Thompson Sampling for Linear-Quadratic Control Problems

Marc Abeille **Alessandro Lazaric**
Inria Lille - Nord Europe, Team SequeL

Abstract

We consider the exploration-exploitation tradeoff in linear quadratic (LQ) control problems, where the state dynamics is linear and the cost function is quadratic in states and controls. 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, we show that when studying the frequentist regret TS in control problems, we need to trade-off the frequency of sampling optimistic parameters and the frequency of switches in the control policy. This results in an overall regret of $O(T^{2/3})$, which is significantly worse than the regret $O(\sqrt{T})$ achieved by the optimism-in-face-of-uncertainty algorithm in LQ control problems. OFU

1 Introduction

One of the most challenging problems in reinforcement learning (RL) is how to effectively trade off exploration and exploitation in an unknown environment. A number of learning methods has been proposed in finite Markov decision processes (MDPs) and they have been analyzed in the PAC-MDP (see e.g., [13]) and the regret framework (see e.g., [8]). The two most popular approaches to address the exploration-exploitation trade-off are the optimism-in-face-of-uncertainty (OFU) principle, where optimistic policies are selected according to upper-confidence bounds on the true MDP paramaters, and the Thompson sampling (TS) strategy[1], where random MDP parameters are selected from a posterior distribution and the corresponding optimal policy is executed. Despite their success in finite MDPs, extensions of these methods and their analyses to continuous state-action spaces are still rather limited. Osband et al. [11] study how to randomize the parameters of a linear function approximator to induce exploration and prove regret guarantees in the finite MDP case. Osband and Van Roy [10] develops a specific TS method applied to the more complex case of neural architectures with significant empirical improvements over alternative exploration strategies, although with no theoretical guarantees. In this paper, we focus on a specific family of continuous state-action MDPs, the linear quadratic (LQ) control problems, 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 and Campi [6] and Campi and Kumar [7], 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 [1] proposed a learning algorithm (OFU-LQ) with $O(\sqrt{T})$ cumulative regret. Abbasi-Yadkori and Szepesvári [2] further studied how the TS strategy, could be adapted to work in the LQ control problem. Under the assumption that the true parameters of the model are drawn from a known prior, they show that the so-called Bayesian regret matches the $O(\sqrt{T})$ bound of OFU-LQ.

In this paper, we analyze the regret of TS in LQ problems in the more challenging frequentist case, where θ is a fixed parameter, with no prior assumption of its value. 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

[1] In RL literature, TS has been introduced by Strens [14] and it is often referred to as posterior-sampling for reinforcement learning (PSRL).

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

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 Bayesian regret analysis of TS in [2] does not apply in this case, since no prior is available on θ. Furthermore, we show that existing frequentist regret analysis for TS in linear bandit [5] cannot be generalized to the LQ case. This requires deriving a novel line of proof in which 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 process to obtain optimistic models and if this is not done *frequently enough*, the regret can grow unbounded. This forces TS to favor short episodes and we prove that this leads to an overall regret of order $O(T^{2/3})$ in the one-dimensional case (i.e., both states and controls are scalars), which is significantly worse than the $O(\sqrt{T})$ regret of OFU-LQ.

2 Preliminaries

The control problem. We consider the discrete-time infinite-horizon linear quadratic (LQ) control problem. Let $x_t \in \mathbb{R}^n$ be the state of the system and $u_t \in \mathbb{R}^d$ be the control at time t; an LQ problem is characterized by linear dynamics and a quadratic cost function

$$\begin{aligned} x_{t+1} &= A_* x_t + B_* u_t + \epsilon_{t+1}, \\ c(x_t, u_t) &= x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t, \end{aligned} \tag{1}$$

where A_* and B_* are *unknown* matrices and Q and R are *known* positive definite matrices of appropriate dimension. We summarize the unknown parameters in $\theta_*^\mathsf{T} = (A_*, B_*)$. The noise process ϵ_{t+1} is zero-mean and it satisfies the following assumption.

Assumption 1. *$\{\epsilon_t\}_t$ is a $\mathcal{F}_t$−martingale difference sequence, where $\mathcal{F}_t$ is the filtration which represents the information knowledge up to time t.*

In LQ, the objective is to design a closed-loop control policy $\pi : \mathbb{R}^n \to \mathbb{R}^d$ mapping states to controls that minimizes the average expected cost

$$J_\pi(\theta_*) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T} c(x_t, u_t)\Big], \tag{2}$$

with $x_0 = 0$ and $u_t = \pi(x_t)$. Standard theory for LQ control guarantees that the optimal policy is linear in the state and that the corresponding average expected cost is the solution of a Riccati equation.

Proposition 1 (Thm.16.6.4 in [9])**.** *Under Asm. 1 and for any LQ system with parameters $\theta^\mathsf{T} = (A, B)$ such that (A, B) is stabilizable[2], and p.d. cost matrices Q and R, the optimal solution of Eq. 2 is given by*

$$\begin{aligned} \pi(\theta) &= K(\theta)x_t, \quad J(\theta) = Tr(P(\theta)), \\ K(\theta) &= -(R + B^\mathsf{T} P(\theta)B)^{-1} B^\mathsf{T} P(\theta)A, \\ P(\theta) &= Q + A^\mathsf{T} P(\theta)A + A^\mathsf{T} P(\theta)BK(\theta) \end{aligned} \tag{3}$$

where $\pi(\theta)$ is the optimal policy, $J(\theta)$ is the corresponding average expected cost, $K(\theta)$ is the optimal gain, and $P(\theta)$ is the unique solution to the Riccati equation associated with the control problem. Finally, we also have that $A + BK(\theta)$ is asymptotically stable.

