Reinforcement Learning with Fast Stabilization in Linear Dynamical Systems

Sahin Lale
Caltech

Kamyar Azizzadenesheli
Purdue University

Babak Hassibi
Caltech

Anima Anandkumar
Caltech

Abstract

In this work, we study model-based reinforcement learning (RL) in unknown stabilizable linear dynamical systems. When learning a dynamical system, one needs to stabilize the unknown dynamics in order to avoid system blow-ups. We propose an algorithm that certifies fast stabilization of the underlying system by effectively exploring the environment with an improved exploration strategy. We show that the proposed algorithm attains $\tilde{\mathcal{O}}(\sqrt{T})$ regret after T time steps of agent-environment interaction. We also show that the regret of the proposed algorithm has only a polynomial dependence in the problem dimensions, which gives an exponential improvement over the prior methods. Our improved exploration method is simple, yet efficient, and it combines a sophisticated exploration policy in RL with an isotropic exploration strategy to achieve fast stabilization and improved regret. We empirically demonstrate that the proposed algorithm outperforms other popular methods in several adaptive control tasks.

1 INTRODUCTION

We study the problem of reinforcement learning (RL) in linear dynamical systems, in particular in linear quadratic regulators (LQR). LQR is the canonical setting for linear dynamical systems with quadratic regulatory costs and observable state evolution. For a known LQR model, the optimal control policy is given by a stabilizing linear state feedback controller (Bertsekas, 1995). When the underlying model is unknown, the learning agent needs to learn the dynamics in order to (1) stabilize the system and (2) find the optimal control policy. This online control task is one of the core challenges in RL and control theory.

Learning LQR models from scratch: The ultimate goal in online control is to design learning agents that can autonomously adapt to the unknown environment with minimal information and also enjoy finite-time stability and performance guarantees. This problem has sparked a flurry of research interest in the control and RL communities. However, there are only a few approaches that provide a complete treatment of the problem and strive for learning from scratch with no initial model estimates (Abbasi-Yadkori and Szepesvári, 2011; Abeille and Lazaric, 2018; Chen and Hazan, 2020). Other than these, the prior works focus either on the problem of finding a stabilizing policy while ignoring the control costs (Faradonbeh et al., 2018a), or on achieving low control costs while assuming access to an initial stabilizing controller (Abeille and Lazaric, 2020; Simchowitz and Foster, 2020).

Lack of stabilization and its consequences: The existing works (Abbasi-Yadkori and Szepesvári, 2011; Abeille and Lazaric, 2017, 2018) that learn from scratch in LQRs aim to minimize the regret, which is the additional cumulative control cost of an agent compared to the expected cumulative cost of the optimal policy. These algorithms suffer from regret that has an exponential dependence in the LQR dimensions since they do not assume access to an initial stabilizing policy. They also face system blow-ups due to unstable system dynamics. Besides poor regret performance, the uncontrolled dynamics prevent the deployment of these learning algorithms in practice.

Joint goals of fast stabilization and low regret: In this paper, we design an RL agent for online LQRs that achieves low regret and fast stabilization. To design stabilizing policies without prior knowledge, the agent needs to effectively explore the environment and estimate the system dynamics. However, in order to achieve low regret, the agent should also strategically exploit the gathered knowledge. Thus, the agent re-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

Table 1: Comparison with the prior works.

Work	**Regret**	**Setting**	**Stabilizing Controller**
Dean et al. (2018)	$\text{poly}(n, d)T^{2/3}$	Controllable	Required
Mania et al. (2019)	$\text{poly}(n, d)\sqrt{T}$	Controllable	Required
Simchowitz and Foster (2020)	$\text{poly}(n, d)\sqrt{T}$	Stabilizable	Required
Abbasi-Yadkori and Szepesvári (2011)	$(n + d)^{n+d}\sqrt{T}$	Controllable	Not required
Chen and Hazan (2020)	$\text{poly}(n, d)\sqrt{T}$	Controllable	Not required
This work	$\text{poly}(n, d)\sqrt{T}$	Stabilizable	Not required

quires to balance exploration and exploitation such that it designs stabilizing policies to avoid dire consequences of unstable dynamics and minimize the regret.

Optimism in the face of uncertainty (OFU) principle: One of the most prominent methods to effectively balance exploration and exploitation is the OFU principle (Lai and Robbins, 1985). An agent that follows the OFU principle deploys the optimal policy of the model with the lowest optimal cost within the set of plausible models. This guarantees the asymptotic convergence to the optimal policy for the LQR (Bittanti et al., 2006).

Failure of OFU to achieve stabilization: Using the OFU principle, the learning algorithm of (Abbasi-Yadkori and Szepesvári, 2011) attains order-optimal $\tilde{\mathcal{O}}(\sqrt{T})$ regret after T time steps, but the regret upper bound suffers from an *exponential* dependence in the LQR model dimensions. This is due to the fact that the OFU principle relies heavily on the confidence-set constructions. An agent following the OFU principle mostly explores parts of state-space with the lowest expected cost and with higher uncertainty. When the agent does not have reliable model estimates, this may cause a lack of exploration in certain parts of the state-space that are important in designing stabilizing policies. This problem becomes more evident in the early stages of agent-environment interactions due to lack of reliable knowledge about the system. This highlights the need for an improved exploration in the early stages. Note that this issue is unique to control problems and not as common in other RL settings, e.g. bandits and gameplay.

The restricted LQR settings in the prior works: In designing our learning agent for the online LQR problem, we consider the stabilizable LQR setting. Stabilizability is the necessary and sufficient condition to have a well-defined online LQR problem, *i.e.* it guarantees the existence of a policy that stabilizes the system (Kailath et al., 2000). In contrast, the prior works that learn from scratch in LQRs only guarantee low regret in the controllable or contractive LQR settings (Abbasi-Yadkori and Szepesvári, 2011; Abeille and Lazaric, 2017, 2018; Chen and Hazan, 2020), which form a narrow subclass of stabilizable LQR problems. These conditions significantly simplify the identification and regulation of the unknown dynamics. However, they are violated in many practical systems, e.g., physical systems with non-minimal representation due to complex dynamics (Friedland, 2012). In contrast, most of the real-world control systems are stabilizable.

Contributions:

Based on the above observations and shortcomings, we propose a novel **Stab**ilizing **L**earning algorithm, StabL, for the online LQR problem and study its performance both theoretically and empirically.

1) We carefully prescribe an early exploration strategy and a policy update rule in the design of StabL. We show that StabL quickly stabilizes the underlying system, and henceforth certifies the stability of the dynamics with high probability in the stabilizable LQRs.

2) We show that StabL attains $\tilde{\mathcal{O}}(\text{poly}(n, d)\sqrt{T})$ regret in the online control of unknown stabilizable LQRs. Here $\tilde{\mathcal{O}}(\cdot)$ presents the order up to logarithmic terms, n is the state and d is the input dimensions respectively. This makes StabL the first RL algorithm to achieve order-optimal regret in all stabilizable LQRs without a given initial stabilizing policy. This result completes an important part of the picture in designing autonomous learning agents for the online LQR problem (See Table 1).

3) We empirically study the performance of StabL in various adaptive control tasks. We show that StabL achieves fast stabilization and consequently enjoys orders of magnitude improvement in regret compared to the existing certainty equivalent and optimism-based learning from scratch methods. Further, we study the statistics of the control inputs and highlight the effect of strategic exploration in achieving this improved performance.

The design of StabL is motivated by the importance of stabilizing the unknown dynamics and the need for exploration in the early stages of agent-environment interactions. StabL deploys the OFU principle to balance exploration vs. exploitation trade-off. Due to lack of reliable estimates in the early stages of learning, an optimistic controller, guided by OFU, neither provides sufficient exploration required to achieve stabilizing controllers, nor achieves sub-linear regret. Therefore, StabL uses isotropic exploration along with the optimistic controller in the early stages to achieve an improved exploration strategy. This allows StabL to excite all dimensions of the system uniformly as well as the dimensions that have more promising impact on the control performance. By carefully adjusting the early improved exploration, we guarantee that the inputs of StabL are persistently exciting the system under the sub-Gaussian process noise. We show that using this improved exploration quickly results in stabilizing policies with high probability, therefore a much smaller regret in the long term.

We conduct extensive experiments to verify the theoretical claims about StabL. In particular, we empirically show that the improved exploration strategy of StabL persistently excites the system in the early stages and achieves effective system identification required for stabilization. In contrast, we observe that the optimism-based learning algorithm of Abbasi-Yadkori and Szepesvári (2011) fails to achieve effective exploration in the early stages and suffers from unstable dynamics and high regret. We also demonstrate that, once StabL obtains reliable model estimates for stabilization, the balanced strategy prescribed by the OFU principle effectively guides StabL to regret minimizing policies, resulting in a significant improved regret performance in all settings.

2 PRELIMINARIES

Notation: We denote the Euclidean norm of a vector x as $\|x\|$. For a given matrix A, $\|A\|$ denotes the spectral norm, $\|A\|_F$ denotes the Frobenius norm, $A^\top$ is the transpose, $\mathrm{Tr}(A)$ 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 maximum and minimum singular values of A are denoted as $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ respectively.

Consider a discrete time linear time-invariant system,

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

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 the process noise at time t. We consider the systems with sub-Gaussian noise.

Assumption 2.1 (Sub-Gaussian Noise). *The process noise w_t is a martingale difference sequence with respect to the filtration $(\mathcal{F}_{t-1})$. Moreover, it is component-wise conditionally σ_w^2-sub-Gaussian and isotropic such that for any $s \in \mathbb{R}$, $\mathbb{E}\left[\exp\left(s w_{t,j}\right)|\mathcal{F}_{t-1}\right] \leq \exp\left(s^2\sigma_w^2/2\right)$ and $\mathbb{E}\left[w_t w_t^\top|\mathcal{F}_{t-1}\right] = \bar{\sigma}_w^2 I$ for some $\bar{\sigma}_w^2 > 0$.*

Note that the results of this paper only require the conditional covariance matrix $W = \mathbb{E}[w_t w_t^\top|\mathcal{F}_{t-1}]$ to be full rank. The isotropic noise assumption is chosen to ease the presentation and similar results can be obtained with upper and lower bounds on W, *i.e.*, $W_{up} > \sigma_{\max}(W) \geq \sigma_{\min}(W) > W_{low} > 0$.

At each time step t, the system is at state x_t. After observing x_t, 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_*. This problem is the canonical example for the control of linear dynamical systems and termed as linear quadratic regulator (LQR). The system (1) can be represented as $x_{t+1} = \Theta_*^\top z_t + w_t$, where $\Theta_*^\top = [A_* \;\; B_*]$ and $z_t = [x_t^\top \;\; u_t^\top]^\top$. Knowing Θ_*, the optimal control policy, is a linear state feedback control $u_t = K(\Theta_*)x_t$ with $K(\Theta_*) = -(R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_*$, where P_* is the unique solution to the discrete-time algebraic Riccati equation (DARE) (Bertsekas, 1995):

$$P_* = A_*^\top P_* A_* + Q - A_*^\top P_* B_* (R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_*. \tag{2}$$

The optimal cost for Θ_* is denoted as $J_* = \mathrm{Tr}(\bar{\sigma}_w^2 P_*)$. When the model parameters, A_* and B_*, are unknown, the learning 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 the designer's choice and given. After T time steps, we evaluate the regret of the learning agent as $\mathrm{R}(T) = \sum_{t=0}^T (c_t - J_*)$, which is the difference between the performance of the agent and the expected performance of the optimal controller. In this work, unlike the controllable LQR setting of the prior adaptive control algorithms without a stabilizing controller (Abbasi-Yadkori and Szepesvári, 2011; Chen and Hazan, 2020), we study the online LQR problem in the general setting of *stabilizable* LQR.

Definition 2.1 (Stabilizability vs. Controllability). *The linear dynamical system Θ_* is stabilizable if there exists K such that $\rho(A_* + B_* K) < 1$. On the other hand, the linear dynamical system Θ_* is controllable if the controllability matrix $[B_* \;\; A_* B_* \;\; A_*^2 B_* \;\; \ldots \;\; A_*^{n-1} B_*]$ has full row rank.*

Note that the stabilizability condition is the minimum requirement to define the optimal control problem. It is *strictly weaker than controllability*, *i.e.*, all controllable systems are stabilizable but the converse is not true (Bertsekas, 1995). Similar to Cohen et al. (2019), we quantify the stabilizability of Θ_* for the finite-time analysis.

Definition 2.2 ((κ,γ)-Stabilizability)**.** *The linear dynamical system Θ_* is (κ,γ)-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 merely a quantification of stabilizability. In other words, any stabilizable system is also (κ,γ)-stabilizable for some κ and γ and conversely (κ,γ)-stabilizability implies stabilizability (See Appendix A). Thus, we consider (κ,γ)-stabilizable LQRs.

Assumption 2.2 (Stabilizable Linear Dynamical System)**.** *The unknown parameter Θ_* is a member of the set $\mathcal{S}$ such that $\mathcal{S} = \big\{\Theta' = [A', B'] \;\big|\; \Theta'$ is (κ,γ)-stabilizable, $\|\Theta'\|_F \leq S\big\}$*

Notice that $\mathcal{S}$ denotes the set of all bounded systems that are (κ,γ)-stabilizable, where Θ_* is an element of, and the membership to $\mathcal{S}$ can be easily verified. Moreover, the proposed algorithm in this work only requires the upper bounds on these relevant control-theoretic quantities κ, γ, and S, which are also standard in prior works, e.g. (Abbasi-Yadkori and Szepesvári, 2011; Cohen et al., 2019). In practice, when there is a total lack of knowledge about the system, one can start with conservative upper bounds and adjust these based on the behavior of the system, *e.g.*, the growth of the state.

From (κ,γ)-stabilizability, we have that $\rho(A' + B'K(\Theta')) \leq 1-\gamma$, and $\sup\{\|K(\Theta')\| \mid \Theta' \in \mathcal{S}\} \leq \kappa$. The following lemma shows that for any (κ,γ)-stabilizable system the solution of (2) is bounded.

Lemma 2.1 (Bounded DARE Solution)**.** *For any Θ that is (κ,γ)-stabilizable and has bounded regulatory cost matrices, i.e., $\|Q\|, \|R\| < \overline{\alpha}$, the solution of (2), P, is bounded as $\|P\| \leq D \coloneqq \overline{\alpha}\gamma^{-1}\kappa^2(1+\kappa^2)$*

3 STABL

In this section, we present StabL, a sample efficient stabilizing RL algorithm for the online stabilizable LQR problem. The algorithmic outline is provided in Algorithm 1. StabL only requires the minimal information about the stabilizability of the underlying system and *does not* need a stabilizing controller. Therefore, along the ultimate goal of minimizing the regret, StabL puts its primary focus on achieving stabilizing controllers for the unknown system dynamics.

