# Explore More and Improve Regret in Linear Quadratic Regulators

Sahin Lale[1], Kamyar Azizzadenesheli[2], Babak Hassibi[1], Anima Anandkumar[2]
[1] Department of Electrical Engineering
[2] Department of Computing and Mathematical Sciences
California Institute of Technology, Pasadena
{alale,kazizzad,hassibi,anima}@caltech.edu



**Abstract**

Stabilizing the unknown dynamics of a control system and minimizing regret in control of an unknown system are among the main goals in control theory and reinforcement learning. In this work, we pursue both these goals for adaptive control of linear quadratic regulators (LQR). Prior works accomplish either one of these goals at the cost of the other one. The algorithms that are guaranteed to find a stabilizing controller suffer from high regret, whereas algorithms that focus on achieving low regret assume the presence of a stabilizing controller at the early stages of agent-environment interaction. In the absence of such stabilizing controller, at the early stages, the lack of reasonable model estimates needed for $(i)$ strategic exploration and $(ii)$ design of controllers that stabilize the system, results in regret that scales exponentially in the problem dimensions. We propose a framework for adaptive control that exploits the characteristics of linear dynamical systems and deploys additional exploration in the early stages of agent-environment interaction to guarantee sooner design of stabilizing controllers. We show that for the classes of controllable and stabilizable LQRs, where the latter is a generalization of prior work, these methods achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret with a polynomial dependence in the problem dimensions.

## 1 Introduction

**Linear quadratic regulator (LQR):** Linear dynamical systems are general and fundamental continuous control systems that, due to their unique characteristics, have been vastly used in real-world problems [Zarchan and Musoff, 2013]. Among linear dynamical systems, LQRs are the canonical settings with quadratic regulatory costs to design desirable controllers. In LQRs, when the model of dynamics is given to the decision-making agent, the problem of finding the optimal control to minimize cumulative costs results in a stabilizing linear controller [Bertsekas, 1995].

**The problem of unknown dynamics:** The study of LQRs becomes more challenging when the environment dynamics are unknown. The agent needs to learn the dynamics in order to (1) stabilize the system and (2) find the optimal controller. This is one of the core challenges in reinforcement learning and control theory termed as adaptive control. The agent interacts with the environment, explores it, estimates the system dynamics, and strategically exploits these estimates for further exploration-exploitation. Due to the agent's possible sub-optimal decisions during exploration, the agent's cumulative cost may increase significantly. The agent needs to balance the exploration and exploitation such that it reduces the cumulative cost in long term. The performance of the agent is evaluated based on the notion of regret, which quantifies the difference between the cumulative cost encountered by the agent, and the expected cumulative cost of the optimal controller.

**Optimism in the face of uncertainty principle (OFU):** In order to minimize regret, the principle of OFU [Lai and Robbins, 1985] is proposed as an effective strategy for exploration and exploitation in the study of sequential decision making. OFU principle suggests to estimate model parameters up to their confidence intervals, and then act according to the policy/controller prescribed by a model in the confidence set with the lowest optimal cost, known as the optimistic model.

**Prior work and motivation:** For a couple of decades, the statistical aspect of regret minimization problem has been investigated from the lens of asymptotic optimality [Lai et al., 1982, Lai and Wei, 1987]. Recently, a set of novel techniques have been proposed to develop learning algorithms with finite-time performance guarantees in linear models [Peña et al., 2009]. The seminal work by Abbasi-Yadkori and Szepesvári [2011] uses OFU principle and proposes an algorithm, OFULQ, to balance the exploration and exploitation in the presence of sub-Gaussian disturbances. After $T$ time steps of interaction, OFULQ achieves regret of $\sqrt{T}$ in controllable LQRs, a subset to stabilizable LQRs. However, in the early stage of interactions, when the agent's model estimate may not be good enough, optimism may not provide a sound and strategic exploration, causing a possible blow up in the system state. This uncontrolled state explosion results in a regret upper bound with exponential dependency in LQR dimensions [Abbasi-Yadkori and Szepesvári, 2011], which further highlights the need for improved exploration in the early stages of interactions. In order to circumvent state explosion and have graceful transition to exploitation, most of the prior works assume access to a stabilizing controller during early stages, which may not be possible in many applications.

**Controllability vs. stabilizability:** The controllability assumption implies that the state of the system can be brought to any desirable state in finite-time (Definition 2.2). However, this condition can be too stringent for practical systems. A weaker notion is the stabilizability, which states that there exists a controller that makes the system stable (Definition 2.1). This condition is the necessary and sufficient condition for the optimal control problem to be well-defined [Bertsekas, 1995].

**Stabilizable but not controllable system:** Consider the following simple linear dynamical system:

$$x_{t+1} = \begin{bmatrix} -2 & 0 & 1.1 \\ 1.5 & 0.9 & 1.3 \\ 0 & 0 & 0.5 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} u_t.$$

The state $x_t$ and the input $u_t$ are 3 and 2 dimensional vectors respectively. First two elements in the state vector correspond to controllable modes of the system since any initial value can be brought to any desired value via inputs. However, the input has no control over the third element, thus the system is not controllable. The third element of the state vector evolves with the dynamics of $x_{t+1,3} = 0.5x_{t,3}$. Notice that it is stable, *i.e.* $x_{t,3}$ decays over time. Thus, the system is stabilizable. This example shows that the class of stabilizable systems includes a fairly larger number of systems, including any practical system where the inner dynamics that user cannot directly control are stable. Therefore, in this work we consider the general case of stabilizable systems. Based on these, in this paper, we address the following questions:

1. *Can we provide regret guarantees in stabilizable setting?*

2. *Can we avoid the exponential dimension dependence (at least in long term) in the regret bound by utilizing unique characteristics of linear dynamical systems?*

Table 1: Comparison of current results with prior works.
$\star :=$ No specified dimension dependency, $\dagger :=$ ExpOpt without additional exploration

| **Work** | **Regret** | **Setting** | **Initial Stabilizing Controller** |
|---|---|---|---|
| [Abbasi-Yadkori and Szepesvári, 2011] | $d^d\sqrt{T}$ | Controllable | Does not require |
| [Faradonbeh et al., 2017] | $\sqrt{T}$ $\star$ | Stabilizable | Requires |
| [Dean et al., 2018] | $\text{poly}(d)T^{2/3}$ | Controllable | Requires |
| [Faradonbeh et al., 2018] | $\sqrt{T}$ $\star$ | Stabilizable | Requires |
| [Mania et al., 2019] | $\text{poly}(d)\sqrt{T}$ | Controllable | Requires |
| [Cohen et al., 2019] | $\text{poly}(d)\sqrt{T}$ | Controllable | Requires |
| [Simchowitz et al., 2020] | $\text{poly}(d)\sqrt{T}$ | Stabilizable | Requires |
| [Simchowitz and Foster, 2020] | $\text{poly}(d)\sqrt{T}$ | Stabilizable | Requires |
| **Theorem 1**$^\dagger$ | $d^d\sqrt{T}$ | Stabilizable | Does not require |
| **Theorem 2** | $\text{poly}(d)\sqrt{T}$ | Controllable | Does not require |
| **Theorem 3** | $\text{poly}(d)\sqrt{T}$ | Stabilizable | Does not require |

**Contributions:** In this work, we give affirmative answers to these questions. First, we extend the prior work on controllable settings to the more general case of stabilizable LQRs. We show that, in stabilizable LQRs, when using optimism to balance exploration and exploitation, choosing proper time steps to update model parameters and controller plays a crucial role in minimizing regret. We propose an algorithm with a carefully designed choice for updating rule and show that it achieves the regret of $\tilde{\mathcal{O}}(d^d\sqrt{T})$ in stabilizable LQRs with sub-Gaussian disturbances. Here $d$ is the problem dimension of LQR and $\tilde{\mathcal{O}}(\cdot)$ presents the order up to logarithmic terms.

In order to mitigate the exponential dependency in the regret upper bounds, we further investigate the behavior of optimism-based agents, specifically, in the early stages of agent-environment interactions. In these early stages, the agent has inaccurate model estimates. When the agent strategizes upon its inaccurate estimates using optimism to directly minimize regret, the committed actions may not provide sufficient exploration required to achieve stabilizing controllers. This insufficiency results in regret with exponential dependence in the problem dimension.

To address this challenge, we suggest to further exploit the unique characteristics of linear dynamical systems. In the early stages, any strategy may result in linear regret. Therefore, instead of directly aiming to minimize regret in these stages, we propose to carefully adjust the early exploration to also guarantee that stabilizing controllers are soon achieved. Achieving stabilizing controllers for the unknown systems assures that the agent can bring the system states under control, avoid the explosion of dynamics, and attain a better long term regret. We accompany the OFU-based controller with an additional random exploration in the early, when the optimism alone may not provide any better exploration strategy. We show that this additional exploration imposes slight additional constant regret in the short stages of early interaction, but the resulting stabilizing controllers ensure stable behavior, therefore much smaller regret in the long term. Using these principles, we propose two algorithms, respectively, for controllable and stabilizable LQRs with sub-Gaussian disturbances, both with regret upper bound of $\tilde{\mathcal{O}}(\text{poly}(d)\sqrt{T})$. The results in this work can be considered as a generalization of the prior work. Table 1 provides the comparison of our results with prior works in terms of rate, system characteristic, and the existence of a initial stabilizing controller.

## 2 Preliminaries

We denote the Euclidean norm of a vector $x$ as $\|x\|_2$. For a given matrix $A$, $\|A\|_2$ denotes the spectral norm, $\|A\|_F$ denotes the Frobenius norm, while $A^\top$ is the transpose, $A^\dagger$ is the Moore-Penrose inverse and for square matrices, $\mathrm{Tr}(A)$ gives the trace of matrix $A$ and $\rho(A)$ denotes the spectral radius of $A$, *i.e.* largest absolute value of $A$'s eigenvalues. The j-th singular value of a rank-$n$ matrix $A$ is denoted by $\sigma_j(A)$, where $\sigma_{\max}(A) := \sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_{\min}(A) := \sigma_n(A)$. $I$ is the identity matrix with relevant dimensions. Consider a discrete time linear time-invariant system characterized as,

$$x_{t+1} = A_* x_t + B_* u_t + w_t. \tag{1}$$

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 exact state information. Then, the agent applies a control input $u_t$ and the system evolves to $x_{t+1}$ at time $t+1$. At each time step $t$, the agent pays a cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$, where $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{d\times d}$ are positive definite matrices such that $\|Q\|, \|R\| < \overline{\alpha}$ and $\sigma_{\min}(Q), \sigma_{\min}(R) > \underline{\alpha}$. The problem is to design control inputs based on past observations in order to minimize the average expected cost:

$$J_\star = \lim_{T\to\infty} \min_{u=[u_1,\ldots,u_T]} \frac{1}{T}\mathbb{E}\Big[\sum\nolimits_{t=1}^{T} x_t^\top Q x_t + u_t^\top R u_t\Big] \tag{2}$$

This problem is the canonical example for the control of linear dynamical systems and termed as linear quadratic regulator (LQR). One can represent the underlying system (1) as

$$x_{t+1} = \Theta_*^\top z_t + w_t$$

where $\Theta_*^\top = [A_* \;\; B_*]$ and $z_t = [x_t^\top \;\; u_t^\top]^\top$. Knowing $\Theta_*$, the solution of (2), the optimal control law, is a linear feedback control $u_t = K(\Theta_*)x_t$ with

$$K(\Theta_*) = -\left(R + B_*^\top P(\Theta_*)B_*\right)^{-1} B_*^\top P(\Theta_*)A_*, \tag{3}$$

where $P(\Theta_*)$ is the unique positive definite solution to the discrete-time algebraic Riccati equation:

$$P(\Theta_*) = A_*^\top P(\Theta_*)A_* + Q - A_*^\top P(\Theta_*)B_* \left(R + B_*^\top P(\Theta_*)B_*\right)^{-1} B_*^\top P(\Theta_*)A_*. \tag{4}$$

The optimal cost for $\Theta_*$, $J_\star = \mathrm{Tr}(P(\Theta_*)W)$ where $W = \mathbb{E}[w_t w_t^\top | \mathcal{F}_{t-1}]$ for a corresponding filtration $\mathcal{F}_t$. When the model parameters, $A_*$, and $B_*$, are unknown, the agent interacts with the environment to learn these parameters and aims to minimize the cumulative cost $\sum_{t=1}^T c_t$. Note that the cost matrices $Q$ and $R$ are designer's choice and given. After $T$ time steps, we evaluate the regret in agent's performance, i.e.,

$$\mathrm{REGRET}(T) = \sum\nolimits_{t=0}^{T}(c_t - J_*),$$

which is the difference between the performance of the agent and the expected performance of the optimal controller. We have the following definitions for the linear dynamical system governed by $A_*$ and $B_*$.

**Definition 2.1** (Stabilizability)**.** *The linear dynamical system* $\Theta_*$ *is stabilizable if there exists* $K$ *such that* $\rho(A_* + B_*K) < 1$.

**Definition 2.2** (Controllability)**.** *The linear dynamical system* $\Theta_*$ *is controllable if the controllability matrix* $[B_* \ A_*B_* \ A_*^2B_* \ \dots \ A_*^{n-1}B_*]$ *has full row rank.*

Note that stabilizability is weaker than controllability, *i.e.*, all controllable systems are stabilizable but the converse is not true. Under both conditions, it is guaranteed to have a unique positive definite solution to (4) [Kučera, 1972]. Similar to Cohen et al. [2019], we need quantitative version of stabilizability for the finite-time analysis.

**Definition 2.3** ($(\kappa,\gamma)$-Stabilizability)**.** *The linear dynamical system* $\Theta_*$ *is* $(\kappa,\gamma)$*-stabilizable for* ($\kappa \geq 1$ *and* $0 < \gamma \leq 1$) *if* $\|K(\Theta_*)\| \leq \kappa$ *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 this is just a quantification of stabilizability. In other words, any stabilizable system is also $(\kappa,\gamma)$-stabilizable for some $\kappa$ and $\gamma$ and the conversely $(\kappa,\gamma)$-stabilizability implies stabilizability (Appendix H). In this work, we provide an array of results for various settings of LQR. We consider the problem setups with sub-Gaussian process noise $w_t$ and two different system characteristics.

**Assumption 2.1** (General Sub-Gaussian Noise)**.** *There exists a filtration* $(\mathcal{F}_t)$ *such that* $z_t, x_t$ *are* $\mathcal{F}_t$*-measurable and for all* $t \geq 0$*, and* $j \in [0,\dots,n]$*,* $w_{t,j}s$ *are* $\sigma_w^2$*-sub-Gaussian, i.e., for any* $\gamma \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\gamma w_{t,j}\right)|\mathcal{F}_{t-1}\right] \leq \exp\left(\gamma^2\sigma_w^2/2\right) \quad \textit{and} \quad \mathbb{E}\left[w_t w_t^\top|\mathcal{F}_{t-1}\right] = \bar{\sigma}_w^2 I \textit{ for some } \bar{\sigma}_w^2 > 0.$$

Note that the assumption of having isotropic $w_t$ is only used to provide cleaner analysis and the analysis works without that assumption similar to Abbasi-Yadkori and Szepesvári [2011].

**Assumption 2.2** (Controllable Linear Dynamical System)**.** *The unknown parameter* $\Theta_*$ *is a member of a set* $\mathcal{S}_c$ *such that*

$$\mathcal{S}_c \subseteq \left\{\Theta' = [A', B'] \in \mathbb{R}^{n\times(n+d)} \;\middle|\; \Theta' \textit{ is controllable}, \|A' + B'K(\Theta')\| \leq \Upsilon < 1, \ \|\Theta'\|_F \leq S\right\}$$

*Following the controllability and the boundedness of* $\mathcal{S}_c$*, we have finite numbers* $D$ *and* $\kappa \geq 1$ *s.t.,* $\sup\{\|P(\Theta')\| \mid \Theta' \in \mathcal{S}_c\} \leq D$ *and* $\sup\{\|K(\Theta')\| \mid \Theta' \in \mathcal{S}_c\} \leq \kappa$.

**Assumption 2.3** (Stabilizable Linear Dynamical System)**.** *The unknown parameter* $\Theta_*$ *is a member of a set* $\mathcal{S}_s$ *such that*

$$\mathcal{S}_s \subseteq \left\{\Theta' = [A', B'] \in \mathbb{R}^{n\times(n+d)} \;\middle|\; \Theta' \textit{ is } (\kappa,\gamma)\textit{-stabilizable}, \ \|\Theta'\|_F \leq S\right\}$$

*From* $(\kappa,\gamma)$*-stabilizability and the boundedness of* $\mathcal{S}_s$*, we have that* $\rho(A' + B'K(\Theta')) \leq 1-\gamma$*, and we have finite numbers* $D$ *and* $\kappa \geq 1$ *s.t.,* $\sup\{\|P(\Theta')\| \mid \Theta' \in \mathcal{S}_c\} \leq D$ *and* $\sup\{\|K(\Theta')\| \mid \Theta' \in \mathcal{S}_c\} \leq \kappa$.

In the following, under Assumption 2.1, we provide three different algorithms with regret guarantees that are tailored for the systems that satisfy Assumption 2.2 or Assumption 2.3.

