Improved Regret Bounds for Thompson Sampling in Linear Quadratic Control Problems

Marc Abeille [1] **Alessandro Lazaric** [2]

Abstract

Thompson sampling (TS) is an effective approach to trade off exploration and exploration in reinforcement learning. Despite its empirical success and recent advances, its theoretical analysis is often limited to the Bayesian setting, finite state-action spaces, or finite-horizon problems. In this paper, we study an instance of TS in the challenging setting of the infinite-horizon linear quadratic (LQ) control, which models problems with continuous state-action variables, linear dynamics, and quadratic cost. In particular, we analyze the regret in the frequentist sense (i.e., for a fixed unknown environment) in one-dimensional systems. We derive the first $O(\sqrt{T})$ frequentist regret bound for this problem, thus significantly improving the $O(T^{2/3})$ bound of Abeille & Lazaric (2017) and matching the frequentist performance derived by Abbasi-Yadkori & Szepesvári (2011) for an optimistic approach and the Bayesian result of Ouyang et al. (2017). We obtain this result by developing a novel bound on the regret due to policy switches, which holds for LQ systems of any dimensionality and it allows updating the parameters and the policy at each step, thus overcoming previous limitations due to lazy updates. Finally, we report numerical simulations supporting the conjecture that our result extends to multi-dimensional systems.

1. Introduction

Designing algorithms to properly trade off *exploring* an unknown environment and *exploiting* the estimated optimal control policy, is one of the most important challenges towards scaling reinforcement learning (RL) (Sutton & Barto, 1998) to problems with large or continuous state and action spaces. The exploration-exploitation dilemma in RL has been mostly addressed following two main approaches: optimism-in-face-of-uncertainty (OFU) and Thompson sampling (TS) (also referred to as posterior-sampling in RL, PSRL (Strens, 2000)). Following the OFU approach (see e.g., Jaksch et al., 2010), we first construct the set of admissible environments whose parameters are within confidence intervals constructed on the samples collected so far. Then the optimal policy w.r.t. the *optimistic* admissible environment (i.e., the one maximizing the optimal value function) is executed. PSRL is a Bayesian algorithm that executes the optimal policy w.r.t. a environment drawn at *random* from the posterior computed from a given prior and the samples observed so far. The flexibility of applying PSRL to any parametric MDP (as soon as sampling from a posterior can be done efficiently) and the difficulty of deriving tight confidence intervals and computing the optimistic policy in OFU, often make PSRL the more effective approach. Empirical evidence of the performance of PSRL ranges from multi-armed bandit (Chapelle & Li, 2011) to RL with linear and non-linear function approximation (Osband et al., 2016b;a).

On the theoretical side, several results exist on PSRL in the episodic setting under the regret framework, where the cumulative reward collected by the learning agent is compared to the performance of the optimal policy. Given the Bayesian nature of PSRL, its regret is often analyzed in expectation w.r.t. the prior over the environments (i.e., the so-called *Bayesian* regret). Osband et al. (2013) proved the first regret bound for PSRL in finite MDPs of order $O(S\sqrt{AT})$. The more general setting of learning in parameterized MDPs is studied in (Osband & Van Roy, 2014), where it is shown that the regret of PSRL depends on the dimensionality of the space of parameters rather than its cardinality. Osband & Roy (2016) showed how leveraging over posterior sampling in PSRL actually leads to improved regret bounds w.r.t. OFU in finite MDPs. Finally, Gopalan & Mannor (2015) proved regret bounds in a slightly more general non-episodic setting under the assumption that the MDP is ergodic and an initial state is positive recurrent under any policy. While these results match the frequentist bounds of OFU, when moving from the episodic to infinite horizon setting the results of TS are still limited. While most of the results for OFU hold in both cases (see e.g. (Bartlett & Tewari, 2009; Jaksch et al., 2010; Abbasi-Yadkori &

[1]Criteo, Paris, France [2]Facebook AI Research, Paris, France. Correspondence to: Marc Abeille <m.abeille@criteo.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Szepesvári, 2011)), Osband & 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 & Szepesvári, 2015). While Osband & Van Roy (2016) conjecture that standard $O(\sqrt{T})$ bounds should hold for TS in infinite horizon, Agrawal & Jia (2017) only recently proved that a specific *optimistic* version of PSRL is able to achieve a $O(\sqrt{T})$ regret in finite MDPs, while Kim (2017) proved problem-dependent $\log T$ regret for ergodic finite parametric MDPs. On the other hand, for continuous state-action spaces, Ouyang et al. (2017) proved a $O(\sqrt{T})$ regret only in the Bayesian setting for infinite-horizon linear quadratic (LQ) control, while Abeille & Lazaric (2017) showed that TS may suffer a frequentist regret $O(T^{2/3})$, which is significantly worse than the $O(\sqrt{T})$ result proved by Abbasi-Yadkori & Szepesvári (2011) and Faradonbeh et al. (2017) for OFU. Abeille & Lazaric (2017) justify this result by an unfavourable trade-off between lazy updates and number of optimistic steps, which leads to shorter episodes than in OFU. While increasing the number of episodes also increases the chance of selecting optimistic environments, which is critical for the functioning of TS, this leads to a larger regret due to the continuous switching between different policies.

In this paper, we build on the result of Abeille & Lazaric (2017) and we prove that their bound can be reduced to $O(\sqrt{T})$. This result is obtained using a novel bound on the regret suffered at the switch between two episodes (the *consistency regret* in (Abeille & Lazaric, 2017)). We show that the regret incurred at policy switches is related to expected absolute deviation of the solution to the Riccati equation under the TS distribution, which is cumulatively bounded as $O(\sqrt{T})$. 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. While the novel bound on the consistency regret is derived for the general case, our final regret bound relies on a lower-bound on the probability of optimistic sampling derived by Abeille & Lazaric (2017), which only holds for one-dimensional systems. We conjecture that a similar bound should hold for any dimension and we provide preliminary numerical simulations to support such conjecture.

2. Preliminaries

Most of the material in this section is borrowed from Abbasi-Yadkori & Szepesvári (2011) and Abeille & Lazaric (2017).

Notation. For any matrix $A \in \mathbb{R}^{n\times m}$, we denote as $A[i,j]$ its i,j component. For any vector x and matrix A of appropriate dimensions we define the following norms $\|x\| = \sqrt{x^\mathsf{T} x}$, $\|A\|_F = \mathrm{Tr}(AA^\mathsf{T})^{1/2}$, $\|A\|_2 = \sup_{\|x\|=1}\|Ax\|$ and for any positive definite matrix V, $\|A\|_V = \|V^{1/2}A\|_F$. We use $\overset{d}{=}$ to denote equality in distribution.

