

# Thompson Sampling Achieves $\tilde{O}(\sqrt{T})$ Regret in Linear Quadratic Control

Taylan Kargin<sup>\*,1</sup>

Sahin Lale<sup>\*,1</sup>

Kamyar Azizzadenesheli<sup>2</sup>

Anima Anandkumar<sup>1</sup>

Babak Hassibi<sup>1</sup>

*\*Equal contribution*

<sup>1</sup>California Institute of Technology

<sup>2</sup>Purdue University

TKARGIN@CALTECH.EDU

ALALE@CALTECH.EDU

KAMYAR@PURDUE.EDU

ANIMA@CALTECH.EDU

HASSIBI@CALTECH.EDU

**Editors:** Po-Ling Loh and Maxim Raginsky

## Abstract

Thompson Sampling (TS) is an efficient method for decision-making under uncertainty, where an action is sampled from a carefully prescribed distribution which is updated based on the observed data. In this work, we study the problem of adaptive control of stabilizable linear-quadratic regulators (LQRs) using TS, where the system dynamics are unknown. Previous works have established that  $\tilde{O}(\sqrt{T})$  frequentist regret is optimal for the adaptive control of LQRs. However, the existing methods either work only in restrictive settings, require a priori known stabilizing controllers, or utilize computationally intractable approaches. We propose an efficient TS algorithm for the adaptive control of LQRs, **TS-based Adaptive Control**, TSAC, that attains  $\tilde{O}(\sqrt{T})$  regret, even for multidimensional systems, thereby solving the open problem posed in [Abeille and Lazaric \(2018\)](#). TSAC does not require a priori known stabilizing controller and achieves fast stabilization of the underlying system by effectively exploring the environment in the early stages. Our result hinges on developing a novel lower bound on the probability that the TS provides an optimistic sample. By carefully prescribing an early exploration strategy and a policy update rule, we show that TS achieves order-optimal regret in adaptive control of multidimensional stabilizable LQRs. We empirically demonstrate the performance and the efficiency of TSAC in several adaptive control tasks.

**Keywords:** Thompson sampling, adaptive control, linear quadratic control, regret

## 1. Introduction

There has been a significant development in data-driven methods for controlling dynamical systems in recent years due to the development of novel reinforcement learning approaches and techniques ([Hou and Wang, 2013](#)). Adaptive control of unknown linear dynamical systems has been the main focus due to its simplicity and its ability to capture the crux of the problem and give insights on more challenging tasks ([Recht, 2019](#)). Among linear dynamical systems, Linear Quadratic Regulators (LQRs) are the canonical settings with quadratic regulatory costs to design desirable controllers and have been studied in an array of prior works ([Abbasi-Yadkori and Szepesvári, 2011](#); [Faradonbeh et al., 2017](#); [Abeille and Lazaric, 2018](#); [Mania et al., 2019](#); [Simchowitz and Foster, 2020](#); [Chen and Hazan, 2020](#); [Lale et al., 2022](#)). These works provide finite-time performance guarantees of adaptive control algorithms in terms of *regret*, which is the difference between the attained cumulative cost and the expected cost of the optimal controller. In particular, they show that  $\tilde{O}(\sqrt{T})$  regret after

$T$  time steps is optimal in adaptive control of LQRs. They utilize several different paradigms for algorithm design such as Certainty Equivalence, Optimism or Thompson Sampling, yet, they suffer either from the inherent algorithmic drawbacks or limited applicability in practice.

**Certainty equivalent control and its challenges:** Certainty equivalent control (CEC) is one of the most straightforward paradigms for control design in adaptive control of dynamical systems. In CEC, an agent obtains a nominal estimate of the system, and executes the optimal control law for this estimated system. Even though Mania et al. (2019); Simchowicz and Foster (2020) show that this simple approach attains optimal regret in LQRs, the proposed algorithms have several drawbacks. First and foremost, CEC is sensitive to model mismatch and requires significantly small model estimation error to a point that exploration of the system dynamics is not required. Since this level of refinement is challenging to obtain for an unknown system, these methods rely on access to an initial stabilizing controller to enable a long exploration. In practice, such a priori known controllers may not be available, which hinders the deployment of these algorithms.

**Optimism-based control and its challenges:** Optimism is one of the most prominent methods to effectively balance exploration and exploitation in adaptive control (Bittanti et al., 2006). In optimism-based control, an agent executes the optimal policy for the model with the lowest cost within a set of plausible models. In Abbasi-Yadkori and Szepesvári (2011); Faradonbeh et al. (2017), the authors use optimism-based control design to achieve  $\tilde{O}(\sqrt{T})$  regret with exponential dimension dependency. Both algorithms solve a non-convex optimization problem to find the optimistic controllers, which is an NP-hard problem in general (Agrawal, 2019). Unfortunately, this computational inefficiency severely limits their practicality. Recently, Abeille and Lazaric (2020) proposed a relaxation to the optimistic controller computation, which makes the optimism-based controllers efficient. However, their algorithm also requires a significantly well-refined model estimate and a given initial stabilizing policy, similar to CEC.

**Restricted LQR settings in the prior works:** In our work, we study the stabilizable multi-dimensional LQR setting. Stabilizability is necessary and sufficient condition to have a well-posed LQR control problem (Kailath et al., 2000). On the contrary, prior works usually consider the controllable LQR setting, which is a subclass of stabilizable LQRs (Cohen et al., 2019; Chen and Hazan, 2020). While the controllability condition simplifies the learning and control problem, it is also often violated in many real-world control systems (Friedland, 2012). Recently, Lale et al. (2022) proposed an adaptive control algorithm that does not need an initial stabilizing controller and achieves optimal regret in stabilizable LQRs. However, their method relies on optimism, and unfortunately inherits the aforementioned computational complexity of optimistic methods.

**Thompson Sampling and its challenges:** Thompson Sampling (TS) is one of the oldest strategies to balance the exploration vs. exploitation trade-off (Thompson, 1933). In TS, the agent samples a model from a distribution computed based on prior control input and observation pairs, and then takes the optimal action for this sampled model and updates the distribution based on its novel observation. Since it relies solely on sampling, this approach provides polynomial-time algorithms for adaptive control. Therefore, it is a promising alternative to overcome the computational burden faced in optimistic control design. For this reason, Abeille and Lazaric (2017, 2018) propose adaptive control algorithms using TS. In particular, Abeille and Lazaric (2018) provide the first TS-based adaptive control algorithm for LQRs that attains optimal regret of  $\tilde{O}(\sqrt{T})$ . However, their result *only holds for scalar* stabilizable systems, since they were able to show that TS samples optimistic parameters with constant probability in only scalar systems. Further, they conjecture that this is true

Work	Setting	Stabilizing Controller	Computation
(Abeille and Lazaric, 2018)	Stabilizable <sup>†</sup>	Not Required	Feasible
(Mania et al., 2019)	Controllable	Required	Feasible
(Simchowitz and Foster, 2020)	Stabilizable	Required	Feasible
(Chen and Hazan, 2020)	Controllable	Not required	Feasible
(Lale et al., 2022)	Stabilizable	Not required	Infeasible
<b>This work</b>	Stabilizable	Not required	Feasible

Table 1: Comparison with the prior works that attain  $\tilde{O}(\sqrt{T})$  regret on LQR, <sup>†</sup> = 1-dim LQRs

in multidimensional systems as well and TS-based adaptive control can provide optimal regret in multidimensional LQRs, and provide a simple numerical example to support their claims.

## Contributions

In this work, we give an affirmative answer to the conjecture posed in Abeille and Lazaric (2018):

- 1) We propose an efficient adaptive control algorithm, **Thompson Sampling-based Adaptive Control (TSAC)**, that attains  $\tilde{O}(\sqrt{T})$  regret in multidimensional stabilizable LQRs. This makes TSAC the first efficient adaptive control algorithm to achieve order-optimal regret in all stabilizable LQRs without the prior knowledge of a stabilizing policy (Table 1).
- 2) We empirically demonstrate the performance of TSAC and compare to the optimism (heuristic) and TS-based methods that do not require initial stabilizing policy in flight control of Boeing 747 with linearized dynamics. We show that TSAC effectively explores the system to find a stabilizing policy and achieves the competitive regret performance, while being computationally feasible.

The design of TSAC and our regret guarantee hinge on three important pieces missing in prior works: Fixed policy update rule, improved exploration in early stages of adaptive control, and a novel lower bound that shows TS samples optimistic parameters with non-zero probability in multidimensional LQRs. Unlike the frequent policy update rule of Abeille and Lazaric (2018) in scalar LQRs, TSAC updates its policy with fixed time periods. This policy update rule prevents fast policy changes that would cause state blow-ups in stabilizable LQRs. In the beginning of agent-environment interaction, TSAC focuses on quickly finding a stabilizing controller to avoid state blow-ups due to lack of a known initial stabilizing policy. By using isotropic exploration in the early stages along with the exploration of TS policy, we show that TSAC achieves fast stabilization.

After stabilizing the unknown system dynamics, TSAC relies on the effective exploration of the TS to find desirable controllers. In particular, we show that the TS samples optimistic parameters with a constant probability in any LQR setting. This novel lower bound shows that the TS is an efficient alternative to optimism in all adaptive control problems in LQRs. Combining this lower bound with the fixed policy update rule, we derive the optimal regret guarantee for TSAC.

## 2. Preliminaries

**Notation:** We denote the Euclidean norm of a vector  $x$  as  $\|x\|_2$ . For a matrix  $A \in \mathbb{R}^{n \times d}$ , we denote  $\rho(A)$  as the spectral radius of  $A$ ,  $\|A\|_F$  as its Frobenius norm and  $\|A\|$  as its spectral norm.  $\text{tr}(A)$  denotes its trace,  $A^\top$  is the transpose. For any positive definite matrix  $V$ ,  $\|A\|_V = \|V^{1/2}A\|_F$ . For matrices  $A, B \in \mathbb{R}^{n \times d}$ ,  $A \bullet B = \text{tr}(AB^\top)$  denotes their Frobenius inner product. The  $j$ -th singular value of a rank- $n$  matrix  $A$  is  $\sigma_j(A)$ , where  $\sigma_{\max}(A) := \sigma_1(A) \geq \dots \geq \sigma_{\min}(A) := \sigma_n(A)$ .  $I$

represents the identity matrix with the appropriate dimensions.  $\mathcal{M}_n = \mathbb{R}^{n \times n}$  denotes the set of  $n$ -dimensional square matrices.  $\mathcal{N}(\mu, \Sigma)$  denotes normal distribution with mean  $\mu$  and covariance  $\Sigma$ .  $Q(\cdot)$  denotes the Gaussian  $Q$ -function.  $O(\cdot)$  and  $o(\cdot)$  denote the standard asymptotic notation and  $f(T) = \omega(g(T))$  is equivalent to  $g(T) = o(f(T))$ .  $\tilde{O}(\cdot)$  presents the order up to logarithmic terms.

## 2.1. Setting

Suppose we are given a discrete time linear time-invariant system with the following dynamics,

$$x_{t+1} = A_*x_t + B_*u_t + w_t, \quad (1)$$

where  $x_t \in \mathbb{R}^n$  is the state of the system,  $u_t \in \mathbb{R}^d$  is the control input,  $w_t \in \mathbb{R}^n$  is i.i.d. process noise at time  $t$ . At each time step  $t$ , the system is at state  $x_t$  where the agent observes the state. Then, the agent applies a control input  $u_t$  and the system evolves to  $x_{t+1}$  at time  $t + 1$ . The underlying system (1) can be represented as  $x_{t+1} = \Theta_*^\top z_t + w_t$ , where  $\Theta_*^\top = [A_* \ B_*]$  and  $z_t = [x_t^\top \ u_t^\top]^\top$ . In this work, we consider stabilizable linear dynamical systems  $\Theta_*$ , such that there exists a controller  $K$  where  $\rho(A_* + B_*K) < 1$ . More precisely, the systems with the following property:

**Assumption 1 (Bounded and  $(\kappa, \gamma)$ -stabilizable System)** *The unknown system  $\Theta_*$  is a member of a set  $\mathcal{S}$  such that  $\mathcal{S} \subseteq \{\Theta' = [A' \ B'] \mid \Theta' \text{ is } (\kappa, \gamma)\text{-stabilizable, } \|\Theta'\|_F \leq S\}$  for some  $\kappa \geq 1$  and  $0 < \gamma \leq 1$ . In particular, for the underlying system  $\Theta_*$ , we have  $\|K(\Theta_*)\| \leq \kappa$  and there exists  $L$  and  $H \succ 0$  such that  $A_* + B_*K(\Theta_*) = HLH^{-1}$ , with  $\|L\| \leq 1 - \gamma$  and  $\|H\| \|H^{-1}\| \leq \kappa$ .*

Note that the stabilizability condition is necessary and sufficient condition to define the optimal control problem (Kailath et al., 2000) and it is weaker than the controllability assumption considered in prior works (Abbasi-Yadkori and Szepesvári, 2011; Cohen et al., 2019; Chen and Hazan, 2020). In particular, the set of stabilizable systems subsumes the set of controllable systems. Moreover,  $(\kappa, \gamma)$ -stabilizability is merely a quantification of stabilizability for the finite-time analysis and it is adopted in recent works (Cohen et al., 2018, 2019; Cassel et al., 2020). One can show that any stabilizable system is also  $(\kappa, \gamma)$ -stabilizable for some  $\kappa$  and  $\gamma$ , conversely,  $(\kappa, \gamma)$ -stabilizability implies stabilizability (Lemma B.1 Cohen et al. (2018)). We have the following assumption on  $w_t$ .

**Assumption 2 (Gaussian Process Noise)** *There exists a filtration  $\mathcal{F}_t$  such that for all  $t \geq 0$ ,  $x_t, z_t$  are  $\mathcal{F}_t$ -measurable and  $w_t | \mathcal{F}_t = \mathcal{N}(0, \sigma_w^2 I)$  for some known  $\sigma_w > 0$ .*

Note that this assumption is standard in literature and adopted for simplicity of exposure. The following results can be extended to sub-Gaussian process noise setting using the techniques developed in Lale et al. (2022). At each time step, the regulating cost is  $c_t = x_t^\top Q x_t + u_t^\top R u_t$ , where  $Q \in \mathbb{R}^{n \times n}$  and  $R \in \mathbb{R}^{d \times d}$  are known positive definite matrices such that  $\|Q\|, \|R\| < \bar{\alpha}$  and  $\sigma_{\min}(Q), \sigma_{\min}(R) > \underline{\alpha} > 0$ . The goal is to minimize the average expected cost

$$J(\Theta_*) = \lim_{T \rightarrow \infty} \min_{u=[u_1, \dots, u_T]} \frac{1}{T} \mathbb{E} \left[ \sum_{t=1}^T x_t^\top Q x_t + u_t^\top R u_t \right], \quad (2)$$

by designing control inputs based on past observations. This problem is the canonical infinite horizon linear quadratic regulator (LQR) problem. If the underlying system  $\Theta_*$  is known, the solution of the optimal control problem is a linear feedback control  $u_t = K(\Theta_*)x_t$  with  $K(\Theta_*) = -(R + B_*^\top P(\Theta_*)B_*)^{-1} B_*^\top P(\Theta_*)A_*$ , where  $P(\Theta_*)$  is the unique positive definite solution to

$$P(\Theta_*) = A_*^\top P(\Theta_*)A_* + Q - A_*^\top P(\Theta_*)B_*(R + B_*^\top P(\Theta_*)B_*)^{-1} B_*^\top P(\Theta_*)A_*, \quad (3)$$

*i.e.*, the discrete algebraic Riccati equation (DARE), and  $J(\Theta_*) = \sigma_w^2 \text{tr}(P(\Theta_*))$ . Note that since the system is stabilizable,  $J(\Theta_*) < \infty$ . In fact, using Assumption 1, one can show that  $\|P(\Theta')\| \leq D := \bar{\alpha}\gamma^{-1}\kappa^2(1 + \kappa^2)$  for all  $\Theta' \in \mathcal{S}$ , including  $\Theta_*$  (Lemma 2.1 of Lale et al. (2022)).

## 2.2. Finite-Time Adaptive Control Problem

In this work, we consider the adaptive control setting, where  $\Theta_*$  is *unknown*. The goal in the finite-time adaptive control problem is to minimize the cumulative cost *i.e.*,  $\sum_{t=0}^T c_t$ . In order to design a controller that achieves this goal, the controlling agent needs to interact with the system to learn the  $\Theta_*$  that governs the dynamics. However, due to a lack of knowledge of model dynamics, the agent takes sub-optimal actions. In this work, we use *regret*,  $R_T$ , as the metric to evaluate the finite-time performance of the controlling agent. The regret quantifies the difference between the performance of the agent and the expected performance of the optimal controller,  $R_T = \sum_{t=0}^T (c_t - J(\Theta_*))$ .

## 2.3. Learning the System Dynamics

For any given input and state pairs up to time  $t$ ,  $\Theta_*$  can be estimated using regularized least squares (RLS) for some  $\mu > 0$ :  $\min_{\Theta} \sum_{s=0}^{t-1} \text{tr}((x_{s+1} - \Theta^\top z_s)(x_{s+1} - \Theta^\top z_s)^\top) + \mu \|\Theta\|_F^2$ . The solution is given as  $\hat{\Theta}_t = V_t^{-1} \sum_{s=0}^{t-1} z_s x_{s+1}^\top$  where  $V_t = \mu I + \sum_{s=0}^{t-1} z_s z_s^\top$ . Using Theorem 1 of Abbasi-Yadkori and Szepesvári (2011), for any  $\delta \in (0, 1)$ , for all  $0 \leq t \leq T$ , the underlying parameter  $\Theta_*$  lives in  $\mathcal{E}_t^{\text{RLS}}(\delta)$  with probability at least  $1 - \delta$  where  $\mathcal{E}_t^{\text{RLS}}(\delta) = \{\Theta : \|\Theta - \hat{\Theta}_t\|_{V_t} \leq \beta_t(\delta)\}$  for  $\beta_t(\delta) = \sigma_w \sqrt{2n \log((\det(V_t)^{1/2})/(\delta \det(\mu I)^{1/2}))} + \sqrt{\mu}S$ .

## 3. TSAC Framework

In this section, we present TSAC, a sample efficient TS-based adaptive control algorithm for the unknown stabilizable LQRs. The algorithm is summarized in Algorithm 1. It has two phases: 1) TS with improved exploration and 2) Stabilizing TS.

**TS with Improved Exploration** Due to lack of a priori known stabilizing controller, TSAC focuses on rapidly learning stabilizing controllers in the early stages of the algorithm. To achieve this, TSAC explores the system dynamics effectively in this phase. At any time-step  $t$ , given the RLS estimate  $\hat{\Theta}_t$  and the design matrix  $V_t$  as described in Section 2.3, TSAC samples a perturbed model parameter  $\tilde{\Theta}_t = \mathcal{R}_{\mathcal{S}}(\hat{\Theta}_t + \beta_t(\delta)V_t^{-1/2}\eta_t)$ , where  $\mathcal{R}_{\mathcal{S}}$  denotes the rejection sampling operator associated with the set  $\mathcal{S}$  given in Assumption 1 and  $\eta_t \in \mathbb{R}^{(n+d) \times n}$  is a matrix with independent standard normal entries. Here  $\mathcal{R}_{\mathcal{S}}$  guarantees that  $\tilde{\Theta}_t \in \mathcal{S}$  and  $\beta_t(\delta)V_t^{-1/2}\eta_t$  randomizes the sampled parameter coherently with the RLS estimate and the uncertainty associated with it. Using this sampled model parameter, TSAC constructs the optimal linear controller  $\bar{u}_t = K(\tilde{\Theta}_t)x_t$  for  $\tilde{\Theta}_t$ .

However, to obtain stabilizing controllers for an unknown linear dynamical system, one needs to explore the state-space in all directions (Lemma 4.2 of (Lale et al., 2022)). Unfortunately, due to lack of reliable estimates in the early stages, deploying the policy achieved via TS,  $\bar{u}_t$ , may not achieve such effective exploration. Therefore, in the early stages of interactions with the underlying system, TSAC deploys isotropic perturbations along with the sampled policy. In particular, for the first  $T_w$  time-steps, TSAC uses  $u_t = \bar{u}_t + \nu_t$  as the control input where  $\nu_t \sim \mathcal{N}(0, 2\kappa^2\sigma_w^2 I)$ . This improved exploration policy effectively excites and explores all dimensions of the system to certify the design of stabilizing controllers. TSAC sets  $T_w$  such that all the sampled controllers  $K(\tilde{\Theta}_t)$  are guaranteed to stabilize the underlying system  $\Theta_*$  for all  $t > T_w$  (Appendix B).

---

**Algorithm 1** TSAC
 

---

```

1: Input:  $\kappa, \gamma, Q, R, \sigma_w^2, V_0 = \mu I, \hat{\Theta}_0 = 0$ 
2: for  $i = 0, 1, \dots$  do
3:   Estimate  $\hat{\Theta}_i$  & Sample  $\tilde{\Theta}_i = \mathcal{R}_S(\hat{\Theta}_i + \beta_t V_t^{-1/2} \eta_t)$ 
4:   for  $t = i\tau_0, \dots, (i+1)\tau_0 - 1$  do
5:     if  $t \leq T_w$  then
6:       Deploy  $u_t = K(\tilde{\Theta}_i)x_t + \nu_t$  ▷ TS with Improved Exploration
7:     else
8:       Deploy  $u_t = K(\tilde{\Theta}_i)x_t$  ▷ Stabilizing TS
    
```

---

Unlike most of the popular RL strategies that follow lazy updates, TSAC updates its sampled policy in every fixed  $\tau_0$  steps, *i.e.*, the same sampled policy  $K(\tilde{\Theta}_t)$  is deployed for  $\tau_0$  time-steps. This update rule is carefully chosen such that TSAC samples enough optimistic policies to reduce the cumulative regret and avoids too frequent policy changes which would cause state blow-ups.

**Stabilizing TS** After guaranteeing the design of stabilizing policies with improved exploration in the first phase, TSAC starts the adaptive control with only TS. In particular, for the remaining time-steps, TSAC deploys  $u_t = K(\tilde{\Theta}_t)x_t$  for  $\tilde{\Theta}_t = \mathcal{R}_S(\hat{\Theta}_t + \beta_t(\delta)V_t^{-1/2}\eta_t)$  and updates the sampled model parameter in every  $\tau_0$  time-steps. Note that, even though all the policies during this phase are stabilizing, frequent policy changes can still cause undesirable state growth. TSAC prevents this possibility by applying the same control policy for  $\tau_0$  time-steps in this phase as well. During this phase, TSAC decays the possible state blow-ups in the first phase and maintains stable dynamics.

#### 4. Theoretical Analysis

In this section, we study the theoretical guarantees of TSAC. The following states the first order-optimal frequentist regret bound for TS in multidimensional stabilizable LQRs, our main result.

**Theorem 3 (Regret of TSAC)** *Suppose Assumptions 1 and 2 hold and set  $\tau_0 = 2\gamma^{-1} \log(2\kappa\sqrt{2})$  and  $T_0 = \text{poly}(\log(1/\delta), \sigma_w^{-1}, n, d, \bar{\alpha}, \gamma^{-1}, \kappa)$ . Then, for long enough  $T$ , TSAC achieves the regret  $R_T = \tilde{O}\left((n+d)^{(n+d)}\sqrt{T \log(1/\delta)}\right)$  w.p. at least  $1-10\delta$ , if  $T_w = \max\left(T_0, c_1(\sqrt{T} \log T)^{1+o(1)}\right)$  for a constant  $c_1 > 0$ . Furthermore, if the closed loop matrix of the optimally controlled underlying system,  $A_{c,*} := A_* + B_*K_*$ , is non-singular, w.p. at least  $1-10\delta$ , TSAC achieves the regret  $R_T = \tilde{O}\left(\text{poly}(n, d)\sqrt{T \log(1/\delta)}\right)$  if  $T_w = \max\left(T_0, c_2(\log T)^{1+o(1)}\right)$  for a constant  $c_2 > 0$ .*

This makes TSAC the *first efficient* adaptive control algorithm that achieves optimal regret in adaptive control of all LQRs *without an initial stabilizing policy*. To prove this result, we follow similar approach as the existing methods in literature, and define the high probability joint event  $E_t = \hat{E}_t \cap \tilde{E}_t \cap \bar{E}_t$ , where  $\hat{E}_t$  states that the RLS estimate  $\hat{\Theta}$  concentrates around  $\Theta_*$ ,  $\tilde{E}_t$  states that the sampled parameter  $\tilde{\Theta}$  concentrates around  $\hat{\Theta}$ , and  $\bar{E}_t$  states that the state remains bounded respectively (Appendix C). Conditioned on this event, we decompose the frequentist regret as,  $R_T \mathbb{1}_{E_T} \leq R_{T_w}^{\text{exp}} + R_T^{\text{RLS}} + R_T^{\text{mart}} + R_T^{\text{TS}} + R_T^{\text{gap}}$ , where  $R_{T_w}^{\text{exp}}$  accounts for the regret attained due to improved exploration,  $R_T^{\text{RLS}}$  represents the difference between the value function of the true next state and the predicted next state,  $R_T^{\text{mart}}$  is a martingale with bounded difference,  $R_T^{\text{TS}}$  measures



the difference in optimal average expected cost between the true model  $\Theta_*$  and the sampled model  $\tilde{\Theta}$ , and  $R_T^{\text{gap}}$  measures the regret due to policy changes. The decomposition and expressions are given in Appendix E. In the analysis, we bound each term separately (Appendix F). Before discussing the details of the analysis, we first consider the prior works that use TS for adaptive control of LQRs and discuss their shortcomings. Further, we highlight the challenges in adaptive control of multidimensional stabilizable LQRs using TS and present our approaches to overcome these.

#### 4.1. Prior Work on TS-based Adaptive Control and Challenges

For the frequentist regret minimization problem given in Section 2.2, the state-of-the-art adaptive control algorithm that uses TS is [Abeille and Lazaric \(2018\)](#). They consider the ‘‘contractible’’ LQR systems, i.e.  $|A_* + B_*K(\Theta_*)| < 1$ , and provide  $\tilde{O}(\sqrt{T})$  regret upper bound for scalar LQRs, i.e.  $n = d = 1$ . Notice that the set of contractible systems is a small subset of the set  $\mathcal{S}$  defined in Assumption 1 and they are only equivalent for scalar systems since  $\rho(A_* - B_*K(\Theta_*)) = |A_* - B_*K(\Theta_*)|$ . This simplified setting allow them to reduce the regret analysis into the trade-off between  $R_T^{\text{TS}} = \sum_{t=0}^T \{J(\tilde{\Theta}_t) - J(\Theta_*)\}$  and  $R_T^{\text{gap}} = \sum_{t=0}^T \mathbb{E}[x_{t+1}^\top (P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t)) x_{t+1} \mid \mathcal{F}_t]$ .