**Algorithm 1** ExpOpt

**Input:** Setting, Initial Exploration, $H_0$ minimum duration for a controller
**if** Setting == Controllable **then**
  Choose the controllability set, $\mathcal{S} = \mathcal{S}_c$, set exploration duration $T_w = T_c$, and $\bar{H} = 0$
**elseif** Setting == Stabilizable
  Choose the stabilizability set, $\mathcal{S} = \mathcal{S}_s$, set exploration duration $T_w = T_s$, and $\bar{H} = H_0$
Initialize the optimistic model $\tilde{\Theta}_0$ and controller $K(\tilde{\Theta}_0)$
**for** $t = 0, \dots, T$ **do**
  **if** Determinant of the design matrix is doubled since last controller update,
    **and** More than $\bar{H}$ steps is passed since last controller update **then**
     Estimate the system using regularized least squares in (5)
     Construct the confidence set and find the optimistic parameter $\tilde{\Theta}_t$
  **if** Initial Exploration = True **and** $(t < T_w)$ **then**
     Deploy control input $u_t = K(\tilde{\Theta}_{t-1})x_t + \nu_t$   Optimism + Exploration
  **else**
     Deploy control input $u_t = K(\tilde{\Theta}_{t-1})x_t$   Optimism
  Observe $x_{t+1}$ and update the design matrix

# 3 Algorithm

We propose ExpOpt, whose pseudocode is provided in Algorithm 1. The algorithm is applicable to both controllable and stabilizable LQRs. If the system is controllable, then ExpOpt chooses the controllability set $\mathcal{S} = \mathcal{S}_c$ described in Assumption 2.2 and if the system is stabilizable, then the stabilizability set $\mathcal{S} = \mathcal{S}_s$ described in Assumption 2.3 is chosen for the search of optimal controller.

ExpOpt uses regularized least squares estimates obtained using the past input-output(state) pairs,

$$\hat{\Theta}_t = \arg\min_{\Theta} \sum\nolimits_{s=0}^{t-1} \mathrm{Tr}\left(\left(x_{s+1} - \Theta^\top z_s\right)\left(x_{s+1} - \Theta^\top z_s\right)^\top\right) + \lambda \|\Theta\|_F^2. \tag{5}$$

Using this, ExpOpt constructs a high probability confidence set $\mathcal{C}_t(\delta)$ that contains the underlying parameter $\Theta_*$ with high probability. For $\delta \in (0,1)$, at time step $t$, $\mathcal{C}_t(\delta)$ is defined as,

$$\mathcal{C}_t(\delta) = \left\{\Theta : \sqrt{\mathrm{Tr}\left\{(\Theta - \hat{\Theta}_t)^\top V_t (\Theta - \hat{\Theta}_t)\right\}} \leq \beta_t(\delta)\right\} \quad \text{for} \quad \beta_t(\delta) = \sigma_w \sqrt{2n \log\left(\frac{\det(V_t)^{1/2}}{\delta \det(\lambda I)^{1/2}}\right)} + \sqrt{\lambda} S,$$

and $V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i^\top$. The guarantee that $\Theta_* \in \mathcal{C}_t(\delta)$ with probability at least $1-\delta$ for all time steps $t$ is obtained in Abbasi-Yadkori and Szepesvári [2011]. The confidence set above provides a self-normalized bound on the model parameter estimates via regularized design matrix $V_t$. At all time steps of execution, ExpOpt deploys optimism in the face of uncertainty (OFU) principle in order to design the controller. ExpOpt chooses an optimistic parameter $\tilde{\Theta}_t$ from $\mathcal{C}_t \cap \mathcal{S}$ such that,

$$J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) + 1/\sqrt{t},$$

and constructs the optimal linear controller $K(\tilde{\Theta}_t)$ for the chosen parameter $\tilde{\Theta}_t$. The key idea in OFU principle is to choose the model whose average expected cost is smallest among the set of plausible models. This allows ExpOpt to obtain a good balance between exploration and exploitation. As the confidence set shrinks, the performance of ExpOpt improves over time. For technical reasons utilized in Appendix C, the algorithm uses the one step prior optimistic controller at each time step, *i.e.* at time $t$, the optimal linear controller $K(\tilde{\Theta}_{t-1})$ for $\tilde{\Theta}_{t-1}$ is used.

ExpOpt avoids frequent updates in the system estimates and the controller. It uses the same controller at least for a fixed time period of $H_0$ if the system is stabilizable and also waits for significant refinement in the estimates whether the system is controllable or stabilizable. The latter is achieved by updating the controller if the determinant of the design matrix is doubled since the last update.

For both systems, the algorithm has two options, applying additional random exploration in first $T_w$ steps or avoiding additional exploration. The additional exploration is dedicated to obtain better estimates of the model at a faster rate in the expense of slightly larger regret in the early stages of ExpOpt. To this end, an i.i.d. Gaussian vector, $\nu_t \sim \mathcal{N}(0, \sigma_\nu^2 I)$ where $\sigma_\nu^2 = 2\kappa^2 \bar{\sigma}_w^2$, is injected besides the control input, $K(\tilde{\Theta}_{t-1})x_t$, at each time step of the additional exploration period.

The additional noise excites the system uniformly which enables to find a stabilizing neighborhood around the underlying parameter $\Theta_*$ faster. If the enforced exploration is chosen, ExpOpt continues injecting $\nu_t$ until the system parameter estimates are guaranteed to be close enough to the underlying parameters which is determined by the enforced exploration duration of $T_w$. Define $T_c$ and $T_s$ as

$$T_c := poly(\sigma_w, \sigma_\nu^{-1}, n+d, (1-\Upsilon)^{-1}, \kappa), \quad T_s := poly(\sigma_w, \sigma_\nu^{-1}, n+d, (1-\gamma/2)^{-1}, \kappa, \overline{\alpha}, \underline{\alpha}). \qquad (6)$$

$T_w$ is equal to $T_c$ if the system is controllable and equal to $T_s$ if the system is stabilizable. These durations are chosen such that after $T_w$ time steps, the agent has the guarantee that the linear controller $K(\tilde{\Theta}_{t-1})$ produces stable dynamics when applied to the underlying system $\Theta_*$. Thus, if initial exploration is desired, then we have

$$u_t = K(\tilde{\Theta}_{t-1})x_t + \nu_t \text{ for } t \leq T_w, \quad \text{and} \quad u_t = K(\tilde{\Theta}_{t-1})x_t \text{ for } t > T_w.$$

Otherwise for all $t$, ExpOpt applies $u_t = K(\tilde{\Theta}_{t-1})x_t$. Note that for controllable systems with no additional exploration, ExpOpt yields OFULQ of Abbasi-Yadkori and Szepesvári [2011] where the authors provide $\tilde{\mathcal{O}}((n+d)^{n+d}\sqrt{T})$ regret upper bound. In the following, we first generalize OFULQ to stabilizable systems and then study ExpOpt with exploration in both controllable and stabilizable setting. The precise description of ExpOpt is in Appendix B.

# 4 Analysis

## 4.1 Generalization to Stabilizable LQR

We first consider ExpOpt without additional exploration in stabilizable LQR. The setting is more challenging compared to its controllable counterpart considered in Abbasi-Yadkori and Szepesvári [2011]. Recall Assumption 2.3 that states the system is $(\kappa, \gamma)$-stabilizable, which yields $\rho(A_\star + B_\star K(\Theta_*)) \leq 1-\gamma$ for the optimal controller $K(\Theta_*) \leq \kappa$. Therefore, even if the optimal controller of the underlying system is chosen from $\mathcal{S}_s$, it may not produce contractive closed-loop system, *i.e.*, we can have $\rho(A_\star + B_\star K(\Theta_*)) < 1 < \|A_\star + B_\star K(\Theta_*)\|$ since for any matrix $M$, $\rho(M) \leq \|M\|$.

From the definition of stabilizability in Definitions 2.1 and 2.3, we know that for any stabilizing controller $K'$, there exists a similarity transformation $H' \succ 0$ such that it makes the closed loop system contractive, *i.e.* $A_* + B_* K' = H' L H'^{-1}$, with $\|L\| < 1$. However, even if all the policies that ExpOpt execute stabilize the underlying system, these different similarity transformations of different policies can further cause an explosion of state during the policy changes. If policy changes happen frequently, this may even lead to linear scaling of the state over time.

In order to remedy this situation, ExpOpt carefully designs the timing of policy updates and applies all the policies long enough, so that the state stays well controlled. Therefore, ExpOpt applies the same policy at least for $H_0 = O(\gamma^{-1}\log(\kappa))$ time steps. Using this, we show the boundedness of the state during the execution of ExpOpt (see Appendix E.2 for the proof). Then decomposing regret similarly with Abbasi-Yadkori and Szepesvári [2011], we provide the following generalization of their result for stabilizable systems whose proof is provided in Appendix F-G.

**Theorem 1** (Regret of ExpOpt in stabilizable system using only OFU)**.** *Suppose Assumptions 2.1 and 2.3 hold for the given* LQR. *Then, for* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,* ExpOpt *without additional exploration achieves regret of* $\tilde{\mathcal{O}}((n+d)^{n+d}\sqrt{T\log(1/\delta)})$.

## 4.2 Regret Upper Bound with Early Exploration

Next we analyze the benefit of the early additional exploration of ExpOpt in controllable and stabilizable LQRs. Define $\sigma_\star > 0$ where $\sigma_\star$ is a problem and in particular $\bar{\sigma}_w$, $\sigma_w$, $\sigma_\nu$ dependent constant (please refer to Appendix C for precise definition). Adding the random exploration $\nu_t \sim \mathcal{N}(0, \sigma_\nu^2)$ for $\sigma_\nu^2 = 2\kappa^2\bar{\sigma}_w^2$ enables to guarantee the consistency of the parameter estimation and provides the following bound.

**Lemma 4.1** (Spectral norm of parameter estimation error)**.** *Suppose Assumption 2.1 holds. For* $T \geq poly(\sigma_w^2, \sigma_\nu^2, n, \log(1/\delta))$ *using additional exploration for the systems that satisfy Assumption 2.2 or Assumption 2.3, with probability at least* $1-4\delta$*, we have*

$$\|\hat{\Theta}_T - \Theta_*\|_2 \leq \frac{1}{\sigma_\star\sqrt{T}}\left(\sigma_w\sqrt{2n\log\left(\frac{\det(V_T)^{1/2}}{\delta\det(\lambda I)^{1/2}}\right)} + \sqrt{\lambda}S\right). \tag{7}$$

The proof is given in Appendix C & D. We show that for both systems, additional exploration provides the persistence of excitation of the inputs. In other words, for long enough additional exploration we show that the smallest eigenvalue of the design matrix $V_t$ scales linearly over time. Using the confidence set construction of ExpOpt, we derive the advertised result in Lemma 4.1.

In Appendix A, we show that there exists a stabilizing neighborhood around $\Theta_*$, *i.e.*, $\|\Theta' - \Theta_*\| \leq \epsilon$, such that $K(\Theta')$ stabilizes $\Theta_*$ for any $\Theta'$ in this neighborhood. By the choice of $T_c$ and $T_s$ in (6), ExpOpt guarantees to find this stabilizing neighborhood sooner via additional exploration of first $T_w = T_c$ steps in controllable systems and first $T_w = T_s$ steps in stabilizable systems. For $t \geq T_w$, ExpOpt starts redressing the possible state explosion due to unstable controllers and the perturbation in the early stages. Define $T_{r,c}$ and $T_{r,s}$ as,

$$T_{r,c} \coloneqq T_c + O\left(\frac{(n+d)\log(n+d)}{\log(2/(1+\Upsilon))}\right), \quad T_{r,s} \coloneqq T_s + O\left(\frac{(n+d)\log(n+d)}{\frac{\gamma}{2} - \frac{2}{H_0}\log\kappa}\right). \tag{8}$$

Recall that $H_0$ is the minimum duration for a controller such that the state is well-controlled despite the policy changes. In the following, we show that for $T > T_{r,c}$ in controllable and for $T > T_{r,s}$ in stabilizable systems, the stabilizing controllers are applied long enough that the state stays bounded.

**Lemma 4.2** (Boundedness of state with additional exploration)**.**
*1) Under Assumptions 2.1 & 2.2, if* ExpOpt *runs with additional exploration for* $T_c$ *time steps, with probability at least* $1-2\delta$*, for* $t \leq T_{r,c}$*,* $\|x_t\| = O((n+d)^{n+d})$*. On the other hand, for* $T \geq t > T_{r,c}$*,* $\|x_t\| \leq \overline{X}_c := poly(\sigma_w, \sqrt{n}, (1-\Upsilon)^{-1}, \log(1/\delta))$ *with probability at least* $1-3\delta$*.*
*2) Under Assumptions 2.1 & 2.3, if* ExpOpt *runs with additional exploration for* $T_s$ *time steps, with probability at least* $1-2\delta$*, for* $t \leq T_{r,s}$*,* $\|x_t\| = O((n+d)^{n+d})$*. On the other hand, for* $T \geq t > T_{r,s}$*,* $\|x_t\| \leq \overline{X}_c := poly(\sigma_w, \sqrt{n}, \kappa, \gamma, \log(1/\delta))$ *with probability at least* $1-3\delta$*.*

In the proof of lemma in Appendix E, we use the fact that the policies seldom change via determinant doubling condition on the design matrix or the lower bound of $H_0$ on the duration of each controller application. Thus, we show that exponential decay of the closed-loop system due to stabilizing controller brings the state to an equilibrium. After showing the boundedness of state for ExpOpt with additional exploration, we can finally present the regret results for these settings. Using the general regret decomposition for adaptive control of LQR in Abbasi-Yadkori and Szepesvári [2011], we show that the extra regret suffered from additional exploration is tolerable in the upcoming stages via the guaranteed stabilizing controller, yielding polynomial dimension dependency in regret.

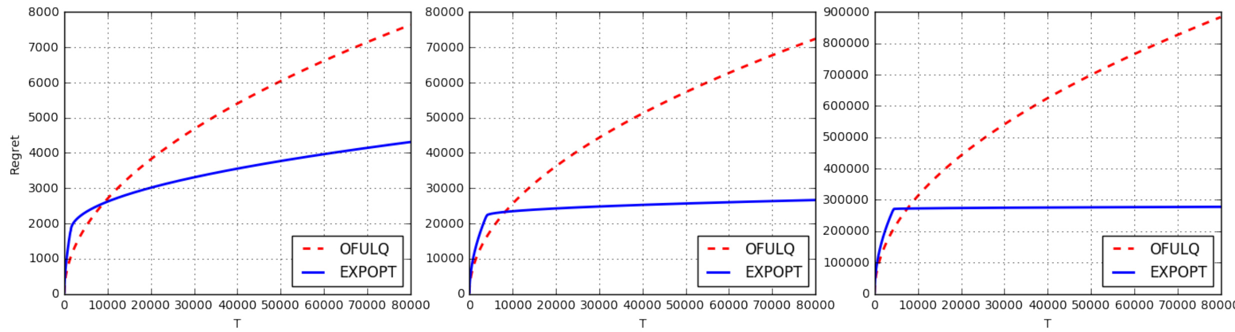
**Theorem 2** (Regret of ExpOpt with additional exploration in controllable systems)**.** *Under Assumptions 2.1 & 2.2, for* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, if* ExpOpt *uses additional exploration for first* $T_c$ *time-steps then it achieves regret of* $\tilde{\mathcal{O}}(poly(n+d)\sqrt{T\log(1/\delta)})$*, for long enough* $T$*.*

**Theorem 3** (Regret of ExpOpt with additional exploration in stabilizable systems)**.** *Suppose Assumptions 2.1 and 2.3 hold. For* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, if* ExpOpt *uses additional exploration for first* $T_s$ *time-steps then* ExpOpt *achieves regret of* $\tilde{\mathcal{O}}\big(poly(n+d)\sqrt{T\log(1/\delta)}\big)$*, for long enough* $T$*.*

The proofs and the exact expressions are presented in Appendix G. Roughly the exact regret expressions have a constant regret term due to additional exploration for $T_w$ time steps with exponential dimension dependency and a term that scales with square root of the remaining time with polynomial dimension dependency, *i.e.* $(n+d)^{n+d}T_w + poly(n+d)\sqrt{T-T_w}$. Note that $T_c$ and $T_s$ are problem dependent expressions. Therefore, for large enough $T$, we derive the advertised regret bounds.

## 4.3 Comparison of Theoretical Upper Bounds

To further observe the nature of the stated regret upper bounds, we plot the regret guarantee of current result with respect to regret upper bound of OFULQ Abbasi-Yadkori and Szepesvári [2011] in Figure 1. Figures 1(a), (b) and (c) show the theoretical regret upper bounds of both algorithms in 3, 4 and 5-dimensional LQR settings respectively. Notice that the additional explorations in the beginning of ExpOpt until guaranteeing the construction of stabilizing controller, increases the regret upper bound by some small margin. Thus for short time period, OFULQ performs better than ExpOpt. However, once it is guaranteed that all the optimistically chosen controllers stabilize the underlying system, regret upper bound of ExpOpt starts scaling gracefully with the rate of $\sqrt{T-T_c}$ with polynomial dimension dependency whereas OFULQ continues with exponential dependency. Therefore, eventually regret upper bound of ExpOpt becomes tighter than OFULQ. Notice that as the problem dimension increase, the benefit of early additional exploration becomes more apparent.




