q) > 0$. Noticing that in the absence of noise, the cost function is equal, in term of portfolio allocation to

$$\tilde{J}(q_0, q_1, \dots) = \sum_{t=0}^{\infty} -PnL_{t,t+1}$$

proves Lem. 5.2.1. Because, in the absence of noise, there is no price predictability, this criterion states that every round-trip must be non-profitable so that impact effects are modeled to act in an adverse manner. Under such guarantee, a unique solution to the portfolio allocation exists.

5.B Proof of Lem. 5.5.1.

As discuss in Sec. 5.5.2, one just need to prove that, for any $\theta \in \mathcal{S}$, the state process generated by the trading policy $q = K(\theta)x$ admits a stationary distribution, with positive definite variance. Since $A^\star$ and $B^\star$ are known, by definition, for any $\theta \in \mathcal{S}$, $A_c(\theta) = A^\star + B^\star K(\theta)$ is stable. Therefore, x_t admits a stationary distribution. Denoting as x_∞ the random variable which follows this stationary distribution, one has $\mathbb{E}(x_\infty) = 0$ and $\mathbb{V}(x_\infty) = \Sigma^\infty$ where

$$\Sigma^\infty = A_c(\theta)\Sigma^\infty A_c^\mathsf{T}(\theta) + \Sigma_x, \text{ with } \Sigma_x = CC^\mathsf{T} \text{ and } C^\mathsf{T} = \begin{pmatrix} 0 & I & 0 \end{pmatrix}.$$

Further, we show the pair $\left(C^\mathsf{T}, A_c^\mathsf{T}(\theta)\right)$ is observable. Writing $K(\theta) = \left(K_Q(\theta), K_\alpha(\theta), K_I(\theta)\right)$ and using Def. 5.B.1, this is provided if the matrix $\mathcal{O}^\mathsf{T}$ is of full row-rank. Formally, one has

$$\mathcal{O}^\mathsf{T} = \begin{pmatrix} 0 & K_\alpha(\theta) & (I + K_Q(\theta))K_\alpha(\theta) + K_\alpha(\theta)\Phi_\alpha^\star + K_I(\theta)K_\alpha(\theta) \\ I & \Phi_\alpha^\star & \Phi_\alpha^{2,\star} \\ 0 & K_\alpha(\theta) & K_Q(\theta)K_\alpha(\theta) + K_\alpha(\theta)\Phi_\alpha^\star + (\Phi_I^\star + K_I(\theta))K_\alpha(\theta) \end{pmatrix},$$

which, is full-rank if and only if

$$\begin{pmatrix} K_\alpha(\theta) & (I + K_Q(\theta))K_\alpha(\theta) + K_\alpha(\theta)\Phi_\alpha^\star + K_I(\theta)K_\alpha(\theta) \\ K_\alpha(\theta) & K_Q(\theta)K_\alpha(\theta) + K_\alpha(\theta)\Phi_\alpha(\theta) + (\Phi_I^\star + K_I(\theta))K_\alpha(\theta) \end{pmatrix}$$

is full-rank. Algebraic manipulations ensure that it is equivalent to consider the matrix

$$\begin{pmatrix} K_\alpha(\theta) & (I + K_Q(\theta))K_\alpha(\theta) + K_\alpha(\theta)\Phi_\alpha^\star + K_I(\theta)K_\alpha(\theta) \\ 0 & (I - \Phi_I^\star)K_\alpha(\theta) \end{pmatrix}$$

By Asm. 5.2.2, $\Phi_I^\star$ is stable so $I - \Phi_I^\star$ is invertible and by Asm. 5.2.3, $K_\alpha(\theta)$ is full-rank which implies that $\mathcal{O}$ is full column-rank. Therefore, by Prop. 5.B.1, Σ^∞ is positive definite thus, $\mathcal{SC}$ is such that

$$\mathcal{SC} = \left\{ \theta \text{ s.t } \mathrm{Tr}\Big((\theta^\star - \theta)^\mathsf{T} \begin{pmatrix} & K(\theta) & \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \Sigma_\infty \begin{pmatrix} & 0 & 0 \\ K^\mathsf{T}(\theta) & I & 0 \\ & 0 & I \end{pmatrix} (\theta^\star - \theta)\Big) = 0 \right\}.$$

Definition 5.B.1. *Let A and C be two matrices of size $n \times n$ and $d \times n$. The pair (C, A) is said to be observable if the observability matrix $\mathcal{O} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}$ is of rank n (full column rank).*

Proposition 5.B.1 (Th. 5.3.5 in (Lancaster and Rodman, 1995))**.** *Let A and C be two matrices of size $n \times n$ and $d \times n$. Let S be the solution of the Lyapunov equation*

$$S = A^{\mathsf{T}} S A + C^{\mathsf{T}} C.$$

Then, provided that the matrix A is stable, the solution is unique, symmetric and positive semi-definite. Further, if the pair (C, A) is observable, the solution is positive definite.