These regret terms are central in the analysis of several adaptive control algorithms. In the certainty equivalent control approaches,  $R_T^{\text{TS}}$  is bounded by the quadratic scaling of model estimation error after a significantly long exploration with a known stabilizing controller ([Mania et al., 2019](#); [Simchowitz and Foster, 2020](#)). In the optimism-based algorithms,  $R_T^{\text{TS}}$  is bounded by 0 by design ([Abbasi-Yadkori and Szepesvári, 2011](#); [Faradonbeh et al., 2017](#)). Similarly, in Bayesian regret setting, ([Ouyang et al., 2017](#)) assume that the underlying parameter  $\Theta_*$  comes from a known prior that the expected regret is computed with respect to. This true prior yields  $\mathbb{E}[R_T^{\text{TS}}] = 0$  in certain restrictive LQRs. The conventional approach in the analysis of  $R_T^{\text{gap}}$  is to have lazy policy updates, i.e.,  $O(\log T)$  policy changes, via doubling the determinant of  $V_t$  ([Abeille and Lazaric, 2017](#); [Lale et al., 2022](#)) or exponentially increasing epoch durations ([Faradonbeh et al., 2020](#); [Cassel et al., 2020](#)).

On the other hand, [Abeille and Lazaric \(2018\)](#) bound  $R_T^{\text{TS}}$  by showing that TS samples the optimistic parameters,  $\tilde{\Theta}_t$  such that  $J(\tilde{\Theta}_t) \leq J(\Theta_*)$ , with a constant probability, which reduces the regret of non-optimistic steps. Unlike the conventional policy update approaches, the key idea in [Abeille and Lazaric \(2018\)](#) is to update the control policy every time-steps via TS, which increases the amount of optimistic policies during the execution. They show that while this frequent update rule reduces  $R_T^{\text{TS}}$ , it only results with  $R_T^{\text{gap}} = \tilde{O}(\sqrt{T})$ . However, they were only able to show that this constant probability of optimistic sampling holds for scalar LQRs.

The difficulty of the analysis for the probability of optimistic parameter sampling lies in the challenging characterization of the optimistic set. Since  $J(\tilde{\Theta}) = \sigma_w^2 \text{tr}(P(\tilde{\Theta}))$ , one needs to consider the spectrum of  $P(\tilde{\Theta})$  to define optimistic models, which makes the analysis difficult. In particular, decreasing the cost along one direction may be result in an increase in other directions. However, for the scalar LQR setting considered in [Abeille and Lazaric \(2018\)](#),  $J(\tilde{\Theta}) = P(\tilde{\Theta})$  and using standard perturbation results on DARE suffices. As mentioned in [Abeille and Lazaric \(2018\)](#), one can naively consider the surrogate set of being optimistic in all directions, i.e.  $P(\tilde{\Theta}) \preceq P(\Theta_*)$ . Nevertheless, this would result in probability that decays linear in time and does not yield sub-linear regret. In this work, we propose new surrogate sets to derive a lower bound on the probability of having optimistic samples, and show that TS in fact samples optimistic model parameters with constant probability.

In designing TS-based adaptive control algorithms for multidimensional stabilizable LQRs, one needs to maintain bounded state. In bounding the state, [Abeille and Lazaric \(2018\)](#) rely on the fact that the underlying system is contractive,  $\|\tilde{A} + \tilde{B}K(\tilde{\Theta})\| < 1$ . However, under Assumption

1, even if the optimal policy of the underlying system is chosen by the learning agent, the closed-loop system may not be contractive since for any symmetric matrix  $M$ ,  $\rho(M) \leq \|M\|$ . Thus, to avoid dire consequences of unstable dynamics, TS-based adaptive control algorithms should focus on finite-time stabilization of the system dynamics in the early stages.

Moreover, the lack of contractive closed-loop mappings in stabilizable LQRs, prevent frequent policy changes used in Abeille and Lazaric (2018). From the definition of  $(\kappa, \gamma)$ -stabilizability (Assumption 1), for any stabilizing controller  $K'$ , we have that  $A_* + B_*K' = H' LH'^{-1}$ , with  $\|L\| < 1$  for some similarity transformation  $H'$ . Thus, even if all the policies are stabilizing, changing the policies at every time step could cause couplings of these similarity transformations and result in linear growth of state over time. Thus, TS-based adaptive control algorithms need to find the balance in rate of policy updates, so that frequent policy switches are avoided, yet, enough optimistic policies are sampled. In light of these observations, our results hinge on the following:

- 1) Improved exploration of TSAC, which allows fast stabilization of the system dynamics,
- 2) Fixed policy update rule of TSAC, which prevents state blow-up and reduces  $R_T^{\text{gap}}$  and  $R_T^{\text{TS}}$ ,
- 3) A novel result that shows TS samples optimistic model parameters with a constant probability for multidimensional LQRs and gives a novel bound on  $R_T^{\text{TS}}$ .

#### 4.2. Details of the analysis

The improved exploration along with TS in the early stages allows TSAC to effectively explore the state-space in all directions. The following shows that for a long enough improved exploration phase, TSAC achieves consistent model estimates and guarantees the design of stabilizing policies.

**Lemma 4 (Model Estimation Error and Stabilizing Policy Design)** *Suppose Assumptions 1 and 2 hold. For  $t \geq 200(n+d) \log \frac{12}{\delta}$  time-steps of TS with improved exploration, with probability at least  $1 - 2\delta$ , TSAC obtains model estimates such that  $\|\hat{\Theta}_t - \Theta_*\|_2 \leq 7\beta_t(\delta)/(\sigma_w \sqrt{t})$ . Moreover, after  $T_w \geq T_0 := \text{poly}(\log(1/\delta), \sigma_w^{-1}, n, d, \bar{\alpha}, \gamma^{-1}, \kappa)$  length TS with improved exploration phase, with probability at least  $1 - 3\delta$ , TSAC samples controllers  $K(\tilde{\Theta}_t)$  such that the closed-loop dynamics on  $\Theta_*$  is  $(\kappa\sqrt{2}, \gamma/2)$  strongly stable for all  $t > T_w$ , i.e. there exists  $L$  and  $H \succ 0$  such that  $A_* + B_*K(\tilde{\Theta}_t) = H L H^{-1}$ , with  $\|L\| \leq 1 - \gamma/2$  and  $\|H\| \|H^{-1}\| \leq \kappa\sqrt{2}$ .*

The proof and the precise expression of  $T_w$  can be collected in Appendix B. In the proof, we show that the inputs  $u_t = K(\tilde{\Theta}_t)x_t + \nu_t$  for  $\nu_t \sim \mathcal{N}(0, 2\kappa^2\sigma_w^2 I)$  guarantees persistence of excitation with high probability, i.e., the smallest eigenvalue of the design matrix  $V_t$  scales linearly over time. Combining this result, with the confidence set construction given in Section 2.3, we derive the first result. Using the first result and the fact that there exists a stabilizing neighborhood around the model parameter  $\Theta_*$ , such that all the optimal linear controllers of the models within this region stabilize  $\Theta_*$ , we derive the final result. Due to early improved exploration, TSAC stabilizes the system dynamics after  $T_w$  samples and starts stabilizing adaptive control with only TS. Using the stabilizing controllers for fixed  $\tau_0 = 2\gamma^{-1} \log(2\kappa\sqrt{2})$  time-steps, TSAC decays the state magnitude and remedy possible state blow-ups in the first phase. To study the boundedness of state, define  $T_r = T_w + (n+d)\tau_0 \log(n+d)$ . The following shows that the state is bounded and well-controlled.

**Lemma 5 (Bounded states)** *Suppose Assumptions 1 & 2 hold. For given  $T_w$  and  $T_r$ , TSAC controls the state such that  $\|x_t\| = O((n+d)^{n+d})$  for  $t \leq T_r$ , with probability at least  $1 - 3\delta$  and  $\|x_t\| \leq (12\kappa^2 + 2\kappa\sqrt{2})\gamma^{-1}\sigma_w\sqrt{2n \log(n(t-T_w)/\delta)}$  for  $T \geq t > T_r$ , with probability at least  $1 - 4\delta$ .*



The proof is given in Appendix C, but here we provide a proof sketch. To bound the state for  $t \leq T_r$ , we show that deploying the same policy for  $\tau_0$  time-steps in the first phase maintains a well-controlled state except  $n + d$  time-steps, under the high probability event of  $\tilde{E}_t \cap \bar{E}_t$ . Moreover, we show that this slow policy change prevents further state blow-ups due to non-contractive system dynamics in stabilizable systems. To bound the state for  $t > T_r$ , we show that, with the given choice of  $\tau_0$ , all the controllers during the stabilizing TS phase halves the magnitude of the state at the end of their control period. Thus, we prove that after  $(n + d) \log(n + d)$  policy updates the state is well-controlled and brought to an equilibrium as shown in Lemma 5. This result shows that the joint event  $E_t = \hat{E}_t \cap \tilde{E}_t \cap \bar{E}_t$  holds with probability at least  $1 - 4\delta$  for all  $t \leq T$ .

Conditioned on this event, we individually analyze the regret terms individually (Appendix F). We show that with probability at least  $1 - \delta$ ,  $R_{T_w}^{\text{exp}}$  yields  $\tilde{O}((n + d)^{n+d} T_w)$  regret due to isotropic perturbations.  $R_T^{\text{RLS}}$  and  $R_T^{\text{mart}}$  are  $\tilde{O}((n + d)^{n+d} \sqrt{T_r} + \text{poly}(n, d) \sqrt{T - T_r})$  with probability at least  $1 - \delta$  due to standard arguments based on the event  $E_T$ . More importantly, conditioned on the event  $E_T$ , we prove that  $R_T^{\text{gap}} = \tilde{O}((n + d)^{n+d} \sqrt{T_r} + \text{poly}(n, d) \sqrt{T - T_r})$  with probability at least  $1 - 2\delta$ , and  $R_T^{\text{TS}} = \tilde{O}(n T_w + \text{poly}(n, d) \sqrt{T - T_w})$  with probability at least  $1 - 2\delta$ , whose analyses require several novel fundamental results.

To bound on  $R_T^{\text{gap}}$ , we extend the results in Abeille and Lazaric (2018) to multidimensional stabilizable LQRs and incorporate the slow update rule and the early improved exploration. We show that while TSAC enjoys well-controlled state with polynomial dimension dependency on regret due to slow policy updates, it also maintains the desirable  $\tilde{O}(\sqrt{T})$  regret of frequent updates with only a constant  $\tau_0$  scaling. As discussed in Section 4.1, bounding  $R_T^{\text{TS}}$  requires selecting optimistic models with constant probability, which has been an open problem in the literature for multidimensional systems. In this work, we provide a solution to this problem and show that TS indeed selects optimistic model parameters with a constant probability for multidimensional LQRs. The precise statement of this result and its proof outline are given in Section 5. Leveraging this result, we derive the upper bound on  $R_T^{\text{TS}}$ . Combining all these terms yields the regret upper bound of TSAC given in Theorem 3.

## 5. Proof Outline of Sampling Optimistic Models with Constant Probability

In this section, we provide the precise statement that the probability of sampling an optimistic parameter is lower bounded by a fixed constant with high probability. Then we give the proof outline with the main steps. The complete proof with the intermediate results are given in Appendix D.

**Theorem 6 (Optimistic probability)** *Let  $\mathcal{F}_t^{\text{cnt}} := \sigma(F_{t-1}, x_t)$  be the information available to the controller up to time  $t$ . Denote the optimistic set by  $\mathcal{S}^{\text{opt}} := \{\Theta \in \mathbb{R}^{(n+d) \times n} \mid J(\Theta) \leq J(\Theta_*)\}$ . If  $T_w = cn^2(\sqrt{T} \log T)^{1+o(1)}$  for a constant  $c > 0$ , then under the event  $E_T$  for large enough  $T$ , we have that  $p_t^{\text{opt}} := \mathbb{P}\left\{\tilde{\Theta}_t \in \mathcal{S}^{\text{opt}} \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\right\} \geq \frac{Q(1)}{1+o(1)}$  for any  $T_r < t \leq T$ . Furthermore, if the closed-loop matrix,  $A_{c,*} = A_* + B_* K_*$ , is non-singular, then the bound above still holds when  $T_w = c(\log T)^{1+o(1)}$  for a constant  $c > 0$ .*

### 5.1. Surrogate Set Definition

First, we define a surrogate subset  $\mathcal{S}^{\text{sur}}$  to the optimistic set  $\mathcal{S}^{\text{opt}}$ . The construction of  $\mathcal{S}^{\text{sur}}$  is important as the geometry of  $\mathcal{S}^{\text{opt}}$  is complicated to study due to (3) that controls the spectrum of  $P(\Theta)$ .

**Lemma 7 (Surrogate set)** *Let  $J(\Theta, K) := \text{tr}((Q + K^\top R K)\Sigma(\Theta, K))$  be the expected average cost of controlling a system  $\Theta \in \mathcal{S}$  by a fixed stabilizing control policy  $K \in \mathbb{R}^{d \times n}$  where  $\Sigma(\Theta, K) := \lim_{t \rightarrow \infty} \mathbb{E}[x_t x_t^\top]$  is the covariance of the state. The following surrogate set is a subset of  $\mathcal{S}^{\text{opt}}$ :*

$$\mathcal{S}^{\text{surr}} := \left\{ \Theta = (A, B)^\top \in \mathbb{R}^{(n+d) \times n} \mid J(\Theta, K(\Theta_*)) \leq J(\Theta_*, K(\Theta_*)) = J(\Theta_*) \right\} \subset \mathcal{S}^{\text{opt}}. \quad (4)$$

Note that  $\Sigma(\Theta, K)$  satisfies the Lyapunov equation  $\Sigma(\Theta, K) - \Theta^\top H_K \Sigma(\Theta, K) H_K^\top \Theta = \sigma_w^2 I$ , where  $H_K^\top := [I, K^\top]$ , and  $\Theta^\top H_K = A + BK$ , given that  $K$  stabilizes the system  $\Theta$ . We can analytically express  $\Sigma(\Theta, K)$  as a converging infinite sum  $\Sigma(\Theta, K) = \sigma_w^2 \sum_{t=0}^{\infty} (A+BK)^t (A^\top + K^\top B^\top)^t$  (Kailath et al., 2000). Using the properties of the trace operator, one can write  $J(\Theta, K(\Theta_*)) = L(\Theta^\top H_*)$ , where  $L(A_c) := \sigma_w^2 \sum_{t=0}^{\infty} \|A_c^t\|_{Q_*}^2$  for any stable matrix  $A_c$ ,  $Q_* := Q + K(\Theta_*)^\top R K(\Theta_*)$ , and  $H_*^\top := [I, K(\Theta_*)^\top]$ . Therefore, we can lower bound the probability of being optimistic as

$$\begin{aligned} p_t^{\text{opt}} &\geq \mathbb{P} \left\{ \tilde{\Theta}_t \in \mathcal{S}^{\text{surr}} \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t \right\} = \mathbb{P} \left\{ L(\tilde{\Theta}_t^\top H_*) \leq L(\Theta_*^\top H_*) \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t \right\} \\ &\geq \min_{\tilde{\Theta}_t \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t \{ L(\hat{\Theta}_t^\top H_* + \eta^\top \beta_t V_t^{-\frac{1}{2}} H_*) \leq L(\Theta_*^\top H_*) \} \end{aligned} \quad (5)$$

$$= \min_{\tilde{\Theta}_t \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t \{ L(\hat{\Theta}_t^\top H_* + \Xi \sqrt{F_t}) \leq L(\Theta_*^\top H_*) \} \quad (6)$$

where  $\mathbb{P}_t \{ \cdot \} := \mathbb{P} \{ \cdot \mid \mathcal{F}_t^{\text{cnt}} \}$ ,  $F_t := \beta_t^2 H_*^\top V_t^{-1} H_*$  and  $\Xi$  is a matrix of size  $n \times n$  with iid  $\mathcal{N}(0, 1)$  entries. Here (5) considers the worst possible estimate within  $\mathcal{E}_t^{\text{RLS}}$  and (6) is the whitening transformation.

## 5.2. Reformulation in Terms of Closed-Loop Matrix

In the second step, we reformulate the probability of sampling optimistic parameters in terms of closed-loop system matrix  $\tilde{A}_c := \tilde{\Theta}^\top H_* = \tilde{A} + \tilde{B}K(\Theta_*)$  of the sampled system  $\tilde{\Theta} = (\tilde{A}, \tilde{B})^\top$  driven by the policy  $K(\Theta_*)$ . Transitioning to the closed-loop formulation allows tighter bounds on the optimistic probability. To complete this reformulation, we need to construct an estimation confidence set for the closed-loop system matrix  $\hat{A}_c := \hat{\Theta}^\top H_* = \hat{A} + \hat{B}K(\Theta_*)$  of the RLS-estimated system  $\hat{\Theta} = (\hat{A}, \hat{B})^\top$  and show that the constructed confidence set is a super set to  $\mathcal{E}_t^{\text{RLS}}$ .

**Lemma 8 (Closed-loop confidence)** *Let  $F_t(\delta) := \beta_t^2(\delta) H_*^\top V_t^{-1} H_*$ . For any  $t \geq 0$ , define by*

$$\mathcal{E}_t^{\text{cl}}(\delta) := \left\{ \hat{\Theta} \in \mathbb{R}^{(n+d) \times n} \mid \text{tr} \left[ (\hat{\Theta}^\top H_* - \Theta_*^\top H_*) F_t^{-1}(\delta) (\hat{\Theta}^\top H_* - \Theta_*^\top H_*)^\top \right] \leq 1 \right\}. \quad (7)$$

*the closed-loop confidence set. Then, for all times  $t \geq 0$  and  $\delta \in (0, 1)$ , we have that  $\mathcal{E}_t^{\text{RLS}}(\delta) \subseteq \mathcal{E}_t^{\text{cl}}(\delta)$ .*

Note that the definition of  $\mathcal{E}_t^{\text{cl}}(\delta)$  only involves closed-loop matrices  $\hat{A}_c := \hat{\Theta}^\top H_*$  and  $A_{c,*} := \Theta_*^\top H_*$ . We can use the result of Lemma 8 to reformulate the probability of sampling optimistic parameters,  $\tilde{\Theta} = (\tilde{A}, \tilde{B})$ , as sampling optimistic closed-loop system matrices,  $\tilde{A}_c$ . We bound  $p_t^{\text{opt}}$  from below as

$$p_t^{\text{opt}} \geq \min_{\tilde{\Theta}_t \in \mathcal{E}_t^{\text{cl}}} \mathbb{P}_t \{ L(\hat{\Theta}_t^\top H_* + \Xi \sqrt{F_t}) \leq L(A_{c,*}) \} \quad (8)$$

$$= \min_{\hat{A}_c : \|\hat{A}_c - A_{c,*}\|_{F_t^{-1}} \leq 1} \mathbb{P}_t \{ L(\hat{A}_c + \Xi \sqrt{F_t}) \leq L(A_{c,*}) \} \quad (9)$$

$$= \min_{\hat{Y} : \|\hat{Y}\|_F \leq 1} \mathbb{P}_t \{ L(A_{c,*} + \hat{Y} \sqrt{F_t} + \Xi \sqrt{F_t}) \leq L(A_{c,*}) \}, \quad (10)$$

where (8) is due to Lemma 8 and (9) follows from the fact that  $H_*$  has full column rank. Observe that, in equation (10),  $\hat{\Upsilon}$  is a unit Frobenius norm matrix of size  $n \times n$  and the term  $A_{c,*} + \hat{\Upsilon}\sqrt{F_t}$  accounts for the confidence ellipsoid for the estimated closed-loop matrix,  $\hat{A}_c$ . The event in (10) corresponds to finding the closed-loop matrix,  $A_{c,*} + (\Xi + \hat{\Upsilon})\sqrt{F_t}$  of the TS sampled system in the sublevel manifold  $\mathcal{M}_* := \{A_c \in \mathcal{M}_n \mid L(A_c) \leq L(A_{c,*})\}$  as illustrated in Figure 1.

### 5.3. Local Geometry of Optimistic Set under Perturbations

Next, we further simplify the form of the probability in (10) by exploiting the local geometric structure of the function  $L : A_c \mapsto \sigma_w^2 \sum_{t=0}^{\infty} \|A_c^t\|_{Q_*}^2$  defined over the set of (Schur-)stable matrices,  $\mathcal{M}_{\text{Schur}} := \{A_c \in \mathcal{M}_n \mid \rho(A_c) < 1\}$ . The following lemma characterizes perturbative properties of  $L$ .

**Lemma 9 (Perturbations)** *The function  $L : \mathcal{M}_{\text{Schur}} \rightarrow \mathbb{R}_+$  defined as  $L(A_c) = \sigma_w^2 \sum_{t=0}^{\infty} \|A_c^t\|_{Q_*}^2$  is smooth in its domain. For any  $A_c \in \mathcal{M}_{\text{Schur}}$ , there exists  $\epsilon > 0$  such that for any perturbation  $\|G\|_F \leq \epsilon$ , the function  $L$  admits a quadratic Taylor expansion as*

$$L(A_c + G) = L(A_c) + \nabla L(A_c) \bullet G + \frac{1}{2}G \bullet \mathcal{H}_{A_c+sG}(G) \quad (11)$$

for an  $s \in [0, 1]$  where  $\mathcal{H}_{A_c} : \mathcal{M}_n \rightarrow \mathcal{M}_n$  is the Hessian operator evaluated at a point  $A_c \in \mathcal{M}_{\text{Schur}}$ . In particular, we have that  $\nabla L(A_{c,*}) = 2P(\Theta_*)A_{c,*}\Sigma_*$ . Furthermore, there exists a constant  $r > 0$  such that  $|G \bullet \mathcal{H}_{A_c+sG}(G)| \leq r\|G\|_F^2$  for any  $s \in [0, 1]$  and  $\|G\|_F \leq \epsilon$ .

Lemma 9 guarantees that if a perturbation is sufficiently small, the perturbed function can be locally expressed as a quadratic function of the perturbation. Since the set of stable matrices,  $\mathcal{M}_{\text{Schur}}$ , is globally non-convex and Taylor's theorem only holds in convex domains, we restrict the perturbations in a ball of radius  $\epsilon > 0$ . The fact that there is a neighborhood of stable matrices around a matrix  $A_c$  enables us to apply Taylor's theorem in this neighborhood.

Given the optimal closed-loop system matrix  $A_{c,*}$ , let  $\epsilon_* > 0$  be chosen such that the expansion in (11) holds for perturbations  $\|G\|_F \leq \epsilon_*$  around  $A_{c,*}$ . Denote the perturbation due to Thompson sampling and estimation error as  $G_t = (\Xi + \hat{\Upsilon})\sqrt{F_t}$  and let  $\|G_t\|_F \leq \epsilon_*$ . Then, we can write

$$\begin{aligned} L(A_{c,*} + G_t) &= L(A_{c,*}) + \nabla L(A_{c,*}) \bullet G_t + \frac{1}{2}G_t \bullet \mathcal{H}_{A_{c,*}+sG_t}(G_t) \\ &\leq L(A_{c,*}) + \nabla L(A_{c,*}) \bullet G_t + \frac{r_*}{2}\|G_t\|_F^2 \end{aligned} \quad (12)$$

where  $r_* > 0$  is a constant due to Lemma 9. Using (12), we have the following lower bound on (10),

$$p_t^{\text{opt}} \geq \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \frac{r_*}{2} \|(\Xi + \hat{\Upsilon})F_t^{\frac{1}{2}}\|_F^2 + \nabla L_* \bullet (\Xi + \hat{\Upsilon})F_t^{\frac{1}{2}} \leq 0, \text{ and } \|(\Xi + \hat{\Upsilon})F_t^{\frac{1}{2}}\|_F \leq \epsilon_* \right\}, \quad (13)$$

where  $\nabla L_* := \nabla L(A_{c,*})$ . The event in (13) corresponds to finding  $A_{c,*} + (\Xi + \hat{\Upsilon})\sqrt{F_t}$  at the intersection of the stable ball  $\mathcal{B}_* := \{A_c \in \mathcal{M}_n \mid \|A_c - A_{c,*}\|_F \leq \epsilon_*\}$  and the sublevel manifold  $\mathcal{M}_*^{\text{qd}} := \{A_c \in \mathcal{M}_n \mid \|A_c - A_{c,*} + r_*^{-1}\nabla L_*\|_F \leq \|r_*^{-1}\nabla L_*\|_F\}$  as illustrated in Figure 1.

The intersection  $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_* \subset \mathcal{M}_*$  serves as another surrogate to sublevel manifold  $\mathcal{M}_*$ . Switching to the new surrogate  $\mathcal{M}_*^{\text{qd}}$  helps us overcome the issue of working with intractable and complicated geometry of  $\mathcal{M}_*$  due to infinite sum in  $L(A_c)$ . We can utilize techniques relating to Gaussian probabilities as the geometry of  $\mathcal{M}_*^{\text{qd}}$  is described by a quadratic form.

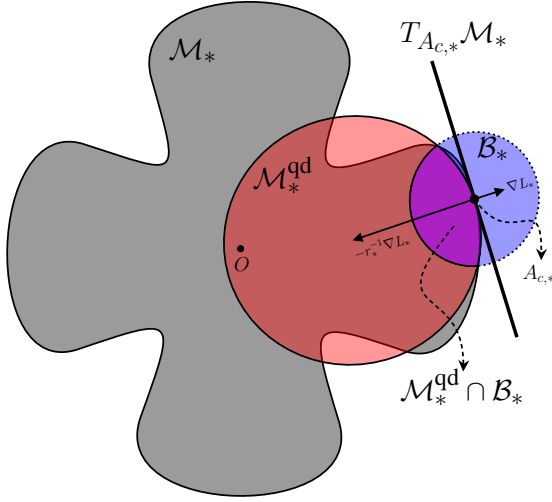


Figure 1: A visual representation of sublevel manifold  $\mathcal{M}_*$ .  $O$  is the origin and  $A_{c,*}$  is the optimal closed-loop system matrix.  $T_{A_{c,*}}\mathcal{M}_*$  is the tangent space to the manifold  $\mathcal{M}_*$  at the point  $A_{c,*}$  and  $\nabla L_*$  is the Jacobian of the function  $L$  at  $A_{c,*}$ .  $\mathcal{M}_*^{\text{qd}}$  is the sublevel manifold of the quadratic approximation to  $L$  and  $\mathcal{B}_*$  is a small ball of stable matrices around  $A_{c,*}$ . The intersection  $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_*$  is a subset of  $\mathcal{M}_*$ .