For notational convenience, we use $H(\theta) = \big(I \; K(\theta)^\mathsf{T}\big)^\mathsf{T}$, so that the closed loop dynamics $A + BK(\theta)$ can be equivalently written as $\theta^\mathsf{T} H(\theta)$. We introduce further assumptions about the LQ systems we consider.

Assumption 2. *We assume that the LQ problem is characterized by parameters (A_*, B_*, Q, R) such that the cost matrices Q and R are symmetric p.d., and $\theta_* \in \mathcal{S}$ where[3] $\mathcal{S} = \{\theta \in \mathbb{R}^{(n+d)\times n}$ s.t. $Tr(P(\theta)) \le D$ and $Tr(\theta\theta^\mathsf{T}) \le S^2\}$.*

While Asm. 1 basically guarantees that the linear model in Eq. 1 is correct, Asm. 2 restricts the control parameters to the admissible set $\mathcal{S}$. This is used later in the learning process and it replaces Asm. A2-4 in [1] in a synthetic way, as shown in the following proposition.

Proposition 2. *Given an admissible set $\mathcal{S}$ as defined in Asm. 2, we have **1)** $\mathcal{S} \subset \{\theta^\mathsf{T} = (A, B)$ s.t. (A,B) is stabilizable$\}$, **2)** $\mathcal{S}$ is compact, and **3)** there exists $\rho < 1$ and $C < \infty$ positive constants such that $\rho = \sup_{\theta\in\mathcal{S}} \|A + BK(A, B)\|_2$ and $C = \sup_{\theta\in\mathcal{S}} \|K(\theta)\|_2$.[4].*

As an immediate result, any system with $\theta \in \mathcal{S}$ is stabilizable, and therefore, Asm. 2 implies that Prop. 1 holds. Finally, we derive a result about the regularity of the Riccati solution, which we later use to relate the regret to the controls performed by TS.

Lemma 1. *Under Asm. 1 and for any LQ with parameters $\theta^\mathsf{T} = (A, B)$ and cost matrices Q and R satisfying Asm. 2, let $J(\theta) = Tr(P(\theta))$ be the optimal solution of*

[2] (A, B) is stabilizable if there exists a control gain matrix K s.t. $A+BK$ is stable (i.e., all eigenvalues are in $(-1, 1)$).

[3] Even if $P(\theta)$ is not defined for every θ, we extend its domain of definition by setting $P(\theta) = +\infty$.

[4] We use $\|\cdot\|$ and $\|\cdot\|_2$ to denote the Frobenius and the 2-norm respectively.

Eq. 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 $J(\theta)$ in a direction $\delta\theta$, denoted as $\nabla J(\theta)^\mathsf{T}\delta\theta$, where $\nabla J(\theta) \in \mathbb{R}^{(n+d)\times n}$ is the gradient of J, is the solution of the Lyapunov equation

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

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

The learning problem. At each time t, the learner chooses a policy π_t, it executes the induced control $u_t = \pi_t(x_t)$ and suffers a cost $c_t = c(x_t, u_t)$. The performance is measured by the cumulative *regret* up to time T as $R_T = \sum_{t=0}^{T}(c_t^{\pi_t} - J_{\pi_*}(\theta_*))$, where at each step the difference between the cost of the controller c^π and the expected average cost $J_{\pi_*}(\theta_*)$ of the optimal controller π_* is measured. Let $(u_0, \dots, u_t)$ be a sequence of controls and $(x_0, x_1, \dots, x_{t+1})$ be the corresponding states, then $\theta^\star$ can be estimated by regularized least-squares (RLS). Let $z_t = (x_t, u_t)^\mathsf{T}$, for any regularization parameter $\lambda \in \mathbb{R}_+^*$, the design matrix and the RLS estimate are defined as

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

For notational convenience, we use $W_t = V_t^{-1/2}$. We recall a concentration inequality for RLS estimates.

Proposition 3 (Thm. 2 in [3])**.** *We assume that ϵ_t are conditionally and component-wise sub-Gaussian of parameter L and that $\mathbb{E}(\epsilon_{t+1}\epsilon_{t+1}^\mathsf{T}|\mathcal{F}_t) = I$. Then 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*

$$Tr\Big((\hat{\theta}_t - \theta_*)^\mathsf{T} V_t(\hat{\theta}_t - \theta_*)\Big) \le \beta_t(\delta)^2, \tag{4}$$

w.p. $1-\delta$ (w.r.t. the noise $\{\epsilon_t\}_t$ and any randomization in the choice of the control), where

$$\beta_t(\delta) = nL\sqrt{2\log\Big(\frac{\det(V_t)^{1/2}}{\det(\lambda I)^{1/2}}\Big)} + \lambda^{1/2}S. \tag{5}$$

Further, when $\|z_t\| \le Z$,

$$\frac{\det(V_t)}{\det(\lambda I)} \le (n+d)\log\big(1 + TZ^2/\lambda(n+d)\big).$$