3.1 Adaptive Control with Improved Exploration

In order to quickly design stabilizing controllers, StabL needs to explore the system dynamics effectively. To this end, StabL solves $\min_\Theta \sum_{s=0}^{t-1} \|x_{s+1} - \Theta^\top z_s\|^2 + \lambda\|\Theta\|_F^2$, using the past state-input pairs to estimate the system dynamics as $\hat{\Theta}_t$. Using this estimate, StabL constructs a high probability confidence set $\mathcal{C}_t(\delta)$ that contains the underlying parameter Θ_* with high probability. In particular, for $\delta \in (0,1)$, at time step t, it forms $\mathcal{C}_t(\delta) = \{\Theta : \|\Theta - \hat{\Theta}_t\|_{V_t} \leq \beta_t(\delta)\}$, for $\beta_t(\delta) = \sigma_w \sqrt{2n \log(\delta^{-1}\sqrt{\det(V_t)\,/\det(\lambda I)})} + \sqrt{\lambda}S$ and $V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i^\top$ such that $\Theta_* \in \mathcal{C}_t(\delta)$ with probability at least $1-\delta$ for all time steps t. Note that this estimation method and the learning guarantee is standard in learning linear dynamical systems since Abbasi-Yadkori and Szepesvári (2011).

Algorithm 1 StabL

```
Input: $\kappa$, $\gamma$, $Q$, $R$, $\sigma_w^2$ $\bar{\sigma}_w^2$, $V_0 = \lambda I$, $\hat{\Theta}_0 = 0$, $\tau = 0$
for $t = 0, \dots, T$ do
    if $(\det(V_t) > 2\det(V_0))$ and $(t - \tau > H_0)$ then
        Estimate $\hat{\Theta}_t$ & find optimistic $\tilde{\Theta}_t \in \mathcal{C}_t(\delta) \cap \mathcal{S}$
        Set $V_0 = V_t$ and $\tau = t$.
    else
        $\tilde{\Theta}_t = \tilde{\Theta}_{t-1}$
    if $t \leq T_w$ then
        $u_t = K(\tilde{\Theta}_{t-1})x_t + \nu_t$    Improved Exploration
    else
        $u_t = K(\tilde{\Theta}_{t-1})x_t$            Stabilizing Control
    Pay cost $c_t$ & Observe $x_{t+1}$
    Update $V_{t+1} = V_t + z_t z_t^\top$ for $z_t = [x_t^\top \; u_t^\top]^\top$
```

The confidence set above provides a self-normalized bound on the model parameter estimates via design matrix V_t. StabL uses the OFU principle in this confidence set to design a policy. In particular, it chooses an optimistic parameter $\tilde{\Theta}_t$ from $\mathcal{C}_t \cap \mathcal{S}$, which has the lowest expected optimal cost, and constructs the optimal linear controller $K(\tilde{\Theta}_t)$ for $\tilde{\Theta}_t$, *i.e.* the optimistic controller. At time t, StabL uses the optimistic controller $K(\tilde{\Theta}_{t-1})$. This choice is for technical reasons to guarantee persistence of excitation (Appendix B).

The optimistic controllers allow StabL to adaptively balance exploration and exploitation. They guide the exploration towards the region of state-space with the lowest expected cost. The key idea in this design is that as the confidence set shrinks, the performance of StabL improves over time (Bittanti et al., 2006).

Due to lack of an initial stabilizing policy, StabL aims to rapidly stabilize the system to avoid the consequences of unstable dynamics. To stabilize an unknown LQR, one requires sufficient exploration in all directions of the state-space (Lemma 4.2). Unfortunately, due to lack of reliable estimates in the early stages, the optimistic policies come short to guarantee such an effective exploration.

Therefore, StabL deploys an adaptive control policy with an improved exploration in the early stages of interactions with the system. In particular, for the first T_w time-steps StabL uses isotropic perturbations along with the optimistic controller. For $t \leq T_w$, it injects an i.i.d. Gaussian vector $\nu_t \sim \mathcal{N}(0, \sigma_\nu^2 I)$ to the system besides the optimistic policy $K(\tilde{\Theta}_{t-1})x_t$, where $\sigma_\nu^2 = 2\kappa^2 \bar{\sigma}_w^2$.

StabL effectively excites and explores all dimensions of the system via this improved exploration strategy (Theorem 4.1). The duration of the adaptive control with improved exploration phase is chosen such that StabL quickly finds a stabilizing controller. In particular, after $T_w := poly(\sigma_w, \sigma_\nu, n, d, \gamma^{-1}, \kappa, \overline{\alpha}, \log(1/\delta))$ time steps, StabL has the guarantee that the linear controllers $K(\tilde{\Theta}_{t-1})$ stabilize Θ_* for all $t \geq T_w$ with high probability (Lemma 4.1 & 4.2).

Moreover, StabL avoids frequent updates in the system estimates and the controller. It uses the same controller at least for a fixed time period of $H_0 = O(\gamma^{-1}\log(\kappa))$ and also waits for a significant improvement in the estimates. The latter is achieved by updating the controller if the determinant of the design matrix V_t is doubled since the last update. This update rule is chosen such that policy changes do not cause unstable dynamics for the stabilizable LQR. The effect of this update rule on maintaining bounded state for StabL are studied in detail in Section 4.1.

3.2 Stabilizing Adaptive Control

After guaranteeing the stabilizing policy design, StabL starts the adaptive control that stabilizes the underlying system. In this phase, StabL stops injecting isotropic perturbations and relies on the balanced exploration and exploitation via the optimistic controller design. The stabilizing optimistic controllers further guide the exploration to adapt the structure of the problem and fine-tune the learning process to achieve optimal performance. However, note that the frequent policy changes can still cause unbounded growth of the state even though the policies are stabilizing. Therefore, StabL continues the same policy update rule in this phase to maintain bounded state.

Unlike the prior works that constitute two distinct phases, StabL has a very subtle two-phase structure. In particular, the same subroutine (optimism) is applied continuously with the aim of balancing exploration and exploitation. An additional isotropic perturbation is only deployed for an improved exploration in the early stages to achieve stable learning for the autonomous agent.

4 THEORETICAL ANALYSIS

In this section, we study the main theoretical contributions of this work. In Section 4.1, we discuss the challenges that the stabilizability setting brings compared to the setting of the prior learning algorithms for the online LQR. We then introduce our approaches to overcome these challenges in the design of StabL. In Section 4.2, we provide the formal statements for the theoretical guarantees of StabL and, finally, we give the regret upper bound of StabL in Section 4.3.

4.1 Challenges in the Online Stabilizable LQR Problem

The main challenge for learning algorithms in control problems is to achieve input-to-state stability (ISS), which requires having well-bounded state in future time steps via using bounded inputs. Achieving this becomes significantly more challenging in the setting of stabilizable LQR compared to their controllable counterpart considered in many recent works (Abbasi-Yadkori and Szepesvári, 2011; Mania et al., 2019; Chen and Hazan, 2020). A controllable system can be brought to $x_t = 0$ in finite time steps. Furthermore, some of these works assume that the underlying system to be closed-loop contractible, *i.e.* $\|A_* - B_* K(\Theta_*)\| < 1$. These facts significantly simplify the overall stabilization problem. Moreover, recalling Definition 2.1, for controllable systems the controllability matrix is full row rank. In prior works, this has been a prominent factor in guaranteeing the persistence of excitation (PE) of the inputs, identifying the system and deriving regret bounds, e.g. (Hazan et al., 2019; Chen and Hazan, 2020).

Unfortunately, we do not have these properties in the general stabilizable LQR setting. Recall Assumption 2.2 that states the system is (κ, γ)-stabilizable, which yields $\rho(A_* + B_* K(\Theta_*)) \leq 1 - \gamma$ for the optimal policy $K(\Theta_*) \leq \kappa$. Therefore, even if the optimal policy of the underlying system is chosen by the learning algorithm, it may not produce contractive closed-loop system, *i.e.*, we can have $\rho(A_* + B_* K(\Theta_*)) < 1 < \|A_* + B_* K(\Theta_*)\|$ since for any matrix M, $\rho(M) \leq \|M\|$.

Moreover, from the definition of stabilizability in Definitions 2.1 and 2.2, 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 StabL 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 growth of the state over time.

In order to resolve these problems, StabL carefully designs the timing of the policy updates and applies all the policies long enough, so that the state stays well controlled, *i.e.*, ISS is achieved. To this end, StabL applies the same policy at least for $H_0 = 2\gamma^{-1}\log(2\kappa\sqrt{2})$ time steps. This particular choice prevents state blow-ups due to policy changes in the optimistic controllers in the stabilizable LQR setting (see Appendix D).

To achieve PE and consistent model estimates under the stabilizability condition, we leverage the early improved exploration strategy which does not require controllability. Using the isotropic exploration in the early stages, we derive a novel lower bound for the smallest eigenvalue of the design matrix V_t in the stabilizable LQR with sub-Gaussian noise setting. Moreover, we derive our regret results using the fast stabilization and the optimistic policy design of StabL. The results only depend on the stabilizability and other trivial model properties such as the LQR dimensions.

4.2 Benefits of Early Improved Exploration

To achieve effective exploration in the early stages, StabL deploys isotropic perturbations along with the optimistic policy for $t \leq T_w$. Define $\sigma_\star > 0$ where $\sigma_\star$ is a problem and in particular $\bar{\sigma}_w, \sigma_w, \sigma_\nu$-dependent constant (See Appendix B for exact definition). The following shows that for a long enough improved exploration, the inputs are persistently exciting the system.

Theorem 4.1 (Persistence of Excitation During the Improved Exploration). *If StabL follows the early improved exploration strategy for* $T \geq poly(\sigma_w^2, \sigma_\nu^2, n, \log(1/\delta))$ *time steps, then with probability at least* $1-\delta$*, StabL has* $\sigma_{\min}(V_T) \geq \sigma_\star^2 T$.

This theorem shows that having isotropic perturbations along with the optimistic controllers provides persistence excitation of the inputs, *i.e.* linear scaling of the smallest eigenvalue of the design matrix V_t. This result is quite technical and its proof is given in Appendix B. At a high-level, we show that isotropic perturbations allow the covariates to have a Gaussian-like tail lower bound even in the stabilizable LQR with sub-Gaussian process noise setting. Using the standard covering arguments, we prove the statement of the theorem. This result guarantees that the inputs excite all dimensions of the state-space and allows StabL to obtain uniformly improving estimates at a faster rate.

Lemma 4.1 (Parameter estimation error). *Suppose Assumptions 2.1 and 2.2 hold. For* $T \geq poly(\sigma_w^2, \sigma_\nu^2, n, \log(1/\delta))$ *time steps of adaptive control with improved exploration, with probability at least* $1-2\delta$*, StabL achieves* $\|\hat{\Theta}_T - \Theta_*\|_2 \leq \beta_t(\delta)/(\sigma_\star\sqrt{T})$.

This lemma shows that early improved exploration strategy using $\nu_t \sim \mathcal{N}(0, \sigma_\nu^2)$ for $\sigma_\nu^2 = 2\kappa^2\bar{\sigma}_w^2$ enables to guarantee the consistency of the parameter estimation. The proof is in Appendix C, where we combine the confidence set construction in Section 3.1 with Theorem 4.1. This bound is utilized to guarantee stabilizing controllers after early improved exploration. However, first we have the following lemma, which shows that there is a stabilizing neighborhood around Θ_*, such that $K(\Theta')$ stabilizes Θ_* for any Θ' in this region.

Lemma 4.2 (Strongly Stabilizable Neighborhood). *For* $D = \bar{\alpha}\gamma^{-1}\kappa^2(1+\kappa^2)$*, let* $C_0 = 142D^8$ *and* $\epsilon = 1/(54D^5)$*. For any* (κ, γ)*-stabilizable system* Θ_* *and for any* $\varepsilon \leq \min\{\sqrt{\bar{\sigma}_w^2 nD/C_0}, \epsilon\}$*, such that* $\|\Theta' - \Theta_*\| \leq \varepsilon$*,* $K(\Theta')$ *produces* (κ', γ')*-stable closed-loop dynamics on* Θ_* *where* $\kappa' = \kappa\sqrt{2}$ *and* $\gamma' = \gamma/2$.

The proof is given in Appendix A. This lemma shows that to guarantee the stabilization of the unknown dynamics a learning agent should have uniformly sufficient exploration in all directions of the state-space. By the choice of T_w (precise expression given in Appendix D) and using Lemma 4.1, StabL guarantees to quickly find this stabilizing neighborhood with high probability due to the adaptive control with improved exploration phase of T_w time steps.

For the remaining time steps, $t \geq T_w$, StabL starts redressing the possible state explosion due to unstable controllers and the perturbations in the early stages. Define T_{base} and T_r such that $T_{base} = (n+d)\log(n+d)H_0$ and $T_r = T_w + T_{base}$. Recall that H_0 is the minimum duration for a controller such that the state is well-controlled despite the policy changes. The following shows that the stabilizing controllers are applied long enough that the state stays bounded for $T > T_r$.

Lemma 4.3 (Bounded states). *Suppose Assumption 2.1 & 2.2 hold. For given* T_w *and* T_{base}*, StabL controls the state such that* $\|x_t\| = O((n+d)^{n+d})$ *for* $t \leq T_r$*, with probability at least* $1-2\delta$ *and* $\|x_t\| \leq (12\kappa^2 + 2\kappa\sqrt{2})\gamma^{-1}\sigma_w\sqrt{2n\log(n(t-T_w)/\delta)}$ *for* $T \geq t > T_r$*, with probability at least* $1-4\delta$.

In the proof (Appendix D), we show the policies seldom change via determinant doubling condition or the lower bound of H_0 for the adaptive control with improved exploration phase to keep the state bounded. For the stabilizing adaptive control, we show that deploying stabilizing policies for at least H_0 time-steps provides an exponential decay on the state and after T_{base} time-steps brings the state to an equilibrium.

4.3 Regret Upper Bound of StabL

After showing the effect of fast stabilization, we can finally present the regret upper bound of StabL.

