

Learning Stabilizing Policies via an Unstable Subspace Representation

Leonardo F. Toso^{*1}, Lintao Ye², and James Anderson¹

¹Department of Electrical Engineering, Columbia University

²School of Artificial Intelligence and Automation, HUST

Abstract

We study the problem of learning to stabilize (LTS) a linear time-invariant (LTI) system. Policy gradient (PG) methods for control assume access to an initial stabilizing policy. However, designing such a policy for an *unknown* system is one of the most fundamental problems in control, and it may be as hard as learning the optimal policy itself. Existing work on the LTS problem requires large data as it scales quadratically with the ambient dimension. We propose a two-phase approach that first learns the *left unstable subspace* of the system and then solves a series of discounted linear quadratic regulator (LQR) problems on the learned unstable subspace, targeting to stabilize only the system’s unstable dynamics and reduce the effective dimension of the control space. We provide non-asymptotic guarantees for both phases and demonstrate that operating on the unstable subspace reduces sample complexity. In particular, when the number of unstable modes is much smaller than the state dimension, our analysis reveals that LTS on the unstable subspace substantially speeds up the stabilization process. Numerical experiments are provided to support this sample complexity reduction achieved by our approach.

1 Introduction

In contrast to traditional model-based control methods, model-free, policy gradient (PG) approaches offer two substantial advantages: (i) they are simple to implement without requiring knowledge of the underlying system dynamics, and (ii) they adapt readily to new tasks with minimal parameter tuning. These methods have been widely used to solve reinforcement learning (RL) tasks in unknown environments [Sutton et al., 1999], with recent work establishing strong optimality guarantees [Agarwal et al., 2021]. As a result, there has been much interest in applying PG methods to optimal control, see the excellent review by Hu et al. [2023] for an overview. Problems of particular relevance to this work includes the linear quadratic regulator (LQR) problem in the offline setting [Fazel et al., 2018, Malik et al., 2019, Gravell et al., 2020, Mohammadi et al., 2021], online setting [Cassel and Koren, 2021], multi-task setting [Wang et al., 2023, Toso et al., 2024a,b, Zhan et al., 2025], and networked setting [Mitra et al., 2024]. A crucial milestone was achieved in Fazel et al. [2018], which showed that the LQR problem exhibits a benign optimization landscape, enabling global convergence of PG methods (with linear rate as shown in Mohammadi et al. [2020]).

^{*}Email addresses: {lt2879,james.anderson}@columbia.edu, yelintao93@hust.edu.cn.

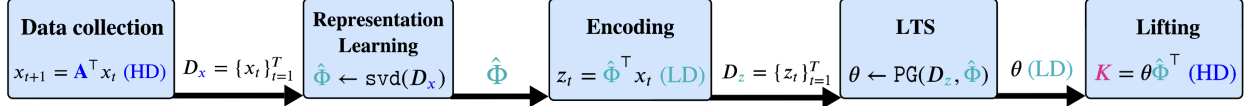


Figure 1: Workflow for learning to stabilize (LTS) a high-dimensional (HD) discrete-time LTI system on its low-dimensional (LD) unstable subspace.

There is however a major obstacle encountered when applying PG methods to control: it is typically assumed that one has access to an initial stabilizing policy. For one of the most fundamental problems in control, that of finding a stabilizing policy for an unknown system, such an assumption precludes the use of PG methods. In particular, learning to stabilize (LTS) a linear system can be as hard as learning the optimal policy itself [Tsiamis et al., 2022, Zeng et al., 2023].

Several solutions to the LTS problem have been proposed, c.f., [Lale et al., 2020, Lamperski, 2020, Chen and Hazan, 2021, Perdomo et al., 2021, Hu et al., 2022, Zhao et al., 2024]. Two notable existing approaches that this work builds on are: discounted methods [Lamperski, 2020, Perdomo et al., 2021, Zhao et al., 2024] and unstable subspace learning [Hu et al., 2022, Zhang et al., 2024a, Werner and Peherstorfer, 2025]. In the first, PG solves discounted LQR problems with a carefully selected sequence of increasing discount factors. Since the policy gradients are estimated from data (i.e., system trajectories), this approach typically suffers from a high sample complexity as it scales quadratically with the ambient problem dimension [Zhao et al., 2024].

On the other hand, since a stabilizing policy only needs to address the system’s unstable dynamics, focusing on stabilizing just the unstable modes reduces the effective dimensionality of the control space and consequently, the sample complexity, as shown in Hu et al. [2022] for the noiseless setting and in Zhang et al. [2024a] for the stochastic setting. However, these works rely on identifying the full unstable dynamics to construct a stabilizing policy on top of the identified model, therefore being model-based and highly sensitive to the model’s estimation accuracy. Furthermore, the analyses in Hu et al. [2022], Zhang et al. [2024a] are restricted to diagonalizable systems.

In contrast, our approach accommodates non-diagonalizable system and combines the strengths of both perspectives: we avoid explicitly identifying the system’s unstable dynamics while using policy gradient to stabilize only the unstable modes. In particular, we solve a sequence of discounted LQR problems by performing policy gradient updates on the left unstable subspace of the system (see Figure 1). Moreover, our work addresses the following questions:

- *To what extent can we guarantee the stability of a high-dimensional system by performing a discount-factor annealing method on its low-dimensional unstable subspace?*
- *How does this approach reduce the sample complexity of learning a stabilizing controller?*
- *What is the sample complexity of estimating the representation of the left unstable subspace?*

1.1 Contributions

- **Sample complexity reduction:** By operating on the unstable subspace, namely, the subspace associated with the system’s $\ell \in \mathbb{N}$ unstable modes, we aim to stabilize only the portion of the system that requires stabilization rather than the full d_X -dimensional state space. We demonstrate the reduction in the sample complexity of finding a stabilizing policy from $\tilde{O}(d_X^2 d_U)$ [Zhao et al., 2024] to $\tilde{O}(\ell^2 d_U)$ (Theorem 5.1), with d_U being the number of inputs, which is significant when the

number of unstable modes is much smaller than the state dimension, i.e., $\ell \ll d_X$.

- **Learning the left unstable subspace:** We demonstrate that operating on the *left* unstable subspace allows for controlling the closed-loop spectral radius in terms of the accuracy of the learned representation. We also provide finite-sample guarantees for learning this representation by sampling data from an adjoint system (Theorem 3.1). Therefore, the closed-loop spectral radius error decreases as more data is collected. This contrasts with prior work of Hu et al. [2022], Zhang et al. [2024a], which recovers a basis of the right unstable subspace. Their error bounds depend on a “coupling term” that arises from decomposing the system’s dynamics into stable and unstable components and inevitably incurs a bias that is significant for non-symmetric system matrices.

- **Non-diagonalizable matrices:** Our results accommodate non-diagonalizable systems. In contrast to Hu et al. [2022], Zhang et al. [2024a], which restrict the analysis to diagonalizable systems, we leverage the Jordan form decomposition and establish that the left unstable subspace representation can be learned with a finite amount of samples (Lemma 3.2 and Theorem 3.1). That is in contrast to Zhang et al. [2024a], where the sample complexity scales inversely with the spectral gap between the unstable modes; this dependence is problematic when the system is non-diagonalizable, as the gap goes to zero and data grows prohibitively large. We prove that it should not be the case.

1.2 Related Work

- **Learning to stabilize with identified model:** A natural idea to find a stabilizing controller for an unknown system is first to identify the system’s model from data and then synthesize a controller on top of it. Chen and Hazan [2021] show that the sample complexity scales exponentially with d_X when learning to stabilize from a single trajectory. However, such scaling is undesirable when d_X is large. To overcome this, Hu et al. [2022] demonstrate that a stabilizing policy can be learned by only identifying the unstable modes of the system, which leads to a much more benign sample complexity that scales with the number of unstable modes $\ell \ll d_X$. In contrast to Hu et al. [2022], this work does not require identifying the unstable dynamics; namely, we identify a basis (or representation) of the left unstable subspace of the system. Moreover, our approach accommodates non-diagonalizable matrices, which is not the case in Hu et al. [2022], Zhang et al. [2024a].

- **Learning to stabilize with policy gradient:** An alternative approach is to learn a stabilizing controller *without* performing system identification. Recent work Lamperski [2020], Perdomo et al. [2021] showed that a reformulation of the LQR problem that involves introducing an additional degree of freedom—a “damping factor”, $\gamma \in (0, 1]$, leads to an intuitive, iterative approach for constructing a stabilizing policy. Initially, setting γ sufficiently small, the trivial zero policy stabilizes the underlying damped system. PG methods solve the damped LQR problem and produce an initial stabilizing policy for the subsequent discounted LQR problem. Once a stabilizing controller is obtained, γ is incrementally increased, and the process is repeated as γ goes to one.

Zhao et al. [2024] provide an explicit update rule for γ , which allows for characterizing the sample complexity of LTS with discounted PG. In particular, it scales as $\mathcal{O}(d_X^2 d_U)$ and becomes prohibitively large for high-dimensional systems where d_X is large. In this work, we only focus on stabilizing the system’s unstable modes which reduces the sample complexity to $\mathcal{O}(\ell^2 d_U)$. We emphasize that Werner and Peherstorfer [2025] consider policy optimization on the unstable subspace to learn a stabilizing policy; however, they do not provide finite-sample guarantees for either the unstable subspace representation learning or the resulting stabilizing policy.

- **Representation learning for control:** We also stress the difference between the control policy

representation considered in this work and the low-rank representation of the system model in Zhang et al. [2024b], Lee et al. [2024]. The low-rank representation of the system model captures the important features to be identified and potentially shared across multiple systems, enabling sample-efficient estimation [Zhang et al., 2024b] and certainty-equivalent control [Lee et al., 2024]. We focus on a policy representation that captures the modes to stabilize (i.e., the unstable modes). In particular, it carries a physical interpretation as it spans the system’s left unstable subspace.

1.3 Notation

We use $\rho(\cdot)$ and $\sigma_{\min}(\cdot)$ to denote the spectral radius and the minimum singular value of a matrix, respectively. $\|\cdot\|$ is the ℓ_2 norm, $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm [Vershynin, 2018], and $\|\cdot\|_F$ is the Frobenius norm of a matrix. $\text{Tr}(\cdot)$ is the trace function. \mathbb{S}^{d-1} denotes the unit sphere in d -dimensional space. $\kappa(A)$ denotes the condition number of the matrix with the eigenvectors of A as columns. We use $\text{col}(A)$ to denote the subspace spanned by the columns of A . We use the big-O notation $\mathcal{O}(\cdot)$ to omit constants and $\tilde{\mathcal{O}}(\cdot)$ to omit logarithmic factors in the argument. Unless otherwise stated, expectation is always taken with respect to the initial state.

2 Problem Formulation

Consider the discrete-time linear time-invariant (LTI) system

$$x_{t+1} = Ax_t + Bu_t, \text{ for } t = 0, 1, 2, \dots, \quad (1)$$

where $x_t \in \mathbb{R}^{d_x}$ denotes the state and $u_t \in \mathbb{R}^{d_u}$ the control input. We assume that the initial state x_0 is drawn according to a zero mean and isotropic distribution, i.e., $\mathbf{E}[x_0] = 0$, $\mathbf{E}[x_0 x_0^\top] = I$, with $\|x_0\| \leq \mu_0$ and $\|x_0\|_{\psi_2} \leq \mu_\psi$. Let $\{\lambda_1, \lambda_2, \dots, \lambda_{d_x}\}$, with $|\lambda_1| \geq \dots \geq |\lambda_\ell| > 1 > |\lambda_{\ell+1}| \geq \dots \geq |\lambda_{d_x}|$, denote the eigenvalues of the drift matrix A . We focus on the setting where the system matrix A is open-loop unstable, with $\ell \ll d_x$ unstable modes $\{\lambda_1, \dots, \lambda_\ell\}$. We assume that (1) is stabilizable, which ensures the existence of a state feedback gain $K \in \mathbb{R}^{d_u \times d_x}$ such that $\rho(A + BK) < 1$.

Goal: Construct a stabilizing controller K that defines a linear policy of the form $u_t = Kx_t$, using policy gradient methods [Fazel et al., 2018], without requiring access to the system matrices (A, B) .

2.1 Discounted Linear Quadratic Regulator Problem

Given a “discount factor” $\gamma \in (0, 1]$, the discounted LQR problem is described as follows:

$$\min_{K \in \mathcal{K}} \left\{ J^\gamma(K) := \mathbf{E} \left[\sum_{t=0}^{\infty} \gamma^t x_t^\top (Q + K^\top R K) x_t \right] \right\}, \text{ subject to (1) with } u_t = Kx_t, \quad (2)$$

where $\mathcal{K} := \{K \mid \rho(A + BK) < 1\}$ denotes the set of stabilizing controllers, and (Q, R) are positive definite matrices. It is important to emphasize that in our problem setup, the cost matrices (Q, R) are “artificial” design parameters that will be used in the implementation of our solution method. Our goal *is not* to learn an optimal control policy with respect to a specific cost, but rather to learn a controller that ensures the stability of (1). By rescaling the state x_t by $\gamma^{t/2}$, i.e., $\tilde{x}_t = \gamma^{t/2} x_t$, the discounted LQR problem (2) is equivalent to

$$\min_{K \in \mathcal{K}^\gamma} \left\{ J^\gamma(K) := \mathbf{E} \left[\sum_{t=0}^{\infty} \tilde{x}_t^\top (Q + K^\top R K) \tilde{x}_t \right] \right\}, \text{ subject to } \tilde{x}_{t+1} = (A^\gamma + B^\gamma K) \tilde{x}_t, \quad (3)$$

where $\mathcal{K}^\gamma := \{K \mid \rho(A^\gamma + B^\gamma K) < 1\}$, with damped system matrices $A^\gamma := \sqrt{\gamma}A$, $B^\gamma := \sqrt{\gamma}B$.

Note that by setting γ sufficiently small, in particular, $\gamma < 1/\rho^2(A)$, the trivial controller $K \equiv 0$ stabilizes the underlying discounted LQR problem. However, such a control gain may not be stabilizing for the original system (i.e., for $\gamma = 1$). In fact, what allows us to design a stabilizing controller by solving a sequence of discounted LQR problems is the appropriate incremental update of γ . Let γ_j denote the discount factor at iteration $j \in \mathbb{N}$. Zhao et al. [2024] showed that by repeating the following process (while $\gamma_{j+1} < 1$):

1. Compute a controller K_{j+1} by solving (2) such that $J^{\gamma_j}(K_{j+1}) \leq \bar{J}$,
2. Update the discount factor: $\gamma_{j+1} = (1 + \xi\alpha_j)\gamma_j$,

a stabilizing controller $K \in \mathcal{K}$ is found within a finite number of iterations of the above process. Here, $\xi \in (0, 1)$ is the decay factor, \bar{J} is a uniform bound of the discount LQR cost, and $\alpha_j > 0$ is the discount factor update rate. We elaborate on the role and selection of each of these quantities in Section 4, where we introduce our method for learning a stabilizing controller on the unstable subspace. For now, it is important to highlight that such explicit discount method comes with a sample complexity that scales quadratically with the system's state dimension, i.e., $\tilde{O}(d_X^2 d_U)$, thus limiting its applicability for high-dimensional systems where data collection is difficult and thus data is scarce (e.g., robot manipulation [Billard and Kragic, 2019]).

However, high-dimensional unstable systems often possess only a small number of unstable modes, as in our setting of interest $\ell \ll d_X$. That observation motivates the following question: *Can we apply the discount method directly on the unstable subspace, aiming to stabilize only the small portion of the state space associated with the unstable dynamics?* We answer this question in the affirmative. For this purpose, we introduce a linear parameterization of K for stabilizing the unstable modes of (1) independently from its stable dynamics.

2.2 Stabilizing Only the Unstable Modes

Let $\Omega := [\Phi \ \Phi_\perp]$ be an orthonormal basis of \mathbb{R}^{d_X} , where the columns of $\Phi \in \mathbb{R}^{d_X \times \ell}$ span the *left* eigenspace corresponding to the unstable modes of A . We refer to this as the “left unstable subspace of A ”, and to Φ as the “unstable subspace representation”. Hence, we can write the following:

$$\Omega^\top A \Omega = \begin{bmatrix} A_u & \\ \Delta & A_s \end{bmatrix}, \text{ with } A_u = \Phi^\top A \Phi, \Delta = \Phi_\perp^\top A \Phi, \text{ and } A_s = \Phi_\perp^\top A \Phi_\perp,$$

where A_u represents the unstable dynamics of A , as it has the spectrum of the Jordan blocks corresponding to the unstable eigenvalues of A . On the other hand, A_s inherits all stable modes of A . The matrix Δ represents the “coupling” of the stable and unstable dynamics arising from the $\text{col}(\Phi) \oplus \text{col}(\Phi_\perp)$ decomposition. We also note that $\Delta \equiv 0$ when A is symmetric.

Controller representation: Suppose that K is linearly decomposed into a low-dimensional control gain $\theta \in \mathbb{R}^{d_U \times \ell}$ and the left unstable subspace representation Φ , namely, $K = \theta \Phi^\top$. The closed-loop system matrix $A + BK$ can then be written as

$$A + BK = \Omega \begin{bmatrix} A_u + B_u \theta & \\ \Delta + B_s \theta & A_s \end{bmatrix} \Omega^\top := \Omega \bar{A} \Omega^\top, \text{ where } B_u = \Phi^\top B \text{ and } B_s = \Phi_\perp^\top B.$$

From the above decomposition, it suffices to stabilize the low-dimensional unstable dynamics described by (A_u, B_u) to guarantee the stability of (A, B) . Hence, one may reduce the problem of

stabilizing (A, B) through designing K , to that of stabilizing (A_u, B_u) by finding a low-dimensional controller θ such that $\rho(A_u + B_u\theta) < 1$. Intuitively, the reduction in the control space should also yield a reduction in the sample complexity of learning the stabilizing controller.