At any step t, we define the ellipsoid $\mathcal{E}_t^{\text{RLS}} = \big\{\theta \in \mathbb{R}^d \mid \|\theta - \hat{\theta}_t\|_{V_t} \le \beta_t(\delta')\big\}$ centered in $\widehat{\theta}_t$ with orientation V_t and radius $\beta_t(\delta')$, with $\delta' = \delta/(4T)$. Finally, we report a standard result of RLS that, together with Prop. 3, shows that the prediction error on the points z_t used to construct the estimator $\widehat{\theta}_t$ is cumulatively small.

```
Input: $\hat{\theta}_0$, $V_0 = \lambda I$, $\delta$, $T$, $\tau$, $t_0 = 0$
Set $\delta' = \delta/(8T)$
for $t = \{0, \dots, T\}$ do
    if $\det(V_t) > 2\det(V_0)$ or $t \ge t_0 + \tau$ then
        while $\widetilde{\theta}_t \notin \mathcal{S}$ do
            Sample $\eta_t \sim \mathcal{D}^{\text{TS}}$ - iid $\mathcal{N}(0,1)$
            Compute $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t$
        end while
        Let $V_0 = V_t$, $t_0 = t$,
    else
        $\widetilde{\theta}_t = \widetilde{\theta}_{t-1}$
    end if
    Execute control $u_t = K(\widetilde{\theta}_t)x_t$
    Move to state $x_{t+1}$, receive cost $c(x_t, u_t)$
    Compute $V_{t+1}$ and $\widehat{\theta}_{t+1}$
end for
```

Figure 1: Thompson sampling algorithm.

Proposition 4 (Lem. 10 in [1])**.** *Let $\lambda \ge 1$, for any arbitrary $\mathcal{F}_t$-adapted sequence $(z_0, z_1, \dots, z_t)$, let V_{t+1} be the corresponding design matrix, then*

$$\sum_{s=0}^{t}\min\big(\|z_s\|^2_{V_s^{-1}}, 1\big) \le 2\log\frac{\det(V_{t+1})}{\det(\lambda I)}. \tag{6}$$

Moreover, when $\|z_t\| \le Z$ for all $t \ge 0$, then

$$\sum_{s=0}^{t}\|z_s\|^2_{V_s^{-1}} \le 2\frac{Z^2}{\lambda}(n+d)\log\Big(1 + \frac{(t+1)Z^2}{\lambda(n+d)}\Big).$$

3 Thompson Sampling for LQR

We introduce a specific instance of TS for learning in LQ problems obtained as a modification of the algorithm proposed in [2], where we replace the Bayesian structure and the Gaussian prior assumption with a generic randomized process and we modify the update rule. The algorithm is summarized in Alg. 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. Denoting as $\mathcal{R}_\mathcal{S}$ the rejection sampling operator associated with the admissible set $\mathcal{S}$, we define $\widetilde{\theta}_t$ as

$$\widetilde{\theta}_t = \mathcal{R}_\mathcal{S}(\widehat{\theta}_t + \beta_t(\delta')W_t\eta_t), \tag{7}$$

where $W_t = V_t^{-1/2}$ and every coordinate of the matrix $\eta_t \in \mathbb{R}^{(n+d)\times(n+d)}$ is a random sample drawn i.i.d. from $\mathcal{N}(0,1)$. We refer to this distribution as $\mathcal{D}^{\text{TS}}$. Notice that such sampling does not need to be associated with an actual posterior over $\theta^\star$ but it just need to randomize parameters coherently with the RLS estimate and the uncertainty captured in V_t. Let $\gamma_t(\delta) = \beta_t(\delta')n\sqrt{2(n+d)\log\big(2n(n+d)/\delta\big)}$, then the high-probability TS ellipsoid $\mathcal{E}_t^{\mathrm{TS}} = \{\theta \in \mathbb{R}^d \mid \|\theta - \widehat{\theta}_t\|_{V_t} \leq \gamma_t(\delta')\}$ is defined so that any parameter $\widetilde{\theta}_t$ belongs to it with $1-\delta/8$ probability.

Given the parameter $\widetilde{\theta}_t$, the gain matrix $K(\widetilde{\theta}_t)$ is computed and the corresponding optimal control $u_t = K(\widetilde{\theta}_t)x_t$ is applied. As a result, the learner observes the cost $c(x_t, u_t)$ and the next state x_{t+1}, and V_t and $\widehat{\theta}_t$ are updated accordingly. Similar to most of RL strategies, the updates are not performed at each step and the same estimated optimal policy $K(\widetilde{\theta}_t)$ is kept constant throughout an *episode*. Let V_0 be the design matrix at the beginning of an episode, then the episode is terminated upon two possible conditions: **1)** the determinant condition of the design matrix is doubled (i.e., $\det(V_t) \geq 2\det(V_0)$) or **2)** a maximum length condition is reached. While the first condition is common to all RL strategies, here we need to force the algorithm to interrupt episodes as soon as their length exceeds τ steps. The need for this additional termination condition is intrinsically related to the TS nature and it is discussed in detail in the next section.

4 Theoretical analysis

We prove the first frequentist regret bound for TS in LQ systems of dimension 2 ($n = 1$, $d = 1$). 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.