The control problem. At any time t, given state $x_t \in \mathbb{R}^n$ and control $u_t \in \mathbb{R}^d$, in a linear quadratic (LQ) control problem, the next state and cost are obtained as

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

where A_*, B_*, Q, R are matrices of appropriate dimension and $\{\epsilon_{t+1}\}_t$ is a zero-mean process. The dynamics parameters are summarized in $\theta_*^\mathsf{T} = (A_*, B_*)$. The solution to an LQ is a stationary deterministic policy $\pi : \mathbb{R}^n \to \mathbb{R}^d$ mapping states to controls minimizing the infinite-horizon average expected cost

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

with $x_0 = 0$ and $u_t = \pi(x_t)$. We denote as $\pi_*(\theta_*)$ the optimal policy of the LQ parametrized by θ_*.

Let $\mathcal{F}_t = \sigma(x_0, u_0, \dots, x_t, u_t)$ be the filtration that represents the knowledge up to time t. We also define the "extended" filtration $\mathcal{F}_t^x = (\mathcal{F}_{t-1}, x_t)$. We impose the following assumptions over the noise process and the linear system of Eq. 1.

Assumption 1 (Noise)**.** *The noise $\{\epsilon_t\}_t$ is a $\mathcal{F}_t-$martingale difference sequence and it is conditionally Gaussian, i.e., $\epsilon_t|\mathcal{F}_t \sim \mathcal{N}(0, I)$ for all $t \le T$.*

Assumption 2 (LQ)**.** *The cost matrices Q and R are symmetric p.d. and (A_*, B_*) is stabilizable.*[1]

Under Asm. 1 and 2, Thm.16.6.4 in (Lancaster & Rodman, 1995) guarantees the existence and uniqueness of an optimal policy $\pi_*(\theta_*) = K(\theta_*)x$, where,

$$\begin{aligned} K(\theta_*) &= -\big(R + B^\mathsf{T} P(\theta_*) B\big)^{-1} B^\mathsf{T} P(\theta_*) A, \\ P(\theta_*) &= Q + A^\mathsf{T} P(\theta_*) A + A^\mathsf{T} P(\theta_*) B K(\theta_*). \end{aligned} \tag{3}$$

The optimal average cost is $J_* = J_{\pi_*}(\theta_*) = \mathrm{Tr}(P(\theta_*))$. For notational convenience, we introduce $H(\theta_*)^\mathsf{T} = \begin{pmatrix} I & K(\theta_*)^\mathsf{T} \end{pmatrix}$ and we have that the closed-loop matrix $A_* + B_* K(\theta_*) = \theta_*^\mathsf{T} H(\theta_*)$ is asymptotically stable. Finally, we recall the definition of *differential* value function in state x as $V(x;\theta_*) = x^\mathsf{T} P(\theta_*) x$, which measures the (limit) difference in cumulative cost between executing the optimal policy from x and the average cost $J(\theta_*)$. Finally, we recall a result about the regularity of the Riccati solution.

Proposition 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 $P(\theta)$ be the unique solution of Eq. 3. Then, for any compact set*

$$\mathcal{S}_0 \subset \{\theta \in \mathbb{R}^{(n+d)\times n} \text{ s.t. } \theta^\mathsf{T} = (A, B) \text{ stabilizable}\},$$

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

the mapping $\theta \in \mathcal{S}_0 \to P(\theta)$ is continuously differentiable. Furthermore, let $A_c(\theta) = \theta^\mathsf{T} H(\theta)$ be the closed-loop matrix, then the directional derivative[2] of $P(\theta)$ in a direction $\delta\theta$, denoted as $dP(\theta)(\delta\theta)$, is the solution of the Lyapunov equation

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

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

The learning problem. Following the setting of Abbasi-Yadkori & Szepesvári (2011), we assume that Q and R are known, while θ_* needs to be estimated from data. 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. 1. The learning performance is measured by the cumulative regret over T steps defined as

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

Exploiting the linearity of the dynamics, the unknown parameter θ_* can be directly estimated from data by regularized least-squares (RLS). For any sequence of controls $(u_0, \dots, u_t)$ and the induced states $(x_0, x_1, \dots, x_{t+1})$, let $z_t = (x_t, u_t)^\mathsf{T}$, then the RLS estimator with regularization parameter $\lambda \in \mathbb{R}_+^*$ is computed as

$$\widehat{\theta}_t = V_t^{-1} \sum_{s=0}^{t-1} z_s x_{s+1}^\mathsf{T}, \text{ with } V_t = \lambda I + \sum_{s=0}^{t-1} z_s z_s^\mathsf{T}. \quad (4)$$

Proposition 2 (Thm. 2 in Abbasi-Yadkori et al. 2011)**.** *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*

$$\|\theta_* - \widehat{\theta}_t\|_{V_t} \le \beta_t(\delta),$$
$$\beta_t(\delta) = n\sqrt{2\log\Big(\frac{\det(V_t)^{1/2}}{\det(\lambda I)^{1/2}\delta}\Big)} + \lambda^{1/2} Tr(\theta_* \theta_*^\mathsf{T}), \quad (5)$$

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

Finally, we recall this standard result of RLS.

Proposition 3 (Lem. 10 in Abbasi-Yadkori & Szepesvári 2011)**.** *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*

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

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

$$\sum_{s=0}^{t} \|z_s\|^2_{V_s^{-1}} \le 2\frac{Z^2}{\lambda}(n+d)\log\Big(1 + \frac{(t+1)Z^2}{\lambda(n+d)}\Big). \quad (6)$$

[2]This proposition fixes Lem. 1 of Abeille & Lazaric (2017), which incorrectly reports that the inner product $\nabla J(\theta)^\mathsf{T} \delta\theta$ is the solution of the Lyapunov equation. As they eventually consider one-dimensional systems, their derivation is still correct.

3. Thompson Sampling for LQR

Input: $\hat{\theta}_0, V_0 = \lambda I, \delta, T, t_0 = 0$
1: Set β_t according to Eq. 5
2: **for** $t = \{0, \dots, T\}$ **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
7: **end for**

Figure 1: Thompson sampling algorithm for LQ.

Following Abeille & Lazaric (2017), we define TS for LQ as a randomized algorithm that samples a system $\widetilde{\theta}$, computes the corresponding optimal policy $\pi(\widetilde{\theta})$ and performs the associated controls $K(\widetilde{\theta})x$. As θ_* is initially unknown and TS may sample systems that are not stabilizable, we first need to characterize the set of "admissible" systems that the learner may use. We first recall the following result.