Remark 2.1. *One might naturally ask: “Why not decompose K with respect to the right unstable subspace of A instead?” We emphasize that doing so introduces the coupling term Δ in the top-right block of the decomposition of A , as it appears in [Hu et al. \[2022\]](#), [Zhang et al. \[2024a\]](#). This disrupts the triangular structure of \bar{A} and thus Δ incurs a bias in the spectral radius of the closed-loop system matrix. As a result, the condition of stabilizing (A, B) via the stabilization of (A_u, B_u) would only be guaranteed if $\|\Delta\|$ is sufficiently small. Therefore, if $\|\Delta\|$ is large, its inevitable effect in $\rho(\bar{A})$ due to the right unstable subspace parameterization would lead to an inflation in the sample complexity or it may even prevent us from stabilizing the (1), as seen in [Hu et al. \[2022\]](#), [Zhang et al. \[2024a\]](#). That is not the case in this work since we operate with the left unstable subspace.*

2.3 Low-Dimensional Discounted LQR Problem

Given the left unstable subspace representation Φ , let $z_t \in \mathbb{R}^\ell$ denote the low-dimensional state that represents x_t on the subspace spanned by the columns of Φ , i.e., $x_t = \Phi z_t$. The low-dimensional unstable dynamics of (1) evolve according to the system

$$z_{t+1} = A_u z_t + B_u u_t, \quad \text{for } t = 0, 1, 2, \dots, \quad (4)$$

where z_0 is also drawn from a zero mean and isotropic distribution since Φ is orthonormal. We can now write the discounted LQR problem on the unstable subspace in the form of (3) as follows:

$$\min_{\theta \in \Theta^\gamma} \left\{ J^\gamma(\theta, \Phi) := \mathbf{E} \left[\sum_{t=0}^{\infty} \gamma^t z_t^\top (\Phi^\top Q \Phi + \theta^\top R \theta) z_t \right] \right\}, \quad \text{subject to } z_{t+1} = (A_u^\gamma + B_u^\gamma \theta) z_t, \quad (5)$$

where $\Theta^\gamma := \{\theta \mid \sqrt{\gamma} \rho(A_u^\gamma + B_u^\gamma \theta) < 1\}$ is the set of stabilizing controllers for the damped unstable dynamics $A_u^\gamma := \sqrt{\gamma} A_u$ and $B_u^\gamma := \sqrt{\gamma} B_u$. Let $\nabla J^\gamma(\theta, \Phi)$ be the gradient with respect to θ , then

$$\nabla J^\gamma(\theta, \Phi) = \nabla J^\gamma(\theta \Phi^\top) \Phi = 2E_\theta \Sigma_\theta,$$

with

$$E_\theta := (R + B_u^{\gamma\top} P_\theta^\gamma B_u^\gamma) \theta + B_u^{\gamma\top} P_\theta^\gamma A_u^\gamma, \quad \text{where } P_\theta^\gamma = \Phi^\top Q \Phi + \theta^\top R \theta + (A_u^\gamma + B_u^\gamma \theta)^\top P_\theta^\gamma (A_u^\gamma + B_u^\gamma \theta),$$

and closed-loop state covariance $\Sigma_\theta := \mathbf{E} [\sum_{t=0}^{\infty} \gamma^t z_t z_t^\top]$. With a slight abuse of notation, we write $J^\gamma(\theta) := J^\gamma(\theta, \Phi)$ and note that the discounted LQR cost can be written as $J^\gamma(\theta) = \text{Tr}(P_\theta^\gamma)$.

Definition 2.1. *Given a discount factor $\gamma \in (0, 1]$ and scalar $\mu_s > 0$. Let $\mathcal{S}_\theta^\gamma$ denote a sublevel set of Θ^γ , $\mathcal{S}_\theta^\gamma \subseteq \Theta^\gamma$, with $\mathcal{S}_\theta^\gamma := \{\theta \mid J^\gamma(\theta) - J^\gamma(\theta^*) \leq \mu_s (J^\gamma(\theta_0) - J^\gamma(\theta^*))\}$, where θ^* is the optimal controller of the underlying low-dimensional discounted LQR problem (5).*

Similarly, \mathcal{S}_K^γ denotes the sublevel set of \mathcal{K}^γ for the high-dimensional LQR problem (2). We use J_\star^γ to denote the optimal cost. Let $\phi, \nu_\theta, L_\theta, L_K$ and μ_{PL} be positive constants. The following properties of $J^\gamma(\theta)$ and $J^\gamma(K)$ hold in their respective stabilizing sublevel sets, $\mathcal{S}_\theta^\gamma$ and \mathcal{S}_K^γ .

Lemma 2.1. *Given high-dimensional and low-dimensional stabilizing controllers $K, K' \in \mathcal{S}_K^\gamma$ and $\theta, \theta' \in \mathcal{S}_\theta^\gamma$, respectively. It holds that $\|\nabla J^\gamma(K)\| \leq \phi$, $\|\theta\| \leq \nu_\theta$, and*

$$\|\nabla J^\gamma(\theta) - \nabla J^\gamma(\theta')\|_F \leq L_\theta \|\theta - \theta'\|_F, \quad \|\nabla J^\gamma(K) - \nabla J^\gamma(K')\|_F \leq L_K \|K - K'\|_F.$$

Lemma 2.2. *Given a stabilizing controller $\theta \in \mathcal{S}_\theta^\gamma$. It holds that $\|\nabla J^\gamma(\theta)\|_F^2 \geq \mu_{PL}(J^\gamma(\theta) - J^\gamma(\theta_\star^\gamma))$.*

Remark 2.2. *Lemmas 2.1 and 2.2 were originally proved by Fazel et al. [2018] and subsequently revisited by Gravel et al. [2020], where the explicit expression of the problem dependent constants ϕ , and ν_θ are provided. We define here ϕ , ν_θ , L_θ , L_K , and μ_{PL} as the uniform bound over the set of all stabilizing controllers, i.e., either $\mathcal{S}_\theta^\gamma$ or \mathcal{S}_K^γ , for any $\gamma \in (0, 1)$.*

We conclude this section by recalling that our setting is model-free, and therefore the left unstable subspace representation Φ cannot be accessed directly. In the following section, we show that an accurate estimate of Φ , denoted by $\hat{\Phi}$, can be recovered when a sufficient amount of trajectory data is collected. The accuracy of this estimate is quantified using the subspace distance between the column spaces of $\hat{\Phi}$ and Φ , as defined in Stewart and Sun [1990].

Definition 2.2. *Let $\hat{\Pi} = \hat{\Phi}\hat{\Phi}^\top$ and $\Pi = \Phi\Phi^\top$ be orthogonal projectors onto the column spaces of $\hat{\Phi}$ and Φ , respectively. The subspace distance between Φ and $\hat{\Phi}$ is $d(\hat{\Phi}, \Phi) \triangleq \|\hat{\Phi}^\top \Phi_\perp\| = \|\hat{\Pi} - \Pi\|$.*

3 Learning the Left Unstable Representation

Sampling from the adjoint system: To learn an estimate of the left unstable subspace representation, we proceed by first collecting data from the autonomous adjoint system of (1), i.e., $x_{t+1} = A^\top x_t$ [Kouba and Bernstein, 2020]. To do so, we perform element-wise computations with the adjoint operator while forward simulating (1) accordingly. Note that for any real-valued matrix $A \in \mathbb{R}^{d_x \times d_x}$ and vectors $x, y \in \mathbb{R}^{d_x}$, we have $\langle Ax, y \rangle = \langle x, A^\top y \rangle$. Therefore, by playing (1) with zero input $u_0 \equiv 0$ and initial condition $x_0 = e_i$, where $\{e_i\}_{i=1}^{d_x}$ is the canonical basis of \mathbb{R}^{d_x} , we collect and store $e_i^+ = Ae_i$ to obtain

$$(A^\top x)_i := \langle e_i, A^\top x \rangle = \langle e_i^+, x \rangle, \forall i \in \{1, 2, \dots, d_x\},$$

which implies that the next adjoint state is $x_{t+1} = [x_t^\top e_1^+ \dots x_t^\top e_{d_x}^+]^\top$. Hence, the next adjoint state is derived from the previous state x_t and samples $\{e_i^+\}_{i=1}^{d_x}$ collected by interacting with (1).

Goal: Construct an estimation for the left unstable subspace of A from T data samples collected from the autonomous adjoint system $D = [x_1, x_2, \dots, x_T] \in \mathbb{R}^{d_x \times T}$.

Estimating the left unstable subspace: We proceed by computing the singular value decomposition $D = U\Sigma V^\top$. An estimation of the orthonormal basis of the right unstable subspace of A^\top (or left unstable subspace of A) is obtained from the range of the top ℓ columns of U , i.e., $\hat{\Phi} = [u_1, \dots, u_\ell]$. We now show that $d(\hat{\Phi}, \Phi)$ becomes sufficiently small as the trajectory length T increases. To establish this result, we leverage a similar approach to [Zhang et al., 2024a, Theorem 5.1], with two key distinctions: our setting accommodates *non-diagonalizable* system matrices A , and our estimation focuses on the *left unstable subspace representation*.

Let $\Psi \in \mathbb{R}^{d_x \times (d_x - \ell)}$ denote an orthonormal basis for the left stable subspace of A , and define $\Xi = [\Phi \ \Psi]$, which contains the left eigenvectors corresponding to the unstable and stable modes of

A . These may include generalized eigenvectors, accounting for A to be non-diagonalizable. Hence, there exists matrices $\Lambda_u \in \mathbb{R}^{\ell \times \ell}$ and $\Lambda_s \in \mathbb{R}^{(d_X - \ell) \times (d_X - \ell)}$ with the same spectrum of the Jordan blocks corresponding to the unstable and stable modes of A , respectively. As a result, we can write

$$A^\top [\Phi \ \Psi] = [\Phi \ \Psi] \begin{bmatrix} \Lambda_u & \\ & \Lambda_s \end{bmatrix}, \text{ and define } \Xi^{-1} := S = [S_1^\top \ S_2^\top]^\top \text{ to obtain}$$

$$D = \Xi S D = [\Phi \ \Psi] \begin{bmatrix} S_1 D \\ S_2 D \end{bmatrix} = \Phi D_1 + \Psi D_2 = D_u + D_s,$$

where $D_1 = S_1 D$ and $D_2 = S_2 D$. We note that $D = D_u + D_s$ is composed of $D_u = \Phi D_1$ that comes from the unstable dynamics of A and $D_s = \Psi D_2$ that depends on the stable counterpart.

Let us first analyze D_u by using the singular value decomposition of D_1 , i.e., $D_u = \Phi D_1 = \Phi U_1 \Sigma_1 V_1^\top$, with $U_1 \in \mathbb{R}^{\ell \times \ell}$, $\Sigma_1 \in \mathbb{R}^{\ell \times \ell}$, and $V_1 \in \mathbb{R}^{T \times d_X}$. Note that $\hat{\Pi}$ is the projector onto the subspace spanned by the top ℓ columns of U , whereas Π projects onto the subspace spanned by the columns of ΦU_1 . The following lemma characterizes the distance between these subspaces.

Lemma 3.1. *Let σ_ℓ be the ℓ -th singular value of D_u and $\hat{\sigma}_{\ell+1}$ the $\ell + 1$ -th singular value of D . Then,*

$$d(\hat{\Phi}, \Phi) \leq \frac{\sqrt{2\ell}\sqrt{T}(d_X - \ell)\mu_0}{(\sigma_\ell - \hat{\sigma}_{\ell+1})(1 - |\lambda_{\ell+1}|)},$$

where $d(\cdot)$ is the subspace distance as defined in Definition 2.2.

The proof follows directly from Davis-Kahan theorem [Davis and Kahan, 1970] along with the following upper bound on $\|D_2\|$:

$$\hat{\sigma}_{\ell+1} \leq \|D_2\| \leq \sqrt{T} \sum_{i=\ell+1}^{d_X} \sum_{t=1}^T |\lambda_i|^t \|x_0\| \leq \frac{\sqrt{T}(d_X - \ell)\mu_0}{1 - |\lambda_{\ell+1}|}.$$

We refer the reader to Appendix G for more details. It remains to characterize the scaling of σ_ℓ with respect to the trajectory length T .

Lemma 3.2. *Suppose that the number of samples collected from the adjoint system scales according to $T = \mathcal{O}(\log(\ell^7/\delta_\sigma^3)/\log(|\lambda_\ell|))$ for some $\delta_\sigma \in (0, 1)$. Then, it holds that*

$$\sigma_\ell \geq \frac{\sqrt{C_\sigma} |\lambda_\ell|^T \delta_\sigma}{2\sqrt{2} C_\psi \ell^{5/2} T^{3/2}}, \text{ with probability } 1 - 4\delta_\sigma, \text{ where } C_\sigma = \mathcal{O}(1) \text{ and } C_\psi = \mathcal{O}(1).$$

We detail the proof in Appendix G. For now, it is important to note that if $|\lambda_\ell| \gg 1$, then as $T \rightarrow \infty$, the subspace distance $d(\hat{\Phi}, \Phi) = \frac{\mathcal{O}(T^2)}{\mathcal{O}(|\lambda_\ell|^T) - \mathcal{O}(T^2)}$ goes to zero, with high probability. Below, we formalize the non-asymptotic guarantees of learning the left unstable subspace representation.

Theorem 3.1. *Suppose that the amount of trajectory data for learning the left unstable subspace representation scales according to $T = \mathcal{O}\left(\log\left(\frac{\ell^7(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\varepsilon\delta_\sigma^3}\right) / \log(|\lambda_\ell|)\right)$, for some small accuracy $\varepsilon > 0$ and $\delta_\sigma \in (0, 1)$. Then, it holds that $d(\hat{\Phi}, \Phi) \leq \varepsilon$, with probability $1 - 4\delta_\sigma$.*

Let us now take a moment to explain this result. First, observe that the required number of samples T depends only *logarithmically* on the problem ambient dimension d_X and the number of unstable modes ℓ . The main bottleneck in learning the left unstable subspace arises when the least unstable mode is close to marginal stability, i.e., $|\lambda_\ell| \approx 1$. Conversely, Theorem 3.1 states that the estimation becomes easier as the system becomes more explosive (i.e., $|\lambda_\ell| \gg 1$).

In addition, while the constant C_σ does not scale with ℓ or T , it is sensitive to the spectral properties of the system. In particular, it depends on the spectral norm of the Jordan matrix $\Lambda = \text{blkdiag}(\Lambda_1, \dots, \Lambda_n)$, with n being the number of distinct eigenvalues of A . Each Jordan block takes the form $\Lambda_i = \text{diag}(\lambda_i, \dots, \lambda_i) + N_i$, where N_i is a nilpotent matrix with ones on the first superdiagonal, if the geometric multiplicity of λ_i , denoted by $\text{gm}(\lambda_i)$, is equal to one. As discussed in Sarkar and Rakhlin [2019], the estimation of the unstable component becomes inconsistent when the geometric multiplicity of the unstable eigenvalues is greater than one. In our setting, this effect deflates C_σ which in turn increases the number of samples T when A contains unstable modes with geometric multiplicity greater than one.

Figure 2, illustrates these trends for a simple example with $d_X = 3$ states and $\ell = 2$ unstable modes. The plot depicts the mean and the standard deviation for 10 different random initial conditions. Notably, learning the unstable subspace for a diagonalizable matrix (blue curve) requires roughly the same amount of data as for a non-diagonalizable case with $\text{gm}(\lambda) = 1$ (green curve). In contrast, as the least unstable mode gets close to one, successful estimation becomes infeasible.

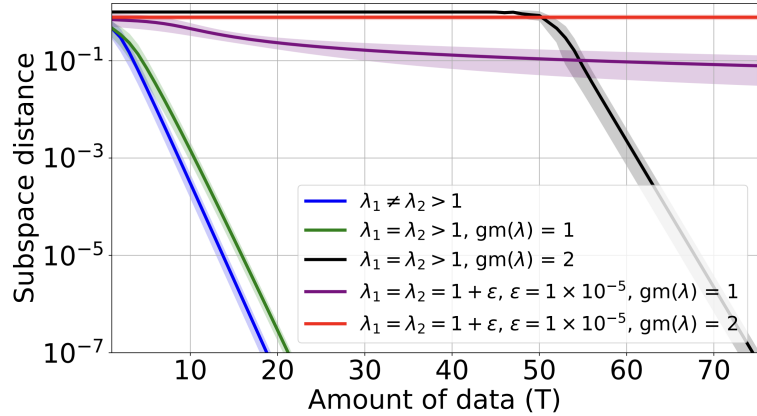


Figure 2: Subspace distance $d(\hat{\Phi}, \Phi)$ with respect to data (T) .

Remark 3.1. We emphasize that the guarantees for learning the right unstable subspace of A presented in Zhang et al. [2024a] could not be directly applied in our setting. This is because their order of T depends inversely on the gap between the unstable modes, which becomes problematic when the system matrix is non-diagonalizable, as the gap goes to zero and the amount of data grows prohibitively large. Moreover, such a dependence on the spectral gap appears to be counterintuitive and does not align with the results illustrated in Figure 2.

4 Learning to Stabilize on the Unstable Subspace

We now introduce our approach for learning to stabilize (LTS) by operating on the system’s unstable subspace. This method combines the unstable subspace representation learning, discussed in the previous section, with the discounted LQR method applied directly on the learned subspace. Specifically, our goal is to learn a low-dimensional controller $\theta \in \mathcal{S}_\theta^1$ that stabilizes the unstable dynamics (A_u, B_u) . This is accomplished by solving a series of discounted LQR problems with PG.

Recall that PG requires access to the gradient $\nabla J^\gamma(\theta, \Phi)$. However, because we operate in a model-free setting, this gradient cannot be computed directly. To address this, we use a zeroth-order gradient estimation method [Flaxman et al., 2004, Spall, 2005], which yields a gradient estimate denoted by $\hat{\nabla} J^\gamma(\theta, \hat{\Phi})$. This estimation is performed by collecting trajectory data from the original system (1), projected onto the estimated left unstable subspace via $\hat{\Phi}$, i.e., using $z_t = \hat{\Phi}^\top x_t$.

Before introducing the zeroth-order gradient estimation and its finite-sample guarantees, we first provide an upper bound on the error between $\nabla J^\gamma(\theta, \Phi)$ and $\nabla J^\gamma(\theta, \hat{\Phi})$.

Lemma 4.1. *Suppose that $\theta \in \mathcal{S}_\theta^\gamma$. Then,*

$$\left\| \nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi}) \right\|_F \leq C_\Phi d(\hat{\Phi}, \Phi), \text{ with } C_\Phi = \sqrt{\ell} \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right).$$

Note that the error in the gradient incurred by the learned representation, i.e., $\hat{\Phi}$, can be made arbitrarily small, provided that $d(\hat{\Phi}, \Phi)$ is sufficiently small. The proof of this lemma follows from Lemmas 2.1 and 2.2, combined with the upper bound $\|\hat{\Phi} - \Phi\| \leq \sqrt{2\ell} d(\hat{\Phi}, \Phi)$ from [Hu et al., 2022, Corollary 5.3]. Additional details can be found in Appendix E.