Theorem 1. *Consider the LQ system in Eq. 1 of dimension $n = 1$ and $d = 1$. Under Asm. 1 and 2 for any $0 < \delta < 1$, the cumulative regret of* TS *over T steps is bounded w.p. at least $1-\delta$ as* [5]

$$R(T) = \widetilde{O}\left(T^{2/3}\sqrt{\log(1/\delta)}\right).$$

This result is in striking contrast with previous results in multi-armed and linear bandit where the frequentist regret of TS is $O(\sqrt{T})$ and the Bayesian analysis of TS in control problems where the regret is also $O(\sqrt{T})$. As discussed in the introduction, the frequentist regret analysis in control problems introduces a critical trade-off between the frequency of selecting optimistic models, which guarantees small regret in bandit problems, and the reduction of the number of policy switches, which leads to small regret in control problems. Unfortunately, this trade-off cannot be easily balanced and this leads to a final regret of $O(T^{2/3})$. Sect. 4.2 provides a more detailed discussion on the challenges of bounding the frequentist regret of TS in LQ problems.

4.1 Setting the Stage

Concentration events. We introduce the following high probability events.

Definition 1. *Let $\delta \in (0,1)$ and $\delta' = \delta/(8T)$ and $t \in [0,T]$. We define the event (RLS estimate concentration) $\widehat{E}_t = \big\{\forall s \leq t,\ \|\widehat{\theta}_s - \theta^\star\|_{V_s} \leq \beta_s(\delta')\big\}$ and the event (parameter $\widetilde{\theta}_s$ concentrates around $\widehat{\theta}_s$) $\widetilde{E}_t = \big\{\forall s \leq t,\ \|\widetilde{\theta}_s - \widehat{\theta}_s\|_{V_s} \leq \gamma_s(\delta')\big\}$.*

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

Definition 2. *Let $\delta \in (0,1)$, X, X' be two problem dependent positive constants and $t \in [0,T]$. We define the event (bounded states) $\bar{E}_t = \big\{\forall s \leq t,\ \|x_s\| \leq X \log \frac{X'}{\delta}\big\}$.*

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 $\bar{E} := \bar{E}_T \subset \cdots \subset \bar{E}_1$. We show that these events do hold with high probability.

Lemma 2. $\mathbb{P}(\widehat{E} \cap \widetilde{E}) \geq 1 - \delta/4$.

Corollary 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. 2 leverages Prop. 3 and the sampling distribution $\mathcal{D}^{\mathrm{TS}}$ to ensure that $\widehat{E} \cap \widetilde{E}$ holds w.h.p. Furthermore, Corollary 1 ensures that the states remains 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 directly decompose the regret and bound it on this event as [1, Sect. 4.2]

$$\begin{aligned} R(T) = & \underbrace{\sum_{t=0}^{T} \big\{J(\widetilde{\theta}_t) - J(\theta_*)\big\}\mathbb{1}\{E_t\}}_{R^{\mathrm{TS}}} \\ & + \underbrace{(R_1^{\mathrm{RLS}} + R_2^{\mathrm{RLS}} + R_3^{\mathrm{RLS}})\mathbb{1}\{E_t\}}_{R^{\mathrm{RLS}}} \end{aligned} \tag{8}$$

where R^{RLS} is decomposed into the three components

$$R_1^{\mathrm{RLS}} = \sum_{t=0}^{T} \big\{\mathbb{E}(x_{t+1}^\mathsf{T} P(\widetilde{\theta}_{t+1})x_{t+1}|\mathcal{F}_t) - x_t^\mathsf{T} P(\widetilde{\theta}_t)x_t\big\},$$

$$R_2^{\mathrm{RLS}} = \sum_{t=0}^{T} \mathbb{E}\big[x_{t+1}^\mathsf{T}(P(\widetilde{\theta}_t) - P(\widetilde{\theta}_{t+1}))x_{t+1}|\mathcal{F}_t\big],$$

$$R_3^{\mathrm{RLS}} = \sum_{t=0}^{T} \big\{z_t^\mathsf{T}\widetilde{\theta}_t P(\widetilde{\theta}_t)\widetilde{\theta}_t^\mathsf{T} z_t - z_t^\mathsf{T}\theta_* P(\widetilde{\theta}_t)\theta_*^\mathsf{T} z_t\big\}.$$

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

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

[6] This non-trivial result is directly collected from the bounding-the-state section of [1].

4.2 Related Work and Challenges

Since the RLS estimator is the same in both TS and OFU, the regret terms R_1^{RLS} and R_3^{RLS} can be bounded as in [1]. In fact, R_1^{RLS} is a martingale by construction and it can be bounded by Azuma inequality. The term R_3^{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_3^{RLS}. Finally, the R_2^{RLS} 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 $R_{2,t}^{\mathrm{RLS}}$ 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 $R_{2,t}^{\mathrm{RLS}}$ 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, it is easy to see that the number of episodes is bounded by $O(\log(T))$, which bounds R_2^{RLS} as well (with a constant depending on the bounding of the state X and other parameters specific of the LQ system).[7]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 in [2] has the same lazy update as OFUL and thus it directly controls R_2^{RLS} by a small number of episodes. 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 $\mathbb{E}[R_t^{\mathrm{TS}}] = 0$.