Theorem 4.2 (Regret of StabL). *Suppose Assump-*

Table 2: Regret Performance After 200 Time Steps in Marginally Unstable Laplacian System. StabL outperfoms other algorithms by a significant margin

Algo.	Avg.	Top	Top	Top
	Regret	**90%**	**75%**	**50%**
StabL	$\mathbf{1.5{\times}10^4}$	$\mathbf{1.3{\times}10^4}$	$\mathbf{1.1{\times}10^4}$	$\mathbf{8.9{\times}10^3}$
OFULQ	$6.2{\times}10^{10}$	$4.0{\times}10^6$	$3.5{\times}10^5$	$4.7{\times}10^4$
CEC-Fix	$3.7{\times}10^{10}$	$2.1{\times}10^4$	$1.9{\times}10^4$	$1.7{\times}10^4$
CEC-Dec	$4.6{\times}10^4$	$4.0{\times}10^4$	$3.5{\times}10^4$	$2.8{\times}10^4$

Table 3: Maximum State Norm in the Laplacian System. StabL keeps the state smallest

Algo.	Avg.	Worst	Worst	Worst
	$\max\|x\|_2$	**5%**	**10%**	**25%**
StabL	$\mathbf{1.3{\times}10^1}$	$\mathbf{2.2{\times}10^1}$	$\mathbf{2.1{\times}10^1}$	$\mathbf{1.9{\times}10^1}$
OFULQ	$9.6{\times}10^3$	$1.8{\times}10^5$	$9.0{\times}10^4$	$3.8{\times}10^4$
CEC-Fix	$3.3{\times}10^3$	$6.6{\times}10^4$	$3.3{\times}10^4$	$1.3{\times}10^4$
CEC-Dec	$2.0{\times}10^1$	$3.5{\times}10^1$	$3.3{\times}10^1$	2.9×10^1

tions 2.1 and 2.2 hold. For the given choices of T_w and T_{base}, with probability at least $1-4\delta$, StabL 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 F. Here, we provide a proof sketch. The regret decomposition leverages the optimistic controller design. Recall that for the early improved exploration, StabL applies independent perturbations through the controller yet still deploys the optimistic policy. Thus, we consider this external perturbation as a part of the underlying system and study the regret obtained by the improved exploration strategy separately.

In particular, denote the system evolution noise at time t as ζ_t. For $t \leq T_w$, system evolution noise can be considered as $\zeta_t = B_* \nu_t + w_t$ and for $t > T_w$, $\zeta_t = w_t$. We denote the optimal average cost of system $\tilde{\Theta}$ under ζ_t as $J_*(\tilde{\Theta}, \zeta_t)$. Using the Bellman optimality equation for LQR (Bertsekas, 1995), we consider the system evolution of the optimistic system $\tilde{\Theta}_t$ using the optimistic controller $K(\tilde{\Theta}_t)$ in parallel with the true system evolution of Θ_* under $K(\tilde{\Theta}_t)$ such that they share the same process noise (See details in Appendix F). Using the confidence set construction, optimistic policy, Lemma 4.3, Assumption 2.2 and Lemma 2.1, we get a regret decomposition and bound each term separately.

At a high-level, the exact regret expression has a constant regret term due to early additional exploration for T_w time-steps with exponential dimension dependency and a term that scales with square root of the duration of stabilizing adaptive control with polynomial dimension dependency, *i.e.* $(n+d)^{n+d}T_w + \mathrm{poly}(n,d)\sqrt{T-T_w}$. Note that T_w is a problem dependent expression. Thus, for large enough T, the polynomial dependence dominates, giving Theorem 4.2.

5 EXPERIMENTS

In this section, we evaluate the performance of StabL in four adaptive control tasks: **(1)** a marginally unstable Laplacian system (Dean et al., 2018), **(2)** the longitudinal flight control of Boeing 747 with linearized dynamics (Ishihara et al., 1992), **(3)** unmanned aerial vehicle (UAV) that operates in a 2-D plane (Zhao et al., 2021), and **(4)** a stabilizable but not controllable linear dynamical system. For each task, we compare StabL with three RL algorithms: (i) OFULQ of Abbasi-Yadkori and Szepesvári (2011); (ii) certainty equivalent controller with fixed isotropic perturbations (CEC-Fix), which is the standard baseline in control theory; and (iii) certainty equivalent controller with decaying isotropic perturbations (CEC-Dec), which is shown to *achieve optimal regret with a given initial stabilizing policy* (Simchowitz and Foster, 2020; Dean et al., 2018; Mania et al., 2019). In the implementation of CEC-Fix and CEC-Dec, the optimal control policies of the estimated model are deployed. Furthermore, in finding the optimistic parameters for StabL and OFULQ, we use projected gradient descent within the confidence sets. We perform 200 independent runs for each algorithm for 200 time steps starting from $x_0 = 0$. We present the performance of best parameter choices for each algorithm. For further details and the experimental results please refer to Appendix I.

Before discussing the experimental results, we would like to highlight the baselines choices. Unfortunately, there are only a few works in literature that consider RL in LQRs without a stabilizing controller. These works are OFULQ of (Abbasi-Yadkori and Szepesvári, 2011), (Abeille and Lazaric, 2018), and (Chen and Hazan, 2020). Among these, (Chen and Hazan, 2020) considers LQRs with adversarial noise setting and deploys *impractically large inputs*, e.g. 10^{28} for task **(1)**, whereas the algorithm of (Abeille and Lazaric, 2018) only works in scalar setting. These prohibit meaningful regret and stability comparisons, thus, we compare StabL against the only relevant comparison of OFULQ among these. Moreover, there are only a few and limited experimental studies in the literature of RL in LQRs. Among these, (Dean et al., 2018; Faradonbeh et al., 2018b, 2020) highlight the superior performance of CEC-Dec. Therefore, we compare StabL against CEC-Dec with the best-performing parameter choice, as well as the standard control baseline of CEC-Fix.

(1) Laplacian system (Appendix I.1). Table 2 provides the regret performance for the average, top 90%, top 75% and top 50% of the runs of the algo-

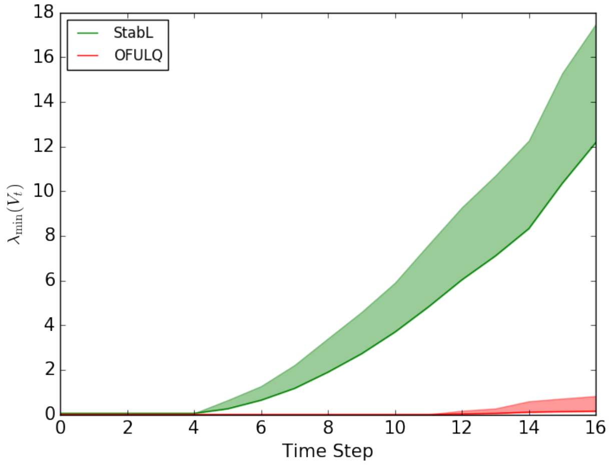

Figure 1: Evolution of the smallest eigenvalue of the design matrix for StabL and OFULQ in Laplacian system. The solid line is the mean and the shaded region is one standard deviation. StabL attains linear scaling whereas OFULQ suffers from lack of early exploration.

rithms. We observe that StabL attains at least an order of magnitude improvement in regret over OFULQ and CECs. This setting combined with the unstable dynamics is challenging for the *solely* optimism-based learning algorithms. Our empirical study indicates that, at the early stages of learning, the smallest eigenvalue of the design matrix V_t for OFULQ is much smaller than that of StabL as shown in Figure 1. The early improved exploration strategy helps StabL achieve linear scaling in $\lambda_{\min}(V_t)$, thus persistence of excitation and identification of stabilizing controllers. In contrast, the only OFU-based controllers of OFULQ fail to achieve persistence of excitation and accurate estimate of the model parameters. Therefore, due to lack of reliable estimates and the skewed cost, OFULQ cannot design effective strategies to learn model dynamics and results in unstable dynamics (see Table 3). Table 3 displays the stabilization capabilities of the deployed RL algorithms. In particular, it provides the averages of the maximum norms of the states for all runs, the worst 5%, 10% and 25% runs. Of all algorithms, StabL keeps the state smallest.

(2) Boeing 747 (Appendix I.2). In practice, nonlinear systems, like Boeing 747, are modeled via local linearizations which hold as long as the states are within a certain region. Thus, to maintain the validity of such linearizations, the state of the underlying system must be well-controlled, *i.e.*, stabilized. Table 4 provides the regret performances and Table 5 displays the stabilization capabilities of the deployed RL algorithms similar to **(1)**. Once more, among all algorithms, StabL maintains the maximum norm of the state smallest and operates within the smallest radius around the linearization point of origin. This observation is consistent among tasks **(3)** and **(4)**, which shows that StabL maintains tightly bounded state with high probability. The specifics of the maximum state results on **(3)** and **(4)** are given in the Appendix I.3 and I.4 respectively.

Table 4: Regret Performance After 200 Time Steps in Boeing 747 Flight Control. StabL outperfoms others.

Algo.	Avg. Regret	Top 90%	Top 75%	Top 50%
StabL	$\mathbf{1.3{\times}10^4}$	$\mathbf{9.6{\times}10^3}$	$\mathbf{7.6{\times}10^3}$	$\mathbf{5.3{\times}10^3}$
OFULQ	$1.5{\times}10^8$	$9.9{\times}10^5$	$5.6{\times}10^4$	$8.9{\times}10^3$
CEC-Fix	$4.8{\times}10^4$	$4.5{\times}10^4$	$4.3{\times}10^4$	$3.9{\times}10^4$
CEC-Dec	$2.9{\times}10^4$	$2.5{\times}10^4$	$2.2{\times}10^4$	$1.9{\times}10^4$

Table 5: Maximum State Norm in Boeing 747 Control. StabL keeps the state smallest.

Algo.	Avg. $\max\|x\|_2$	Worst 5%	Worst 10%	Worst 25%
StabL	$\mathbf{3.4{\times}10^1}$	$\mathbf{7.5{\times}10^1}$	$\mathbf{7.0{\times}10^1}$	$\mathbf{5.2{\times}10^1}$
OFULQ	$1.6{\times}10^3$	$2.2{\times}10^4$	$1.4{\times}10^4$	$6.3{\times}10^3$
CEC-Fix	$5.0{\times}10^1$	$7.8{\times}10^1$	$7.3{\times}10^1$	$6.5{\times}10^1$
CEC-Dec	$4.6{\times}10^1$	$8.0{\times}10^1$	$7.3{\times}10^1$	$6.3{\times}10^1$

(4) Stabilizable but not controllable system (Appendix I.4). Besides StabL, which is tailored for the general stabilizable setting, other algorithms perform poorly in this challenging setting. In particular, CEC-Fix drastically blows up the state due to significantly unstable dynamics for the uncontrollable part of the system. Therefore, the regret performances of only StabL, OFULQ and CEC-Dec are presented in Figure 2. Figure 2 is in semi-log scale and StabL provides an order of magnitude improved regret compared to the best performing state-of-art baseline CEC-Dec.

6 RELATED WORK

Finite-time regret guarantees: Prior works study the problem of regret minimization in LQRs and achieve sublinear regret using CECs (Mania et al., 2019; Faradonbeh et al., 2018b, 2020), robust controllers (Dean et al., 2018), the OFU principle (Abeille and Lazaric, 2020), Thompson sampling (Abeille and Lazaric, 2018) and an SDP relaxation (Cohen et al., 2019) with a lower bound provided in Simchowitz and Foster (2020). These works all assume that an initial stabilizing policy is given and do not design autonomous learning agents which is the main focus of this paper. Among these, Simchowitz and Foster (2020) provide the tight regret guarantee for the setting with known initial stabilizing policy. Their proposed algorithm follows the given non-adaptive ini-

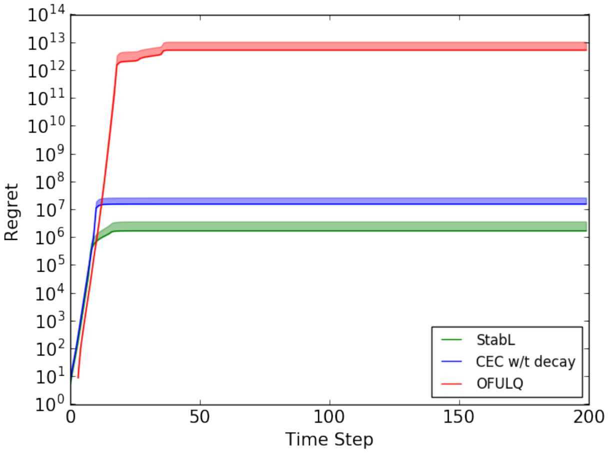

Figure 2: Regret Comparison of three algorithms in controlling a stabilizable but not controllable system. The solid lines are the average regrets and the shaded regions are the quarter standard deviations.

tial stabilizing policy for a *long* period of time with isotropic perturbations. Thus, they provide an order-optimal theoretical regret upper bound with an additional large constant regret. However, in many applications, e.g. medical, such constant regret, and non-adaptive controllers are not tolerable. StabL aims to address these challenges and provide an adaptive algorithm that can be deployed in practice. Moreover, StabL achieves significantly improved performance over the prior baseline RL algorithms in various adaptive control tasks (Section 5).

Finding a stabilizing controller: Similar to the regret minimization, there has been a growing interest in finite-time stabilization of linear dynamical systems (Dean et al., 2019; Faradonbeh et al., 2018a, 2019). Among these works, Faradonbeh et al. (2018a) is the closest to our work. However, there are significant differences in the methods and the span of the results. In Faradonbeh et al. (2018a), random linear controllers are used solely for finding a stabilizing set without a control goal. This results in the explosion of state, presumably exponentially in time, leading to a regret that scales exponentially in time. The proposed method provides many insightful aspects for finding a stabilizing set in finite-time, yet a cost analysis of this process or an adaptive control policy are not provided. Moreover, the stabilizing set in Faradonbeh et al. (2018a) relates to the minimum value that satisfies a specific condition for the roots of a polynomial. This results in a somewhat implicit sample complexity for constructing such a set. On the other hand, in this work, we provide a complete study of an autonomous learning algorithm for the online LQR problem. Among our results, we give an explicit formulation of the stabilizing set and a sample complexity that only relates to the minimal stabilizability information of the system.