Proposition 4 (Cor. 12 in (Klamka, 2016))**.** *For any dimensions n and d, the set of controllable dynamical systems is open and dense in the space $\mathbb{R}^{n(n+d)}$ of all dynamical systems of the form* (1)*.*

Since controllability implies stabilizability, the previous result implies that the set of non-stabilizable systems is of zero Lebesgue measure. Since for a non-stabilizable system (A, B), there exists no control K that can make the dynamics stable, then the state process x_t diverges exponentially and the associated "optimal" average cost is $J(\theta) = +\infty$. This property has two major implications: *1)* 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; *2)* for any sampling distribution absolutely continuous w.r.t the Lebesgue measure, the associated optimal average cost is finite with probability one. However, TS may still sample parameters that are *almost* non-stabilizable (and hence of large cost), which is still harmful from both theoretical and practical perspective. This motivates the introduction of the constraint set

$$\mathcal{S} = \{\theta^\mathsf{T} \in \mathbb{R}^{n(n+d)}, \|\theta^\mathsf{T} H(\theta)\|_2 \le \rho < 1 \text{ and } \mathrm{Tr}(\theta\theta^\mathsf{T}) \le S^2\}.$$

This definition directly implies the following guarantees.

Proposition 5. *$\mathcal{S}$ is a compact set. For any $\theta \in \mathcal{S}$, θ is a stabilizable pair and there exist $D < \infty$ and $C < \infty$ positive constants such that $D = \sup_{\theta\in\mathcal{S}} J(\theta)$ and $C = \sup_{\theta\in\mathcal{S}} \|K(\theta)\|_2$.*

We now introduce the TS instance that we study. At each step t, given the RLS-estimate $\widehat{\theta}_t$, TS samples a perturbation matrix $\eta \in \mathbb{R}^{(n+d)\times n}$ from a component-wise *normal* distribution and computes the *perturbed* parameter $\widetilde{\theta}_t$ as

$$\widetilde{\theta}_t \overset{d}{=} \mathcal{R}_{\mathcal{S}}(\widehat{\theta}_t + \beta_t V_t^{-1/2}\eta), \tag{7}$$

where $\beta_t = \beta_t(\delta/8T)$ is defined as in (5) with $\mathrm{Tr}(\theta_*\theta_*^\mathsf{T})$ replaced by S and $\mathcal{R}_{\mathcal{S}}$ is the rejection sampling w.r.t. the admissible set $\mathcal{S}$. 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 $\widehat{\theta}_t$ and V_t.

Unlike many previous instances, we consider the most basic implementation of the TS principle. Alg. 1 does not require any initial exploration phase (see e.g., Faradonbeh et al. 2017 in the case of OFU) and it directly executes the control associated with the sampled environment, instead of using optimistic sampling as in (Agrawal & Jia, 2017) for finite MDPs. Unlike the instances of TS studied by Abeille & Lazaric (2017), Abbasi-Yadkori & Szepesvári (2015), and Ouyang et al. (2017), we consider a "frequent-update" version of TS, where a new parameter $\widetilde{\theta}_t$ is sampled at *each step*, without the need of introducing any notion of episode and stopping criterion. Furthermore, in order to avoid almost non-stabilizable systems, we only require a rejection sampling step over $\mathcal{S}$, which is always computationally feasible. This is in contrast with requiring a stabilizing controller as Abbasi-Yadkori & Szepesvári (2015), assuming recurrent states as Gopalan & Mannor (2015), stating by assumption that any sampled $\widetilde{\theta}$ is stabilizable as Ouyang et al. (2017), or assuming that the dynamics under different parameters is easily distinguishable (Kim, 2017). Finally, notice that the computational cost of frequent updates could be mitigated by sampling every *fixed* number of steps.

4. Theoretical analysis

In order to make the learning problem well-posed, we first state the following assumption[3].

Assumption 3. *Let $\mathcal{S}$ be the set of admissible systems, then $\theta_* \in \mathcal{S}$.*

In the rest of the paper, we prove this novel bound of the frequentist regret.

Theorem 1. *Consider the LQ system in Eq. 1 of dimension $n = 1$ and arbitrary d. Under assumptions 1, 2 and 3, for any $0 < \delta < 1$, the cumulative regret of* TS *(Alg. 1) over T steps is bounded w.p. at least $1-\delta$ as*[4] *$R_T = \widetilde{O}(\sqrt{T})$ where $\widetilde{O}$ hides logarithmic factors in T and $1/\delta$, and problem dependent constants.*

[3]We discuss the validity of this assumption in App. E.

[4]Explicit numerical and problem dependent constants can be collected from the proof.

As discussed in Sect. 3, here we analyze a very *basic* instance of TS, which removes any restriction on the algorithm. This result greatly improves the previous analysis by Abeille & Lazaric (2017) by reducing the regret from $O(T^{2/3})$ down to $O(\sqrt{T})$, thus matching the performance of OFU for LQ problems (Abbasi-Yadkori & Szepesvári, 2011; Faradonbeh et al., 2017). Together with the recent result of Agrawal & Jia (2017), Thm. 1 confirms the conjecture of Osband & Van Roy (2016) that $O(\sqrt{T})$ regret guarantees of TS extends to learning in the infinite-horizon setting. Furthermore, this result shows that the frequentist regret is as small as existing Bayesian regret guarantees for (variants) of TS (Ouyang et al., 2017). Finally, beside the improvement on the rate, Thm. 1 also removes the (somewhat unnatural) inverse dependency on $J(\theta_*)$ in the previous bound (see Eq.11 in Abeille & Lazaric, 2017). The main limitation of this result lies on the restriction to $n = 1$ dimensional problems. This limitation is due to the proof lower-bounding the probability of being optimistic (Lem.3 in Abeille & Lazaric, 2017), which we could not extend beyond $n = 1$. On the other hand, as illustrated in the next section, all the new results we develop in this paper (in particular the bound on the regret of policy switches) hold in the general case $n \geq 1$, $d \geq 1$. See also Sect. 6 for further discussion.

4.1. Challenges and Main Tools

In this section we discuss the technical challenges in proving the final $O(\sqrt{T})$ regret bound and we report the core results that enable the final result in Thm. 1.

Regret decomposition. We introduce the same events as in (Abeille & Lazaric, 2017).