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 [1] exploits the linear structure of LQ problems to reuse arguments originally developed in the linear bandit setting. Similarly, we could leverage on the analysis of TS for linear bandit by Agrawal and Goyal [5] to derive a frequentist regret bound. Agrawal and Goyal [5] 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. 3. 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 $|J_\pi(\theta_*) - J_\pi(\widehat{\theta}_t)|$. Despite the concentration inequality in Prop. 3, 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 [5] 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. 3 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 a small consistency regret. In fact, we need optimistic parameters $\widetilde{\theta}_t$ to be sampled *often enough.* On the other hand, bounding the consistency regret R_2^{RLS} 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 ingredient **1)** constant probability of optimism, **2)** connection between optimism and R^{TS} without using the saturated and unsaturated argument, **3)** a suitable trade-off between lazy updates to bound the consistency regret and frequent updates to guarantee small optimality regret.

[7] Notice that the consistency regret is not specific to LQ systems but it is common to all regret analyses in RL (see e.g., UCRL [8]) 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).

4.3 Bounding the Optimality Regret $R^{\mathbf{TS}}$

$R^{\mathbf{TS}}$ decomposition. We define the "extended" filtration $\mathcal{F}_t^x = (\mathcal{F}_{t-1}, x_t)$. Let K be the (random) number of episodes up to time T, $\{t_k\}_{k=1}^K$ be the steps when the policy is updated, i.e., when a new parameter $\widetilde{\theta}$ is sampled, and let T_k be the associated length of each episode, then we can further decompose R^{TS} as

$$R^{\mathrm{TS}} = \sum_{k=0}^{K} T_k \underbrace{\Big(J(\widetilde{\theta}_{t_k}) - \mathbb{E}[J(\widetilde{\theta}_{t_k})|\mathcal{F}_{t_k}^x, E_{t_k}]\Big)\mathbb{1}_{E_{t_k}}}_{R_{t_k}^{\mathrm{TS},1}} + \sum_{k=0}^{K} T_k \underbrace{\{\mathbb{E}[J(\widetilde{\theta}_{t_k})|\mathcal{F}_{t_k}^x, E_{t_k}] - J(\theta_*)\}\mathbb{1}_{E_{t_k}}}_{R_{t_k}^{\mathrm{TS},2}}. \tag{9}$$

We focus on the second regret term that we redefine $R_{t_k}^{\mathrm{TS},2} = \Delta_t$ for any $t = t_k$ for notational convenience.

Optimism and expectation. 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 Δ_t is bounded by:

$$\begin{aligned}\Delta_t &\leq \Big(\mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t] - J(\theta)\Big)\mathbb{1}_{E_t}, \\ &\leq \Big|J(\theta) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\Big|\mathbb{1}_{E_t}, \text{ which implies that}\end{aligned}$$

$$\Delta_t \leq \mathbb{E}\Big[\big|J(\widetilde{\theta}) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\big|\mathbb{1}_{\widetilde{E}_t}\,|\,\mathcal{F}_t^x, \widehat{E}_t, \bar{E}_t, \widetilde{\theta} \in \Theta^{\mathrm{opt}}\Big],$$

where we use first the definition of the optimistic parameter set, then bounding the resulting quantity by its absolute value, and finally switch to the expectation over the optimistic set, since the inequality is true for any $\widetilde{\theta} \in \Theta^{\mathrm{opt}}$. While this inequality is true for any sampling distribution, it is convenient to select it equivalent to the sampling distribution of TS. Thus, we set $\widetilde{\theta} = \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t(\delta')W_t\eta)$ with η is component wise Gaussian $\mathcal{N}(0,1)$ and obtain

$$\begin{aligned}\Delta_t &\leq \mathbb{E}\Big[\big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\big|\mathbb{1}_{\widetilde{E}_t} \mid \mathcal{F}_t^x, \widehat{E}_t, \bar{E}_t, \widetilde{\theta}_t \in \Theta^{\mathrm{opt}}\Big], \\ &\leq \frac{\mathbb{E}\Big[\big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\big|\mathbb{1}_{\widetilde{E}_t} \mid \mathcal{F}_t^x, \widehat{E}_t, \bar{E}_t\Big]}{\mathbb{P}\big(\widetilde{\theta}_t \in \Theta^{\mathrm{opt}} \mid \mathcal{F}_t^x, \widehat{E}_t\big)}.\end{aligned}$$

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 result is proved in the following lemma.

Lemma 3. *Let $\Theta^{opt} := \{\theta \in \mathbb{R}^d \mid J(\theta) \leq J(\theta^\star)\}$ be the set of optimistic parameters and $\widetilde{\theta}_t = \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t(\delta')W_t\eta)$ with η be component-wise normal $\mathcal{N}(0,1)$, then in the one-dimensional case ($n{=}1$ and $d{=}1$)*

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

where p is a strictly positive constant.

Integrating this result into the previous expression gives

$$\Delta_t \leq \frac{1}{p}\mathbb{E}\Big[\Big|J(\widetilde{\theta}_t) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]\Big| \mid \mathcal{F}_t^x, E_t\Big]. \tag{10}$$

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 *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. In fact, we show later that at each step TS chooses a sampling distribution that tends to minimize the expected absolute deviations of J, thus contributing to reduce the deviations in R_t^{TS}.