#### 5.4. Final Bound

Equipped with the preceding results, we can bound the optimism probability tractably from below by the probability of a TS sampled closed-loop system matrix lying inside the intersection of two balls  $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_*$  as given in (13). By bounding the weighted Frobenius norms in (13) from above by  $\lambda_{\max,t}$ , the maximum eigenvalue of  $F_t$ , and normalizing the matrix  $\nabla L_* \sqrt{F_t}$ , we can write

$$\begin{aligned} p_t^{\text{opt}} &\geq \min_{\|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \frac{r_*}{2} \lambda_{\max,t} \|\Xi + \hat{\Upsilon}\|_F^2 + (\nabla L_* \sqrt{F_t}) \bullet (\Xi + \hat{\Upsilon}) \leq 0, \text{ and } \lambda_{\max,t} \|\Xi + \hat{\Upsilon}\|_F^2 \leq \epsilon_*^2 \right\} \\ &= \min_{\|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \frac{(\nabla L_* F_t^{1/2}) \bullet (\Xi + \hat{\Upsilon})}{\|\nabla L_* F_t^{1/2}\|_F} \leq \frac{-\lambda_{\max,t} r_* \|\Xi + \hat{\Upsilon}\|_F^2}{2\|\nabla L_* F_t^{1/2}\|_F}, \text{ and } \|\Xi + \hat{\Upsilon}\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\}. \end{aligned} \quad (14)$$

Observe that the inner product  $(\nabla L_* F_t^{1/2}) \bullet \hat{\Upsilon}$  is maximized by  $\Upsilon_{\#} := \frac{(\nabla L_* F_t^{1/2})}{\|\nabla L_* F_t^{1/2}\|_F}$  subject to  $\|\hat{\Upsilon}\|_F \leq 1$ . Since the probability distribution of  $\|\Xi + \hat{\Upsilon}\|_F^2$  is invariant under orthogonal transformation of  $\Xi$  and  $\hat{\Upsilon}$ , (14) also attains its minimum at  $\Upsilon_{\#}$ . Thus, we can rewrite (14) as

$$\begin{aligned} p_t^{\text{opt}} &\geq \mathbb{P}_t \left\{ \frac{(\nabla L_* F_t^{1/2}) \bullet \Xi}{\|\nabla L_* F_t^{1/2}\|_F} + 1 \leq \frac{-\lambda_{\max,t} r_*}{2\|\nabla L_* F_t^{1/2}\|_F} \|\Xi + \Upsilon_{\#}\|_F^2, \text{ and } \|\Xi + \Upsilon_{\#}\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\} \\ &= \mathbb{P}_t \left\{ \xi + 1 \leq -\frac{\lambda_{\max,t} r_*}{2\|\nabla L_* F_t^{1/2}\|_F} ((\xi + 1)^2 + X), \text{ and } (\xi + 1)^2 + X \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\}, \end{aligned} \quad (15)$$

where  $\xi \sim \mathcal{N}(0, 1)$  and  $X \sim \chi_{n^2-1}^2$  are independent standard normal and chi-squared distributions, and (15) is derived by rotating  $\Xi$  so that its first element is along the direction of  $\nabla L_* F_t^{1/2}$ . We use the following lemma to characterize the eigenvalues of  $F_t$  and control the lower bound (15) on  $p_t^{\text{opt}}$ .

**Lemma 10 (Bounded eigenvalues)** *Suppose  $T_w = O((\sqrt{T})^{1+o(1)})$ . Denote the minimum and maximum eigenvalues of  $F_t$  by  $\lambda_{\min,t}$  and  $\lambda_{\max,t}$ , respectively. Under the event  $E_T$ , for large enough  $T$ , we have that  $\lambda_{\max,t} \leq C \frac{\log T}{T_w}$  and  $\frac{\lambda_{\max,t}}{\lambda_{\min,t}} \leq C \frac{T \log T}{T_w}$  for any  $T_r < t \leq T$  for a constant  $C = \text{poly}(n, d, \log(1/\delta))$ .*

Lemma 10 states that maximum eigenvalue and the condition number of  $F_t$  are controlled inversely by the length of initial exploration phase  $T_w$  and proportionally by  $\log T$  and  $T \log T$  given that exploration time is bounded by a certain amount. The length of initial exploration  $T_w$  relative to the horizon  $T$  is critical in guaranteeing asymptotically constant optimistic probability  $p_t^{\text{opt}}$ . Although more lengthy initial exploration will lead to better convergence to constant optimistic probability, it also incurs higher asymptotic regret due to linear scaling of exploration regret with  $T_w$ .

Using the relation  $\|\nabla L_* F_t^{\frac{1}{2}}\|_F \geq \max(\sigma_{\min,*} \|F_t^{\frac{1}{2}}\|_F, \lambda_{\min,t}^{\frac{1}{2}} \|\nabla L_*\|_F)$  where  $\sigma_{\min,*}$  is the minimum singular value of  $\nabla L_*$ , we can further bound (15) from below. From Lemma 9, we can write  $\nabla L_* = 2P(\Theta_*)A_{c,*}\Sigma_*$  where  $P(\Theta_*) \succ 0$  is the solution to the DARE in (3) and  $\Sigma_* = \Sigma(\Theta_*, K_*) \succ 0$  is the stationary state covariance matrix. Notice that the minimum singular value of  $\nabla L_*$  is positive (*i.e.*  $\nabla L_*$  is full-rank) if and only if the closed-loop system matrix,  $A_{c,*}$ , is non-singular.

In general,  $A_{c,*}$  can be singular. Assuming that  $T_w = O((\sqrt{T})^{1+o(1)})$ , under the event  $E_T$ , we can use  $\|\nabla L_* F_t^{\frac{1}{2}}\|_F \geq \sqrt{\lambda_{\min,t}} \|\nabla L_*\|_F$  to obtain the following lower bound on  $p_t^{\text{opt}}$  for  $T_r < t \leq T$ :

$$\begin{aligned} p_t^{\text{opt}} &\geq \mathbb{P}_t \left\{ \xi + 1 \leq -\frac{\sqrt{\lambda_{\max,t}}}{2\rho_*} \sqrt{\frac{\lambda_{\max,t}}{\lambda_{\min,t}}} ((\xi + 1)^2 + X), \text{ and } (\xi + 1)^2 + X \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\}, \\ &\geq \mathbb{P} \left\{ \xi + 1 \leq -\frac{C}{2\rho_*} \frac{\sqrt{T} \log T}{T_w} ((\xi + 1)^2 + X), \text{ and } (\xi + 1)^2 + X \leq \frac{\epsilon_*^2 T_w}{C \log T} \right\}, \end{aligned}$$

where  $\rho_* := \|r_*^{-1} \nabla L_*\|_F$ . Choosing the exploration time as  $T_w = \omega(\sqrt{T} \log T)$  makes the coefficients  $\frac{\sqrt{T} \log T}{T_w} = o(1)$  to be very small and  $\frac{T_w}{\log T}$  to be very large, leading to constant lower bound on limiting optimistic probability  $\liminf_{T \rightarrow \infty} p_T^{\text{opt}} \geq \mathbb{P}\{\xi + 1 \leq 0\} =: Q(1)$ .

On the other hand, if  $A_{c,*}$  is non-singular, then we can use the alternative bound  $\|\nabla L_* \sqrt{F_t}\|_F \geq \sigma_{\min,*} \|\sqrt{F_t}\|_F \geq \sigma_{\min,*} \sqrt{\lambda_{\max,t}}$  to obtain the following lower bound for  $T_r < t \leq T$ :

$$\begin{aligned} p_t^{\text{opt}} &\geq \mathbb{P}_t \left\{ \xi + 1 \leq -\frac{\sqrt{\lambda_{\max,t}}}{2\sigma_{\min,*}} ((\xi + 1)^2 + X), \text{ and } (\xi + 1)^2 + X \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\}, \\ &\geq \mathbb{P} \left\{ \xi + 1 \leq -\frac{\sqrt{C}}{2\sigma_{\min,*}} \sqrt{\frac{\log T}{T_w}} ((\xi + 1)^2 + X), \text{ and } (\xi + 1)^2 + X \leq \frac{\epsilon_*^2 T_w}{C \log T} \right\}. \end{aligned}$$

Similarly, choosing the exploration time as  $T_w = \omega(\log T)$  makes the coefficients  $\sqrt{\frac{\log T}{T_w}} = o(1)$  to be very small and  $\frac{T_w}{\log T} = \omega(1)$  to be very large, leading to constant lower bound on limiting optimistic probability  $\liminf_{T \rightarrow \infty} p_T^{\text{opt}} \geq Q(1)$ .

In both cases, the optimistic probability achieves a constant lower bound for large enough  $T$  as  $p_T^{\text{opt}} \geq Q(1)(1 + o(1))^{-1}$ . This result can be interpreted in a geometric way as follows. As the time passes, the estimates of the system become more accurate in the sense that the confidence region of the estimate shrinks very quickly as controlled by the eigenvalues of  $F_t$ . Similarly, the high-probability region of TS samples also shrink very fast controlled by the covariance matrix  $F_t$ . Therefore, for large enough  $T$ , the confidence region of the model estimate and the high-probability region of TS samples get significantly smaller compared to the surrogate optimistic set  $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_*$ . This size difference effectively reduces the probability of finding a sampled system in  $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_*$  to the probability of finding a sampled system in the half-space separated by the tangent space  $T_{A_{c,*}} \mathcal{M}_*$ .



## 6. Numerical Experiments

Algorithm	Average	Top 95%	Top 90%	Average	Top 95%	Top 90%
	Regret			$\max \ x\ _2$		
TSAC	$4.58 \times 10^7$	$1.43 \times 10^5$	$9.49 \times 10^4$	$1.23 \times 10^3$	$1.07 \times 10^2$	$9.77 \times 10^1$
StabL	$1.34 \times 10^4$	$1.05 \times 10^3$	$9.60 \times 10^3$	$3.38 \times 10^1$	$3.14 \times 10^1$	$2.98 \times 10^1$
OFULQ	$1.47 \times 10^8$	$4.19 \times 10^6$	$9.89 \times 10^5$	$1.62 \times 10^3$	$5.21 \times 10^2$	$2.78 \times 10^2$
TS-LQR	$5.63 \times 10^{11}$	$3.07 \times 10^7$	$5.33 \times 10^6$	$6.26 \times 10^4$	$1.08 \times 10^3$	$6.39 \times 10^2$

Table 2: Regret and Maximum State Norm After 200 Time Steps in Boeing 747 Flight Control

Finally, we evaluate the performance of TSAC in longitudinal flight control of Boeing 747 with linearized dynamics (Ishihara et al., 1992). We compare TSAC with three adaptive control algorithms in literature that do not require an initial stabilizing policy: (i) OFULQ of Abbasi-Yadkori and Szepesvári (2011); (ii) TS-LQR of Abeille and Lazaric (2018); (iii) StabL of Lale et al. (2022). We perform 200 independent runs for 200 time-steps for each algorithm and report their average, top 95% and top 90% regret and maximum state norm performances. Note that, since optimistic control design is computationally intractable, we use projected gradient descent to *heuristically* find optimistic models in OFULQ and StabL. For fair comparison, we also adopt slow policy updates in OFULQ and TS-LQR and report the best results of each algorithm. Further details are in Appendix H. The results are presented in Table 2. Notice that TSAC achieves the second best performance after StabL. As expected, StabL outperforms TSAC since it performs much heavier computations to find the optimistic controller in the confidence set, whereas TSAC samples optimistic parameters only with some fixed probability. However, TSAC compares favorably against both OFULQ and TS-LQR, making it the best performing computationally efficient algorithm.

## 7. Conclusion and Future Directions

We present the first efficient adaptive control algorithm, TSAC, that attains optimal regret of  $\tilde{O}(\sqrt{T})$  in stabilizable LQRs without an initial stabilizing policy. We design TSAC to quickly stabilize the system and avoid state blow-ups via careful policy updates. Building on these design choices, the main technical contribution of this work is to show that TS samples optimistic parameters with constant probability in all LQRs, thereby resolving the conjecture in Abeille and Lazaric (2018).

This result highlights that a simple sampling strategy provides effective exploration to recover low-cost achieving controllers in adaptive control of LQRs which yields order optimal regret. An important future direction is to investigate whether TS achieves optimal regret in partially observable LTI systems, e.g. (Lale et al., 2020, 2021). Moreover, to obtain constant probability of sampling optimistic parameters for general LQRs, TSAC requires  $T_w = \omega(\sqrt{T} \log T)$  time-steps of improved exploration (Theorem 6), which causes the regret to be dominated by this phase. This long exploration is avoided in LQRs with non-singular optimal closed-loop matrix, which results in regret that scales polynomially in system dimensions (Theorem 3). It remains an open problem whether this polynomial dimension dependency in regret can be achieved via TS in general LQRs.

## References

- Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.
- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.
- Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. *arXiv preprint arXiv:1703.08972*, 2017.
- Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.
- Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. *arXiv preprint arXiv:2007.06482*, 2020.
- Shipra Agrawal. Recent advances in multiarmed bandits for sequential decision making. *Operations Research & Management Science in the Age of Analytics*, pages 167–188, 2019.
- Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 2. Athena scientific Belmont, MA, 1995.
- Sergio Bittanti, Marco C Campi, et al. Adaptive control of linear time invariant systems: the “bet on the best” principle. *Communications in Information & Systems*, 6(4):299–320, 2006.
- Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv preprint arXiv:2002.08095*, 2020.
- Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. 2020.
- Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. *arXiv preprint arXiv:1806.07104*, 2018.
- Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only  $\sqrt{T}$  regret. *arXiv preprint arXiv:1902.06223*, 2019.
- Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.
- Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.
- Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear-quadratic regulators. *Automatica*, 117:108982, 2020.
- Bernard Friedland. *Control system design: an introduction to state-space methods*. Courier Corporation, 2012.

- Zhong-Sheng Hou and Zhuo Wang. From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, 235:3–35, 2013.
- Tadashi Ishihara, Hai-Jiao Guo, and Hiroshi Takeda. A design of discrete-time integral controllers with computation delays via loop transfer recovery. *Automatica*, 28(3):599–603, 1992.
- Thomas Kailath, Ali H Sayed, and Babak Hassibi. Linear estimation, 2000.
- Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.
- Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Finite-time system identification and adaptive control in autoregressive exogenous systems. In *Learning for Dynamics and Control*, pages 967–979. PMLR, 2021.
- Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pages 5354–5390. PMLR, 2022.
- Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.
- Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.
- Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.
- Oron Sabag, Gautam Goel, Sahin Lale, and Babak Hassibi. Regret-optimal full-information control. *arXiv preprint arXiv:2105.01244*, 2021.
- Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

# Appendix

In Appendix A, we provide the notation tables for the paper. In Appendix B, we provide the system identification and stabilization guarantees of TSAC. In particular, we give the proof of Lemma 4 and give the precise duration of the TS with improved exploration phase  $T_w$ . In Appendix C, we show that under the joint event of  $\hat{E}_t \cap \tilde{E}_t$  the state stays bounded as described in Lemma 5 with high probability. In Appendix E, we provide the precise regret decomposition and discuss the individual terms in the regret upper bound. In Appendix D, we provide the complete proof of Theorem 6, as well as the intermediate results discussed in the main text. Appendix F comprises the analysis of individual terms in the regret decomposition. In particular, Appendix F.1 studies  $R_{T_w}^{\text{exp}}$ , Appendix F.2 studies  $R_T^{\text{RLS}}$ , Appendix F.3 studies  $R_T^{\text{mart}}$ , Appendix F.4 considers  $R_T^{\text{TS}}$ , Appendix F.5 bounds  $R_T^{\text{gap}}$ , and finally we combine these results to prove the regret upper bound of TSAC in Appendix F.6. In Appendix G, we give the technical theorems and lemmas used in the proofs. Finally, in Appendix H, we give the implementation details of all algorithms. Before proceeding the next section, we define the following high probability events which are standard in TS-based algorithms. First recall the RLS confidence ellipsoid given in Section 2.3:

$$\mathcal{E}_t^{\text{RLS}}(\delta) = \{\Theta : \|\Theta - \hat{\Theta}_t\|_{V_t} \leq \beta_t(\delta)\},$$

for  $\beta_t(\delta) = \sigma_w \sqrt{2n \log((T \det(V_t)^{1/2}) / (\delta \det(\mu I)^{1/2}))} + \sqrt{\mu} S$ . Further define

$$\mathcal{E}_t^{\text{TS}}(\delta) = \{\Theta : \|\Theta - \hat{\Theta}_t\|_{V_t} \leq v_t(\delta)\},$$

for  $v_t(\delta) = \beta_t(\delta) n \sqrt{(n+d) \log(n(n+d)/\delta)}$ . Define the events

$$\hat{E}_t = \{\forall s \leq t, \Theta_* \in \mathcal{E}_s^{\text{RLS}}(\delta)\} \tag{16}$$

$$\tilde{E}_t = \{\forall s \leq t, \tilde{\Theta}_s \in \mathcal{E}_s^{\text{TS}}(\delta)\}. \tag{17}$$

As described in Section 4,  $\hat{E}_t$  defines the event that RLS estimates  $\hat{\Theta}_t$  concentrate around  $\Theta_*$  and  $\tilde{E}_t$  defines the event that the sampled model parameter concentrates around  $\hat{\Theta}_t$ . From standard Gaussian tail bound and the self-normalized estimation error, we have that  $\hat{E} \cap \tilde{E}$  for all  $t \leq T$ , with probability at least  $1 - 2\delta$ . Here the time dependency dropped since  $\hat{E} := \hat{E}_T \subset \dots \subset \hat{E}_1$  and  $\tilde{E} := \tilde{E}_T \subset \dots \subset \tilde{E}_1$ . These events will be key in providing all the technical results starting from stabilization guarantees to final regret upper bound.

## Appendix A. Notation

This section contains two tables which list the notations used throughout the paper for improving readability. In particular, Table 3 provides the system dependent notations and the useful notations for presenting the design of TSAC. In Table 4, we present the notation used in deriving theoretical results, namely, the regret analysis and the lower bound on the probability of selecting optimistic parameters. Further details are also referenced to the related parts of the paper.

System Not.	Definition
$\Theta_*$	Unknown discrete-time LTI system with dynamics of (1); $[A_* \ B_*]^\top$
$x_t$	State of the system $\in \mathbb{R}^n$
$u_t$	Input to the system $\in \mathbb{R}^d$
$w_t$	Process noise as defined in Assumption 2; $\mathcal{N}(0, \sigma_w^2 I)$
$z_t$	Stack of current state and input; $[x_t^\top \ u_t^\top]^\top$
$Q, R$	Known cost matrices; $\ Q\ , \ R\  < \bar{\alpha}$ and $\sigma_{\min}(Q), \sigma_{\min}(R) > \underline{\alpha} > 0$
$c_t$	Quadratic cost at time $t$ ; $x_t^\top Q x_t + u_t^\top R u_t$
$\mathcal{S}$	Set of $(\kappa, \gamma)$ -stabilizable and bounded systems that $\Theta_*$ belongs (Assumption 1)
$P(\Theta)$	Unique p.d. solution to DARE (3) for a stabilizable system $\Theta = [A \ B]^\top$
$K(\Theta)$	Optimal controller for $\Theta$ ; $-(R + B^\top P(\Theta)B)^{-1}B^\top P(\Theta)A$
$J(\Theta)$	Average expected cost of system $\Theta$ ; $\sigma_w^2 \text{tr}(P(\Theta))$
$\kappa$	Bound over all possible optimal controllers in $\mathcal{S}$ ; $\sup_{\Theta \in \mathcal{S}} K(\Theta)$
$D$	Bound over all possible solutions to (3) in $\mathcal{S}$ ; $\bar{\alpha}\gamma^{-1}\kappa^2(1 + \kappa^2)$
TSAC Not.	
$\hat{\Theta}$	Least squares estimate of $\Theta_*$ using the history of inputs and states; $[\hat{A} \ \hat{B}]^\top$
$\mu$	Regularizer for least squares; set to $(1 + \kappa^2)X_s^2$
$V_t$	Regularized design matrix; $\mu I + \sum_{s=0}^{t-1} z_s z_s^\top$
$\eta_t$	Random matrix with iid standard normal entries used for sampling systems
$\mathcal{R}_{\mathcal{S}}(\cdot)$	Rejection sampling to make sure that sampled system belongs to $\mathcal{S}$
$\tilde{\Theta}$	System obtained via TS; $\mathcal{R}_{\mathcal{S}}(\hat{\Theta}_t + \beta_t(\delta)V_t^{-1/2}\eta_t)$
$\nu_t$	Improved exploration; $u_t = K(\tilde{\Theta}_t)x_t + \nu_t$ for $\nu_t \sim \mathcal{N}(0, 2\kappa^2\sigma_w^2 I)$
Quantities	
$\delta$	Fixed probability to define high probability events; $(0, 1)$
$T$	Time horizon
$T_w$	Duration of TS with improved exploration; defined in Theorem 3
$X_s$	Upper bound on state after stabilization; $\ x_t\  \leq X_s$ for $t > T_r$ w.h.p.
$S$	Upper bound on the Frobenius norm of $\Theta_*$
$\beta_t(\delta)$	Size of the RLS confidence ellipsoid at time $t$ ; $\sigma_w \sqrt{2n \log \left( \frac{\det(V_t)^{1/2}}{\delta \det(\mu I)^{1/2}} \right)} + \sqrt{\mu}S$
$\nu_t(\delta)$	Size of the sampling ellipsoid at time $t$ ; $\beta_t(\delta)n \sqrt{(n+d) \log(n(n+d)/\delta)}$
$\tau_0$	Fixed duration for each sampled policy; $2\gamma^{-1} \log(2\kappa\sqrt{2})$
$T_0$	Number of samples required to identify a stabilizing controller; (19)
$T_r$	Time required to control the state w.h.p.; $T_w + (n+d)\tau_0 \log(n+d)$

Table 3: Useful Notations for the Design of TSAC



Regret Analy.	
$R_T$	Regret of TSAC at until time $T$ ; $R_T = \sum_{t=0}^T (c_t - J(\Theta_*))$
$\mathcal{F}_t$	Filtration such that for all $t \geq 0$ , $x_t, z_t$ are $\mathcal{F}_t$ -measurable
$\mathcal{F}_t^{\text{cnt}}$	Information available to the controller up to time $t$ ; $\sigma(F_{t-1}, x_t)$
$R_T^{\text{exp}}$	Regret attained due to improved exploration (Appendix E)
$R_T^{\text{RLS}}$	Cost-to-go difference of the true and predicted next states (Appendix E)
$R_T^{\text{mart}}$	Martingale with bounded difference (Appendix E)
$R_T^{\text{TS}}$	Difference in $J(\Theta_*)$ and $J(\hat{\Theta})$ (Appendix E)
$R_T^{\text{gap}}$	Regret due to policy changes (Appendix E)
$\mathcal{E}_t^{\text{RLS}}(\delta)$	Regularized least squares confidence ellipsoid; $\{\Theta : \ \Theta - \hat{\Theta}_t\ _{V_t} \leq \beta_t(\delta)\}$
$\mathcal{E}_t^{\text{TS}}(\delta)$	Confidence ellipsoid for sampled system; $\{\Theta : \ \Theta - \hat{\Theta}_t\ _{V_t} \leq v_t(\delta)\}$
$\hat{E}_t$	Event of $\{\forall s \leq t, \Theta_* \in \mathcal{E}_t^{\text{RLS}}(\delta)\}$
$\tilde{E}_t$	Event of $\{\forall s \leq t, \tilde{\Theta}_s \in \mathcal{E}_t^{\text{TS}}(\delta)\}$
$\bar{E}_t$	Event of $\{\forall t \leq T_r, \ x_t\  \leq c'(n+d)^{n+d} \text{ and } \forall t > T_r, \ x_t\  \leq X_s\}$
$E_t$	$E_t = \hat{E}_t \cap \tilde{E}_t \cap \bar{E}_t$
Optimism Analy.	
$\mathcal{S}^{\text{opt}}$	Optimistic set; $\{\Theta = (A, B)^\top \in \mathbb{R}^{(n+d) \times n} \mid J(\Theta) \leq J(\Theta_*)\}$
$p_t^{\text{opt}}$	Probability of selecting optimistic system; $\mathbb{P}\{\tilde{\Theta}_t \in \mathcal{S}^{\text{opt}} \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\}$
$J(\Theta, K)$	Average expected cost of controlling $\Theta$ with a stabilizing controller $K$
$\Sigma(\Theta, K)$	Covariance matrix of the state in system $\Theta$ under controller $K$
$H_*^\top$	Concatenation of identity and optimal controller $K(\Theta_*)$ ; $[I, K(\Theta_*)^\top]$
$Q_*$	$Q + K(\Theta_*)^\top R K(\Theta_*)$
$L(A_c)$	Function that maps any stable matrix $A_c$ to $\sigma_w^2 \sum_{t=0}^{\infty} \ A_c^t\ _{Q_*}^2$
$F_t$	Confidence interval for estimated closed-loop system; $\beta_t^2 H_*^\top V_t^{-1} H_*$
$\lambda_{\max, t}$	Maximum eigenvalue of $F_t$
$\lambda_{\min, t}$	Minimum eigenvalue of $F_t$
$\Xi$	Random matrix of size $n \times n$ with iid $\mathcal{N}(0, 1)$ entries
$A_{c,*}$	Closed-loop system matrix of the $\Theta_*$ driven by $K(\Theta_*)$ ; $\Theta_*^\top H_*$
$\hat{A}_c$	Closed-loop system matrix of the $\hat{\Theta}$ driven by $K(\Theta_*)$ ; $\hat{\Theta}^\top H_*$
$\tilde{A}_c$	Closed-loop system matrix of the $\tilde{\Theta}$ driven by $K(\Theta_*)$ ; $\tilde{\Theta}^\top H_*$
$\mathcal{E}_t^{\text{cl}}(\delta)$	Closed-loop confidence set that is super set to $\mathcal{E}_t^{\text{RLS}}(\delta)$ ; (7)
$\hat{\Upsilon}$	Unit Fro. norm matrix s.t. $\hat{\Upsilon} \sqrt{F_t}$ is the h.p. confidence ellipsoid on $A_{c,*}$
$\mathcal{M}_n$	Manifold of square matrices of dimension $n$ ; $\mathbb{R}^{n \times n}$
$\mathcal{M}_{\text{Schur}}$	Manifold of (Schur-)stable matrices in $\mathcal{M}_n$ ; $\{A_c \in \mathcal{M}_n \mid \rho(A_c) < 1\}$
$\mathcal{M}_*$	Sublevel manifold in $\mathcal{M}_{\text{Schur}}$ s.t. $\{A_c \in \mathcal{M}_{\text{Schur}} \mid L(A_c) \leq L(A_{c,*})\}$
$G_t$	Perturbation around $A_{c,*}$ ; $(\Xi + \hat{\Upsilon}) \sqrt{F_t}$
$\nabla L_*$	Jacobian operator of $L(\cdot)$ evaluated at $A_c \in \mathcal{M}_{\text{Schur}}$
$\sigma_{\min,*}$	Minimum singular value of $\nabla L_*$
$\mathcal{H}_{A_c}$	Hessian operator of $L(\cdot)$ evaluated at $A_c \in \mathcal{M}_{\text{Schur}}$
$\mathcal{B}_*$	Stable ball for some constant $\epsilon_*$ ; $\{A_c \in \mathcal{M}_n \mid \ A_c - A_{c,*}\ _F \leq \epsilon_*\}$
$\mathcal{M}_*^{\text{qd}}$	Sublevel manifold; $\{A_c \in \mathcal{M}_n \mid \ A_c - A_{c,*} + r_*^{-1} \nabla L_*\ _F \leq \ r_*^{-1} \nabla L_*\ _F\}$

Table 4: Useful Notations for the Analysis

## Appendix B. System Identification and Stabilization Guarantees

In this section, we show that improved exploration of TSAC provides persistently exciting inputs, which will be used to enable reaching a stabilizing neighborhood around  $\Theta_*$ . From Assumption 2, we have that  $\mathbb{E}[x_{t+1}x_{t+1}^\top | \mathcal{F}_t] \succcurlyeq \sigma_w^2 I$ . Thus, with the input  $u_t = K(\tilde{\Theta}_t)x_t + v_t$  for  $v_t \sim \mathcal{N}(0, 2\kappa^2\sigma_w^2 I)$ , we have that  $\mathbb{E}[z_{t+1}z_{t+1}^\top | \mathcal{F}_t] \succcurlyeq \frac{\sigma_w^2}{2} I$ . Using Lemma 31, we have that  $V_t \succcurlyeq t \frac{\sigma_w^2}{40} I$  for  $t \geq 200(n+d) \log \frac{1}{\delta}$  with probability at least  $1 - \delta$ . Using the RLS estimate error bound given in Section 2.3, *i.e.*, under the event of  $\hat{E}_t$  we have

$$\|\hat{\Theta}_t - \Theta_*\|_2 \leq \frac{\beta_t}{\sqrt{\lambda_{\min}(V_t)}} \quad (18)$$

with probability at least  $1 - \delta$ . Plugging in the  $\lambda_{\min}(V_t)$  in its place yields the first result.