Chapter 6
Summary and Future Work

This chapter provides a summary of methods and analyses presented in this thesis, highlights some open questions that lay ground for future work and discusses the Linear Quadratic Gaussian (LQG) extension that consists in adding partial observability to the LQ problem of Ch. 4.

1 Summary

This thesis has been motivated by the study of Reinforcement Learning (RL) algorithms in linearly parametrized systems, that address the *exploration-exploitation* trade-off in sequential decision making. While linear models seem somehow limited to reflect the true dynamics of real systems, they remain accurate enough for lots of practical applications and offer many advantages that balance the induced approximation error. First, as stressed in the specific example on portfolio construction (see Ch. 5), they consist in robust, yet flexible models and thus can encode complicated features (e.g., dynamical predictability and impact effects) while maintaining tractability of both the solution of the control problem and of the estimation problem. Second, as a specific instance of parametrized system, they can handle problems with large and/or continuous state and action spaces, for which designing algorithms that properly trade-off exploration and exploitation stands as one of the main challenges in RL. Third, from a theoretical perspective, their relative simplicity allows one to focus on the underlying difficulty in the analysis of algorithms and to exhibit the key aspect of their functioning. Finally, they are the cornerstone of generic parametrized problems, and a good understanding of analyses in those settings is a prerequisite to any further developments.
Two popular principles have been introduced to tackle the exploration-exploitation trade-off: one based on *optimism* and one based on *Thompson sampling* (TS). While the former have been intensively studied over the last decades and gave rise to many algorithms for which theoretical guarantees have been provided, the latter has recently generated significant interest due to the impressive empirical performance of the induced algorithms. We focused in this thesis on Thompson sampling-based algorithms and analyzed their performance in the Linear Bandit (LB) of Linear Quadratic (LQ) control problems. While TS is originally build on *Bayesian* ideas, we studied it in the *frequentist* setting and stressed the randomized nature of its functioning.

In Ch. 3, we analyzed the regret of TS in the LB setting, and derived an alternative proof that sheds new light on the functioning of the algorithm. In particular, we leveraged the structure of the problem to show how the regret is related to the sensitivity of the objective function, and how the structure of the algorithm (i.e., selecting optimal

arm w.r.t. a chosen parameter) takes this sensitivity into account. Then, we explain how the random nature of TS selects arms that control the sensitivity of the objective function, and hence the regret. Additionally, our analysis holds for any appropriate sampling distribution, which stresses that randomization is the key feature of TS while prior/posterior Bayesian update only stands as a convenient tool to obtain the sampling distribution. Further, our proof relies on the property of the objective function and can be readily applied to problems whose objective function shares the same structure (e.g., generalized linear model and regularized linear optimization).

In Ch. 4, we leveraged our novel analysis in LB and extended it to the LQ control problem. In this more complicated setting, we showed that the functioning of the algorithm is similar to the one of Ch. 3. We stressed the link between the actual actions chosen by TS and the Jacobian of the optimal value function, and showed how the randomization of TS induces actions that do control the sensitivity of the optimal value function. Further, we exhibited the need to trade-off the frequency of sampling parameters and the frequency of switches in the control policy, and showed that standard lazy update schemes induce at best an overall regret of $O(T^{2/3})$. We overcame this issue by deriving a novel bound on the regret due to policy switches, thus allowing to update parameters and the policy at each step and overcome the limitations due to lazy updates. As a result, we proved $O(\sqrt{T})$ regret bound for the regret of TS in LQ.

In Ch. 5, we presented an application example for portfolio allocation problems. We highlighted the interest of the LQ framework, by showing that it is possible to cast a complicated dynamical allocation problem, that takes into account the main features of the prices dynamics, into a LQ control problem. We investigated the exploration-exploitation trade-off in this specific setting and compared the performance of TS with the optimistic-based algorithm for LQ (OFU-LQ) and with a naive greedy strategy. While TS and OFU-LQ achieve a $O(\sqrt{T})$ regret as expected from the theory, we showed that depending on whether or not risk constraint is considered, the greedy strategy suffers a $O(T)$ or a $O(\log T)$ regret. We discussed this surprising result and showed that the greedy strategy may or may not be consistent. Further, we exhibited the support of the distribution of the limit estimates and explained the failure or the success of the greedy strategy depending on the shape of this set.