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

$$\mathcal{E}_t^{\mathrm{RLS}} = \{\theta \in \mathbb{R}^{(n+d)n} \text{ s.t. } \|\theta - \widehat{\theta}_t\|_{V_t} \leq \beta_t\},$$
$$\mathcal{E}_t^{\mathrm{TS}} = \{\theta \in \mathbb{R}^{(n+d)n} \text{ s.t. } \|\theta - \widehat{\theta}_t\|_{V_t} \leq \gamma_t\},$$

where $\gamma_t := n\sqrt{2(n+d)\log\left(2n(n+d)/\delta'\right)}\beta_t$ and introduce the event (RLS estimate concentration) $\widehat{E}_t = \left\{\forall s \leq t,\ \widehat{\theta}_s \in \mathcal{E}_s^{\mathrm{RLS}}\right\}$ and the event (parameter $\widetilde{\theta}_s$ concentrates around $\widehat{\theta}_s$) $\widetilde{E}_t = \left\{\forall s \leq t,\ \widetilde{\theta}_s \in \mathcal{E}_t^{\mathrm{TS}}\right\}$. Furthermore, let X, X' be two problem dependent positive constants, we define the event (bounded state) $\bar{E}_t = \left\{\forall s \leq t,\ \|x_s\| \leq X\log\frac{X'}{\delta}\right\}$. Finally, we define $E_t = \widehat{E}_t \cap \widetilde{E}_t \cap \bar{E}_t$.

Proposition 6 (Cor.1 in (Abeille & Lazaric, 2017))**.** *The events in Def. 1 jointly hold with high probability, i.e., $\mathbb{P}(\widehat{E}_T \cap \widetilde{E}_T \cap \bar{E}_T) \geq 1 - \delta/2$.*

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 (Sect. 4.2 Abbasi-Yadkori & Szepesvári, 2011) as

$$R_T \mathbb{1}\{E_T\} \leq$$

$$\sum_{t=0}^{T} \{J(\widetilde{\theta}_t) - J(\theta_*)\} \mathbb{1}_{\{E_t\}} \qquad (R^{\text{TS}})$$

$$+ \sum_{t=0}^{T} \{z_t^\intercal \widetilde{\theta}_t P(\widetilde{\theta}_t) \widetilde{\theta}_t^\intercal z_t - z_t^\intercal \theta_* P(\widetilde{\theta}_t) \theta_*^\intercal z_t\} \mathbb{1}_{\{E_t\}} \qquad (R^{\text{RLS}})$$

$$+ \sum_{t=0}^{T} \{V(x_t; \widetilde{\theta}_t) \mathbb{1}_{\{E_t\}} - \mathbb{E}\big[V(x_{t+1}; \widetilde{\theta}_{t+1}) \mathbb{1}_{\{E_{t+1}\}} | \mathcal{F}_t\big]\} \quad (R^{\text{mart}})$$

$$+ \sum_{t=0}^{T} \mathbb{E}\big[(V(x_{t+1}; \widetilde{\theta}_{t+1}) - V(x_{t+1}; \widetilde{\theta}_t)) \mathbb{1}_{\{E_{t+1}\}} | \mathcal{F}_t\big]. \qquad (R^{\text{gap}})$$

As discussed in (Abeille & Lazaric, 2017), while R^{RLS} and R^{mart} can be easily bounded using Prop. 3 and Azuma's inequality, the main technical challenge is to prove the boundedness of R^{TS} and R^{gap}.[5] In fact, the former requires TS to sample *optimistic* parameters $\widetilde{\theta}_t$ as often as possible, so that $J(\widetilde{\theta}_t) \leq J(\theta_*)$ and R^{TS} is cumulatively small. As Abeille & Lazaric (2017) (Lemma 3) proved that TS has a fixed probability to sample optimistic parameters, this suggests implementing a frequent update scheme, where a new parameter is sampled at each time step (as in Alg. 1). Nonetheless, this may cause a linear regret in R^{gap}, as the difference between the differential value function at state x_{t+1} for any two different parameters $\widetilde{\theta}, \widetilde{\theta}'$ is bounded by a constant. In fact, let assume $\|x_{t+1}\| \leq X$ and $\widetilde{\theta}, \widetilde{\theta}' \in \mathcal{S}$, then

$$x_{t+1}^\intercal \big(P(\theta') - P(\theta)\big) x_{t+1} = V(x_{t+1}; \widetilde{\theta}') - V(x_{t+1}; \widetilde{\theta}) \leq X^2 \|P(\widetilde{\theta}) - P(\widetilde{\theta}')\|_2 \leq 2X^2 D, \qquad (8)$$

where we use the fact that $\|P(\theta)\|_2 \leq \text{Tr} P(\theta) = J(\theta) \leq D$ for any θ in $\mathcal{S}$. Based on this observation, Abeille & Lazaric (2017) claimed that TS should trade off between frequent (and thus optimistic) updates and lazy updates to avoid regret associated to policy changes. Unfortunately, the resulting balance leads to an overall $O(T^{2/3})$ regret. Nonetheless, while the bound in Eq. 8 is correct in the *worst case*, in the next lemma, we prove that the *expected* difference in the regret decomposition is such that R^{gap} is cumulatively bounded by $O(\sqrt{T})$, even when a new parameter is sampled at each step.

Lemma 1. *Consider the LQ system in Eq. 1 of dimension n and d. Under Asm. 1 and 2, for any $0 < \delta < 1$, the regret R^{gap} incurred by running the* TS *(Alg, 1) is bounded w.p. at least $1 - \delta/6$ as*[6] *$R^{gap} = \widetilde{O}(\sqrt{T})$.*

While the full proof of the lemma is postponed to Sect. 5, we provide a first intuition behind this result. By inspecting the definition of R^{gap} and Eq. 8 we notice that if $\widetilde{\theta}$ and $\widetilde{\theta}'$ were generated by the same distribution (i.e., they were independent realizations of the TS distribution at time t), then the expectation of the norm $\|P(\widetilde{\theta}) - P(\widetilde{\theta}')\|_2$ could be simply upper-bounded by twice the expected absolute deviation, which is bounded in the next lemma (the need for the two slightly different definitions is detailed in Sect. 5).