For the second result, we use Lemma 4.2 of Lale et al. (2022). Recall that  $D = \bar{\alpha}\gamma^{-1}\kappa^2(1 + \kappa^2)$ . Lemma 4.2 of Lale et al. (2022) states that for any  $(\kappa, \gamma)$ -stabilizable system  $\Theta_*$  and for any  $\varepsilon \leq \min\{\sqrt{(\sigma_w^2 n)/(142D^7)}, 1/(54D^5)\}$ , such that  $\|\Theta' - \Theta_*\| \leq \varepsilon$ ,  $K(\Theta')$  produces  $(\kappa\sqrt{2}, \gamma/2)$ -stable closed-loop dynamics on  $\Theta_*$  such that there exists  $L$  and  $H \succ 0$  such that  $A_* + B_*K(\Theta') = H' L H'^{-1}$ , with  $\|L\| \leq 1 - \gamma/2$  and  $\|H'\| \|H'^{-1}\| \leq \kappa\sqrt{2}$ . Under the event of  $\hat{E} \cap \tilde{E}$ , we have  $\|\tilde{\Theta}_t - \Theta_*\|_2 \leq \frac{\beta_t(\delta) + v_t(\delta)}{\sqrt{\lambda_{\min}(V_T)}}$ . Under the event of  $\hat{E} \cap \tilde{E}$ , this yields  $\|\tilde{\Theta}_t - \Theta_*\|_2 \leq \frac{7(\beta_t(\delta) + v_t(\delta))}{\sigma_w \sqrt{t}}$  with probability  $1 - \delta$ . Combining this result with the required  $\varepsilon$  for finding the stabilizing neighborhood, for TS with exploration duration of

$$T_w \geq T_0 := \frac{49(\beta_T(\delta) + v_T(\delta))^2}{\sigma_w \min\{(\sigma_w^2 n)/(142D^7), 1/(54^2 D^{10})\}}, \quad (19)$$

TSAC achieves  $(\kappa\sqrt{2}, \gamma/2)$ -stable closed-loop dynamics on  $\Theta_*$ , with probability at least  $1 - 3\delta$ .

## Appendix C. Boundedness of State, Proof of Lemma 5

In this section, we show that under the joint event of  $\hat{E} \cap \tilde{E}$  and the stabilization guarantee of the previous section, the state is bounded at all times during TSAC and it is well-controlled during the stabilizing TS phase, *i.e.* provide the proof of Lemma 5. We first consider the evolution of state for  $t \leq T_w$ . To bound the state for the first phase, we adapt the state bounding strategy given in Section 4.1 of Abbasi-Yadkori and Szepesvári (2011) for contractible systems to the stabilizable systems via the slow policy changes of TSAC. To this end, define the following

$$\bar{\alpha}_t = \frac{18\kappa^3}{\gamma(8\kappa - 1)} \bar{\eta}^{n+d} \left[ G Z_t^{\frac{n+d}{n+d+1}} \beta_t(\delta)^{\frac{1}{2(n+d+1)}} + (\|B_*\| \sigma_\nu + \sigma_w) \sqrt{2n \log \frac{nt}{\delta}} \right],$$

for

$$\bar{\eta} \geq \sup_{\Theta \in \mathcal{S}} \|A_* + B_*K(\Theta)\|, \quad Z_T = \max_{1 \leq t \leq T_r} \|z_t\|, \quad U = \frac{U_0}{H}$$

$$G = 2 \left( \frac{2S(n+d)^{n+d+1/2}}{\sqrt{U}} \right)^{1/(n+d+1)}, \quad U_0 = \frac{1}{16^{n+d-2} \max(1, S^{2(n+d-2)})}$$

and where  $H$  is any number satisfying

$$H > \max \left( 16, \frac{4S^2M^2}{(n+d)U_0} \right), \text{ where } M = \sup_{Y \geq 1} \frac{\left( \sigma_w \sqrt{n(n+d) \log \left( \frac{1+TY/\lambda}{\delta} \right)} + \lambda^{1/2} S \right)}{Y}.$$

Under the joint event of  $\hat{E}_t \cap \tilde{E}_t$ , [Abbasi-Yadkori and Szepesvári \(2011\)](#) show that the norm of the state is well-controlled except  $n+d$  times at most in any horizon  $T_r$ . Denoting the set of time-steps that the state is not well-controlled by  $\mathcal{T}_t$ , the following lemma formalizes this argument:

**Lemma 11 (Lemma 18 of [Abbasi-Yadkori and Szepesvári \(2011\)](#))** *We have that for any  $0 \leq t \leq T$ ,*

$$\max_{s \leq t, s \notin \mathcal{T}_t} \left\| (\Theta_* - \tilde{\Theta}_s)^\top z_s \right\| \leq G Z_t^{\frac{n+d}{n+d+1}} (\beta_t(\delta) + \nu(\delta))^{\frac{1}{2(n+d+1)}}.$$

Notice that this lemma is updated for TS. Moreover, it does not depend neither on the contractibility of the underlying system on the standard basis nor on the stabilizability. Equipped with this result, we write the closed loop system as

$$x_{t+1} = \Gamma_t x_t + r_t$$

where

$$\Gamma_t = \begin{cases} \tilde{A}_{t-1} + \tilde{B}_{t-1}K(\tilde{\Theta}_{t-1}) & t \notin \mathcal{T}_{T_w} \\ A_* + B_*K(\tilde{\Theta}_{t-1}) & t \in \mathcal{T}_{T_w} \end{cases} \text{ and } r_t = \begin{cases} (\Theta_* - \tilde{\Theta}_{t-1})^\top z_t + B_*\nu_t + w_t & t \notin \mathcal{T}_{T_w} \\ B_*\nu_t + w_t & t \in \mathcal{T}_{T_w} \end{cases} \quad (20)$$

Starting from  $x_0 = 0$ , we obtain the following roll out for the state,

$$\begin{aligned} x_t &= \Gamma_{t-1}x_{t-1} + r_{t-1} = \Gamma_{t-1}(\Gamma_{t-2}x_{t-2} + r_{t-2}) + r_t \\ &= \Gamma_{t-1}\Gamma_{t-2}\Gamma_{t-3}x_{t-3} + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} + \Gamma_{t-1}r_{t-1} + r_t \\ &= \Gamma_{t-1}\Gamma_{t-2} \cdots \Gamma_{t-(t-1)}r_1 + \cdots + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} + \Gamma_{t-1}r_{t-1} + r_t \\ &= \sum_{k=1}^t \left( \prod_{s=k}^{t-1} \Gamma_s \right) r_k \end{aligned} \quad (21)$$

Recall that the sampled model is an element of  $\mathcal{S}$  due to rejection sampling, thus, it is  $(\kappa, \gamma)$ -stabilizable by its optimal controller ([Assumption 1](#)):

$$1 - \gamma \geq \max_{t \leq T} \rho \left( \tilde{A}_t + \tilde{B}_tK(\tilde{\Theta}_t) \right). \quad (22)$$

Notice that multiplication of the closed-loop system matrices are not guaranteed to be contractive without a similarity transformation. Therefore, unlike [Abbasi-Yadkori and Szepesvári \(2011\)](#) that bounds the rollout terms via contractive mappings due to their assumption of contractive systems, we need to make sure that the policy changes does not cause unexpected growth in the magnitude of the state. The slow policy update schedule, *i.e.*, using all the sampled controllers

for fixed  $\tau_0$  time-steps, allows us to prevent such undesirable outcomes, In particular, by setting  $\tau_0 = 2\gamma^{-1} \log(2\kappa\sqrt{2})$ , we have that

$$\|x_t\| \leq \frac{18\kappa^3\bar{\eta}^{n+d}}{\gamma(8\kappa-1)} \left( \max_{1 \leq k \leq t} \|r_k\| \right) \quad (23)$$

Moreover, we have that  $\|r_k\| \leq \left\| (\Theta_* - \tilde{\Theta}_{k-1})^\top z_k \right\| + \|B_*\nu_k + w_k\|$  when  $k \notin \mathcal{T}_T$ , and  $\|r_k\| = \|B_*\nu_k + w_k\|$ , otherwise. Hence,

$$\max_{k \leq t} \|r_k\| \leq \max_{k \leq t, k \notin \mathcal{T}_T} \left\| (\Theta_* - \tilde{\Theta}_{k-1})^\top z_k \right\| + \max_{k \leq t} \|B_*\nu_k + w_k\|$$

The first term is bounded by the Lemma 11. The second term involves summation of independent  $\|B_*\| \sigma_\nu$  and  $\sigma_w$  Gaussian vectors. Using standard Gaussian tail inequalities, for all  $k \leq t$ , we have  $\|B_*\nu_k + w_k\| \leq (\|B_*\| \sigma_\nu + \sigma_w) \sqrt{2n \log \frac{nt}{\delta}}$  with probability at least  $1 - \delta$ . Therefore, on the joint event of  $\hat{E} \cap \tilde{E}$ ,

$$\|x_t\| \leq \frac{18\kappa^3\bar{\eta}^{n+d}}{\gamma(8\kappa-1)} \left[ GZ_t^{\frac{n+d}{n+d+1}} (\beta_t(\delta) + \nu(\delta))^{\frac{1}{2(n+d+1)}} + (\|B_*\| \sigma_\nu + \sigma_w) \sqrt{2n \log \frac{nt}{\delta}} \right] \quad (24)$$

for  $t \leq T_w$  with probability  $1 - \delta$ . Notice that this bound depends on  $Z_t$  and  $\beta_t(\delta)$  which in turn depends on  $x_t$ . Using Lemma 5 of [Abbasi-Yadkori and Szepesvári \(2011\)](#), one can obtain the following bound

$$\|x_t\| \leq c'(n+d)^{n+d}. \quad (25)$$

for some constant  $c'$  for all  $t \leq T_w$  with probability  $1 - 3\delta$ , which gives the first advertised result.

To bound the state for  $t > T_w$ , we show that, with the given choice of  $\tau_0$ , all the controllers during the stabilizing TS phase halves the magnitude of the state at the end of their control period. In particular, during the stabilizing TS phase, the closed-loop system dynamics can be written as  $x_{t+1} = (A_* + B_*K(\tilde{\Theta}_t))x_t + w_t = \Theta_*^\top H_{K(\tilde{\Theta}_t)} x_t + w_t$ . From the choice of  $T_w$  for the stabilizable systems, we have that  $\Theta_*^\top H_{K(\tilde{\Theta}_t)}$  is  $(\kappa\sqrt{2}, \gamma/2)$ -strongly stable. Thus, we have  $\rho(\Theta_*^\top H_{K(\tilde{\Theta}_t)}) \leq 1 - \gamma/2$  for all  $t > T_w$  and  $\|H_t\| \|H_t^{-1}\| \leq \kappa\sqrt{2}$  for  $H_t \succ 0$ , such that  $\|L_t\| \leq 1 - \gamma/2$  for  $\Theta_*^\top H_{K(\tilde{\Theta}_t)} = H_t L_t H_t^{-1}$ . Then for  $T > t > T_w$ , if the same policy,  $\Theta_*^\top H_{K(\tilde{\Theta}_t)}$  is applied starting from the state  $x_{T_w}$ , we have the following state roll-out on the event of  $\hat{E}_t \cap \tilde{E}_t$

$$\|x_t\| = \left\| \prod_{i=T_w+1}^t \Theta_*^\top H_{K(\tilde{\Theta}_i)} x_{T_w} + \sum_{i=T_w+1}^t \left( \prod_{s=i}^{t-1} \Theta_*^\top H_{K(\tilde{\Theta}_s)} \right) w_i \right\| \quad (26)$$

$$\leq \kappa\sqrt{2}(1 - \gamma/2)^{t-T_w} \|x_{T_w}\| + \max_{T_w < i \leq T} \|w_i\| \left( \sum_{i=T_w+1}^t \kappa\sqrt{2}(1 - \gamma/2)^{t-i+1} \right) \quad (27)$$

$$\leq \kappa\sqrt{2}(1 - \gamma/2)^{t-T_w} \|x_{T_w}\| + \frac{2\kappa\sigma_w\sqrt{2}}{\gamma} \sqrt{2n \log(n(t - T_w)/\delta)} \quad (28)$$

with probability at least  $1 - \delta$ . Since  $\tau_0 = 2\gamma^{-1} \log(2\kappa\sqrt{2})$ , we have  $\kappa\sqrt{2}(1 - \gamma/2)^{\tau_0} \leq 1/2$ . Therefore, at the end of each controller period the effect of previous state is halved. Using this fact,

at the  $i$ th policy change after  $T_w$ , we get

$$\begin{aligned} \|x_{t_i}\| &\leq 2^{-i}\|x_{T_w}\| + \sum_{j=0}^{i-1} 2^{-j} \frac{2\kappa\sigma_w\sqrt{2}}{\gamma} \sqrt{2n \log(n(t - T_w)/\delta)} \\ &\leq 2^{-i}\|x_{T_w}\| + \frac{4\kappa\sigma_w\sqrt{2}}{\gamma} \sqrt{2n \log(n(t - T_w)/\delta)} \end{aligned}$$

For all  $i > (n + d) \log(n + d) - \log(\frac{2\kappa\sigma_w\sqrt{2}}{\gamma} \sqrt{2n \log(n(t - T_w)/\delta)})$ , at policy change  $i$ , we get

$$\|x_{t_i}\| \leq \frac{6\kappa\sigma_w\sqrt{2}}{\gamma} \sqrt{2n \log(n(t - T_w)/\delta)}.$$

Finally, from (28), we have that

$$\|x_t\| \leq \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma} \sqrt{2n \log(n(t - T_w)/\delta)}, \quad (29)$$

with probability  $1 - 4\delta$  for all  $t > T_r := T_w + ((n + d) \log(n + d)) \tau_0$ . Based on this result, let  $X_s = \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma} \sqrt{2n \log(n(T - T_w)/\delta)}$ . We define our final good event,

$$\bar{E}_t = \{\forall t \leq T_r, \|x_t\| \leq c'(n + d)^{n+d} \text{ and } \forall t > T_r, \|x_t\| \leq X_s\}. \quad (30)$$

Notice that the joint event  $E_t = \hat{E}_t \cap \tilde{E}_t \cap \bar{E}_t$  holds with probability at least  $1 - 4\delta$ . This event will be the key conditioning in the regret decomposition and the analysis.

## Appendix D. Constant Probability of Sampling Optimistic Models, Proof of Theorem 6

In this section, we give the proof of the main technical contribution of this work, showing that TS samples optimistic model parameters with constant probability (Theorem 6). The proof follows the outline provided in Section 5. We first provide the proofs of each lemma in Section 5. In particular Lemma 7 is proven in Appendix D.1, Lemma 8 is studied in Appendix D.2, Lemma 9 in Appendix D.3, and Lemma 10 in D.4. Finally, we combine these results to prove Theorem 6 in Appendix D.5.

### D.1. Proof of Lemma 7

Given a stabilizable system  $\Theta = (A, B)^\top$ , and a stabilizing linear feedback controller  $K$ , we can find the LQR cost as follows

$$J(\Theta, K) = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[ \sum_{t=1}^T x_t^\top Q x_t + u_t^\top R u_t \right], \quad (31)$$

$$= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[ \sum_{t=1}^T \text{tr}((Q + K^\top R K) x_t x_t^\top) \right], \quad (32)$$

$$= \lim_{T \rightarrow \infty} \text{tr} \left( (Q + K^\top R K) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [x_t x_t^\top] \right), \quad (33)$$

$$= \text{tr}((Q + K^\top R K) \Sigma(\Theta, K)) \quad (34)$$



where  $\Sigma(\Theta, K) := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[x_t x_t^\top]$  is the stationary state covariance of the closed-loop system. In (32), we used the feedback control policy relation  $u_t = Kx_t$  and trace trick for inner products of vectors. Note that the closed-loop system evolves as

$$x_{t+1} = (A + BK)x_t + w_t. \quad (35)$$

The covariance of the state at time  $t$  can be written as a recursive relation

$$\mathbb{E}[x_{t+1} x_{t+1}^\top] = \mathbb{E}[(A + BK)x_t + w_t][(A + BK)x_t + w_t]^\top \quad (36)$$

$$= (A + BK) \mathbb{E}[x_t x_t^\top] (A + BK)^\top + \mathbb{E}[w_t w_t^\top] \quad (37)$$

$$= (A + BK) \mathbb{E}[x_t x_t^\top] (A + BK)^\top + \sigma_w^2 I \quad (38)$$

where (37) is because  $\mathbb{E}[w_t] = 0$  and  $w_t$  and  $x_t$  are independent. Since  $\rho(A + BK) < 1$ , the above iteration converges to a finite fixed-point. Furthermore, we have the following relation

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[x_{t+1} x_{t+1}^\top] = (A + BK) \frac{1}{T} \sum_{t=1}^T \mathbb{E}[x_t x_t^\top] (A + BK)^\top + \sigma_w^2 I \quad (39)$$

Denoting by  $\Sigma_T(\Theta, K) := \frac{1}{T} \sum_{t=1}^T \mathbb{E}[x_t x_t^\top]$  the finite averaged state covariance, we have the following

$$\Sigma_T(\Theta, K) + \frac{\mathbb{E}[x_{T+1} x_{T+1}^\top] - \mathbb{E}[x_1 x_1^\top]}{T} = (A + BK) \Sigma_T(\Theta, K) (A + BK)^\top + \sigma_w^2 I \quad (40)$$

Taking the limit of both sides as  $T \rightarrow \infty$  and noting that  $\mathbb{E}[x_{T+1} x_{T+1}^\top]$  has a finite value at the limit, we obtain the following Lyapunov equation

$$\Sigma(\Theta, K) = (A + BK) \Sigma(\Theta, K) (A + BK)^\top + \sigma_w^2 I \quad (41)$$

whose solution is given by the following convergent infinite sum

$$\Sigma(\Theta, K) = \sum_{t=0}^{\infty} (A + BK)^t \sigma_w^2 I ((A + BK)^\top)^t \quad (42)$$

It is well known that the optimal control policy of infinite-horizon LQR systems can be achieved by stationary linear feedback controllers Bertsekas (1995). Therefore, we can find the optimal LQR cost of a stabilizable system by minimizing its closed-loop cost among all stabilizing stationary linear feedback controllers.

Suppose  $\Theta \in \mathcal{S}^{\text{sur}}$ , i.e.,  $J(\Theta, K(\Theta_*)) \leq J(\Theta_*, K(\Theta_*))$ . Then, the optimal LQR cost of  $\Theta$  is given as

$$J(\Theta) = J(\Theta, K(\Theta)) = \min_{K \in \mathbb{R}^{d \times n}} J(\Theta, K) \quad (43)$$

$$\leq J(\Theta, K(\Theta_*)) \stackrel{(a)}{\leq} J(\Theta_*, K(\Theta_*)) = J(\Theta_*) \quad (44)$$

where (a) is due to  $\Theta \in \mathcal{S}^{\text{sur}}$ . Thus,  $\Theta \in \mathcal{S}^{\text{opt}}$ . ■

## D.2. Proof of Lemma 8

The following lemma will be used as the backbone for Lemma 8.

**Lemma 12** *Let  $V_1, V_2 \in \mathbb{R}^{n \times n}$  be symmetric positive semi-definite matrices. Define two ellipsoids as*

$$\mathcal{E}_1 := \{\Theta \in \mathbb{R}^{n \times m} \mid \text{tr}(\Theta^\top V_1 \Theta) \leq 1\} \quad \text{and} \quad \mathcal{E}_2 := \{\Theta \in \mathbb{R}^{n \times m} \mid \text{tr}(\Theta^\top V_2 \Theta) \leq 1\} \quad (45)$$

Then,  $\mathcal{E}_1 \subseteq \mathcal{E}_2$  if and only if  $V_1 \succcurlyeq V_2$ .

**Proof** For the forward direction, assume  $V_1 - V_2$  has a negative eigenvalue, *i.e.*, there exist  $\lambda < 0$  and a unit vector  $\theta \in \mathbb{R}^n / \{0\}$  such that  $(V_1 - V_2)\theta = \lambda\theta$ . Construct  $\Theta = [\theta, \theta, \dots, \theta] \in \mathbb{R}^{n \times m}$ . Observe that  $\text{tr}(\Theta^\top V_1 \Theta) = m\theta^\top V_1 \theta$  and  $\text{tr}(\Theta^\top V_2 \Theta) = m\theta^\top V_2 \theta$ . Therefore, we have the relationship  $\text{tr}(\Theta^\top V_2 \Theta) = \text{tr}(\Theta^\top V_1 \Theta) - m\lambda$ .

If  $V_1 \theta = 0$ , then  $\text{tr}(\Theta^\top V_1 \Theta) = 0 \leq 1$  and therefore for any scalar  $\alpha > 0$ ,  $\alpha\Theta \in \mathcal{E}_1$ . On the other hand,  $\text{tr}(\Theta^\top V_2 \Theta) = -m\lambda > 0$  and therefore, one can find a scalar  $\alpha > 0$  such that  $\text{tr}((\alpha\Theta)^\top V_2 (\alpha\Theta)) = -m\lambda\alpha^2 > 1$ , *i.e.*  $\alpha\Theta \notin \mathcal{E}_2$ . If  $V_1 \theta \neq 0$ , then define  $\Theta' = \frac{1}{\sqrt{m\theta^\top V_1 \theta}} \Theta$  and observe that  $\text{tr}(\Theta'^\top V_1 \Theta') = 1$ , *i.e.*,  $\Theta' \in \mathcal{E}_1$ . On the other hand,  $\text{tr}(\Theta'^\top V_2 \Theta') = 1 - \frac{\lambda}{\theta^\top V_1 \theta} > 1$ , *i.e.*,  $\Theta' \notin \mathcal{E}_2$ . Therefore, we have that if  $\mathcal{E}_1 \subseteq \mathcal{E}_2$  then  $V_1 \succcurlyeq V_2$ .

For the reverse direction, assume that  $V_1 \not\succeq V_2$  and  $\Theta \in \mathcal{E}_1$ . Then,  $\text{tr}(\Theta^\top (V_1 - V_2) \Theta) \geq 0$  and  $\text{tr}(\Theta^\top V_2 \Theta) \leq \text{tr}(\Theta^\top V_1 \Theta) \leq 1$ . Therefore,  $\Theta \in \mathcal{E}_2$ .  $\blacksquare$

**Proof of Lemma 8.** Let us rewrite the the ellipsoids. For the time being, we will drop  $\delta$  dependence for simplicity.

$$\mathcal{E}_t^{\text{RLS}} = \left\{ \hat{\Theta} \in \mathbb{R}^{(n+d) \times n} \mid \text{tr} \left( (\hat{\Theta} - \Theta_*)^\top \beta_t^{-1} V_t (\hat{\Theta} - \Theta_*) \right) \leq 1 \right\}, \quad (46)$$

$$\mathcal{E}_t^{\text{cl}} = \left\{ \hat{\Theta} \in \mathbb{R}^{(n+d) \times n} \mid \text{tr} \left( (\hat{\Theta} - \Theta_*)^\top H_* F_t^{-1} H_*^\top (\hat{\Theta} - \Theta_*) \right) \leq 1 \right\}. \quad (47)$$

In order to prove the lemma, it is necessary and sufficient to show  $\beta_t^{-1} V_t \succcurlyeq H_* F_t^{-1} H_*^\top$  by Lemma 12. Eliminating  $b_t$  terms from both sides and multiplying by  $V_t^{-\frac{1}{2}}$  from left and right, we obtain the equivalent condition,

$$I \succcurlyeq V_t^{-\frac{1}{2}} H_* (H_*^\top V_t^{-1} H_*)^{-1} H_*^\top V_t^{-\frac{1}{2}} = V_t^{-\frac{1}{2}} H_* (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}}, \quad (48)$$

In other words, we have that  $\mathcal{E}_t^{\text{RLS}} \subseteq \mathcal{E}_t^{\text{cl}}$  if and only if  $\|(H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}}\|_2 \leq 1$ . Notice that

$$\|(H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}}\|_2^2 = \sigma_1 \left( (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}} \right)^2, \quad (49)$$

$$= \lambda_{\max} \left( V_t^{-\frac{1}{2}} H_* (H_*^\top V_t^{-1} H_*)^{-1} H_*^\top V_t^{-\frac{1}{2}} \right), \quad (50)$$

$$= \lambda_{\max} \left( (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-1} H_* (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} \right), \quad (51)$$

$$= \lambda_{\max}(I) = 1, \quad (52)$$

where we used the fact that  $\sigma_1(A) = \sqrt{\lambda_{\max}(A^\top A)} = \sqrt{\lambda_{\max}(AA^\top)}$ . This is true for any time  $t$  and  $\delta$  and therefore completes the proof.  $\blacksquare$

### D.3. Proof of Lemma 9

The following lemma guarantees existence of a stable neighborhood around any stable matrix.