2 Future Work

The analyses provided in this thesis open a number of interesting questions and research directions. In particular, we discussed in Sec. 3.7 the need for optimism that is at the core of the TS analysis, and hinted that this is a sufficient condition rather than a necessary one. Investigating this aspect of TS is highly challenging because it requires to quantify how 'informative" actions induced by non-optimistic samples are, but may improve the regret bound by $\sqrt{d}$ and thus prove the empirical evidence that TS offers similar performance as OFUL. Additionally, it would allow to extend the results of Ch. 4 to the n-dimensional setting, since as discussed in Sec. 4.6, the need for optimism

raises technical difficulties that would be overcome using a generic approach that relies on the shape of the objective function over the whole ellipsoid (and not only on the optimistic subset). Finally, we highlighted in Ch. 5 that a greedy strategy may guarantee consistent estimation, and thus achieve a $O(\log T)$ regret, depending on the structure of the problem. A formal characterization of this matter would be of major interest, from both theoretical and practical perspective. In particular, we have collected preliminary theoretical and empirical evidence that the greedy strategy achieves a $O(\log T)$ regret in the Linear Quadratic Gaussian (LQG) control problem.

The LQG control problem is a natural extension of the LQ problem of Ch. 4, where the agent no longer observes the current state of the system x_t, but a noisy linear transformation y_t of it as $y_t = C_* x_t + \epsilon_t^y$, where $\{\epsilon_t^y\}_t$ is a zero-mean noise process, and C_* parametrizes the unknown observation model. As a result, it can be seen as a linearly parametrized Partially Observable Markov Decision Process (POMDP). From a practical perspective, it is very useful to model systems that do suffer from partial observability or that follow a more complicated linear dynamics than the one of LQ, e.g., autoregressive-moving-average with exogenous inputs (ARMAX) model versus autoregressive with exogenous inputs (ARX) model. From a theoretical perspective, the derivation of the optimal policy have been intensively studied (Bertsekas, 1995) and relies on the *separation principle* that states that the optimal policy is linear in the Kalman filter estimate of the internal state, where the linear map is given by the LQ control matrix. One of the main challenge towards the analysis of exploration-exploitation algorithms in LQG is the lack of theoretical guarantee for the estimation of the unknown parameters of the dynamics. Several methods have been proposed, such as the Subspace Method (SM) and the Prediction Error Method (PEM). The former is more popular in system identification, and gave rise to the celebrated N4SID algorithm while the latter is more popular in statistics, as it is based on Maximum Likelihood (ML). Despite this limitation, it is possible to sketch a regret analysis, assuming that confidence bounds do hold, in the same flavor as in Ch. 4. In particular, the regret decomposition is very similar to Eq. 4.6, and, up to technical difficulties, a regret analysis may be derived following the same proof structure. On the other hand, numerical experiments suggest that the greedy strategy, which consists in following the optimal policy w.r.t. the current estimate, achieves a $O(\log T)$ regret, and thus, is optimal. Using ML to estimate the parameters in the 1-dimensional case (i.e., the state is 1-dimensional) and running the greedy strategy starting from different initial parameters, one retrieve similar results as in Fig. 5.3 which suggests that the greedy strategy induces consistent estimates, as in the portfolio example of Ch. 5. From a theoretical perspective, inspired from the discussions in Sec. 5.5, one should show that the self-coherent parameters set is reduced to a singleton, that is the true parameter. Since ML aims at minimizing the prediction error, the sequence of estimates are designed so that the predicted observation be equal to the true observation in average. However, this equality is more difficult to characterize compared to LQ as it involves the whole dynamics of the predictions, and not only the stationary distribution. While using the z-transform theory in the 1-dimensional case and equalizing the transfert functions

addresses this issue and proves that the self-coherent parameters set is reduced to a singleton, how to properly define and characterize this set in the general $n-$dimensional case, and proving that the N4SID or ML estimates of the greedy strategy converge to this set, are open questions that lay ground for future work and might provide good intuitions about exploration-exploitation issue in similar settings, such as POMDPs.

Bibliography

Y. Abbasi-Yadkori and C. Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, pages 1–26, 2011. (→ pages 21, 25, 27, 28, 60, 61, 62, 64, 65, 66, 67, 71, 80, 85, 101, 115, and 116.)