(a) 3-dimensional LQR (b) 4-dimensional LQR (c) 5-dimensional LQR

Figure 1: Regret Upper Bound Comparison of ExpOpt vs. OFULQ in Adaptive Control of LQR

# 5 Related Work

**Asymptotic results in adaptive control:** Over the last decade, a large body of literature attempt to address the controlling and minimizing costs in unknown dynamical systems. Interactive learning-based methods are proposed to explore and estimate the system dynamics and exploit the estimates to design even better controllers. For such methods, a set of fundamental studies provide in asymptotic convergence guarantees [Lai et al., 1982, Lai and Wei, 1987, Fiechter, 1997]. These works show that as the number of interactions incrase, the gap between a proposed policy and the optimal one closes.

**Finite time regret guarantees:** The finite-time study of LQR in adaptive setting is one of the most popular research directions recently. Some of the works analyze the suboptimality gap in adaptive control of LQR using certainty equivalence controller [Mania et al., 2019, Faradonbeh et al., 2018] where they show sublinear regret, when the estimates are close enough. In Dean et al. [2018], it is shown that $\epsilon$-greedy exploration with a robust controller achieves $\tilde{\mathcal{O}}(poly(d)T^{2/3})$ regret. In Abeille and Lazaric [2018] and Ouyang et al. [2017], authors use Thompson sampling to show $\sqrt{T}$ frequentist regret for scalar systems and $\sqrt{T}$ bayesian regret respectively. Cohen et al. [2019] proposes a new SDP formulation for OFU principle and with an initial stabilizing controller shows $\sqrt{T}$ regret can be obtained. More recently, Cassel et al. [2020] provides logarithmic regret if only $A$ or $B$ are unknown in LQR and Simchowitz and Foster [2020] provides a regret lower bound and shows the optimal polynomial dependency when a stabilizing controller is given initially.

**OFU based works:** For control problems, OFU principle was first used by Campi and Kumar [1998]. Besides adaptive control of LQR, OFU principle is widely used in a variety of decision making paradigm, such as multi-arm bandit [Auer, 2002], linear bandit [Abbasi-Yadkori et al., 2011], Markov Decision Processes [Jaksch et al., 2010, Azizzadenesheli et al., 2016], and adaptive control of partially observable linear quadratic control systems [Lale et al., 2020b,c].

**Generalized settings:** Besides the classical setting of adaptive control of LQR considered in this paper, there are various works that consider more general settings. One line of research considers the partially observable counterpart of LQR termed as LQG. It is the setting where instead of exact state vector, a noisy linear combination of the state vector is observed. In this setting Mania et al. [2019], Simchowitz et al. [2020], Lale et al. [2020c] achieve $\sqrt{T}$ regret and Lale et al. [2020a] introduces the algorithm that achieves first polylogarithmic regret in this setting. Another line of research considers adaptive control under adversarial noise disturbances [Hazan et al., 2019, Simchowitz et al., 2020]. All these works either consider the existence of a stabilizing controller or open-loop stable system dynamics. We believe extending the idea of first finding a stabilizing controller without these assumptions is an important future direction.

# 6 Conclusion

In this paper, we propose an algorithm framework, ExpOpt, that follows OFU principle to balance between exploration and exploitation in interaction with LQRs. We show that if an additional random exploration is enforced in the early stages of the agent's interaction with the environment, ExpOpt has the guarantee to design a stabilizing controller sooner. We then show that while the agent enjoys the benefit of stable dynamics in further stages, the additional exploration does not alter the early performance of the agent considerably. Finally, we prove that the regret upper bound of ExpOpt is $\tilde{\mathcal{O}}(\sqrt{T})$ with polynomial dependence in the problem dimensions of the LQRs in both controllable and stabilizable systems. The benefit of additional exploration in finding the stabilizing neighborhood earlier and significantly reducing the regret in adaptive control of LQRs suggests to study this phenomenon in other adaptive control problems such as adaptive control in partially observable setting.

# 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. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. *arXiv preprint arXiv:1602.07764*, 2016.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 2. Athena scientific Belmont, MA, 1995.

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

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv preprint arXiv:2002.08095*, 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. Input perturbations for adaptive regulation and learning. *arXiv preprint arXiv:1811.04258*, 2018.

Claude-Nicolas Fiechter. Pac adaptive control of linear systems. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 72–80. Citeseer, 1997.

Elad Hazan, Sham M Kakade, and Karan Singh. The nonstochastic control problem. *arXiv preprint arXiv:1911.12178*, 2019.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Vladimír Kučera. The discrete riccati equation of optimal control. *Kybernetika*, 8(5):430–447, 1972.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tze Leung Lai and Ching-Zong Wei. Asymptotically efficient self-tuning regulators. *SIAM Journal on Control and Optimization*, 25(2):466–481, 1987.

Tze Leung Lai, Ching Zong Wei, et al. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1): 154–166, 1982.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *arXiv preprint arXiv:2003.11227*, 2020a.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable linear quadratic control. *arXiv preprint arXiv:2002.00082*, 2020b.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret bound of adaptive control in linear quadratic gaussian (lqg) systems. *arXiv preprint arXiv:2003.05999*, 2020c.

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.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2009.

Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. *arXiv preprint arXiv:2001.09254*, 2020.

Paul Zarchan and Howard Musoff. *Fundamentals of Kalman filtering: a practical approach*. American Institute of Aeronautics and Astronautics, Inc., 2013.

# Appendix

## A Stabilizing Neighborhood Around The System Parameters

In this section, we first show that the given systems (both controllable and stabilizable) Discrete Algebraic Riccati Equation has unique positive definite solution. Then, we show that combining two prior results, there exists a stabilizing neighborhood round the system parameters that any controller designed using parameters in that neighborhood stabilizes the system.

**Theorem 4** (Unique Positive Definite Solution to DARE, [Bertsekas, 1995])**.** *For $\Theta_* = (A_*, B_*)$, If $(A_*, B_*)$ is stabilizable and $(C, A_*)$ is observable for $Q = C^\top C$, or $Q$ is positive definite, then there exists a unique, bounded solution, $P(\Theta_*)$, to the DARE:*

$$P(\Theta_*) = A_*^\top P(\Theta_*) A_* + Q - A_*^\top P(\Theta_*) B_* \left(R + B_*^\top P(\Theta_*) B_*\right)^{-1} B_*^\top P(\Theta_*) A_*. \tag{9}$$

*The controller $K(\Theta_*) = -\left(R + B_*^\top P(\Theta_*) B_*\right)^{-1} B_*^\top P(\Theta_*) A_*$ produces stable closed-loop system, $\rho(A_* + B_* K(\Theta_*)) < 1$.*

This result shows that, for we get unique positive definite solution to DARE for both controllable and stabilizable systems.

Let $J_\star \leq \mathcal{J}$. The following lemma is introduced in Mania et al. [2019] and shows that if the estimation error on the system parameters is small enough, then the performance of the optimal controller synthesized by these model parameter estimates scales quadratically with the estimation error.

**Lemma A.1** ([Mania et al., 2019])**.** *There are explicit constants $C_0, \epsilon = \text{poly}\left(\underline{\alpha}^{-1}, \overline{\alpha}, \|A_*\|, \|B_*\|, \bar{\sigma}_w^2, D, n, d\right)$ such that, for any $0 \leq \varepsilon \leq \epsilon$ and for $\|\Theta' - \Theta_*\| \leq \varepsilon$, the infinite horizon performance of the policy $K(\Theta')$ on $\Theta_*$ obeys the following,*

$$J(K(\Theta'), A_*, B_*, Q, R) - J_\star \leq C_0 \varepsilon^2.$$

This result shows that there exists a $\epsilon$-neighborhood around the system parameters that stabilizes the system. This result further extended to quantify the stability in Cassel et al. [2020].

**Lemma A.2** (Lemma 41 in Cassel et al. [2020])**.** *Suppose $J(K(\Theta'), A_*, B_*, Q, R) \leq \mathcal{J}'$ for the* LQR *under Assumption 2.1, then $K(\Theta')$ produces $(\kappa', \gamma')$-stable closed-loop dynamics where $\kappa' = \sqrt{\frac{\mathcal{J}'}{\underline{\alpha}\bar{\sigma}_w^2}}$ and $\gamma' = 1/2\kappa'^2$.*

Combining these results, we obtain the following lemma which will be useful in defining the exploration duration and the regret results.

**Lemma A.3** (Strongly Stabilizable Neighborhood)**.** *Under Assumptions 2.1 & 2.3, for any $\varepsilon \leq \min\{\sqrt{\bar{\sigma}_w^2 n D / C_0}, \epsilon\}$, such that $\|\Theta' - \Theta_*\| \leq \varepsilon$ for any $(\kappa, \gamma)$-stabilizable system $\Theta_*$, $K(\Theta')$ produces $(\kappa', \gamma')$-stable closed-loop dynamics on $\Theta_*$ where $\kappa' = \kappa\sqrt{2}$ and $\gamma' = \gamma/2$.*

*Proof.* Under Assumptions 2.1 & 2.3, for the given choice of $\varepsilon$, we have $\varepsilon \leq \min\{\sqrt{\mathcal{J}/C_0}, \epsilon\}$, thus we obtain $J(K(\Theta'), A_*, B_*, Q, R) \leq 2\mathcal{J}$. Plugging this into Lemma A.2 gives the presented result. □

# B ExpOpt

In this section, we provide the pseudocode of ExpOpt explicitly. The delay in the controller is a technical consiquence in order to lower bound the smallest singular value of the regularized design matrix $V_t$ in Appendix C.

**Algorithm 2** ExpOpt

**Input:** $S > 0$, $\delta > 0$, $\lambda > 0$, $Q$, $R$, $\sigma_w$, $\sigma_\nu$, $H_0$, Setting, Initial Exploration
**if** Setting = Controllable **then**
  Set $\mathbb{1}_{con} = 1$, choose $\mathcal{S} = \mathcal{S}_c$ & set $T_w = T_c$
**else**
  Set $\mathbb{1}_{con} = 0$, choose $\mathcal{S} = \mathcal{S}_s$ & set $T_w = T_s$
**if** Initial Exploration = Yes **then**
  Set $\mathbb{1}_{exp} = 1$, **else** set $\mathbb{1}_{exp} = 0$
Set $V_0 = \lambda I$, $\hat{\Theta}_0 = 0$, $\tau = 0$
$\tilde{\Theta}_0 = \arg\min_{\Theta \in \mathcal{C}_0 \cap \mathcal{S}} J(\Theta)$
**for** $t = 0, \ldots, T$ **do**
  **if** $(\det(V_t) > 2\det(V_0))$ **and** $(\mathbb{1}_{con}$ **or** $(t - \tau > H_0))$ **then**
    Estimate $\hat{\Theta}_t$ using (5)
    Find $\tilde{\Theta}_t$ such that $J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) + \frac{1}{\sqrt{t}}$
    Set $V_0 = V_t$ and $\tau = t$.
  **else**
    $\tilde{\Theta}_t = \tilde{\Theta}_{t-1}$
  **if** $\mathbb{1}_{exp}$ **and** $(t < T_w)$ **then**
    Deploy control input $u_t = K(\tilde{\Theta}_{t-1})x_t + \nu_t$ Optimism + Exploration
  **else**
    Deploy control input $u_t = K(\tilde{\Theta}_{t-1})x_t$ Optimism
  Observe $x_{t+1}$ and set $V_{t+1} = V_t + z_t z_t^\top$ for $z_t = [x_t^\top \; u_t^\top]^\top$

# C Smallest Singular Value of Regularized Design Matrix $V_t$

In this section, we show that during the additional exploration, ExpOpt provides persistently exciting inputs, which will be used to enable reaching a stabilizing neighborhood around the system parameters. In other words, we will lower bound the smallest eigenvalue of the regularized design matrix, $V_t$. The analysis generalizes the lower bound on smallest eigenvalue of the sample covariance matrix in Theorem 20 of [Cohen et al., 2019] for the general case of subgaussian noise.

For the state $x_t$, and input $u_t$, we have:

$$x_t = A_* x_{t-1} + B_* u_{t-1} + w_{t-1}, \quad \textit{and} \quad u_t = K(\tilde{\Theta}_{t-1})x_t + \nu_t \tag{10}$$

Let $\xi_t = z_t - \mathbb{E}\left[z_t | \mathcal{F}_{t-1}\right]$. Using the equalities in (10), and the fact that $w_t$ and $\nu_t$ are $\mathcal{F}_t$ measurable, we write $\mathbb{E}\left[\xi_t \xi_t^\top | \mathcal{F}_{t-1}\right]$ as follows.

$$
\begin{aligned}
\mathbb{E}\left[\xi_t\xi_t^\top|\mathcal{F}_{t-1}\right] &= \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix} \mathbb{E}\left[w_t w_t^\top|\mathcal{F}_{t-1}\right] \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{E}\left[\nu_t\nu_t^\top|\mathcal{F}_{t-1}\right] \end{pmatrix} \\
&= \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix} (\bar{\sigma}_w^2 I) \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 \\ 0 & \sigma_\nu^2 I \end{pmatrix} \quad (11) \\
&= \begin{pmatrix} \bar{\sigma}_w^2 I & \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1})^\top \\ \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1}) & \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1})K(\tilde{\Theta}_{t-1})^\top + 2\kappa^2\bar{\sigma}_w^2 I \end{pmatrix} \quad (12) \\
&\succeq \bar{\sigma}_w^2 \begin{pmatrix} I & K(\tilde{\Theta}_{t-1})^\top \\ K(\tilde{\Theta}_{t-1}) & 2K(\tilde{\Theta}_{t-1})K(\tilde{\Theta}_{t-1})^\top + I/2 \end{pmatrix} \quad (13) \\
&= \frac{\bar{\sigma}_w^2}{2} I + \bar{\sigma}_w^2 \begin{pmatrix} \frac{1}{\sqrt{2}} I \\ \sqrt{2}K(\tilde{\Theta}_{t-1}) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} I \\ \sqrt{2}K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top \quad (14) \\
&\succeq \frac{\bar{\sigma}_w^2}{2} I \quad (15)
\end{aligned}
$$

where (12) follows from $\sigma_\nu^2 = 2\kappa^2\bar{\sigma}_w^2$ and (13) follows from the fact that $\kappa \geq 1$ and $\|K(\tilde{\Theta}_{t-1})\| \leq \kappa$ for all $t$. Let $s_t = v^\top \xi_t$ for any unit vector $v \in \mathbb{R}^{n+d}$. (15) shows that that $\mathrm{Var}\left[s_t|\mathcal{F}_{t-1}\right] \geq \frac{\bar{\sigma}_w^2}{2}$.

**Lemma C.1.** *Suppose the system is stabilizable and we use stabilizable variant of* ExpOpt *with enforced exploration. Denote $s_t = v^\top \xi_t$ where $v \in \mathbb{R}^{n+d}$ is any unit vector. For a given positive $\sigma_1^2$, let $E_t$ be an indicator random variable that equals 1 if $s_t^2 > \sigma_1^2$ and 0 otherwise. Then for any positive $\sigma_1^2$, and $\sigma_2^2$, such that $\sigma_1^2 \leq \sigma_2^2$, we have*

$$
\mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] \geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_\nu^2(1 + \frac{\sigma_2^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})}{\sigma_2^2} \quad (16)
$$

Note that, for any $\bar{\sigma}_\nu \geq \bar{\sigma}_w$, there is a pair $(\sigma_1^2, \sigma_2^2)$ such that the right hand side of (16) is positive.