4.1 Gradient and Cost Estimation

The zeroth-order gradient estimation method is standard and has been widely adopted for policy gradient estimation in model-free LQR [Fazel et al., 2018, Malik et al., 2019]. Next, we define the two-point zeroth-order estimation.

$$\hat{\nabla} J^\gamma(\theta, \hat{\Phi}) := \frac{1}{2rn_s} \sum_{i=1}^{n_s} (V^{\gamma, \tau}(\theta_{1,i}, z_0^i) - V^{\gamma, \tau}(\theta_{2,i}, z_0^i)) U_i,$$

where U_i is randomly drawn from a uniform distribution on the sphere $\sqrt{\ell d_U} \mathbb{S}^{\ell d_U - 1}$. In addition, we have that $\theta_{1,i} = \theta + rU_i$, $\theta_{2,i} = \theta - rU_i$, with $r > 0$ denoting the smoothing radius and n_s the number of rollouts (or trajectories). Let $\tau > 0$ denote the time horizon. The finite-horizon value function $V^{\gamma, \tau}(\theta, z_0)$ is defined as follows:

$$V^{\gamma, \tau}(\theta, z_0) := \sum_{t=0}^{\tau-1} \gamma^t z_t^\top \left(\hat{\Phi}^\top Q \hat{\Phi} + \theta^\top R \theta \right) z_t,$$

with $\{z_t\}_{t=0}^{\tau-1} = \{\hat{\Phi}^\top x_t\}_{t=0}^{\tau-1}$ and $\{x_t\}_{t=0}^{\tau-1}$ being the trajectory data from (1) with $u_t = \theta \hat{\Phi}^\top x_t$.

Lemma 4.2. *Let ζ and ε_τ be positive scalars. Suppose that $n_s = \mathcal{O}(\zeta^4 \mu_\psi^4 \log^6(\ell))\ell$, $\tau = \mathcal{O}(\log(1/\varepsilon_\tau))$ and $r = \mathcal{O}(\sqrt{\varepsilon_\tau})$. It holds with probability at least $1 - c_1(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-c_2 n_s})$ that*

$$\|\hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F^2 \leq C_{est,1} \|\nabla J^\gamma(\theta)\|_F^2 + C_{est,1} C_\Phi^2 d(\hat{\Phi}, \Phi)^2 + \varepsilon_\tau^2,$$

$$\langle \nabla J^\gamma(\theta), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle \geq C_{est,2} \|\nabla J^\gamma(\theta)\|_F^2 - C_{est,3} C_\Phi^2 d(\hat{\Phi}, \Phi)^2 - C_{est,4} \varepsilon_\tau^2,$$

with positive scalars c_1 and c_2 . $C_{est,1}$, $C_{est,3}$, and $C_{est,4}$ scale as $\mathcal{O}(d_U \ell \log^2(\ell))$, and $C_{est,2} = \mathcal{O}(1)$.

The proof for this lemma follows from [Mohammadi et al., 2020, Section V] and Lemma 4.1. Lemma 4.2 states that if Φ is accurately estimated, the smoothing parameter r is chosen sufficiently small, and both the time horizon τ and the number of rollouts n_s are sufficiently large,

then $\|\widehat{\nabla} J^\gamma(\theta, \widehat{\Phi})\|_F^2 = \mathcal{O}(\|\nabla J^\gamma(\theta)\|_F^2)$ and $\langle \nabla J^\gamma(\theta, \Phi), \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \rangle = \mathcal{O}(\|\nabla J^\gamma(\theta, \Phi)\|_F^2)$, with high probability. This result is crucial to establish the linear convergence of PG for (2).

Moreover, let $\widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) = \frac{1}{n_c} \sum_{i=1}^{n_c} V^{\gamma, \tau}(\theta, z_0^i)$ be the estimated cost with n_c rollouts. We provide the following lemma to control $|J^\gamma(\theta) - \widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi})|$; the proof is deferred to Appendix F.

Lemma 4.3. *Given a stabilizing controller $\theta \in \mathcal{S}_\theta^\gamma$ and $\delta_\tau \in (0, 1)$. Suppose that*

$$\tau \geq \tau_0 := \frac{J^\gamma(\theta, \widehat{\Phi})}{\sigma_{\min}(Q)} \log \left(\frac{8(J^\gamma(\theta, \widehat{\Phi}))^2 \mu_0^2}{\sigma_{\min}(Q) J^\gamma(\theta)} \right), n_c \geq 8\mu_0^2 \log(2/\delta_\tau), \text{ and } d(\widehat{\Phi}, \Phi) \leq J^\gamma(\theta)/(4\ell\sqrt{\ell}C_{\text{cost}}).$$

Then, it holds that $|\widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) - J^\gamma(\theta)| \leq \frac{1}{2}J^\gamma(\theta)$, with probability $1 - \delta_\tau$. C_{cost} is polynomial in the problem parameters $\|A\|, \|B\|, \|Q\|, \|R\|$, and ν_θ .

4.2 Discounted LQR on the Unstable Subspace

With the gradient and cost estimation results in place, we are now ready to present our discounted LQR method on the unstable subspace for learning to stabilize the system's unstable dynamics. As a starting point, we assume access to an upper bound on the largest eigenvalue, namely, $|\lambda_1| \leq \bar{\lambda}_1$. This assumption is necessary to initialize the discount factor as $\gamma_0 < 1/\bar{\lambda}_1^2$, which ensures that the initial controller $\theta_0 \equiv 0$ stabilizes the corresponding damped system.

To ensure that the discount factor reaches one within a finite number of iterations, we adopt the explicit discount scheme proposed in Zhao et al. [2024]. In particular, $\gamma_{j+1} = (1 + \xi\alpha_j)\gamma_j$, where $\xi \in (0, 1)$ is the decay factor and the update rate α_j is given by

$$\alpha_j = \frac{3\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta_j^\top R \theta_j)}{\frac{4}{3}\widehat{J}^{\gamma_j, \tau}(\theta_j, \widehat{\Phi}) - 3\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta_j^\top R \theta_j)}. \quad (6)$$

The update rule for the discount factor follows from the Lyapunov stability analysis of the low-dimensional damped system. Let $V(z_t) = z_t^\top P_\theta^\gamma z_t$ be a quadratic Lyapunov function for the damped dynamics $z_{t+1} = \sqrt{\gamma_{j+1}}(A_u + B_u \theta)z_t$, and define $\Delta V = V(z_{t+1}) - V(z_t)$. Hence, we have

$$\Delta V = z_t^\top \left(\frac{\gamma_{j+1}}{\gamma_j} (P_\theta^\gamma - \Phi^\top Q \Phi - \theta^\top R \theta) - P_\theta^\gamma \right) z_t,$$

and thus $\frac{\gamma_{j+1}}{\gamma_j} (P_\theta^\gamma - \Phi^\top Q \Phi - \theta^\top R \theta) - P_\theta^\gamma \prec 0$ yields $\sqrt{\gamma_{j+1}}\rho(A_u + B_u \theta) < 1$. Sufficiently, we write

$$1 - \frac{\gamma_j}{\gamma_{j+1}} \leq \sigma_{\min}(\Phi^\top Q \Phi + \theta^\top R \theta) / \text{Tr}(P_\theta^\gamma) \leq \frac{3}{2}\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta) / J^\gamma(\theta),$$

where the last inequality follows from Bauer-Fike theorem [Bauer and Fike, 1960] and making the subspace distance to satisfy $d(\widehat{\Phi}, \Phi) \leq \frac{\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)}{4\|Q\|\sqrt{2\ell\kappa}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)}$. Therefore, by invoking Lemma 4.3, we recover (6). As also discussed in Zhao et al. [2024], the decay factor $\xi \in (0, 1)$ ensures that the updated controller θ_{j+1} “strongly” stabilizes the damped system $(A_u^{\gamma_{j+1}}, B_u^{\gamma_{j+1}})$. In the following section, we show the role of ξ in providing a uniform bound for $\sqrt{\gamma_{j+1}}\rho(A_u^{\gamma_{j+1}} + B_u^{\gamma_{j+1}}\theta_{j+1})$.

We conclude this section by presenting the algorithm for learning a stabilizing controller for system (1) by operating on its unstable subspace. As previously discussed, we initialize the discount

factor with $\gamma_0 < 1/\bar{\lambda}_1^2$ and choose $\xi \in (0, 1)$. The data D collected from the adjoint system is used in line 2 of Algorithm 1 to estimate the left unstable subspace representation, which in turn defines the discounted LQR problem over the learned subspace. The algorithm proceeds by solving a sequence of low-dimensional discounted LQR problems, while $\gamma_j < 1$ (lines 4–8). In particular, given a stabilizing controller θ_j for the damped system with factor γ_j , N policy gradient iterations using the estimated gradient $\widehat{\nabla} J^{\gamma_j}(\theta, \widehat{\Phi})$ are performed (lines 5 and 6). The number of iterations N is set to ensure that $J^{\gamma_j}(\theta) \leq \bar{J}$ (this is made explicit in the next section). The discount factor is updated with (6) (line 8). Finally, Algorithm 1 returns $K = \theta_{j+1} \widehat{\Phi}^\top \in \mathcal{S}_K^1$.

Algorithm 1 Learning to Stabilize on the Unstable Subspace

- 1: **Input:** $\gamma_0, \xi, N, \eta, D$
 - 2: **Compute** $D = U\Sigma V^\top$ and let $\widehat{\Phi} = [u_1, \dots, u_\ell]$ be the top ℓ columns of U
 - 3: **Initialize** $\theta_0 = 0$ and $j = 0$
 - 4: **While** $\gamma_j < 1$ **do**
 - 5: **Initialize** $\bar{\theta}_0 = \theta_j$ and **for** $n = 0, 1, \dots, N - 1$ **do**
 - 6: $\bar{\theta}_{n+1} = \bar{\theta}_n - \eta \widehat{\nabla} J^{\gamma_j}(\bar{\theta}_n, \widehat{\Phi})$
 - 7: **Let** $\theta_{j+1} = \bar{\theta}_N$ and **compute** α_j as in Eq. (6)
 - 8: **Update** $\gamma_{j+1} = (1 + \xi\alpha_j)\gamma_j$ and $j \leftarrow j + 1$
 - 9: **Output:** $K = \theta_{j+1} \widehat{\Phi}^\top$
-

Next, we provide the condition on n_s, n_c, r, τ, T, N , and η to guarantee that $K \in \mathcal{S}_K^1$, namely, K is a stabilizing controller for the original system (1).

5 Sample Complexity Analysis

We now present our main results. We first establish the conditions under which the lifted controller $K = \theta_{j+1} \widehat{\Phi}^\top$ stabilizes (1). We then quantify the sample complexity reduction achieved by our approach. To facilitate a clear presentation, we introduce the following key quantities.

$$\varepsilon_\tau := \sqrt{\frac{J_\star^1}{\mu_{\text{PL}} (d_U(\ell \log^2 \ell))}}, \bar{\lambda}_\theta := \sqrt{1 - \frac{3(1-\xi)\sigma_{\min}(Q)}{2\bar{J}}}, \bar{J} := \max\{2J_\star^1, J^{\gamma_0}(0)\},$$

and $\underline{\alpha} := 3\sigma_{\min}(Q)/(2\bar{J} - 3\sigma_{\min}(Q))$.

Theorem 5.1. *Given $\delta_\tau \in (0, 1)$, $\delta_\sigma \in (0, 1)$, and $\zeta > 0$. Suppose that $n_s = \mathcal{O}(\zeta^4 \mu_\psi^4 \log^6(\ell))\ell$, $n_c = \mathcal{O}(\log(1/\delta_\tau))$, $\tau = \mathcal{O}(\log(1/\varepsilon_\tau) + \tau_0)$, $r = \mathcal{O}(\sqrt{\varepsilon_\tau})$, and $T = \mathcal{O}\left(\log\left(\frac{\ell^7(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\varepsilon_{\text{dist}}\delta_\sigma^3}\right)\right)$, with*

$$\varepsilon_{\text{dist}} := \min\left\{\frac{(1 - \max\{\bar{\lambda}_\theta, |\lambda_{\ell+1}|\})^\ell}{C_{\text{dist},1}}, \sqrt{\frac{J_\star^1}{C_{\text{dist},2}}}\right\}.$$

In addition, suppose that the number of PG iterations and step-size satisfy

$$N \geq \frac{\mu_{\text{PL}}}{\eta} \log\left(\frac{2\bar{J}^2}{(1-\xi)\sigma_{\min}(Q)J_\star^1}\right), \quad \eta = \tilde{\mathcal{O}}(1/(d_U\ell)).$$

Then, Algorithm 1 returns $K \in \mathcal{S}_K^1$ with $\rho(A + BK) < \bar{\lambda}_\theta$, within $j = \log(1/\gamma_0)/\log(1 + \xi\underline{\alpha})$ iterations, with probability $1 - \bar{\delta}$, where $\bar{\delta} := \delta_\sigma + j(\delta_\tau + \bar{c}_1 N(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-\bar{c}_2 n_s}))$.

Note that \bar{c}_1 and \bar{c}_2 above are positive constants and the quantities $C_{\text{dist},1}$ and $C_{\text{dist},2}$ are polynomial in the problem parameters $\|A\|$, $\|B\|$, $\|Q\|$, ν_θ , L_θ , L_K , ϕ , ℓ and d_U .

Theorem 5.1 characterizes the convergence of Algorithm 1 to a stabilizing controller of (1). In particular, when the learned unstable subspace representation $\hat{\Phi}$ is sufficiently accurate and, the number of rollouts n_s , n_c , time horizon τ and number of iterations N are set large enough, with r and η small enough, our algorithm produces a low-dimensional controller that stabilizes the system’s unstable dynamics, i.e., $\theta_{j+1} \in \mathcal{S}_\theta^1$. When lifted through $\hat{\Phi}$, this controller stabilizes (1). Our results also highlight that learning to stabilize on the unstable subspace becomes more demanding (as it requires more data) when the least stable mode, $|\lambda_{\ell+1}|$, approaches marginal stability (i.e., $|\lambda_{\ell+1}| \approx 1$). It is also important to emphasize that, in contrast to Werner and Peherstorfer [2025], our work is the first to provide the non-asymptotic guarantees for learning to stabilize LTI systems via the unstable subspace representation with policy gradient. We present the proof of Theorem 5.1 in Appendix H.3. Next, we briefly discuss the idea of the proof.

Proof idea: The first step of the proof is to guarantee that $J^{\gamma_j}(\theta) \leq \bar{J}$ for every iteration. To do so, we use Lemmas 2.1, 2.2, and 4.2) to determine the number of PG iterations N to ensure $J^{\gamma_j}(\theta) \leq \bar{J}$. A preliminary condition on $d(\hat{\Phi}, \Phi)$, and on the estimation parameters n_s, n_c, τ , and r , comes from this step, where we set their corresponding error terms to scale as $\mathcal{O}(\bar{J} - J_\star^\gamma)$. We note that such a uniform bound on the cost implies that α_j is uniformly lower bounded as $\alpha_j \geq \underline{\alpha}$. Hence, γ_j reaches one within $\log(1/\gamma_0)/\log(1 + \xi\underline{\alpha})$ iterations. Moreover, since $\sqrt{(1 + \alpha_j)\gamma_j\rho(A_u + B_u\theta_{j+1})} < 1$, then it holds that $\sqrt{(1 + \xi\alpha_j)\gamma_j\rho(A_u + B_u\theta_{j+1})} < \bar{\lambda}_\theta$. We emphasize that $\bar{\lambda}_\theta$ depends on ξ , which is set within $(0, 1)$ to guarantee that the spectral radius of the closed-loop low-dimensional system is much smaller than one. It then follows from an induction step that $\rho(A_u + B_u\theta_{j+1}) < \bar{\lambda}_\theta$.

The final step of the proof is to demonstrate that $K = \theta_{j+1}\hat{\Phi}^\top$ stabilizes (1). For this, we note that $A + BK$ is equivalent to

$$\Omega \left(\begin{bmatrix} A_u + B_u\theta_{j+1}\hat{\Phi}^\top\Phi & B_u\theta_{j+1}\hat{\Phi}^\top\Phi_\perp \\ \Delta + B_s\theta_{j+1}\hat{\Phi}^\top\Phi & A_s + B_s\theta_{j+1}\hat{\Phi}^\top\Phi_\perp \end{bmatrix} \right) \Omega^\top,$$

where its spectral radius can be controlled by using the block perturbation bound from Mathias [1998] and the generalized Bauer-Fike theorem [Golub and Van Loan, 2013]. In particular, it is important to remark that the exponential dependence on ℓ showing up in $\varepsilon_{\text{dist}}$ follows from the generalized Bauer-Fike theorem, due to the fact that $A_u + B_u\theta_{j+1}$ and A_s are non-diagonalizable.

We are now in place to characterize the sample complexity of Algorithm 1. To do so, we first quantify the sample complexity of the discounted LQR method (i.e., lines 4-8 of Algorithm 1) as the total number of system rollouts, denoted by $\mathcal{S}_c := j(n_c + n_sN)$ as in Zhao et al. [2024].

Corollary 5.1. *Let the arguments of Theorem 5.1 hold. Then, Algorithm 1 returns a stabilizing policy for system (1) within $\mathcal{S}_c = \log(\rho(A))\tilde{\mathcal{O}}(\ell^2 d_U)$ trajectories collected from (1).*

This result demonstrates the sample complexity reduction achieved by learning to stabilize on the unstable subspace. In contrast to Zhao et al. [2024], where it scales as $\log(\rho(A))\tilde{\mathcal{O}}(d_X^2 d_U)$, our approach significantly improves scalability by requiring a number of rollouts that depends on the number of unstable modes, rather than the full state dimension. It is also important to highlight that Algorithm 1 also collects samples from the adjoint system, which scales with T and d_X , where the latter is due to the element-wise computations with the adjoint operator. However, we note that for a “regular” system where the least unstable and stable modes are strictly away from

marginal stability, and $\text{gm}(\lambda) = 1$ for the unstable modes, the order of T is negligible. As a result, $\tilde{\mathcal{O}}(\ell^2 d_U) + \mathcal{O}(d_X)$ is much smaller than $\tilde{\mathcal{O}}(d_X^2 d_U)$ when $\ell \ll d_X$ (i.e., our regime of interest).

6 Numerical Validation

We now present numerical experiments to validate and illustrate our theoretical guarantees¹. Additional experimental results and implementation details are provided in Appendix B.