Y. Abbasi-Yadkori and C. Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2015. (→ pages 24, 28, 60, 64, 65, and 66.)

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011a. (→ pages 16, 18, 19, and 33.)

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011b. (→ pages 16, 17, and 45.)

M. Abeille and A. Lazaric. Linear thompson sampling revisited. In *AISTATS 2017-20th International Conference on Artificial Intelligence and Statistics*, 2017a. (→ page 29.)

M. Abeille and A. Lazaric. Thompson sampling for linear-quadratic control problems. In *AISTATS*, 2017b. (→ page 59.)

M. Abeille, E. Serie, A. Lazaric, and X. Brokmann. Lqg for portfolio optimization. *arXiv preprint arXiv:1611.00997*, 2016. (→ pages 105 and 107.)

J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, pages 2197–2205, 2015. (→ page 50.)

G. Acosta and R. G. Durán. An optimal poincaré inequality in l 1 for convex domains. *Proceedings of the american mathematical society*, pages 195–202, 2004. (→ pages 68, 73, 97, 98, and 99.)

R. Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995. (→ page 8.)

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012a. (→ pages 14, 15, and 30.)

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*, 2012b. (→ pages 16, 18, 19, 30, 31, 32, 34, 39, 45, 47, 50, 60, 66, and 67.)

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of AI&Stats*, 2013. (→ pages 15, 19, and 52.)

S. Agrawal and R. Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017. (→ page 24.)

R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001. (→ page 106.)

J.-Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *ALT*, volume 4754, pages 150–165. Springer, 2007. (→ pages 12 and 13.)

P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007. (→ page 23.)

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a. (→ pages 8, 12, 13, and 16.)

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b. (→ page 15.)

P. L. Bartlett and A. Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009. (→ page 22.)

D. P. Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA, 1995. (→ pages 20, 25, 127, and 133.)

D. Bertsimas and A. W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, 1998. (→ page 106.)

S. Bittanti, M. Campi, et al. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, 6(4):299–320, 2006. (→ page 60.)

J.-P. Bouchaud, J. Farmer, and F. Lillo. How markets slowly digest changes in supply and demand. *arXiv.org*, 2008. (→ page 106.)

X. Brokmann, J. Kockelkoren, J.-P. Bouchau, and E. Sérié. Slow decay of impact in equity markets. *Available at SSRN 2471528*, 2014. (→ page 106.)

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. (→ pages 8 and 15.)

S. Bubeck and C.-Y. Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems 26*, pages 638–646. 2013. (→ page 15.)

M. C. Campi and P. Kumar. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998. (→ page 60.)

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013. (→ pages 12 and 13.)

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. (→ page 9.)

S.-H. Chang, P. C. Cosman, and L. B. Milstein. Chernoff-type bounds for the gaussian error function. *Communications, IEEE Transactions on*, 59(11):2939–2944, 2011. (→ page 54.)

O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011. (→ pages 12 and 14.)

C.-P. Chen and F. Qi. Completely monotonic function associated with the gamma functions and proof of wallis' inequality. *Tamkang Journal of Mathematics*, 36(4): 303–307, 2005. (→ page 58.)

G. Chun-hua. Newtons method for discrete algebraic riccati equations when the closed-loop matrix has eigenvalues on the unit circle. *SIAM J. Matrix Anal. Appl*, pages 279–294, 1998. (→ page 26.)

G. M. Constantinides. Multiperiod consumption and investment behavior with convex transactions costs. *Management Science*, 25(11):1127–1137, 1979. (→ page 106.)

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008. (→ pages 16 and 18.)

V. De La Pena, T. L. Lai, and Q.-M. Shao. Self-normalized processes: Limit theory and statistical applications, 2009. (→ page 17.)

J. Donier, J. Bonart, I. Mastromatteo, and J.-P. . Bouchaud. A fully consistent, minimal model for non-linear market impact. *Minimal Model for Non-Linear Market Impact (November 29, 2014)*, 2014. (→ page 106.)

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993. (→ page 106.)

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594, 2010. (→ pages 20, 48, and 49.)

N. Gârleanu. Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory*, 144(2):532–564, 2009. (→ page 106.)

N. Gârleanu and L. Pedersen. Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, 68(6):2309–2340, 2013. (→ page 106.)

J. Gatheral. No-dynamic-arbitrage and market impact. *Quantitative finance*, 10(7): 749–759, 2010. (→ pages 106 and 126.)