*Proof.* Using the lower bound on the variance of $s_t$, we have,

$$
\begin{aligned}
\frac{\bar{\sigma}_w^2}{2} &\leq \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 < \sigma_1^2)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] \\
&\leq \sigma_1^2 + \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right]
\end{aligned}
$$

Now, deploying the fact that both $\nu_t$ and $w_t$, for any t, are sub-Gaussian given $\mathcal{F}_{t-1}$, have that $\xi_t$ is also sub-Gaussian vector. Therefore, $s_t$ is a sub-Gaussian random variable with parameter $\bar{\sigma}_\nu$, where $\bar{\sigma}_\nu := ((1+\kappa)^2 + 2\kappa^2)\sigma_w^2$.

$$
\begin{aligned}
\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 &\leq \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_2^2 \geq s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_2^2)|\mathcal{F}_{t-1}\right] \quad (17)
\end{aligned}
$$

For the second term in the right hand side of the (17), under the considerations of Fubini's and Radon–Nikodym theorems, we derive the following equality,

$$\begin{aligned}
\int_{s^2\geq\sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1})ds^2 &= \int_{s^2\geq\sigma_2^2}\int_{s'^2\geq s^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2}ds'^2 ds^2 \\
&= \int_{s'^2\geq\sigma_2^2}\int_{s'^2\geq s^2\geq\sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2}ds'^2 ds^2 \\
&= \int_{s'^2\geq\sigma_2^2}\int_{s'^2\geq s^2\geq\sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2}ds^2 ds'^2 \\
&= \int_{s'^2\geq\sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2}(s'^2-\sigma_2^2) ds'^2 \\
&= \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2\geq\sigma_2^2)|\mathcal{F}_{t-1}\right] - \sigma_2^2\int_{s'^2\geq\sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2}ds'^2 \\
&= \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2\geq\sigma_2^2)|\mathcal{F}_{t-1}\right] - \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}),
\end{aligned}$$

resulting in the following equality,

$$\mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2\geq\sigma_2^2)|\mathcal{F}_{t-1}\right] = \int_{s^2\geq\sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}). \tag{18}$$

Using this equality, we extend the (17) as follows,

$$\begin{aligned}
\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 &\leq \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_2^2\geq s_t^2\geq\sigma_1^2)|\mathcal{F}_{t-1}\right] + \int_{s^2\geq\sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}) \\
&\leq \sigma_2^2\ \mathbb{E}\left[\mathbb{1}(\sigma_2^2\geq s_t^2\geq\sigma_1^2)|\mathcal{F}_{t-1}\right] + \int_{s^2\geq\sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}) \\
&\leq \sigma_2^2\ \mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] + \int_{s^2\geq\sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}). \qquad (19)
\end{aligned}$$

Rearranging this inequality, we have,

$$\begin{aligned}
\mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] &\geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - \int_{s^2\geq\sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1})ds^2 - \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1})}{\sigma_2^2} \\
&\geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 2\int_{s^2\geq\sigma_2^2} \exp(\frac{-s^2}{2\bar{\sigma}_\nu^2})ds^2 - 2\sigma_2^2\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})}{\sigma_2^2} \\
&\geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_\nu^2\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2}) - 2\sigma_2^2\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})}{\sigma_2^2} \\
&= \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_\nu^2(1+\frac{\sigma_2^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})}{\sigma_2^2} \qquad (20)
\end{aligned}$$

The inequality in (20) holds for any $\sigma_1^2 \leq \sigma_2^2$, therefore, the stated lower-bound on $\mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right]$ in the main statement holds.

□

For the choices of $\sigma_1^2$ and $\sigma_2^2$ that makes right hand side of (16), let $c_p$ denote the right hand side of (16), $c_p = \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_\nu^2(1+\frac{\sigma_2^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})}{\sigma_2^2}$.

**Lemma C.2.** *Consider $\bar{s}_t = v^\top z_t$ where $v \in \mathbb{R}^{n+d}$ is any unit vector. Let $\bar{E}_t$ be an indicator random variable that equal 1 if $\bar{s}_t^2 > \sigma_1^2/4$ and 0 otherwise. Then, there exist a positive pair $\sigma_1^2$, and $\sigma_2^2$, and a constant $c_p > 0$, such that $\mathbb{E}\left[\bar{E}_t|\mathcal{F}_{t-1}\right] \geq c_p' > 0$.*

*Proof.* Using the Lemma C.1, we know that for $s_t = v^\top \xi_t$, we have $|s_t| \geq \sigma_1$ with a non-zero probability $c_p$. On the other hand, we have that,

$$\bar{s}_t = v^\top z_t = v^\top \xi_t + v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right] = s_t + v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]$$

Therefore, we have, $|\bar{s}_t| = \left|s_t + v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]\right|$. Using this equality, if $\left|v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]\right| \leq \sigma_1/2$, since $|s_t| \geq \sigma_1$ with probability $c_p$, we have $|\bar{s}_t| \geq \sigma_1/2$ with probability $c_p$.

In the following, we consider the case where $\left|v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]\right| \geq \sigma_1/2$. For a constant $\sigma_3$, using a similar derivation as in (18) and (19), we have

$$\begin{aligned}
\mathbb{E}\left[s_t^2|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(s_t^2 \geq \sigma_3^2)|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + 4\bar{\sigma}_\nu^2(1+\frac{\sigma_2^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})
\end{aligned}$$

Using the lower bound in the variance results in,

$$\frac{\bar{\sigma}_w^2}{2} \leq \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + 4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})$$

Therefore,

$$\begin{aligned}
\frac{\bar{\sigma}_w^2}{2} - 4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}) &\leq \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] \\
&= \sigma_3^2\left(\mathbb{E}\left[\frac{s_t^2}{\sigma_3^2}\mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{s_t^2}{\sigma_3^2}\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right]\right) \\
&\leq \sigma_3^2\left(\mathbb{E}\left[\frac{|s_t|}{\sigma_3}\mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right]\right)
\end{aligned} \tag{21}$$

Note the for a large enough $\sigma_3$, the second term on the left hand side vanishes. Since we have $\mathbb{E}\left[s_t|\mathcal{F}_{t-1}\right] = 0$, we write the following, to further analyze the right hand side of (21),

$$\begin{aligned}
&\mathbb{E}\left[s_t|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[s_t\mathbb{1}(s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t\mathbb{1}(s_t > 0)|\mathcal{F}_{t-1}\right] = 0 \\
&\rightarrow \mathbb{E}\left[|s_t|\mathbb{1}(s_t < 0)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[s_t\mathbb{1}(s_t > 0)|\mathcal{F}_{t-1}\right]
\end{aligned}$$

Note that, since $s_t$ is sub-Gaussian variable, and has bounded away from zero variance, we have $\mathbb{E}\left[\mathbb{1}(s_t<0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[\mathbb{1}(s_t>0)|\mathcal{F}_{t-1}\right]$ is bounded away from zero. We write this equality as follows:

$$\begin{aligned}&\mathbb{E}\left[|s_t|\mathbb{1}(-\sigma_3<s_t<0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[|s_t|\mathbb{1}(s_t\le-\sigma_3)|\mathcal{F}_{t-1}\right]\\&\qquad=\mathbb{E}\left[s_t\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[s_t\mathbb{1}(s_t\ge\sigma_3)|\mathcal{F}_{t-1}\right]\end{aligned}$$

With rearranging this equality, and upper bounding the first term on the left hand side, we have

$$\begin{aligned}\mathbb{E}\left[|s_t|\mathbb{1}(-\sigma_3<s_t<0)|\mathcal{F}_{t-1}\right]&\le\mathbb{E}\left[s_t\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[s_t\mathbb{1}(s_t\ge\sigma_3)|\mathcal{F}_{t-1}\right]\\&\le\mathbb{E}\left[s_t\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\end{aligned}\tag{22}$$

similarly we have

$$\mathbb{E}\left[s_t\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]\le\mathbb{E}\left[|s_t|\mathbb{1}(-\sigma_3<s_t<0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\tag{23}$$

Using the inequality (22) on the right hand side of (21), we have

$$\begin{aligned}\frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{\sigma_3^2}&\le\mathbb{E}\left[\frac{|s_t|}{\sigma_3}\mathbb{1}(-\sigma_3<s_t<0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]\\&\le2\mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\\&\le2\mathbb{E}\left[\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\\&\le2\mathbb{E}\left[\mathbb{1}(s_t>0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\end{aligned}$$

Similarly, using (22) on the right hand side of (21) we have

$$\begin{aligned}\frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{\sigma_3^2}&\le\mathbb{E}\left[\frac{|s_t|}{\sigma_3}\mathbb{1}(-\sigma_3<s_t<0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]\\&\le2\mathbb{E}\left[\mathbb{1}(s_t<0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\end{aligned}$$

Therefore, it results in the two following lower bounds,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}(s_t<0)|\mathcal{F}_{t-1}\right]&\ge\frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{2\sigma_3^2}-0.5\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\\\mathbb{E}\left[\mathbb{1}(s_t>0)|\mathcal{F}_{t-1}\right]&\ge\frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{2\sigma_3^2}-0.5\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\end{aligned}\tag{24}$$

Choosing $\sigma_3$ sufficiently large results in the right hand sides in inequalities (24) to be positive and bounded away form zero. Let $c''_p > 0$ denote the right hand sides in the (24). We use this fact to analyze $\bar{s}_t$ when $\left|v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]\right| \geq \sigma_1/2$.

When $v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right] \geq \sigma_1/2$, since probability $c''_p$, $s_t$ is positive, therefore, $|\bar{s}_t| \geq \sigma_1/2$ with probability $c''_p$. When $v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right] \leq -\sigma_1/2$, since probability $c''_p$, $s_t$ is negative, therefore, $|\bar{s}_t| \geq \sigma_1/2$ with probability $c''_p$.

Therefore, overall, with probability $c'_p := \min\{c_p, c''_p\}$, we have that $|\bar{s}_t| \geq \sigma_1/2$, resulting in the statement of the lemma.

□

**Lemma C.3** (Persistence of Excitation During the Extra Exploration)**.** *When the exploration duration* $T_w \geq \frac{6n}{c'_p}\log(12/\delta)$*, then with probability at least* $1-\delta$*,* ExpOpt *has*

$$\lambda_{\min}(V_{T_w}) \geq \sigma_\star^2 T_w,$$

*for* $\sigma_\star^2 = \frac{c'_p\sigma_1^2}{16}$.

*Proof.* Let $U_t = \bar{E}_t - \mathbb{E}_t\left[\bar{E}_t|\mathcal{F}_{t-1}\right]$. Then $U_t$ is a martingale difference sequence with $|U_t| \leq 1$. Applying Azuma's inequality, we have that with probability at least $1-\delta$

$$\sum_{t=1}^{T_w} U_t \geq -\sqrt{2T_w \log\frac{1}{\delta}}$$

Using the Lemma C.2, we have

$$\begin{aligned}\sum_{t}^{T_w} \bar{E}_t &\geq \sum_{t}^{T_w} \mathbb{E}_t\left[\bar{E}_t|\mathcal{F}_{t-n}\right] - \sqrt{2T_w\log\frac{1}{\delta}}\\ &\geq c'_p T_w - \sqrt{2T_w\log\frac{1}{\delta}}\end{aligned}$$

where for $T_w \geq 8\log(1/\delta)/c'^2_p$, we have $\sum_t^{T_w}\bar{E}_t \geq \frac{c'_p}{2}T_w$. Now, for any unit vector $v$, define $\bar{s}_t = v^\top z_t$, therefore from the definition of $\bar{E}_t$ we have,

$$v^\top V_{T_w} v = \sum_t^{T_w} \bar{s}_t^2 \geq \bar{E}_t\sigma_1^2/4 \geq \frac{c'_p\sigma_1^2}{8}T_w$$

This inequality hold for a given $v$. In the following we show a similar inequality for all $v$ together. Similar to the Theorem 20 in [Cohen et al., 2019], consider a $1/4$-net of $\mathbb{S}^{n+d-1}$, $\mathcal{N}(1/4)$ and set $M_{T_w} := \{V_{T_w}^{-1/2}v/\|V_{T_w}^{-1/2}v\| : v \in \mathcal{N}(1/4)\}$. These two sets have at most $12^{n+d-1}$ members. Using union bound over members of this set, when $T_w \geq \frac{20}{c'^2_p}((n+d)+\log(1/\delta))$, we have that $v^\top V_{T_w} v \geq \frac{c'_p\sigma_1^2}{8}T_w$ for all $v \in M_{T_w}$ with a probability at least $1-\delta$. Using the definition of members in $M_{T_w}$, for each $v \in \mathcal{N}(1/4)$, we have $v^\top V_{T_w}^{-1} v \leq \frac{8}{T_w c'_p\sigma_1^2}$. Let $v_n$ denote the eigenvector of the largest eigenvalue of $V_{T_w}^{-1}$, and a vector $v' \in \mathcal{N}(1/4)$ such that $\|v_n - v'\| \leq 1/4$. Then we have

$$\begin{aligned}\|V_{T_w}^{-1}\| &= v_n^\top V_{T_w}^{-1} v_n = v'^\top V_{T_w}^{-1} v' + (v_n - v')^\top V_{T_w}^{-1}(z_n + v')\\ &\leq \frac{8}{T_w c'_p\sigma_1^2} + \|v_n - v'\|\|V_{T_w}^{-1}\|\|z_n + v'\| \leq \frac{8}{T_w c'_p\sigma_1^2} + \|V_{T_w}^{-1}\|/2\end{aligned}$$

Rearranging, we get that $\|V_{T_w}^{-1}\| \leq \frac{16}{T_w c_p' \sigma_1^2}$. Therefore, the advertised bound holds for $T_w \geq \frac{20}{c_p'^2}((n+d) + \log(1/\delta))$ with probability at least $1-\delta$. ☐

# D System Identification and Confidence Set Construction, Proof Lemma 4.1

To have completeness, for the proof of Lemma 4.1 we first provide the proof for confidence set construction borrowed from Abbasi-Yadkori and Szepesvári [2011], since Lemma 4.1 builds upon this confidence set construction.

*Proof.* Define $\Theta_*^\top = [A, B]$ and $z_t = \left[x_t^\top u_t^\top\right]^\top$. The system in (1) can be characterized equivalently as

$$x_{t+1} = \Theta_*^\top z_t + w_t$$

Given a single input-output trajectory $\{x_t, u_t\}_{t=1}^T$, one can rewrite the input-output relationship as,

$$X_T = Z_T \Theta_* + W_T \tag{25}$$

for

$$X_T = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_{T-1}^\top \\ x_T^\top \end{bmatrix} \in \mathbb{R}^{T\times n} \quad Z_T = \begin{bmatrix} z_1^\top \\ z_2^\top \\ \vdots \\ z_{T-1}^\top \\ z_T^\top \end{bmatrix} \in \mathbb{R}^{T\times (n+d)} \quad W_T = \begin{bmatrix} w_1^\top \\ w_2^\top \\ \vdots \\ w_{T-1}^\top \\ w_T^\top \end{bmatrix} \in \mathbb{R}^{T\times n}. \tag{26}$$

Then, we estimate $\Theta_*$ by solving the following least square problem,

$$\begin{aligned}
\hat{\Theta}_T &= \arg\min_X ||X_T - Z_T X||_F^2 + \lambda ||X||_F^2 \\
&= (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top X_T \\
&= (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top W_T + (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top Z_T \Theta_* + \lambda (Z_T^\top Z_T + \lambda I)^{-1}\Theta_* - \lambda (Z_T^\top Z_T + \lambda I)^{-1}\Theta_* \\
&= (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top W_T + \Theta_* - \lambda (Z_T^\top Z_T + \lambda I)^{-1}\Theta_*
\end{aligned}$$

The confidence set is obtained using the expression for $\hat{\Theta}_T$ and subgaussianity of the $w_t$,

$$\begin{aligned}
|\operatorname{Tr}((\hat{\Theta}_T - \Theta_*)^\top X)| &= |\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \lambda I)^{-1} X) - \lambda \operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \lambda I)^{-1} X)| \\
&\leq |\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \lambda I)^{-1} X)| + \lambda |\operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \lambda I)^{-1} X)| \\
&\leq \sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T + \lambda I)^{-1} X) \operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top W_T)} \\
&\quad + \lambda \sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T + \lambda I)^{-1} X) \operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \lambda I)^{-1} \Theta_*)}, \qquad (27) \\
&= \sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T + \lambda I)^{-1} X)} \left[\sqrt{\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top W_T)} + \lambda \sqrt{\operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \lambda I)^{-1} \Theta_*)}\right]
\end{aligned}$$

where (27) follows from $|\operatorname{Tr}(A^\top BC)| \leq \sqrt{\operatorname{Tr}(A^\top BA)\operatorname{Tr}(C^\top BC)}$ for square positive definite B due to Cauchy Schwarz (weighted inner-product). For $X = (Z_T^\top Z_T + \lambda I)(\hat{\Theta}_T - \Theta_*)$, we get

$$\sqrt{\operatorname{Tr}((\hat{\Theta}_T - \Theta_*)^\top (Z_T^\top Z_T + \lambda I)(\hat{\Theta}_T - \Theta_*))} \leq \sqrt{\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top W_T)} + \sqrt{\lambda}\sqrt{\operatorname{Tr}(\Theta_*^\top \Theta_*)}$$

Let $\mathcal{S}_T = Z_T^\top W_T \in \mathbb{R}^{(n+d)\times n}$ and $s_i$ denote the columns of it. Also, let $V_T = (Z_T^\top Z_T + \lambda I)$. Thus,

$$\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top W_T) = \operatorname{Tr}(\mathcal{S}_T^\top V_T^{-1} \mathcal{S}_T) = \sum_{i=1}^{n} s_i^\top V_T^{-1} s_i = \sum_{i=1}^{n} \|s_i\|^2_{V_T^{-1}}. \tag{28}$$

Notice that $s_i = \sum_{j=1}^{T} w_{j,i} z_j$ where $w_{j,i}$ is the $i$'th element of $w_j$. From Assumption 2.1, we have that $w_{j,i}$ is $\sigma_w$-subgaussian, thus we can use Theorem 5 to show that,

$$\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top W_T) \leq 2n\sigma_w^2 \log\left(\frac{\det(V_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right). \tag{29}$$

with probability $1-\delta$. From Assumptions 2.2 or 2.3, we also have that $\sqrt{\operatorname{Tr}(\Theta_*^\top \Theta_*)} \leq S$. Combining these gives the self-normalized confidence set or the model estimate:

$$\operatorname{Tr}((\hat{\Theta}_T - \Theta_*)^\top V_T (\hat{\Theta}_T - \Theta_*)) \leq \left(\sigma_w \sqrt{2n\log\left(\frac{\det(V_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \sqrt{\lambda}S\right)^2. \tag{30}$$