Lemma 2. *For arbitrary dimensions n and d, for any $t \geq 1$, let $\overline{\theta}_t$ be the un-rejected sampling parameter, i.e., $\overline{\theta}_t \overset{d}{=} \widehat{\theta}_t + \beta_t V_t^{-1/2} \eta$, where $\eta \in \mathbb{R}^{(n+d)\times n}$ is component-wise normal. We introduce two expected values for P*

$$P_t = \mathbb{E}\big(P(\widetilde{\theta}_t) | \mathcal{F}_t^x, E_t\big); \quad \overline{P}_t = \mathbb{E}\big(P(\bar{\theta}_t) \mathbb{1}_{\mathcal{S}}(\bar{\theta}_t) | \mathcal{F}_t^x, E_t\big),$$

and we define the corresponding expected absolute deviation

$$\Delta_t = \mathbb{E}\left(\|P(\widetilde{\theta}_t) - P_t\|_2 \big| \mathcal{F}_t^x, E_t\right); \qquad (9)$$

$$\overline{\Delta}_t = \mathbb{E}\left(\|P(\widetilde{\theta}_t) - \bar{P}_t\|_2 \big| \mathcal{F}_t^x, E_t\right). \qquad (10)$$

Then for any $0 < \delta < 1$, w.p. at least $1 - \delta/12$,

$$\sum_{t=0}^{T} \Delta_t = \widetilde{O}(\sqrt{T}) \quad \text{and} \quad \sum_{t=0}^{T} \bar{\Delta}_t = \widetilde{O}(\sqrt{T}).$$

Intuitively, the absolute deviation measures the expected *diameter* of the TS sampling ellipsoid using the ℓ_2-norm of the Riccati solution $P(\theta)$ as a metric. The fact that this is cumulative small indicates that the TS generates distributions (i.e., the sampling ellipsoid) that are *well-adapted* to the sensitivity of the Riccati solution, so that deviations in $P(\theta)$ are kept small. The impact of such behavior is even clearer if we consider the average gain $J(\theta) = \text{Tr}(P(\theta))$. Since $\|A\|_2 \leq \text{Tr}(A)$, we have that Lemma 2 directly implies that

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

Notice that while this quantity was previously bounded by Abeille & Lazaric (2017), the result of Lem. 2 is more general and it also removes *problem-dependent* quantities such as $J(\theta_*)$.

Unfortunately, Lem. 2 is not sufficient to bound R^{gap}, as $\widetilde{\theta}_t$ and $\widetilde{\theta}_{t+1}$ follow different distributions. Then we also need to control how the evolution of the distribution of $\widetilde{\theta}$ over two consecutive time steps influences the value of $P(\widetilde{\theta})$. By carefully decomposing R^{gap}, we only need to measure the effect of shifting distributions for "un-rejected" parameters (i.e., obtained by removing the rejection sampling step), which is bounded in the next lemma.

Lemma 3. *Let $\overline{P}_t$ as defined in Lem. 2, then the difference between two consecutive steps is bounded almost surely as*

$$\sum_{t=0}^{T} \|\overline{P}_{t+1} - \overline{P}_t\|_F \mathbb{1}_{\{E_{t+1}\}} \leq \widetilde{O}(\sqrt{T}).$$

[5] For a more thorough discussion of the challenges of proving frequentist regret bounds for TS in LQ problems, refer to (Sect. 4.2, Abeille & Lazaric, 2017).

[6] Explicit constant is provided in App. D.2.

The combination of lemmas 2 and 3 finally leads to Lem. 1. The final regret bound is then obtained following similar steps as in Abeille & Lazaric (2017) to bound the remaining terms (notably R^{TS}).

5. Proof

In this section we prove the lemmas reported in the previous section, while technical details are left to the supplement.

5.1. Bounding the average absolute deviation

Lem. 2 bounds two different instances of the *average absolute deviation* Δ_t and $\bar{\Delta}_t$, which only differ in the way the rejection sampling is treated. We focus here on bounding Δ_t and postpone the other part of the proof to App. B.1. We follow similar steps as Abeille & Lazaric (2017):[7] **1)** we use the weighted Poincaré inequality to relate Δ_t to the average gradient norm, **2)** we link this gradient to the control $K(\widetilde{\theta}_t)$ selected by TS over time *on average*, **3)** we introduce the state x_t to obtain a bound that depends on $\|z_t\|_{V_t^{-1}}$ which is cumulatively bounded by Prop. 3.

First, we use the relationship $\widetilde{\theta}_t \overset{d}{=} \bar{\theta}_t | \mathcal{S}$ to deal with the rejection sampling and we use the fact that for any matrix $A \in \mathbb{R}^{n\times n}$, $\|A\|_F \leq \sum_{i,j=1}^n |A[i,j]|$ to rewrite Δ_t as

$$\begin{aligned}\Delta_t &= \frac{\mathbb{E}\left[\|P(\bar{\theta}_t) - P_t\|_F \mathbb{1}_{\mathcal{S}}(\bar{\theta}_t) \middle| \mathcal{F}_t^x, E_t\right]}{\mathbb{P}(\bar{\theta}_t \in \mathcal{S} | \mathcal{F}_t^x, E_t)} \\ &\leq \frac{\sum_{i,j=1}^n \Delta_t^{ij}}{\mathbb{P}(\bar{\theta}_t \in \mathcal{S} | \mathcal{F}_t^x, E_t)},\end{aligned}$$

where $\Delta_t^{ij} = \mathbb{E}\left[|P(\bar{\theta}_t)[i,j] - P_t[i,j]| \mathbb{1}_{\mathcal{S}}(\bar{\theta}_t) \middle| \mathcal{F}_t^x, E_t\right]$. We apply the following Poincaré inequality.

Proposition 7 (Lemma. 4 in (Abeille & Lazaric, 2017))**.** *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*

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

Recalling the definition of $\bar{\theta}_t \overset{d}{=} \widehat{\theta}_t + \beta_t V_t^{-1/2}\eta$, we can rewrite Δ_t^{ij} as a function of the random variable η as $\Delta_t^{ij} = \mathbb{E}_\eta\left[|f_t^{ij}(\eta)| \middle| \mathcal{F}_t^x, E_t\right]$ where f_t^{ij} is properly defined. From Prop. 1, we have that $\theta \to P(\theta)$ is continuously differentiable on $\mathcal{S}$, which implies that f_t^{ij} is continuously differentiable almost everywhere on $\mathbb{R}^{n(n+d)}$. Furthermore, the conditioning on E_t imposes that $\widetilde{\theta}_t \in \mathcal{E}_t^{\text{TS}}$ and thus $f_t^{ij} \in W^{1,1}(\Omega)$ where $\Omega = \{\eta \in \mathbb{R}^{n(n+d)} \text{ s.t. } \|\eta\| \leq \gamma_t/\beta_t\}$. Let Φ_η be the pdf of η. On Ω, up to a normalization factor, Φ_η is the product of standard gaussian pdf and hence log-concave and twice differentiable. Finally, it is easy to check that $\mathbb{E}(f_t^{ij}(\eta)|\mathcal{F}_t^x, E_t) = 0$. Therefore, we can apply Prop. 7 as

$$\begin{aligned}\Delta_t^{ij} &\leq 2\gamma_t/\beta_t \mathbb{E}\left[\|\nabla f_t^{ij}(\eta)\|_F \middle| \mathcal{F}_t^x, E_t\right], \\ &= 2\gamma_t/\beta_t \mathbb{E}\left[\beta_t \|V_t^{-1/2}\nabla P(\bar{\theta}_t)[i,j]\|_F \mathbb{1}_{\mathcal{S}}(\bar{\theta}_t) \middle| \mathcal{F}_t^x, E_t\right].\end{aligned}$$