Variance 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 as

$$f_t(\eta) = J(\widehat{\theta}_t + \beta_t(\delta')W_t\eta) - \mathbb{E}[J(\widetilde{\theta}_t)|\mathcal{F}_t^x, E_t]$$

where the restriction on the ball is here to meet the $\mathcal{E}_t^{\mathrm{TS}}$ confidence ellipsoid of the sampling. Since the perturbation $\eta \sim \mathcal{D}^{\mathrm{TS}}$ is independent of the past, we can rewrite Eq. 10 as

$$\Delta_t \leq \mathbb{E}_{\eta\sim\mathcal{D}^{\mathrm{TS}}}\big[|f_t(\eta)|\big|\eta \in \mathcal{B}(0,d'), \widehat{\theta}_t + \beta_t(\delta')W_t\eta \in \mathcal{S}\big].$$

We now need to show that this formulation of the regret is strictly related to the policy executed by TS. We prove the following result (proof in the supplement).

Lemma 4. *Let $\Omega \subset \mathbb{R}^d$ be a convex domain with finite diameter* diam*. Let p be a non-negative log-concave function on Ω with continuous derivative up to the second order. Then, for all $u \in W^{1,1}(\Omega)$*[8] *such that $\int_\Omega u(z)p(z)dz = 0$ one has*

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

Before using the previous result, we relate the gradient of f_t to the gradient of J. Since for any η and any $\theta = \widehat{\theta}_t + \beta_t(\delta')W_t\eta$, we have

$$\nabla f_t(\eta) = \beta_t(\delta')W_t\nabla J(\theta)$$

To obtain a bound on the norm of ∇f_t, we apply Prop. 5 (derived from Lem. 1) to get a bound on $\|\nabla J(\theta)\|_{W_t^2}$:

$$\begin{aligned}\|\nabla J(\theta)\|_{W_t^2} \leq& \|A_c(\theta)\|_2^2\|\nabla J(\theta)\|_{W_t^2} \\ &+ 2\|P(\theta)\|\|A_c(\theta)\|_2\|H(\theta)\|_{W_t^2}.\end{aligned}$$

[8] $W^{1,1}(\Omega)$ is the Sobolev space of order 1 in $L^1(\Omega)$.

Making use of $\|M\| \leq \operatorname{Tr}(M)$ for any positive definite matrix together with $\operatorname{Tr}(P(\theta)) \leq D$ (Asm. 2) and $\|A_c(\theta)\|_2 \leq \rho$ (Prop. 2),

$$\|\nabla J(\theta)\|_{W_t^2} \leq \rho^2 \|\nabla J(\theta)\|_{W_t^2} + 2D\rho\|H(\theta)\|_{W_t^2},$$

which leads to

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

We are now ready to use the weighted Poincaré inequality of Lem. 4 to link the expectation of $|f_t|$ to the expectation of its gradient. From Lem. 1, we have $f_t \in W^{1,1}(\Omega)$ and its expectation is zero by construction. On the other hand, the rejection sampling procedure impose that we conditioned the expectation with $\widehat{\theta}_t + \beta_t(\delta')W_t\eta \in \mathcal{S}$ which is unfortunately not convex. However, we can still apply Lem. 4 considering the function $\tilde{f}_t(\eta) = f_t(\eta)\mathbb{1}(\widehat{\theta}_t + \beta_t(\delta')W_t\eta \in \mathcal{S})$ and diameter $diam = d'$. As a result, we finally obtain

$$\Delta_t \leq \gamma\mathbb{E}\Big[\big\|H(\widetilde{\theta}_t)\big\|_{W_t^2}|\mathcal{F}_t^x\Big],$$

where $\gamma = 8\sqrt{n(n+d)}\beta_T(\delta')D\rho/(p(1-\rho^2))$.

From gradient to actions. Recalling the definition of $H(\theta) = \big(I \;\; K(\theta)^{\mathsf{T}}\big)^{\mathsf{T}}$ 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 R^{TS} regret 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 $K(\widetilde{\theta}_t)$. We first plug the bound on Δ_t back into Eq. 9 as

$$R^{\mathrm{TS}} \leq \sum_{k=1}^{K} T_k\bigg(R^{\mathrm{TS},1}_{t_k} + \gamma\mathbb{E}\Big[\big\|H(\widetilde{\theta}_{t_k})\big\|_{V_{t_k}^{-1}}|\mathcal{F}^x_{t_k}\Big]\bigg)\mathbb{1}_{E_{t_k}}.$$

We remove the expectation by adding and subtracting the actual realizations of $\widetilde{\theta}_{t_k}$ as

$$R^{\mathrm{TS},3}_{t_k} = \mathbb{E}\Big[\big\|H(\widetilde{\theta}_{t_k})\big\|_{V_{t_k}^{-1}}|\mathcal{F}^x_{t_k}\Big] - \big\|H(\widetilde{\theta}_{t_k})\big\|_{V_{t_k}^{-1}}.$$

Thus, one obtains

$$R^{\mathrm{TS}} \leq \sum_{k=1}^{K} T_k\Big(R^{\mathrm{TS},1}_{t_k} + R^{\mathrm{TS},3}_{t_k} + \gamma\big\|H(\widetilde{\theta}_{t_k})\big\|_{V_{t_k}^{-1}}\Big)\mathbb{1}_{E_{t_k}}.$$