Generalized LQR setting: Another line of research considers the generalizations of the online LQR problem under partial observability (Lale et al., 2020a,b,c; Mania et al., 2019; Simchowitz et al., 2020) or adversarial disturbances (Hazan et al., 2019; Chen and Hazan, 2020). These works either assume a given stabilizing controller or open-loop stable system dynamics, except Chen and Hazan (2020). Independently and concurrently, the recent work by Chen and Hazan (2020) designs an autonomous learning algorithm and regret guarantees that are similar to the current work. However, the approaches and the settings have major differences. Chen and Hazan (2020) considers the restrictive setting of *controllable* systems, yet with adversarial disturbances and general cost functions. They inject *significantly* big inputs, *exponential in system parameters*, with a pure exploration intent to guarantee the recovery of system parameters and stabilization. This negatively affects the practicality of the algorithm. On the other hand, in this work, we inject isotropic Gaussian perturbations to improve the exploration in the stochastic (sub-Gaussian process noise) *stabilizable* LQR while still aiming to control, *i.e.* no pure exploration phase. This yields a practical RL algorithm StabL that attains state-of-the-art performance.

7 CONCLUSION

In this paper, we propose an RL framework, StabL, 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, StabL 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 StabL is $\tilde{\mathcal{O}}(\sqrt{T})$ with polynomial dependence in the problem dimensions of the LQRs in stabilizable systems.

Our results highlight the benefit of early improved exploration to achieve improved regret at the expense of a slight increase in regret in the early stages. An important future direction is to study this phenomenon in more challenging online control problems in linear systems, e.g., under partially observability. Another interesting direction is to combine this mindset with the existing state-of-the-art model-based RL approaches for the general systems and study their performance.

References

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

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.

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.

Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. *arXiv preprint arXiv:2007.06650*, 2020.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time adaptive stabilization of lq systems. *arXiv preprint arXiv:1807.09120*, 2018a.

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. *arXiv preprint arXiv:2007.06482*, 2020.

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

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. *arXiv preprint arXiv:1703.08972*, 2017.

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.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

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

Sergio Bittanti, Marco C Campi, et al. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, 6(4):299–320, 2006.

Thomas Kailath, Ali H Sayed, and Babak Hassibi. Linear estimation, 2000.

Bernard Friedland. *Control system design: an introduction to state-space methods.* Courier Corporation, 2012.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

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

Tadashi Ishihara, Hai-Jiao Guo, and Hiroshi Takeda. A design of discrete-time integral controllers with computation delays via loop transfer recovery. *Automatica*, 28(3):599–603, 1992.

Feiran Zhao, Keyou You, and Tamer Basar. Infinite-horizon risk-constrained linear quadratic regulator with average cost. *arXiv preprint arXiv:2103.15363*, 2021.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive regulation and learning. *arXiv preprint arXiv:1811.04258*, 2018b.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear–quadratic regulators. *Automatica*, 117:108982, 2020.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Randomized algorithms for data-driven stabilization of stochastic linear systems. In *2019 IEEE Data Science Workshop (DSW)*, pages 170–174. IEEE, 2019.

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

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*, 2020b.

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

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

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. *arXiv preprint arXiv:1806.07104*, 2018.

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv preprint arXiv:2002.08095*, 2020.

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.

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.

Arthur Becker, P Kumar, and Ching-Zong Wei. Adaptive control with the stochastic approximation algorithm: Geometry and convergence. *IEEE Transactions on Automatic Control*, 30(4):330–338, 1985.

PR Kumar. Convergence of adaptive control schemes using least-squares parameter estimates. *IEEE Transactions on Automatic Control*, 35(4):416–424, 1990.

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Regret bounds for model-free linear quadratic control. *arXiv preprint arXiv:1804.06021*, 2018.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 5005–5014. PMLR, 2018.

Supplementary Material: Reinforcement Learning with Fast Stabilization in Linear Dynamical Systems

In Appendix A, we first provide a discussion on how stabilizability and (κ,γ)-stabilizable systems are equivalent. Then we prove that for (κ,γ)-stabilizable systems the unique positive definite solution to the DARE given in (2) is bounded. Finally, we show that there exists a stabilizing neighborhood (ball) around the true model parameters in which the optimal controllers of the models within this ball stabilize the underlying system in Appendix A. In Appendix B, we show that due to improved exploration strategy, the regularized design matrix V_t has its minimum eigenvalue scaling linearly over time, which guarantees the persistently exciting inputs for finding the stabilizing neighborhood and stabilizing controllers after the adaptive control with improved exploration phase. The exact definition of $\sigma_\star$ is also given in Lemma B.3 in Appendix B. We provide the system identification and confidence set constructions with their guarantees (both in terms of self-normalized and spectral norm) in Appendix C. In Appendix D, we provide the boundedness guarantees for the system's state throughout the execution of StabL and provide the proof of Lemma 4.3. The precise definition of T_w, which was omitted in the main text, is also given in (31) in Appendix D. We provide the regret decomposition in Appendix E and we analyze each term in this decomposition and give the proof of the main result of the paper in Appendix F. In Appendix G, we compare the results with Abbasi-Yadkori and Szepesvári (2011) and show it subsumes and improves the prior work. Appendix H provides the technical theorems and lemmas that are utilized in the proofs. Finally, in Appendix I, we provide the details on the experiments including the dynamics of the adaptive control tasks, parameter choices for the algorithms and additional experimental results.

A STABILIZABILITY OF THE UNDERLYING SYSTEM

In this section, we first show that (κ,γ)-stabilizability is merely a quantification of stabilizability. Then, we show that the given systems (both controllable and stabilizable) DARE has a unique positive definite solution. Finally, 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.

A.1 (κ,γ)-stabilizability

Any stabilizable system is also (κ,γ)-stabilizable for some κ and γ and the conversely (κ,γ)-stabilizability implies stabilizability. In particular, for all stabilizable systems, by setting $1-\gamma=\rho(A_*+B_*K(\Theta_*))$ and κ to be the condition number of $P(\Theta_*)^{1/2}$ where $P(\Theta_*)$ is the positive definite matrix that satisfies the following Lyapunov equation:

$$(A_*+B_*K(\Theta_*))^\top P(\Theta_*)(A_*+B_*K(\Theta_*)) \preceq P(\Theta_*), \tag{3}$$

one can show that $A_*+B_*K(\Theta_*)=HLH^{-1}$, where $H=P(\Theta_*)^{-1/2}$ and $L=P(\Theta_*)^{1/2}(A_*+B_*K(\Theta_*))P(\Theta_*)^{-1/2}$ with $\|H\|\|H^{-1}\|\leq\kappa$, and $\|L\|\leq 1-\gamma$ (Lemma B.1 of Cohen et al. (2018)).

A.2 Bound on the Solution of DARE for (κ, γ)-Stabilizable Systems, Proof of Lemma 2.1

Proof of Lemma 2.1: Recall the DARE given in (2). The solution of this equation corresponds to recursively applying the following

$$\begin{aligned}
\|P_*\| &= \| \sum\nolimits_{t=0}^{\infty} \left((A_* + B_* K(\Theta_*))^t\right)^\top \left(Q + K(\Theta_*)^\top R K(\Theta_*)\right) \left(A_* + B_* K(\Theta_*)\right)^t \| \\
&= \| \sum\nolimits_{t=0}^{\infty} \left(HL^t H^{-1}\right)^\top \left(Q + K(\Theta_*)^\top R K(\Theta_*)\right) \left(HL^t H^{-1}\right) \| \\
&\leq \overline{\alpha}(1 + \|K(\Theta_*)\|^2)\|H\|^2\|H^{-1}\|^2 \sum\nolimits_{t=0}^{\infty} \|L\|^{2t} \qquad (4) \\
&\leq \overline{\alpha}\gamma^{-1}\kappa^2(1+\kappa^2) \qquad (5)
\end{aligned}$$

where (4) follows from the upper bound on $\|Q\|, \|R\| \leq \overline{\alpha}$ and (5) follows from the definition of (κ, γ)-stabilizability. ■

A.3 Stabilizing Neighborhood Around the System Parameters

Theorem A.1 (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:*

$$P(\Theta_*) = A_*^\top P(\Theta_*) A_* + Q - A_*^\top P(\Theta_*) B_* \left(R + B_*^\top P(\Theta_*) B_*\right)^{-1} B_*^\top P(\Theta_*) A_*. \qquad (6)$$

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 stabilizable systems. Let $J_* \leq \mathcal{J}$. The following lemma is introduced in Simchowitz and Foster (2020) 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 ((Simchowitz and Foster, 2020)). *For constants $C_0 = 142\|P_*\|^8$ and $\epsilon = \frac{54}{\|P_*\|^5}$, 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 Θ_* obeys the following,*

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

This result shows that there exists a ϵ-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 (κ', γ')-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 proof of Lemma 4.2.

Proof of Lemma 4.2: Under Assumptions 2.1 & 2.2, for $\varepsilon \leq \min\{\sqrt{\mathcal{J}/C_0}, \epsilon\}$, we obtain $J(K(\Theta'), A_*, B_*, Q, R) \leq 2\mathcal{J}$. Plugging this into Lemma A.2 gives the presented result.

■

B SMALLEST SINGULAR VALUE OF REGULARIZED DESIGN MATRIX V_t

In this section, we show that improved exploration of StabL 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:

$$x_t = A_* x_{t-1} + B_* u_{t-1} + w_{t-1}, \quad and \quad u_t = K(\tilde{\Theta}_{t-1}) x_t + \nu_t \tag{7}$$

Let $\xi_t = z_t - \mathbb{E}\left[z_t | \mathcal{F}_{t-1}\right]$. Using the equalities in (7), and the fact that w_t and ν_t are $\mathcal{F}_t$ measurable, we write $\mathbb{E}\left[\xi_t \xi_t^\top | \mathcal{F}_{t-1}\right]$ as follows.

$$\begin{aligned}
\mathbb{E}\left[\xi_t \xi_t^\top | \mathcal{F}_{t-1}\right] &= \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix} \mathbb{E}\left[w_t w_t^\top | \mathcal{F}_{t-1}\right] \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{E}\left[\nu_t \nu_t^\top | \mathcal{F}_{t-1}\right] \end{pmatrix} \\
&= \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix} (\bar{\sigma}_w^2 I) \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 \\ 0 & \sigma_\nu^2 I \end{pmatrix} && (8) \\
&= \begin{pmatrix} \bar{\sigma}_w^2 I & \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1})^\top \\ \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1}) & \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1}) K(\tilde{\Theta}_{t-1})^\top + 2\kappa^2 \bar{\sigma}_w^2 I \end{pmatrix} && (9) \\
&\succeq \bar{\sigma}_w^2 \begin{pmatrix} I & K(\tilde{\Theta}_{t-1})^\top \\ K(\tilde{\Theta}_{t-1}) & 2K(\tilde{\Theta}_{t-1}) K(\tilde{\Theta}_{t-1})^\top + I/2 \end{pmatrix} && (10) \\
&= \frac{\bar{\sigma}_w^2}{2} I + \bar{\sigma}_w^2 \begin{pmatrix} \frac{1}{\sqrt{2}} I \\ \sqrt{2} K(\tilde{\Theta}_{t-1}) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} I \\ \sqrt{2} K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top && (11) \\
&\succeq \frac{\bar{\sigma}_w^2}{2} I && (12)
\end{aligned}$$

where (9) follows from $\sigma_\nu^2 = 2\kappa^2 \bar{\sigma}_w^2$ and (10) 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 $v \in \mathbb{R}^{n+d}$. (12) shows that that $\mathrm{Var}\left[s_t | \mathcal{F}_{t-1}\right] \geq \frac{\bar{\sigma}_w^2}{2}$.

Lemma B.1. *Suppose the system is stabilizable and we are in adaptive control with improved exploration phase of StabL. Denote $s_t = v^\top \xi_t$ where $v \in \mathbb{R}^{n+d}$ is any unit vector. Let $\bar{\sigma}_\nu := ((1+\kappa)^2 + 2\kappa^2)\sigma_w^2$. For a given positive σ_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 σ_1^2, and σ_2^2, such that $\sigma_1^2 \leq \sigma_2^2$, we have*

$$\mathbb{E}\left[E_t | \mathcal{F}_{t-1}\right] \geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_\nu^2 (1 + \frac{\sigma_2^2}{2\bar{\sigma}_\nu^2}) \exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})}{\sigma_2^2} \tag{13}$$

Note that, for any $\bar{\sigma}_\nu \geq \bar{\sigma}_w$, there is a pair (σ_1^2, σ_2^2) such that the right hand side of (13) 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 ν_t and w_t, for any t, are sub-Gaussian given $\mathcal{F}_{t-1}$, have that ξ_t is also sub-Gaussian vector. Therefore, s_t is a sub-Gaussian random variable with parameter $\bar{\sigma}_\nu$, where $\bar{\sigma}_\nu := ((1+\kappa)^2 + 2\kappa^2)\sigma_w^2$.

$$\begin{aligned}
\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 &\leq \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_1^2) | \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_2^2 \geq s_t^2 \geq \sigma_1^2) | \mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_2^2) | \mathcal{F}_{t-1}\right] && (14)
\end{aligned}$$

For the second term in the right hand side of the (14), 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,

$$\mathbb{E}\left[s_t^2\mathbb{1}(s_t^2\geq\sigma_2^2)|\mathcal{F}_{t-1}\right] = \int_{s^2\geq\sigma_2^2}\mathbb{P}(s_t^2\geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}). \tag{15}$$

Using this equality, we extend the (14) as follows,

$$\begin{aligned}
\frac{\bar\sigma_w^2}{2} - \sigma_1^2 &\leq \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_2^2\geq s_t^2\geq\sigma_1^2)|\mathcal{F}_{t-1}\right] + \int_{s^2\geq\sigma_2^2}\mathbb{P}(s_t^2\geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}) \\
&\leq \sigma_2^2\ \mathbb{E}\left[\mathbb{1}(\sigma_2^2\geq s_t^2\geq\sigma_1^2)|\mathcal{F}_{t-1}\right] + \int_{s^2\geq\sigma_2^2}\mathbb{P}(s_t^2\geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}) \\
&\leq \sigma_2^2\ \mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] + \int_{s^2\geq\sigma_2^2}\mathbb{P}(s_t^2\geq s^2|\mathcal{F}_{t-1})ds^2 + \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1}). \qquad (16)
\end{aligned}$$

Rearranging this inequality, we have,

$$\begin{aligned}
\mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] &\geq \frac{\frac{\bar\sigma_w^2}{2} - \sigma_1^2 - \int_{s^2\geq\sigma_2^2}\mathbb{P}(s_t^2\geq s^2|\mathcal{F}_{t-1})ds^2 - \sigma_2^2\ \mathbb{P}(s_t^2\geq\sigma_2^2|\mathcal{F}_{t-1})}{\sigma_2^2} \\
&\geq \frac{\frac{\bar\sigma_w^2}{2} - \sigma_1^2 - 2\int_{s^2\geq\sigma_2^2}\exp(\frac{-s^2}{2\bar\sigma_\nu^2})ds^2 - 2\sigma_2^2\exp(\frac{-\sigma_2^2}{2\bar\sigma_\nu^2})}{\sigma_2^2} \\
&\geq \frac{\frac{\bar\sigma_w^2}{2} - \sigma_1^2 - 4\bar\sigma_\nu^2\exp(\frac{-\sigma_2^2}{2\bar\sigma_\nu^2}) - 2\sigma_2^2\exp(\frac{-\sigma_2^2}{2\bar\sigma_\nu^2})}{\sigma_2^2} \\
&= \frac{\frac{\bar\sigma_w^2}{2} - \sigma_1^2 - 4\bar\sigma_\nu^2(1+\frac{\sigma_2^2}{2\bar\sigma_\nu^2})\exp(\frac{-\sigma_2^2}{2\bar\sigma_\nu^2})}{\sigma_2^2} \qquad (17)
\end{aligned}$$

The inequality in (17) 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 σ_1^2 and σ_2^2 that makes right hand side of (13) positive, let c_p denote the right hand side of (13), $c_p = \frac{\frac{\bar\sigma_w^2}{2}-\sigma_1^2-4\bar\sigma_\nu^2(1+\frac{\sigma_2^2}{2\bar\sigma_\nu^2})\exp(\frac{-\sigma_2^2}{2\bar\sigma_\nu^2})}{\sigma_2^2}$.