Notice that we have $\operatorname{Tr}((\hat{\Theta}_T - \Theta_*)^\top V_T (\hat{\Theta}_T - \Theta_*)) \geq \lambda_{\min}(V_T)\|\hat{\Theta}_T - \Theta_*\|_F^2$. Therefore,

$$\|\hat{\Theta}_T - \Theta_*\|_2 \leq \frac{1}{\sqrt{\lambda_{\min}(V_T)}}\left(\sigma_w \sqrt{2n\log\left(\frac{\det(V_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \sqrt{\lambda}S\right) \tag{31}$$

To complete the proof, we need a lower bound on $\lambda_{\min}(V_{T_w})$. Using Lemma C.3, we obtain the following with probability at least $1-2\delta$:

$$\|\hat{\Theta}_{T_w} - \Theta_*\|_2 \leq \frac{1}{\sigma_\star\sqrt{T_w}}\left(\sigma_w \sqrt{2n\log\left(\frac{\det(V_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \sqrt{\lambda}S\right).$$

From Lemma 4.2 for both controllable and stabilizable systems, for $t \leq T_w$, we have that $\|z_t\| \leq c(n+d)^{n+d}$ with probability at least $1-2\delta$, for some constant $c$. Combining this with Lemma I.1,

$$\|\hat{\Theta}_{T_w} - \Theta_*\|_2 \leq \frac{\kappa_e}{\sqrt{T_w}} := \frac{1}{\sqrt{T_w}}\left(\frac{\sigma_w}{\sigma_\star}\sqrt{n(n+d)\log\left(1 + \frac{cT(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda(n+d)}\right) + 2n\log\frac{1}{\delta}} + \sqrt{\lambda}S\right). \tag{32}$$

□

## E Boundedness of States, Proof of Lemma 4.2

In this section, we will provide bounds on states for controllable and stabilizable systems with and without additional exploration. First define the following. For controllable systems, let

$$T_c = \frac{4(1+\kappa)^2\kappa_e^2}{(1-\Upsilon)^2}$$

such that for $T_w > T_c$, we have $\|\hat{\Theta}_{T_w} - \Theta_*\|_2 \leq \frac{1-\Upsilon}{2(1+\kappa)}$ with probability at least $1-2\delta$. Similarly for stabilizable systems, let

$$T_s = \frac{\kappa_e^2}{\min\{\bar{\sigma}_w^2 nD/C_0, \epsilon^2\}}$$

such that for $T_w > T_s$, we have $\|\hat{\Theta}_{T_w} - \Theta_*\|_2 \leq \min\{\sqrt{\bar{\sigma}_w^2 nD/C_0}, \epsilon\}$ with probability at least $1-2\delta$. Notice that due to Lemma A.3 and as shown in the following section for controllability, these guarantee the stability of the closed-loop dynamics for deploying optimistic controller for the remaining part of ExpOpt.

### E.1 Controllable System

In this section we will first recall the boundedness of state result from Abbasi-Yadkori and Szepesvári [2011]. Since our input is $u_t = K(\tilde{\Theta}_{t-1})x_t + \nu_t$ for $t \leq T_c$ and $u_t = K(\tilde{\Theta}_{t-1})x_t$ for $t > T_c$, we will first include the effect of additional uniform exploration in the bound of the state. Then we will provide a new stability analysis for the states at $t > T_c$.

$\mathbf{t \leq T_c}$ :

Choose an error probability, $\delta > 0$. The following events are modified from Abbasi-Yadkori and Szepesvári [2011]. In the probability space $\Omega$:

- The event that the confidence sets hold for $s = 0, \dots, T$,

$$\mathcal{E}_t = \{\omega \in \Omega : \forall s \leq T, \quad \Theta_* \in \mathcal{C}_s(\delta)\}$$

- The event that the state vector stays "small" for $s = 0, \dots, T_w$,

$$\mathcal{F}_t^{[c]} = \{\omega \in \Omega : \forall s \leq T_w, \quad \|x_s\| \leq \alpha_t\}$$

where

$$\alpha_t = \frac{1}{1-\Upsilon}\left(\frac{\eta}{\Upsilon}\right)^{n+d}\left[GZ_t^{\frac{n+d}{n+d+1}}\beta_t(\delta)^{\frac{1}{2(n+d+1)}} + (\|B_*\|\sigma_\nu + \sigma_w)\sqrt{2n\log\frac{nt}{\delta}}\right]$$

$$\eta = \max\left(1, \sup_{\Theta\in\mathcal{S}}\|A_* + B_*K(\Theta)\|\right), \qquad Z_T = \max_{1\leq t\leq T}\|z_t\|$$

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

and $H$ is any number satisfying

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

Notice that $\mathcal{E}_1 \supseteq \mathcal{E}_2 \supseteq \ldots \supseteq \mathcal{E}_T$ and $\mathcal{F}_1^{[c]} \supseteq \mathcal{F}_2^{[c]} \supseteq \ldots \supseteq \mathcal{F}_{T_c}^{[c]}$. This means considering the probability of last event is sufficient in lower bounding all event happening simultaneously. In Abbasi-Yadkori and Szepesvári [2011], an argument regarding projection onto subspaces is constructed to show that the norm of the state is well-controlled except $n+d$ times at most in any horizon $T$. The set of time steps that is not well-controlled are denoted as $\mathcal{T}_t$. The given lemma shows how well controlled $\|(\Theta_* - \hat{\Theta}_t)^\top z_t\|$ is besides $\mathcal{T}_t$.

**Lemma E.1** (Abbasi-Yadkori and Szepesvári [2011]). *We have that for any* $0 \leq t \leq T$,

$$\max_{s\leq t, s\notin T_t} \left\|(\Theta_* - \hat{\Theta}_s)^\top z_s\right\| \leq GZ_t^{\frac{n+d}{n+d}} \beta_t(\delta/4)^{\frac{1}{2(n+d+1)}}.$$

Building upon this result we will bound the state during the exploration phase. This bound follows the proof of Lemma 4 in [Abbasi-Yadkori and Szepesvári, 2011]. One can write the state update as

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

where

$$\Gamma_t = \begin{cases} \tilde{A}_{t-1} + \tilde{B}_{t-1}K(\tilde{\Theta}_{t-1}) & t \notin \mathcal{T}_T \\ A_* + B_*K(\tilde{\Theta}_{t-1}) & t \in \mathcal{T}_T \end{cases} \quad \text{and} \quad r_t = \begin{cases} (\Theta_* - \tilde{\Theta}_{t-1})^\top z_t + B_*\nu_t + w_t & t \notin \mathcal{T}_T \\ B_*\nu_t + w_t & t \in \mathcal{T}_T \end{cases} \tag{33}$$

Thus, using the fact that $x_0 = 0$, we can obtain the following roll out for $x_t$,

$$\begin{aligned} x_t &= \Gamma_{t-1}x_{t-1} + r_{t-1} = \Gamma_{t-1}\left(\Gamma_{t-2}x_{t-2} + r_{t-2}\right) + r_t \\ &= \Gamma_{t-1}\Gamma_{t-2}\Gamma_{t-3}x_{t-3} + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} + \Gamma_{t-1}r_{t-1} + r_t \\ &= \Gamma_{t-1}\Gamma_{t-2}\ldots\Gamma_{t-(t-1)}r_1 + \cdots + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} + \Gamma_{t-1}r_{t-1} + r_t \\ &= \sum_{k=1}^{t}\left(\prod_{s=k}^{t-1}\Gamma_s\right) r_k \end{aligned} \tag{34}$$

Recall the following expressions,

$$\eta \geq \max_{t\leq T}\left\|A_* + B_*K(\tilde{\Theta}_t)\right\|, \quad \Upsilon \geq \max_{t\leq T}\left\|\left(\tilde{A}_t + \tilde{B}_tK(\tilde{\Theta}_t)\right)\right\|.$$

Using these, we have that

$$\begin{aligned} \|x_t\| &\leq \left(\frac{\eta}{\Upsilon}\right)^{n+d}\sum_{k=1}^{t}\Upsilon^{t-k+1}\|r_k\| \\ &\leq \frac{1}{1-\Upsilon}\left(\frac{\eta}{\Upsilon}\right)^{n+d}\max_{1\leq k\leq t}\|r_k\| \end{aligned}$$

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 E.1. The second term involves summation of independent $\|B_*\|\sigma_\nu$ and $\sigma_w$ subgaussian vectors. Using Lemma I.2 with a union bound argument, for all $k \leq t$, $\|B_*\nu_k + w_k\| \leq (\|B_*\|\sigma_\nu + \sigma_w)\sqrt{2n \log \frac{nt}{\delta}}$ with probability at least $1-\delta$. Therefore, on the event of $\mathcal{E}$,

$$\|x_t\| \leq \frac{1}{1-\Upsilon}\left(\frac{\eta}{\Upsilon}\right)^{n+d}\left[G Z_t^{\frac{n+d}{n+d}} \beta_t(\delta)^{\frac{1}{2(n+d+1)}} + (\|B_*\|\sigma_\nu + \sigma_w)\sqrt{2n\log\frac{nt}{\delta}}\right] \tag{35}$$

for $t \leq T_w$. Using union bound, we can deduce that $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$ holds with probability at least $1-2\delta$. 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

$$\|x_t\| \leq c'(n+d)^{n+d}. \tag{36}$$

for some large enough constant $c'$.

$\mathbf{t > T_c}$ :

Recall that once $t \geq T_c$, the controller stops using the exploratory component $\nu_t$. Thus, the state has the following dynamics,

$$\begin{aligned} x_{t+1} &= (A_* + B_*K(\tilde{\Theta}_{t-1}))x_t + w_t \\ &= \left(A_* - \tilde{A}_{t-1} + \tilde{A}_{t-1} + B_*K(\tilde{\Theta}_{t-1}) - \tilde{B}_{t-1}K(\tilde{\Theta}_{t-1}) + \tilde{B}_{t-1}K(\tilde{\Theta}_{t-1})\right)x_t + w_t. \end{aligned} \tag{37}$$

Hence, it propagates according to the linear system given in equation (37) with closed loop dynamics $\mathbf{M_t} = \left(A_* - \tilde{A}_{t-1} + (B_* - \tilde{B}_{t-1})K(\tilde{\Theta}_{t-1}) + \tilde{A}_{t-1} + \tilde{B}_{t-1}K(\tilde{\Theta}_{t-1})\right)$ driven by the process $w_t$ with $x_{T_c}$ as the initial state. With the Assumption 2.2, for the given $T_c$, if the event of $\mathcal{E}$ holds, we have $\|\mathbf{M_t}\| < \frac{1+\Upsilon}{2} < 1$ for all $t > T_c$. Then for $t > T_c$,

$$\|x_t\| = \left\| \prod_{i=T_c+1}^{t} \mathbf{M_i} x_{T_c} + \sum_{i=T_c+1}^{t}\left(\prod_{s=i}^{t-1}\mathbf{M_s}\right) w_i \right\| \tag{38}$$

$$\leq \left(\frac{1+\Upsilon}{2}\right)^{t-T_c}\|x_{T_c}\| + \max_{T_c < i \leq t}\|w_i\| \left(\sum_{i=T_c+1}^{t}\left(\frac{1+\Upsilon}{2}\right)^{t-i}\right) \tag{39}$$

$$\leq \left(\frac{1+\Upsilon}{2}\right)^{t-T_c}\|x_{T_c}\| + \frac{2}{1-\Upsilon}\max_{T_c < i \leq t}\|w_i\| \tag{40}$$

Using Lemma I.2 with a union bound argument, we get $\|w_i\| \leq \sigma_w\sqrt{2n\log(n(t-T_c)/\delta)}$ with probability $1-\delta$, for all $t > T_c$. Using (43), for all $T \geq t > T_c$, with probability $1-3\delta$,

$$\|x_t\| \leq \underbrace{\frac{2\sigma_w}{1-\Upsilon}\sqrt{2n\log\frac{n(T-T_c)}{\delta}}}_{\mathcal{X}_s} + \left(\frac{1+\Upsilon}{2}\right)^{t-T_c} c'(n+d)^{n+d} \tag{41}$$

Notice that for $t > T_c + \frac{(n+d)\log(n+d)+\log(c')-\log(\mathcal{X}_s)}{\log\frac{2}{1+\Upsilon}} := T_{r,c}$, the second term in (41) is equal to $\mathcal{X}_s$ which is the same as if the noise is driving the stable system starting at an initial state $x_0 = 0$, *i.e.* the effect of unstable controllers during the exploration is removed. Therefore, for all $T \geq t > T_{r,c}$ we have $\|x_t\| \leq 2\mathcal{X}_s = \frac{4\sigma_w}{1-\Upsilon}\sqrt{2n\log\frac{n(T-T_c)}{\delta}}$.

## E.2 Stabilizable System

Notice that Lemma E.1 does not depend on controllability or the stabilizability of the system. Thus, we will again use Lemma E.1 for all $1 \leq t \leq T$ for the variant of ExpOpt that deploys only OFU principle and for $t \leq T_w$ for the variant of ExpOpt with additional exploration. Then we consider the effect of stabilizing controller for the additional exploration variant of ExpOpt in stabilizable systems.

Recall the following events in the probability space $\Omega$:

- The event that the confidence sets hold for $s = 0, \ldots, T$,

$$\mathcal{E}_t = \{\omega \in \Omega : \forall s \leq T, \quad \Theta_* \in \mathcal{C}_s(\delta)\}$$

- The event that the state vector stays "small" for $s = 0, \ldots, T_w$,

$$\mathcal{F}_t^{[s]} = \{\omega \in \Omega : \forall s \leq T_w, \quad \|x_s\| \leq \bar{\alpha}_t\}$$

where

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

for $\bar{\eta}$ defined in (42) and the rest are the same with controllable setting.

**ExpOpt with only OFU:** Following similarly with the controllable system, we have the same state update in (33) and same roll out for $x_t$ in (34). However, the controller is optimistically designed from set of parameters are $(\kappa, \gamma)$-strongly stabilizable by their optimal controllers. Therefore, we now have

$$\bar{\eta} \geq \max_{t\leq T}\left\|A_* + B_*K(\tilde{\Theta}_t)\right\|, \quad 1-\gamma \geq \max_{t\leq T}\rho\left(\tilde{A}_t + \tilde{B}_tK(\tilde{\Theta}_t)\right). \tag{42}$$

During the exploration phase, since the estimates of $\Theta$ are not refined enough, the closed loop matrix is not stable. In order to have the previous proof go through, we need to satisfy that the epochs that we use a particular optimistic controller is long enough that the state doesn't scale too badly during the exploration. By choosing $H_0 = 2\gamma^{-1}\log(2\kappa\sqrt{2})$, adopting Lemma 39 of Cassel et al. [2020], we guarantee that

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

Again using Lemma 5 of Abbasi-Yadkori and Szepesvári [2011], one can obtain the following bound

$$\|x_t\| \leq c'(n+d)^{n+d}. \tag{43}$$

for some large enough constant $c'$.