A. Gopalan and S. Mannor. Thompson sampling for learning parameterized markov decision processes. In *Proceedings of The 28th Conference on Learning Theory*, 2015. (→ page 24.)

R. Grinold. Signal weighting. *The Journal of Portfolio Management*, 36(4):24–34, 2010. (→ page 106.)

O. Guéant. Optimal execution and block trade pricing: a general framework. *arXiv preprint arXiv:1210.6372*, 2012. (→ page 106.)

G. Huberman and W. Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 72(4):1247–1275, 2004. (→ page 106.)

V. Ionescu, C. Oara, and M. Weiss. General matrix pencil techniques for the solution of algebraic riccati equations: a unified approach. *Automatic Control, IEEE Transactions on*, 42(8):1085–1097, 1997. (→ page 126.)

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, Aug. 2010. (→ pages 21, 22, 23, 24, and 66.)

M. C. Jensen, F. Black, and M. S. Scholes. The capital asset pricing model: Some empirical tests. 1972. (→ page 106.)

K.-S. Jun, A. Bhargava, R. Nowak, and R. Willett. Scalable generalized linear bandits: Online computation and hashing. *arXiv preprint arXiv:1706.00136*, 2017. (→ page 20.)

J. Kallsen and J. Muhle-Karbe. The general structure of optimal investment and consumption with small transaction costs. *Swiss Finance Institute Research Paper*, (13-15), 2013. (→ page 106.)

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT 2012)*, pages 199–213, 2012. (→ page 15.)

J. Klamka. Controllability of dynamical systems. *Mathematica Applicanda*, 36(50/09): 57–75, 2016. (→ page 62.)

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems 26*, pages 1448–1456, 2013. (→ page 15.)

S. G. Krantz and H. R. Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2012. (→ page 92.)

P. R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015. (→ page 114.)

A. Kyle. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335, 1985. (→ page 106.)

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985. (→ pages 9 and 15.)

P. Lancaster and L. Rodman. *Algebraic riccati equations*. Oxford University Press, 1995. (→ pages 25, 26, 110, and 130.)

J. D. Lataillade, C. Deremble, M. Potters, and J.-P. Bouchaud. Optimal trading with linear costs. *arXiv preprint arXiv:1203.5957*, 2012. (→ page 106.)

T. Lattimore and C. Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737, 2017. (→ page 20.)

A. J. Laub. A schur method for solving algebraic riccati equations. *Automatic Control, IEEE Transactions on*, 24(6):913–921, 1979. (→ page 26.)

A. J. Laub. Invariant subspace methods for the numerical solution of riccati equations. In *The Riccati Equation*, pages 163–196. Springer, 1991. (→ page 26.)

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010. (→ page 16.)

L. Li, Y. Lu, and D. Zhou. Provable optimal algorithms for generalized linear contextual bandits. *arXiv preprint arXiv:1703.00048*, 2017. (→ page 20.)

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011. (→ page 57.)

O.-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. *arXiv preprint arXiv:1105.5820*, 2011. (→ pages 12 and 13.)

H. Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952. (→ page 106.)

I. Mastromatteo, B. Toth, and J.-P. Bouchaud. Agent-based models for latent liquidity and concave price impact. *Physical Review E*, 89(4):042805, 2014. (→ page 106.)

B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, 13(1): 2069–2106, 2012. (→ page 15.)

B. P. Molinari. The stabilizing solution of the discrete algebraic riccati equation. *Automatic Control, IEEE Transactions on*, 20(3):396–399, Jun 1975. (→ pages 111 and 126.)

L. Moreau, J. Muhle-Karbe, and H. M. Soner. Trading with small price impact. *Swiss Finance Institute Research Paper*, (14-17), 2014. (→ page 106.)

A. J. Morton and S. R. Pliska. Optimal portfolio management with fixed transaction costs. *Mathematical Finance*, 5(4):337–356, 1995. (→ page 106.)

C. Niculescu and L.-E. Persson. *Convex functions and their applications: a contemporary approach*. Springer Science & Business Media, 2006. (→ page 56.)

A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32, 2013. (→ page 106.)

I. Osband and B. V. Roy. Near-optimal reinforcement learning in factored mdps. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 604–612. Curran Associates, Inc., 2014. (→ page 24.)

I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1466–1474. Curran Associates, Inc., 2014. (→ page 24.)

I. Osband and B. Van Roy. Posterior sampling for reinforcement learning without episodes. *arXiv preprint arXiv:1608.02731*, 2016. (→ pages 5, 24, 28, 59, 61, and 66.)