Plugging everything together and reintegrating the conditioning on $\mathcal{S}$, one obtains:

$$\Delta_t \leq 2\gamma_t \sum_{i,j=1}^n \mathbb{E}\left[\|\nabla P(\widetilde{\theta}_t)[i,j]\|_{V_t^{-1}} \middle| \mathcal{F}_t^x, E_t\right]. \quad (11)$$

We use the following property which links the gradient of $P(\theta)[i,j]$ to the optimal controller $H(\theta)$ (proof in App. A).

Proposition 8. *For any $\theta \in \mathcal{S}$, for any i, j,*

$$\|\nabla P(\theta)[i,j]\|_{V_t^{-1}} \leq \frac{2\rho D}{1-\rho^2} \|H(\theta)\|_{V_t^{-1}}.$$

Using Prop. 8 in Eq. 11, one obtains:

$$\Delta_t \leq \frac{4\rho D \gamma_t n^2}{1-\rho^2} \mathbb{E}\left[\|H(\widetilde{\theta}_t)\|_{V_t^{-1}} \middle| \mathcal{F}_t^x, E_t\right]. \quad (12)$$

Finally, the last step of the proof consists in integrating the state x_t in Eq. 12, in order to obtain a bound that depends on $\|z_t\|_{V_t^{-1}}$ using the fact that $z_t = H(\widetilde{\theta}_t)x_t$. We rely on the following proposition (proof in App. B.2).

Proposition 9. *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,*

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

Further, if $\bar{\theta}_t \in \mathcal{E}^{\text{TS}}$, $\mathbb{1}_{\{\|x_t\| \leq \alpha_t\}} \leq \mathbb{1}_{\{\|x_t\| \leq \alpha\}}$ where

$$\alpha = (1+1/\beta_t^2)\left(\sqrt{2n\log(3n)} + \gamma_t + (1+C)SX\right).$$

Since on E_t, $\bar{\theta}_t \in \mathcal{E}_t^{\text{TS}}$ and $\widetilde{\theta}_t = \bar{\theta}_t | \mathcal{S}$, we can apply Prop. 9 to the conditional expectation in Eq. 12 to bound

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

[7] Despite the similarity in the structure, every step reported here follows slightly different path to e.g., properly deal with the rejection set and integrating the state x_t more carefully so as to avoid the dependency on $J(\theta^*)$.

and finally obtain $\Delta_t \leq \frac{8\alpha}{1+\frac{1}{\beta_t}} \frac{4\rho D \gamma_t n^2}{1-\rho^2} Y_t$, where

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

By the law of iterated expectation, $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = \mathbb{E}(\|z_t\|_{V_t^{-1}} \mathbb{1}_{\{\|x_t\| \leq \alpha\}} | \mathcal{F}_{t-1})$ and $\|z_t\|_{V_t^{-1}} \mathbb{1}_{\{\|x_t\| \leq \alpha\}} \leq (1+C)\alpha/\lambda$. As a result, $\{Y_t - \|z_t\|_{V_t^{-1}} \mathbb{1}_{\{\|x_t\| \leq \alpha\}}\}_{t \geq 0}$ is a $\mathcal{F}_{t-1}$ bounded martingale difference sequence and by Azuma's inequality and Prop. 3 w.p. at least $1 - \delta/12$

$$\sum_{t=0}^{T} \Delta_t \leq \gamma_{abs}\Big(\sum_{t=0}^{T} \|z_t\|_{V_t^{-1}} + \frac{2\alpha}{\lambda}(1+C)\sqrt{2T \log\big(\frac{12}{\delta}\big)}\Big),$$

which concludes the proof.

5.2. Bounding the consistency regret

We now bound the *consistency regret* in term of the average absolute deviation $\bar{\Delta}_t$ and the KL divergence between two subsequent sampling distributions. Let $R_t^{\text{gap}} = x_{t+1}^\mathsf{T}\big(P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\big)x_{t+1}\mathbb{1}_{E_{t+1}}$ (i.e., an element in R^{gap} without the expectation). On E_{t+1}, $\|x_{t+1}\| \leq X$ and thus

$$\begin{aligned} R_t^{\text{gap}} &\leq X^2 \|P(\widetilde{\theta}_{t+1}) - P(\widetilde{\theta}_t)\|_F \mathbb{1}_{E_{t+1}} \\ &\leq X^2 (\|P(\widetilde{\theta}_{t+1}) - \bar{P}_{t+1}\|_F + \|P(\widetilde{\theta}_t) - \bar{P}_t\|_F \\ &\qquad + \|\bar{P}_{t+1} - \bar{P}_t\|_F \mathbb{1}_{E_{t+1}}) \end{aligned}$$

where $\bar{P}_t$ is defined in the statement of the lemma. Plugging this inequality into $R^{\text{gap}} = \sum_{t=0}^{T} \mathbb{E}(R_t^{\text{gap}}|\mathcal{F}_t)$, using the law of iterated expectation, and Azuma's inequality, we obtained that w.p. at least $1 - \delta/12$

$$\begin{aligned} R^{\text{gap}} \leq & X^2 \sum_{t=0}^{T} \mathbb{E}(\bar{\Delta}_{t+1}|\mathcal{F}_t) + X^2 \sum_{t=0}^{T} \mathbb{E}(\bar{\Delta}_t|\mathcal{F}_t) \\ & + X^2 \sum_{t=0}^{T} \mathbb{E}(\|\bar{P}_{t+1} - \bar{P}_t\|_F \mathbb{1}_{E_{t+1}})|\mathcal{F}_t) \\ \leq & X^2 \sum_{t=0}^{T} (\bar{\Delta}_{t+1} + \Delta_t) + X^2 \sum_{t=0}^{T} \|\bar{P}_{t+1} - \bar{P}_t\|_F \mathbb{1}_{E_{t+1}} \\ & + 12\sqrt{n} D X^2 \sqrt{2T \log(12/\delta)}. \end{aligned}$$

While the first two terms can be directly bounded by Lem. 2, we need to prove Lem. 3 to conclude the proof of Lem. 1.