**Lemma 13** *Let  $A_c \in \mathcal{M}_{\text{Schur}}$ , i.e.,  $\rho(A_c) < 1$ . Then, there exists  $\epsilon > 0$  such that for any  $\Delta \in \mathcal{M}_n$  with  $\|\Delta\|_F \leq 1$ , we have that  $A_c + \epsilon\Delta \in \mathcal{M}_{\text{Schur}}$ , i.e.,  $\rho(A_c + \epsilon\Delta) < 1$ .*

**Proof** Per Gelfand's formula, we have that for any  $\delta > 0$ , there exists  $N_\delta \in \mathbb{N}$  such that

$$\rho(A_c) \leq \|A_c^k\|_F^{1/k} < \rho(A_c) + \delta \quad (53)$$

for any  $k \geq N_\delta$ . Since the mapping  $A_c \mapsto \|A_c^k\|_F^{1/k}$  is smooth for any  $k \in \mathbb{N}$ , we can write the following expansion by Taylor's theorem for any  $t \in \mathbb{R}$

$$\|(A_c + t\Delta)^k\|_F^{1/k} = \|A_c^k\|_F^{1/k} + t \frac{d}{dt} \|(A_c + t\Delta)^k\|_F^{1/k} \Big|_{\lambda t} \quad (54)$$

where  $\lambda \in [0, 1]$ . For a given  $t \in \mathbb{R}$ , there exists a constant  $M_{k,t} > 0$  such that for any  $\|\Delta\|_F \leq 1$ , we have that  $\left| \frac{d}{dt} \|(A_c + t\Delta)^k\|_F^{1/k} \right| \leq M_{k,t}$  by Taylor's theorem. Then, we can write the following upper bound

$$\|(A_c + t\Delta)^k\|_F^{1/k} \leq \|A_c^k\|_F^{1/k} + |t| M_{t,k} \quad (55)$$

Using the relation (53) and the upper bound (55), we have that for any  $\delta > 0$ ,  $t > 0$ , and  $\|\Delta\|_F \leq 1$ , there exists  $N_\delta \in \mathbb{N}$  and  $M_{t,N_\delta} > 0$  such that

$$\rho(A_c + t\Delta) \leq \|(A_c + t\Delta)^{N_\delta}\|_F^{1/N_\delta} \leq \|A_c^{N_\delta}\|_F^{1/N_\delta} + t M_{t,N_\delta} \quad (56)$$

$$< \rho(A_c) + \delta + t M_{t,N_\delta} \quad (57)$$

Fix a  $\delta > 0$  such that  $\rho(A_c) + \delta < 1$  and fix a  $t > 0$ . Then, we can find  $0 < \epsilon \leq t$  such that  $\rho(A_c) + \delta + \epsilon M_{t,N_\delta} < 1$  and thus

$$\rho(A_c + \epsilon\Delta) < \rho(A_c) + \delta + \epsilon M_{t,N_\delta} < 1 \quad (58)$$

for any  $\|\Delta\|_F \leq 1$  by (57). ■

**Proof of Lemma 9.** For any  $A_c \in \mathcal{M}_{\text{Schur}}$ , there exists a constant  $\epsilon > 0$ , such that for any  $\|\Delta\|_F \leq 1$ , we have that  $A_c + \epsilon\Delta \in \mathcal{M}_{\text{Schur}}$  by Lemma 13. To see smoothness of  $L$ , we write  $A_t := A_c + t\Delta$  and  $L(A_t) = \text{tr}(Q_*\Sigma_t)$  for any  $|t| \leq \epsilon$  and  $\|\Delta\|_F \leq 1$  where  $\Sigma_t$  solves the following Lyapunov equation

$$\Sigma_t - A_t \Sigma_t A_t^\top = \sigma_w^2 I \text{ and } \Sigma_0 - A_c \Sigma_0 A_c^\top = \sigma_w^2 I \quad (59)$$

Note that,  $\rho(A_t) < 1$  for any  $|t| \leq \epsilon$  and therefore both equations in (59) have unique solutions for any  $|t| \leq \epsilon$ . The Jacobian  $\nabla L(A_c) \in \mathcal{M}_n$  satisfies  $\nabla L(A_c) \bullet \Delta = \frac{d}{dt} L(A_t) \Big|_{t=0} = \text{tr}(Q_* \dot{\Sigma}_0)$  for any  $\|\Delta\|_F \leq 1$  where  $\dot{\Sigma}_t$  is the derivative of  $\Sigma_t$  and satisfies the following Lyapunov equation

$$\dot{\Sigma}_t - A_t \dot{\Sigma}_t A_t^\top = \Delta \Sigma_t A_t^\top + A_t \Sigma_t \Delta^\top \text{ and } \dot{\Sigma}_0 - A_c \dot{\Sigma}_0 A_c^\top = \Delta \Sigma_0 A_c^\top + A_c \Sigma_0 \Delta^\top \quad (60)$$

Similarly, both equations in (60) have unique solutions for any  $|t| \leq \epsilon$  and therefore  $\nabla L(A_c)$  exists for any  $A_c$ . To find the Jacobian, we have that  $\dot{\Sigma}_0 = \sum_{k=0}^{\infty} A_c^k (\Delta \Sigma_0 A_c^\top + A_c \Sigma_0 \Delta^\top) (A_c^\top)^k$  and

$$\text{tr}(Q_* \dot{\Sigma}_0) = \text{tr} \left( Q_* \sum_{k=0}^{\infty} A_c^k (\Delta \Sigma_0 A_c^\top + A_c \Sigma_0 \Delta^\top) (A_c^\top)^k \right) \quad (61)$$

$$= 2 \text{tr} \left( \sum_{k=0}^{\infty} (A_c^\top)^k Q_* A_c^k A_c \Sigma_0 \Delta^\top \right) = 2 \sum_{k=0}^{\infty} (A_c^\top)^k Q_* A_c^k A_c \Sigma_0 \bullet \Delta \quad (62)$$

Therefore,  $\nabla L(A_c) = 2 \sum_{k=0}^{\infty} (A_c^\top)^k Q_* A_c^k A_c \Sigma_0$ . In particular, in the case of  $A_{c,*}$ , we have that  $\sum_{k=0}^{\infty} (A_{c,*}^\top)^k Q_* A_{c,*}^k = P_*$ , the solution to the Riccati equation, and thus  $\nabla L(A_{c,*}) = 2P_* A_{c,*} \Sigma_*$ . Repeating the same process, one can see that  $L(A_t)$  is infinitely differentiable and thus we conclude  $L$  is a smooth function.

Denote by  $\mathcal{B}_\epsilon := \{A \in \mathcal{M}_n \mid \|A - A_c\|_F \leq \epsilon\} \subset \mathcal{M}_{\text{Schur}}$  the ball of radius  $\epsilon > 0$  around  $A_c \in \mathcal{M}_{\text{Schur}}$ . Consider the function  $L$  restricted to the domain  $\mathcal{B}_\epsilon$ . Since  $\mathcal{B}_\epsilon$  is a convex set, we can apply Taylor's theorem to  $L$  around  $A_c$  in this domain to obtain

$$L(A_c + \epsilon \Delta) = L(A_c) + \nabla L(A_c) \bullet \epsilon \Delta + \frac{1}{2} \epsilon \Delta \bullet \mathcal{H}_{A_c + s \Delta}(\epsilon \Delta) \quad (63)$$

for  $\|\Delta\|_F \leq 1$  and for some  $s \in [0, \epsilon]$ . Here,  $\mathcal{H}_{A_c} : \mathcal{M}_n \rightarrow \mathcal{M}_n$  is the Hessian operator evaluated at a point  $A_c \in \mathcal{M}_{\text{Schur}}$  and satisfies the following relationship

$$\Delta \bullet \mathcal{H}_{A_c}(\Delta) = \left. \frac{d^2}{dt^2} L(A_c + t \Delta) \right|_{t=0} \quad (64)$$

for any  $\|\Delta\|_F \leq 1$ . Finally, there exists a constant  $r > 0$ , such that for any  $G \in \mathcal{M}_n$ , we have that  $|G \bullet \mathcal{H}_{A_c + s \Delta}(G)| \leq r \|G\|_F^2$  for any  $s \in [0, \epsilon]$  and  $\|\Delta\|_F \leq 1$  by Taylor's theorem.  $\blacksquare$

#### D.4. Proof of Lemma 10

In this section, we will assume that Assumptions 1 & 2 hold. First, we need to show the boundedness of the stacked state and control input vector,  $z_t$ .

**Lemma 14** *Define the terms*

$$Z'_{T_w} := (1 + \kappa) c' (n + d)^{n+d} + \kappa \sigma_w \sqrt{4d \log(d T_w / \delta)} \quad (65)$$

$$Z''_T := (1 + \kappa) (12\kappa^2 + 2\kappa\sqrt{2}) \gamma^{-1} \sigma_w \sqrt{2n \log(n(T - T_w) / \delta)} \quad (66)$$

Then, the following holds w.p. at least  $1 - 4\delta$ ,

$$\|z_t\| \leq \begin{cases} Z'_{T_w}, & \text{for } t \leq T_r \\ Z''_T, & \text{for } T_r < t \leq T \end{cases} \quad (67)$$

**Proof** From Lemma 5, we know that  $\|x_t\| \leq c' (n + d)^{n+d}$  with  $c' > 0$  a constant for  $t \leq T_r$  and  $\|x_t\| \leq (12\kappa^2 + 2\kappa\sqrt{2}) \gamma^{-1} \sigma_w \sqrt{2n \log(n(t - T_w) / \delta)}$  for all  $T_r < s \leq T$  w.p. at least  $1 - 4\delta$ . Furthermore, under the event of  $E_t$ , we have that  $\|u_t\| \leq \kappa \|x_t\| + \|v_t\| \leq \kappa \|x_t\| + \kappa \sigma_w \sqrt{4d \log(d T_w / \delta)}$  for all  $0 \leq t \leq T_w$ . Observing that  $\|z_t\| = \sqrt{\|x_t\|^2 + \|u_t\|^2} \leq \|x_t\| +$

$\|u_t\|$ , one can reach the desired result by substituting the appropriate bounds on  $\|x_t\|$  and  $\|u_t\|$  and considering the maximal case achieved when  $t = T$ .  $\blacksquare$

The following lemma will be used to bound  $V_t$ .

**Lemma 15** *Let  $V_t = \mu I + \sum_{s=0}^{t-1} z_s z_s^\top$ . On the event of  $E_T = \hat{E} \cap \tilde{E} \cap \bar{E}$ , we have*

$$\lambda_{\max}(V_t) \leq \begin{cases} \mu + tZ_{T_w}'^2, & \text{for } t \leq T_r \\ \mu + T_r Z_{T_w}'^2 + (t - T_r)Z_T''^2, & \text{for } T_r < t \leq T \end{cases} \quad (68)$$

$$\text{and } \lambda_{\min}(V_t) \geq \begin{cases} \mu + t\frac{\sigma_w^2}{40}, & \text{for } 200(n+d)\log\frac{12}{\delta} \leq t \leq T_w \\ \mu + T_w\frac{\sigma_w^2}{40}, & \text{for } T_w < t \leq T \end{cases} \quad (69)$$

**Proof** Recall that on the event  $E_T$ , the RLS estimates, TS sampled systems are concentrated and the state is bounded, *i.e.*, Lemma 5. Conditioned on this event, we will start with bounding  $\lambda_{\max}(V_t)$ . For any time  $0 \leq t \leq T$ , triangle inequality gives  $\lambda_{\max}(V_t) = \|\mu I + \sum_{s=0}^{t-1} z_s z_s^\top\|_2 \leq \mu + \sum_{s=0}^{t-1} \|z_s\|^2$ . Using the bounds on  $\|z_t\|$  given in Lemma 14, we can write  $\lambda_{\max}(V_t) \leq \mu + tZ_{T_w}'^2$  for  $t \leq T_r$  and  $\lambda_{\max}(V_t) \leq \mu + T_r Z_{T_w}'^2 + (t - T_r)Z_T''^2$  for  $T_r < t \leq T$ . For the lower bound, note that we have that  $\mathbb{E}[z_{t+1} z_{t+1}^\top | \mathcal{F}_t] \succcurlyeq \frac{\sigma_w^2}{2} I$ . Using Lemma 31, on the event  $E_T$ , we have that  $V_t \succcurlyeq \mu I + t\frac{\sigma_w^2}{40} I$  for  $200(n+d)\log\frac{12}{\delta} \leq t \leq T_w$ . Since  $V_{t+1} = V_t + z_t z_t^\top$ , we have that  $V_t \succcurlyeq V_{T_w} \succcurlyeq \mu I + T_w\frac{\sigma_w^2}{40} I$  for  $T_w < t \leq T$ .  $\blacksquare$

Finally, we will use the following lemma to bound  $\beta_t(\delta) = \sigma_w \sqrt{2n \log\left(\frac{\det(V_t)^{1/2}}{\delta \det(\mu I)^{1/2}}\right)} + \sqrt{\mu} S$

**Lemma 16** *On the event of  $E_T$ , we have the following upper bound on  $\beta_T(\delta)$ :*

$$\beta_T(\delta) \leq 4\sigma_w^2 n \log\left(\frac{1}{\delta}\right) + 2\sigma_w^2 n(n+d) \log\left(1 + \frac{T_r Z_{T_w}'^2 + (T - T_r)Z_T''^2}{(n+d)\mu}\right) + 2\mu S^2 \quad (70)$$

**Proof** Following a similar approach pursued in Lemma 10 of Abbasi-Yadkori and Szepesvári (2011), we can bound the log-determinant of  $V_t$  as

$$\log \frac{\det(V_T)}{\det(\mu I)} \leq (n+d) \log\left(1 + \frac{T_r Z_{T_w}'^2 + (T - T_r)Z_T''^2}{(n+d)\mu}\right)$$

by Lemma 15. This leads to the following upper bound on  $\beta_t(\delta)$

$$\begin{aligned} \beta_T(\delta)^2 &\leq \left( \sigma_w \sqrt{2n \log\left(\frac{1}{\delta}\right) + n(n+d) \log\left(1 + \frac{T_r Z_{T_w}'^2 + (T - T_r)Z_T''^2}{(n+d)\mu}\right)} + \sqrt{\mu} S \right)^2 \\ &\leq 4\sigma_w^2 n \log\left(\frac{1}{\delta}\right) + 2\sigma_w^2 n(n+d) \log\left(1 + \frac{T_r Z_{T_w}'^2 + (T - T_r)Z_T''^2}{(n+d)\mu}\right) + 2\mu S^2. \end{aligned}$$



**Proof of Lemma 10.** We will first show the desired bounds on  $\lambda_{\min}(F_t)$  and  $\lambda_{\max}(F_t)$ . Recall that the event  $E_T$  holds with probability at least  $1 - 4\delta$ . Noting that  $H_*^\top H_* = I + K_*^\top K_*$ , it is clear that  $F_t \succcurlyeq \beta_t^2 \lambda_{\min}(V_t^{-1}) H_*^\top H_* \succcurlyeq \frac{\beta_t^2}{\lambda_{\max}(V_t)} I$ . Thus, from Lemma 15, for  $T_r < t \leq T$ , we have that  $\lambda_{\min,t} \geq \frac{\beta_t^2}{\lambda_{\max}(V_t)} \geq \frac{\beta_t^2}{\mu + T_r Z_{T_w}^2 + (t - T_r) Z_T^2}$ .

On the other hand,  $F_t \preccurlyeq \beta_t^2 \lambda_{\max}(V_t^{-1}) H_*^\top H_* \preccurlyeq \frac{\beta_t^2 (1 + \kappa^2)}{\lambda_{\min}(V_t)} I$ . Again using Lemma 15, for  $T_r < t \leq T$ , we have that  $\lambda_{\max,t} \leq \frac{(1 + \kappa^2) \beta_t^2}{\lambda_{\min}(V_t)} \leq \frac{(1 + \kappa^2) \beta_t^2}{\mu + T_w \frac{\sigma_w^2}{40}}$ . Since  $t \mapsto \beta_t$  is increasing,  $t \mapsto \lambda_{\max,t}$  is increasing as well. The condition number  $\kappa_t := \frac{\lambda_{\max,t}}{\lambda_{\min,t}} \leq \frac{\mu + T_r Z_{T_w}^2 + (t - T_r) Z_T^2}{(1 + \kappa^2)^{-1} (\mu + T_w \frac{\sigma_w^2}{40})}$  is increasing for  $T_r < t \leq T$ .

If  $T_w = O(\sqrt{T}^{1+o(1)})$ , then we have that  $\lambda_{\max}(V_T) \leq O(\text{poly}(n, d, \log(1/\delta)) T \log T)$  and  $\beta_T(\delta) \leq O(\text{poly}(n, d, \log(1/\delta)) \log T)$ . Thus, there are positive constants  $C = \text{poly}(n, d, \log(1/\delta))$  and  $c = \text{poly}(n, d, \log(1/\delta))$  such that  $\lambda_{\max,T} \leq C \frac{\log T}{T_w}$  and  $\kappa_t = \frac{\lambda_{\max,t}}{\lambda_{\min,t}} \leq c \frac{T \log T}{T_w}$  for  $T_r < t \leq T$  for large enough  $T$ . Choosing the larger between  $C$  and  $c$  yields the desired result.  $\blacksquare$

## D.5. Proof of Theorem 6

Defining by  $p_t^{\text{opt}} := \mathbb{P}\{\tilde{\Theta}_t \in \mathcal{S}^{\text{opt}} \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\}$  the optimistic probability, and by  $\mathbb{P}_t\{\cdot\} := \mathbb{P}\{\cdot \mid \mathcal{F}_t^{\text{cnt}}\}$  conditional probability measure, we can write

$$p_t^{\text{opt}} \geq \mathbb{P}\{\tilde{\Theta}_t \in \mathcal{S}^{\text{surr}} \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\} \quad (71)$$

$$= \mathbb{P}\{L(\tilde{\Theta}_t^\top H_*) \leq L(\Theta_*^\top H_*) \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\} \quad (72)$$

$$\geq \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \eta^\top \beta_t V_t^{-\frac{1}{2}} H_*) \leq L(\Theta_*^\top H_*)\} \quad (73)$$

$$= \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \Xi \sqrt{F_t}) \leq L(\Theta_*^\top H_*)\} \quad (74)$$

where (71) is by Lemma 7, (73) is a worst-case estimation bound within high-probability confidence region, and (74) is because  $\eta^\top \beta_t V_t^{-\frac{1}{2}} H_*$  and  $\Xi \sqrt{F_t}$  have the same distributions with  $\eta \in \mathbb{R}^{(n+d) \times n}$  and  $\Xi \in \mathbb{R}^{n \times n}$  being i.i.d. standard normal random matrices.

The bound in (74) can be further lower bounded by minimizing over a larger confidence set as

$$p_t^{\text{opt}} \geq \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{cl}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \Xi \sqrt{F_t}) \leq L(A_{c,*})\} \quad (75)$$

$$= \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t\{L(A_{c,*} + (\Xi + \hat{\Upsilon}) \sqrt{F_t}) \leq L(A_{c,*})\}, \quad (76)$$

where (75) is by Lemma (8) and (76) is because  $H_*$  is full column rank and therefore we can minimize over closed-loop matrices instead of open-loop system parameters.

Denoting by  $G_t = (\Xi + \hat{\Upsilon}) \sqrt{F_t}$  the perturbation due to estimation and sampling, Lemma 9 suggests

that there exists constants  $\epsilon_* > 0$  and  $r_* > 0$  such that

$$L(A_{c,*} + G_t) = L(A_{c,*}) + \nabla L_* \bullet G_t + \frac{1}{2} G_t \bullet \mathcal{H}_{A_{c,*} + sG_t}(G_t) \quad (77)$$

$$\leq L(A_{c,*}) + \nabla L_* \bullet G_t + \frac{r_*}{2} \|G_t\|_F^2 \quad (78)$$

whenever  $\|G_t\|_F \leq \epsilon_*$ . Substituting (78) into (76) leads to the following lower bound

$$p_t^{\text{opt}} \geq \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \{L(A_{c,*} + G_t) \leq L(A_{c,*})\} \quad (79)$$

$$\geq \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{l} L(A_{c,*}) + \nabla L_* \bullet G_t + \frac{r_*}{2} \|G_t\|_F^2 \leq L(A_{c,*}), \\ \text{and } \|G_t\|_F \leq \epsilon_* \end{array} \right\} \quad (80)$$

$$= \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{l} \frac{r_*}{2} \|(\Xi + \hat{\Upsilon})\sqrt{F_t}\|_F^2 + \nabla L_* \bullet (\Xi + \hat{\Upsilon})\sqrt{F_t} \leq 0, \\ \text{and } \|(\Xi + \hat{\Upsilon})\sqrt{F_t}\|_F \leq \epsilon_* \end{array} \right\} \quad (81)$$

Noting that  $\|(\Xi + \hat{\Upsilon})\sqrt{F_t}\|_F \leq \sqrt{\lambda_{\max,t}} \|\Xi + \hat{\Upsilon}\|_F$  where  $\lambda_{\max,t} := \lambda_{\max}(F_t)$ , we can further relax the lower bound (81) as

$$p_t^{\text{opt}} \geq \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{l} \frac{\lambda_{\max,t} r_*}{2} \|\Xi + \hat{\Upsilon}\|_F^2 + (\nabla L_* \sqrt{F_t}) \bullet (\Xi + \hat{\Upsilon}) \leq 0, \\ \text{and } \sqrt{\lambda_{\max,t}} \|\Xi + \hat{\Upsilon}\|_F \leq \epsilon_* \end{array} \right\} \quad (82)$$

$$= \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{l} \left\| \Xi + \hat{\Upsilon} + \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F^2 \leq \left\| \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F^2, \\ \text{and } \|\Xi + \hat{\Upsilon}\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\} \quad (83)$$

where (83) is obtained by completion of squares. Let  $\mathcal{U} : \mathcal{M}_n \rightarrow \mathcal{M}_n$  be an orthogonal transformation such that  $\mathcal{U} \left( \hat{\Upsilon} + \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right) = \left\| \hat{\Upsilon} + \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F E_{11}$  where  $E_{11} \in \mathcal{M}_n$  has 1 in its (1, 1) entry and zeros elsewhere. Since Frobenius norm and the probability density of  $\Xi$  are invariant under orthogonal transformations, (83) can be rewritten as

$$p_t^{\text{opt}} \geq \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{l} \left\| \mathcal{U} \left( \Xi + \hat{\Upsilon} + \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right) \right\|_F^2 \leq \left\| \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F^2, \\ \text{and } \|\mathcal{U}(\Xi + \hat{\Upsilon})\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\} \quad (84)$$

$$= \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{l} \left\| \Xi + \left\| \hat{\Upsilon} + \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F E_{11} \right\|_F^2 \leq \left\| \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F^2, \\ \text{and } \|\Xi + \mathcal{U}(\hat{\Upsilon})\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\} \quad (85)$$

$$= \min_{\hat{\Upsilon} : \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{l} \left( \Xi_{11} + \left\| \hat{\Upsilon} + \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F \right)^2 + \sum_{i,j \neq 1,1} \Xi_{ij}^2 \leq \left\| \frac{\nabla L_* \sqrt{F_t}}{\lambda_{\max,t} r_*} \right\|_F^2, \\ \text{and } \|\Xi + \mathcal{U}(\hat{\Upsilon})\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\} \quad (86)$$

Notice that the probability in (86) is described by the intersection of two balls whose centers are far apart by  $\frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*}$  and hence the intersection has a fixed shape. Choosing  $\hat{\Upsilon}$  along the direction of  $\|\nabla L_* \sqrt{F_t}\|_F$  moves the center of the first ball furthest possible from the origin which leads to the intersection of the balls to move furthest away from the origin as well. Therefore, the probability in

(86) attains its minimum at  $\hat{\Upsilon}_\# := \frac{\nabla L_* \sqrt{F_t}}{\|\nabla L_* \sqrt{F_t}\|_F}$  and (86) can be equivalently expressed by

$$p_t^{\text{opt}} \geq \mathbb{P}_t \left\{ \begin{array}{l} \left( \Xi_{11} + 1 + \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max, t} r_*} \right)^2 + \sum_{i, j \neq 1, 1} \Xi_{ij}^2 \leq \frac{\|\nabla L_* \sqrt{F_t}\|_F^2}{\lambda_{\max, t}^2 r_*^2}, \\ \text{and } \|\Xi + E_{11}\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max, t}} \end{array} \right\} \quad (87)$$

$$= \mathbb{P}_t \left\{ \begin{array}{l} \left( \xi + 1 + \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max, t} r_*} \right)^2 + X \leq \frac{\|\nabla L_* \sqrt{F_t}\|_F^2}{\lambda_{\max, t}^2 r_*^2}, \\ \text{and } (\xi + 1)^2 + X \leq \frac{\epsilon_*^2}{\lambda_{\max, t}} \end{array} \right\} \quad (88)$$

where  $\xi \sim \mathcal{N}(0, 1)$  and  $X \sim \chi_{n^2-1}^2$  are independent normal and chi-squared random variables, respectively. Denoting by  $a_t := \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max, t} r_*}$  and  $b_t = \frac{\epsilon_*}{\sqrt{\lambda_{\max, t}}}$  the radii of the balls, we can rewrite (88) as

$$p_t^{\text{opt}} \geq \mathbb{P}_t \left\{ (\xi + 1 + a_t)^2 + X \leq a_t^2, \text{ and } (\xi + 1)^2 + X \leq b_t^2 \right\} \quad (89)$$

$$= \mathbb{P}_t \left\{ \begin{array}{l} |\xi + 1 + a_t| \leq \sqrt{a_t^2 - X}, \text{ and } |\xi + 1| \leq \sqrt{b_t^2 - X}, \\ \text{and } X \leq \min(a_t^2, b_t^2) \end{array} \right\} \quad (90)$$

$$= \int_0^{\min(a_t^2, b_t^2)} \mathbb{P}_t \left\{ |\xi + 1 + a_t| \leq \sqrt{a_t^2 - x}, \text{ and } |\xi + 1| \leq \sqrt{b_t^2 - x} \right\} f_{n^2-1}(x) dx \quad (91)$$

$$= \int_0^{\min(a_t^2, b_t^2)} \mathbb{P}_t \left\{ \begin{array}{l} 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + a_t + \sqrt{a_t^2 - x}, \\ \text{and } 1 - \sqrt{b_t^2 - x} \leq \xi \leq 1 + \sqrt{b_t^2 - x}, \end{array} \right\} f_{n^2-1}(x) dx \quad (92)$$

where  $f_k(x) := \left(2^{\frac{k}{2}} \Gamma(\frac{k}{2})\right)^{-1} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$  is the probability density function of the chi-squared distribution with  $k \in \mathbb{N}$  degrees of freedom. (91) is derived from law of total probability. Notice that the probability inside the integral in (92) is determined by the intersection of two intervals. This probability will have a non-zero value only for a fixed interval of  $x$  depending on the relation between  $a_t$  and  $b_t$ . We will investigate three cases:

**i.  $0 \leq b_t \leq \sqrt{2}a_t$  :** There is a non-empty intersection if and only if  $0 \leq x \leq b_t^2 \left(1 - \frac{b_t^2}{4a_t^2}\right)$  and the integral (92) becomes

$$p_t^{\text{opt}} \geq \int_0^{b_t^2 \left(1 - \frac{b_t^2}{4a_t^2}\right)} \mathbb{P}_t \left\{ 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + \sqrt{b_t^2 - x} \right\} f_{n^2-1}(x) dx \quad (93)$$

$$= \int_0^{b_t^2 \left(1 - \frac{b_t^2}{4a_t^2}\right)} \left[ Q \left( 1 + a_t - \sqrt{a_t^2 - x} \right) - Q \left( 1 + \sqrt{b_t^2 - x} \right) \right] f_{n^2-1}(x) dx \quad (94)$$

where  $Q$  is the Gaussian  $Q$ -function. Notice that for fixed values of  $b_t$ , (94) is monotonically increasing with respect to  $a_t$  and vice versa.