Consider the cartpole dynamics as our nominal system (A_0, B_0) with four states and single input where (A_0, B_0) are obtained by linearizing (around the origin) and discretizing with Euler's method the following equations:

$$(m_c + m_p)\ddot{x} + m_p \ell_p (\ddot{\zeta} \cos(\zeta) - \dot{\zeta}^2 \sin(\zeta)) = u, \text{ and } m_p(\ddot{x} \cos(\zeta) + \ell_p \ddot{\zeta} - g \sin(\zeta)) = 0, \quad (7)$$

where x is the position of the cart and ζ denotes the angle of the pendulum. In addition, $m_c = 1$, $m_p = 1$, and $\ell_p = 1$ denote the mass of the cart, the mass of the pole, and the length of the pole, respectively. We set the gravitational constant to $g = 10$ and the discretization step-size to 0.25. The resulting discrete-time LTI dynamics (A_0, B_0) has $\ell = 3$ unstable modes with $|\lambda_1| = |\lambda_2| = 1$ and $|\lambda_3| = 2.12$, where the geometric multiplicity of λ_1 is equal to one. This nominal system is then augmented by adding random stable modes, resulting in a higher-dimensional system with $d_X = 30$ states and single input, while preserving the original three unstable modes of (A_0, B_0) . We adopt $T = 40$ samples from the adjoint system to learn the left unstable subspace representation. Figures 3 and 4 depict the mean and standard deviation across five runs.

Figure 3 illustrates the closed-loop spectral radius $\rho(A + BK_j)$ with respect to the iteration count j , for two cases: 1) (green curve) Algorithm 1, where we learn an unstable subspace representation and perform discounted LQR method to stabilize only the unstable modes the high-dimensional system; 2) (blue curve) applying the discounted LQR method [Zhao et al., 2024] to stabilize the full dynamics of the high-dimensional system. We note that, by stabilizing only the unstable dynamics while operating on the unstable subspace, Algorithm 1 can significantly reduce the number of iterations and thus the amount of samples required to find a stabilizing policy.

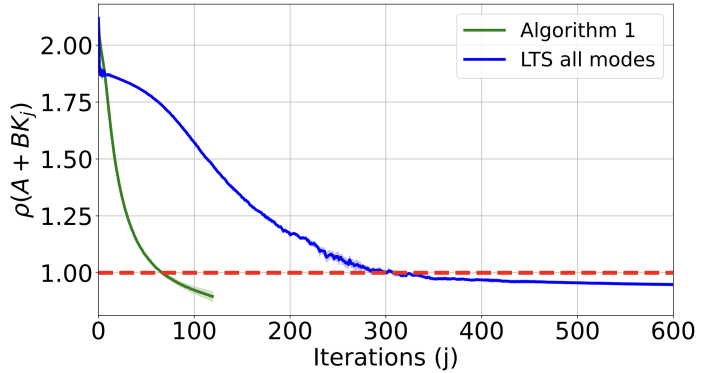


Figure 3: Closed-loop spectral radius w.r.t. iterations.

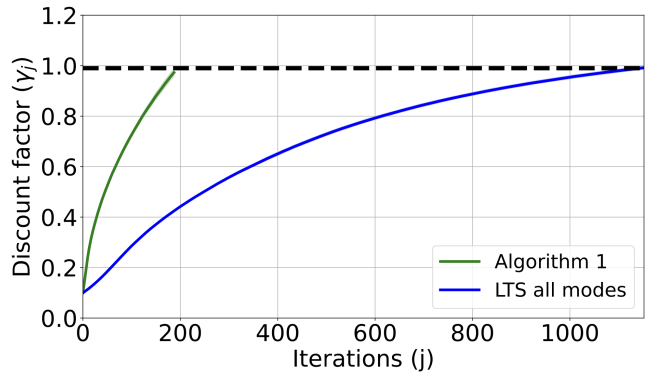


Figure 4: Discount factor w.r.t. iterations.

¹Code is available at <https://github.com/jd-anderson/LTS-unstable-representation>.

This trend is even more pronounced in Figure 4, which shows the evolution of the discount factor γ_j as a function of the iteration count j . We observe that Algorithm 1 reaches $\gamma_j = 1$ in approximately 200 iterations, whereas the approach that stabilizes all modes, as in Zhao et al. [2024], requires around 1200 iterations. These results support our theoretical guarantees (i.e., Theorem 5.1 and Corollary 5.1) which predict the sample complexity reduction achieved by restricting policy gradient updates to the left unstable subspace in the discounted LQR setting.

7 Conclusions and Future Work

We studied the problem of learning to stabilize an LTI system. To solve this problem, we proposed a sample efficient method to learn the left unstable space of the system with finite-sample guarantees. We then applied a discount LQR method based on the learned left unstable subspace representation of the system. We proved that when the unstable subspace representation is accurately recovered, the discount method on the unstable subspace returns a stabilizing policy for the original system within a number of iterations that is much smaller than that of learning to stabilize all modes. Compared to existing works, our approach accommodates non-diagonalizable systems and reveal the sample complexity reduction of LTS on the unstable subspace. Future work includes studying the LTS problem for multiple systems with “similar” unstable subspaces and learning the representation online where we continuously update the learned unstable subspace as more data becomes available.

8 Acknowledgments

The authors thank Bruce D. Lee for instructive discussions at the initial stage of this work. Leonardo F. Toso is funded by the Center for AI and Responsible Financial Innovation (CAIRFI) Fellowship and by the Columbia Presidential Fellowship. James Anderson is partially funded by NSF grants ECCS 2144634 and 2231350. Lintao Ye is supported in part by NSFC grant 62203179.

References

- Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.
- Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a Theoretical Foundation of Policy Optimization for Learning Control Policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123–158, 2023.
- Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

- Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd international conference on artificial intelligence and statistics*, pages 2916–2925. PMLR, 2019.
- Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020.
- Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Convergence and sample complexity of gradient methods for the model-free linear-quadratic regulator problem. *IEEE Transactions on Automatic Control*, 67(5):2435–2450, 2021.
- Asaf B Cassel and Tomer Koren. Online policy gradient for model free learning of linear quadratic regulators with \sqrt{T} regret. In *International Conference on Machine Learning*. PMLR, 2021.
- Han Wang, Leonardo F Toso, Aritra Mitra, and James Anderson. Model-free Learning with Heterogeneous Dynamical Systems: A Federated LQR Approach. *arXiv preprint arXiv:2308.11743*, 2023.
- Leonardo Felipe Toso, Donglin Zhan, James Anderson, and Han Wang. Meta-learning linear quadratic regulators: a policy gradient maml approach for model-free lqr. In *6th Annual Learning for Dynamics & Control Conference*, pages 902–915. PMLR, 2024a.
- Leonardo F. Toso, Han Wang, and James Anderson. Asynchronous heterogeneous linear quadratic regulator design. *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 801–808, 2024b.
- Donglin Zhan, Leonardo F Toso, and James Anderson. Coreset-based task selection for sample-efficient meta-reinforcement learning. *arXiv preprint arXiv:2502.02332*, 2025.
- Aritra Mitra, Lintao Ye, and Vijay Gupta. Towards model-free lqr control over rate-limited channels. In *6th Annual Learning for Dynamics & Control Conference*, pages 1253–1265. PMLR, 2024.
- Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R Jovanović. On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, 5(3):989–994, 2020.
- Anastasios Tsiamis, Ingvar M Ziemann, Manfred Morari, Nikolai Matni, and George J Pappas. Learning to control linear systems can be hard. In *Conference on Learning Theory*, pages 3820–3857. PMLR, 2022.
- Xiong Zeng, Zexiang Liu, Zhe Du, Necmiye Ozay, and Mario Sznaier. On the hardness of learning to stabilize linear systems. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 6622–6628. IEEE, 2023.
- Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in linear quadratic regulators. *arXiv preprint arXiv:2007.12291*, 31:32, 2020.

- Andrew Lamperski. Computing stabilizing linear controllers via policy iteration. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1902–1907. IEEE, 2020.
- Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. In *Conference on Learning Theory*, pages 1114–1143. PMLR, 2021.
- Juan Perdomo, Jack Umenberger, and Max Simchowitz. Stabilizing dynamical systems via policy gradient methods. *Advances in neural information processing systems*, 34:29274–29286, 2021.
- Yang Hu, Adam Wierman, and Guannan Qu. On the sample complexity of stabilizing lti systems on a single trajectory. *Advances in Neural Information Processing Systems*, 35:16989–17002, 2022.
- Feiran Zhao, Xingyun Fu, and Keyou You. Convergence and sample complexity of policy gradient methods for stabilizing linear systems. *IEEE Transactions on Automatic Control*, 2024.
- Ziyi Zhang, Yorie Nakahira, and Guannan Qu. Learning to Stabilize Unknown LTI Systems on a Single Trajectory under Stochastic Noise. *arXiv preprint arXiv:2406.00234*, 2024a.
- Steffen WR Werner and Benjamin Peherstorfer. System stabilization with policy optimization on unstable latent manifolds. *Computer Methods in Applied Mechanics and Engineering*, 433: 117483, 2025.
- Thomas TCK Zhang, Leonardo Felipe Toso, James Anderson, and Nikolai Matni. Sample-efficient linear representation learning from non-IID non-isotropic data. In *The Twelfth International Conference on Learning Representations*, 2024b.
- Bruce D Lee, Leonardo F Toso, Thomas T Zhang, James Anderson, and Nikolai Matni. Regret analysis of multi-task representation learning for linear-quadratic adaptive control. *arXiv preprint arXiv:2407.05781*, 2024.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446): eaat8414, 2019.
- Gilbert W Stewart and Ji-guang Sun. Matrix perturbation theory. *Academic press*, 1990.
- Omran Kouba and Dennis S Bernstein. What is the adjoint of a linear system?[lecture notes]. *IEEE Control Systems Magazine*, 40(3):62–70, 2020.
- Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.
- Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.
- Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

- James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005.
- Friedrich L Bauer and Charles T Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, 1960.
- Roy Mathias. Quadratic residual bounds for the hermitian eigenvalue problem. *SIAM journal on matrix analysis and applications*, 19(2):541–550, 1998.
- Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Appendix

Table of Contents

A	Appendix Roadmap	20
B	Additional Experiments	20
C	Auxiliary Results	21
D	Discounted LQR Problem	23
E	Linear Decomposition of the Control Policy	23
E.1	Gradient Error Due to Misspecified Representation	24
F	Gradient and Cost Estimation	25
F.1	Cost Estimation Error	27
G	Learning the Left Unstable Subspace Representation	28
H	Stabilizing Only the Unstable Modes	32
H.1	Lifting the Controller	34
H.2	Policy Gradient Per-iteration Stability Analysis	36
H.3	Sample Complexity Reduction	37

A Appendix Roadmap

The appendix is organized as follows. Section B provides additional experiments and further details on the experimental setup used in Section 6 to validate our theoretical guarantees. Next, in Section C, we revisit several auxiliary results crucial for deriving the convergence guarantees of Algorithm 1, including the Davis-Kahan theorem [Davis and Kahan, 1970] and the generalized Bauer-Fike theorem [Golub and Van Loan, 2013], which are used to control the subspace distance $d(\hat{\Phi}, \Phi)$ and the closed-loop spectral radius $\rho(A + B\theta_{j+1}\hat{\Phi}^\top)$, respectively. We then re-state the discounted LQR problem and the linear decomposition of the high-dimensional control gain K in Sections D and E, respectively, where we derive an upper bound on $\|\nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi})\|$ in Section E.1. The gradient and cost estimation guarantees are presented in Section F.

Section G is dedicated to establishing finite-sample guarantees for learning the left unstable subspace representation, which are then leveraged in Section H to derive conditions on the problem parameters under which Algorithm 1 returns a stabilizing policy for system (1).

B Additional Experiments

For the numerical experiments presented in Section 6, we consider the cartpole dynamics (7) and obtain the following nominal system matrices:

$$A_0 = \begin{bmatrix} 1 & 0.25 & 0 & 0 \\ 0 & 1 & -2.5 & 0 \\ 0 & 0 & 1 & 0.25 \\ 0 & 0 & 5 & 1 \end{bmatrix}, B_0 = \begin{bmatrix} 0 \\ 0.25 \\ 0 \\ -0.25 \end{bmatrix}.$$

Augmenting the nominal system: $A = \text{blkdiag}(A_0, \frac{1}{2}(\tilde{A} + \tilde{A}^\top)/\|\tilde{A} + \tilde{A}^\top\|)$ where $\tilde{A} \in \mathbb{R}^{d_x-4 \times d_x-4}$ is a random matrix with entries drawn from a normal distribution. In addition, $B = [B_0^\top \ \frac{1}{2}\tilde{B}^\top/\|\tilde{B}\|]^\top$ where $\tilde{B} \in \mathbb{R}^{(d_x-4) \times d_u}$ has also i.i.d. normal distributed entries.

Problem parameters: We use $T = 40$ for the number of samples collected from the adjoint system to learn the left unstable subspace representation. The zeroth-order estimation and Algorithm 1 parameters are set to $n_s = 20$, $n_c = 100$, $\tau = 50$, $r = 1 \times 10^{-3}$, $\gamma_0 = 0.1$, and $\xi = 0.9$. We refer the reader to our code² for additional details.

Inverted Pendulum: We also provide numerical results for the linearized (around the origin) and discretized (with Euler’s method) inverted pendulum dynamics given by

$$A_0 = \begin{bmatrix} 1 & dt \\ \frac{g}{\ell} & 1 \end{bmatrix}, B_0 = \begin{bmatrix} 0 \\ \frac{d_t}{m\ell_p^2} \end{bmatrix},$$

where $g = 10$, $m = 1$, $\ell_p = 1$, and $d_t = 0.25$. We augment this nominal system as discussed previously, where the ambient problem dimension is set to $d_x = 20$. The inverted pendulum has a single unstable mode $\lambda_1 = 1.79$ and it is easier to stabilize compared to the cartpole system (7). The problem parameters are set as follows: $T = 40$, $n_s = 20$, $n_c = 100$, $\tau = 50$, $r = 1 \times 10^{-3}$, $\gamma_0 = 0.1$, and $\xi = 0.9$. Figure 5 shows the closed-loop spectral radius $\rho(A + BK_j)$ (left) and the

²Code is available at <https://github.com/jd-anderson/LTS-unstable-representation>.

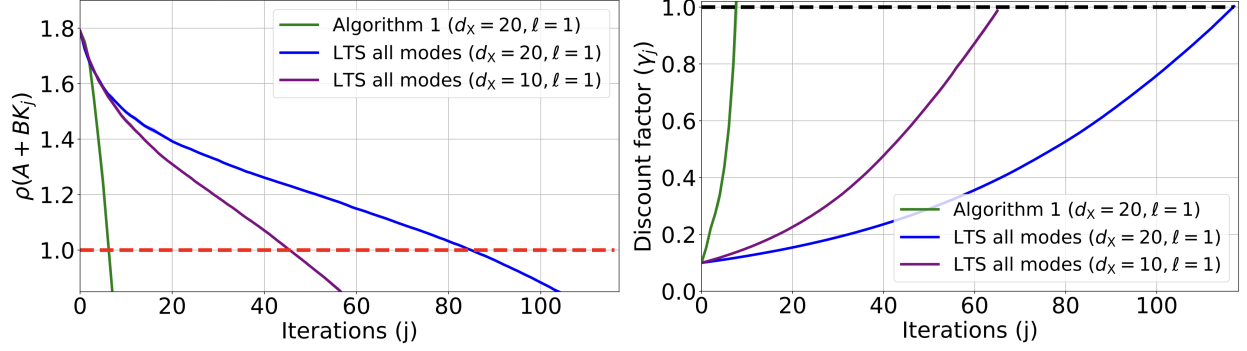


Figure 5: Closed-loop spectral radius (left) and discount factor (right) w.r.t the iteration count.

discount factor γ_j (right) with respect to the iteration count j , for three different cases: 1) (green curve) Algorithm 1 to stabilize a system with $d_X = 20$ states; 2) (blue curve) discounted LQR method, as in [Zhao et al., 2024], to stabilize the full dynamics of a system with $d_X = 20$ states; and 3) (purple curve) also learning to stabilize all modes but with a system with $d_X = 10$ states.

As predicted by our theoretical guarantees (Theorem 5.1 and Corollary H.1), Algorithm 1 significantly reduces the number of iterations and thus the number of samples required to find a stabilizing controller. Remarkably, this reduction persists even when compared to LTS all modes of a system with only half the state dimension of the ambient system to which Algorithm 1 is applied.

Random System with Multiple Inputs: We also provide experimental results for the setting where the system has multiple inputs. In particular, we generate an open-loop unstable system (A_0, B_0) with $A_0 = 2(\tilde{A} + \tilde{A}^\top) / \|\tilde{A} + \tilde{A}^\top\|$ and $B_0 = \tilde{B}^\top / \|\tilde{B}\|$ with \tilde{A} and \tilde{B} having entries randomly drawn from a normal distribution. We note that (A_0, B_0) is controllable with probability one. In particular, we randomly generate the following system matrices:

$$A_0 = \begin{bmatrix} 0.68 & 0.68 & -0.16 & 0.49 & 0.45 \\ 0.68 & 0.45 & -0.04 & -0.01 & 0.39 \\ -0.16 & -0.04 & 0.62 & 0.77 & 0.47 \\ 0.49 & -0.01 & 0.77 & 1.41 & -0.34 \\ 0.45 & 0.39 & 0.47 & -0.34 & -0.67 \end{bmatrix}, B_0 = \begin{bmatrix} -0.20 & 0.21 & 0.19 \\ 0.00 & -0.17 & 0.31 \\ 0.04 & -0.23 & -0.25 \\ -0.01 & 0.11 & 0.29 \\ 0.49 & -0.54 & 0.11 \end{bmatrix},$$

which are augmented by following the same procedure discussed previously.

Similar to previous results, Figure 6 illustrates the reduction in the number of iterations and overall sample complexity achieved by Algorithm 1, compared to the approach that stabilizes all modes as in Zhao et al. [2024]. These results further validate our theoretical guarantees and highlight the efficiency of the proposed method for learning a stabilizing controller for LTI systems.