Lemma B.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 σ_1^2, and σ_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 B.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 σ_3, using a similar derivation as in (15) and (16), 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,

$$\begin{aligned}
\frac{\bar{\sigma}_w^2}{2} - 4\bar{\sigma}_\nu^2(1 + \frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}) &\leq \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] \\
&= \sigma_3^2\left(\mathbb{E}\left[\frac{s_t^2}{\sigma_3^2}\mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{s_t^2}{\sigma_3^2}\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right]\right) \\
&\leq \sigma_3^2\left(\mathbb{E}\left[\frac{|s_t|}{\sigma_3}\mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right]\right)
\end{aligned} \tag{18}$$

Note the for a large enough σ_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 (18),

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

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:

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

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

$$\begin{aligned}
\mathbb{E}\left[|s_t|\mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[s_t\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t\mathbb{1}(s_t \geq \sigma_3)|\mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[s_t\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + \bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})
\end{aligned} \tag{19}$$

similarly we have

$$\mathbb{E}\left[s_t\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[|s_t|\mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}) \tag{20}$$

Using the inequality (19) on the right hand side of (18), we have

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

Similarly, using (19) on the right hand side of (18) we have

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

Therefore, it results in the two following lower bounds,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}(s_t < 0)|\mathcal{F}_{t-1}\right] &\geq \frac{\frac{\bar{\sigma}_w^2}{2} - 4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{2\sigma_3^2} - 0.5\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}) \\ \mathbb{E}\left[\mathbb{1}(s_t > 0)|\mathcal{F}_{t-1}\right] &\geq \frac{\frac{\bar{\sigma}_w^2}{2} - 4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{2\sigma_3^2} - 0.5\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}) \qquad (21)\end{aligned}$$

Choosing σ_3 sufficiently large results in the right hand sides in inequalities (21) to be positive and bounded away form zero. Let $c''_p > 0$ denote the right hand sides in the (21). 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 B.3 (Precise version of Lemma 4.1, Persistence of Excitation During the Extra Exploration)**.** *If the duration of the adaptive control with improved exploration $T_w \geq \frac{6n}{c'_p}\log(12/\delta)$, then with probability at least $1-\delta$, StabL 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 B.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 $T_w \geq 8\log(1/\delta)/c_p'^2$, 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 $T_w \geq \frac{20}{c_p'^2}((n+d)+\log(1/\delta))$, 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 $1-\delta$. □

C SYSTEM IDENTIFICATION & CONFIDENCE SET CONSTRUCTION

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. First let

$$\kappa_e = \left(\frac{\sigma_w}{\sigma_\star} \sqrt{n(n+d)\log\left(1 + \frac{cT(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda(n+d)}\right) + 2n\log\frac{1}{\delta}} + \sqrt{\lambda}S \right) \tag{22}$$

Proof. Define $\Theta_*^\top = [A, B]$ and $z_t = \begin{bmatrix} x_t^\top u_t^\top \end{bmatrix}^\top$. 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,

$$X_T = Z_T \Theta_* + W_T \tag{23}$$

for

$$X_T = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_{T-1}^\top \\ x_T^\top \end{bmatrix} \in \mathbb{R}^{T\times n} \quad Z_T = \begin{bmatrix} z_1^\top \\ z_2^\top \\ \vdots \\ z_{T-1}^\top \\ z_T^\top \end{bmatrix} \in \mathbb{R}^{T\times(n+d)} \quad W_T = \begin{bmatrix} w_1^\top \\ w_2^\top \\ \vdots \\ w_{T-1}^\top \\ w_T^\top \end{bmatrix} \in \mathbb{R}^{T\times n}. \tag{24}$$

Then, we estimate Θ_* 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,

$$\begin{aligned}
|\operatorname{Tr}((\hat{\Theta}_T-\Theta_*)^\top X)| &= |\operatorname{Tr}(W_T^\top Z_T(Z_T^\top Z_T+\lambda I)^{-1}X) - \lambda \operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T+\lambda I)^{-1}X)| \\
&\leq |\operatorname{Tr}(W_T^\top Z_T(Z_T^\top Z_T+\lambda I)^{-1}X)| + \lambda |\operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T+\lambda I)^{-1}X)| \\
&\leq \sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T+\lambda I)^{-1}X)\operatorname{Tr}(W_T^\top Z_T(Z_T^\top Z_T+\lambda I)^{-1}Z_T^\top W_T)} \\
&\quad + \lambda\sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T+\lambda I)^{-1}X)\operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T+\lambda I)^{-1}\Theta_*)}, \qquad (25) \\
&= \sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T+\lambda I)^{-1}X)}\left[\sqrt{\operatorname{Tr}(W_T^\top Z_T(Z_T^\top Z_T+\lambda I)^{-1}Z_T^\top W_T)}+\lambda\sqrt{\operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T+\lambda I)^{-1}\Theta_*)}\right]
\end{aligned}$$

where (25) 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,

$$\operatorname{Tr}(W_T^\top Z_T(Z_T^\top Z_T+\lambda I)^{-1}Z_T^\top W_T) = \operatorname{Tr}(\mathcal{S}_T^\top V_T^{-1}\mathcal{S}_T) = \sum_{i=1}^{n} s_i^\top V_T^{-1} s_i = \sum_{i=1}^{n} \|s_i\|^2_{V_T^{-1}}. \tag{26}$$

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 σ_w-subgaussian, thus we can use Theorem H.1 to show that,

$$\operatorname{Tr}(W_T^\top Z_T(Z_T^\top Z_T+\lambda I)^{-1}Z_T^\top W_T) \leq 2n\sigma_w^2 \log\left(\frac{\det(V_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right). \tag{27}$$

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

$$\operatorname{Tr}((\hat{\Theta}_T-\Theta_*)^\top V_T(\hat{\Theta}_T-\Theta_*)) \leq \left(\sigma_w\sqrt{2n\log\left(\frac{\det(V_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)}+\sqrt{\lambda}S\right)^2. \tag{28}$$

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,

$$\|\hat{\Theta}_T-\Theta_*\|_2 \leq \frac{1}{\sqrt{\lambda_{\min}(V_T)}}\left(\sigma_w\sqrt{2n\log\left(\frac{\det(V_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)}+\sqrt{\lambda}S\right) \tag{29}$$

To complete the proof, we need a lower bound on $\lambda_{\min}(V_{T_w})$. Using Lemma 4.1, we obtain the following with probability at least $1-2\delta$:

$$\|\hat{\Theta}_{T_w}-\Theta_*\|_2 \leq \frac{\beta_t(\delta)}{\sigma_\star\sqrt{T_w}}.$$

From Lemma 4.3, 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 H.1,

$$\|\hat{\Theta}_{T_w}-\Theta_*\|_2 \leq \frac{\kappa_e}{\sqrt{T_w}}. \tag{30}$$

□

D BOUNDEDNESS OF STATES

In this section, we will provide the proof of Lemma 4.3, *i.e.* bounds on states for the adaptive control with improved exploration and stabilizing adaptive control phases. First define the following.

Let

$$T_w = \frac{\kappa_e^2}{\min\{\bar{\sigma}_w^2 nD/C_0, \epsilon^2\}} \tag{31}$$

such that for $T > T_w$, we have $\|\hat{\Theta}_T - \Theta_*\|_2 \leq \min\{\sqrt{\bar{\sigma}_w^2 nD/C_0}, \epsilon\}$ with probability at least $1 - 2\delta$. Notice that due to Lemma 4.2 and as shown in the following, these guarantee the stability of the closed-loop dynamics for deploying optimistic controller for the remaining part of StabL.

Choose an error probability, $\delta > 0$. Consider the following events, in the probability space Ω:

- 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

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

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

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 $\mathcal{F}_1^{[s]} \supseteq \mathcal{F}_2^{[s]} \supseteq \ldots \supseteq \mathcal{F}_{T_s}^{[s]}$. 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 D.1 (Abbasi-Yadkori and Szepesvári (2011))**.** *We have that for any* $0 \leq t \leq T$*,*

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

Notice that Lemma D.1 does not depend on controllability or the stabilizability of the system. Thus, we will use Lemma D.1 for $t \leq T_w$ for the adaptive control with improved exploration phase of StabL. Then we consider the effect of stabilizing controllers for the remaining time steps.

D.1 State Bound for the Adaptive Control with Improved Exploration Phase

One can write the state update as

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

where

$$\Gamma_t = \begin{cases} \tilde{A}_{t-1} + \tilde{B}_{t-1} K(\tilde{\Theta}_{t-1}) & t \notin \mathcal{T}_T \\ A_* + B_* K(\tilde{\Theta}_{t-1}) & t \in \mathcal{T}_T \end{cases} \quad \text{and} \quad r_t = \begin{cases} (\Theta_* - \tilde{\Theta}_{t-1})^\top z_t + B_* \nu_t + w_t & t \notin \mathcal{T}_T \\ B_* \nu_t + w_t & t \in \mathcal{T}_T \end{cases} \tag{32}$$

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

$$\begin{aligned} x_t &= \Gamma_{t-1} x_{t-1} + r_{t-1} = \Gamma_{t-1} \left(\Gamma_{t-2} x_{t-2} + r_{t-2} \right) + r_t \\ &= \Gamma_{t-1} \Gamma_{t-2} \Gamma_{t-3} x_{t-3} + \Gamma_{t-1} \Gamma_{t-2} r_{t-2} + \Gamma_{t-1} r_{t-1} + r_t \\ &= \Gamma_{t-1} \Gamma_{t-2} \dots \Gamma_{t-(t-1)} r_1 + \dots + \Gamma_{t-1} \Gamma_{t-2} r_{t-2} + \Gamma_{t-1} r_{t-1} + r_t \\ &= \sum_{k=1}^{t} \left(\prod_{s=k}^{t-1} \Gamma_s \right) r_k \end{aligned} \tag{33}$$

Recall that the controller is optimistically designed from set of parameters are (κ, γ)-strongly stabilizable by their optimal controllers. Therefore, we have

$$1 - \gamma \geq \max_{t \leq T} \rho \left(\tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t) \right). \tag{34}$$

Therefore, multiplication of closed-loop system matrices, $\tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t)$, is not guaranteed to be contractive. In Abbasi-Yadkori and Szepesvári (2011), the authors assume these matrices are contractive under controllability assumption. In order to bound the state similarly, 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 and produces bounded state. Thus, by choosing $H_0 = 2\gamma^{-1} \log(2\kappa\sqrt{2})$ and adopting Lemma 39 of Cassel et al. (2020), we have that

$$\|x_t\| \leq \frac{18\kappa^3 \overline{\eta}^{n+d}}{\gamma(8\kappa - 1)} \left(\max_{1 \leq k \leq t} \|r_k\| \right) \tag{35}$$

Furthermore, 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 D.1. The second term involves summation of independent $\|B_*\|\sigma_\nu$ and σ_w subgaussian vectors. Using Lemma H.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 $\mathcal{E}$,

$$\|x_t\| \leq \frac{18\kappa^3 \overline{\eta}^{n+d}}{\gamma(8\kappa - 1)} \left[G Z_t^{\frac{n+d}{n+d+1}} \beta_t(\delta)^{\frac{1}{2(n+d+1)}} + (\|B_*\|\sigma_\nu + \sigma_w) \sqrt{2n \log \frac{nt}{\delta}} \right] \tag{36}$$

for $t \leq T_w$. Using union bound, we can deduce that $\mathcal{E}_T \cap \mathcal{F}^{[s]}_{T_s}$ 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

$$\|x_t\| \leq c'(n+d)^{n+d}. \tag{37}$$

for some large enough constant c'. The adaptive control with improved exploration phase of StabL has this exponentially dimension dependent state bound for all $t \leq T_w$. In the following section, we show that during the stabilizing adaptive control phase, the bound on state has a polynomial dependency on the dimensions.

D.2 State Bound in Stabilizing Adaptive Control phase

In the stabilizing adaptive control phase, StabL stops using the additive isotropic exploration component ν_t, the state follows the dynamics of

$$x_{t+1} = (A_* + B_* K(\tilde{\Theta}_{t-1}))x_t + w_t \tag{38}$$

Denote $\mathbf{M_t} = A_* + B_* K(\tilde{\Theta}_{t-1})$ as the closed loop dynamics of the system. From the choice of T_w 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 $T > t > T_w$, if the same policy, $\mathbf{M}$ is applied starting from state x_{T_w}, we have

$$\|x_t\| = \left\| \prod_{i=T_w+1}^{t} \mathbf{M} x_{T_w} + \sum_{i=T_w+1}^{t} \left(\prod_{s=i}^{t-1} \mathbf{M} \right) w_i \right\| \tag{39}$$

$$\leq \kappa\sqrt{2}(1-\gamma/2)^{t-T_w} \|x_{T_w}\| + \max_{T_w < i \leq T} \|w_i\| \left(\sum_{i=T_w+1}^{t} \kappa\sqrt{2}(1-\gamma/2)^{t-i+1} \right) \tag{40}$$

$$\leq \kappa\sqrt{2}(1-\gamma/2)^{t-T_w} \|x_{T_w}\| + \frac{2\kappa\sigma_w\sqrt{2}}{\gamma} \sqrt{2n\log(n(t-T_w)/\delta)} \tag{41}$$

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 ith policy change after T_w, we get

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

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

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

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

Moreover, due to stability of the synthesized controller, the worst possible controller update scheme is to update the controller every H_0 time-steps, *i.e.,* invoking the condition of $t - \tau > H_0$ in the update rule. Notice that this update rule considers the worst effect of similarity transformation on the growth of the state, since otherwise applying the same controller for longer periods would have further reduction on the state due to the contraction that the stabilizing controller brings. Thus, from (41) we have that

$$\|x_t\| \leq \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma} \sqrt{2n\log(n(t-T_w)/\delta)}, \tag{42}$$

for all $t > T_r \coloneqq T_w + T_{base}$ where $T_{base} = ((n+d)\log(n+d))\, H_0$.