**ExpOpt with additional exploration:** The bound for $t \leq T_w$ follows exactly from previous section. We know consider the state after $T_w$. Since once $t \geq T_w$, the controller stops using the exploratory component $\nu_t$, the state follows the dynamics of

$$x_{t+1} = (A_* + B_* K(\tilde{\Theta}_{t-1}))x_t + w_t \tag{44}$$

Similar to controllable setting, denote $\mathbf{M_t} = A_* + B_* K(\tilde{\Theta}_{t-1})$ as the closed loop dynamics of the system. From the choice of $T_s$ for the stabilizable systems, we have that $\mathbf{M_t}$ is $(\kappa\sqrt{2}, \gamma/2)$-strongly stable. Thus, we have $\rho(\mathbf{M_t}) \leq 1-\gamma/2$ for all $t > T_s$ and $\|H_t\|\|H_t^{-1}\| \leq \kappa\sqrt{2}$ for $H_t \succ 0$, such that $\|L_t\| \leq 1-\gamma/2$ for $\mathbf{M_t} = H_t L_t H_t^{-1}$. Then for $T > t > T_s$, if the same policy, $\mathbf{M}$ is applied starting from state $x_{T_s}$, we have

$$\|x_t\| = \left\| \prod_{i=T_s+1}^{t} \mathbf{M} x_{T_s} + \sum_{i=T_s+1}^{t} \left( \prod_{s=i}^{t-1} \mathbf{M} \right) w_i \right\| \tag{45}$$

$$\leq \kappa\sqrt{2}(1-\gamma/2)^{t-T_s}\|x_{T_s}\| + \max_{T_s < i \leq T} \|w_i\| \left( \sum_{i=T_s+1}^{t} \kappa\sqrt{2}(1-\gamma/2)^{t-i+1} \right) \tag{46}$$

$$\leq \kappa\sqrt{2}(1-\gamma/2)^{t-T_s}\|x_{T_s}\| + \frac{2\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)} \tag{47}$$

Note that $H_0 = 2\gamma^{-1}\log(2\kappa\sqrt{2})$. This gives that $\kappa\sqrt{2}(1-\gamma/2)^{H_0} \leq 1/2$. 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 $T_s$, we get

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

$$\leq 2^{-i}\|x_{T_s}\| + \frac{4\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)}$$

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

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

Finally, from (47), we have that

$$\|x_t\| \leq \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)}, \tag{48}$$

for all $t > T_{r,s}$ where

$$T_{r,s} = T_s + \left( (n+d)\log(n+d) - \log(\frac{2\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)}) \right) H_0.$$

## F Regret Decomposition

**ExpOpt without Additional Exploration** The regret of ExpOpt using only OFU yields the same regret decomposition in Section 4.2 of Abbasi-Yadkori and Szepesvári [2011], since the underlying system dynamics is the same. Let $J_*(\Theta_*, w_t)$ denote the optimal average expected cost of an LQR, $\Theta_*$, with $w_t$ disturbances obtained by its optimal controller. Therefore, under the event $\mathcal{E}_T \cap \mathcal{F}_T^{[s]}$ for ExpOpt without additional exploration, we have

$$\textsc{Regret}(T) = \sum_{t=0}^{T} \left( x_t^\top Q x_t + u_t^\top R u_t \right) - T J_*(\Theta_*, w_t) \leq R_1 - R_2 - R_3 + 2\sqrt{T}$$

where

$$R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_{t-1}) x_t - \mathbb{E}\left[ x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \middle| \mathcal{F}_{t-1} \right] \right\} \tag{49}$$

$$R_2 = \sum_{t=0}^{T} \mathbb{E}\left[ x_{t+1}^\top \left( P(\tilde{\Theta}_{t-1}) - P(\tilde{\Theta}_t) \right) x_{t+1} \middle| \mathcal{F}_{t-1} \right] \tag{50}$$

$$R_3 = \sum_{t=0}^{T} \left\{ \left( \tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t \right)^\top P(\tilde{\Theta}_{t-1}) \left( \tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t \right) - \left( A_* x_t + B_* u_t \right)^\top P(\tilde{\Theta}_{t-1}) \left( A_* x_t + B_* u_t \right) \right\}. \tag{51}$$

**ExpOpt with Additional Exploration (Controllable or Stabilizable)** Since for the additional exploration ExpOpt applies independent external perturbation through the controller but still designs the optimistic controller (optimal controller for the optimistically chosen system), one can consider the external perturbation as a component of the underlying system and consider the regret obtained by using the external perturbation separately.

Denote the system evolution noise at time $t$ as $\zeta_t$. For $t \leq T_w$, during the additional exploration, system evolution noise can be considered as $\zeta_t = B_* \nu_t + w_t$ and for $t > T_w$, $\zeta_t = w_t$. Denote the optimal average cost of system $\tilde{\Theta}$ under $\zeta_t$ as $J_*(\tilde{\Theta}, \zeta_t)$. The regret of the algorithm can be decomposed as follows:

$$\text{REGRET}(T) = \sum_{t=0}^{T} \left( x_t^\top Q x_t + u_t^\top R u_t + 2\nu_t^\top R u_t + \nu_t^\top R \nu_t \right) - T J_*(\Theta_*, w_t) \tag{52}$$

where $u_t$ is the optimal controller input for the optimistic system $\tilde{\Theta}_{t-1}$, $\nu_t$ is the noise injected and $x_t$ is the state of the system $\tilde{\Theta}_{t-1}$ with the system evolution noise of $\zeta_t$. From Bellman optimality equations for LQR, [Bertsekas, 1995], we can write the following for any LQR,

$$\begin{aligned} J_*(\tilde{\Theta}_{t-1}, \zeta_t) + x_t^\top P(\tilde{\Theta}_{t-1}) x_t &= x_t^\top Q x_t + u_t^\top R u_t \\ &+ \mathbb{E}\left[ \left( \tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t + \zeta_t \right)^\top P(\tilde{\Theta}_{t-1}) \left( \tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t + \zeta_t \right) \middle| \mathcal{F}_{t-1} \right], \end{aligned}$$

where we considered the optimistic system, $\tilde{\Theta}_{t-1}$. Following the decomposition used in without additional exploration, we get,

$$
\begin{aligned}
&J_*(\tilde{\Theta}_{t-1}, \zeta_t) + x_t^\top P(\tilde{\Theta}_{t-1}) x_t \\
&\qquad = x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[\left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right)^\top P(\tilde{\Theta}_{t-1})\left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right)\big|\mathcal{F}_{t-1}\right] \\
&\qquad\quad + \mathbb{E}\left[\zeta_t^\top P(\tilde{\Theta}_{t-1})\zeta_t \big| \mathcal{F}_{t-1}\right] \\
&\qquad = x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[\left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right)^\top P(\tilde{\Theta}_{t-1})\left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right)\big|\mathcal{F}_{t-1}\right] \\
&\qquad\quad + \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_{t-1}) x_{t+1} \big| \mathcal{F}_{t-1}\right] - \mathbb{E}\left[\left(A_* x_t + B_* u_t\right)^\top P(\tilde{\Theta}_{t-1})\left(A_* x_t + B_* u_t\right)\big|\mathcal{F}_{t-1}\right] \\
&\qquad = x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_{t-1}) x_{t+1} \big| \mathcal{F}_{t-1}\right] \\
&\qquad\quad + \left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right)^\top P(\tilde{\Theta}_{t-1})\left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right) - \left(A_* x_t + B_* u_t\right)^\top P(\tilde{\Theta}_{t-1})\left(A_* x_t + B_* u_t\right)
\end{aligned}
$$

where in the one before the last equality we use $x_{t+1} = A_* x_t + B_* u_t + \zeta_t$ and the martingale property of the noise. Hence,

$$
\sum_{t=0}^{T} J_*(\tilde{\Theta}_{t-1}, \zeta_t) + R_1^\zeta = \sum_{t=0}^{T}\left(x_t^\top Q x_t + u_t^\top R u_t\right) + R_2^\zeta + R_3^\zeta
$$

where

$$
R_1^\zeta = \sum_{t=0}^{T}\left\{x_t^\top P(\tilde{\Theta}_{t-1}) x_t - \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \big| \mathcal{F}_{t-1}\right]\right\} \tag{53}
$$

$$
R_2^\zeta = \sum_{t=0}^{T}\mathbb{E}\left[x_{t+1}^\top\left(P(\tilde{\Theta}_{t-1}) - P(\tilde{\Theta}_t)\right)x_{t+1}\big|\mathcal{F}_{t-1}\right] \tag{54}
$$

$$
R_3^\zeta = \sum_{t=0}^{T}\left\{\left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right)^\top P(\tilde{\Theta}_{t-1})\left(\tilde{A}_{t-1}x_t + \tilde{B}_{t-1}u_t\right) - \left(A_* x_t + B_* u_t\right)^\top P(\tilde{\Theta}_{t-1})\left(A_* x_t + B_* u_t\right)\right\} \tag{55}
$$

Therefore, on $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$,

$$
\begin{aligned}
\sum_{t=0}^{T}\left(x_t^\top Q x_t + u_t^\top R u_t\right) &= \sum_{t=0}^{T} J_*(\tilde{\Theta}_{t-1}, \zeta_t) + R_1^\zeta - R_2^\zeta - R_3^\zeta \\
&= \sum_{t=0}^{T_w}\left(\sigma_\nu^2 \operatorname{Tr}(P(\tilde{\Theta}_{t-1})B_* B_*^\top)\right) + \sum_{t=0}^{T}\left(\bar{\sigma}_w^2 \operatorname{Tr}(P(\tilde{\Theta}_{t-1}))\right) + R_1^\zeta - R_2^\zeta - R_3^\zeta
\end{aligned}
$$

where the last equality follows from the fact that, $J_*(\tilde{\Theta}_{t-1}, \zeta_t) = \operatorname{Tr}(P(\tilde{\Theta}_{t-1})W)$ where $W = \mathbb{E}[\zeta_t \zeta_t^\top | \mathcal{F}_{t-1}]$ for a corresponding filtration $\mathcal{F}_t$. The optimistic choice of $\tilde{\Theta}_t$ provides that $\bar{\sigma}_w^2 \operatorname{Tr}(P(\tilde{\Theta}_{t-1})) = J_*(\tilde{\Theta}_{t-1}, w_t) \leq J_*(\Theta_*, w_t) = \bar{\sigma}_w^2 \operatorname{Tr}(P(\Theta_*))$. Thus we get,

$$\sum_{t=0}^{T}\left(x_t^\top Q x_t + u_t^\top R u_t\right) - TJ_*(\Theta_*, w_t) \leq T_w \max_{0\leq t\leq T_w}\left\{\sigma_\nu^2 \operatorname{Tr}(P(\tilde{\Theta}_{t-1})B_* B_*^\top)\right\} + R_1^\zeta - R_2^\zeta - R_3^\zeta$$

Combining this with (52) and Assumption 2.2 or 2.3,

$$\text{REGRET}(T) \leq \sigma_\nu^2 T_w D \|B_*\|_F^2 + \sum_{t=0}^{T_w}\left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right) + R_1^\zeta - R_2^\zeta - R_3^\zeta.$$

In the next section, we will bound each term individually.

# G Regret Analysis

In this section, we provide the bounds on each term in regret decomposition for both with and without additional exploration. Notice that ExpOpt without additional exploration in stabilizable setting yields the same regret decomposition with Abbasi-Yadkori and Szepesvári [2011] and the bound on the state throughout the algorithm is again $c'(n+d)^{n+d}$. Therefore, the main difference compared to OFULQ which is tailored for controllable systems is the update rule due to maintaining boundedness of states in stabilizable setting. This reflects its' effect on $R_2$ (50) and $R_3$ (51) regret terms and the analysis of $R_1$ in Abbasi-Yadkori and Szepesvári [2011] directly applies.

For additional exploration in LQRs, the update rule is still doubling of the determinant of regularized design matrix, thus the additional regret of exploration and the benefit of stabilization in regret are considered.

## G.1 Regret of ExpOpt without additional exploration in stabilizable LQRs

### G.1.1 Bounding $R_1$

The following lemma from Abbasi-Yadkori and Szepesvári [2011] bounds $R_1$ with high probability.

**Lemma G.1** (R1 in Abbasi-Yadkori and Szepesvári [2011])**.** *Let $R_1$ be as defined by (49). With probability at least $1-\delta/2$, under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$,*

$$R_1 \leq 2D\mathcal{W}^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{\mathcal{B}_{\delta,R_1}}$$

*where $\mathcal{W} = \sigma_w\sqrt{2n\log(8nT/\delta)}$ and*

$$\mathcal{B}_{\delta,R_1} = (1 + TD^2S^2(n+d)^{2(n+d)}(1+\kappa^2))\log\left(\frac{4n}{\delta}\sqrt{1 + TD^2S^2(n+d)^{2(n+d)}(1+\kappa^2)}\right)$$

### G.1.2 Bounding $|R_2|$

This term can be bounded by showing that ExpOpt rarely changes policy and besides policy change instances all the terms are 0.

**Lemma G.2** (Number of Policy Changes in Stabilizable LQR without additional exploration)**.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, in a stabilizable* LQR*,* ExpOpt *without additional exploration changes the policy at most*

$$\min \left\{ T/H_0, (n+d)\log_2\left(1 + \frac{T(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda}\right)\right\} \tag{56}$$

*Proof.* Changing policy $K$ times up to time $T$ requires $\det(V_T) \geq \lambda^{n+d}2^K$. We also have that

$$\lambda_{\max}(V_T) \leq \lambda + \sum_{t=0}^{T-1} \|z_t\|^2 \leq \lambda + T(n+d)^{2(n+d)}(1+\kappa^2)$$

Thus, $\lambda^{n+d}2^K \leq (\lambda + T(n+d)^{2(n+d)}(1+\kappa^2))^{n+d}$. Solving for K gives

$$K \leq (n+d)\log_2\left(1 + \frac{T(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda}\right).$$

Moreover, the number of policy changes is also controlled by the lower bound $H_0$ on the duration of each controller. This policy update method would give at most $T/H_0$ policy changes. Since for the policy update, ExpOpt requires both conditions to be met, the upper bound on the number of policy changes is minimum of these. □

**Lemma G.3** (Bounding $R_2$)**.** *Let $R_2$ be as defined by (50). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, we have*

$$|R_2| \leq 2D(n+d)^{2(n+d)+1}\log\left(1 + \frac{T(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda}\right)$$

*Proof.* Each non-zero term in the summation of $R_2$ is bounded by $2D(n+d)^{2(n+d)}$. Using Lemma G.2 as the upper bound on the number of changes we get the result. □

### G.1.3 Bounding $|R_3|$

The proofs in this section are adapted from Abbasi-Yadkori and Szepesvári [2011]. First consider the following lemma.

**Lemma G.4.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, in a stabilizable* LQR*, the following holds,*

$$\sum_{t=0}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq 8\max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \frac{\beta_T^2(\delta)(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\log\left(\frac{\det(V_T)}{\det(\lambda I)}\right)$$

*Proof.* Let $s_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t$ and $\tau \leq t$ be the time step that the last policy change happened. We have the following using triangle inequality,

$$\|s_t\| \leq \|(\Theta_* - \hat{\Theta}_t)^\top z_t\| + \|(\hat{\Theta}_t - \tilde{\Theta}_t)^\top z_t\|.$$

For all $\Theta \in \mathcal{C}_\tau(\delta)$, we have

$$\|(\Theta - \hat{\Theta}_t)^\top z_t\| \leq \|V_t^{1/2}(\Theta - \hat{\Theta}_t)\| \|z_t\|_{V_t^{-1}} \tag{57}$$

$$\leq \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\| \sqrt{\frac{\det(V_t)}{\det(V_\tau)}} \|z_t\|_{V_t^{-1}} \tag{58}$$

$$\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}}\right\} \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\| \|z_t\|_{V_t^{-1}} \tag{59}$$

$$\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}}\right\} \beta_\tau(\delta) \|z_t\|_{V_t^{-1}} \tag{60}$$

where (57) follows from Cauchy-Schwarz, (58) follows from Lemma 11 of Abbasi-Yadkori and Szepesvári [2011]. For (59) consider the following,

$$\det(V_t)/\det(V_\tau) = \prod_{i=\tau+1}^{t} (1 + \|z_i\|^2_{V_{i-1}^{-1}}).$$

ExpOpt has policy update when the determinant of the regularized design matrix is doubled and $H_0$ time steps has passed since the last update. Therefore this ratio is either 2 or it is upper bounded by $\prod_{i=\tau+1}^{\tau+H_0}(1 + \|z_i\|^2_{V_{i-1}^{-1}}) \leq \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}$, where we use the fact that $\|z_i\|^2_{V_{i-1}^{-1}} \leq (1+\kappa^2)(n+d)^{2(n+d)}/\lambda$, which gives (59). Finally at (60), we use the fact that $\lambda_{\max}(M) \leq \mathrm{Tr}(M)$ for $M \succeq 0$. Using this result, we obtain,

$$\sum_{t=0}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2$$
$$\leq 8 \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \frac{\beta_T^2(\delta)(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda} \log\left(\frac{\det(V_T)}{\det(\lambda I)}\right)$$

where we use Lemma I.1. □

**Lemma G.5** (Bounding $R_3$). *Let $R_3$ be as defined by (51). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, we have*

$$|R_3| \leq (1+\kappa^2)(n+d)^{2(n+d)} SD\beta_T(\delta) \sqrt{\frac{32T}{\lambda} \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \log\left(\frac{\det(V_T)}{\det(\lambda I)}\right)}$$

*Proof.*

$$
\begin{aligned}
|R_3| &\leq \sum_{t=0}^{T} \left| \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\|^2 \right| \\
&\leq \left( \sum_{t=0}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \left( \sum_{t=0}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&\leq \left( \sum_{t=0}^{T} \left\| P(\tilde{\Theta}_t)^{1/2} \left( \tilde{\Theta}_t - \Theta_* \right)^\top z_t \right\|^2 \right)^{1/2} \left( \sum_{t=0}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&\leq \sqrt{8D \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \frac{\beta_T^2(\delta)(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda} \log\left(\frac{\det(V_T)}{\det(\lambda I)}\right)} \\
&\quad \times \sqrt{4TD(1+\kappa^2)(n+d)^{2(n+d)} S^2} \\
&\quad \leq (1+\kappa^2)(n+d)^{2(n+d)} SD\beta_T(\delta) \sqrt{\frac{32T}{\lambda} \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \log\left(\frac{\det(V_T)}{\det(\lambda I)}\right)} \\
&\quad \leq (1+\kappa^2)(n+d)^{2(n+d)} SD\beta_T(\delta) \times \\
&\quad \sqrt{\frac{32T(n+d)}{\lambda} \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \log\left(1 + \frac{T(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda(n+d)}\right)}
\end{aligned}
$$

□

### G.1.4 Combining Terms for Final Regret Upper Bound

**Proof of Theorem 1:** Combining Lemmas G.1, G.3 and G.5 we obtain the following,

$$
\begin{aligned}
&\textsc{Regret}(T) \leq R_1 - R_2 - R_3 + 2\sqrt{T} \\
&\leq 2D\mathcal{W}^2 \sqrt{2T \log\frac{8}{\delta}} + n\sqrt{\mathcal{B}_{\delta,R_1}} + 2D(n+d)^{2(n+d)+1} \log\left(1 + \frac{T(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda}\right) + 2\sqrt{T} \\
&+ (1+\kappa^2)(n+d)^{2(n+d)} SD\beta_T(\delta) \times \\
&\sqrt{\frac{32T(n+d)}{\lambda} \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \log\left(1 + \frac{T(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda(n+d)}\right)}
\end{aligned}
$$

for $\mathcal{W} = \sigma_w \sqrt{2n \log(8nT/\delta)}$ and

$$
\mathcal{B}_{\delta,R_1} = (1 + TD^2S^2(n+d)^{2(n+d)}(1+\kappa^2)) \log\left(\frac{4n}{\delta}\sqrt{1 + TD^2S^2(n+d)^{2(n+d)}(1+\kappa^2)}\right).
$$

This gives the advertised regret upper bound of ExpOpt for the stabilizable system without using additional exploration.