I. Osband and B. Van Roy. On optimistic versus randomized exploration in reinforcement learning. *arXiv preprint arXiv:1706.04241*, 2017. (→ page 24.)

I. Osband, B. Van Roy, and D. Russo. (more) efficient reinforcement learning via posterior sampling. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3003–3011, USA, 2013. Curran Associates Inc. (→ page 24.)

B. Park and B. Van Roy. Adaptive execution: Exploration and learning of price impact. *Operations Research*, 63(5):1058–1076, 2015. (→ pages 106 and 112.)

L. E. Payne and H. F. Weinberger. An optimal poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960. (→ page 99.)

J. W. Polderman. On the necessity of identifying the true parameter in adaptive lq control. *Systems & control letters*, 8(2):87–91, 1986. (→ page 116.)

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. (→ page 22.)

D. Russo, D. Tse, and B. Van Roy. Time-sensitive bandit learning and satisficing thompson sampling. *arXiv preprint arXiv:1704.09028*, 2017. (→ page 15.)

W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk*. *The journal of finance*, 19(3):425–442, 1964. (→ page 106.)

M. J. A. Strens. A bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 943–950, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2. (→ page 23.)

R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981. (→ pages 20 and 60.)

M. Taksar, M. J. Klass, and D. Assaf. A diffusion model for optimal portfolio selection in the presence of brokerage fees. *Mathematics of Operations Research*, 13(2):277–294, 1988. (→ page 106.)

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933. (→ pages 12 and 30.)

P. Van Dooren. A generalized eigenvalue approach for solving riccati equations. *SIAM Journal on Scientific and Statistical Computing*, 2(2):121–135, 1981. (→ pages 26 and 126.)

Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1729–1736, 2009. (→ page 16.)

H. K. Wimmer. The algebraic riccati equation: conditions for the existence and uniqueness of solutions. *Linear algebra and its applications*, 58:441–452, 1984. (→ page 126.)

Appendix

A Concentration inequalities

Lemma A.1 (Hoeffding's inequality). *Let X be a bounded r.v. in $[a,b]$, of zero mean. Then, for any $t \in \mathbb{R}$,*

$$\mathbb{E}\Big[e^{tX}\Big] \leq e^{t^2(b-a)^2/8}.$$

Lemma A.2 (Chernoff bound for Gaussian r.v.). *Let $X \sim \mathcal{N}(0,1)$. For any $t \geq 0$, then,*

$$\mathbb{P}(|X| \geq t) \leq 2\exp\Big(-\frac{t^2}{2}\Big).$$

Theorem A.1 (Chernoff-Hoeffding's inequality). *Let $(X_1,\dots,X_n)$ be bounded independent r.v. such that $X_i \in [a_i,b_i]$ and $\mu_i = \mathbb{E}(X_i)$ for all $i = 1,\dots,n$. Then, for any $t \in \mathbb{R}$,*

$$\mathbb{E}\bigg(\Big|\sum_{i=1}^{n} X_i - \mu_i\Big| \geq t\bigg) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}.$$

Theorem A.2 (Azuma's inequality). *Let $\{M_s\}_{s\geq 0}$ be a super-martingale such that $|M_s - M_{s-1}| \leq c_s$ almost surely. Then, for all $t > 0$ and all $\epsilon > 0$,*

$$\mathbb{P}\Big(|M_t - M_0| \geq \epsilon\Big) \leq 2\exp\bigg(\frac{-\epsilon^2}{2\sum_{s=1}^{t} c_s^2}\bigg).$$

B Convergence of random variables

Definition B.1 (Convergence in distribution). *A sequence $(X_1, X_2, \dots)$ of real-valued random variables is said to converge in distribution or weakly to a random variable X if*

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for all $x \in \mathbb{R}$ at which F is continuous, where F_n and F are the cumulative distribution functions of r.v. X_n and X, respectively. This convergence is denoted as $X_n \xrightarrow{d} X$.

Definition B.2 (Convergence in probability). *A sequence $(X_1, X_2, \dots)$ of real-valued random variables is said to converge in probability to a random variable X if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

This convergence is denoted as $X_n \xrightarrow{p} X$.

Definition B.3 (Almost sure convergence). *A sequence $(X_1, X_2, \dots)$ of real-valued random variables is said to converge almost surely to a random variable X if,*

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

This convergence is denoted as $X_n \xrightarrow{as} X$.