E REGRET DECOMPOSITION

The regret decomposition leverages the OFU principle. Since during the adaptive control with improved exploration period StabL applies independent isotropic perturbations through the controller but still designs the optimistic controller, one can consider the external perturbation as a component of the underlying system. With this way, we consider the regret obtained by using the improved exploration separately.

First noted that based on the definition of OFU principle, StabL solves $J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) + 1/\sqrt{t}$ to find the optimistic parameter. This search is done over only $\mathcal{C}_t(\delta)$ in the stabilizing adaptive control phase. Denote the system evolution noise at time t as ζ_t. For $t \leq T_w$, 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 ζ_t as $J_*(\tilde{\Theta}, \zeta_t)$. The regret of the StabL can be decomposed as

$$\sum_{t=0}^{T} x_t^\top Q x_t + u_t^\top R u_t + 2\nu_t^\top R u_t + \nu_t^\top R \nu_t - J_*(\Theta_*, w_t) \tag{43}$$

where u_t is the optimal controller input for the optimistic system $\tilde{\Theta}_{t-1}$, ν_t is the noise injected and x_t is the state of the system $\tilde{\Theta}_{t-1}$ with the system evolution noise of ζ_t. From Bellman optimality equation for LQR, (Bertsekas, 1995), we can write the following for the optimistic system, $\tilde{\Theta}_{t-1}$,

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

where $\tilde{P}_{t-1}$ is the solution of DARE for $\tilde{\Theta}_{t-1}$. Following the decomposition used in without additional exploration (Abbasi-Yadkori and Szepesvári, 2011), we get,

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

where we use the fact that $x_{t+1} = A_* x_t + B_* u_t + \zeta_t$, the martingale property of the noise and the conditioning on the filtration $\mathcal{F}_{t-1}$. Hence, summing up over time, we get

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

for

$$R_1^\zeta = \sum\nolimits_{t=0}^{T} \left\{ x_t^\top \tilde{P}_{t\text{-}1} x_t - \mathbb{E}\left[x_{t+1}^\top \tilde{P}_t x_{t+1}\big|\mathcal{F}_{t-1}\right]\right\} \tag{44}$$

$$R_2^\zeta = \sum\nolimits_{t=0}^{T} \mathbb{E}\left[x_{t+1}^\top \left(\tilde{P}_{t-1} - \tilde{P}_t\right) x_{t+1}\big|\mathcal{F}_{t-1}\right] \tag{45}$$

$$R_3^\zeta = \sum_{t=0}^{T} \bar{x}_{t+1,\tilde{\Theta}_{t-1}}^\top \tilde{P}_{t-1} \bar{x}_{t+1,\tilde{\Theta}_{t-1}} - \bar{x}_{t+1,\Theta_*}^\top \tilde{P}_{t-1} \bar{x}_{t+1,\Theta_*} \tag{46}$$

where $\bar{x}_{t+1,\tilde{\Theta}_{t-1}} = \tilde{A}_{t-1}x_t {+} \tilde{B}_{t-1}u_t$ and $\bar{x}_{t+1,\Theta_*} = A_* x_t {+} B_* u_t$.

Therefore, when we jointly have that $\Theta_* \in \mathcal{C}_t(\delta)$ for all time steps t and the state is bounded as shown in Lemma 4.3,

$$\sum_{t=0}^{T} (x_t^\top Q x_t {+} u_t^\top R u_t) {=} \sum_{t=0}^{T_w} \sigma_\nu^2 \operatorname{Tr}(\tilde{P}_{t-1} B_* B_*^\top) {+} \sum_{t=0}^{T} \bar{\sigma}_w^2 \operatorname{Tr}(\tilde{P}_{t-1}) {+} R_1^\zeta {-} R_2^\zeta {-} R_3^\zeta$$

where the equality follows from the fact that, $J_*(\tilde{\Theta}_{t-1},\zeta_t) = \operatorname{Tr}(\tilde{P}_{t-1} W)$ 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

$$\bar{\sigma}_w^2 \operatorname{Tr}(\tilde{P}_{t-1}) = J_*(\tilde{\Theta}_{t-1}, w_t) \leq J_*(\Theta_*, w_t) + 1/\sqrt{t} = \bar{\sigma}_w^2 \operatorname{Tr}(P_*) + 1/\sqrt{t}.$$

Combining this with (43) and Assumption 2.2, we obtain the following expression for the regret of StabL:

$$\mathrm{R}(T) \leq \sigma_\nu^2 T_w D \|B_*\|_F^2 + R_1^\zeta - R_2^\zeta - R_3^\zeta + \sum_{t=0}^{T_w} 2\nu_t^\top R u_t + \nu_t^\top R \nu_t. \tag{47}$$

F REGRET ANALYSIS

In this section, we provide the bounds on each term in the regret decomposition separately. We show that the regret suffered from the improved exploration is tolerable in the upcoming stages via the guaranteed stabilizing controller, yielding polynomial dimension dependency in regret.

F.1 Direct Effect of Improved 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 $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$

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

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

$$\sum_{t=0}^{T_w} \left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right) \leq d\sigma_\nu \sqrt{B_\delta} + d\|R\|\sigma_\nu^2 \left(T_w + \sqrt{T_w}\log\frac{4dT_w}{\delta}\sqrt{\log\frac{4}{\delta}}\right) \tag{48}$$

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 H.1 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 $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$ 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 $1-\frac{\delta}{2}$,

$$\sum_{t=0}^{T_w} 2\nu_t^\top R u_t \leq \\ d\sqrt{8\sigma_\nu^2\left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right)\log\left(\frac{4d}{\delta}\left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right)^{1/2}\right)} \tag{49}$$

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

$$\sum_{t=0}^{T_w}\nu_t^\top R\nu_t \leq dT_w\sigma_\nu^2\|R\| + 2\|R\|W^2\sqrt{2T_w\log\frac{4}{\delta}}. \tag{50}$$

Combining (49) and (50) gives the statement of lemma for the regret of external exploration noise.

□

F.2 Bounding R_1^ζ in the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$ or $\mathcal{E}_T \cap \mathcal{F}_{T_c}^{[c]}$

In this section, we state the bound on R_1^ζ given in (44). We first provide high probability bound on the system noise.

Lemma F.2 (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^ζ for stabilizable systems.

Lemma F.3 (Bounding R_1^ζ for StabL). *Let* R_1^ζ *be as defined by (44). Under the event of* $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$*, with probability at least* $1-\delta/2$*, using StabL for* $t > T_r$*, we have*

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

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

Proof. Assume that the event $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$ 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 H.1 on some event $G_{\delta,i}$ that holds with probability at least $1-\delta/(4n)$, for any $t \geq 0$,

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

Notice that StabL stops additional isotropic perturbation after $t = T_w$, and the state starts decaying until $t = T_r$. For simplicity of presentation we treat the time between T_w and T_r as exploration sacrificing the tightness of the result. On $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$, $\|v_k\| \leq DS(n+d)^{n+d}\sqrt{1+\kappa^2}$ for $k \leq T_r$ and $\|v_k\| \leq$

$\frac{(12\kappa^2+2\kappa\sqrt{2})DS\sigma_w\sqrt{1+\kappa^2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$ for $k > T_r$. Thus, $v_{k,i} \leq DS(n+d)^{n+d}\sqrt{1+\kappa^2}$ and $v_{k,i} \leq \frac{(12\kappa^2+2\kappa\sqrt{2})DS\sigma_w\sqrt{1+\kappa^2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$ respectively for $k \leq T_r$ and $k > T_r$. Using union bound we get, for probability at least $1-\frac{\delta}{4}$, for $t > T_r$,

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

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 H.2 with union bound and Theorem H.2, 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. □

F.3 Bounding $|R_2^\zeta|$ on the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$

In this section, we will bound $|R_2^\zeta|$ given in (45). We first provide a bound on the maximum number of policy changes.

Lemma F.4 (Number of Policy Changes for StabL)**.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[c]}$, StabL changes the policy at most*

$$\min\left\{T/H_0, (n+d)\log_2\left(1+\frac{\lambda+T_r(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_r)(1+\kappa^2)X_s^2}{\lambda}\right)\right\}, \tag{51}$$

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

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

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

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

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

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 of StabL 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^ζ are 0. Therefore, we have the following results for stabilizable systems.

Lemma F.5 (Bounding R_2^ζ for StabL)**.** *Let R_2^ζ be as defined by (45). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$, using StabL, we have*

$$\begin{aligned}|R_2^\zeta| \leq\ & 2D(n+d)^{2(n+d)+1}\log_2\left(1+\frac{T_r(n+d)^{2(n+d)}(1+\kappa^2)}{\lambda}\right)\\ &+2DX_s^2(n+d)\log_2\left(1+\frac{T_r(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_r)(1+\kappa^2)X_s^2}{\lambda}\right)\end{aligned}$$

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

Proof. On the event $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$, we know the maximum number of policy changes up to T_r and T using Lemma F.4. Using the fact that $\|x_t\| \leq (n+d)^{n+d}$ for $t \leq T_r$ and $\|x_t\| \leq \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$, we obtain the statement of the lemma. □

F.4 Bounding $|R_3^\zeta|$ on the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$

Before bounding R_3^ζ, first consider the following for stabilizable LQRs.

Lemma F.6. *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$, using StabL in a stabilizable LQR, the following holds,*

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

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

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 $\tau \leq T_r$, we have

$$
\begin{aligned}
\|(\Theta - \hat{\Theta}_t)^\top z_t\| &\leq \|V_t^{1/2}(\Theta - \hat{\Theta}_t)\| \|z_t\|_{V_t^{-1}} &(52)\\
&\leq \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\| \sqrt{\frac{\det(V_t)}{\det(V_\tau)}} \|z_t\|_{V_t^{-1}} &(53)\\
&\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}}\right\} \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\| \|z_t\|_{V_t^{-1}} &(54)\\
&\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}}\right\} \beta_\tau(\delta) \|z_t\|_{V_t^{-1}}. &(55)
\end{aligned}
$$

Similarly, for for all $\Theta \in \mathcal{C}_\tau(\delta)$, for $\tau > T_r$, 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,

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

where we use Lemma H.1. □

Using Lemma F.6, we bound R_3^ζ as follows.

Lemma F.7 (Bounding R_3^ζ for StabL)**.** *Let R_3^ζ be as defined by (46). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}^{[s]}$, using StabL with the choice of $\lambda = (1+\kappa^2)X_s^2$, we have*

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

Proof. Let $Y_1 = \frac{8(1+\kappa^2)\beta_T^2(\delta)}{\lambda}(n+d)^{2(n+d)} \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda}\right)^{H_0}\right\} \log\frac{\det(V_{T_r})}{\det(\lambda I)}$ and $Y_2 = \frac{8(1+\kappa^2)\beta_T^2(\delta)}{\lambda} X_s^2 \max\left\{2, \left(1 + \frac{(1+\kappa^2)X_s^2}{\lambda}\right)^{H_0}\right\} \log\frac{\det(V_T)}{\det(V_{T_r})}$ for $X_s = \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$. The following uses triangle inequality and Cauchy Schwarz inequality and again triangle inequality to give:

$$\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} \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}^{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} \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} \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}^{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}^{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} \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} \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}^{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}^{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{Y_1} \sqrt{4 T_r D (1+\kappa^2) S^2 (n+d)^{2(n+d)}} + \sqrt{Y_2} \sqrt{4 (T - T_r) D (1+\kappa^2) S^2 X_s^2} \\
&\leq \frac{\max\left\{ 8, 4\sqrt{2} \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\lambda} \right)^{H_0/2} \right\} DS(1+\kappa^2)\beta_T(\delta)(n+d)^{2(n+d)}}{\sqrt{\lambda}} \times \\
&\qquad \sqrt{T_r (n+d) \log\left(1 + \frac{T_r (1+\kappa^2)(n+d)^{2(n+d)}}{\lambda(n+d)} \right)} \\
&\quad + \frac{\max\left\{ 8, 4\sqrt{2} \left(1 + \frac{(1+\kappa^2) X_s^2}{\lambda} \right)^{H_0/2} \right\} DS(1+\kappa^2)\beta_T(\delta)}{\sqrt{\lambda}} X_s^2 \times \\
&\qquad \sqrt{(T - T_r)(n+d) \log\left(1 + \frac{T_r (1+\kappa^2)(n+d)^{2(n+d)} + (T - T_r) X_s^2}{\lambda(n+d)} \right)}
\end{aligned}$$

Examining the first term, it has the dimension dependency of $(n+d)^{(n+d)H_0} \times \sqrt{n(n+d)} \times (n+d)^{2(n+d)} \times \sqrt{n+d}$ where $\sqrt{n(n+d)}$ is due to $\beta_T(\delta)$. For the second term, with the choice of $\lambda = (1+\kappa^2)X_s^2$, the exponential dependency on the dimension with H_0 can be converted to a scalar multiplier, *i.e.*, $\left(1 + \frac{(1+\kappa^2)X_s^2}{\lambda}\right)^{H_0/2} = \sqrt{2}^{H_0}$ and $(1+\kappa^2)X_s^2/\sqrt{\lambda} = \sqrt{(1+\kappa^2)}X_s$. Therefore, for the second term, we have the dimension dependency of $\sqrt{n(n+d)} \times \sqrt{n} \times \sqrt{n+d}$ which gives the advertised bound.