Now we want to relate the cumulative sum of the last regret term to $\sum_{t=1}^{T}\|z_t\|_{V_t^{-1}}$. This quantity represents the prediction error of the RLS, and we know from Prop. 6 that it is bounded w.h.p. We now focus on the one-dimensional case, where x_t is just a scalar value. Noticing that $\|z_t\|_{V_t^{-1}} = |x_t|\|H(\widetilde{\theta}_t)\|_{V_t^{-1}}$, one has:

$$\sum_{t=0}^{T}\|z_t\|_{V_t^{-1}} = \sum_{k=1}^{K}\Big(\sum_{t=t_k}^{t_{k+1}-1}|x_t|\Big)\|H(\widetilde{\theta}_{t_k})\|_{V_t^{-1}}.$$

Intuitively, it means that over each episode, the more states are excited (e.g., the larger $\sum_{t=t_k}^{t_{k+1}-1}|x_t|$), the more V_t^{-1} reduces in the direction $H(\widetilde{\theta}_{t_k})$. As a result, to ensure that the term $\sum_{k=1}^{K}T_k\|H(\widetilde{\theta}_{t_k})\|_{V_t^{-1}}$ in R^{TS} is small, it would be sufficient ti show that $\sum_{t=t_k}^{t_{k+1}-1}|x_t| \sim T_k$, i.e., that the states provides enough information to learn the system in each chosen direction $H(\widetilde{\theta}_{t_k})$. More formally, let assume that there exists a constant α such that $T_k \leq \alpha\sum_{t=t_k}^{t_{k+1}-1}|x_t|$ for all $k \leq K$. Then,

$$\sum_{k=1}^{K}T_k\|H(\widetilde{\theta}_{t_k})\|_{V_{t_k}^{-1}} \leq \alpha\sum_{t=0}^{T}\|z_t\|_{V_{t_k}^{-1}} \leq 2\alpha\sum_{t=0}^{T}\|z_t\|_{V_t^{-1}},$$

where we use that $\det(V_t) \leq 2\det(V_{t_k})$ as guaranteed by the termination condition. Unfortunately, the intrinsic randomness of x_t (triggered by the noise ξ_t) is such that the assumption above is violated w.p. 1. However, in the one-dimensional case, the regret over the episode k can be conveniently written as

$$R_k(T) = \Big(\sum_{t=t_k}^{t_{k+1}-1}|x_t|^2\Big)\big(Q + K(\theta_{t_k})^2R\big) - T_kJ(\theta_*).$$

As a result, if we set

$$\alpha := X\frac{Q+RC^2}{J(\theta_*)} \geq X\frac{Q+RK(\theta_{t_k})^2}{J(\theta_*)}, \qquad (11)$$

whenever $\sum_{t=t_k}^{t_{k+1}-1}\|x_t\| \leq \frac{1}{\alpha}T_k$ then we can directly conclude that $R_k(T)$ is zero. On the other hand, in the opposite case, we have $T_k \leq \alpha\sum_{t=t_k}^{t_{k+1}-1}|x_t|$ and thus we can upper bound the last term in R^{TS} as

$$R^{\mathrm{TS}} \leq \sum_{k=1}^{K}T_k\Big(R^{\mathrm{TS},1}_{t_k} + R^{\mathrm{TS},3}_{t_k}\Big)\mathbb{1}_{E_{t_k}} + 2\gamma\alpha\sum_{t=0}^{T}\|z_t\|_{V_t^{-1}}.$$

4.4 Final bound

Bounding $R_1^{\mathbf{RLS}}$ and $R_3^{\mathbf{RLS}}$. These two terms can be bounded following similar steps as in [1]. We report the detailed derivation in the supplement while here we simply report the final bounds

$$R_1^{\mathrm{RLS}} \leq \underbrace{2DX^2\sqrt{2\log(4/\delta)}}_{:=\gamma_1}\sqrt{T},$$

and

$$R_3^{\mathrm{RLS}} \leq \underbrace{4SD\sqrt{(1+C^2)X^2}\mu_T(\delta')}_{:=\gamma_3}\sum_{t=0}^{T}\|z_t\|_{V_t^{-1}}\mathbb{1}_{E_t},$$

where $\mu_T(\delta') = \beta_T(\delta') + \gamma_T(\delta')$.

Bounding $R_2^{\mathbf{RLS}}$. Since the policy is updated from time to time, the difference of the optimal values

$P(\widetilde{\theta}_t) - P(\widetilde{\theta}_{t+1})$ is zero unless when the parameters are updated. When it is the case, thanks to the rejection sampling procedure which ensures that every parameters belong to the set $\mathcal{S}$ of Asm. 2, it is trivially bounded by $2D$. Therefore, on event E, one has:

$$R_2^{\mathrm{RLS}} \leq 2X^2 DK,$$

where K is the (random) number of episodes. By definition of TS, the updates are triggered either when the $\det(V_t)$ increases by a factor 2 or when the length of the episode is greater than τ. Hence, the number of update can be split into $K = K^{det} + K^{len}$, where K^{det} and K^{len} are the number of updates triggered by the two conditions respectively. From Cor. 2, one gets:

$$K \leq \big(T/\tau + (n+d)\log_2(1 + TX^2(1+C^2)/\lambda)\big),$$

and thus

$$R_2^{\mathrm{RLS}} \leq \underbrace{2X^2 D(n+d)\log_2(1+TX^2(1+C^2)/\lambda)}_{:=\gamma_2} T/\tau.$$