C Auxiliary Results

Lemma C.1 (Young’s inequality). *Given any two real-valued matrices $A, B \in \mathbb{R}^{n \times m}$. It holds that*

$$\|A + B\|_2^2 \leq (1 + \beta)\|A\|_2^2 + \left(1 + \frac{1}{\beta}\right)\|B\|_2^2 \leq (1 + \beta)\|A\|_F^2 + \left(1 + \frac{1}{\beta}\right)\|B\|_F^2, \quad (8)$$

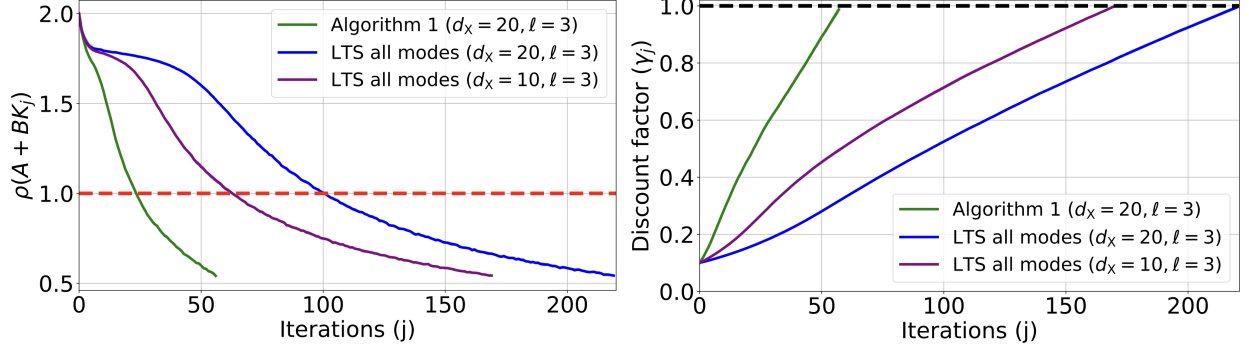


Figure 6: Closed-loop spectral radius (left) and discount factor (right) w.r.t the iteration count.

for any positive scalar $\beta > 0$. In addition, we have

$$\langle A, B \rangle \leq \frac{\beta}{2} \|A\|_2^2 + \frac{1}{2\beta} \|B\|_2^2 \leq \frac{\beta}{2} \|A\|_F^2 + \frac{1}{2\beta} \|B\|_F^2. \quad (9)$$

Theorem C.1 (Davis and Kahan [1970]). Let Σ and $\Sigma + \Delta$ be two $n \times n$ symmetric matrices with spectral decomposition

$$\Sigma = \sum_{j=1}^n \lambda_j u_j u_j^\top, \text{ and } \Sigma + \Delta = \sum_{j=1}^n \mu_j v_j v_j^\top,$$

we also let $\Pi = \sum_{j=1}^\ell u_j u_j^\top$ and $\Pi' = \sum_{j=1}^\ell v_j v_j^\top$ denote the projectors onto the subspace spanned by the top ℓ eigenvectors of Σ and $\Sigma + \Delta$, respectively. Then, it holds that

$$\|\Pi - \Pi'\| \leq \frac{\sqrt{2\ell} \|\Delta\|}{\delta},$$

where the eigengap $\delta := \inf \{|\lambda_i - \mu_j|\}, \forall i \in \{1, \dots, \ell\}, j \in \{\ell + 1, \dots, n\}$.

Theorem C.2 (Generalized Bauer-Fike [Golub and Van Loan, 2013]). Let $Q^\top A Q = D + N$ be the Schur decomposition of $A \in \mathbb{R}^{d \times d}$, where D is diagonal and N upper triangular with zeros in the diagonal. Then, it holds that

$$|\rho(A + \Delta) - \rho(A)| \leq \max \left\{ \|\Delta\| C_{bf}, (\|\Delta\| C_{bf})^{1/d} \right\}, \text{ where } C_{bf} = \sum_{i=0}^{d-1} \|N\|^i.$$

Lemma C.2 (Block perturbation bound). For any 2-by-2 block matrices M and N in the form

$$M = \begin{bmatrix} M_1 & \\ & M_2 \end{bmatrix}, N = \begin{bmatrix} & N_1 \\ N_2 & \end{bmatrix},$$

it holds that $|\rho(M + N) - \rho(M)| \leq C_{gap} \|N_1\| \|N_2\|$, where $C_{gap} = \frac{\kappa(M) \kappa(M+N)}{\min_i \{gap_i(M)\}}$.

In the lemma above, $\kappa(M)$ and $\kappa(M + N)$ denote the condition number of M and $M + N$, respectively. In addition, gap_i is the (bipartite) spectral gap around λ_i with respect to M , i.e.,

$$\text{gap}_i(M) := \begin{cases} \min_{\lambda_j \in \lambda(M_2)} |\lambda_i - \lambda_j| & \lambda_i \in \lambda(M_1) \\ \min_{\lambda_j \in \lambda(M_1)} |\lambda_i - \lambda_j| & \lambda_i \in \lambda(M_2) \end{cases}$$

with $\lambda(M_j)$ being the set of eigenvalues of M_j for $j \in \{1, 2\}$.

Proof. The proof follows directly from the quadratic residual bounds of non-symmetric matrices from [Mathias, 1998, Theorem 5]. \square

D Discounted LQR Problem

We recall that the the discounted LQR problem is defined as follows:

$$\text{minimize}_{K \in \mathcal{K}} \left\{ J^\gamma(K) := \mathbf{E} \left[\sum_{t=0}^{\infty} \gamma^t x_t^\top (Q + K^\top R K) x_t \right] \right\}, \text{ subject to (1) with } u_t = K x_t, \quad (10)$$

where the expectation is taken with respect to the randomness of the initial state. Moreover, the above discounted LQR problem is equivalent to solve

$$\text{minimize}_{K \in \mathcal{K}^\gamma} \left\{ J^\gamma(K) := \mathbf{E} \left[\sum_{t=0}^{\infty} \tilde{x}_t^\top (Q + K^\top R K) \tilde{x}_t \right] \right\}, \text{ subject to } \tilde{x}_{t+1} = (A^\gamma + B^\gamma K) \tilde{x}_t, \quad (11)$$

where $\tilde{x}_t := \gamma^{t/2} x_t$, $A^\gamma := \sqrt{\gamma} A$, $B^\gamma := \sqrt{\gamma} B$.

Definition D.1 (Set of stabilizing controllers). *Given a discount factor $\gamma \in (0, 1]$, the set of stabilizing controllers of the damped system (A^γ, B^γ) is $\mathcal{K}^\gamma := \{K \mid \sqrt{\gamma} \rho(A + BK) < 1\}$.*

Given a discount factor $\gamma \in (0, 1]$ and stabilizing controller $K \in \mathcal{K}^\gamma$ the discounted LQR cost and its gradient are given by

$$J^\gamma(K) := \text{Tr} \left(\Sigma_K^\gamma (Q + K^\top R K) \right), \quad \nabla J^\gamma(K) := 2E_K^\gamma \Sigma_K^\gamma, \quad \Sigma_K^\gamma := \mathbf{E} \left[\sum_{t=0}^{\infty} x_t x_t^\top \right], \quad (12)$$

with $E_K^\gamma := (R + B^{\gamma\top} P_K^\gamma B^\gamma) K + B^{\gamma\top} P_K^\gamma A^\gamma$, where P_K^γ is the solution of the closed-loop Lyapunov equation $P_K^\gamma = Q + K^\top R K + (A^\gamma + B^\gamma K)^\top P_K^\gamma (A^\gamma + B^\gamma K)$. Note that the discounted LQR cost can also be written as $J^\gamma(K) = \text{Tr} (P_K^\gamma)$.

E Linear Decomposition of the Control Policy

We consider the linear decomposition of the controller as $K = \theta \Phi^\top$, where $\theta \in \mathbb{R}^{d_u \times \ell}$ is a low-dimensional control gain and $\Phi \in \mathbb{R}^{d_x \times \ell}$ is the so-called representation. In addition, Φ has orthonormal columns. In particular, the columns of Φ form a basis for the left eigenspace of A

corresponding to its unstable modes. Let $z_t \in \mathbb{R}^\ell$ be a low-dimensional state that represents x_t in the subspace spanned by the columns of Φ , i.e., $x_t = \Phi z_t$. Therefore, we write

$$z_{t+1} = A_u z_t + B_u u_t, \text{ where } A_u = \Phi^\top A \Phi, B_u = \Phi^\top B, \text{ and } u_t = \theta z_t, \quad (13)$$

and state the low-dimensional LQR problem as follows:

$$\text{minimize}_{\theta \in \Theta} \left\{ J^\gamma(\theta, \Phi) := \mathbf{E} \left[\sum_{t=0}^{\infty} \gamma^t z_t^\top (\Phi^\top Q \Phi + \theta^\top R \theta) z_t \right] \right\}, \text{ subject to } z_{t+1} = (A_u + B_u \theta) z_t, \quad (14)$$

or equivalently

$$\text{minimize}_{\theta \in \Theta^\gamma} \left\{ J^\gamma(\theta, \Phi) := \mathbf{E} \left[\sum_{t=0}^{\infty} \tilde{z}_t^\top (\Phi^\top Q \Phi + \theta^\top R \theta) \tilde{z}_t \right] \right\}, \text{ subject to } \tilde{z}_{t+1} = (A_u^\gamma + B_u^\gamma \theta) \tilde{z}_t, \quad (15)$$

with $\tilde{z}_t = \gamma^{t/2} z_t$, $A_u^\gamma = \Phi^\top A^\gamma \Phi$, and $B_u^\gamma = \Phi^\top B^\gamma$. In addition, the sets of low-dimensional stabilizing controllers are defined as $\Theta := \{\theta \mid \rho(A_u + B_u \theta) < 1\}$, and $\Theta^\gamma := \{\theta \mid \sqrt{\gamma} \rho(A_u + B_u \theta) < 1\}$.

Let $\nabla J^\gamma(\theta, \Phi)$ denote the gradient with respect to θ . Therefore, we can write

$$\begin{aligned} \nabla J^\gamma(\theta, \Phi) &= \nabla J^\gamma(\theta \Phi^\top) \Phi = 2 \left((R + B^\gamma P_K^\gamma B^\gamma) K + B^\gamma P_K^\gamma A^\gamma \right) \mathbf{E} \left[\sum_{t=0}^{\infty} x_t x_t^\top \right] \Phi \\ &= 2 \left((R + B^\gamma P_K^\gamma B^\gamma) \theta + B^\gamma P_K^\gamma A^\gamma \Phi \right) \mathbf{E} \left[\sum_{t=0}^{\infty} z_t z_t^\top \right], \end{aligned}$$

with Lyapunov equation satisfying

$$\Phi^\top P_K^\gamma \Phi = \Phi^\top Q \Phi + \theta^\top R \theta + \Phi^\top (A^\gamma + B^\gamma \theta \Phi^\top)^\top P_K^\gamma (A^\gamma + B^\gamma \theta \Phi^\top) \Phi,$$

where $P_K^\gamma = \Phi P_\theta^\gamma \Phi^\top$, and thus we have

$$\nabla J^\gamma(\theta, \Phi) = 2 \left((R + B_u^\gamma P_\theta^\gamma B_u^\gamma) \theta + B_u^\gamma P_\theta^\gamma A_u^\gamma \right) \mathbf{E} \left[\sum_{t=0}^{\infty} z_t z_t^\top \right],$$

with $P_\theta^\gamma = \Phi^\top Q \Phi + \theta^\top R \theta + (A_u^\gamma + B_u^\gamma \theta)^\top P_\theta^\gamma (A_u^\gamma + B_u^\gamma \theta)$.

E.1 Gradient Error Due to Misspecified Representation

We proceed to control $\left\| \nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi}) \right\|$, where $\hat{\Phi}$ is an estimation of Φ . To do so, we have

$$\begin{aligned} \left\| \nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi}) \right\| &= \left\| \nabla J(\theta \hat{\Phi}^\top) \hat{\Pi} \hat{\Phi} - \nabla J(\theta \hat{\Phi}^\top) \Pi \hat{\Phi} + \nabla J(\theta \hat{\Phi}^\top) \Pi \hat{\Phi} - \nabla J(\theta \Phi^\top) \Pi \Phi \right\| \\ &\leq \left\| \nabla J(\theta \hat{\Phi}^\top) \hat{\Pi} \hat{\Phi} - \nabla J(\theta \hat{\Phi}^\top) \Pi \hat{\Phi} \right\| + \left\| \nabla J(\theta \hat{\Phi}^\top) \Pi \hat{\Phi} - \nabla J(\theta \Phi^\top) \Pi \Phi \right\| \end{aligned}$$

$$\begin{aligned}
&\leq \|\nabla J(\theta\hat{\Phi}^\top)\| \|\hat{\Pi} - \Pi\| + \left\| \nabla J(\theta\hat{\Phi}^\top)\Pi\hat{\Phi} - \nabla J(\theta\Phi^\top)\Pi\Phi \right\| \\
&\stackrel{(i)}{\leq} \phi d(\hat{\Phi}, \Phi) + \left\| \nabla J(\theta\hat{\Phi}^\top)\Pi\hat{\Phi} - \nabla J(\theta\Phi^\top)\Pi\Phi \right\| \\
&\leq \phi d(\hat{\Phi}, \Phi) + \left\| \nabla J(\theta\hat{\Phi}^\top)\Pi\hat{\Phi} - \nabla J(\theta\Phi^\top)\Pi\hat{\Phi} \right\| \\
&\quad + \left\| \nabla J(\theta\Phi^\top)\Pi\hat{\Phi} - \nabla J(\theta\Phi^\top)\Pi\Phi \right\| \\
&\stackrel{(ii)}{\leq} \phi d(\hat{\Phi}, \Phi) + L_K \nu_\theta \|\hat{\Phi} - \Phi\| + \phi \|\hat{\Phi} - \Phi\|,
\end{aligned}$$

where (i) follows from Lemma 2.1 and Definition 2.2. Moreover, (ii) also follows from Lemma 2.1. By leveraging [Hu et al., 2022, Corollary 5.3] we have that $\|\hat{\Phi} - \Phi\| \leq \sqrt{2\ell} d(\hat{\Phi}, \Phi)$, which implies

$$\left\| \nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi}) \right\| \leq \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right) d(\hat{\Phi}, \Phi), \quad (16)$$

or in the Frobenius norm, we have the following:

$$\left\| \nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi}) \right\|_F \leq \sqrt{\ell} \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right) d(\hat{\Phi}, \Phi). \quad (17)$$

F Gradient and Cost Estimation

Recall that we operate in model-free, namely, we do not have access to the system matrices (A, B) , and thus we cannot directly compute the gradient. Hence, we need to estimate $\nabla J^\gamma(\theta, \hat{\Phi})$. We proceed by defining the two-point zeroth-order estimation and presenting its guarantees.

$$\hat{\nabla} J^\gamma(\theta, \hat{\Phi}) := \frac{1}{2rn_s} \sum_{i=1}^{n_s} (V^{\gamma, \tau}(\theta_{1,i}, z_0^i) - V^{\gamma, \tau}(\theta_{2,i}, z_0^i)) U_i,$$

where U_i is drawn from a uniform distribution on the sphere $\sqrt{\ell d_U} \mathbb{S}^{\ell d_U - 1}$. In addition, $\theta_{1,i} = \theta + rU_i$, $\theta_{2,i} = \theta - rU_i$. Note that the initial condition of the low-dimensional system, z_0^i , is also distributed according to a zero-mean isotropic distribution, since Φ has orthonormal columns. Let $\tau > 0$ denote the time horizon. The finite-horizon value function $V^{\gamma, \tau}(\theta, z_0)$ is defined as follows:

$$V^{\gamma, \tau}(\theta, z_0) := \sum_{t=0}^{\tau-1} \gamma^t z_t^\top \left(\hat{\Phi}^\top Q \hat{\Phi} + \theta^\top R \theta \right) z_t,$$

where $\{z_t\}_{t=0}^{\tau-1} = \{\hat{\Phi}^\top x_t\}_{t=0}^{\tau-1}$ and $\{x_t\}_{t=0}^{\tau-1}$ is the trajectory data of (1) with $u_t = \theta \hat{\Phi}^\top x_t$. Moreover, let $\tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) := \frac{1}{n_s} \sum_{i=1}^{n_s} \langle \nabla V^\gamma(\theta, z_0^i), U_i \rangle U_i$ be the unbiased estimate of $\nabla J^\gamma(\theta, \hat{\Phi})$ where the infinite horizon cost is given by $V^\gamma(\theta, z_0) := \sum_{t=0}^{\infty} \gamma^t z_t^\top \left(\hat{\Phi}^\top Q \hat{\Phi} + \theta^\top R \theta \right) z_t$. Therefore, it is evident that $\mathbf{E}[\tilde{\nabla} J^\gamma(\theta, \hat{\Phi})] = \nabla J^\gamma(\theta, \hat{\Phi})$ [Mohammadi et al., 2020, Section IV].

Lemma F.1 (Zeroth-order Estimation Bias [Mohammadi et al., 2020]). *Suppose that $\tau = \mathcal{O}(\log(1/\varepsilon))$ and $r \leq \mathcal{O}(\sqrt{\varepsilon})$. Then, it holds that $\|\tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi})\|_F \leq \varepsilon$.*

Lemma F.2 (Propositions 3 and 4 of [Mohammadi et al., 2020]). *Let μ_1 and μ_2 be two positive scalars, and \mathcal{E}_1 and \mathcal{E}_2 be the following events:*

$$\mathcal{E}_1 := \left\{ \left\langle \nabla J^\gamma(\theta, \hat{\Phi}), \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\rangle \geq \mu_1 \left\| \nabla J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 \right\}, \mathcal{E}_2 := \left\{ \left\| \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 \leq \mu_2 \left\| \nabla J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 \right\}.$$

Suppose that $n_s = \mathcal{O}(\zeta^4 \mu_\psi^4 \log^6(\ell) \ell)$ for some positive scalar ζ . Then, the events \mathcal{E}_1 and \mathcal{E}_2 hold with probability $1 - c_1(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-c_2 n_s})$.