□

F.5 Combining Terms for Final Regret Upper Bound

Proof of Theorem 4.2: Recall that

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

Combining Lemma F.1 for $\sum_{t=0}^{T_w} \left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t \right)$, Lemma F.3 for R_1^ζ, Lemma F.5 for $|R_2^\zeta|$ and Lemma F.7 for $|R_3^\zeta|$, we get the advertised regret bound. ■

G CONTROLLABILITY ASSUMPTION IN ABBASI-YADKORI AND SZEPESVARI (2011)

In Abbasi-Yadkori and Szepesvári (2011), the authors derive their results for the following setting:

Assumption G.1 (Controllable Linear Dynamical System)**.** *The unknown parameter Θ_* is a member of a set $\mathcal{S}_c$ such that*

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

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 $\sup\{\|K(\Theta')\| \mid \Theta' \in \mathcal{S}_c\} \leq \kappa$.

Our results are strict generalizations of these since stabilizable systems subsume controllable systems and all closed-loop contractible systems considered with Assumption G.1 is a subset of general stable closed-loop systems considered in this work. For the setting in Assumption G.1, we can bound the state following similar steps to stabilizable case but since the closed-loop system is contractible we do not need minimum length on epoch of an optimistic controller since the state would always shrink. Adopting the proofs provided in this work to Assumption G.1, one can obtain the similar polynomial dimension dependency via additional exploration of StabL. This shows that with additional exploration the result of Abbasi-Yadkori and Szepesvári (2011) could be directly improved.

H TECHNICAL THEOREMS AND LEMMAS

Theorem H.1 (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 η_k is conditionally sub-Gaussian with constant R. Consider the martingale*

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

and the matrix-valued processes

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

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 H.2 (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 H.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

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

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

I IMPLEMENTATION DETAILS OF EXPERIMENTS AND ADDITIONAL RESULTS

In this section, we provide the simulation setups, with the parameter settings for each algorithm and the details of the adaptive control tasks. The implementations and further system considerations are available at https://github.com/SahinLale/StabL. In the experiments, we use four adaptive control tasks:

(1) A marginally unstable Laplacian system (Dean et al., 2018)

(2) The longitudinal flight control of Boeing 747 with linearized dynamics (Ishihara et al., 1992)

(3) Unmanned aerial vehicle (UAV) that operates in a 2-D plane (Zhao et al., 2021)

(4) A stabilizable but not controllable linear dynamical system.

For each setting we deploy 4 different algorithms:

(i) Our algorithm StabL,

(ii) OFULQ of Abbasi-Yadkori and Szepesvári (2011),

(iii) Certainty equivalent controller (CEC) with fixed isotropic perturbations,

(iv) CEC with decaying isotropic perturbations.

For each algorithm there are different varying parameters. For each adaptive control task, we tune each parameter in terms of regret performance and present the performance of the best performing parameter choices since the regret analysis for each algorithm considers the worst case scenario. In each setting, we will specify these parameters choices for each algorithm. We use the actual errors $\|\hat{\Theta}_t - \Theta_*\|_2$ rather than bounds or bootstrap estimates for each algorithm, since we observe that the overall effect is negligible as mentioned in Dean et al. (2018). The following gives the implementation details of each algorithm.

(i) StabL We have σ_ν, H_0 and T_w as the varying parameters. In the implementation of optimistic parameter search we deploy projected gradient descent (PGD), which works efficiently for the small dimensional problems. The implementation follows Section G.1 of (Dean et al., 2018). Note that this approach, hence the optimistic parameter choice, can be computationally challenging for higher dimensional systems. We pick the regularizer $\lambda = 0.05$ for all adaptive control tasks.

(ii) OFULQ We deploy a slight modification on the implementation of OFULQ given by (Abbasi-Yadkori and Szepesvári, 2011). Similar to StabL, we add an additional minimum policy duration constraint to the general switching constraints of OFULQ, *i.e.*, the standard determinant doubling of V_t. This prevents too frequent changes in the beginning of the algorithm and dramatically improves the regret performance. This minimum duration H_0^{OFU} is the only varying parameter for OFULQ. For the optimistic parameter search we also implement PGD. We pick the regularizer $\lambda = 0.001$ for all adaptive control tasks.

(iii) CEC w/t fixed perturbations This algorithm is the standard baseline in control theory. In the implementation, the optimal infinite-horizon LQR controller for the estimated system is deployed and fixed isotropic perturbations $\mathcal{N}(0, \sigma_{\text{exp}}^2 I)$ are injected throughout the implementation. The isotropic perturbations are injected since it is well-known that certainty equivalent controllers can result with drastically incorrect parameter estimates (Lai et al., 1982; Becker et al., 1985; Kumar, 1990) due to lack of exploration. The policy changes happen in epochs with linear scaling, *i.e.*, each epoch i is of iH_{ep} length. This growth is observed to be preferable over the standard exponentially increasing epoch lengths adopted in theoretical analyses of the worst case regret guarantees. Thus, the varying parameters for CEC w/t fixed perturbations are σ_{exp} and H_{ep}. We pick the regularizer $\lambda = 0.5$ for all adaptive control tasks.

(iv) CEC w/t decaying perturbations The implementation of this algorithm is similar to **(iii)**. The difference is that the injected perturbations have decaying variance over epochs. We adopt the decay of $1/\sqrt{i}$ for each epoch i, *i.e.* $\sigma_{i,exp} = \sigma_{exp}^{dec}/\sqrt{i}$ for some initial σ_{exp}^{dec} such that isotropic perturbations are injected in each epoch. Based on the extensive experimental study, we deduced that this decay performs better than the decay of $i^{-1/3}$ as given in Dean et al. (2018) or $2^{-i/2}$ as given in Simchowitz and Foster (2020). The varying parameters for this algorithm are σ_{exp}^{dec} and H_{ep}^{dec} in which the latter defines the first epoch length in the linear scaling of epochs. We pick the regularizer $\lambda = 0.05$ for all adaptive control tasks.

In each experiment, the system starts from $x_0 = 0$ to reduce variance over runs. For each setting, we run 200 independent runs with the duration of 200 time steps. Note that we do not compare StabL with the adaptive control algorithms provided in (Simchowitz and Foster, 2020; Simchowitz et al., 2020; Dean et al., 2018; Cohen et al., 2019; Faradonbeh et al., 2018b, 2020) which all require a given initial stabilizing policy or stable open-loop dynamics and (Abeille and Lazaric, 2017) which is tailored for scalar systems. Moreover, Chen and Hazan (2020) deals with adversarial LQR setting and uses "significantly" large inputs to identify the model dynamics which causes orders of magnitude worse regret.

I.1 Marginally Unstable Laplacian System

The LQR problem is given as

$$A_* = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \;\; B_* = I_{3\times 3}, \;\; Q = 10I, \;\; R = I, \;\; w \sim \mathcal{N}(0, I). \tag{56}$$

This system dynamics have been studied in (Dean et al., 2018, 2019; Abbasi-Yadkori et al., 2018; Tu and Recht, 2018) and it corresponds a Laplacian system with weakly connected adjacent nodes. Notice that the inputs have less cost weight than the states. This skewed cost combined with the unstable dynamics severely hinders the design of effective strategies for OFU-based methods.

Algorithmic Setups: For StabL, we set $H_0 = 15$, $T_w = 35$ and $\sigma_\nu = 1.5$. For CEC with decaying perturbation, we set $H_{ep}^{dec} = 20$, and $\sigma_{exp}^{dec} = 2$. For CEC with fixed perturbation, we set $\sigma_{exp} = 1.3$ and $H_{ep} = 15$. For OFULQ, we set $H_0^{OFU} = 6$.

Regret After 200 Time Steps: In Table 6, we provide the regret performance of the algorithms after 200 time steps of adaptive control in the Laplacian system. As expected the regret performance of OFULQ suffers the most regret due to unstable dynamics and skewed cost, which makes it difficult to design effective policies for the OFU-based algorithms. Even though StabL uses OFU principle, it overcomes the difficulty to design effective policies via the improved exploration in the early stages and achieves the best regret performance.

Table 6: Regret After 200 Time Steps in Marginally Unstable Laplacian System

Algorithm	**Average Regret**	**Best 95%**	**Best 90%**	**Best 75%**	**Best 50%**
StabL	$\mathbf{1.55 \times 10^4}$	$\mathbf{1.42 \times 10^4}$	$\mathbf{1.32 \times 10^4}$	$\mathbf{1.12 \times 10^4}$	$\mathbf{8.89 \times 10^3}$
OFULQ	6.17×10^{10}	4.57×10^7	4.01×10^6	3.49×10^5	4.70×10^4
CEC w/t Fixed	3.72×10^{10}	2.23×10^5	2.14×10^4	1.95×10^4	1.73×10^4
CEC w/t Decay	4.63×10^4	4.27×10^4	4.03×10^4	3.51×10^4	2.84×10^4

Figure 3 gives the regret comparison between StabL and CEC with decaying isotropic perturbations which performs the second best in the given Laplacian system. Note that we did not include OFULQ and CEC w/t fixed perturbations in the figure since they perform orders of magnitude worse that StabL and CEC w/t decaying perturbations.

Maximum State Norm: In Table 7, we display the stabilization capabilities of the algorithms by providing the averages of the maximum ℓ_2 norms of the states in 200 independent runs. We also include the worst case state magnitudes which demonstrates how controlled the states are during the entire adaptive control task. The results show that StabL maintains the smallest magnitude of the state and thus, the most stable dynamics. We also verify that after the first policy change which happens after 15 time steps, the spectral radius of the closed-loop system formed via StabL is always stable, *i.e.* $\rho(A_* + B_* K(\tilde{\Theta}_t)) < 1$ for $t > 15$.

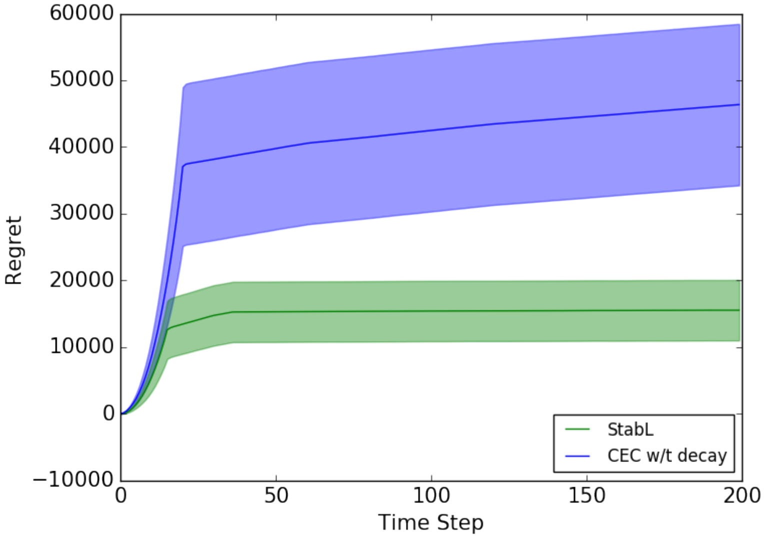

Figure 3: Regret of StabL vs CEC with decaying isotropic perturbations. The solid lines are the average regrets for 200 independent runs and the shaded regions are the half standard deviations.

Table 7: Maximum State Norm in Marginally Unstable Laplacian System

Algorithm	**Average** $\max \|x\|_2$	**Worst 5%**	**Worst 10%**	**Worst 25%**
StabL	$\mathbf{1.35 \times 10^1}$	$\mathbf{2.24 \times 10^1}$	$\mathbf{2.15 \times 10^1}$	$\mathbf{1.95 \times 10^1}$
OFULQ	9.59×10^3	1.83×10^5	9.04×10^4	3.81×10^4
CEC w/t Fixed	3.33×10^3	6.64×10^4	3.32×10^4	1.33×10^4
CEC w/t Decay	2.04×10^1	3.46×10^1	3.27×10^1	2.87×10^1

Persistence of Excitation via StabL: In order to further highlight the benefit of improved exploration strategy, we empirically study the smallest eigenvalue of the regularized design matrix V_t for StabL and OFULQ. The evolution of the $\lambda_{\min}(V_t)$ is shown for both algorithms in Figure 4. From the figure, one can see that improved exploration strategy of StabL achieves linear scaling of $\lambda_{\min}(V_t)$, *i.e.*, persistence of excitation. Thus, it finds the stabilizing neighborhood after the first epoch. On the other hand, the control inputs of OFULQ fail to excite the system uniformly, thus it cannot quickly find a stabilizing policy. This results in unstable dynamics and significantly more regret on average (Table 6).

I.2 Longitudinal Flight Control of Boeing 747

The LQR problem is given as

$$A_* = \begin{bmatrix} 0.99 & 0.03 & -0.02 & -0.32 \\ 0.01 & 0.47 & 4.7 & 0 \\ 0.02 & -0.06 & 0.4 & 0 \\ 0.01 & -0.04 & 0.72 & 0.99 \end{bmatrix}, \; B_* = \begin{bmatrix} 0.01 & 0.99 \\ -3.44 & 1.66 \\ -0.83 & 0.44 \\ -0.47 & 0.25 \end{bmatrix}, \; Q = I, \; R = I, \; w \sim \mathcal{N}(0, I). \tag{57}$$

This problem is the longitudinal flight control of Boeing 747 with linearized dynamics and introduced in (Ishihara et al., 1992). The given linear dynamical system corresponds to the dynamics for level flight of Boeing 747 at the altitude of 40000ft with the speed of 774ft/sec, for a discretization of 1 second. The first state element is the velocity of aircraft along body axis, the second is the velocity of aircraft perpendicular to body axis, the third is the angle between body axis and horizontal and the final element is the angular velocity of aircraft. The first input element is the elevator angle and the second one is the thrust. The process noise corresponds to the external wind conditions.