Plugging everything together. We are now ready to bring all the regret terms together and obtain

$$R(T) \leq (2\gamma\alpha + \gamma_3)\sum_{t=0}^{T} \|z_t\|_{V_t^{-1}} \mathbb{1}_{E_t} + \gamma_2 T/\tau$$
$$+ \gamma_1\sqrt{T} + \sum_{k=1}^{K} T_k\big(R_{t_k}^{\mathrm{TS},1} + R_{t_k}^{\mathrm{TS},3}\big)\mathbb{1}_{E_{t_k}}$$

At this point, the regret bound is decomposed into several parts: 1) the first term can be bounded as $\sum_{t=0}^{T} \|z_t\|_{V_t^{-1}} = \tilde{O}(\sqrt{T})$ on E using Prop. 4 (see App. E for details) 2) two terms which are already conveniently bounded as T/τ and $\sqrt{T}$, and 3) two remaining terms from R^{TS} that are almost exact martingales. In fact, T_k is random w.r.t. $\mathcal{F}_{t_k}$ and thus the terms $T_k R_{t_k}^{\mathrm{TS},1}$ and $T_k R_{t_k}^{\mathrm{TS},3}$ are not proper martingale difference sequences. However, we can leverage on the fact that on most of the episodes, the length T_k is not random since the termination of the episode is triggered by the (deterministic) condition $T_k \leq \tau$. Let $\alpha_k = (R_{t_k}^{\mathrm{TS},1} + R_{t_k}^{\mathrm{TS},3})\mathbb{1}_{E_{t_k}}$, $\mathcal{K}^{\det}$ and $\mathcal{K}^{\mathrm{len}}$ two set of indexes of cardinality $K^{\det}$ and K^{len} respectively, which correspond to the episodes terminated following the determinant or the limit condition respectively. Then, we can write

$$\sum_{k=1}^{K} T_k\alpha_k = \sum_{k\in\mathcal{K}^{\det}} T_k\alpha_k + \tau\sum_{k\in\mathcal{K}^{\mathrm{len}}} \alpha_k$$
$$\leq \sum_{k\in\mathcal{K}^{\det}} T_k\alpha_k + \sum_{k\in\mathcal{K}^{\mathrm{len}}} \tau\alpha_k + \sum_{k\in\mathcal{K}^{\det}} \tau\alpha_k + \sum_{k\in\mathcal{K}^{\det}} \tau\|\alpha_k\|$$
$$\leq 2\tau\sum_{k\in\mathcal{K}^{\det}} \|\alpha_k\| + \tau\sum_{k=1}^{K}\alpha_k.$$

The first term can be bounded using Lem. 6, which implies that the number of episodes triggered by the determinant condition is only logarithmic. On the other hand the remaining term $\sum_{k=1}^{K}\alpha_k$ is now a proper martingale and, together with the boundedness of α_k on event E, Azuma inequality directly holds. We obtain

$$\sum_{k=1}^{K} T_k\big(R_{t_k}^{\mathrm{TS},1} + R_{t_k}^{\mathrm{TS},3}\big)\mathbb{1}_{E_{t_k}} = \tilde{O}(\tau\sqrt{K}).$$

w.p. $1-\delta/2$. Grouping all higher-order terms w.r.t. to T and applying Cor. 2 to bound K, we finally have

$$R(T) \leq C_1\frac{T}{\tau} + C_2\tau\sqrt{T/\tau},$$

where C_1 and C_2 are suitable problem-dependent constants. This final bound is optimized for $\tau = O(T^{1/3})$ and it induces the final regret bound $R(T) = O(T^{2/3})$. More details are reported in App. E.

5 Discussion

We derived the first 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 new line of proof that directly relates the regret of TS and the controls executed over time. Furthermore, we show that TS has to solve a trade-off between frequently updating the policy to guarantee enough optimistic samples and reducing the number of policy switches to limit the regret incurred at each change. This gives rise to a final bound of $O(T^{2/3})$. This opens a number of questions. **1)** The current analysis is derived in the general n/d-dimensional case except for Lem. 3 and the steps leading to the introduction of the state in Sect. 4.4, where we set $n = d = 1$. We believe that these steps can be extended to the general case without affecting the final result. **2)** The final regret bound is in striking contrast with previous results for TS. While we provide a rather intuitive reason on the source of this extra regret, it is an open question whether a different TS or analysis could allow to improve the regret to $O(\sqrt{T})$ or whether this result reveals an intrinsic limitation of the randomized approach of TS.

Acknowledgement This research is supported in part by a grant from CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, CRIStAL (Centre de Recherche en Informatique et Automatique de Lille), and the French National Research Agency (ANR) under project ExTra-Learn n.ANR-14-CE24-0010-01.

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, pages 1–26, 2011.

[2] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2015.

[3] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

[4] Gabriel Acosta and Ricardo G Durán. An optimal poincaré inequality in l 1 for convex domains. *Proceedings of the american mathematical society*, pages 195–202, 2004.

[5] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*, 2012.

[6] S Bittanti and MC Campi. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, 6(4):299–320, 2006.

[7] Marco C Campi and PR Kumar. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

[8] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010.

[9] P. Lancaster and L. Rodman. *Algebraic riccati equations*. Oxford University Press, 1995.

[10] Ian Osband and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

[11] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2377–2386, 2016.

[12] Lawrence E Payne and Hans F Weinberger. An optimal poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960.

[13] Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite mdps: Pac analysis. *J. Mach. Learn. Res.*, 10:2413–2444, December 2009.

[14] Malcolm J. A. Strens. A bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950, 2000.