In the lemma above, c_1 and c_2 are positive constants, and the initial condition satisfies $\|z_0\|_{\psi_2} \leq \mu_\psi$. Therefore, by combining Lemmas F.1 and F.2, we obtain the following:

$$\begin{aligned} \left\| \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 &\leq \mu_2 \left\| \nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi) + \nabla J^\gamma(\theta, \Phi) \right\|_F^2 \\ &\stackrel{(i)}{\leq} 2\mu_2 \left\| \nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi) \right\|_F^2 + 2\mu_2 \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 \\ &\stackrel{(ii)}{\leq} 2\mu_2 \ell \left((L\nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + 2\mu_2 \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2, \end{aligned}$$

where (i) follows from Young's inequality (8) with $\beta = 1$ and (ii) from (17). We also use (8) to write $\left\| \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\| \geq -\left\| \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 + \frac{1}{2} \left\| \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2$, which implies that

$$\begin{aligned} \left\| \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 &\leq 4\mu_2 \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 + 2 \left\| \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 \\ &\quad + 4\mu_2 \ell \left((L_K \nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 \\ &\stackrel{(i)}{\leq} 4\mu_2 \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 + 2\varepsilon^2 + 4\mu_2 \ell \left((L_K \nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2, \end{aligned} \quad (18)$$

where (i) is due to Lemma F.1. Similarly, we can write

$$\begin{aligned} \left\langle \nabla J^\gamma(\theta, \hat{\Phi}), \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\rangle &\geq \mu_1 \left\| \nabla J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 \\ &\geq \frac{\mu_1}{2} \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 - \left\| \nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi) \right\|_F^2 \\ &\geq \frac{\mu_1}{2} \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 - \ell \left((L\nu_\theta + \phi)\sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2, \end{aligned} \quad (19)$$

along with $T_{\text{grad}} := \langle \nabla J^\gamma(\theta, \hat{\Phi}), \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle$,

$$\begin{aligned} T_{\text{grad}} &= \langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle + \langle \nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle \\ &\quad + \langle \nabla J^\gamma(\theta, \hat{\Phi}), \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle \\ &\leq \langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle + \frac{\beta}{2} \left\| \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 + \frac{1}{2\beta} \left\| \nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi) \right\|_F^2 \\ &\quad + \frac{\beta}{2} \left\| \nabla J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 + \frac{1}{2\beta} \left\| \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 \\ &\leq \langle \nabla J^\gamma(\theta, \Phi), \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \rangle + \frac{\beta}{2} \left\| \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 + \frac{1}{2\beta} \left\| \nabla J^\gamma(\theta, \hat{\Phi}) - \nabla J^\gamma(\theta, \Phi) \right\|_F^2 \\ &\quad + \beta \left\| \nabla J^\gamma(\theta, \Phi) \right\|_F^2 + \beta \left\| \nabla J^\gamma(\theta, \Phi) - \nabla J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 + \frac{1}{2\beta} \left\| \tilde{\nabla} J^\gamma(\theta, \hat{\Phi}) - \hat{\nabla} J^\gamma(\theta, \hat{\Phi}) \right\|_F^2 \end{aligned}$$

$$\begin{aligned}
&\leq \langle \nabla J^\gamma(\theta, \Phi), \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \rangle + \frac{\beta}{2} \|\widehat{\nabla} J^\gamma(\theta, \widehat{\Phi})\|_F^2 + \frac{\ell}{2\beta} \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2 \\
&+ \beta \|\nabla J^\gamma(\theta, \Phi)\|_F^2 + \beta \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2 + \frac{\varepsilon^2}{2\beta} \\
&\stackrel{(i)}{\leq} \langle \nabla J^\gamma(\theta, \Phi), \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \rangle + \beta(2\mu_2 + 1) \|\nabla J^\gamma(\theta, \Phi)\|_F^2 + \varepsilon^2 \left(1 + \frac{1}{2\beta} \right) \\
&+ \left(2\mu_2\beta + \beta + \frac{1}{2\beta} \right) \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2,
\end{aligned} \tag{20}$$

where (i) is due to (18). Hence, with $\beta = \frac{\mu_1}{4(2\mu_2+1)}$ and applying (20) in (19), we have Lemma F.3.

Lemma F.3. *Given positive scalars μ_1 , μ_2 and ζ . Suppose that we have $n_s = \mathcal{O}(\zeta^4 \mu_\psi^4 \log^6(\ell)\ell)$, $\tau = \mathcal{O}(\log(1/\varepsilon))$ and $r = \mathcal{O}(\sqrt{\varepsilon})$. Then, it holds that*

$$\begin{aligned}
\|\widehat{\nabla} J(\theta, \widehat{\Phi})\|_F^2 &\leq 4\mu_2 \|\nabla J^\gamma(\theta, \Phi)\|_F^2 + 4\mu_2 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2 + 2\varepsilon^2, \\
\langle \nabla J^\gamma(\theta, \Phi), \widehat{\nabla} J^\gamma(\theta, \widehat{\Phi}) \rangle &\geq \frac{\mu_1}{4} \|\nabla J^\gamma(\theta, \Phi)\|_F^2 - c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2 - c_5 \varepsilon^2,
\end{aligned}$$

with probability $1 - c_2(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-c_3 n_s})$, where

$$c_4 = 1 + \frac{2(2\mu_2 + 1)}{\mu_1}, \text{ and } c_5 = \frac{2\mu_2\mu_1}{4(2\mu_2 + 1)} + \frac{2(2\mu_2 + 1)}{\mu_1} + \frac{\mu_1}{4(2\mu_2 + 1)}.$$

F.1 Cost Estimation Error

We conclude this section by revisiting Lemma 5 from Zhao et al. [2024], i.e., the one that controls the error in the cost estimation, and we adapt it to our setting of performing PG on the unstable subspace. Let $\widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) = \frac{1}{n_c} \sum_{i=1}^{n_c} V^{\gamma, \tau}(\theta, z_0^i)$ be the estimated cost with n_c samples and z_0^i denoting a random draw of the low-dimensional initial state.

Lemma F.4. *Given $\theta \in \mathcal{S}_\theta^\gamma$ and $\delta_\tau \in (0, 1)$. Suppose that the time horizon τ , number of rollouts n_c , and subspace distance $d(\widehat{\Phi}, \Phi)$ satisfy*

$$\tau \geq \tau_0 := \frac{J^\gamma(\theta, \widehat{\Phi})}{\sigma_{\min}(Q)} \log \left(\frac{8(J^\gamma(\theta, \widehat{\Phi}))^2 \mu_0^2}{\sigma_{\min}(Q) J^\gamma(\theta)} \right), n_c \geq 8\mu_0^2 \log(2/\delta_\tau), \text{ and } d(\widehat{\Phi}, \Phi) \leq J^\gamma(\theta)/(4\sqrt{\ell} C_{\text{cost}}),$$

then, it holds that $|\widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) - J^\gamma(\theta)| \leq \frac{1}{2} J^\gamma(\theta)$, with probability $1 - \delta_\tau$, where C_{cost} is polynomial in the problem parameters $\|A\|$, $\|B\|$, $\|Q\|$, $\|R\|$ and ν_θ .

Proof. The proof follows from first writing

$$\begin{aligned}
\left| \widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) - J^\gamma(\theta) \right| &= \left| \widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) - J^\gamma(\theta, \widehat{\Phi}) + J^\gamma(\theta, \widehat{\Phi}) - J^\gamma(\theta) \right| \\
&\leq \left| \widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) - J^\gamma(\theta, \widehat{\Phi}) \right| + \left| J^\gamma(\theta, \widehat{\Phi}) - J^\gamma(\theta) \right|,
\end{aligned}$$

where we use [Zhao et al., 2024, Lemma 5] to control the first term. Namely, if τ and n_c are set according to the conditions of Lemma F.4, we have that $|\hat{J}^{\gamma,\tau}(\theta, \hat{\Phi}) - J^\gamma(\theta, \hat{\Phi})| \leq \frac{J^\gamma(\theta)}{4}$, with probability $1 - \delta_\tau$. On the other hand, for the second term, we have $|J^\gamma(\theta, \hat{\Phi}) - J^\gamma(\theta)| \leq \ell \|\hat{P}_\theta^\gamma - P_\theta^\gamma\|$. We then use perturbation bound of the Lyapunov equation, presented in [Toso et al., 2024a, Proof of Lemma 4] to obtain the following:

$$\|\hat{P}_\theta^\gamma - P_\theta^\gamma\| \leq C_{\text{cost},1}(\|\hat{A}_u^\gamma - A_u^\gamma\| + \|\hat{B}_u^\gamma - B_u^\gamma\|) + C_{\text{cost},2}\|\hat{\Phi}^\top Q \hat{\Phi} - \Phi^\top Q \Phi\|,$$

where $\hat{A}_u^\gamma = \hat{\Phi}^\top A^\gamma \hat{\Phi}$ and $\hat{B}_u^\gamma = \hat{\Phi}^\top B^\gamma$. In addition, $C_{\text{cost},1}$ and $C_{\text{cost},2}$ are polynomials the problem parameters $\|A\|, \|B\|, \|Q\|, \|R\|, \nu_\theta$. By using [Hu et al., 2022, Corollary 5.3], we can write $|J^\gamma(\theta, \hat{\Phi}) - J^\gamma(\theta)| \leq C_{\text{cost}} \ell \sqrt{\ell} d(\hat{\Phi}, \Phi)$. The proof is completed by setting $d(\hat{\Phi}, \Phi) \leq \frac{J^\gamma(\theta)}{4\ell\sqrt{\ell}C_{\text{cost}}}$. \square

G Learning the Left Unstable Subspace Representation

With the data of the adjoint system collected and stored in $D = [x_1, x_2, \dots, x_T] \in \mathbb{R}^{d_x \times T}$, we proceed by taking its singular value decomposition $D = U \Sigma V^\top$, where $U = [u_1, u_2, \dots, u_{d_x}] \in \mathbb{R}^{d_x \times d_x}$, $V = [v_1, v_2, \dots, v_{d_x}] \in \mathbb{R}^{T \times d_x}$, and $\Sigma = \text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_{d_x}) \in \mathbb{R}^{d_x \times d_x}$. The orthonormal basis for the left unstable subspace is constructed with the first ℓ columns of U , i.e., $\hat{\Phi} = [u_1, \dots, u_\ell]$. Let $\hat{\Pi} = \hat{\Phi} \hat{\Phi}^\top$ and $\Pi = \Phi \Phi^\top$ denote the projectors onto the column spaces of $\hat{\Phi}$ and Φ , respectively.

Goal: Prove that $d(\hat{\Phi}, \Phi) = \|\hat{\Pi} - \Pi\|$ is sufficiently small when T is sufficiently large.

To do so, we follow a similar derivation as presented in [Zhang et al., 2024a, Theorem 5.1], where the main differences in our setting is that we accommodate for non-diagonalizable system matrices A , as well as we construct the basis for the left unstable subspace of A rather than the right unstable subspace as in Zhang et al. [2024a]. Since A is assumed to be real-valued (with potential complex conjugate eigenvalues and eigenvectors) there always exist real basis matrices $\tilde{\Phi} \in \mathbb{R}^{d_x \times \ell}$ and $\tilde{\Psi} \in \mathbb{R}^{d_x \times d_x - \ell}$, for the left unstable and stable eigenspaces of A , respectively. Hence, we have

$$A^\top \tilde{P} = \tilde{P} \begin{bmatrix} \tilde{T}_u & 0 \\ 0 & \tilde{T}_s \end{bmatrix}, \text{ with } \tilde{P} = [\tilde{\Phi} \ \tilde{\Psi}] \in \mathbb{R}^{d_x \times d_x}, \tilde{T}_u \in \mathbb{R}^{\ell \times \ell}, \text{ and } \tilde{T}_s \in \mathbb{R}^{d_x - \ell \times d_x - \ell},$$

where \tilde{T}_u has the same spectrum of the Jordan blocks corresponding to the unstable eigenvalues of A , whereas \tilde{T}_s has the spectrum of the stable counterpart. By orthonormalizing the basis matrices $\tilde{\Phi}$ and $\tilde{\Psi}$ with a thin QR decomposition we obtain the following:

$$A^\top [\Phi \ \Psi] = [\Phi \ \Psi] \begin{bmatrix} R_\Phi & 0 \\ 0 & R_\Psi \end{bmatrix} \begin{bmatrix} \tilde{T}_u & 0 \\ 0 & \tilde{T}_s \end{bmatrix} \begin{bmatrix} R_\Phi^{-1} & 0 \\ 0 & R_\Psi^{-1} \end{bmatrix} = [\Phi \ \Psi] \begin{bmatrix} T_u & 0 \\ 0 & T_s \end{bmatrix},$$

with R_Φ and R_Ψ being the upper triangular matrices for the QR decomposition of $\tilde{\Phi}$ and $\tilde{\Psi}$, respectively. Their inverses exist due to the fact that $\tilde{\Phi}$ and $\tilde{\Psi}$ have full column rank. Moreover, we note that \tilde{T}_u and T_u have identical spectrum as well as \tilde{T}_s and T_s . Let $\Xi = [\Phi \ \Psi]$ be composed

by the orthonormal basis of the left unstable and stable subspaces of A , and denote its inverse by $S = [S_1^\top \ S_2^\top]^\top := \Xi^{-1}$. Therefore, we can write

$$D = \Xi S D = [\Phi \ \Psi] \begin{bmatrix} S_1 D \\ S_2 D \end{bmatrix} = \Phi D_1 + \Psi D_2 = D_u + D_s,$$

where $D_1 = S_1 D$ and $D_2 = S_2 D$. Note that D is composed of $D_u = \Phi D_1$ that comes from the left unstable subspace of A , whereas $D_s = \Psi D_2$ depends on the left stable subspace of A . The main idea is to collect enough data such that the unstable part dominates the stable one, i.e., the data sufficiently represents the unstable dynamics of (1).

We proceed our analysis by first considering D_u and writing the singular value decomposition of D_1 , namely, $D_u = \Phi D_1 = \Phi U_1 \Sigma_1 V_1^\top$, with $U_1 \in \mathbb{R}^{\ell \times \ell}$, $\Sigma_1 \in \mathbb{R}^{\ell \times \ell}$, and $V_1 \in \mathbb{R}^{T \times d_X}$. Note that $\hat{\Pi}$ is the projector onto the subspace spanned by the first ℓ columns of U , whereas Π is the projector onto the columns of ΦU_1 . In order to leverage Davis-Kahan theorem (i.e., Theorem C.1) to control the subspace distance, we first write the following symmetric matrices:

$$\mathcal{D}_u = \begin{bmatrix} 0 & D_u^\top \\ D_u & 0 \end{bmatrix} = \begin{bmatrix} 0 & V_1 \Sigma_1 U_1^\top \Phi^\top \\ \Phi U_1 \Sigma_1 V_1^\top & 0 \end{bmatrix}, \mathcal{D}_s = \begin{bmatrix} 0 & D_s^\top \\ D_s & 0 \end{bmatrix}, \mathcal{D} = \mathcal{D}_u + \mathcal{D}_s = \begin{bmatrix} 0 & D^\top \\ D & 0 \end{bmatrix},$$

and observe that the eigenvalues and eigenvectors of \mathcal{D} are $\hat{\lambda}_i = \pm \hat{\sigma}_i$ and $[v_i^\top \ \pm u_i^\top]^\top \ \forall i \in [d_X]$. Let $\{\sigma_j\}_{j=1}^\ell$ denote the top ℓ eigenvalues of \mathcal{D}_u which are the singular values of D_u . Therefore, we use Theorem C.1 to write

$$d(\hat{\Phi}, \Phi) = \|\hat{\Pi} - \Pi\| \leq \frac{\sqrt{2\ell} \|D_s\|}{\sigma_\ell - \hat{\sigma}_{\ell+1}} = \frac{\sqrt{2\ell} \|\Psi D_2\|}{\sigma_\ell - \hat{\sigma}_{\ell+1}} \leq \frac{\sqrt{2\ell} \|D_2\|}{\sigma_\ell - \hat{\sigma}_{\ell+1}}, \quad (21)$$

where we control $\|D_2\|$ as follows:

$$\|D_2\| \leq \sqrt{T} \|D_2\|_1 \leq \sqrt{T} \sum_{i=\ell+1}^{d_X} \sum_{t=1}^T |\lambda_i|^t \|x_0\| \leq \sqrt{T} \sum_{i=\ell+1}^{d_X} \sum_{t=1}^\infty |\lambda_i|^t \mu_0 \stackrel{(i)}{\leq} \frac{\sqrt{T}(d_X - \ell) \mu_0}{1 - |\lambda_{\ell+1}|}, \quad (22)$$

where (i) is due to the fact that $\{\lambda_i\}_{i=\ell+1}^{d_X}$ are the stable modes of A with $|\lambda_{\ell+1}| \geq \dots \geq |\lambda_{d_X}|$. Note that the second inequality follows from the fact that $\|D_2\|$ captures the stable dynamics in $\|D\|$, which is in the order of $|\lambda_i|^t \|x_0\|$ for any stable mode $i \in \{\ell+1, \dots, d_X\}$ of A .

By combining (21) and (22), we obtain

$$d(\hat{\Phi}, \Phi) \leq \frac{\sqrt{2\ell} \sqrt{T} (d_X - \ell) \mu_0}{(\sigma_\ell - \hat{\sigma}_{\ell+1})(1 - |\lambda_{\ell+1}|)}, \quad (23)$$

where we now proceed to control σ_ℓ and $\hat{\sigma}_{\ell+1}$. First recall that σ_ℓ is the ℓ -th top singular value of D_1 . The following lemma provides a high probability lower bound on σ_ℓ .

Lemma G.1. *Suppose that $T = \mathcal{O}(\log(\ell^7/\delta_\sigma^3)/\log(|\lambda_\ell|))$ for some $\delta_\sigma \in (0, 1)$. Then, it holds that*

$$\sigma_\ell \geq \frac{\sqrt{C_\sigma} |\lambda_\ell|^T \delta_\sigma}{2\sqrt{2} C_\psi \ell^{5/2} T^{3/2}}, \text{ with probability } 1 - 4\delta_\sigma, \text{ where } C_\sigma = \mathcal{O}(1).$$

Proof. To prove this lemma, we first define the following quantities:

$$\phi(A_u, T) := \sqrt{\inf_{v \in S_\ell(1)} \sigma_{\min} \left(\sum_{t=0}^T \Lambda_u^{-t+1} v v^\top \Lambda_u^{-t+1, \top} \right)},$$

where $S_\ell(1) := \{v \in \mathbb{R}^\ell \mid \min_{1 \leq i \leq \ell} |v_i| \geq 1\}$ is the outbox set [Sarkar and Rakhlin, 2019, Definition 3], and $A_u = P^{-1} \Lambda_u P$ is the Jordan decomposition of A_u , with $P = [P_1 \ P_2 \ \dots \ P_\ell]^\top$.

$$\psi(A_u, T) := \frac{1}{2\ell \sup_{1 \leq i \leq \ell} C_{|P_i^\top x_0|}},$$

where $C_{|P_i^\top x_0|}$ is the essential supremum of the pdf of $|P_i^\top x_0|$.