For any $t \geq 0$, denote by Φ_t and ϕ_t the pdf of $\bar{\theta}_t|\mathcal{F}_t^x, E_t$ and $\bar{\theta}_t|\mathcal{F}_t^x$ respectively. Since the conditioning on E_t is equivalent to $\bar{\theta}_t \in \mathcal{E}_t^{\text{TS}}$, one has $\Phi_t(\theta) = \phi_t(\theta)\mathbb{1}_{\mathcal{E}_t^{\text{TS}}}(\theta)/\mathbb{P}(\bar{\theta}_t \in \mathcal{E}_t^{\text{TS}})$. Rewriting $\bar{P}_t$ and $\bar{P}_{t+1}$ using integrals gives

$$\begin{aligned} \|\bar{P}_{t+1} - \bar{P}_t\|_F &= \|\int_{\mathcal{S}} P(\theta)\Phi_{t+1}(\theta)d\theta - \int_{\mathcal{S}} P(\theta)\Phi_t(\theta)d\theta\|_F \\ &\leq \int_{\mathcal{S}} \|P(\theta)\|_F |\Phi_{t+1}(\theta) - \Phi_t(\theta)|d\theta \\ &\leq \sqrt{n} D \int_{\mathcal{S}} |\Phi_{t+1}(\theta) - \Phi_t(\theta)|d\theta. \end{aligned}$$

Furthermore

$$\begin{aligned} \int_{\mathcal{S}} |\Phi_{t+1}(\theta) - \Phi_t(\theta)|d\theta &\leq \int |\Phi_{t+1}(\theta) - \phi_{t+1}(\theta)|d\theta \\ &+ \int_{\mathcal{S}} |\phi_{t+1}(\theta) - \phi_t(\theta)|d\theta + \int |\phi_t(\theta) - \Phi_t(\theta)|d\theta, \end{aligned}$$

and algebraic manipulations show that

$$\begin{aligned} &\int |\phi_t(\theta) - \Phi_t(\theta)|d\theta \leq 2\left(1 - \mathbb{P}(\bar{\theta}_t \in \mathcal{E}_t^{\text{TS}})\right) \leq 2\delta', \\ &\int |\phi_{t+1}(\theta) - \Phi_{t+1}(\theta)|d\theta \leq 2\left(1 - \mathbb{P}(\bar{\theta}_{t+1} \in \mathcal{E}_{t+1}^{\text{TS}})\right) \leq 2\delta'. \end{aligned}$$

Finally, using Pinsker's inequality, we have

$$\int_{\mathcal{S}} |\phi_{t+1}(\theta) - \phi_t(\theta)|d\theta \leq \sqrt{2\text{KL}(\phi_t||\phi_{t+1})},$$

which means that the *consistency regret* is kept under control as long as the KL divergence between two subsequent unrejected distributions is cumulatively small. This is provided by the following proposition (proof in App. C).

Proposition 10. *For any $t \geq 0$, on E_{t+1},*

$$\text{KL}(\phi_t||\phi_{t+1}) \leq \gamma_{\text{KL}} \|z_t\|^2_{V_t^{-1}}.$$

Using Prop. 3, we obtain that w.p. $1 - \delta/6$, $R^{gap} = \widetilde{O}(\sqrt{T})$.

6. Discussion

While Thm. 1 reports the first $O(\sqrt{T})$ frequentist regret bound for *standard* TS in infinite-horizon LQ problems, the final result is still limited to $n = 1$ systems because of the lower bound on the probability of being optimistic. In this section we provide preliminary evidence that the difficulty in extending the lower bound to $n \geq 1$ is not related to an intrinsic difference in the "nature" of the problem[8], but rather to technical challenges in the proof. We then conjecture that the $O(\sqrt{T})$ bound holds in systems of arbitrary dimension.

Stacked systems. Consider a n-dimensional system consisting in n independent 1-dimensional problems. The structure of the system is then diagonal, i.e., $A_* = \text{diag}(a_*^1, \dots, a_*^n)$, $A_* = \text{diag}(b_*^1, \dots, b_*^n)$ (similar for Q and R). Let 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, as $\text{Tr}(P(\widetilde{\theta}_t)) = \sum_{i=1}^{n} J(\widetilde{\theta}_t^i)$,

[8] While for $n = 1$, $J(\theta) = P(\theta)$ and any analysis on $P(\theta)$ directly transitions to guarantees for the average cost, for $n > 1$ we have $J(\theta) = \text{Tr}(P(\theta))$ and we need to study the *spectrum* of $P(\theta)$ to provide guarantees on the performance. In general, this may indeed make the problem considerably more difficult.

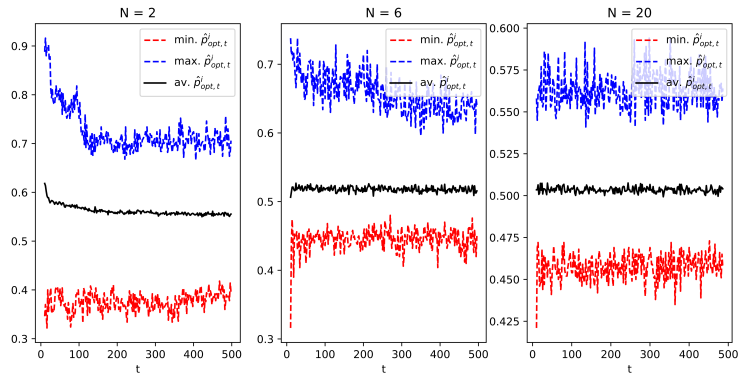

Figure 2: Evolution of $p_{\text{opt},t}$ for three systems of dimension $N = n(n+d)$.

n/d	1	2	3	4
1	0.021204	0.018000	0.032000	0.015400
2	0.038444	0.054600	0.021800	0.043009
3	0.010481	0.048926	0.098315	0.162633
4	0.030162	0.046220	0.060961	0.046920

Table 1: Worst case values for $p_{\text{opt},t}$.