Notice that the dynamics are linearized around a certain point and it is important to guarantee that the linearization is valid. To this end, an RL policy should stabilize the system and keep the state small in order to

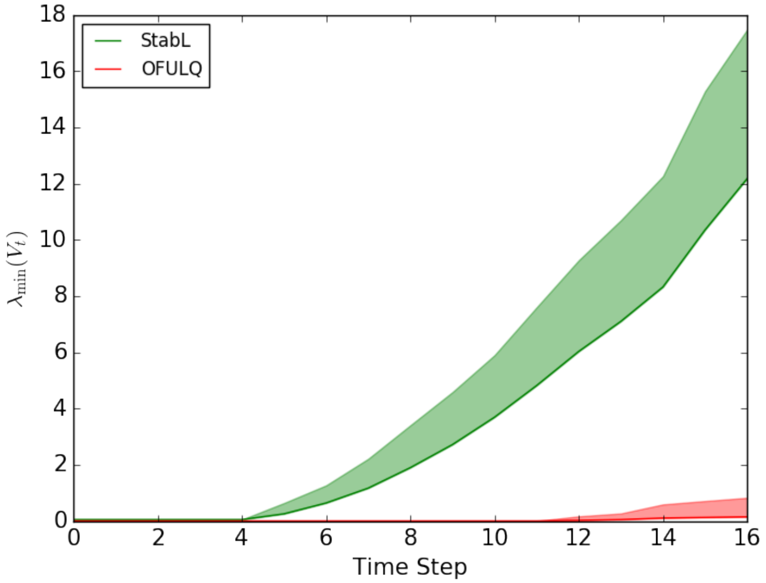

Figure 4: Scaling of the smallest eigenvalue of the design matrix for StabL and OFULQ. The solid line denotes the mean and the shaded region denotes one standard deviation. The early improved exploration strategy helps StabL achieve linear scaling in $\lambda_{\min}(V_t)$, thus persistence of excitation. The only OFU-based controllers of OFULQ fail to achieve persistence of excitation.

not lead the system to the unmodeled nonlinear dynamics.

Algorithmic Setups: For StabL, we set $H_0 = 10$, $T_w = 35$ and $\sigma_\nu = 2$. For CEC with decaying perturbation, we set $H_{ep}^{dec} = 30$, and $\sigma_{exp}^{dec} = 2$. For CEC with fixed perturbation, we set $\sigma_{exp} = 2.5$ and $H_{ep} = 25$. For OFULQ, we set $H_0^{OFU} = 7$.

Regret After 200 Time Steps: In Table 8, we give the regret performance of the algorithms after 200 time steps in Boeing 747 flight control. In terms of average regret, StabL attains half of the regret of CEC with decay and performs orders of magnitude better than OFULQ. Also, consider Figure 5. Notice that until the third policy update, OFULQ is still working towards further exploration and is not designing effective controllers to regulate the system dynamics. This is due to the higher dimensions of the Boeing 747 control system which prevents quick and effective exploration. This results in unstable system dynamics in the early stages and poorly scaling of the regret. On the other hand, the early improved exploration strategy helps StabL to maintain stable dynamics with the expense of an additional regret in the early stages compared to OFULQ. However, as it can be seen from Figure 5, this improved exploration strategy yields significantly lower regret in the later stages.

Table 8: Regret After 200 Time Steps in Boeing 747 Flight Control

Algorithm	**Average Regret**	**Top 95%**	**Top 90%**	**Top 75%**	**Top 50%**
StabL	$\mathbf{1.34 \times 10^4}$	$\mathbf{1.05 \times 10^3}$	$\mathbf{9.60 \times 10^3}$	$\mathbf{7.58 \times 10^3}$	$\mathbf{5.28 \times 10^3}$
OFULQ	1.47×10^8	4.19×10^6	9.89×10^5	5.60×10^4	8.91×10^3
CEC w/t Fixed	4.79×10^4	4.62×10^4	4.51×10^4	4.25×10^4	3.88×10^4
CEC w/t Decay	2.93×10^4	2.61×10^4	2.48×10^4	2.22×10^4	1.86×10^4

Maximum State Norm: Similar to Laplacian system, StabL controls the state well and provides the lowest average maximum norm.

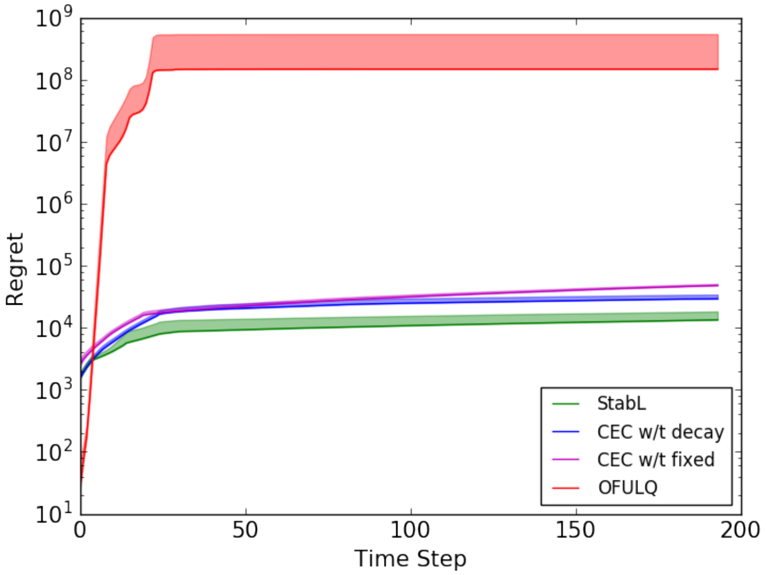

Figure 5: Regret Comparison of all algorithms in Boeing 747 flight control. The solid lines are the average regrets for 200 independent runs and the shaded regions are the quarter standard deviations.

Table 9: Maximum State Norm in Boeing 747 Control

Algorithm	**Average** $\max \|x\|_2$	**Worst 5%**	**Worst 10%**	**Worst 25%**
StabL	$\mathbf{3.38 \times 10^1}$	$\mathbf{8.02 \times 10^1}$	$\mathbf{7.01 \times 10^1}$	$\mathbf{5.23 \times 10^1}$
OFULQ	1.62×10^3	2.25×10^4	1.37×10^4	6.26×10^3
CEC w/t Fixed	4.97×10^1	7.78×10^1	7.31×10^1	6.48×10^1
CEC w/t Decay	4.60×10^1	7.96×10^1	7.25×10^1	6.31×10^1

I.3 Unmanned Aerial Vehicle (UAV) in 2-D plane

The LQR problem is given as

$$A_* = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_* = \begin{bmatrix} 0.125 & 0 \\ 0.5 & 0 \\ 0 & 0.125 \\ 0 & 0.5 \end{bmatrix}, Q = diag(1, 0.1, 2, 0.2), \quad R = I, \quad w \sim \mathcal{N}(0, I) \tag{58}$$

This problem is the linearized model of a UAV which operates in a 2-D plane (Zhao et al., 2021). Notice that it corresponds to the model of double integrator. The first and third state elements correspond to the position, whereas the second and fourth state elements are velocity components. The inputs are the acceleration. The process noise corresponds to the external wind conditions. Similar to Boeing 747, the dynamics are linearized and keeping the state vector small is critical in order to maintain the validity of the linearization.

Algorithmic Setups: For StabL, we set $H_0 = 20$, $T_w = 55$ and $\sigma_\nu = 4$. For CEC with decaying perturbation, we set $H_{ep}^{dec} = 30$, and $\sigma_{exp}^{dec} = 3.5$. For CEC with fixed perturbation, we set $\sigma_{exp} = 3$ and $H_{ep} = 35$. For OFULQ, we set $H_0^{OFU} = 7$.

Regret After 200 Time Steps: In Table 10, we give the regret performance of the algorithms after 200 time steps in UAV control control task. Once more, StabL performs significantly better than other RL methods. The evolution of the average regret is also given in Figure 6. As suggested by the theory, by paying a linear regret cost for a short period of time in the early stages, StabL guarantees stabilizing the underlying system and achieves the best regret performance.

Table 10: Regret After 200 Time Steps in UAV Control

Algorithm	**Average Regret**	**Top 95%**	**Top 90%**	**Top 75%**	**Top 50%**
StabL	$\mathbf{1.53 \times 10^5}$	$\mathbf{1.05 \times 10^5}$	$\mathbf{9.23 \times 10^4}$	$\mathbf{6.85 \times 10^4}$	$\mathbf{4.47 \times 10^4}$
OFULQ	5.06×10^7	1.75×10^6	1.03×10^6	2.46×10^5	5.82×10^4
CEC w/t Fixed	4.52×10^5	3.80×10^5	3.35×10^5	2.50×10^5	1.64×10^5
CEC w/t Decay	3.24×10^5	2.70×10^5	2.37×10^5	1.75×10^5	1.03×10^5

Maximum State Norm:

Table 11: Maximum State Norm in UAV Control

Algorithm	**Average** $\max \|x\|_2$	**Worst 5%**	**Worst 10%**	**Worst 25%**
StabL	$\mathbf{8.46 \times 10^1}$	$\mathbf{2.51 \times 10^2}$	$\mathbf{2.00 \times 10^2}$	$\mathbf{1.50 \times 10^2}$
OFULQ	5.61×10^2	6.35×10^3	3.78×10^3	1.90×10^3
CEC w/t Fixed	1.45×10^2	3.12×10^2	2.91×10^2	2.42×10^2
CEC w/t Decay	1.26×10^2	2.71×10^2	2.48×10^2	2.12×10^2

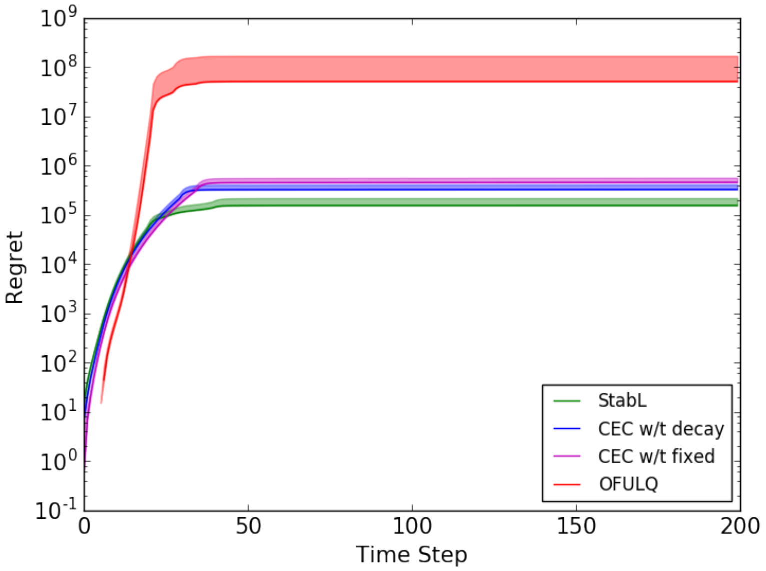

Figure 6: Regret Comparison of all algorithms in UAV control task. The solid lines are the average regrets for 200 independent runs and the shaded regions are the quarter standard deviations.

I.4 Stabilizable but Not Controllable System

The LQR problem is given as

$$A_* = \begin{bmatrix} -2 & 0 & 1.1 \\ 1.5 & 0.9 & 1.3 \\ 0 & 0 & 0.5 \end{bmatrix}, \quad B_* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, Q = I, \quad R = I, \quad w \sim \mathcal{N}(0, I) \tag{59}$$

This problem is particularly challenging in terms of system identification and controller design since the system is not controllable but stabilizable. As expected besides StabL which is tailored for the general stabilizable setting, other algorithms perform poorly. In fact, CEC with fixed noise significantly blows up due to significantly unstable dynamics for the controllable part of the system. Therefore, we only present the remaining three algorithms.

Algorithmic Setups: For StabL, we set $H_0 = 8$, $T_w = 20$ and $\sigma_\nu = 2.5$. For CEC with decaying perturbation, we set $H_{ep}^{dec} = 30$, and $\sigma_{exp}^{dec} = 3$. For OFULQ, we set $H_0^{OFU} = 6$.

Regret After 200 Time Steps: Table 12 provides the regret of the algorithms after 200 time steps. This setting is where OFULQ fails dramatically due to not being tailored for the stabilizable systems. Compared to CEC with decaying perturbation, StabL also provides an order of magnitude improvement (Figure 7)

Table 12: Regret After 200 Time Steps in Stabilizable but Not Controllable System (59)

Algorithm	Average Regret	Top 95%	Top 90%	Top 75%	Top 50%
StabL	$\mathbf{1.68 \times 10^6}$	$\mathbf{9.56 \times 10^5}$	$\mathbf{7.21 \times 10^5}$	$\mathbf{3.72 \times 10^5}$	$\mathbf{1.29 \times 10^5}$
OFULQ	5.20×10^{12}	1.74×10^{12}	8.27×10^{11}	2.13×10^{11}	4.51×10^{10}
CEC w/t Decay	1.56×10^7	1.17×10^7	9.75×10^6	5.96×10^6	2.33×10^6

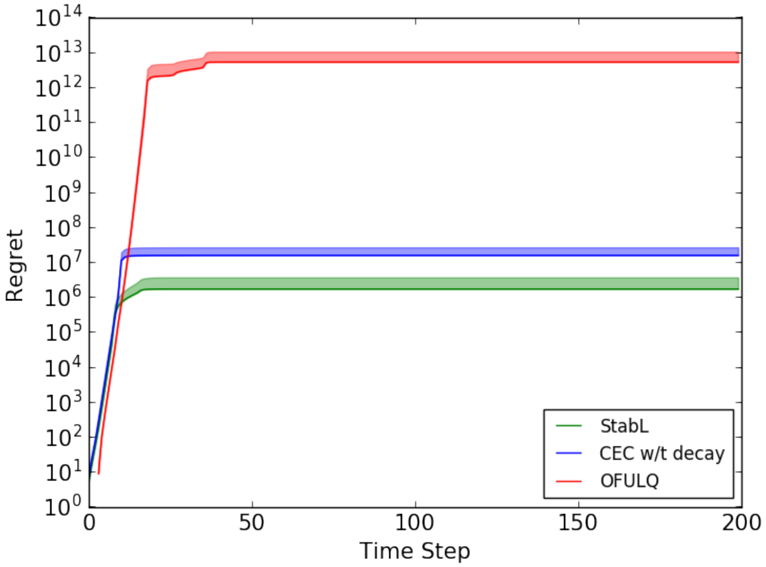

Figure 7: Regret Comparison of three algorithms in controlling (59). The solid lines are the average regrets for 200 independent runs and the shaded regions are the quarter standard deviations.

Maximum State Norm:

Table 13: Maximum State Norm in the Control of Stabilizable but Not Controllable System (59)

Algorithm	Average $\max \|x\|_2$	Worst 5%	Worst 10%	Worst 25%
StabL	$\mathbf{3.02 \times 10^2}$	$\mathbf{1.04 \times 10^3}$	$\mathbf{8.88 \times 10^2}$	$\mathbf{6.68 \times 10^2}$
OFULQ	4.39×10^5	3.10×10^6	2.40×10^6	1.39×10^6
CEC w/t Decay	1.37×10^3	4.07×10^3	3.54×10^3	2.78×10^3