Lemma G.2 ([Sarkar and Rakhlin, 2019]). *Given $\delta_\sigma \in (0, 1)$ and suppose that T satisfy*

$$4T^2 \sigma_{\max} \left(A_u^{-(T+1)\varepsilon_T} \right) \text{Tr} \left(\Gamma_T(A_u^{-1}) \right) + \frac{T \text{Tr} \left(A_u^{-T-1} \Gamma_T(A_u^{-1}) (A_u^{-T-1})^\top \right)}{\delta_\sigma} \leq \frac{\phi^2(A_u, T) \psi^2(A_u, T) \delta_\sigma^2}{2}, \quad (24)$$

where we pick ε_T such that $\varepsilon_T(T+1) = \lfloor \frac{T+1}{2} \rfloor$, and $\Gamma_T(A_u) = \sum_{t=0}^T A_u^t (A_u^t)^\top$. Then, it holds that $\sigma_\ell \geq \frac{\phi(A_u, T) \psi(A_u, T) \delta_\sigma |\lambda_\ell|^T}{\sqrt{2}}$, with probability $1 - 4\delta_\sigma$.

• **Lower bounding $\phi(A_u, T)$:** Let $H(v) = [v \ \Lambda_u^{-1}v \ \Lambda_u^{-2}v, \dots, \Lambda_u^{-T+1}v] = \tilde{H}\tilde{V}$, where we define $\tilde{H} = [I \ \Lambda_u^{-1} \ \Lambda_u^{-2}, \dots, \Lambda_u^{-T+1}]$ with \tilde{V} being an $\ell T \times T$ matrix with $v \in S_\ell(1)$ placed accordingly. Therefore, we can write

$$\phi(A_u, T) = \sqrt{\inf_{v \in S_\ell(1)} \sigma_{\min} (H(v)H(v)^\top)},$$

which implies the following:

$$\phi(A_u, T) = \sqrt{\inf_{v \in S_\ell(1)} \sigma_{\min} (H(v)H(v)^\top)} = \inf_{v \in S_\ell(1)} \frac{1}{\|\tilde{H}^\dagger\|} \geq \frac{1}{\|\tilde{H}^\dagger\|(\ell T)^{3/2}},$$

with $\|\tilde{H}^\dagger\| = 1/\sqrt{\sigma_{\min}(\tilde{H}\tilde{H}^\top)}$ and $\sigma_{\min}(\tilde{H}\tilde{H}^\top) = \sigma_{\min} \left(\sum_{t=0}^{T-1} \Lambda_u^{-t} (\Lambda_u^{-t})^\top \right) \geq \sum_{t=0}^{T-1} \lambda_{\min} (\Lambda_u^{-t} (\Lambda_u^{-t})^\top)$.

$$\sigma_{\min}(\tilde{H}\tilde{H}^\top) \geq \sum_{t=0}^{T-1} \frac{1}{\sigma_{\max}(\Lambda_u)^{2t}} \geq \sum_{t=0}^{T-1} \left(\frac{1}{|\lambda_1| + 1} \right)^{2t} := C_\sigma = \mathcal{O}(1),$$

and thus $\|\tilde{H}^\dagger\| \leq 1/\sqrt{C_\sigma}$, which yields $\phi(A_u, T) \geq \frac{\sqrt{C_\sigma}}{(\ell T)^{3/2}}$.

Note that to lower bound $\psi(A_u, T)$, we simply need to upper bound $C_{|P_i^\top x_0|}$. We recall that x_0 is distributed according to a zero-mean and isotropic distribution (e.g., sub-Gaussian distribution),

which implies that $C_{|P_i^\top x_0|} \leq C_\psi$ for some sufficiently large constant C_ψ . Therefore, we have that $\psi(A_u, T) \geq \frac{1}{2\ell C_\psi}$, which can be used to obtain

$$\sigma_\ell \geq \frac{\phi(A_u, T)\psi(A_u, T)\delta_\sigma|\lambda_\ell|^T}{\sqrt{2}} \geq \frac{\sqrt{C_\sigma}|\lambda_\ell|^T}{2\sqrt{2}C_\psi\ell^{5/2}T^{3/2}}.$$

We complete the proof by analyzing the conditions on T to satisfy (24). Let us first write

$$4T^2\sigma_{\max}\left(A_u^{-(T+1)\varepsilon_T}\right)\text{Tr}\left(\Gamma_T(A_u^{-1})\right) + \frac{T\text{Tr}\left(A_u^{-T-1}\Gamma_T(A_u^{-1})(A_u^{-T-1})^\top\right)}{\delta_\sigma} \leq \frac{\phi^2(A_u, T)\psi^2(A_u, T)\delta_\sigma^2}{2},$$

which implies that

$$4T^3\ell|\lambda_\ell|^{-2(T+1)\varepsilon_T} + \frac{T^2\ell\sum_{i=1}^\ell|\lambda_i|^{-2(T+1)}}{\delta_\sigma} \leq \frac{\phi^2(A_u, T)\psi^2(A_u, T)\delta_\sigma^2}{2},$$

then we have $4T^3\ell|\lambda_\ell|^{-2(T+1)\varepsilon_T} \leq \frac{\phi^2(A_u, T)\psi^2(A_u, T)\delta_\sigma^2}{4}$ and $T^2\ell\sum_{i=1}^\ell|\lambda_i|^{-2(T+1)} \leq \frac{\phi^2(A_u, T)\psi^2(A_u, T)\delta_\sigma^3}{4}$,

which yields $T \geq -\frac{\log\left(\frac{\phi^2(A_u, T)\psi^2(A_u, T)\delta_\sigma^3}{4\ell^2}\right)}{\log(|\lambda_\ell|)} \geq \frac{\log(16\ell^7C_\psi^2/(C_\sigma\delta_\sigma^3))}{\log(|\lambda_\ell|)}$ and completes the proof. \square

Recall that $\hat{\sigma}_{\ell+1}$ denotes the $(\ell+1)$ -th singular value of D , which corresponds to its stable component $D_s = \Psi D_2$. Consequently, $\hat{\sigma}_{\ell+1}$ is upper bounded by the largest singular value of D_s , which in turn is bounded by the largest singular value of D_2 . This leads to the following:

$$\hat{\sigma}_{\ell+1} \leq \|D_2\| \leq \frac{\sqrt{T}(d_X - \ell)\mu_0}{1 - |\lambda_{\ell+1}|}, \quad (25)$$

where the second inequality follows from (22). By combining Lemma G.1 and (25) in (23), we have

$$\begin{aligned} d(\hat{\Phi}, \Phi) &\leq \frac{\sqrt{2\ell}\sqrt{T}(d_X - \ell)\mu_0/(1 - |\lambda_{\ell+1}|)}{\frac{\sqrt{C_\sigma}|\lambda_\ell|^T\delta_\sigma}{2\sqrt{2}C_\psi\ell^{5/2}T^{3/2}} - \sqrt{T}(d_X - \ell)\mu_0/(1 - |\lambda_{\ell+1}|)} = \frac{4\ell^3T^2(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\sqrt{C_\sigma}|\lambda_\ell|^T\delta_\sigma - 2\sqrt{2}T^2\ell^{5/2}(d_X - \ell)\mu_0} \\ &\stackrel{(i)}{\leq} \frac{8\ell^3T^2(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\sqrt{C_\sigma}|\lambda_\ell|^T\delta_\sigma}, \end{aligned}$$

where (i) is due to the selection of T according to $T \geq \log\left(\frac{4\sqrt{2}\ell^{5/2}(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\sqrt{C_\sigma}\delta_\sigma}\right)/\log|\lambda_\ell|$. We conclude by determining the condition of T that guarantees $d(\hat{\Phi}, \Phi) \leq \varepsilon$, for some small accuracy ε . Namely,

$$\frac{8\ell^3T^2(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\sqrt{C_\sigma}|\lambda_\ell|^T\delta_\sigma} \leq \varepsilon, \text{ which implies } T \geq \log\left(\frac{8\ell^3(d_X - \ell)\mu_0}{(1 - |\lambda_{\ell+1}|)\sqrt{C_\sigma}\delta_\sigma\varepsilon}\right)/\log|\lambda_\ell|,$$

with probability $1 - 4\delta_\sigma$.

H Stabilizing Only the Unstable Modes

Given an estimation of the left unstable subspace representation, $\widehat{\Phi}$, we now turn our attention to design a low-dimensional control gain θ that stabilizes the low-dimensional unstable dynamics (A_u, B_u) . To do so, we leverage the explicit discount LQR method from Zhao et al. [2024]. Our goal is to guarantee that for every iteration j of the Algorithm 1 the cost remains uniformly upper bounded, i.e., $J^{\gamma_j}(\theta_j) := J^{\gamma_j}(\theta_j, \Phi) \leq \bar{J}$, for some positive scalar \bar{J} . In addition, the updated discount factor needs to ensure that $\sqrt{\gamma_{j+1}}\rho(A_u + B_u\theta_{j+1}) < 1$ while $\gamma_{j+1} > \gamma_j$.

Lemma H.1. *Given a discount factor $\gamma \in (0, 1]$, a decay factor $\xi \in (0, 1)$, and a low-dimensional controller θ , such that $\sqrt{\gamma}\rho(A_u + B_u\theta) < 1$. In addition, suppose that τ and n_c satisfy the conditions of Lemma F.4, and suppose $\gamma_+ = (1 + \xi\alpha)\gamma$, with*

$$\alpha = \frac{3\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)}{\frac{4}{3}\widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) - 3\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)},$$

then, it holds that $\sqrt{\gamma_+}\rho(A_u + B_u\theta) < 1$.

Proof. Consider the quadratic Lyapunov function $V(z_t) = z_t^\top P_\theta^\gamma z_t$ for the corresponding low-dimensional damped system $z_{t+1} = \sqrt{\gamma_+}(A_u + B_u\theta)z_t$. Therefore, we can write

$$\begin{aligned} V(z_{t+1}) - V(z_t) &= \gamma_+ z_t^\top (A_u + B_u\theta)^\top P_\theta^\gamma (A_u + B_u\theta) z_t - z_t^\top P_\theta^\gamma z_t \\ &\stackrel{(i)}{=} z_t^\top \left(\frac{\gamma_+}{\gamma} (P_\theta^\gamma - \Phi^\top Q \Phi - \theta^\top R \theta) - P_\theta^\gamma \right) z_t, \end{aligned}$$

where (i) follows from the definition of P_θ^γ . Hence, $\frac{\gamma_+}{\gamma} (P_\theta^\gamma - \Phi^\top Q \Phi - \theta^\top R \theta) - P_\theta^\gamma \prec 0$ ensures that $\sqrt{\gamma_+}\rho(A_u + B_u\theta) < 1$. By applying the trace function on both sides, we have

$$1 - \frac{\gamma}{\gamma_+} \leq \sigma_{\min}(\Phi^\top Q \Phi + \theta^\top R \theta) / \text{Tr}(P_\theta^\gamma) \leq \frac{3}{2} \sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta) / \text{Tr}(P_\theta^\gamma),$$

where the last inequality follows from applying Bauer-Fike theorem [Bauer and Fike, 1960] along with setting T accordingly to guarantee $d(\widehat{\Phi}, \Phi) \leq \frac{\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)}{4\|Q\|\sqrt{2\ell\kappa}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)}$. In particular, since the distribution of the initial state is isotropic, we have $J^\gamma(\theta, \Phi) = \text{Tr}(P_\theta^\gamma)$, which implies

$$\begin{aligned} \gamma_+ &\leq \left(1 + \frac{\frac{3}{2}\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)}{J^\gamma(\theta, \Phi) - \frac{3}{2}\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)} \right) \gamma \\ &\stackrel{(i)}{\leq} \left(1 + \frac{3\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)}{\frac{4}{3}\widehat{J}^{\gamma, \tau}(\theta, \widehat{\Phi}) - 3\sigma_{\min}(\widehat{\Phi}^\top Q \widehat{\Phi} + \theta^\top R \theta)} \right) \gamma = (1 + \alpha)\gamma, \end{aligned} \tag{26}$$

where (i) is due to Lemma F.4. As discussed in [Zhao et al., 2024, Section III], the decay factor $\xi \in (0, 1)$ is necessary to guarantee that $\sqrt{\gamma_+}\rho(A_u + B_u\theta)$ is strictly away from one. \square

We now proceed to show that for a sufficiently large amount of PG iterations N , the LQR cost is uniformly bounded according to $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$. Given $\theta_j = \bar{\theta}_0 \in \mathcal{S}_\theta^\gamma$, we use Lemma 2.1 to write

$$\begin{aligned}
J^\gamma(\bar{\theta}_{n+1}) - J^\gamma(\bar{\theta}_n) &\leq \langle \nabla J^\gamma(\bar{\theta}_n, \Phi), \bar{\theta}_{n+1} - \bar{\theta}_n \rangle + \frac{L_\theta}{2} \|\bar{\theta}_{n+1} - \bar{\theta}_n\|_F^2 \\
&\leq -\eta \langle \nabla J^\gamma(\bar{\theta}_n, \Phi), \hat{\nabla} J^\gamma(\bar{\theta}_n, \hat{\Phi}) \rangle + \frac{L_\theta \eta^2}{2} \|\hat{\nabla} J^\gamma(\bar{\theta}_n, \hat{\Phi})\|_F^2 \\
&\stackrel{(i)}{\leq} -\frac{\eta \mu_1}{4} \|\nabla J^\gamma(\bar{\theta}_n, \Phi)\|_F^2 + \eta c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + \eta c_5 \varepsilon^2 \\
&\quad + \frac{L_\theta \eta^2}{2} \left(4\mu_2 \|\nabla J^\gamma(\bar{\theta}_n, \Phi)\|_F^2 + 4\mu_2 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + 2\varepsilon^2 \right) \\
&\stackrel{(ii)}{\leq} -\frac{\eta \mu_1}{8} \|\nabla J^\gamma(\bar{\theta}_n, \Phi)\|_F^2 + 2\eta c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + 2\eta c_5 \varepsilon^2 \\
&\stackrel{(iii)}{\leq} -\frac{\eta \mu_1}{8\mu_{\text{PL}}} (J^\gamma(\bar{\theta}_n) - J_\star^\gamma) + 2\eta c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + 2\eta c_5 \varepsilon^2,
\end{aligned}$$

where $J_\star^\gamma = J^\gamma(\theta_\star^\gamma)$, with θ_\star^γ being the optimal controller of the corresponding discounted LQR problem with discount factor γ . In addition, (i) is due to Lemma F.3 and (ii) follows from selecting the step-size according to $\eta \leq \min \left\{ \frac{\mu_1}{16\mu_2 L_\theta}, \frac{c_4}{2L_\theta \mu_2}, \frac{c_5}{L_\theta} \right\}$. (iii) follows from Lemma 2.2. Therefore, by adding and subtracting $J^\gamma(\bar{\theta}^\star)$ from both sides, we obtain

$$J^\gamma(\bar{\theta}_{n+1}) - J_\star^\gamma \leq \left(1 - \frac{\eta \mu_1}{8\mu_{\text{PL}}} \right) (J^\gamma(\bar{\theta}_n) - J_\star^\gamma) + 2\eta c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + 2\eta c_5 \varepsilon^2,$$

and by unrolling the above expression over N iterations, we have

$$J^\gamma(\bar{\theta}_N) - J_\star^\gamma \leq \left(1 - \frac{\eta \mu_1}{8\mu_{\text{PL}}} \right)^N (J^\gamma(\bar{\theta}_0) - J_\star^\gamma) + \frac{16\mu_{\text{PL}} c_4 \ell}{\mu_1} \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + \frac{16\mu_{\text{PL}} c_5 \varepsilon^2}{\mu_1},$$

where we can select ε , $d(\hat{\Phi}, \Phi)$, and N according to

$$\varepsilon \leq \sqrt{\frac{\mu_1(\bar{J} - J_\star^\gamma)}{48\mu_{\text{PL}} c_5}}, \quad d(\hat{\Phi}, \Phi) \leq \sqrt{\frac{\mu_1(\bar{J} - J_\star^\gamma)}{48\mu_{\text{PL}} c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2}}, \quad N \geq \frac{8\mu_{\text{PL}}}{\eta \mu_1} \log \left(\frac{3(J^\gamma(\bar{\theta}_0) - J_\star^\gamma)}{\bar{J} - J_\star^\gamma} \right), \quad (27)$$

to obtain $J^\gamma(\bar{\theta}_N) = J^\gamma(\theta_{j+1}) \leq \bar{J}$. Therefore, given that $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$, for any iteration j of Algorithm 1, and supposing that we select τ and n_c according to Lemma F.4 to ensure $\left| \hat{J}^{\gamma_j, \tau}(\theta_{j+1}, \hat{\Phi}) - J^{\gamma_j}(\theta_{j+1}) \right| \leq \frac{1}{2} J^{\gamma_j}(\theta_{j+1})$, we obtain

$$\begin{aligned}
\alpha_j &= \frac{3\sigma_{\min}(\hat{\Phi}^\top Q \hat{\Phi} + \theta^\top R \theta)}{\frac{4}{3} \hat{J}^{\gamma, \tau}(\theta, \hat{\Phi}) - 3\sigma_{\min}(\hat{\Phi}^\top Q \hat{\Phi} + \theta^\top R \theta)} \geq \frac{3\sigma_{\min}(\hat{\Phi}^\top Q \hat{\Phi})}{\frac{4}{3} \hat{J}^{\gamma, \tau}(\theta, \hat{\Phi}) - 3\sigma_{\min}(\hat{\Phi}^\top Q \hat{\Phi})} \geq \frac{3\sigma_{\min}(Q)}{\frac{4}{3} \hat{J}^{\gamma, \tau}(\theta, \hat{\Phi}) - 3\sigma_{\min}(Q)} \\
&\geq \frac{3\sigma_{\min}(Q)}{2\bar{J} - 3\sigma_{\min}(Q)} := \underline{\alpha},
\end{aligned}$$

where we can this lower bound on α_j to unroll the discount factor update over M iterations of Algorithm 1 and obtain

$$\gamma_M = \gamma_0 \prod_{j=0}^{M-1} (1 + \xi \alpha_j) \geq \gamma_0 \prod_{j=0}^{M-1} (1 + \xi \underline{\alpha}) = \gamma_0 (1 + \xi \underline{\alpha})^M,$$

which implies that Algorithm 1 finds a stabilizing controller $\theta_M \in \mathcal{S}_\theta^1$ (i.e., $\gamma_M = 1$), within $\frac{\log(1/\gamma_0)}{\log(1+\xi\alpha)}$ iterations. Moreover, (26) implies that $\sqrt{(1+\alpha_j)\gamma_j}\rho(A_u + B_u\theta_{j+1}) < 1$, which yields

$$\begin{aligned}\sqrt{(1+\xi\alpha_j)\gamma_j}\rho(A_u + B_u\theta_{j+1}) &= \frac{\sqrt{(1+\xi\alpha_j)\gamma_j}}{\sqrt{(1+\alpha_j)\gamma_j}} \sqrt{(1+\alpha_j)\gamma_j}\rho(A_u + B_u\theta_{j+1}) < \frac{\sqrt{(1+\xi\alpha_j)\gamma_j}}{\sqrt{(1+\alpha_j)\gamma_j}} \\ &< \sqrt{1 - \frac{3(1-\xi)\sigma_{\min}(Q)}{2\bar{J}}},\end{aligned}$$

and it guarantees that after M iterations of the discounted LQR method, it returns a low-dimensional stabilizing controller $\theta \in \mathcal{S}_\theta^1$ with $\rho(A_u + B_u\theta) < \bar{\lambda}_\theta := \sqrt{1 - \frac{3(1-\xi)\sigma_{\min}(Q)}{2\bar{J}}}$. We complete our analysis showing that as long as $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$ and $|\hat{J}^{\gamma_j, \tau}(\theta_{j+1}, \hat{\Phi}) - J^{\gamma_j}(\theta_{j+1})| \leq \frac{1}{2}J^{\gamma_j}(\theta_{j+1})$ hold for the j -th iteration of Algorithm 1, then they also hold for the subsequent iterations, with high probability. This guarantees that $\alpha_j \geq \underline{\alpha}$ holds for every iteration, and thus $\rho(A_u + B_u\theta_{j+1}) < \bar{\lambda}_\theta$.