$\text{Tr}(P(\theta_*)) = \sum_{i=1}^n J(\theta_*^i)$. As a consequence, the probability of being optimistic can be lower bounded as

$$\mathbb{P}\big(\text{Tr}(P(\widetilde{\theta}_t)) \le \text{Tr}(P(\theta_*))\big)$$
$$\ge \mathbb{P}\big(J(\widetilde{\theta}_t^i) \le J(\theta_*^i),\ \forall i = 1..n\big) \ge \prod_{i=1}^n \mathbb{P}\big(J(\widetilde{\theta}_t^i) \le J(\theta_*^i)\big).$$

By independence, the probability of being optimistic reduces to the probability of being jointly optimistic *in each* direction. In order to compensate for the reduction of probability due to the product in the previous expression, it is sufficient to slightly increase the "size" of the TS distribution by a $\sqrt{n}$ factor. While Prop. 2 can be written as

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

we need to sample parameters component-wise as $\widetilde{\theta}_t^i = \mathcal{R}_{\mathcal{S}}\big(\widehat{\theta}_t^i + \widetilde{\beta}_t^i V_t^{i,-1/2} \eta_t^i\big)$, with $\widetilde{\beta}_t^i \ge \sqrt{n}\beta_t^i$ for all $i \in [1, \dots, n]$. This $\sqrt{n}$ over-sampling ensures the joint probability to be constant (Lem. 3 of Abeille & Lazaric (2017)). Despite the lack of generality due to the diagonal structure, this example stresses the intuition that over-sampling (coming from $\widetilde{\beta}$) prevents the probability of being optimistic to poorly scale with the dimension, and that the difficulty in the analysis lies in the characterization of the optimistic set Θ^{opt}. While in the diagonal case one can focus on the joint probability of each system being optimistic, this is no longer possible in the general case, as increasing the cost along one direction may be compensated by decreasing it in another direction. In fact, as $J(\theta) = \text{Tr}(P(\theta))$, we need to control the overall sum of the eigenvalues of $P(\theta)$. Unfortunately, we are not aware of any perturbation theory for the trace of Riccati/Lyapunov solutions, which makes the analysis challenging. A more restrictive analysis may rather focus on a super-set of optimistic parameters $\Theta^{\text{opt},+}$ such that $P(\theta) \preccurlyeq P(\theta_*)$, thus requiring optimism in *every* direction. A preliminary study of this problem shows that it is possible to derive a lower bound on the mass of $\Theta^{\text{opt},+}$ which depends on t, thus failing to provide an "any-time" lower bound on the probability of being optimistic as required by the current proof.

Numerical simulations. Since all the results in Sect. 5 hold for $n \ge 1$, we try to simulate several random LQ systems of variable dimensionality and numerically estimate the probability of being optimistic (p_{opt}) in each of them. We first analyze the "evolution" of p_{opt} over time. We construct $\mathcal{S}$ by setting $D = 20 J(\theta_*)$. For different values of n and d, we sample θ_* as $\theta_*[i,j] \sim \mathcal{N}(0,1)$ independently and we run multiple trajectories of TS of length $T = 500$ steps. At each step t, we sample 1000 $\widetilde{\theta}_t$ from the TS distribution (with rejection), we compute the corresponding average cost $J(\widetilde{\theta}_t)$ and we compare it to $J(\theta_*)$. Since the early steps of the trajectories mostly depend on the initialization of the RLS, we wait for $t \ge 10$ to estimate p_{opt}. The fraction of $\widetilde{\theta}_t$ with $J(\widetilde{\theta}_t) \le J(\theta_*)$ corresponds to our estimate of $p_{\text{opt},t}$. In Fig. 2 we report the minimum, maximum, and average value of $p_{\text{opt},t}$ across 100 trajectories. We notice that for all values of n and d, p_{opt} rapidly increases and converges to a fixed value. Even considering the minimum value obtained across all trajectories at each step t, p_{opt} is always lower bounded away from 0. In order to better validate the fact that p_{opt} is lower bounded, we report additional results in Tab. 1, where we report the minimum value of $p_{\text{opt},t}$ at time $\bar{t} = 100$ across 100 trajectories and 5000 values of θ_* for many combinations of n and d.[9] Notice that this analysis is over-pessimistic since in order to prove the $O(\sqrt{T})$ regret we only need to lower bounded the probability of being optimistic *conditioned* on the high-probability event related to the RLS ellipsoid, which is somehow ignored here. Nonetheless, Tab. 1 illustrates that $p_{\text{opt},t}$ is indeed strictly bigger than zero, thus providing a numerical support to the conjecture that Lem.3 in (Abeille & Lazaric, 2017) does extend to LQ systems of arbitrary dimension.

[9] The choice of such an early $\bar{t}$ is that from Fig. 2, p_{opt} seems to converge fast and its smallest values are usually at early stages of the learning process.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, pp. 1–26, 2011.

Abbasi-Yadkori, Y. and Szepesvári, C. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2015.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

Abeille, M. and Lazaric, A. Thompson sampling for linear-quadratic control problems. In *AISTATS*, 2017.

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NIPS*, pp. 1184–1194, 2017.

Bartlett, P. L. and Tewari, A. 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.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2249–2257. Curran Associates, Inc., 2011.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time analysis of optimal adaptive policies for linear-quadratic systems. *CoRR*, abs/1711.07230, 2017. URL `http://arxiv.org/abs/1711.07230`.

Gopalan, A. and Mannor, S. Thompson sampling for learning parameterized markov decision processes. In *Proceedings of The 28th Conference on Learning Theory*, 2015.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010.

Kim, M. J. Thompson sampling for stochastic control: The finite parameter case. *IEEE Transactions on Automatic Control*, 62 (12):6415–6422, Dec 2017. ISSN 0018-9286. doi: 10.1109/TAC.2017.2653942.

Klamka, J. Controllability of dynamical systems. *Mathematica Applicanda*, 36(50/09):57–75, 2016.

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

Osband, I. and Roy, B. V. Why is posterior sampling better than optimism for reinforcement learning, 2016.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 1466–1474. Curran Associates, Inc., 2014.

Osband, I. and Van Roy, B. Posterior sampling for reinforcement learning without episodes. *arXiv preprint arXiv:1608.02731*, 2016.

Osband, I., Roy, B. V., and Russo, D. (more) efficient reinforcement learning via posterior sampling. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pp. 3003–3011, USA, 2013. Curran Associates Inc.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4026–4034. Curran Associates, Inc., 2016a.

Osband, I., Van Roy, B., and Wen, Z. 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*, pp. 2377–2386, 2016b.

Ouyang, Y., Gagrani, M., and Jain, R. Learning-based control of unknown linear systems with thompson sampling. *CoRR*, abs/1709.04047, 2017. URL `http://arxiv.org/abs/1709.04047`.

Strens, M. J. A. A bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 943–950, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2.

Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.