*ii.*  $\sqrt{2}a_t \leq b_t \leq 2a_t$  : There is a non-empty intersection if and only if  $0 \leq x \leq a_t^2$  and the integral (92) becomes

$$p_t^{\text{opt}} \geq \int_0^{a_t^2} \mathbb{P}_t \left\{ 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + \sqrt{b_t^2 - x} \right\} f_{n^2-1}(x) dx \quad (95)$$

$$= \int_0^{a_t^2} \left[ Q \left( 1 + a_t - \sqrt{a_t^2 - x} \right) - Q \left( 1 + \sqrt{b_t^2 - x} \right) \right] f_{n^2-1}(x) dx \quad (96)$$

Notice that for fixed values of  $b_t$ , (96) is monotonically increasing with respect to  $a_t$  and vice versa.

*iii.*  $2a_t \leq b_t$  : There is a non-empty intersection if and only if  $0 \leq x \leq a_t^2$  and the integral (92) becomes

$$p_t^{\text{opt}} \geq \int_0^{a_t^2} \mathbb{P}_t \left\{ 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + a_t + \sqrt{a_t^2 - x} \right\} f_{n^2-1}(x) dx \quad (97)$$

$$= \int_0^{a_t^2} \left[ Q \left( 1 + a_t - \sqrt{a_t^2 - x} \right) - Q \left( 1 + a_t + \sqrt{a_t^2 - x} \right) \right] f_{n^2-1}(x) dx \quad (98)$$

Notice that for fixed values of  $b_t$ , (98) is monotonically increasing with respect to  $a_t$  and vice versa.

As seen from all three case, the integral in (92) is monotonically increasing with respect to both  $a_t$ , and  $b_t$  regardless of their relative relation. Therefore, we will consider tight lower bounds of  $a_t = \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*}$  so that the relation  $b_t \geq 2a_t$  holds for large enough  $t \geq 0$ . Noting that  $\nabla L_* = 2P_* A_{c,*} \Sigma_*$  by Lemma 9 and  $P_* \succ 0$ ,  $\Sigma_* \succ 0$ , we will consider two cases.

**1. Singular  $A_{c,*}$  :** In this case, the Jacobian matrix  $\nabla L_*$  becomes singular as well. Then, we can bound  $a_t$  from below as  $a_t = \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*} \geq \sqrt{\lambda_{\min,t}} \frac{\|\nabla L_*\|_F}{\lambda_{\max,t} r_*} = \sqrt{\frac{\lambda_{\min,t}}{\lambda_{\max,t}}} \frac{\|r_*^{-1} \nabla L_*\|_F}{\sqrt{\lambda_{\max,t}}}$ . Furthermore, choosing  $T_w = O((\sqrt{T})^{1+o(1)})$ , we can use upper bounds for  $\frac{\lambda_{\max,t}}{\lambda_{\min,t}}$  and  $\lambda_{\max,t}$  from Lemma 10 to write down,  $a_t \geq \frac{T_w}{\sqrt{T} \log T} \frac{\|\nabla L_*\|_F}{Cr_*} =: a_{1,T}$  and  $b_t \geq \sqrt{\frac{T_w}{\log T}} \frac{\epsilon_*}{\sqrt{C}} =: b_{1,T}$  for all  $T_r < t \leq T$  under the event  $E_T$  for large enough  $T$ . Therefore, replacing  $a_t$  and  $b_t$  with  $a_{1,T}$  and  $b_{1,T}$  in (92) gives a lower bound to (92). Noting that the ratio  $\frac{b_{1,T}}{a_{1,T}} = \sqrt{\frac{T \log T}{T_w}} \frac{\epsilon_* r_* \sqrt{C}}{\|\nabla L_*\|_F}$  can be made to be greater than or equal to 2 by an appropriate choice of  $T_w$  leading to the case (iii) bound

$$p_t^{\text{opt}} \geq \int_0^{a_{1,T}^2} \left[ Q \left( 1 + a_{1,T} - \sqrt{a_{1,T}^2 - x} \right) - Q \left( 1 + a_{1,T} + \sqrt{a_{1,T}^2 - x} \right) \right] f_{n^2-1}(x) dx \quad (99)$$

for all  $T_r < t \leq T$  for large enough  $T$ .

**2. Nonsingular  $A_{c,*}$  :** In this case, the Jacobian matrix  $\nabla L_*$  becomes nonsingular as well. Then, we can bound  $a_t$  from below as  $a_t = \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*} \geq \sigma_{\min,*} \frac{\|\sqrt{F_t}\|_F}{\lambda_{\max,t} r_*} \geq \frac{\sigma_{\min,*}}{r_* \sqrt{\lambda_{\max,t}}}$ . Choosing  $T_w = O((\sqrt{T})^{1+o(1)})$ , we can use the upper bound for  $\lambda_{\max,t}$  from Lemma 10 to write the lower bound,  $a_t \geq \sqrt{\frac{T_w}{\log T}} \frac{\min(\sigma_{\min,*}, \epsilon_* r_*/2)}{\sqrt{Cr_*}} =: a_{2,T}$  and  $b_t \geq \sqrt{\frac{T_w}{\log T}} \frac{\epsilon_*}{\sqrt{C}} =: b_{2,T}$  for all  $T_r < t \leq T$  under the event  $E_T$  for large enough  $T$ . Therefore, replacing  $a_t$  and  $b_t$  with  $a_{2,T}$  and  $b_{2,T}$  in (92) gives a lower

bound to (92) for  $T_r < t \leq T$ . Noting that the ratio  $\frac{\beta_{2,T}}{a_{2,T}} = \frac{\epsilon_* r_*}{\min(\sigma_{\min,*}, \epsilon_* r_*/2)} = \max\left(\frac{\epsilon_* r_*}{\sigma_{\min,*}}, 2\right) \geq 2$ , we can use the case (iii) bound

$$p_t^{\text{opt}} \geq \int_0^{a_{2,T}^2} \left[ Q\left(1 + a_{2,T} - \sqrt{a_{2,T}^2 - x}\right) - Q\left(1 + a_{2,T} + \sqrt{a_{2,T}^2 - x}\right) \right] f_{n^2-1}(x) dx \quad (100)$$

for all  $T_r < t \leq T$  for large enough  $T$ .

In both cases, our focus will be on the following probability with a parameters  $a > 0$ , and  $k \in \mathbb{N}$

$$p_k(a) := \int_0^{a^2} \left[ Q(1 + a - \sqrt{a^2 - x}) - Q(1 + a + \sqrt{a^2 - x}) \right] f_k(x) dx \quad (101)$$

The following lemma summarizes some of the important properties of the function  $a \mapsto p_k(a)$ .

**Lemma 17** *The non-negative real valued function  $a \mapsto p_k(a)$  is monotonically increasing with respect to  $a \geq 0$ . Furthermore, we have that  $\frac{1}{p_k(a)} \leq \frac{1}{Q(1)} \left(1 + \frac{C_k}{a^{1/2}}\right)$  for  $a \geq ck$  for problem independent constants  $c, C > 0$ .*

**Proof** Notice that for a fixed value of  $0 \leq x \leq a^2$ , the functions  $a \mapsto 1 + a - \sqrt{a^2 - x}$  and  $a \mapsto 1 + a + \sqrt{a^2 - x}$  are monotonically decreasing and monotonically increasing, respectively. As  $Q$ -function is monotonically decreasing, the function  $a \mapsto Q(1 + a - \sqrt{a^2 - x}) - Q(1 + a + \sqrt{a^2 - x})$  is monotonically increasing for fixed  $0 \leq x \leq a^2$ . Therefore, the function  $a \mapsto p_k(a)$  is also monotonically increasing.

In order to obtain the desired asymptotic bound, let  $\epsilon \in (0, 1)$  and we can write

$$\begin{aligned} p_k(a) &= \int_0^{a^2} \left[ Q(1 + a - \sqrt{a^2 - x}) - Q(1 + a + \sqrt{a^2 - x}) \right] f_k(x) dx \\ &\geq \int_0^{\epsilon a^2} \left[ Q(1 + a - \sqrt{a^2 - x}) - Q(1 + a + \sqrt{a^2 - x}) \right] f_k(x) dx \\ &\geq \int_0^{\epsilon a^2} \min_{0 \leq x' \leq \epsilon a^2} \left[ Q(1 + a - \sqrt{a^2 - x'}) - Q(1 + a + \sqrt{a^2 - x'}) \right] f_k(x) dx \\ &= \left[ Q(1 + a(1 - \sqrt{1 - \epsilon})) - Q(1 + a(1 + \sqrt{1 - \epsilon})) \right] F_k(\epsilon a^2) \end{aligned}$$

where  $F_k(x) := 1 - \frac{\Gamma(k/2, x/2)}{\Gamma(k/2)}$  is the cumulative distribution function of chi-square distribution and  $(s, x) \mapsto \Gamma(s, x) := \int_x^\infty t^{s-1} e^{-t} dt$  and  $s \mapsto \Gamma(s) := \int_0^\infty t^{s-1} e^{-t} dt$  are upper incomplete Gamma and ordinary Gamma functions respectively. Notice that the functions  $(s, x) \mapsto \Gamma(s, x)$  and  $x \mapsto Q(x)$  are monotonically decreasing with increasing  $x > 0$ . Therefore, for large enough  $\epsilon a^2 \gg 1$  and large enough  $a \gg 1$ , we can claim that  $\Gamma(k/2, \epsilon a^2/2) \ll 1$  and  $Q(1 + a) \ll 1$  are small enough. Furthermore, for small enough  $\epsilon \ll 1$ , we can use Taylor expansion to see that  $1 - \sqrt{1 - \epsilon} = \frac{\epsilon}{2} \sum_{k=0}^\infty \frac{\epsilon^k}{2^k} (2k - 1)!! \leq c_1 \epsilon$  for a problem independent constant  $c_1 > 0$ . Then, for small enough  $\epsilon \ll 1$ , we have that

$$\begin{aligned} p_k(a) &\geq \left[ Q(1 + a(1 - \sqrt{1 - \epsilon})) - Q(1 + a(1 + \sqrt{1 - \epsilon})) \right] \left( 1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)} \right) \\ &\geq \left[ Q(1 + c_1 \epsilon a) - Q(1 + a) \right] \left( 1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)} \right) \end{aligned}$$

Furthermore, for small enough  $\epsilon a \ll 1$ , we have that  $Q(1 + c_1 \epsilon a) \geq Q(1) - c_2 \epsilon a$  by Taylor's theorem where  $c_2$  is a problem independent constant. Using these bounds, we can bound the inverse of  $p_k(a)$  from above for small enough  $\epsilon \ll 1$ , small enough  $\epsilon a \ll 1$ , large enough  $a \gg 1$  and large enough  $\epsilon a^2 \gg 1$  as

$$\begin{aligned} \frac{1}{p_k(a)} &\leq \frac{1}{Q(1) - c_2 \epsilon a - Q(1 + a)} \frac{1}{1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)}} \\ &= \frac{1}{Q(1)} \frac{1}{(1 - c_2 \epsilon a - Q(1 + a)) \left(1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)}\right)} \\ &\leq \frac{1}{Q(1)} \left[1 + 2C \left(c_2 \epsilon a + Q(1 + a) + \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)}\right)\right] \end{aligned} \quad (102)$$

where we used the Taylor expansion  $\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \leq 1 + Cx$  for small enough  $x \ll 1$  with  $C > 0$  being a problem independent constant.

The assumption  $\epsilon a^2 \gg 1$  can be used to write the asymptotic expansion of incomplete Gamma function  $\Gamma(k/2, \epsilon a^2/2) = (\epsilon a^2/2)^{k/2-1} e^{-\epsilon a^2/2} [1 + O((\epsilon a^2/2)^{-1})]$ . Noting that the Q function is always bounded as  $Q(1 + a) \leq \frac{e^{-(1+a)^2}}{\sqrt{2\pi(1+a)}}$ , we claim that choosing  $\epsilon = \frac{k}{2ea^{1+1/2}}$ , for  $\alpha \geq c''k$  with a constant  $c'' > 0$  guarantees that  $\epsilon a = \frac{k}{2e} a^{-1/2} \ll 1$  and  $\epsilon a^2 = \frac{k}{2e} a^{1-1/2} \gg 1$ . Therefore, the upper bound (102) is valid for  $\alpha \geq c''k$ . Furthermore, the term  $\epsilon a$  decays slower than both  $Q(1 + a)$  and  $\frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)}$  and thus  $\epsilon a$  dominates as

$$\frac{1}{p_k(a)} \leq \frac{1}{Q(1)} \left(1 + \frac{Ck}{2e} a^{-1/2}\right)$$

for a problem independent constant  $C > 0$ . ■

Based on Lemma 17, the integrals in (99) and (100) are asymptotically constant if both  $a_{1,T}$  and  $a_{2,T}$  are asymptotically large enough. This can be achieved if  $a_{1,T} = \frac{T_w}{\sqrt{T} \log T} \frac{\|\nabla L_*\|_F}{Cr_*} = \omega(1)$  for singular  $A_{c,*}$  and  $a_{2,T} = \sqrt{\frac{T_w}{\log T} \frac{\min(\sigma_{\min,*}, \epsilon_* r_*/2)}{\sqrt{Cr_*}}} = \omega(1)$  for non-singular  $A_{c,*}$ . In other words, choosing  $T_w = n^2 \omega(\sqrt{T} \log T)$  for singular  $A_{c,*}$  and  $T_w = n^2 \omega(\log T)$  for non-singular  $A_{c,*}$  yields the desired bound

$$p_t^{\text{opt}} \geq \frac{Q(1)}{1 + o(1)}.$$

for  $T_r < t \leq T$  for large enough  $T$ . Combined with the upper  $T_w = O((\sqrt{T})^{1+o(1)})$ , the proposed choices of  $T_w$  satisfy the asymptotic conditions.

## Appendix E. Regret Decomposition

Denote the optimal expected average cost of an LQR system  $\Theta$  with process noise covariance  $W$  by  $J_*(\Theta, W) = \text{tr}(P(\Theta)W)$ . Note that during the initial exploration period, we have that  $u_t = \bar{u}_t + \nu_t$



for  $t \leq T_w$  and after the initial exploration, we have that  $u_t = \bar{u}_t$  for  $t > T_w$  where we denote by  $\bar{u}_t := K(\tilde{\Theta}_t)x_t$  the optimal control action assuming the system  $\tilde{\Theta}_t$ . Since initial exploration period injects independent random perturbations through the optimal control input,  $\bar{u}_t$ , for sampled system,  $\tilde{\Theta}_t$ , the state dynamics can be reformulated in order to take the external perturbations into account by adding it to the process noise:

$$x_{t+1} = A_*x_t + B_*\bar{u}_t + \zeta_t, \quad (103)$$

where  $\bar{u}_t = K(\tilde{\Theta}_t)x_t$ ,  $\zeta_t = B_*\nu_t + w_t$  for  $t \leq T_w$ , and  $\zeta_t = w_t$  for  $t > T_w$ . We can write the regret explicitly as

$$R_T = \sum_{t=0}^T \{x_t^\top Qx_t + u_t^\top Ru_t - J_*(\Theta_*, \sigma_w^2 I)\} = R_{T_w}^{\text{exp}} + R_T^{\text{noexp}}, \quad (104)$$

where

$$R_{T_w}^{\text{exp}} := \sum_{t=0}^{T_w} (2\bar{u}_t^\top R\nu_t + \nu_t^\top R\nu_t), \quad \& \quad R_T^{\text{noexp}} := \sum_{t=0}^T \{x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\}$$

Since  $E_s \subset E_t$  for any  $0 \leq s \leq t$ , we have that

$$\begin{aligned} R_T^{\text{noexp}} \mathbb{1}_{E_T} &= \sum_{t=0}^T \{x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\} \mathbb{1}_{E_T} \\ &\leq \sum_{t=0}^T \{x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\} \mathbb{1}_{E_t}, \end{aligned} \quad (105)$$

$$R_{T_w}^{\text{exp}} \mathbb{1}_{E_T} = \sum_{t=0}^{T_w} (2\bar{u}_t^\top R\nu_t + \nu_t^\top R\nu_t) \mathbb{1}_{E_T} \leq \sum_{t=0}^{T_w} (2\bar{u}_t^\top R\nu_t + \nu_t^\top R\nu_t) \mathbb{1}_{E_t}. \quad (106)$$

From Bellman optimality equations (Bertsekas, 1995), we obtain

$$\begin{aligned} &J_*(\tilde{\Theta}_t, \text{Cov}[\zeta_t]) + x_t^\top P(\tilde{\Theta}_t)x_t \\ &= \min_u \left\{ x_t^\top Qx_t + u^\top Ru + \mathbb{E} \left[ (\tilde{A}_t x_t + \tilde{B}_t u + \zeta_t)^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t u + \zeta_t) \mid \mathcal{F}_t \right] \right\}, \\ &= x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t + \mathbb{E} \left[ (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t + \zeta_t)^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t + \zeta_t) \mid \mathcal{F}_t \right], \\ &= x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t + \mathbb{E} \left[ (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t)^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t) \mid \mathcal{F}_t \right] + \mathbb{E} \left[ \zeta_t^\top P(\tilde{\Theta}_t) \zeta_t \mid \mathcal{F}_t \right], \\ &= x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t + \mathbb{E} \left[ (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t)^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t) \mid \mathcal{F}_t \right] \\ &\quad + \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mid \mathcal{F}_t \right] - \mathbb{E} \left[ (A_* x_t + B_* \bar{u}_t)^\top P(\tilde{\Theta}_t) (A_* x_t + B_* \bar{u}_t) \mid \mathcal{F}_t \right], \\ &= x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t + \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mid \mathcal{F}_t \right] \\ &\quad + (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t)^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t \bar{u}_t) - (A_* x_t + B_* \bar{u}_t)^\top P(\tilde{\Theta}_t) (A_* x_t + B_* \bar{u}_t), \\ &= x_t^\top Qx_t + \bar{u}_t^\top R\bar{u}_t + \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mid \mathcal{F}_t \right] + \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t) \tilde{\Theta}_t^\top \bar{z}_t - \bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t) \Theta_*^\top \bar{z}_t, \end{aligned}$$

where  $\bar{z}_t^\top = [x_t^\top, \bar{u}_t^\top]$ . Rearranging the terms and subtracting the optimal expected average cost of the true system, we obtain the following for each term in (105),

$$\begin{aligned} & \{x_t^\top Q x_t + \bar{u}_t^\top R \bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\} \mathbb{1}_{E_t} \\ &= \left\{ J_*(\tilde{\Theta}_t, \text{Cov}[\zeta_t]) - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_t} + \left\{ \bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t) \Theta_*^\top \bar{z}_t - \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t) \tilde{\Theta}_t^\top \bar{z}_t \right\} \mathbb{1}_{E_t}, \\ & \quad + x_t^\top P(\tilde{\Theta}_t) x_t \mathbb{1}_{E_t} - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \mid \mathcal{F}_t \right]. \end{aligned}$$

Note that,  $\mathbb{1}_{E_t} \mathbb{1}_{E_{t+1}} = \mathbb{1}_{E_{t+1}}$  since  $E_{t+1} \subset E_t$ . Since  $P(\tilde{\Theta}_t) \succ 0$ , we obtain

$$\begin{aligned} & \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \mid \mathcal{F}_t \right] \\ &= \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \left( \mathbb{1}_{E_{t+1}} + \mathbb{1}_{E_{t+1}^c} \right) \mid \mathcal{F}_t \right], \\ &= \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right] + \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \mathbb{1}_{E_{t+1}^c} \mid \mathcal{F}_t \right], \\ &\geq \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right], \\ &= \mathbb{E} \left[ x_{t+1}^\top \left( P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1}) \right) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right] + \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \{x_t^\top Q x_t + \bar{u}_t^\top R \bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\} \mathbb{1}_{E_t} &\leq \left\{ J_*(\tilde{\Theta}_t, \text{Cov}[\zeta_t]) - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_t}, \\ & \quad + \left\{ \bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t) \Theta_*^\top \bar{z}_t - \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t) \tilde{\Theta}_t^\top \bar{z}_t \right\} \mathbb{1}_{E_t}, \\ & \quad + \left\{ x_t^\top P(\tilde{\Theta}_t) x_t \mathbb{1}_{E_t} - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right] \right\}, \\ & \quad + \mathbb{E} \left[ x_{t+1}^\top \left( P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right] \quad (107) \end{aligned}$$

Notice that  $\text{Cov}[\zeta_t] = \sigma_\nu^2 B_* B_*^\top + \sigma_w^2 I$  for  $t \leq T_w$  and  $\text{Cov}[\zeta_t] = \sigma_w^2 I$  for  $t > T_w$  and therefore

$$J_*(\tilde{\Theta}_t, \text{Cov}[\zeta_t]) = \text{tr}(P(\tilde{\Theta}_t) \text{Cov}[\zeta_t]) = \begin{cases} \sigma_\nu^2 \text{tr}(P(\tilde{\Theta}_t) B_* B_*^\top) + \sigma_w^2 \text{tr}(P(\tilde{\Theta}_t)) & t \leq T_w \\ \sigma_w^2 \text{tr}(P(\tilde{\Theta}_t)) & t > T_w \end{cases} \quad (108)$$

Summing the terms in (107) upto time T and adding the  $R_{T_w}^{\text{exp}}$  term, we obtain

$$R_T \mathbb{1}_{E_T} = R_{T_w}^{\text{exp}} \mathbb{1}_{E_T} + R_T^{\text{noexp}} \mathbb{1}_{E_T} \leq R_{T_w}^{\text{exp},1} + R_{T_w}^{\text{exp},2} + R_T^{\text{TS}} + R_T^{\text{RLS}} + R_T^{\text{mart}} + R_T^{\text{gap}} \quad (109)$$

where

$$R_{T_w}^{\text{exp},1} = \sum_{t=0}^{T_w} (2\bar{u}_t^\top R \nu_t + \nu_t^\top R \nu_t) \mathbb{1}_{E_t}, \quad (110)$$

$$R_{T_w}^{\text{exp},2} = \sum_{t=0}^{T_w} \sigma_\nu^2 \text{tr}(P(\tilde{\Theta}_t) B_* B_*^\top) \mathbb{1}_{E_t}, \quad (111)$$

$$R_T^{\text{TS}} = \sum_{t=0}^T \left\{ J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_t}, \quad (112)$$

$$R_T^{\text{RLS}} = \sum_{t=0}^T \left\{ \bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t) \Theta_*^\top \bar{z}_t - \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t) \tilde{\Theta}_t^\top \bar{z}_t \right\} \mathbb{1}_{E_t}, \quad (113)$$

$$R_T^{\text{mart}} = \sum_{t=0}^T \left\{ x_t^\top P(\tilde{\Theta}_t) x_t \mathbb{1}_{E_t} - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right] \right\}, \quad (114)$$

$$R_T^{\text{gap}} = \sum_{t=0}^T \mathbb{E} \left[ x_{t+1}^\top \left( P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right]. \quad (115)$$

In the next section, we will give upper bounds to each term.

## Appendix F. Regret Analysis

In this section, we bound each term in regret decomposition individually. In particular,  $R_{T_w}^{\text{exp}}$  is studied in Appendix F.1,  $R_T^{\text{RLS}}$  is studied in Appendix F.2,  $R_T^{\text{mart}}$  in Appendix F.3,  $R_T^{\text{TS}}$  in Appendix F.4, and  $R_T^{\text{gap}}$  in Appendix F.5. Finally, in Appendix F.6, we combine these results to obtain the regret upper bound of TSAC as stated in Theorem 3.

### F.1. Bounding $R_{T_w}^{\text{exp},1}$ and $R_{T_w}^{\text{exp},2}$

The following gives an upper bound on the regret attained due to isotropic perturbations in the TS with improved exploration phase of TSAC.

**Lemma 18 (Direct Effect of Improved Exploration on Regret)** *The following holds with probability at least  $1 - \delta$ ,*

$$R_{T_w}^{\text{exp},1} = \sum_{t=0}^{T_w} \{2\bar{u}_t^\top R \nu_t + \nu_t^\top R \nu_t\} \mathbb{1}_{E_t} \leq d\sigma_\nu \sqrt{B_\delta} + d\|R\|\sigma_\nu^2 \left( T_w + \sqrt{T_w} \log \frac{4dT_w}{\delta} \sqrt{\log \frac{4}{\delta}} \right)$$

where

$$B_\delta = 8 \left( 1 + T_w \kappa^2 \|R\|^2 (n+d)^{2(n+d)} \right) \log \left( \frac{4d}{\delta} \left( 1 + T_w \kappa^2 \|R\|^2 (n+d)^{2(n+d)} \right)^{1/2} \right).$$

Furthermore, we have  $R_{T_w}^{\text{exp},2} \leq \sigma_\nu^2 D \|B_*\|_F^2 T_w$ .