Lemma H.2. *Suppose that $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$ and $|\hat{J}^{\gamma_j, \tau}(\theta_{j+1}, \hat{\Phi}) - J^{\gamma_j}(\theta_{j+1})| \leq \frac{1}{2}J^{\gamma_j}(\theta_{j+1})$. Then, it holds that*

$$\alpha_j \leq \underline{\alpha}, \quad J^{\gamma_{j+1}}(\theta_{j+1}) \leq \frac{2\bar{J}^2}{3(1-\xi)\sigma_{\min}(Q)}.$$

Proof. The proof is similar to [Zhao et al., 2024, Lemma 7] with our definitions of $\underline{\alpha}$ and $\bar{\lambda}_\theta$. \square

Suppose that $\bar{J} > 2J_\star^1$. Then, by the definition we have that $J_\star^{\gamma_0} \leq J^{\gamma_j}(\theta_j)$ which implies that $\bar{J} - J_\star^{\gamma_j} \geq 2J_\star^1 - J_\star^1 = J_\star^1$. Therefore, according to (27) we know that $J^{\gamma_j}(\theta_{j+1}) \leq \bar{J}$ if

$$N \geq \frac{8\mu_{\text{PL}}}{\eta\mu_1} \log \left(\frac{2\bar{J}^2}{(1-\xi)\sigma_{\min}(Q)J_\star^1} \right), \quad \epsilon \leq \sqrt{\frac{\mu_1 J_\star^1}{48\mu_{\text{PL}}c_5}}, \quad d(\hat{\Phi}, \Phi) \leq \sqrt{\frac{\mu_1 J_\star^1}{48\mu_{\text{PL}}c_4l \left((L\nu_\theta + \phi)\sqrt{2\bar{\ell}} + \phi \right)^2}},$$

with probability $1 - (\delta + c_1N(\ell^{-\zeta} + n_s^{-\zeta} - n_se^{-\ell/8} - e^{-c_2n_s}))$. The proof is completed by union bounding over all iterations j of Algorithm 1.

H.1 Lifting the Controller

Given θ that stabilizes (A_u, B_u) , we now demonstrate that $\rho(A + B\theta\hat{\Phi}^\top) < 1$. To do so, we write

$$A + B\theta\hat{\Phi}^\top = \Omega \left(\Omega^\top A \Omega + \Omega^\top B\theta\hat{\Phi}^\top \Omega \right) \Omega^\top = \Omega \left(\begin{bmatrix} A_u + B_u\theta\hat{\Phi}^\top \Phi & B_u\theta\hat{\Phi}^\top \Phi_\perp \\ \Delta + B_s\theta\hat{\Phi}^\top \Phi & A_s + B_s\theta\hat{\Phi}^\top \Phi_\perp \end{bmatrix} \right) \Omega^\top,$$

and leverage Lemma C.2 to obtain

$$\begin{aligned}\rho(A + B\theta\hat{\Phi}) &\leq \max \left\{ \rho(A_u + B_u\theta\hat{\Phi}^\top \Phi), \rho(A_s + B_s\theta\hat{\Phi}^\top \Phi_\perp) \right\} + C_{\text{gap}} \|B_u\theta\hat{\Phi}^\top \Phi_\perp\| \|\Delta + B_s\theta\hat{\Phi}^\top \Phi\| \\ &\leq \max \left\{ \rho(A_u + B_u\theta\hat{\Phi}^\top \Phi), \rho(A_s + B_s\theta\hat{\Phi}^\top \Phi_\perp) \right\} + C_{\text{gap}} \|B\| \nu_\theta (\|A\| + \|B\| \nu_\theta) d(\hat{\Phi}, \Phi).\end{aligned}\tag{28}$$

Observe that the second term in the above expression is in the order of the subspace distance. Therefore, we can make it sufficiently small to guarantee that the spectral radius of the closed-loop matrix is less than one. That is a benefit of learning to stabilize on the left unstable subspace instead of the right unstable subspace of A . For instance, if the columns of Φ formed the basis of the right unstable subspace of A , the decomposition above would lead to an error term that scales as $\mathcal{O}(\|\Delta\| + d(\hat{\Phi}, \Phi))$, where $\|\Delta\|$ is only sufficiently small for A “almost symmetric” (i.e., where A is easily decomposable into the stable and unstable modes). We now proceed to control the spectral radius: $\rho(A_u + B_u \theta \hat{\Phi}^\top \Phi)$ and $\rho(A_s + B_s \theta \hat{\Phi}^\top \Phi_\perp)$.

$$\begin{aligned} \rho(A_u + B_u \theta \hat{\Phi}^\top \Phi) &= \rho(A_u + B_u \theta + B_u \theta (\hat{\Phi}^\top \Phi - I)) \\ &\stackrel{(i)}{\leq} \rho(A_u + B_u \theta) + \max \left\{ \|B_u \theta (\hat{\Phi}^\top \Phi - I)\| C_{\text{bf},1}, \left(\|B_u \theta (\hat{\Phi}^\top \Phi - I)\| C_{\text{bf}} \right)^{1/\ell} \right\} \\ &\stackrel{(ii)}{\leq} \rho(A_u + B_u \theta) + (\|B\| \nu_\theta C_{\text{bf},1})^{1/\ell} d(\hat{\Phi}, \Phi)^{1/\ell} \leq \bar{\lambda}_\theta + (\|B\| \nu_\theta C_{\text{bf},1})^{1/\ell} d(\hat{\Phi}, \Phi)^{1/\ell}, \end{aligned} \quad (29)$$

where $\bar{\lambda}_\theta := \rho(A_u + B_u \theta)$. (i) follows from Lemma C.2 with $C_{\text{bf},1}$ being a constant that depends on the Schur decomposition of $A_u + B_u \theta$. (ii) is due to Lemma 2.1 and $d(\hat{\Phi}, \Phi) \leq \frac{1}{\|B\| \nu_\theta C_{\text{bf},1}}$.

Similarly, we can write

$$\rho(A_s + B_s \theta \hat{\Phi}^\top \Phi_\perp) \leq |\lambda_{\ell+1}| + (\|B\| \nu_\theta C_{\text{bf},2})^{1/\ell} d(\hat{\Phi}, \Phi)^{1/\ell}, \quad (30)$$

where $C_{\text{bf},2}$ depends on the Schur decomposition of A_s . In addition, we require the subspace distance to satisfy $d(\hat{\Phi}, \Phi) \leq \frac{1}{\|B\| \nu_\theta C_{\text{bf},2}}$. By combining (29) and (30) in (28) to obtain

$$\rho(A + B \theta \hat{\Phi}^\top) \leq \max\{\bar{\lambda}_\theta, \lambda_{\ell+1}\} + \left(C_{\text{gap}} \|B\| \nu_\theta (\|A\| + \|B\| \nu_\theta) + (\|B\| \nu_\theta)^{1/\ell} \left(C_{\text{bf},1}^{1/\ell} + C_{\text{bf},2}^{1/\ell} \right) \right) d(\hat{\Phi}, \Phi)^{1/\ell},$$

which implies that

$$d(\hat{\Phi}, \Phi) < \frac{(1 - \max\{\bar{\lambda}_\theta, |\lambda_{\ell+1}|\})^\ell}{\left(C_{\text{gap}} \|B\| \nu_\theta (\|A\| + \|B\| \nu_\theta) + (\|B\| \nu_\theta)^{1/\ell} \left(C_{\text{bf},1}^{1/\ell} + C_{\text{bf},2}^{1/\ell} \right) \right)^\ell},$$

to guarantee that $\rho(A + B \theta \hat{\Phi}^\top) < 1$.

Theorem H.1 (Main Result). *Given positive scalars $\delta_\tau \in (0, 1)$, $\delta_\sigma \in (0, 1)$ and $\zeta > 0$. Suppose that the problem parameters are selected as follows:*

- Gradient and cost estimation parameters:

$$\begin{aligned} n_s &= \mathcal{O}(\ell \zeta^4 \log^6 \ell), \quad n_c = \mathcal{O}(\log(1/\delta_\tau)), \quad \varepsilon' := \sqrt{\frac{J_\star^1}{\mu_{PL}(d_U(\ell \log^2 \ell))}}, \\ r &= \mathcal{O}(\sqrt{\varepsilon'}), \quad \text{and } \tau = \mathcal{O}(\log(1/\varepsilon') + \tau_0). \end{aligned}$$

- Subspace distance: $d(\widehat{\Phi}, \Phi) \leq \varepsilon_{dist}$ with

$$\varepsilon_{dist} := \min \left\{ \frac{(1 - \max\{\bar{\lambda}_\theta, |\lambda_{\ell+1}|\})^\ell}{C_{dist,1}}, \sqrt{\frac{J_\star^1}{C_{dist,2}}}, \frac{1}{\|B\|\nu_\theta \max\{C_{bf,1}, C_{bf,1}\}}, \frac{J_\star^1}{4\ell\sqrt{I}C_{cost}}, \frac{\sigma_{\min}(\widehat{\Phi}^\top Q\widehat{\Phi})}{4\|Q\|\sqrt{2\ell}\kappa(\widehat{\Phi}^\top Q\widehat{\Phi})} \right\},$$

and $C_{dist,1} = \text{poly}(\|A\|, \|B\|, \nu_\theta)$ and $C_{dist,2} = \text{poly}(\nu_\theta, L, \mu_{PL}, \phi, \ell, d_U)$.

- Time horizon:

$$T = \mathcal{O} \left(\log \left(\frac{\ell^7 (d_X - \ell) \mu_0}{(1 - |\lambda_{\ell+1}|) \varepsilon_{dist} \delta_\sigma^3} \right) / \log(|\lambda_\ell|) \right).$$

- Algorithm 1 parameters:

$$N \geq \frac{32\mu_{PL}}{\eta} \log \left(\frac{2\bar{J}^2}{(1 - \xi)\sigma_{\min}(Q)J_\star^1} \right), \quad \eta = \mathcal{O}(1/(d_U \ell \log^2 \ell)), \quad \gamma_0 \leq 1/\rho^2(A), \quad \text{and } \xi \in (0, 1),$$

with $\bar{J} := \max\{2J_\star^1, J^{\gamma_0}(0)\}$. Therefore, within $M \geq \frac{\log(1/\gamma_0)}{\log(1+\xi\alpha)}$ iterations, Algorithm 1 returns $K = \theta_M \widehat{\Phi}^\top \in \mathcal{K}$, with probability $1 - \bar{\delta}$, where $\bar{\delta} := \delta_\sigma + M(\delta_\tau + \bar{c}_1 N(\ell^{-\zeta} + n_s^{-\zeta} - n_s e^{-\ell/8} - e^{-\bar{c}_2 n_s}))$.

H.2 Policy Gradient Per-iteration Stability Analysis

Given a discount factor $\gamma \in (0, 1]$, we proceed to demonstrate that $\bar{\theta}_n \in \mathcal{S}_\theta^\gamma$, for any policy gradient update $n \in \{0, 1, \dots, N-1\}$ of Algorithm 1 (i.e., line 6). As discussed previously, by carefully incrementing γ , the low-dimensional control gain $\theta_+ = \bar{\theta}_N$ is stabilizing for the underlying low-dimensional damped system (4). Therefore, it remains to show that for any $\bar{\theta}_0 \in \mathcal{S}_\theta^\gamma$, $\bar{\theta}_n$ stays within $\mathcal{S}_\theta^\gamma$. To do so, we can first show that $\bar{\theta}_1$ does not leave $\mathcal{S}_\theta^\gamma$, with high probability, as long as the problem parameters are set accordingly. Finally, we use an induction step to extend this conclusion for any iteration $n \in \{0, 1, \dots, N-1\}$.

As previously, we use Lemma 2.1 to write

$$J^\gamma(\bar{\theta}_1) - J_\star^\gamma \leq \left(1 - \frac{\eta\mu_1}{8\mu_{PL}}\right) (J^\gamma(\bar{\theta}_0) - J_\star^\gamma) + 2\eta c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\widehat{\Phi}, \Phi)^2 + 2\eta c_5 \varepsilon^2,$$

where the step-size is set according to $\eta \leq \min \left\{ \frac{\mu_1}{16\mu_2 L_\theta}, \frac{c_4}{2L_\theta \mu_2}, \frac{c_5}{L_\theta} \right\}$. Therefore, as long as the subspace distance $d(\widehat{\Phi}, \Phi)$, time horizon τ and smoothing radius r are set according to

$$d(\widehat{\Phi}, \Phi) \leq \frac{\sqrt{\mu_1}}{8 \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right) \sqrt{c_4 \ell \mu_{PL}}}, \quad \tau = \mathcal{O}(\log(1/\varepsilon)), \quad \text{and } r = \mathcal{O}(\sqrt{\varepsilon}), \quad \text{respectively,}$$

with $\varepsilon \leq \frac{1}{8} \sqrt{\mu_1 / c_5 \mu_{PL}}$. Hence, we obtain $J^\gamma(\bar{\theta}_1) - J_\star^\gamma \leq \left(1 - \frac{\eta\mu_1}{16\mu_{PL}}\right) (J^\gamma(\bar{\theta}_0) - J_\star^\gamma)$, which implies that $\bar{\theta}_1$ is stabilizing for the underlying damped system, i.e., $\bar{\theta}_1 \in \mathcal{S}_\theta^\gamma$. Let the base case and inductive hypothesis be defined as follows:

Base case: $J^\gamma(\bar{\theta}_1) - J_\star^\gamma \leq J^\gamma(\bar{\theta}_0) - J_\star^\gamma$,

Inductive hypothesis: $J^\gamma(\bar{\theta}_n) - J_\star^\gamma \leq J^\gamma(\bar{\theta}_0) - J_\star^\gamma$,

which combined with the aforementioned conditions on the problem parameters yields

$$\begin{aligned} J^\gamma(\bar{\theta}_{n+1}) - J_\star^\gamma &\leq \left(1 - \frac{\eta\mu_1}{8\mu_{\text{PL}}}\right) (J^\gamma(\bar{\theta}_n) - J_\star^\gamma) + 2\eta c_4 \ell \left((L_K \nu_\theta + \phi) \sqrt{2\ell} + \phi \right)^2 d(\hat{\Phi}, \Phi)^2 + 2\eta c_5 \varepsilon^2 \\ &\leq \left(1 - \frac{\eta\mu_1}{16\mu_{\text{PL}}}\right) (J^\gamma(\bar{\theta}_n) - J_\star^\gamma) \leq \left(1 - \frac{\eta\mu_1}{16\mu_{\text{PL}}}\right) (J^\gamma(\bar{\theta}_0) - J_\star^\gamma) \leq J^\gamma(\bar{\theta}_0) - J_\star^\gamma. \end{aligned}$$

H.3 Sample Complexity Reduction

We proceed to characterize the sample complexity of Algorithm 1 and the benefit of learning to stabilize on the unstable subspace. We quantify the sample complexity by the number of data samples x_t we query from the system (1) and its adjoint. Namely, $\mathcal{S}_c := \mathcal{S}_c^1 + \mathcal{S}_c^2$, where $\mathcal{S}_c^1 := M(n_c + n_s N)\tau$ includes the samples used discounted LQR method to learn a low-dimensional control gain that stabilizes the unstable dynamics, and $\mathcal{S}_c^2 := T + d_X$ corresponds to the number of data points needed for estimating the left unstable subspace of A . We emphasize that the extra d_X term comes from sampling data from the adjoint system through element-wise computations via the adjoint operator, as discussed previously in Section 3.

Corollary H.1. *Let the arguments of Theorem H.1 hold. Then, Algorithm 1 returns a stabilizing controller for the original system (1) with*

$$\mathcal{S}_c = \log(\rho(A)) \tilde{\mathcal{O}}(\ell^2 d_U) C_{sc,1} + \mathcal{O} \left(\log \left(\frac{\ell^7 (d_X - \ell) C_{sc,2}}{(1 - |\lambda_{\ell+1}|) (1 - \max\{\bar{\lambda}_\theta, |\lambda_{\ell+1}|\})^\ell} \right) \right) + \mathcal{O}(d_X),$$

where $C_{sc,1} = \text{poly}(\|A\|, \|B\|, \|Q\|, \mu_{PL})$ and $C_{sc,2} = \text{poly}(\|A\|, \|B\|, \nu_\theta, L, \mu_{PL}, \phi, \ell, d_U, 1/J_\star^1, 1/\delta_\sigma)$.

Note that the sample complexity is dominated by $\tilde{\mathcal{O}}(\ell^2 d_U) + \mathcal{O}(d_X)$ which scales much slower than $\tilde{\mathcal{O}}(d_X^2 d_U)$ for the setting where the number of unstable modes is much smaller than the number of states of the system, i.e., our setting of interest with $\ell \ll d_X$.