□

## G.2 Direct Effect of Additional Exploration, Bounding $\sum_{t=0}^{T_w} \left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right)$ in the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$

The following result also holds for stabilizable setting since the state is upper bounded similar to the controllable LQRs, *i.e.*, $\|x_t\| \leq c'(n+d)^{n+d}$ for $t \leq T_w$. Thus, we present the result for generic $T_w$ and specialization to the settings can be obtained by picking $T_c$ or $T_s$.

**Lemma G.6** (Direct Effect of Enforced Exploration on Regret)**.** *If $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$ holds then with probability at least $1-\delta$,*

$$\sum_{t=0}^{T_w} \left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right) \leq d\sigma_\nu \sqrt{B_\delta} + d\|R\|\sigma_\nu^2 \left(T_w + \sqrt{T_w} \log \frac{4dT_w}{\delta} \sqrt{\log \frac{4}{\delta}}\right) \tag{61}$$

*where*

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

*Proof.* Let $q_t^\top = u_t^\top R$. 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 5 on some event $G_{\delta,i}$ that holds with probability at least $1-\delta/(2d)$, for any $t \geq 0$,

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

On $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, $\|q_k\| \leq \kappa\|R\|(n+d)^{n+d}$, thus $q_{k,i} \leq \kappa\|R\|(n+d)^{n+d}$. Using union bound we get, for probability at least $1-\frac{\delta}{2}$,

$$\sum_{t=0}^{T_w} 2\nu_t^\top R u_t \leq$$
$$d\sqrt{8\sigma_\nu^2\left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right)\log\left(\frac{4d}{\delta}\left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right)^{1/2}\right)} \tag{62}$$

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

$$\Pr\left(\sum_{t=1}^{T_w} \Psi_t > 2\|R\|W^2\sqrt{2T_w \log\frac{4}{\delta}}\right) \leq$$

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

Using Lemma I.2 with union bound and Theorem 6, summation of terms on the right hand side is bounded by $\delta/2$. Thus, with probability at least $1-\delta/2$,

$$\sum_{t=0}^{T_w} \nu_t^\top R\nu_t \leq dT_w\sigma_\nu^2\|R\| + 2\|R\|W^2\sqrt{2T_w \log\frac{4}{\delta}}. \tag{63}$$

Combining (62) and (63) gives the statement of lemma for the regret of external exploration noise.

□

## G.3 Regret of ExpOpt with additional exploration in LQRs

### G.3.1 Bounding $R_1^\zeta$ in the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$

In this section, we state the bounds on $R_1^\zeta$ for both controllable and stabilizable systems. We first provide high probability bound on the system noise.

**Lemma G.7** (Bounding sub-Gaussian vector). *With probability $1-\frac{\delta}{8}$, $\|\zeta_k\| \leq (\sigma_w+\|B_*\|\sigma_\nu)\sqrt{2n\log\frac{8nT}{\delta}}$ for $k \leq T_w$ and $\|\zeta_k\| \leq \sigma_w\sqrt{2n\log\frac{8nT}{\delta}}$ for $T_w < k \leq T$.*

*Proof.* From the subgaussianity assumption, we have that for any index $1 \leq i \leq n$ and any time $k$, $|w_{k,i}| \leq \sigma_w\sqrt{2\log\frac{8}{\delta}}$ and $|(B_*\nu_k)_i| < \|B_*\|\sigma_\nu\sqrt{2\log\frac{8}{\delta}}$ with probability $1-\frac{\delta}{8}$. Using the union bound, we get the statement of lemma. □

Using this we state the bound on $R_1^\zeta$ for controllable systems.

**Lemma G.8** (Bounding $R_1^\zeta$ for controllable LQR). *Let $R_1^\zeta$ be as defined by (53). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, with probability at least $1-\delta/2$, for controllable* LQR*, for $t > T_{r,c}$, we have*

$$\begin{aligned} R_1 \leq\ & k_{c,1}(n+d)^{n+d}(\sigma_w + \|B_*\|\sigma_\nu)n\sqrt{T_{r,c}}\log((n+d)T_{r,c}/\delta) \\ &+ k_{c,2}\sigma_w^2\frac{n\sqrt{n}}{(1-\Upsilon)}\sqrt{t-T_{r,c}}\log(n(t-T_c)/\delta) \\ &+ k_{c,3}n\sigma_w^2\sqrt{T-T_w}\log(nT/\delta) + k_{c,4}n(\sigma_w+\|B_*\|\sigma_\nu)^2\sqrt{T_w}\log(nT/\delta), \end{aligned}$$

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

*Proof.* Assume that the event $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$ holds. Let $f_t = A_* x_t + B_* u_t$. One can decompose $R_1$ as

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

Since $P(\tilde{\Theta}_0)$ is positive semidefinite and $x_0 = 0$, the first two terms are bounded above by zero. The second term is decomposed as follows

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

Let $R_{1,1} = \sum_{t=1}^{T} f_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1}$ and $R_{1,2} = \sum_{t=1}^{T} \left(\zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} - \mathbb{E}\left[\zeta_{t-1}^\top \tilde{P}_t \zeta_{t-1} \middle| \mathcal{F}_{t-2}\right]\right)$. Let $v_{t-1}^\top = f_{t-1}^\top P(\tilde{\Theta}_t)$. $R_{1,1}$ 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 5 on some event $G_{\delta,i}$ that holds with probability at least $1 - \delta/(4n)$, for any $t \geq 0$,

$$M_{t,i}^2 \leq 2(\sigma_w^2 + \|B_*\|^2\sigma_\nu^2)\left(1 + \sum_{k=1}^{T_{r,c}} v_{k-1,i}^2\right)\log\left(\frac{4n}{\delta}\left(1 + \sum_{k=1}^{T_{r,c}} v_{k-1,i}^2\right)^{1/2}\right)$$
$$+ 2\sigma_w^2\left(1 + \sum_{k=T_{r,c}+1}^{t} v_{k-1,i}^2\right)\log\left(\frac{4n}{\delta}\left(1 + \sum_{k=T_{r,c}+1}^{t} v_{k-1,i}^2\right)^{1/2}\right) \quad \text{for } t > T_{r,c}.$$

Notice that ExpOpt stops additional exploration after $t = T_w$, and the state starts decaying until $t = T_{r,c}$. For simplicity of presentation we treat the time between $T_w$ and $T_{r,c}$ as exploration sacrificing the tightness of the result. On $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, $\|\nu_k\| \leq DS(n+d)^{n+d}\sqrt{1+\kappa^2}$ for $k \leq T_{r,c}$ and $\|\nu_k\| \leq \frac{4DS\sigma_w\sqrt{1+\kappa^2}}{1-\Upsilon}\sqrt{2n\log\frac{n(T-T_c)}{\delta}}$ for $k > T_{r,c}$. Thus, $v_{k,i} \leq DS(n+d)^{n+d}\sqrt{1+\kappa^2}$ and $v_{k,i} \leq \frac{4DS\sigma_w\sqrt{1+\kappa^2}}{1-\Upsilon}\sqrt{2n\log\frac{n(T-T_c)}{\delta}}$ respectively for $k \leq T_{r,c}$ and $k > T_{r,c}$ . Using union bound we get, for probability at least $1 - \frac{\delta}{4}$, for $t > T_{r,c}$,

$$R_{1,1} \leq n\sqrt{2(\sigma_w^2 + \|B_*\|^2\sigma_\nu^2)\left(1 + T_{r,c}D^2S^2(n+d)^{2(n+d)}(1+\kappa^2)\right)} \times$$
$$\sqrt{\log\left(\frac{4n}{\delta}\left(1 + T_{r,c}D^2S^2(n+d)^{2(n+d)}(1+\kappa^2)\right)^{1/2}\right)}$$
$$+ n\sqrt{2\sigma_w^2\left(1 + 32(t - T_{r,c})D^2S^2\frac{n\sigma_w^2}{(1-\Upsilon)^2}(1+\kappa^2)\log\left(\frac{n(T-T_c)}{\delta}\right)\right)} \times$$
$$\sqrt{\log\left(\frac{4n}{\delta}\left(1 + 32(t - T_{r,c})D^2S^2\frac{n\sigma_w^2}{(1-\Upsilon)^2}(1+\kappa^2)\log\left(\frac{n(T-T_c)}{\delta}\right)\right)\right)}.$$

Let $\mathcal{W}_{exp} = (\sigma_w + \|B_*\|\sigma_\nu)\sqrt{2n\log\frac{8nT}{\delta}}$ and $\mathcal{W}_{noexp} = \sigma_w\sqrt{2n\log\frac{8nT}{\delta}}$. 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 $\tilde{\Psi}_t = \Psi_t \mathbb{I}_{\left\{\Psi_t \leq 2DW_{exp}^2\right\}}$ for $t \leq T_w$ and $\tilde{\Psi}_t = \Psi_t \mathbb{I}_{\left\{\Psi_t \leq 2DW_{noexp}^2\right\}}$ for $t > T_w$ . Notice that $R_{1,2} = \sum_{t=1}^{T} \Psi_t$.

$$
\begin{aligned}
&\Pr\left(\sum_{t=1}^{T_w} \Psi_t > 2DW_{exp}^2\sqrt{2T_w \log\frac{8}{\delta}}\right) + \Pr\left(\sum_{t=T_w+1}^{T} \Psi_t > 2DW_{noexp}^2\sqrt{2(T-T_w)\log\frac{8}{\delta}}\right) \\
&\leq \Pr\left(\max_{1\leq t\leq T_w} \Psi_t > 2DW_{exp}^2\right) + \Pr\left(\max_{T_w+1\leq t\leq T} \Psi_t > 2DW_{noexp}^2\right) \\
&+ \Pr\left(\sum_{t=1}^{T_w} \tilde{\Psi}_t > 2DW_{exp}^2\sqrt{2T_w \log\frac{8}{\delta}}\right) + \Pr\left(\sum_{t=T_w+1}^{T} \tilde{\Psi}_t > 2DW_{noexp}^2\sqrt{2(T-T_w)\log\frac{8}{\delta}}\right)
\end{aligned}
$$

By Lemma I.2 with union bound and Theorem 6, summation of terms on the right hand side is bounded by $\delta/4$. Thus, with probability at least $1-\delta/4$, for $t > T_w$,

$$
R_{1,2} \leq 4nD\sigma_w^2\sqrt{2(t-T_w)\log\frac{8}{\delta}}\log\frac{8nT}{\delta} + 4nD(\sigma_w + \|B_*\|\sigma_\nu)^2\sqrt{2T_w\log\frac{8}{\delta}}\log\frac{8nT}{\delta}.
$$

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

Recall from Lemma 4.2, the bound on the state in stabilizable system is similar to its controllable counterpart in the additional exploration period. Similarly after stabilization the state is bounded as $\|x_t\| \leq \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)}$. Therefore, the same result for $R_1^\zeta$ directly translates to stabilizable setting with change of bounds on the states. We have the following bound in stabilizable systems with additional exploration.

**Lemma G.9** (Bounding $R_1^\zeta$ for stabilizable LQR)**.** *Let $R_1^\zeta$ be as defined by (53). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, with probability at least $1-\delta/2$, for stabilizable* LQR*, for $t > T_{r,s}$, we have*

$$
\begin{aligned}
R_1 &\leq k_{s,1}(n+d)^{n+d}(\sigma_w + \|B_*\|\sigma_\nu)n\sqrt{T_{r,s}}\log((n+d)T_{s,c}/\delta) \\
&+ \frac{k_{s,2}(12\kappa^2+2\kappa\sqrt{2})}{\gamma}\sigma_w^2 n\sqrt{n}\log(n(t-T_s)/\delta) \\
&+ k_{s,3}n\sigma_w^2\sqrt{T-T_w}\log(nT/\delta) + k_{s,4}n(\sigma_w+\|B_*\|\sigma_\nu)^2\sqrt{T_w}\log(nT/\delta),
\end{aligned}
$$

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

### G.3.2 Bounding $|R_2^\zeta|$ on the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$

In this section, we will bound $|R_2^\zeta|$ for both controllable and stabilizable systems. This term is similar to $R_2$ analyzed in Appendix G.1.2. In a controllable LQR policy change is governed only by the determinant of regularized design matrix $V_t$.

**Lemma G.10** (Number of Policy Changes in Controllable LQR with additional exploration)**.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, in a controllable* LQR, ExpOpt *with additional exploration changes the policy at most*

$$(n+d)\log_2\left(1+\frac{T_{r,c}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,c})\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2}\log\frac{n(T-T_c)}{\delta}}{\lambda}\right) \quad (64)$$

*Proof.* Changing policy $K$ times up to time $T$ requires $\det(V_T) \geq \lambda^{n+d}2^K$. We also have that

$$\lambda_{\max}(V_T) \leq \lambda + \sum_{t=0}^{T-1}\|z_t\|^2 \leq \lambda + T_{r,c}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,c})\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2}\log\frac{n(T-T_c)}{\delta}$$

Thus, $\lambda^{n+d}2^K \leq \left(\lambda + T_{r,c}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,c})\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2}\log\frac{n(T-T_c)}{\delta}\right)^{n+d}$. Solving for K gives

$$K \leq (n+d)\log_2\left(1+\frac{T_{r,c}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,c})\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2}\log\frac{n(T-T_c)}{\delta}}{\lambda}\right).$$

□

Similarly we have the following for stabilizable LQR.

**Lemma G.11** (Number of Policy Changes in Stabilizable LQR with additional exploration)**.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, in a stabilizable* LQR, ExpOpt *with additional exploration changes the policy at most*

$$\min\left\{T/H_0, (n+d)\log_2\left(1+\frac{\lambda+T_{r,s}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,s})X_s^2}{\lambda}\right)\right\} \quad (65)$$

*where* $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)}$

*Proof.* Changing policy $K$ times up to time $T$ requires $\det(V_T) \geq \lambda^{n+d}2^K$. We also have that

$$\lambda_{\max}(V_T) \leq \lambda + \sum_{t=0}^{T-1}\|z_t\|^2 \leq \lambda + T_{r,s}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,s})X_s^2$$

Thus, $\lambda^{n+d}2^K \leq \left(\lambda + T_{r,s}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,s})X_s^2\right)^{n+d}$. Solving for K gives

$$K \leq (n+d)\log_2\left(1+\frac{T_{r,s}(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_{r,s})X_s^2}{\lambda}\right).$$

Moreover, the number of policy changes is also controlled by the lower bound $H_0$ on the duration of each controller. This policy update method would give at most $T/H_0$ policy changes. Since for the policy update, ExpOpt requires both conditions to be met, the upper bound on the number of policy changes is minimum of these. □

Notice that besides the policy change instances, all the terms in $R_2^\zeta$ are 0. Therefore, we have the following results for controllable and stabilizable systems respectively.

**Lemma G.12** (Bounding $R_2^\zeta$ for controllable LQR)**.** *Let $R_2^\zeta$ be as defined by* (54)*. Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, for controllable* LQR*, we have*

$$|R_2| \leq 2D(n+d)^{2(n+d)+1} \log_2 \left(1 + \frac{T_{r,c}(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda}\right)$$
$$+ 2D\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2} \log \frac{n(T-T_c)}{\delta}(n+d)\times$$
$$\log_2 \left(1 + \frac{T_{r,c}(n+d)^{2(n+d)}(1+\kappa^2) + (T-T_{r,c})\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2} \log \frac{n(T-T_c)}{\delta}}{\lambda}\right)$$

*Proof.* On the event $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, we know the maximum number of policy changes up to $T_{r,c}$ and $T$ using Lemma G.10. Using the fact that $\|x_t\| \leq (n+d)^{n+d}$ for $t \leq T_{r,c}$ and $\|x_t\| \leq \frac{4\sigma_w}{1-\Upsilon}\sqrt{2n \log \frac{n(T-T_c)}{\delta}}$, we obtain the statement of the lemma. □

**Lemma G.13** (Bounding $R_2^\zeta$ for stabilizable LQR)**.** *Let $R_2^\zeta$ be as defined by* (54)*. Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, for stabilizable* LQR*, we have*

$$|R_2| \leq 2D(n+d)^{2(n+d)+1} \log_2 \left(1 + \frac{T_{r,s}(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda}\right)$$
$$+ 2DX_s(n+d) \log_2 \left(1 + \frac{\lambda + T_{r,s}(n+d)^{2(n+d)}(1+\kappa^2) + (T-T_{r,s})X_s^2}{\lambda}\right)$$

*where* $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n \log(n(t-T_s)/\delta)}$

The proof follows the same with the controllable counterpart.