**Proof** First we will study  $R_{T_w}^{\text{exp},1}$ . Let  $q_t^\top = \bar{u}_t^\top R \mathbb{1}_{E_t}$ . The first term can be written as

$$2 \sum_{t=0}^{T_w} \sum_{i=1}^d q_{t,i} \nu_{t,i} = 2 \sum_{i=1}^d \sum_{t=0}^{T_w} q_{t,i} \nu_{t,i}$$

Let  $M_{t,i} = \sum_{k=0}^t q_{k,i} \nu_{k,i}$ . By Theorem 27 on some event  $G_{\delta,i}$  that holds with probability at least  $1 - \delta/(2d)$ , for any  $t \geq 0$ ,

$$M_{t,i}^2 \leq 2\sigma_\nu^2 \left( 1 + \sum_{k=0}^t q_{k,i}^2 \right) \log \left( \frac{2d}{\delta} \left( 1 + \sum_{k=0}^t q_{k,i}^2 \right)^{1/2} \right)$$

Note that  $\|q_k\| = \|R\bar{u}_k\| \mathbb{1}_{E_k} \leq \kappa \|R\| (n+d)^{n+d}$ , thus  $q_{k,i} \leq \kappa \|R\| (n+d)^{n+d}$ . Using union bound we get, for probability at least  $1 - \frac{\delta}{2}$ ,

$$\begin{aligned} & \sum_{t=0}^{T_w} 2\nu_t^\top R \bar{u}_t \mathbb{1}_{R_t} \leq \\ & d \sqrt{8\sigma_\nu^2 \left( 1 + T_w \kappa^2 \|R\|^2 (n+d)^{2(n+d)} \right) \log \left( \frac{4d}{\delta} \left( 1 + T_w \kappa^2 \|R\|^2 (n+d)^{2(n+d)} \right)^{1/2} \right)} \quad (116) \end{aligned}$$

Let  $W = \sigma_\nu \sqrt{2d \log \frac{4dT_w}{\delta}}$ . Define  $\Psi_t = \nu_t^\top R \nu_t - \mathbb{E}[\nu_t^\top R \nu_t | \mathcal{F}_{t-1}]$  and its truncated version  $\tilde{\Psi}_t = \Psi_t \mathbb{1}_{\{\Psi_t \leq 2DW^2\}}$ .

$$\begin{aligned} \mathbb{P}\left(\sum_{t=1}^{T_w} \Psi_t > 2\|R\|W^2 \sqrt{2T_w \log \frac{4}{\delta}}\right) &\leq \\ &\mathbb{P}\left(\max_{1 \leq t \leq T_w} \Psi_t > 2\|R\|W^2\right) + \mathbb{P}\left(\sum_{t=1}^{T_w} \tilde{\Psi}_t > 2\|R\|W^2 \sqrt{2T_w \log \frac{4}{\delta}}\right) \end{aligned}$$

Using Lemma 30 with union bound and Theorem 28, summation of terms on the right hand side is bounded by  $\delta/2$ . Thus, with probability at least  $1 - \delta/2$ ,

$$\sum_{t=0}^{T_w} \nu_t^\top R \nu_t \leq dT_w \sigma_\nu^2 \|R\| + 2\|R\|W^2 \sqrt{2T_w \log \frac{4}{\delta}}. \quad (117)$$

Combining (116) and (117) gives the statement of lemma for the regret of external exploration noise. Next, we consider  $R_{T_w}^{\text{exp},2}$ . Due to rejection sampling  $\mathcal{R}_S(\cdot)$ , a new model sample is redrawn until it lies on the set  $\mathcal{S}$  at every TS step, *i.e.*,  $\tilde{\Theta}_t \in \mathcal{S}$  for every time step  $t \geq 0$ . By Assumption 1, we have  $\|P(\tilde{\Theta}_t)\|_F \leq D = \bar{\alpha}\gamma^{-1}\kappa^2(1 + \kappa^2)$ . Thus, we have

$$\begin{aligned} R_{T_w}^{\text{exp},2} &= \sum_{t=0}^{T_w} \sigma_\nu^2 \text{tr}(P(\tilde{\Theta}_t) B_* B_*^\top) \mathbb{1}_{E_t}, \\ &\leq \sum_{t=0}^{T_w} \sigma_\nu^2 \|P(\tilde{\Theta}_t)\|_F \|B_*\|_F^2 \mathbb{1}_{E_t}, \\ &\leq \sigma_\nu^2 D \|B_*\|_F^2 \sum_{t=0}^{T_w} \mathbb{1}_{E_t} \leq \sigma_\nu^2 D \|B_*\|_F^2 T_w. \end{aligned} \quad (118)$$

■

## F.2. Bounding $R_T^{\text{RLS}}$

Bounding this term is achieved by manipulating the similar bounds in Abeille and Lazaric (2017); Abbasi-Yadkori and Szepesvári (2011) to our setting and TS algorithm. We first have the following result from regularized least squares estimate.

**Lemma 19** *On the event of  $E_T$ , for  $X_s = \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma} \sqrt{2n \log(n(T - T_w)/\delta)}$ , we have,*

$$\begin{aligned} \sum_{t=0}^T \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 &\leq 2(\beta_T(\delta) + v_T(\delta))^2 \left( \left( 1 + \frac{(1 + \kappa^2)(n + d)^{2(n+d)}}{\mu} \right)^{\tau_0+1} \log \frac{\det(V_{T_r})}{\det(\mu I)} \right. \\ &\quad \left. + \left( 1 + \frac{(1 + \kappa^2)X_s^2}{\mu} \right)^{\tau_0+1} \log \frac{\det(V_T)}{\det(V_{T_r})} \right). \end{aligned}$$

**Proof** Let  $\tau \leq t$  be the last time step before  $t$ , when the policy was updated. Using Cauchy-Schwarz inequality, we have:

$$\sum_{t=0}^T \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \sum_{t=0}^T \|V_t^{1/2}(\tilde{\Theta}_t - \Theta_*)\|^2 \|z_t\|_{V_t^{-1}}^2 \leq \sum_{t=0}^T \frac{\det(V_t)}{\det(V_\tau)} \|V_\tau^{1/2}(\tilde{\Theta}_\tau - \Theta_*)\|^2 \|z_t\|_{V_t^{-1}}^2. \quad (119)$$

Note that  $t - \tau \leq \tau_0$  due to policy update rule. Moreover, we have

$$\det(V_t) = \det(V_\tau) \prod_{i=0}^{t-\tau} (1 + \|z_t\|_{V_{t-i}^{-1}}^2) \leq \det(V_\tau) \left(1 + \frac{\|z_t\|^2}{\mu}\right)^{\tau_0}.$$

Combining this with (119), on the event of  $E_T$ , for  $t \leq T_r$ , we have:

$$\sum_{t=0}^{T_r} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \sum_{t=0}^{T_r} \left(1 + \frac{(1 + \kappa^2)(n + d)^{2(n+d)}}{\mu}\right)^{\tau_0} \|V_\tau^{1/2}(\tilde{\Theta}_\tau - \Theta_*)\|^2 \|z_t\|_{V_t^{-1}}^2 \quad (120)$$

$$\leq \sum_{t=0}^{T_r} \left(1 + \frac{(1 + \kappa^2)(n + d)^{2(n+d)}}{\mu}\right)^{\tau_0} (\beta_T(\delta) + v_T(\delta))^2 \|z_t\|_{V_t^{-1}}^2, \quad (121)$$

$$\leq \frac{2(1 + \kappa^2)(n + d)^{2(n+d)}}{\mu} \left(1 + \frac{(1 + \kappa^2)(n + d)^{2(n+d)}}{\mu}\right)^{\tau_0} (\beta_T(\delta) + v_T(\delta))^2 \log \left(\frac{\det(V_{T_r})}{\det(\mu I)}\right) \quad (122)$$

where in (121) we used the fact that on the event of  $E_T$ , using triangle inequality, we have  $\|\tilde{\Theta}_\tau - \Theta_*\|_{V_\tau} \leq \|\tilde{\Theta}_\tau - \hat{\Theta}_\tau\|_{V_\tau} + \|\hat{\Theta}_\tau - \Theta_*\|_{V_\tau} \leq v_\tau(\delta) + \beta_\tau(\delta) \leq v_T(\delta) + \beta_T(\delta)$  and in (122) we used the upper bound of  $\|z_t\|_{V_t^{-1}}$  to utilize Lemma 10 of [Abbasi-Yadkori and Szepesvári \(2011\)](#). Similarly, on the even of  $E_t$ , for  $t > T_r$ , we get:

$$\sum_{t=T_r+1}^T \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \frac{2(1 + \kappa^2)X_s^2}{\mu} \left(1 + \frac{(1 + \kappa^2)X_s^2}{\mu}\right)^{\tau_0} (\beta_T(\delta) + v_T(\delta))^2 \log \left(\frac{\det(V_T)}{\det(V_{T_r})}\right)$$

■

**Lemma 20 (Bounding  $R_T^{\text{RLS}}$  for TSAC)** Let  $R_T^{\text{RLS}}$  be as defined by (113). Under the event of  $E_T$ , setting  $\mu = (1 + \kappa^2)X_s^2$ , we have

$$|R_T^{\text{RLS}}| = \tilde{O} \left( (n + d)^{(\tau_0+2)(n+d)+1.5} \sqrt{n} \sqrt{T_r} + (n + d)n \sqrt{T - T_r} \right).$$

**Proof**

$$|R_T^{\text{RLS}}| \leq \sum_{t=0}^T \left| \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\|^2 \right| \quad (123)$$

$$\begin{aligned} &= \sum_{t=0}^{T_r} \left| \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\|^2 \right| + \sum_{t=T_r}^T \left| \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\|^2 \right| \\ &\leq \left( \sum_{t=0}^{T_r} \left( \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\| - \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\| \right)^2 \right)^{\frac{1}{2}} \left( \sum_{t=0}^{T_r} \left( \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\| \right)^2 \right)^{\frac{1}{2}} \\ &+ \left( \sum_{t=T_r}^T \left( \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\| - \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\| \right)^2 \right)^{\frac{1}{2}} \left( \sum_{t=T_r}^T \left( \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\| \right)^2 \right)^{\frac{1}{2}} \end{aligned} \quad (124)$$

$$\begin{aligned} &\leq \left( \sum_{t=0}^{T_r} \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} (\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{\frac{1}{2}} \left( \sum_{t=0}^{T_r} \left( \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\| \right)^2 \right)^{\frac{1}{2}} \\ &+ \left( \sum_{t=T_r}^T \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} (\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{\frac{1}{2}} \left( \sum_{t=T_r}^T \left( \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{\frac{1}{2}} \Theta_*^\top z_t \right\| \right)^2 \right)^{\frac{1}{2}} \end{aligned} \quad (125)$$

where (123) and (125) follow from triangle inequality, and (124) follows from Cauchy Schwarz inequality. Note that for  $t \leq T_r$ , we have  $\|z_t\|^2 \leq (1 + \kappa^2)(n + d)^{2(n+d)}$  and for  $t > T_r$  we have  $\|z_t\|^2 \leq (1 + \kappa^2)X_s^2$ . Moreover, since  $\tilde{\Theta}$  belongs to  $\mathcal{S}$  by construction of the rejection sampling, we get

$$\begin{aligned} |R_T^{\text{RLS}}| &\leq \left( D \sum_{t=0}^{T_r} \left\| (\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{\frac{1}{2}} \sqrt{4T_r D (1 + \kappa^2) S^2 (n + d)^{2(n+d)}} \\ &+ \left( D \sum_{t=T_r}^T \left\| (\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{\frac{1}{2}} \sqrt{4(T - T_r) D (1 + \kappa^2) S^2 X_s^2} \\ &\leq \frac{\sqrt{8T_r} D S (1 + \kappa^2) (n + d)^{2(n+d)} (\beta_T(\delta) + v_T(\delta))}{\sqrt{\mu}} \left( 1 + \frac{(1 + \kappa^2)(n + d)^{2(n+d)}}{\mu} \right)^{\frac{\tau_0}{2}} \sqrt{\log \left( \frac{\det(V_{T_r})}{\det(\mu I)} \right)} \\ &+ \frac{\sqrt{8(T - T_r)} D S (1 + \kappa^2) X_s^2 (\beta_T(\delta) + v_T(\delta))}{\sqrt{\mu}} \left( 1 + \frac{(1 + \kappa^2) X_s^2}{\mu} \right)^{\frac{\tau_0}{2}} \sqrt{\log \left( \frac{\det(V_T)}{\det(V_{T_r})} \right)} \end{aligned} \quad (126)$$

From Lemma 10 of [Abbasi-Yadkori and Szepesvári \(2011\)](#), we have that  $\log \left( \frac{\det(V_{T_r})}{\det(\mu I)} \right) \leq (n + d) \log \left( 1 + \frac{T_r (1 + \kappa^2) (n + d)^{2(n+d)}}{\mu (n + d)} \right)$  and  $\log \left( \frac{\det(V_T)}{\det(V_{T_r})} \right) \leq (n + d) \log \left( 1 + \frac{T_r (1 + \kappa^2) (n + d)^{2(n+d)} + (T - T_r) X_s^2}{\mu (n + d)} \right)$ .

After inserting these quantities into (126), we have the dimension dependency of  $(n + d)^{2(n+d)} \times \sqrt{n(n + d)} \times (n + d)^{(n+d)\tau_0} \times (n + d)$  on the first term where  $\sqrt{n(n + d)}$  is due to  $\beta_T(\delta) + v_T(\delta)$ .

For the second term, for large enough  $T$ , we have the dimension dependency of  $n \times \sqrt{n(n+d)} \times n^{(\tau_0/2)} \times \sqrt{n+d}$ , where  $n$  comes from  $X_s^2$ . Thus, we achieve the following bound for  $|R_T^{\text{RLS}}|$ :

$$|R_T^{\text{RLS}}| = \tilde{O} \left( (n+d)^{(\tau_0+2)(n+d)+1.5} \sqrt{n} \sqrt{T_r} + (n+d) n^{1.5+\tau_0/2} \sqrt{T-T_r} \right).$$

With the choice of  $\mu = (1+\kappa^2)X_s^2$ , the dependency of  $n^{(\tau_0/2)}$  on the second term can be converted to a scalar multiplier of  $\sqrt{2}^{\tau_0}$  and reduces the dependency of  $X_s^2$  to  $X_s$ , which gives the advertised bound.  $\blacksquare$

### F.3. Bounding $R_T^{\text{mart}}$

Notice that this term is very similar to corresponding term in [Abbasi-Yadkori and Szepesvári \(2011\)](#); [Abeille and Lazaric \(2018\)](#), besides the difference of early improved exploration. Following the same analysis, while including the effect of improved exploration gives the upper bound on  $R_T^{\text{mart}}$ . A similar analysis is also conducted in [Lale et al. \(2022\)](#), yet we provide it for completeness.

**Lemma 21 (Bounding  $R_T^{\text{mart}}$ )** *Let  $R_T^{\text{mart}}$  be as defined by (114). Under the event of  $E_T$ , with probability at least  $1 - \delta$ , for  $t > T_r$ , we have*

$$\begin{aligned} R_T^{\text{mart}} &\leq k_{s,1}(n+d)^{n+d}(\sigma_w + \|B_*\|\sigma_\nu)n\sqrt{T_r} \log((n+d)T_r/\delta) \\ &\quad + \frac{k_{s,2}(12\kappa^2 + 2\kappa\sqrt{2})}{\gamma} \sigma_w^2 n \sqrt{n} \sqrt{T-T_w} \log(n(T-T_w)/\delta) \\ &\quad + k_{s,3} n \sigma_w^2 \sqrt{T-T_w} \log(nT/\delta) + k_{s,4} n (\sigma_w + \|B_*\|\sigma_\nu)^2 \sqrt{T_w} \log(nT/\delta), \end{aligned}$$

for some problem dependent coefficients  $k_{s,1}, k_{s,2}, k_{s,3}, k_{s,4}$ .

**Proof** Let  $f_t = A_* x_t + B_* u_t$ . One can decompose  $R_T^{\text{mart}}$  as

$$R_1 = x_0^\top P(\tilde{\Theta}_0)x_0 - x_{T+1}^\top P(\tilde{\Theta}_{T+1})x_{T+1} + \sum_{t=1}^T x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_t^\top P(\tilde{\Theta}_t)x_t | \mathcal{F}_{t-2} \right]$$

Since  $P(\tilde{\Theta}_0)$  is positive semidefinite and  $x_0 = 0$ , the first two terms are bounded above by zero. Recall that  $\zeta_t = B_* \nu_t + w_t$  for  $t \leq T_w$ , and  $\zeta_t = w_t$  for  $t > T_w$ . The second term is decomposed as follows

$$\begin{aligned} &\sum_{t=1}^T x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_t^\top P(\tilde{\Theta}_t)x_t | \mathcal{F}_{t-2} \right] \\ &= \sum_{t=1}^T f_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} + \sum_{t=1}^T \left( \zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} - \mathbb{E} \left[ \zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} | \mathcal{F}_{t-2} \right] \right) \end{aligned}$$

Let  $R_{1,1} = \sum_{t=1}^T f_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1}$ ,  $R_{1,2} = \sum_{t=1}^T \left( \zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} - \mathbb{E} \left[ \zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} | \mathcal{F}_{t-2} \right] \right)$ , and  $v_{t-1}^\top = f_{t-1}^\top P(\tilde{\Theta}_t)$ . Then one can write  $R_{1,1}$ . Let can be written as

$$R_{1,1} = \sum_{t=1}^T \sum_{i=1}^n v_{t-1,i} \zeta_{t-1,i} = \sum_{i=1}^n \sum_{t=1}^T v_{t-1,i} \zeta_{t-1,i}.$$



Let  $M_{t,i} = \sum_{k=1}^t v_{k-1,i} \zeta_{k-1,i}$ . By Theorem 27 on some event  $G_{\delta,i}$  that holds with probability at least  $1 - \delta/(2n)$ , for any  $t \geq 0$ ,

$$M_{t,i}^2 \leq 2(\sigma_w^2 + \|B_*\|^2 \sigma_\nu^2) \left( 1 + \sum_{k=1}^{T_r} v_{k-1,i}^2 \right) \log \left( \frac{2n}{\delta} \left( 1 + \sum_{k=1}^{T_r} v_{k-1,i}^2 \right)^{1/2} \right) \\ + 2\sigma_w^2 \left( 1 + \sum_{k=T_r+1}^t v_{k-1,i}^2 \right) \log \left( \frac{2n}{\delta} \left( 1 + \sum_{k=T_r+1}^t v_{k-1,i}^2 \right)^{1/2} \right) \quad \text{for } t > T_r.$$

Notice that TSAC stops additional isotropic perturbation after  $t = T_w$ , and the state starts decaying until  $t = T_r$ . For simplicity of presentation we treat the time between  $T_w$  and  $T_r$  as TS with improved exploration while sacrificing the tightness of the result. On  $E_T$ ,  $\|v_k\| \leq DS(n+d)^{n+d} \sqrt{1+\kappa^2}$  for  $k \leq T_r$  and  $\|v_k\| \leq \frac{(12\kappa^2+2\kappa\sqrt{2})DS\sigma_w\sqrt{1+\kappa^2}}{\gamma} \sqrt{2n \log(n(t-T_w)/\delta)}$  for  $k > T_r$ . Thus,  $v_{k,i} \leq DS(n+d)^{n+d} \sqrt{1+\kappa^2}$  and  $v_{k,i} \leq \frac{(12\kappa^2+2\kappa\sqrt{2})DS\sigma_w\sqrt{1+\kappa^2}}{\gamma} \sqrt{2n \log(n(t-T_w)/\delta)}$  respectively for  $k \leq T_r$  and  $k > T_r$ . Using union bound we get, for probability at least  $1 - \frac{\delta}{2}$ , for  $t > T_r$ ,

$$R_{1,1} \leq n \sqrt{2(\sigma_w^2 + \|B_*\|^2 \sigma_\nu^2) (1 + T_r D^2 S^2 (n+d)^{2(n+d)} (1 + \kappa^2))} \times \\ \sqrt{\log \left( \frac{4n}{\delta} (1 + T_r D^2 S^2 (n+d)^{2(n+d)} (1 + \kappa^2))^{1/2} \right)} \\ + n \sqrt{2\sigma_w^2 \left( 1 + \frac{2(t-T_r)(12\kappa^2+2\kappa\sqrt{2})^2 D^2 S^2 n \sigma_w^2 (1 + \kappa^2)}{\gamma^2} \log(n(T-T_w)/\delta) \right)} \times \\ \sqrt{\log \left( \frac{4n}{\delta} \left( 1 + \frac{2(t-T_r)(12\kappa^2+2\kappa\sqrt{2})^2 D^2 S^2 n \sigma_w^2 (1 + \kappa^2)}{\gamma^2} \log(n(T-T_w)/\delta) \right) \right)}.$$

Let  $\mathcal{W}_{exp} = (\sigma_w + \|B_*\| \sigma_\nu) \sqrt{2n \log \frac{4nT}{\delta}}$  and  $\mathcal{W}_{noexp} = \sigma_w \sqrt{2n \log \frac{4nT}{\delta}}$ . Define  $\Psi_t = \zeta_{t-1}^\top P(\tilde{\Theta}_t) \zeta_{t-1} - \mathbb{E} \left[ \zeta_{t-1}^\top P(\tilde{\Theta}_t) \zeta_{t-1} | \mathcal{F}_{t-2} \right]$  and its truncated version  $\tilde{\Psi}_t = \Psi_t \mathbb{1}_{\{\Psi_t \leq 2DW_{exp}^2\}}$  for  $t \leq T_w$  and  $\tilde{\Psi}_t = \Psi_t \mathbb{1}_{\{\Psi_t \leq 2DW_{noexp}^2\}}$  for  $t > T_w$ . Notice that  $R_{1,2} = \sum_{t=1}^T \Psi_t$ .

$$\mathbb{P} \left( \sum_{t=1}^{T_w} \Psi_t > 2DW_{exp}^2 \sqrt{2T_w \log \frac{4}{\delta}} \right) + \mathbb{P} \left( \sum_{t=T_w+1}^T \Psi_t > 2DW_{noexp}^2 \sqrt{2(T-T_w) \log \frac{4}{\delta}} \right) \\ \leq \mathbb{P} \left( \max_{1 \leq t \leq T_w} \Psi_t > 2DW_{exp}^2 \right) + \mathbb{P} \left( \max_{T_w+1 \leq t \leq T} \Psi_t > 2DW_{noexp}^2 \right) \\ + \mathbb{P} \left( \sum_{t=1}^{T_w} \tilde{\Psi}_t > 2DW_{exp}^2 \sqrt{2T_w \log \frac{4}{\delta}} \right) + \mathbb{P} \left( \sum_{t=T_w+1}^T \tilde{\Psi}_t > 2DW_{noexp}^2 \sqrt{2(T-T_w) \log \frac{4}{\delta}} \right)$$

By Lemma 30 with union bound and Theorem 28, summation of terms on the right hand side is bounded by  $\delta/2$ . Thus, with probability at least  $1 - \delta/2$ , for  $t > T_w$ ,

$$R_{1,2} \leq 4nD\sigma_w^2 \sqrt{2(t - T_w) \log \frac{4}{\delta} \log \frac{4nT}{\delta}} + 4nD(\sigma_w + \|B_*\|\sigma_\nu)^2 \sqrt{2T_w \log \frac{4}{\delta} \log \frac{4nT}{\delta}}.$$

Combining  $R_{1,1}$  and  $R_{1,2}$  gives the statement.  $\blacksquare$

#### F.4. Bounding $R_T^{\text{TS}}$

**Lemma 22 (Bounding  $R_T^{\text{TS}}$  for TSAC)** *Let  $R_T^{\text{TS}}$  be as defined by (112). Under the event of  $E_T$ , we have that*

$$|R_T^{\text{TS}}| \leq \tilde{O} \left( \sqrt{n}T_w + \text{poly}(n, d, \log(1/\delta))\sqrt{T - T_w} \right).$$

with probability at least  $1 - 2\delta$  if  $T_w = \omega(\sqrt{T \log T})$  for singular  $A_{c,*}$  and  $T_w = \omega(\log T)$  for non-singular  $A_{c,*}$ .

**Proof** We decompose  $R_T^{\text{TS}}$  into two pieces as

$$R_T^{\text{TS}} = \underbrace{\sum_{t=0}^{T_w} \left\{ J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_t}}_{R_{T_w}^{\text{TS,exp}}} + \underbrace{\sum_{t=T_w+1}^T \left\{ J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_t}}_{R_T^{\text{TS,noexp}}}$$

Since every sampled system is in set  $\mathcal{S}$ , we have that  $\|P(\tilde{\Theta}_t)\|_F \leq D$  and therefore

$$\begin{aligned} R_{T_w}^{\text{TS,exp}} &\leq \sum_{t=0}^{T_w} \left| J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I) \right| \mathbb{1}_{E_t} \\ &\leq \sum_{t=0}^{T_w} \left( \left| J_*(\tilde{\Theta}_t, \sigma_w^2 I) \right| + \left| J_*(\Theta_*, \sigma_w^2 I) \right| \right) \end{aligned} \quad (127)$$

$$\leq \sqrt{n}\sigma_w^2 \sum_{t=0}^{T_w} \left( \|P(\tilde{\Theta}_t)\|_F + \|P(\Theta_*)\|_F \right) \leq 2\sqrt{n}\sigma_w^2 D T_w \quad (128)$$

where we used the relation  $\text{tr}(P) \leq \sqrt{n}\|P\|_F$  in (127). Considering the number of times a new TS sample is drawn, the second term in  $R_T^{\text{TS}}$  can be written as

$$R_K^{\text{TS,noexp}} = \sum_{k=0}^K \tau_0 \left\{ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_{t_k}}$$

where  $t_k = T_w + 1 + k\tau_0$  and  $K = \left\lceil \frac{T-T_w}{\tau_0} \right\rceil$ . Denoting the information available to the controller up to time  $t \geq 0$  via  $\mathcal{F}_t^{\text{cnt}} := \sigma(\mathcal{F}_{t-1}, x_t)$ ,  $R_K^{\text{TS,noexp}}$  can be further decomposed into two pieces as

$$\begin{aligned} R_K^{\text{TS,noexp}} &= \underbrace{\sum_{k=0}^K \tau_0 \left\{ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) - \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right\}}_{R_K^{\text{TS},1}} \mathbb{1}_{E_t}, \\ &\quad + \underbrace{\sum_{k=0}^K \tau_0 \left\{ \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] - J_*(\Theta_*, \sigma_w^2 I) \right\}}_{R_K^{\text{TS},2}} \mathbb{1}_{E_{t_k}} \end{aligned}$$

We will investigate each term in order under the event of  $E_T$ .

**Bounding  $R_K^{\text{TS},1}$ .** Notice that  $\{R_K^{\text{TS},1}\}_{K \geq 0}$  is a martingale sequence with  $\left| R_K^{\text{TS},1} - R_{K-1}^{\text{TS},1} \right| \leq 2\tau_0 \sigma_w^2 \sqrt{n}D$ . Therefore it can be bounded by Azuma's inequality (Lemma 28) w.p. at least  $1 - \delta$  as

$$R_K^{\text{TS},1} \leq \sigma_w^2 D \sqrt{8n\tau_0^2 K \log(2/\delta)} \leq \sigma_w^2 D \sqrt{8n\tau_0(T - T_w) \log(2/\delta)} \quad (129)$$

**Bounding  $R_K^{\text{TS},2}$ .** Denoting by  $\mathcal{S}^{\text{opt}} := \{\Theta \in \mathbb{R}^{(n+d) \times n} \mid J_*(\Theta, \sigma_w^2 I) \leq J_*(\Theta_*, \sigma_w^2 I)\}$  the set of optimistic parameters and defining  $R_k^{\text{TS},2} := \left\{ \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_{t_k}}$ . Notice that, for any  $\Theta \in \mathcal{S}^{\text{opt}}$ , we can write

$$\begin{aligned} R_k^{\text{TS},2} &\leq \left\{ \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] - J_*(\Theta, \sigma_w^2 I) \right\} \mathbb{1}_{E_{t_k}} \\ &\leq \left| J_*(\Theta, \sigma_w^2 I) - \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right| \mathbb{1}_{E_{t_k}} \end{aligned}$$

As the above bound holds for any  $\Theta \in \mathcal{S}^{\text{opt}}$ , we can replace the right hand side with an expectation over the optimistic set  $\mathcal{S}^{\text{opt}}$ . Specifically, we choose an i.i.d. copy of  $\tilde{\Theta}_{t_k}$ , that is, we choose a random variable  $\tilde{\Theta}'_{t_k}$  which has the same distribution as  $\tilde{\Theta}_{t_k}$  and independent from it. Then, we have that

$$\begin{aligned} R_k^{\text{TS},2} &\leq \mathbb{E} \left[ \left| J_*(\tilde{\Theta}'_{t_k}, \sigma_w^2 I) - \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right| \mathbb{1}_{E_{t_k}} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}, \tilde{\Theta}'_{t_k} \in \mathcal{S}^{\text{opt}} \right] \\ &= \frac{\mathbb{E} \left[ \left| J_*(\tilde{\Theta}'_{t_k}, \sigma_w^2 I) - \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right| \mathbb{1}_{E_{t_k}} \mathbb{1}_{\tilde{\Theta}'_{t_k} \in \mathcal{S}^{\text{opt}}} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right]}{\mathbb{P} \left( \tilde{\Theta}'_{t_k} \in \mathcal{S}^{\text{opt}} \mid \mathcal{F}_{t_k}^{\text{cnt}}, \hat{E}_{t_k} \right)} \end{aligned}$$

Denoting by  $p_t^{\text{opt}} = \mathbb{P} \left( \tilde{\Theta}'_t \in \mathcal{S}^{\text{opt}} \mid \mathcal{F}_t^{\text{cnt}}, \hat{E}_t \right)$  the probability of drawing cost optimistic TS samples, we can write further bounds on  $R_k^{\text{TS},2}$  as

$$\begin{aligned} R_k^{\text{TS},2} &\leq \frac{1}{p_{t_k}^{\text{opt}}} \mathbb{E} \left[ \left| J_*(\tilde{\Theta}'_{t_k}, \sigma_w^2 I) - \mathbb{E} \left[ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right| \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \\ &= \frac{\sigma_w^2}{p_{t_k}^{\text{opt}}} \mathbb{E} \left[ \left| \text{tr} \left( P(\tilde{\Theta}'_{t_k}) - \mathbb{E} \left[ P(\tilde{\Theta}_{t_k}) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right) \right| \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \\ &\leq \frac{n\sigma_w^2}{p_{t_k}^{\text{opt}}} \mathbb{E} \left[ \left\| P(\tilde{\Theta}'_{t_k}) - \mathbb{E} \left[ P(\tilde{\Theta}_{t_k}) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right\|_2 \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \quad (130) \end{aligned}$$

where we used the relation  $|\text{tr}(A)| \leq n\|A\|_2$ . Denoting  $P_k := \mathbb{E} \left[ P(\tilde{\Theta}_{t_k}) \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right]$ , the following definition will be used in the rest of the section to understand the behavior of  $R_k^{TS;2}$

$$\Delta_k := \mathbb{E} \left[ \|P(\tilde{\Theta}_{t_k}) - P_k\|_2 \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \quad (131)$$

The following lemma will be used to bound  $\Delta_k$  from above.

**Lemma 23** *For any  $\Theta \in \mathcal{S}$ , any positive definite matrix  $V \in \mathbb{R}^{(n+d) \times (n+d)}$ , and for any  $i, j \in [n]$ ,*

$$\|\nabla P_{ij}(\Theta)\|_V \leq \Gamma \|H(\Theta)\|_V,$$

where  $\Gamma \geq 0$  is a problem dependent constant.

**Proof** Let  $\delta P(\Theta, \delta\Theta)$  be the differential of  $P(\Theta)$  in the direction  $\delta\Theta$ . Then, we have that

$$\begin{aligned} \delta P(\Theta, \delta\Theta) &= A_c(\Theta)^\top \delta P(\Theta, \delta\Theta) A_c(\Theta) \\ &\quad + A_c(\Theta)^\top P(\Theta) \delta\Theta^\top H(\Theta) + H(\Theta)^\top \delta\Theta P(\Theta) A_c(\Theta) \end{aligned} \quad (132)$$

where  $A_c(\Theta) = \Theta^\top H(\Theta)$  is the closed-loop matrix. We know that  $P(\Theta)$  satisfies the Riccati equation as

$$P - A_c^\top P A_c = Q + K^\top R K \succ 0 \implies \left( P^{\frac{1}{2}} A_c P^{-\frac{1}{2}} \right)^\top P^{\frac{1}{2}} A_c P^{-\frac{1}{2}} \prec I$$

where we dropped  $\Theta$  dependence for simplicity. Therefore, similarity transformation of the closed-loop matrix  $\bar{A}_c := P^{\frac{1}{2}} A_c P^{-\frac{1}{2}}$  is a contraction, i.e.,  $\|P^{\frac{1}{2}} A_c P^{-\frac{1}{2}}\|_2 =: \sigma_\Theta < 1$ . Multiplying both sides of (132) by  $P^{-\frac{1}{2}}$  we obtain

$$\delta \bar{P}(\delta\Theta) = \bar{A}_c^\top \delta \bar{P}(\delta\Theta) \bar{A}_c + \bar{A}_c^\top P^{\frac{1}{2}} \delta\Theta^\top H P^{-\frac{1}{2}} + P^{-\frac{1}{2}} H^\top \delta\Theta P^{\frac{1}{2}} \bar{A}_c$$

where  $\delta \bar{P}(\delta\Theta) = P^{-\frac{1}{2}} \delta P(\delta\Theta) P^{-\frac{1}{2}}$ . Taking the spectral norm of both sides and using submultiplicativity of spectral norm as well as equivalence of matrix norms, we have that

$$\begin{aligned} \|\delta \bar{P}(\delta\Theta)\|_2 &\leq \|A_c\|_2^2 \|\delta \bar{P}(\delta\Theta)\|_2 + 2\|\bar{A}_c\|_2 \|P^{\frac{1}{2}} \delta\Theta^\top H P^{-\frac{1}{2}}\|_2 \\ &\leq \|A_c\|_2^2 \|\delta \bar{P}(\delta\Theta)\|_2 + 2\|\bar{A}_c\|_2 \|P^{\frac{1}{2}} \delta\Theta^\top H P^{-\frac{1}{2}}\|_F \\ &= \sigma_\Theta^2 \|\delta \bar{P}(\delta\Theta)\|_2 + 2\sigma_\Theta \|\delta\Theta^\top H\|_F \end{aligned}$$

By rearranging the inequality and using the property  $\|\delta\Theta^\top H\|_F \leq \|\delta\Theta\|_{V^{-1}} \|H\|_V$ , we obtain

$$\|\delta \bar{P}(\delta\Theta)\|_2 \leq \frac{2\sigma_\Theta}{1 - \sigma_\Theta^2} \|\delta\Theta\|_{V^{-1}} \|H\|_V$$

Observing that  $\|\delta P(\delta\Theta)\|_2 = \|P^{\frac{1}{2}} \delta \bar{P}(\delta\Theta) P^{\frac{1}{2}}\|_2 \leq \|P\|_2 \|\delta \bar{P}(\delta\Theta)\|_2 \leq D \|\delta \bar{P}(\delta\Theta)\|_2$  and noting that  $\|\nabla P_{ij}(\Theta)\|_V = \sup_{\|\delta\Theta\|_{V^{-1}}=1} |\delta P_{ij}(\delta\Theta)| \leq \sup_{\|\delta\Theta\|_{V^{-1}}=1} \|\delta P(\delta\Theta)\|_2$ , one can get

$$\|\nabla P_{ij}(\Theta)\|_V \leq \frac{2D\sigma_\Theta}{1 - \sigma_\Theta^2} \|H(\Theta)\|_V$$

Observing that the function  $\sigma_\Theta : \mathcal{S} \rightarrow \mathbb{R}_+$  is continuous on  $\mathcal{S}$  and  $\sigma_* := \max_{\Theta \in \mathcal{S}} \sigma_\Theta < 1$  as  $\mathcal{S}$  is compact, we can further bound the scalar from above by  $\Theta$  independent constant  $\Gamma = \frac{2D\sigma_*}{1 - \sigma_*^2} > 0$ . ■

The following lemma gives a useful upper bound on  $\Delta_k$ .

**Lemma 24** Let  $\Delta_k$  be defined as in (131). Then, for all  $k \geq 0$ , we have that

$$\Delta_k \leq 2n^2 v_{t_k} \Gamma \mathbb{E} \left[ \|H(\tilde{\Theta}_{t_k})\|_{V_{t_k}^{-1}} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right].$$

**Proof** The proof follows directly from applying the bound in Lemma 23 to Equation 11 in Abeille and Lazaric (2018).  $\blacksquare$

Finally, we are ready to give a bound on the summation of  $\Delta_k$  terms

**Lemma 25** Let  $\Delta_k$  be defined as in (131) for any  $k \geq 0$ . Then, the following bound holds with probability at least  $1 - \delta$

$$\begin{aligned} \sum_{k=0}^K \Delta_k &\leq \frac{16n^2 \alpha v_T \Gamma}{1 + \frac{1}{\beta_T}} \left( \sum_{t=T_w+1}^T \|z_t\|_{V_t^{-1}} + 2\alpha \sqrt{2 \frac{T - T_w}{\tau_0} \frac{1 + \kappa^2}{\mu} \log \left( \frac{2}{\delta} \right)} \right) \\ &\leq \tilde{O}(\text{poly}(n, d, \log(1/\delta)) \sqrt{T - T_w}) \end{aligned}$$

where  $\alpha = (1 + 1/\beta_0^2)(\sqrt{2n \log(3n)} + v_T + (1 + \kappa)S X_s)$ .

**Proof** Define  $\bar{\Theta}_{t_k} = \hat{\Theta}_{t_k} + \beta_{t_k} V_{t_k}^{-\frac{1}{2}} \eta_{t_k}$ . Using Proposition 9 in Abeille and Lazaric (2018), we have that

$$\begin{aligned} \|H(\bar{\Theta}_{t_k})\|_{V_{t_k}^{-1}} &\leq \frac{8}{1 + \frac{1}{\beta_{t_k}}} \left\| H(\bar{\Theta}_{t_k}) \mathbb{E} \left[ x_{t_k} x_{t_k}^\top \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \mid \mathcal{F}_{t_k-1}, E_{t_k-1}, \bar{\Theta}_{t_k} \right] \right\|_{V_{t_k}^{-1}} \\ &\leq \frac{8}{1 + \frac{1}{\beta_{t_k}}} \left\| \mathbb{E} \left[ H(\bar{\Theta}_{t_k}) x_{t_k} x_{t_k}^\top \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \mid \mathcal{F}_{t_k-1}, E_{t_k-1}, \bar{\Theta}_{t_k} \right] \right\|_{V_{t_k}^{-1}} \\ &\leq \frac{8\alpha}{1 + \frac{1}{\beta_{t_k}}} \mathbb{E} \left[ \left\| H(\bar{\Theta}_{t_k}) x_{t_k} \right\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \mid \mathcal{F}_{t_k-1}, E_{t_k-1}, \bar{\Theta}_{t_k} \right] \end{aligned}$$

By Lemma 24 and the preceding bound, we can write

$$\begin{aligned} \Delta_k &\leq 2n^2 v_{t_k} \Gamma \mathbb{E} \left[ \|H(\tilde{\Theta}_{t_k})\|_{V_{t_k}^{-1}} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] = 2n^2 v_{t_k} \Gamma \frac{\mathbb{E} \left[ \|H(\bar{\Theta}_{t_k})\|_{V_{t_k}^{-1}} \mathbb{1}_{\bar{\Theta}_{t_k} \in \mathcal{S}} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right]}{\mathbb{P} \{ \bar{\Theta}_{t_k} \in \mathcal{S} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \}} \\ &\leq \frac{16n^2 \alpha v_{t_k} \Gamma}{1 + \frac{1}{\beta_{t_k}}} \frac{\mathbb{E} \left[ \mathbb{E} \left[ \|H(\bar{\Theta}_{t_k}) x_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \mid \mathcal{F}_{t_k-1}, E_{t_k-1}, \bar{\Theta}_{t_k} \right] \mathbb{1}_{\bar{\Theta}_{t_k} \in \mathcal{S}} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right]}{\mathbb{P} \{ \bar{\Theta}_{t_k} \in \mathcal{S} \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \}} \\ &= \frac{16n^2 \alpha v_{t_k} \Gamma}{1 + \frac{1}{\beta_{t_k}}} \underbrace{\mathbb{E} \left[ \mathbb{E} \left[ \underbrace{\|H(\tilde{\Theta}_{t_k}) x_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha}}_{z_{t_k}} \mid \mathcal{F}_{t_k-1}, E_{t_k-1}, \tilde{\Theta}_{t_k} \right] \mid \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right]}_{=: Y_k} \end{aligned}$$

Notice that  $\mathbb{E} [Y_k \mid \mathcal{F}_{t_k-1}] = \mathbb{E} \left[ \|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \mid \mathcal{F}_{t_k-1} \right]$  by law of iterated expectations and

$$\|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \leq \frac{1}{\sqrt{\mu}} \|H(\tilde{\Theta}_{t_k}) x_{t_k}\| \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \leq \sqrt{\frac{1 + \kappa^2}{\mu}} \alpha.$$

Therefore, the sequence  $\left\{Y_k - \|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha}\right\}_{k \geq 0}$  is a bounded martingale difference sequence. By Azuma's inequality, we have that with probability at least  $1 - \delta$ ,

$$\sum_{k=0}^K \left(Y_k - \|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha}\right) \leq 2\alpha \sqrt{2 \frac{T - T_w}{\tau_0} \frac{1 + \kappa^2}{\mu} \log\left(\frac{2}{\delta}\right)}$$

We can bound the sum of  $\|z_{t_k}\|_{V_{t_k}^{-1}}$  terms using Lemma 10 of [Abbasi-Yadkori et al. \(2011\)](#) and Hölder's inequality as

$$\begin{aligned} \sum_{k=0}^K \|z_{t_k}\|_{V_{t_k}^{-1}} &\leq \sum_{k=0}^K \|z_{t_k}\|_{V_{t_k}^{-1}} + \sum_{k=0}^K \sum_{t=t_k+1}^{t_{k+1}-1} \|z_t\|_{V_t^{-1}} \\ &= \sum_{t=T_w+1}^T \|z_t\|_{V_t^{-1}} \leq \sqrt{T - T_w} \log \frac{\det(V_T)}{\det(V_{T_w})} \end{aligned}$$

Combining these results, we obtain the desired bound

$$\sum_{k=0}^K \Delta_k \leq \frac{16n^2 \alpha \nu_T \Gamma}{1 + \frac{1}{\beta T}} \left( \sum_{t=T_w+1}^T \|z_t\|_{V_t^{-1}} + 2\alpha \sqrt{2 \frac{T - T_w}{\tau_0} \frac{1 + \kappa^2}{\mu} \log\left(\frac{2}{\delta}\right)} \right)$$

■

Now, we are ready to bound  $R_K^{TS,2}$ . Under the event  $E_T$  Theorem 6 suggests that  $1/p_{t^{\text{opt}}} \leq O(1)$  if  $T_w = \omega(\sqrt{T \log T})$  for singular  $A_{c,*}$  and  $T_w = \omega(\log T)$  for non-singular  $A_{c,*}$ . Using this result together with Lemma 25, we have that

$$R_K^{TS,2} = \sum_{k=0}^K \tau_0 R_k^{TS,2} \leq n \sigma_w^2 \tau_0 \sum_{k=0}^K \frac{\Delta_k}{p_{t_k^{\text{opt}}}} \leq \tilde{O}(\text{poly}(n, d) \sqrt{(T - T_w) \log(1/\delta)}) \quad (133)$$

with probability at least  $1 - \delta$ . Combining the above with (128) and (129), we obtain the desired bound. ■

### F.5. Bounding $R_T^{\text{gap}}$

**Lemma 26 (Bounding  $R_T^{\text{gap}}$  for TSAC)** *Let  $R_T^{\text{gap}}$  be as defined by (115). Under the event of  $E_T$ , we have that*

$$|R_T^{\text{gap}}| = \tilde{O}\left(\text{poly}(n, d) \sqrt{T \log(1/\delta)}\right).$$

with probability at least  $1 - 2\delta$  for large enough  $T$ .

**Proof**

$$R_T^{\text{gap}} = \sum_{t=0}^T \mathbb{E} \left[ x_{t+1}^\top \left( P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right) x_{t+1} \mathbb{1}_{E_{t+1}} \mid \mathcal{F}_t \right] \quad (134)$$

$$= \sum_{t=0}^K \mathbb{E} \left[ x_{t_k+1}^\top \left( P(\tilde{\Theta}_{t_k+1}) - P(\tilde{\Theta}_{t_k}) \right) x_{t_k+1} \mathbb{1}_{E_{t_k+1}} \mid \mathcal{F}_{t_k} \right] \quad (135)$$

Separating the duration of TSAC into two parts at  $t = T_r$ , we obtain two same term achieved in (Abeille and Lazaric, 2018). Note that in Abeille and Lazaric (2018), the authors follow frequent update rule and TSAC updates every  $\tau_0$  time-steps. The proof of these terms similarly follow Section 5.2 in Abeille and Lazaric (2018) and using Lemma 25 we obtain  $\tilde{O}((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n, d)\sqrt{T-T_r})$ . Note that there is an additional  $\tau_0$  factor in these bounds, due to “relatively slower” update of TSAC. For large enough  $T$  such that the second term dominates the overall upper bound, we obtain the advertised guarantee.  $\blacksquare$

### F.6. Proof of Theorem 3

Collecting the regret terms derived in subsections of Appendix F, for large enough  $T$ , under the event  $E_T$ , we have that

$$\begin{aligned} R_{T_w}^{\text{exp}} &= \tilde{O}\left((n+d)^{n+d}T_w\right), \quad \text{w.p. } 1 - \delta \\ R_T^{\text{RLS}} &= \tilde{O}\left((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n, d, \log(1/\delta))\sqrt{T-T_r}\right), \\ R_T^{\text{mart}} &= \tilde{O}\left((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n, d, \log(1/\delta))\sqrt{T-T_r}\right), \quad \text{w.p. } 1 - \delta \\ R_T^{\text{gap}} &= \tilde{O}\left((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n, d, \log(1/\delta))\sqrt{T-T_r}\right), \quad \text{w.p. } 1 - 2\delta \end{aligned}$$

and choosing  $T_w = \omega(\sqrt{T \log T})$  for singular  $A_{c,*}$  and  $T_w = \omega(\log T)$  for non-singular  $A_{c,*}$  gives

$$R_T^{\text{TS}} = \tilde{O}\left(\text{poly}(n, d)T_w + \text{poly}(n, d, \log(1/\delta))\sqrt{T-T_r}\right), \quad \text{w.p. } 1 - 2\delta.$$

Recall that the event  $E_T$  is true with probability at least  $1 - 4\delta$ . Combining all these bounds, we have the overall regret bound as

$$R_T = \tilde{O}\left((n+d)^{n+d}T_w + \text{poly}(n, d, \log(1/\delta))\sqrt{T-T_r}\right), \quad \text{w.p. } 1 - 10\delta. \quad (136)$$

Notice that  $R_T$  is linear in the initial exploration time  $T_w$  with an exponential dimension dependency. Also note that  $T_w \geq T_0 := \text{poly}(\log(1/\delta), \sigma_w^{-1}, n, d, \bar{\alpha}, \gamma^{-1}, \kappa)$  guarantees a stabilizing controller by Lemma 4. In order to control the growth of  $R_T$  by  $\tilde{O}(\sqrt{T})$ , the initial exploration time can maximally be in the order of  $(\sqrt{T})^{1+o(1)}$  where  $T^{o(1)}$  hides all multiplicative sub-polynomial growths, *i.e.*,  $T_w = O\left((\sqrt{T})^{1+o(1)}\right) = \tilde{O}(\sqrt{T})$ .

On the other hand, Theorem 6 puts strict lower bounds on the growth of  $T_w$  in order to maintain asymptotically constant optimistic probability. In particular, for singular  $A_{c,*}$ , this condition is stated as  $T_w = \omega(\sqrt{T \log T})$ . Combined with the required upper bound  $O\left((\sqrt{T})^{1+o(1)}\right)$ , it must be that  $T_w = \max\left(T_0, c(\sqrt{T \log T})^{1+o(1)}\right)$  for a constant  $c > 0$  for large enough  $T$ . Inserting this result in (136) gives us

$$R_T = \tilde{O}\left((n+d)^{n+d}\sqrt{T}\right), \quad \text{w.p. } 1 - 10\delta$$

for large enough  $T$ . Observe that exponential dimension dependence is unavoidable in this case as the system is excited with isotropic noise in every direction long enough to dominate with exponential dimension.



For non-singular  $A_{c,*}$ , the lower bound is stated as  $T_w = \omega(\log T)$ . For large enough  $T$ , choosing  $T_w = \max(T_0, c(\log T)^{1+o(1)})$  for a constant  $c > 0$  is sufficient to satisfy both the upper and lower bounds on  $T_w$ . Inserting this result in (136) gives us

$$R_T = \tilde{O}\left(\text{poly}(n, d, \log(1/\delta))\sqrt{T}\right), \quad \text{w.p. } 1 - 10\delta$$

for large enough  $T$ . Observe that the exponential dimension dependence is not dominant anymore since logarithmically large  $T_w$  is sufficient to guarantee asymptotically constant optimistic probability.

## Appendix G. Technical Theorems

**Theorem 27 (Theorem 1 of Abbasi-Yadkori et al. (2011))** *Let  $(\mathcal{F}_t; k \geq 0)$  be a filtration and  $(m_k; k \geq 0)$  be an  $\mathbb{R}^d$ -valued stochastic process adapted to  $(\mathcal{F}_k)$ ,  $(\eta_k; k \geq 1)$  be a real-valued martingale difference process adapted to  $(\mathcal{F}_k)$ . Assume that  $\eta_k$  is conditionally sub-Gaussian with constant  $R$ . Consider the martingale*

$$S_t = \sum_{k=1}^t \eta_k m_{k-1}$$

and the matrix-valued processes

$$V_t = \sum_{k=1}^t m_{k-1} m_{k-1}^\top, \quad \bar{V}_t = V + V_t, \quad t \geq 0$$

Then for any  $0 < \delta < 1$ , with probability  $1 - \delta$

$$\forall t \geq 0, \quad \|S_t\|_{\bar{V}_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta}\right)$$

**Theorem 28 (Azuma's inequality)** *Assume that  $X_s$  is a supermartingale and  $|X_s - X_{s-1}| \leq c_s$  almost surely for  $s \geq 0$ . Then for all  $t > 0$  and all  $\epsilon > 0$ ,*

$$P(|X_t - X_0| \geq \epsilon) \leq 2 \exp\left(\frac{-\epsilon^2}{2 \sum_{s=1}^t c_s^2}\right)$$

**Lemma 29 (Lemma 10 of Abbasi-Yadkori and Szepesvári (2011))** *The following holds for any  $t \geq 1$ :*

$$\sum_{k=0}^{t-1} \left(\|z_k\|_{V_k^{-1}}^2 \wedge 1\right) \leq 2 \log \frac{\det(V_t)}{\det(\lambda I)}$$

Further, when the covariates satisfy  $\|z_t\| \leq c_m, t \geq 0$  with some  $c_m > 0$  w.p. 1 then

$$\log \frac{\det(V_t)}{\det(\lambda I)} \leq (n+d) \log\left(\frac{\lambda(n+d) + tc_m^2}{\lambda(n+d)}\right)$$

**Lemma 30 (Norm of Subgaussian vector)** *Let  $v \in \mathbb{R}^d$  be a entry-wise  $R$ -subgaussian random variable. Then with probability  $1 - \delta$ ,  $\|v\| \leq R\sqrt{2d \log(d/\delta)}$ .*

**Lemma 31 (Theorem 20 of Cohen et al. (2019))** *Let  $z_t \in \mathbb{R}^{n+d}$  for  $t = 0, 1, \dots$  be a sequence random variables that is adapted to a filtration  $\{\mathcal{F}_t\}_{t=0}^\infty$ . Suppose that  $z_t$  are conditionally Gaussian on  $\mathcal{F}_{t-1}$  and that  $\mathbb{E}[z_t z_t^T | \mathcal{F}_{t-1}] \succeq \sigma_z^2 I$  for some fixed  $\sigma_z^2 > 0$ . Then for  $t \geq 200(n+d) \log \frac{12}{\delta}$  we have that with probability at least  $1 - \delta$*

$$\sum_{s=1}^t z_s z_s^T \succeq \frac{t\sigma_z^2}{40} I.$$

## Appendix H. Implementation Details of Numerical Experiments

The LQR problem for the longitudinal flight control of Boeing 747 with linearized dynamics (Ishihara et al., 1992) is given as

$$A_* = \begin{bmatrix} 0.99 & 0.03 & -0.02 & -0.32 \\ 0.01 & 0.47 & 4.7 & 0 \\ 0.02 & -0.06 & 0.4 & 0 \\ 0.01 & -0.04 & 0.72 & 0.99 \end{bmatrix}, B_* = \begin{bmatrix} 0.01 & 0.99 \\ -3.44 & 1.66 \\ -0.83 & 0.44 \\ -0.47 & 0.25 \end{bmatrix}, Q = I, R = I, w \sim \mathcal{N}(0, I). \tag{137}$$

This system has been studied in (Sabag et al., 2021; Lale et al., 2022). It corresponds to the dynamics for level flight of Boeing 747 at the altitude of 40000ft with the speed of 774ft/sec, for a discretization of 1 second. The first element of the state corresponds to the velocity of aircraft along body axis, the second is the velocity of aircraft perpendicular to body axis, the third is the angle between body axis and horizontal and the fourth is the angular velocity of aircraft. The system takes two dimensional inputs, where the first is the elevator angle and the second one is thrust.

For this task we deploy 4 different adaptive control algorithms that do not require initial stabilizing controller: (i) TSAC, (ii) StabL of Lale et al. (2022), (iii) TS-LQR of Abeille and Lazaric (2018), and (iv) OFULQ of Abbasi-Yadkori and Szepesvári (2011). Each algorithm has certain hyperparameters and we tune each parameter in terms of its effect on regret and present the performance of the best performing hyperparameter choices. We use the actual estimation errors in the algorithm design. Note that this has been observed to have negligible effect on the performance (Dean et al., 2018).

To have fair comparison in the regret performance in a stabilizable system like (137), we follow fixed update rule in TS-LQR in parallel with TSAC, and add an additional minimum policy duration constraint to the standard design matrix determinant doubling of OFULQ. Moreover, in the implementation of optimistic parameter search we deploy projected gradient descent (PGD). Even though this approach works efficiently for the small dimensional problems such as (137), it becomes computationally challenging as the dimensionality of the system grows. Nevertheless, our results show that PGD is effective to find optimistic parameters and as observed in Lale et al. (2022) yields the superior performance of StabL with a small margin between TSAC. This difference is in parallel with the predictions of theory. As we show in our analysis, TS samples an optimistic model with a fixed probability. However, an effective way of solving the optimistic control design problem yields optimistic controllers at every time-step and gives more effective control over exploration vs. exploitation trade-off.