### G.3.3 Bounding $|R_3^\zeta|$ on the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$

For $R_3^\zeta$, we will first consider the controllable LQR. The following adapts the proof of Abbasi-Yadkori and Szepesvári [2011] to our setting.

**Lemma G.14.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, in a controllable* LQR*, the following holds,*

$$\sum_{t=0}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \frac{16(1+\kappa^2)\beta_T^2(\delta)}{\lambda}\left((n+d)^{2(n+d)} \log \frac{\det(V_{T_{r,c}})}{\det(\lambda I)} + X_c^2 \log \frac{\det(V_T)}{\det(V_{T_{r,c}})}\right)$$

*where* $X_c^2 = \frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2} \log \frac{n(T-T_c)}{\delta}$

*Proof.* Let $s_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t$ and $\tau \leq t$ be the time step that the last policy change happened. We have the following using triangle inequality,

$$\|s_t\| \leq \|(\Theta_* - \hat{\Theta}_t)^\top z_t\| + \|(\hat{\Theta}_t - \tilde{\Theta}_t)^\top z_t\|.$$

For all $\Theta \in \mathcal{C}_\tau(\delta)$, we have

$$\|(\Theta - \hat{\Theta}_t)^\top z_t\| \leq \|V_t^{1/2}(\Theta - \hat{\Theta}_t)\|z_t\|\|_{V_t^{-1}} \tag{66}$$

$$\leq \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\| \sqrt{\frac{\det(V_t)}{\det(V_\tau)}} \|z_t\|_{V_t^{-1}} \tag{67}$$

$$\leq \sqrt{2}\|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\|\|z_t\|_{V_t^{-1}} \tag{68}$$

$$\leq \sqrt{2}\beta_\tau(\delta)\|z_t\|_{V_t^{-1}} \tag{69}$$

where (66) follows from Cauchy-Schwarz, (67) follows from Lemma 11 of Abbasi-Yadkori and Szepesvári [2011]. We use the update rule for (68) and finally at (69), we use the fact that $\lambda_{\max}(M) \leq \mathrm{Tr}(M)$ for $M \succeq 0$. Using this result, we obtain,

$$\begin{aligned}\sum_{t=0}^{T_{r,c}} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 &\leq \frac{8(1+\kappa^2)\beta_{T_{r,c}}^2(\delta)}{\lambda}(n+d)^{2(n+d)} \sum_{t=0}^{T_{r,c}} \min\{\|z_t\|^2_{V_t^{-1}}, 1\} \\ &\leq \frac{16(1+\kappa^2)\beta_T^2(\delta)}{\lambda}(n+d)^{2(n+d)} \log\frac{\det(V_{T_{r,c}})}{\det(\lambda I)}\end{aligned}$$

$$\begin{aligned}\sum_{t=T_{r,c}}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 &\leq \frac{8(1+\kappa^2)\beta_T^2(\delta)}{\lambda}\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2}\log\frac{n(T-T_c)}{\delta} \sum_{t=T_{r,c}+1}^{T} \min\{\|z_t\|^2_{V_t^{-1}}, 1\} \\ &\leq \frac{16(1+\kappa^2)\beta_T^2(\delta)}{\lambda}\frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2}\log\frac{n(T-T_c)}{\delta}\log\frac{\det(V_T)}{\det(V_{T_{r,c}})}.\end{aligned}$$

where we use Lemma I.1 in the last lines. □

The stabilizable counterpart follows similarly.

**Lemma G.15.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, in a stabilizable* LQR*, the following holds,*

$$\begin{aligned}\sum_{t=0}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq &\frac{8(1+\kappa^2)\beta_T^2(\delta)}{\lambda} \times \\ &\left((n+d)^{2(n+d)} \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \log\frac{\det(V_{T_{r,s}})}{\det(\lambda I)} \right.\\ &\left. + X_s^2 \max\left\{2, \left(1 + \frac{(1+\kappa^2)X_s^2}{\lambda}\right)^{H_0}\right\} \log\frac{\det(V_T)}{\det(V_{T_{r,s}})}\right)\end{aligned}$$

*where* $X_s = \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)}$.

*Proof.* Let $s_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t$ and $\tau \leq t$ be the time step that the last policy change happened. We have the following using triangle inequality,

$$\|s_t\| \leq \|(\Theta_* - \hat{\Theta}_t)^\top z_t\| + \|(\hat{\Theta}_t - \tilde{\Theta}_t)^\top z_t\|.$$

For all $\Theta \in \mathcal{C}_\tau(\delta)$, for $\tau \leq T_{r,s}$, we have

$$\begin{aligned}
\|(\Theta - \hat{\Theta}_t)^\top z_t\| &\leq \|V_t^{1/2}(\Theta - \hat{\Theta}_t)\| \|z_t\|_{V_t^{-1}} &(70)\\
&\leq \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\| \sqrt{\frac{\det(V_t)}{\det(V_\tau)}} \|z_t\|_{V_t^{-1}} &(71)\\
&\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}}\right\} \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\| \|z_t\|_{V_t^{-1}} &(72)\\
&\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}}\right\} \beta_\tau(\delta) \|z_t\|_{V_t^{-1}}. &(73)
\end{aligned}$$

Similarly, for for all $\Theta \in \mathcal{C}_\tau(\delta)$, for $\tau > T_{r,s}$, we have

$$\|(\Theta - \hat{\Theta}_t)^\top z_t\| \leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)X_s^2}{\lambda}\right)^{H_0}}\right\} \beta_\tau(\delta) \|z_t\|_{V_t^{-1}}$$

Using these results, we obtain,

$$\begin{aligned}
&\sum_{t=0}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \\
&\quad \leq 8 \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \frac{\beta_T^2(\delta)(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda} \log\left(\frac{\det(V_{T_{r,s}})}{\det(\lambda I)}\right) \\
&\quad + 8 \max\left\{2, \left(1 + \frac{(1+\kappa^2)X_s^2}{\lambda}\right)^{H_0}\right\} \frac{\beta_T^2(\delta)(1+\kappa^2)X_s^2}{\lambda} \log\left(\frac{\det(V_T)}{\det(V_{T_{r,s}})}\right)
\end{aligned}$$

where we use Lemma I.1. □

**Remark 1.** *Here, we provide another lemma which bounds the quantity in Lemma G.15 more tightly using the system properties and assumptions. Note that Lemma G.15 is more general. After stabilization, updating the policy in doubling epochs, instead of doubling of determinant of regularized design matrix, would remove the dependency on $H_0$ in Lemma G.15 for $t \geq T_{r,s}$, by using the fact that $\|\Theta_* - \tilde{\Theta}_t\| = \tilde{\mathcal{O}}(1/\sqrt{t})$.*

**Lemma G.16.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, in a stabilizable* LQR, *let the* ExpOpt *update its policy in doubling epochs after stabilizing the system where the base epoch length is $T_{r,s}$, i.e., after time-step*

*$T_{r,s}$, first controller is applied for $T_{r,s}$ time steps, second controller is applied for $2T_{r,s}$ time steps, so on. Then the following holds for this new update rule,*

$$\sum_{t=0}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \kappa_e^2(1+\kappa^2)X_s^2 \log(T)$$

*for $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_s)/\delta)}$.*

*Proof.* Let $s_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t$ and $\tau \leq t$ be the time step that the last policy change happened. We have the following using triangle inequality,

$$\|s_t\| \leq \|(\Theta_* - \hat{\Theta}_t)^\top z_t\| + \|(\hat{\Theta}_t - \tilde{\Theta}_t)^\top z_t\|.$$

For all $\Theta \in \mathcal{C}_\tau(\delta)$, for $\tau > T_{r,s}$, we have

$$\begin{aligned}\|(\Theta - \hat{\Theta}_t)^\top z_t\| &\leq \|(\Theta - \hat{\Theta}_t)\|\|z_t\| \\ &\leq \frac{\kappa_e}{\sqrt{\tau}}\sqrt{1+\kappa^2}X_s\end{aligned}$$

Let $\tau_t$ denote the time step that last policy change occured before time $t$. Using the new update rule we obtain,

$$\begin{aligned}\sum_{t=T_{r,s}}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 &\leq \sum_{t=T_{r,s}}^{T} 4\frac{\kappa_e^2(1+\kappa^2)X_s^2}{\tau_t} \\ &\leq T_{r,s}4\frac{\kappa_e^2(1+\kappa^2)X_s^2}{T_{r,s}} + 2T_{r,s}4\frac{\kappa_e^2(1+\kappa^2)X_s^2}{2T_{r,s}} + ... \\ &\leq \kappa_e^2(1+\kappa^2)X_s^2\log(T),\end{aligned}$$

where the last line follows from the fact that there can be at most $\log(T)$ updates in this update scheme. □

Now, we bound $R_3^\zeta$ in controllable systems.

**Lemma G.17** (Bounding $R_3^\zeta$ for controllable LQR)**.** *Let $R_3^\zeta$ be as defined by* (55)*. Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$, for controllable* LQR*, we have*

$$|R_3^\zeta| = \tilde{\mathcal{O}}\left((n+d)^{2(n+d)}\sqrt{T_{r,c}} + (n+d)n^2\sqrt{T - T_{r,c}}\right)$$

*Proof.* The proof follows similar decomposition with Lemma G.5, *i.e.*, after using triangle inequlity and we use Cauchy Schwarz inequality and again triangle inequality, and gives the following result:

$$
\begin{aligned}
\left|R_3^\zeta\right| &\leq \sum_{t=0}^{T} \left| \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\|^2 \right| \\
&= \sum_{t=0}^{T_{r,c}} \left| \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\|^2 \right| + \sum_{t=T_{r,c}}^{T} \left| \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\|^2 \right| \\
&\leq \left( \sum_{t=0}^{T_{r,c}} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \left( \sum_{t=0}^{T_{r,c}} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&+ \left( \sum_{t=T_{r,c}}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \left( \sum_{t=T_{r,c}}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&\leq \left( \sum_{t=0}^{T_{r,c}} \left\| P(\tilde{\Theta}_t)^{1/2} \left( \tilde{\Theta}_t - \Theta_* \right)^\top z_t \right\|^2 \right)^{1/2} \left( \sum_{t=0}^{T_{r,c}} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&\quad + \left( \sum_{t=T_{r,c}}^{T} \left\| P(\tilde{\Theta}_t)^{1/2} \left( \tilde{\Theta}_t - \Theta_* \right)^\top z_t \right\|^2 \right)^{1/2} \left( \sum_{t=T_{r,c}}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&\leq \sqrt{\frac{16D(1+\kappa^2)\beta_T^2(\delta)}{\lambda} (n+d)^{2(n+d)} \log \frac{\det(V_{T_{r,c}})}{\det(\lambda I)}} \sqrt{4T_{r,c} D(1+\kappa^2) S^2 (n+d)^{2(n+d)}} \\
&+ \sqrt{\frac{16D(1+\kappa^2)\beta_T^2(\delta)}{\lambda} \frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2} \log \frac{n(T-T_c)}{\delta} \log \frac{\det(V_T)}{\det(V_{T_{r,c}})}} \times \\
&\qquad \sqrt{4(T-T_{r,c}) D(1+\kappa^2) S^2 \frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2} \log \frac{n(T-T_c)}{\delta}} \\
&\leq \frac{8DS(1+\kappa^2)\beta_T(\delta)(n+d)^{2(n+d)}}{\sqrt{\lambda}} \sqrt{T_{r,c}(n+d) \log \left( 1 + \frac{T_{r,c}(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda(n+d)} \right)} \\
&+ \frac{8DS(1+\kappa^2)\beta_T(\delta)}{\sqrt{\lambda}} \left( \frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2} \log \frac{n(T-T_c)}{\delta} \right) \times \\
&\qquad \sqrt{(T-T_{r,c})(n+d) \log \left( 1 + \frac{T_{r,c}(1+\kappa^2)(n+d)^{2(n+d)} + (T-T_{r,c}) \left( \frac{32n\sigma_w^2(1+\kappa^2)}{(1-\Upsilon)^2} \log \frac{n(T-T_c)}{\delta} \right)}{\lambda(n+d)} \right)}
\end{aligned}
$$

□

The stabilizable counterpart follows similarly with the required changes, thus it's proof is omitted.

**Lemma G.18** (Bounding $R_3^\zeta$ for stabilizable LQR)**.** *Let $R_3^\zeta$ be as defined by (55). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, for stabilizable* LQR, *we have*

$$
|R_3^\zeta| = \tilde{\mathcal{O}} \left( (n+d)^{(H_0+1)(n+d)} \sqrt{T_{r,s}} + (n+d) n^{2+H_0/2} \sqrt{T - T_{r,s}} \right)
$$

If one uses Lemma G.16, then the following result is obtained directly.

**Lemma G.19** (Improved bound on $R_3^\zeta$ for stabilizable LQR using doubling epoch length)**.** *Let $R_3^\zeta$ be as defined by* (55)*. Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_s}^{[s]}$, for stabilizable* LQR*, after stabilization if* ExpOpt *updates its policy in doubling epochs with base epoch length of $T_{r,s}$, we have*

$$|R_3^\zeta| = \tilde{\mathcal{O}}\left((n+d)^{(H_0+1)(n+d)}\sqrt{T_{r,s}} + (n+d)n^2\sqrt{T - T_{r,s}}\right)$$

### G.3.4 Combining Terms for Final Regret Upper Bounds

**Proof of Theorem 2:** Recall that

$$\text{REGRET}(T) \leq \sigma_\nu^2 T_w D \|B_*\|_F^2 + \sum_{t=0}^{T_c}\left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right) + R_1^\zeta - R_2^\zeta - R_3^\zeta.$$

Combining Lemma G.6 for $\sum_{t=0}^{T_c}\left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right)$, Lemma G.8 for $R_1^\zeta$, Lemma G.12 for $|R_2^\zeta|$ and Lemma G.17 for $|R_3^\zeta|$, we get the advertised regret bound.

□

**Proof of Theorem 3:** Recall that

$$\text{REGRET}(T) \leq \sigma_\nu^2 T_w D \|B_*\|_F^2 + \sum_{t=0}^{T_s}\left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right) + R_1^\zeta - R_2^\zeta - R_3^\zeta.$$

- Combining Lemma G.6 for $\sum_{t=0}^{T_s}\left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right)$, Lemma G.9 for $R_1^\zeta$, Lemma G.13 for $|R_2^\zeta|$ and Lemma G.18 for $|R_3^\zeta|$, we get the advertised regret bound.

**Remark 2.** *Note that this bound is in general setting. If the structure and the assumptions of the system is further exploited, then using Lemma G.19 in bounding $R_3^\zeta$, we remove the dependency on $H_0$ in the polynomial in dimension regret bound after stabilization.*

□

# H Stabilizability Discussion

In Assumption 2.3, we are given a set, $\mathcal{S}_s$, that consists of $(\kappa, \gamma)$-stabilizable systems, *i.e.* for all $(A, B) \in \mathcal{S}_0$, $\exists K$ such that $\rho(A + BK) < 1 - \gamma$, $\|K\| \leq \kappa$ and $A + BK$ is $(\kappa, \gamma)$-strongly stable.

**Definition H.1.** *A matrix $M$ is $(\kappa, \gamma)$-strongly stable (for $\kappa \geq 1$ and $0 \leq \gamma \leq 1$) if there exists matrices $H \succ 0$ and $L$ such that $M = HLH^{-1}$ with $\|L\| \leq 1 - \gamma$ and $\|H\|\|H^{-1}\| \leq \kappa$.*

This is a valid assumption since for all stabilizable systems, by setting $1 - \gamma = \rho(A + BK)$ and $\kappa$ to be the condition number of $P^{1/2}$ where $P$ is the positive definite matrix that satisfies the following Lyapunov equation:

$$(A + BK)^\top P (A + BK) \preceq P, \tag{74}$$

one can show that closed-loop system is $(\kappa, \gamma)$-strongly stable Lemma B.1 of Cohen et al. [2018].

# I Technical Theorems and Lemmas

**Theorem 5** (Self-normalized bound for vector-valued martingales [Abbasi-Yadkori et al., 2011])**.** *Let* $(\mathcal{F}_t; k \geq 0)$ *be a filtration,* $(m_k; k \geq 0)$ *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*

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

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

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

**Theorem 6** (Azuma's inequality)**.** *Assume that* $(X_s; s \geq 0)$ *is a supermartingale and* $|X_s - X_{s-1}| \leq c_s$ *almost surely. Then for all* $t > 0$ *and all* $\epsilon > 0$,

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

**Lemma I.1** (Bound on Logarithm of the Determinant of Sample Covariance Matrix [Abbasi-Yadkori et al., 2011])**.** *The following holds for any* $t \geq 1$ *:*

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

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

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

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